博士論文

Image Narrative Generation via
Interactive Visual Question Generation and Answering
（インタラクティブな画像質問生成・応答による
画像物語文の生成）

シン アンドリュー

Image Narrative Generation via Interactive Visual Question Generation and Answering

(インタラクティブな画像質問生成・応答による画像物語文の生成)

by

Andrew Yongsup Shin

Submitted to the Graduate School of Information Science and Technology,
The University of Tokyo
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Information Science and Technology

at the

THE UNIVERSITY OF TOKYO

September 2017

Author . . . . . .
Graduate School of Information Science and Technology, The University of Tokyo
August 14, 2017

Certified by . . . . . .
Tatsuya Harada
Professor
Thesis Supervisor

This thesis is dedicated to my family whose support has been unwavering.

Acknowledgments

I would first like to express my sincere gratitude towards Professor Tatsuya Harada. Not only have I been a beneficiary of his excellence in directing research, but his ethics and attitude towards education and work have inspired me just as much, and will remain an invaluable asset to me for a very long time, perhaps even more so than technical expertise.

Lecturer Yoshitaka Ushiku has been a friendly and readily available advisor on a variety of matters including, but not limited to, technical ones, and I have always felt lucky to have someone with whom I could converse with such ease, while being constructive at the same time.

I would also like to thank Professor Yasuo Kuniyoshi, Professor Kiyoharu Aizawa, and Professor Michitaka Hirose for their acute comments and advices throughout the dissertation of this thesis. Their insights have further solidified the overall integrity of this work.

I would not have been able to survive the plethora of paperworks and procedures for past 3 years, had it not been the help of Ms. Aoi Kaneko. Occasional conversations with her over a wide range of non-technical matters was a sincere pleasure. Periodical talks with Assistant Professor Mamoru Nakamura have also been a worthwhile pleasure.

I feel very much indebted to Dr. Asako Kanezaki's help. Having come from a different research background, I spent my early days at the laboratory struggling to familiarize myself with both the material and the environment. It would have taken me much longer without her help.

My colleagues at this laboratory, past and present, have been an incessant source of information, intimacy, and inspiration. While I am grateful to all of them for being there, I would like to name some of them with whom I have worked together on research projects; Yusuke Mori, Atsushi Kanehira, Hiroharu Kato, Katsunori Ohnishi, Masataka Yamaguchi, Kuniaki Saito, and Yuichiro Kikura. I would also like to thank Yusuke Mukuta and Masatoshi Hidaka who have spent the longest span of time with me in this laboratory, as well as working together as teaching assistants. Kaikai Huang, Haowei Yeh, Tejero de Pablos Antonio, and many other international students have exposed me to cultural diversity, as well as simply being fun to be around. I am just as well grateful to those whom I have not been able to name individually here.

My final gratitude naturally goes to my family. Not a single page of this thesis would have been possible without their support and encouragement.

Image Narrative Generation via Interactive Visual Question Generation and Answering
(インタラクティブな画像質問生成・応答による画像物語文の生成)

by
Andrew Yongsup Shin

Submitted to the Graduate School of Information Science and Technology, The University of Tokyo
on August 14, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Information Science and Technology

Abstract

Image captioning task has enticed an unprecedented amount of attention with the advent of deep learning techniques. However, its objective has been limited to the generation of factual description of the global event in the image. Yet, humans can provide far richer contents than a factual description, including sentiments, inferences, etc. Furthermore, such contents will widely vary depending on the narrator. We examine a novel task of image narrative generation in which we attempt to overcome the limitations of image captioning task. While the sole primary objective of image captioning task is the generation of factual description of the image, image narrative is not restricted by such rigid objective, and may discuss any aspect of the image as long as it can relate to the image, including, but not limited to, further details, sentiments, or inferences about the image. It may even creatively assign story-like characteristics to the image. If such is possible, the level of interaction between vision and language will be elevated to a more eloquent stage with stronger resemblance to human linguistic capability. In addition, we examine the task in an interactive setting, so that each user's distinct preference can be learned and applied to image narrative generation. In this paper, we propose a series of models to examine each prerequisite required to implement image narrative generation. First, in order to capture local details that are difficult to obtain through generic CNN features from the entire image, we employ spatial pyramid and vector of locally aggregated descriptors (VLAD) coding of convolutional neural network (CNN) features. Then, we build a weakly-supervised sentiment dataset, from which we fine-tune a separate neural network that outputs sentiment features, to capture the overall sentiment of the image. Unlike factual description, which can be obtained directly by looking at the image, the elements of image narrative cannot be obtained straightforwardly, and require a reasoning process, in which it first has to decide what to discuss by asking questions first, and subsequently find answers to it, similarly to how humans might perform such task. We exploit visual question generation (VQG) and answering (VQA) techniques to implement such process. Finally, we engage the users with the learning and generation process, thereby providing an interactive model to learn and reflect each user's preference. We experimentally demonstrate that our proposed model for

image narrative generation can generate a highly expressive image description with much wider range of topic contents, which turns out to be difficult to realize via conventional models.

Thesis Supervisor: Tatsuya Harada
Title: Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we will briefly review the historical context of the current landscape of artificial intelligence (AI), and how such context has enabled the advancements in image captioning task in particular. We will then briefly discuss the limitations of current image captioning task, which lays out the foundation for the objective of this thesis, namely image narrative generation. The rest of the chapter will describe how this thesis is organized.

1.1 Background

Emerging after several decades of "AI winter," deep learning techniques have revolutionized not only the field of artificial intelligence, but also a wide array of potential applications, and ever since have attracted an unprecedented amount of attention and exuberance worldwide. Lying at the center of deep learning is the concept known as convolutional neural networks (CNN). While the technique has been around in theory for a few decades, it has only recently been brought into practice with exponentially growing computing power and the advent of internet, from which an unlimited amount of data became freely available. CNN consists of multiple layers, each of which performs a convolution or pooling of given filter size. Over multiple stages of iteration enabled by backpropagation, the parameters "*learn*" to distinguish between different classes of images. As such, image classification task became the first beneficiary of deep learning techniques.

The core benefit of CNN, however, is that it demonstrates high portability to other tasks as well. Image captioning task has proven to be one of many such tasks where CNN features boost the performance. Another deep learning technique that plays a central role in image captioning task is recurrent neural networks (RNN), particularly long short-term memory (LSTM) [34]. Conditioned on image features from CNN and previously generated words, each LSTM unit learns to generate the next likely word, so that multiple stacks of LSTM units generate a complete sentence describing the image. The combination of CNN features and LSTM units has become a de-facto standard for image captioning task, and nearly every recent work on the task has been a variation of such pipeline.

Image captioning task, while highly successful in its own objective, inevitably poses new challenges for learning of vision and language. First, it is mostly limited to the events occurring at the global scale of the image. However, humans may frequently pay attention

to objects or events occurring at the local or secondary scale of the image. Representing and reflecting local elements of the image can enrich the image description both qualitatively and quantitatively.

Second, image captioning has primarily dealt with factual, objective components of the image. Humans on the other hand are able to perceive beyond objective components from the image (*e.g.*, sentiments of the image or inferences about the image), and linguistically express it. In other words, humans can see and express *more than meets the eye*.

Third, conventional image captioning task has assumed that there exists a single correct "gold" description that is applicable to anyone. While this is true for factual description of global event, different people may pay attention to different parts of the image, and yield different interpretations. Such diversity in perspectives cannot be derived with conventional image captioning task.

In this thesis, we introduce a novel task of image narrative generation, in which we attempt to overcome the limitations discussed above. Each limitation necessitates a design of separate module, and these modules are eventually combined for interactive image narrative generation task.

Primary contributions of the thesis can be summarized as following:

- Proposal of a model to represent images that better account for local elements of the image
- Proposal of a model to represent images that reflect subjective elements of the image, particularly sentiments present in the image
- Utilization of visual question generation (VQG) and visual question answering (VQA) techniques to deal with inferential elements, and
- Examination of interactive environment to learn and reflect the user's preference into image narratives.

1.2 Objective

In this thesis, we develop an image narrative generation model, which generates not only a single sentence factual description of the image, but an image narrative whose contents encompass both factual and non-factual elements at global and local scales of the image, enriching the image descriptions both in quality and quantity. Furthermore, we develop such model in an interactive way, namely via Q&A module, which enables us to derive image narratives with diversity, and further reflect the user's preference into new images by customizing. Figure 1-1 describes the overview of the system proposed and developed throughout this thesis.

1.3 Structure of the Thesis

This thesis is organized as shown in Figure 1-2. We have just walked through a high-level introduction of this thesis in Chapter 1. In Chapter 2, we will describe and define our

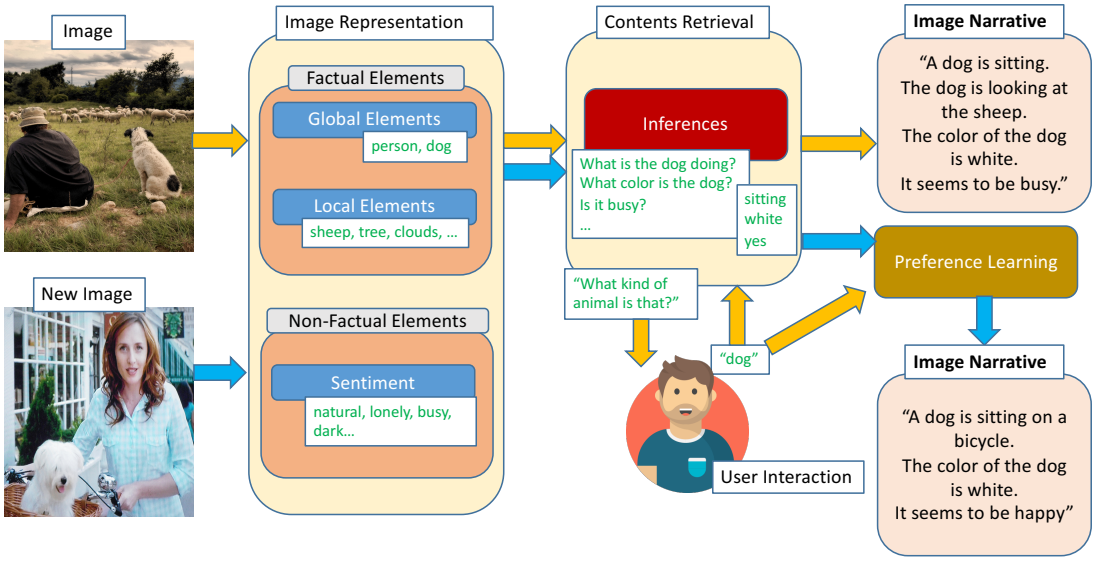


Figure 1-1: System overview of the model developed in the thesis. Yellow arrows correspond to the workflow of interactive image narrative generation for preference learning, and blue arrows correspond to the workflow of automatic image narrative generation reflecting user preference based on the preference learning module, without user interaction.

task more in details, especially in comparison to conventional image captioning task. We discuss some of the limitations in the current research on the learning of vision and language, and specify our objectives thereon, along with necessary modules to realize those objectives. Chapter 3 deals with more specific background, from the classic works of each field and recent related works, which highly correlate to the core methods, concepts, and algorithms employed in this thesis. Chapter 4 deals with our attempt at better reflecting the local elements of the image. Specifically, we apply vector of locally aggregated descriptors (VLAD) coding to convolutional neural network (CNN) features extracted from the regions proposed by selective search. In addition, such coding is performed on multiple grids using spatial pyramid. In Chapter 5, we describe our attempt to deal with another key challenge, namely learning of non-factual concepts, particularly sentiment of the image. We build a weakly-supervised dataset, on top of which we fine-tune a separate convolutional neural network to extract the "sentiment" features from the images, with the aid of simple multi-label learning. In Chapter 6, we mainly discuss our proposed method for visual question generation (VQG) and answering (VQA). In particular, our proposed method for VQA enabled us to win the 1st place in the international challenge of the corresponding task. VQG and VQA tasks have high relevance to our objective, since they go a step further beyond the conventional image captioning task by enabling the AI to both raise questions about the image, and to answer those questions in natural language. Each of Chapter 4, 5, and 6 functions as a distinct module of the primary task to be examined in the later chapters of this thesis. In Chapter 7, we introduce a novel task of single image narrative generation, in which we attempt to generate multiple-sentences captions from single image that consist of both visual and non-visual elements. We accomplish the goal by serializing a number of modules, namely region extraction, image captioning, visual question answering, and

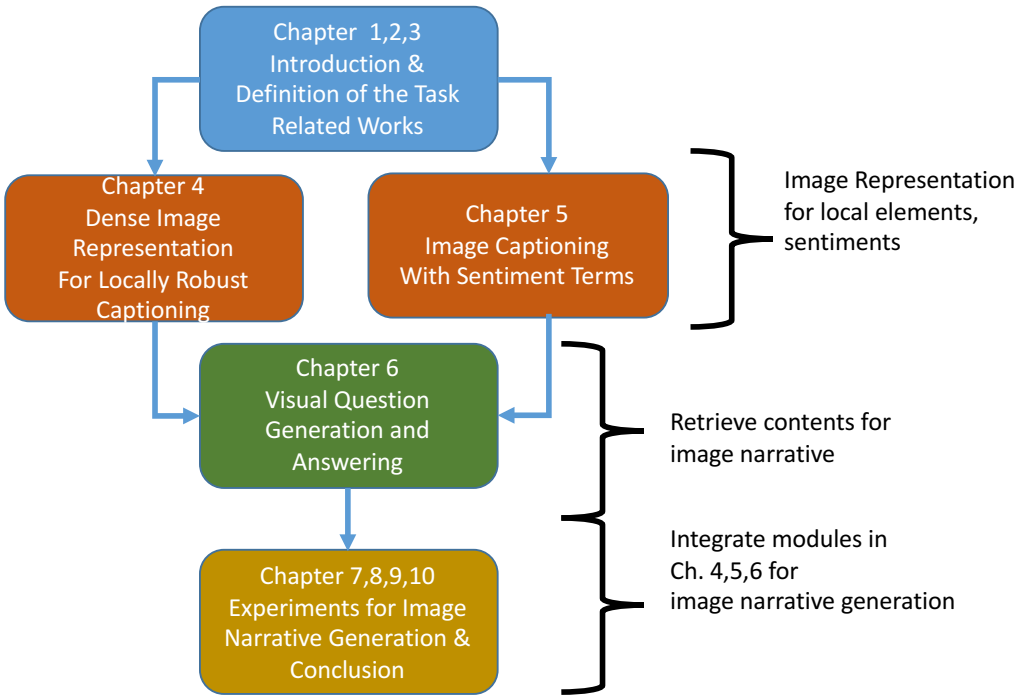


Figure 1-2: Structure of this thesis

simple natural language processing techniques. In particular, individual modules described through Chapter 4 to 6 will be brought into effect. Chapter 8 expands upon single image narrative generation task to involve user interaction. Visual questions that allow for multiple responses from the users are generated via a novel workflow utilizing VQG and VQA. We show that, by collecting and training with user responses from such questions, we can learn the user's interest and apply it to unseen images to generate customized image descriptions. Chapter 9 deals with the setting and results of the experiments for the primary task of this thesis. We will evaluate the results both qualitatively and quantitatively using a variety of evaluation metrics. Finally, we will conclude our thesis and discuss unsolved problems and future work in Chapter 10.

Chapter 2

Defining Image Narrative Generation

We briefly introduced the task of our interest in Chapter 1, but only at an abstract level. In this chapter, we define the task more formally, differentiate it from previous task in multiple aspects, and discuss its objective and key challenges in details.

2.1 Elements of Image Narrative

We first discuss the elements of image narrative, and how they can be categorized. While there can be various ways of categorizing the elements, we focus here on the possibility of varying interpretations. The contents of image narrative should inevitably be derived from, or relate to, the image to which they are assigned. In this sense, an objective, straightforward description of the visual contents of the image qualify as the contents of image narrative. For instance, a simple sentence in a form of subject-verb-object may have to recognize and depict the main object (subject), action (verb), and secondary object (object) occurring in the image. This type of objective contents whose interpretation is unambiguous can be categorized into factual elements. In addition to global/local objects and action taking place, it is also frequently possible to recognize and describe the setting of the image, including place and time, to a fairly unambiguous degree. We may also approach the factual elements from the 5W1H perspective. Each component of 5W1H, with the exception of '*why*' corresponds to one of object, action, or setting, which constitute the factual elements.

On the other hand, there also exist elements that permit a space for varying interpretations, that cannot be visually verified in a straightforward manner, while clearly relating to the image. From the 5W1H perspective discussed above, a question *why* would be an example of such element, as it requires an inferential process beyond simple visual recognition. Sentiments of the image would be another such element, as it is subject to a relatively wide range of interpretations and can frequently be ambiguous. Prediction of what would happen after the event occurring in the image is another example, as it is also open to a wide array of possibilities. We categorize the elements with these characteristics, namely the potential for varying interpretations, as non-factual elements. Figure 2-1 shows the examples for both factual and non-factual elements that together constitute an image narrative.

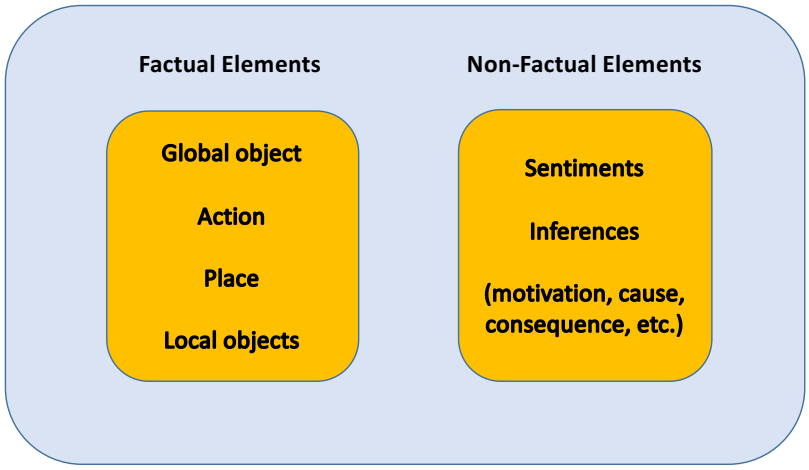


Figure 2-1: The factual and non-factual elements of narrative.

	Factual	Non-Factual
Global	Captioning	Visual Storytelling
Local	Dense Captioning	--

(a) Spectrum of conventional image description tasks.

	Factual	Non-Factual
Global	Image Narrative	
Local		

(b) Specrtrum of image narrative generation.

Figure 2-2: Spectrum of various image description tasks

2.2 Comparison to Previous Tasks

It is necessary to draw clear distinctions between conventional image captioning task and image narrative generation task, which is our primary concern in this thesis.

Image captioning task aims to describe the objective, factual components of the given image. In particular, it is mostly concerned with the primary event occurring at the global scale of the image. Thus, an image caption describing a local, secondary event of minor importance in the image may be considered inappropriate in conventional image captioning task, regardless of its correctness. An image caption with non-factual element that is open to different interpretations and cannot be visually verified in a straightforward manner is also considered inappropriate for conventional image captioning task. Image narrative on the other hand is not restricted by such rigid objective, and may discuss any aspect of the image as long as it can relate to the image. For example, a typical caption may say "an elephant is standing on a field." On the other hand, image narrative may choose to

Table 2.1: Comparing image captioning task and image narrative generation task.

	Image Captioning	**Image Narrative Generation**
Length	mostly single sentence	multiple sentences
Contents	factual description of main event (objects, actions)	any image-relevant aspect (sentiments, inferences, etc.)
Primary Techniques	CNN, LSTM	CNN, LSTM, VQA, VQG, region extraction

discuss sentiments such as how the elephant might be feeling, or inferences such as why the elephant is in such status, or even creatively assign a story-like characteristics such as the name of the elephant, etc. While image narrative may also discuss factual description as in image captions, it is not restricted by it. In other words, image caption in its conventional sense is one of many components that constitute an image narrative.

In addition, image captioning task mostly produces a single sentence per image. This is closely related to the task objective described above. Since the primary objective is to describe the main event in the image, multiple sentences are likely redundant. Some works attempt to produce multiple sentences of image captions. For example, DenseCap [43] produces image caption for each region in the image, and Krause et al. [50] attempt to generate a paragraph of image description. However, their purpose is still restricted by the objective of describing factual components of the main event in the image. For example, local image caption generated by DenseCap [43] for each region in the image will describe the factual event or object in the corresponding region.

Table 2.1 outlines some of the key differences between image captioning task and image narrative generation task. Table 2.2 shows examples of image captions and image narratives for the same images.

As discussed in Section 2.1, factual elements and non-factual elements are elements of a narrative. In light of the relative importance and saliency of the elements, we can also group these elements into global and local elements. With these two criteria, we can further clarify and define the image narrative generation task, especially in comparison to previous tasks concerning image description (Figure 2-2). An image description that describes the factual and global elements of an image would correspond to an image captioning task in its conventionally understood sense. As was discussed in Section 2.2, this is inherently so by definition, as the primary objective of image captioning task is to generate a factual description of the main event occurring in the image.

There have also been attempts to describe the factual elements of the image that are local or secondary, not limited to the primary event of the image, most notable work being [43]. Such works on *dense captioning* enable the AI to have a more complete understanding of the image. However, local descriptions in previous works are mostly limited to naming the object in a phrase or sentence. As in image captioning task, this is also inherently so since its goal is to describe the factual elements of local elements. Note that dense captioning task involves the global elements of the image as well, but its novelty and focus are on local elements, and we categorize it in this thesis as an image description containing local and factual elements for convenience.

Some of the previous works have tackled the task of generating non-factual description

of the images. Most of them were carried out in a form of assigning a story-like description to the primary event in the image; that is, to the global elements of the image. This task of generating non-factual, global description of the image has had less rigid boundary in terms of format and contents. For example, [37] presents a dataset with story-like captions assigned to the images, but requires that the images form a sequence. Thus, a single image and its corresponding caption do not constitute a complete story themselves, but a part of it.

Each task described so far occupies a portion of the spectrum of image description, as shown in Figure 2-2 (a). On the other hand, Image narrative is posited to involve each of the elements tackled by the previous image description tasks. That is, it attempts to account for both factual and non-factual elements of the image at both global and local scales of the image. Thus, an image narrative in principle corresponds to the region occupying the full spectrum of image description, as shown in Figure 2-2 (b).

2.3 Task Definition

Taking into account the spectrum of image description, and the differences from previous image description tasks as described in the previous section, we can define image narrative as an image description passage that encompasses 1) both factual and non-factual elements at 2) both global and local scales within the image. Thus, an image narrative encompasses multiple aspects about the image beyond global, factual aspects, including, but not limited to, sentiments, details, and inferences.

Considering that multiple aspects about the image can be discussed inevitably leads us to question whether there can be multiple image narratives from a single image. In fact, as we will see in depth in Chapter 8, different viewers may attend to different parts of the image given a contextual trigger, leading to different interpretations. In order to examine and expand upon this characteristics of image narrative, we also implement an interactive image narrative generation task, in which image narrative generation process is not fully automated, but involves an interaction with user. More specifically, we generate visual questions that allow for multiple responses so that users can respond in their own unique ways, influencing the outcome. We also examine whether we can learn the user's interest from their responses to the visual questions.

2.4 Key Challenges

Our task objectives as defined above necessitate a number of modules, each of which is a challenging task itself.

First, by definition of the image narrative generation task, we need to generate multiple sentences from a single image. One way to accomplish it would be to examine beam search with different values of n, and output the results in an n-best manner. However, it will mostly result in sentences that are in different wordings but nearly identical semantically. As such, it clearly deviates from our objective of conveying a variety of contents. It follows that a more plausible approach would be to generate sentences from different parts

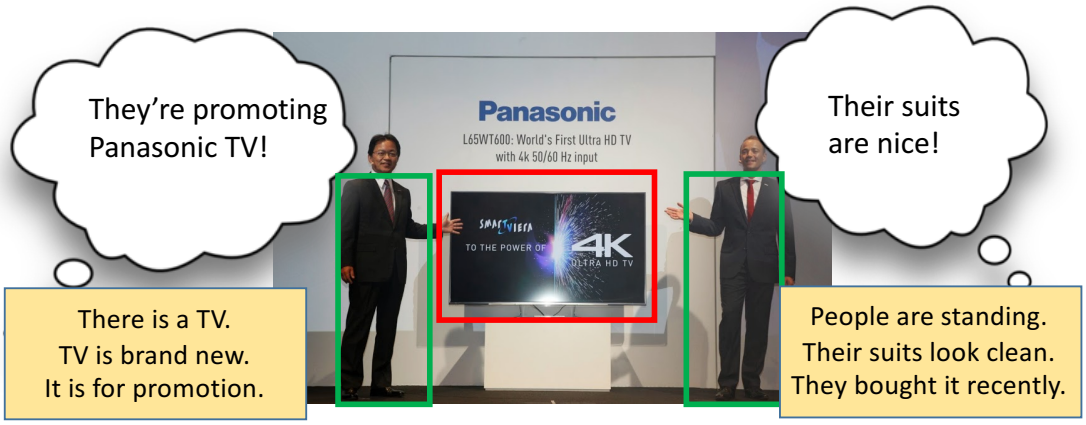


Figure 2-3: Example of varying narratives from a single image depending on the viewer's attention.

of the image, by employing region proposal or attention mechanism. DenseCap [43] is an example of a model that generates multiple captions from different regions of the same image.

Once we can generate multiple sentences from a single image, subsequent challenges will have to do with their contents. As described in previous sections, we would like to account not only for the factual components of the main event occurring in the image, but also for local, secondary components, or subjective components such as sentiment, or even non-visual elements, such as inference or assignment of story-like characteristics. Accounting for each of these elements is a major challenge for our task objective.

Incorporating local components into captions is difficult because convolutional neural network features are originally trained for single object classification. Although CNN features are known for their flexible transferability for different tasks, images containing multiples objects at varying scales inevitably affect the task's viability. As such, previous works on local captioning [43] required a construction of very costly dataset. In this thesis, we propose a model to account for multiple local objects of varying scales without having to rely on additional dataset. Our basic strategy is to "extract multiple CNN features at varying scales," since we are dealing with multiple objects at varying scales. More specifically, we extract CNN features from regions proposed by selective search, and apply vector of locally aggregated descriptor (VLAD) coding to the regional features. We also show that applying this pipeline to multiple grids employing spatial pyramid can further account for local elements.

Another challenge is to incorporate subjective components, primarily sentiments present in the images. This too is a difficult task, since multiple sentiment classes may be applicable to the same image. It is also questionable whether neural networks can be trained to distinguish between such abstract concepts that do not have concrete visual forms, not to mention the lack of dataset specialized on the topic. We examine this challenge by building a weakly-supervised sentiment dataset, from which we fine-tune a separate neural network specialized for sentiment classification. Features extracted from this network are combined with features for object classification, from which image captioning with sentiment terms are generated.

Unlike image captioning task, where the objective is clear and description can be derived in a straightforward manner by looking at the entire image, image narrative has a much wider range of potential topics, and thus has to first decide *what to talk about*, especially to include inferential elements as its contents. We implement such mechanism by generating visual questions first, and then finding answers to the questions, similarly to how a human may perform the same task. VQG and VQA techniques play the role for each part.

While different viewers may have certain patterns or tendencies in their interests and attentions upon viewing the images, it is yet uncertain whether such patterns are consistent or explicit enough to learn. We tackle this problem mainly by two novel ideas. First, we generate and ask a specific image-relevant question about each image, providing a context to which the user can provide a meaningful response, rather than simply picking the most conspicuous object. Second, we train a preference learning module, in which the system is trained to predict the user's response to a new image and new question, given the same user's previous choice on another image and a question.

Dataset and evaluation also pose challenges. To the best of our knowledge, there exists no dataset which perfectly fits our task of image narrative generation; that is, a dataset with multiple captions for single image that describe various aspects of the image. As a workaround, we take advantage of different characteristics of existing datasets, and combine them in a novel way so that the datasets can be utilized for our task. In some cases, we do construct a new dataset or complement an existing dataset for various purposes including fine-tuning, data augmentation, or references. Lack of an existing dataset that alone fits our task also causes a challenge for evaluation. Image captioning task in its conventional sense can count on popular automatic evaluation metrics with reliability. VQA task alone can also be evaluated by simple matching of generated answers and ground truth answers. However, image narratives complicate the evaluation due to both its lengths and range of topic contents. Ground truth image narratives are not provided, and even if they were, it is hard to assert that simply resembling the ground truth image narratives more correlates to better image narratives. We employ a number of evaluation metrics with different characteristics to complement each metric's drawback.

Each of these key challenges will be dealt with in greater details in the chapters to follow.

Table 2.2: Examples of image caption and image narrative for the same image. Image narrative covers a wider range of contents, not limited to factual description, with longer text.

Image	
	
Image Caption	**Image Narrative**
A brown bear is walking through logs.	A brown bear is walking.(factual description) He is living in Africa. (inference/imagination) There are mountains shown behind. (local elements) He seems to be sad. (sentiment) He is headed to get some food. (inference/imagination)
Image	
	
Image Caption	**Image Narrative**
A girl in a blue shirt is riding ski.	A girl is riding ski.(factual description) She is wearing pink ski hat. (local elements) She is visiting here during vacation. (inference/imagination) She seems to be having fun. (sentiment) The ground is slippery and she might slip. (inference/imagination)

Chapter 3

Previous Works on Vision and Language

Image narrative generation task both directly and indirectly involves a number of different tasks concerning vision and language, each of which is a highly competitive and challenging task. In this chapter, we introduce some of the core tasks and concepts that form the basis for our primary task, and review the seminal works and models for each task. We will first review the tasks concerning images only, especially how the images and the objects in the images are classified, detected, and represented. We will then review the tasks involving language,with focus on its how language has been put together with images. Representation and generation of text will be first described, and tasks involving both image and text will be reviewed, with both technical aspects and current trends. Finally, we introduce some of the important complementaries including datasets and evaluation metrics that are frequently employed both in the field and in this thesis.

3.1 Visual Recognition & Representation

3.1.1 Object Classification

Prior to deep learning era, object classification task has relied heavily on hand-crafted features, such as scale-invariant feature transform (SIFT) [61]. Since the emergence of deep learning techniques, however, the task has been dominated by CNN features, nearly without exception. Most of the recent milestones have in fact been achieved through making variations of CNN architectures.

AlexNet [52] first appeared in ISLVRC 2012, and was the first model to demonstrate the discriminative capacity of CNN, overwhelmingly outperforming other models and winning the challenge. AlexNet consists of 8 layers; 5 conv, max-pooling, and dropout layers, and 3 fully-connected layers. VGG [86] introduced a deeper architecture with 19 layers. Filter size is fixed to 3×3, adding simplicity. Due to its significant performance boost enabled by deeper architecture, along with the simplicity of its internal components, VGG has been one of the most popular network architectures in the deep learning era. 16-layer version also exists but the difference in performance is fairly negligible.

GoogleNet [88] introduced inception module, where convolutions of multiple filter sizes occur simultaneously in parallel, instead of sequentially. In particular, GoogleNet

does not contain any full-connected layer, thus significantly reducing the number of overall parameters. ResNet [33] introduced an overwhelmingly deep network architecture of 152 layers, and is currently the deepest and best-performing network at the time of writing. While its internal structure is relatively simple, its uniqueness comes from identity mapping, in which the input is passed on to the rectified linear unit (ReLU) along with the transformed input that goes through convolution.

3.1.2 Region Proposal

Region proposal task is one of the most classical tasks in computer vision, and plays an indispensable role, particularly for object detection task, which will be discuss later. Its main objective is to propose regions that are most likely to contain objects, i.e., regions that contribute to the overall semantics of the image.

Selective search [92] starts by superpixel segmentation, and proceeds with hierarchical grouping of regions in a bottom-up manner, in which neighbouring regions are combined iteratively. Different modes of selective search exist; in which HSV and Lab colorspaces are employed, and other measures, such as size of region and similarity between neighbouring regions, are also taken into consideration. EdgeBoxes [56] is based on the observation that the number of edges and contours within the bounding boxes are highly indicative of the probability of the box containing an object. While the performance gap is minor, it is much faster than selective search.

While selective search and edge boxes are based on simple geometric features or handcrafted features, other region proposal methods employing deep features have also been proposed. DeepProposal [25] extracts region candidates from the feature map of an image, by applying linear SVM trained on annotation bounding boxes at multiple scales, and applying non-maximal suppression. The region candidates then go through inverse cascade from upper, fine layer to lower, coarser layers of CNN, in order to better-localize the detected objects. DeepBox [54] uses 4-layer CNN in a bottom-up manner to rerank the proposals. Their CNN architecture generalizes to unseen categories by directly learning semantic notions of objectness.

3.1.3 Object Detection

While object classification task aims to simply classify the object in the image, object detection task needs to specifically locate the objects in the image, frequently in multiplicity. Such functionality is critical in image narrative generation, as our objective is not only to describe the primary objects in the image, but local, secondary objects as well. It also plays a central role in other modules employed in our model. For example, VQA task frequently needs to answer questions that deal with local, secondary objects.

Deep learning techniques have enabled many important advancements in object detection task as well. R-CNN [27] extracts hundreds or thousands of region proposals using selective search [92], and passes each region through CNN. Features for each region are obtained, and can be classified. Fast R-CNN [26] also accepts region proposals from selective search as input along with the image. Features for each region proposal are obtained

from the feature map of the entire image, instead of passing each region through convolutional layers, thus significantly recudcing the processing time. Faster R-CNN [78] does not rely on selective search or any other previous region proposal model, but instead has its own region proposal network, which proposes regions from the convolutional feature map. Mask R-CNN [32] further further extended Faster R-CNN by implementing object mask prediction branch which runs in parallel with bounding box recognition.

3.1.4 Feature Coding

Coding methods Bag of Visual Words (BoVW) [87] builds a global histogram of the image by finding the nearest visual vocabularies. Fisher Vector (FV) [72] produces each local descriptor from Gaussian mixture model (GMM). It has an advantage of getting high-dimensional features with relatively small size codebook, and can apply linear classifier to feature vectors. However, computational cost is high as it has exponentially larger number of parameters compared to bag of visual words. Vector of Locally Aggregated Descriptors (VLAD) [40] is works with a simplification of Fisher kernel. It searches for the nearest codeword, and computes the difference between local descriptor and nearest codeword descriptor.

Spatial Pyramid [57] is not a feature coding scheme itself, but a mode on which a feature coding techniques is applied. Its basic motivation is to add geometric invariability to coding schemes. Its most elementary format is simple division of the image into evenly distributed cells, $n \times n$, $n \times 1$, $1 \times n$, etc. Local descriptors for each cell are aggregated into a single global feature. While simplistic, it has proven to be a powerful performance booster for a wide array of tasks. Inevitably, its disadvantage is that feature dimensionality gets multiplied in proportion to the number of cells. Also, cell division may not be optimal for image contents. Discriminative Spatial Pyramid [31] makes up for such limitation of spatial pyramid by assigning weights to each local descriptor upon aggregating them to a global feature.

3.2 Vision and Language

3.2.1 Text Representation

Representing text in a format that is understandable to machines is as challenging as it is central, since both structure and semantics must be preserved. For example, Bag-of-Words [44] simply represents text as histogram of each word's frequency. While useful and convenient, it ignores and loses ordering and semantics, necessitating more eloquent representations that retain semantics and context.

Word2Vec [67] extends conventional skip-gram model by sub-sampling frequent words during training and applying a simplified version of noise contrastive estimation [30], which results in better vector representations as well as convenient additive property and extensibility to phrases. Similarly, sentence2Vec [58], in which an unsupervised algorithm learns to predict surrounding words in contexts from paragraph, has also been developed.

3.2.2 Text Generation

Rule-based or template-based text generation had been considered a norm for text generation for decades [75, 76], but this too has rapidly shifted towards data-based paradigm. Success of statistical machine translation [49] has been a paragon of data-based approach to text generation, and such approach is now prevalent in tasks beyond translation.

Long short-term memory (LSTM) [34] has been a pivotal milestone in text generation task, and has become a dominant technique in a variety of relevant tasks, particularly in image captioning task. An LSTM block contains a number of gates, and its distinct feature is the inclusion of *forget* gate f_t, which controls how long an information is supposed to remain in the block, hence affecting dependency. Original equations for activation of each gate as defined in [34, 24] were as following:

$$\begin{aligned} i_t &= \sigma_g(W_{xi}x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma_g(W_{xf}x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma_g(W_{xo}x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \circ \sigma_t(c_t) \end{aligned} \tag{3.1}$$

where x_t, h_t, c_t are input, output, and cell state at *t*, and f_t, i_t, o_t are vectors for forget gate, input gate, and output gate respectively.

In image captioning task, LSTM usually takes in image features as its input, learns to generate one word at a time, conditioned on the image features and previously generated words, until the end token is predicted.

3.2.3 Image Captioning

Image captioning task has become one of the most competitive tasks empowered by deep learning techniques. As such, it is now generally considered to have reached the level of practical usage and have been incorporated into many products or services. Note that image captioning task can be further categorized into multiple tasks; dense captioning, stylistic captioning, etc. In this section, we will mostly review the seminal works in the general image captioning task for single-sentence factual image description. Related works for different categories of image captioning task will be discussed in the relevant chapters to be followed later in the thesis.

First, some of the pioneering works from pre-deep-learning era deserve an honorable mention. BabyTalk [53] used object detectors, attribute classifiers, and prepositional relationship functions to extract information from the image, and constructed conditional random field to predict labels and generate sentences. While this work did not rely on any *deep* techniques in the way the task is normally handled today, it marked a significant paradigm shift from simply retrieving most likely sentence to *generating* one. Ushiku et al. [93] extracted key phrases from an input image using passive aggressive with average pairwise loss, and generated caption by spinning those key phrases with an experimental grammar model.

A majority of recent work on image captioning task have been dominated by the usage

of convolutional and recurrent neural networks for feature extraction and caption generation respectively, although with substantial variations. Karpathy et al. [47] exploited multimodal RNN to generate descriptions of image regions, aided by the alignment model of CNN over image regions and bidirectional RNN over sentences, which are intermingled via a multimodal embedding. Inspired by statistical machine translation, Vinyals et al. [95] built a model in which the encoder RNN for source sentences is replaced by CNN features of images. Long short-term memory (LSTM) was employed as a generative RNN of nonlinear function. Thus, given an input image *I*, a sentence S=(S_0,...,S_N) describing the image is generated as follows:

$$\begin{aligned} x\_1 &= CNN(I) \\ x_t &= W_e S_t, t \in 0...N-1 \\ p_{t+1} &= LSTM(x_t), t \in 0...N-1 \end{aligned} \tag{3.2}$$

Xu et al. [100] took a similar workflow, but introduced attention-based model, which learns to update the saliency while generating corresponding words. Donahue et al. [19] expanded the CNN-LSTM architecture to activity recognition and video recognition by building long-term recurrent convolutional networks (LRCNs). Time-varying inputs are processed by CNN whose outputs are fed to a stack of LSTMs. Fang et al. [21] took a more linguistically inspired approach by training visual detectors for words with multiple instance learning, which learns to extract nouns, verbs, and adjectives from regions in the image. Maximum-entropy language model generates a set of candidates, which are reranked by sentence-level features and deep multimodal similarity model.

3.2.4 Visual Question Answering

Visual question answering (VQA) has escalated the interaction of language and vision to a new stage. In VQA, an image textitI and a question textitQ_i about the image are provided, and the goal is to provide an appropriate answer to the question in natural language. It thus requires understanding of semantics of the questions, and visual clues necessary to answer those questions. It is noteworthy that the questions consist not only of objective, visually verifiable questions, but also of questions that require common-sense, inference, even imagination. This characteristic enables us to obtain both visual and non-visual contents about the image.

A number of different approaches have been proposed to tackle VQA task, but so far, classification approach has been shown to outperform generative approach [1, 45]. [22] proposed multimodal compact bilinear pooling (MCB) to combine multimodal features of visual and text representations. This approach won the 1st place in 2016 VQA Challenge in real images category. [79] proposed DualNet, in which both addition and multiplication of the input features are performed, in order to fully take advantage of the discriminative features in the data. This method won the 1st place in 2016 VQA Challenge in abstract scenes category. [101] was one of the first to propose attention model for VQA. They proposed stacked attention networks (SANs) that utilize question representations to search for most relevant regions in the image. [70] also built an attention-based model, which optimizes the network by minimizing the joint loss from all answering units. They further-proposed

an early stopping strategy, in which overfitting units are disregarded in training. [62] argued that not only visual attention is important, but also question attention is important. Co-attention model was thus proposed to jointly decide where to attend visually and linguistically. [48] introduced multimodal residual network (MRN), which uses element-wise multiplication for joint residual learning of attention models. [82] proposed an attention-based model to select a region from the image based on text query. It is yet arguable as of now whether visual attention is a must-have prerequisite for higher performance [45]. Finally, [97] introduced a model to extract information from general knowledge base to answer image-based questions.

More advanced types of VQA task have also been proposed. Das et al. [16] proposed Visual Dialog task which attempts to go a step further than a single round of visual question answering by considering a dialog history $H_0,...,H_{t-1}$ on top of the image (I) and a question (Q).

3.2.5 Visual Question Generation

Visual Question Generation (VQG) task, as its name suggests, attempts to generate image-relevant questions. Note that the generated questions may not always be answerable solely by visual clues, and may necessitate common sense, inference, or imagination. For example, it may ask the name of the famous person in the image, or what would happen after the event depicted in the image.

VQG is still at its infancy at the point of this writing, and as such, has very few relevant works, with [68] best exemplifying the effort. They attempted to generate visual questions that a human might naturally ask upon seeing the image, instead of visually verifiable questions designed for AI as in VQA tasks. Such design objective enables a generation of intriguing contents that cannot be captured by image captioning or VQA task. They compared different models for the taks, and show that multimodal RNN outperforms maximum entropy language model or machine translation model.

3.3 Complementaries

3.3.1 Dataset

ImageNet [17]: ImageNet was originally constructed to help develop the techniques for object classification task, and ever since has become the criterion to test the effectiveness of various models. While multiple versions of ImageNet exist, most of them consist of 1,000 object classes, each of which in turn contains 1,000 images of the class. Throughout the paper, whenever we speak of CNN features, we assume that the corresponding CNN architecture has been pre-trained on ImageNet, unless noted otherwise.

FlickrStyle [46]: While ImageNet consists of multiple classes of objects, FlickrStyle comprises multiple stages of visual *styles*, ranging from Baroque and Rococo to Impressionism. Note that, while the characteristics of style classification task deviate from those of object classification, it is still visually grounded with fairly unambiguous boundaries between different classes.

Table 3.1: Examples of images and ground truth captions in MS COCO.

Image	Captions
	• a man holding a bag watches a baseball game unfold • there is a baseball game going on • a baseball player holding a bat on top of a field • the baseball player is up to bat at the stadium full of players • a man holding a baseball bat while waiting on a field
Image	**Captions**
	• a man holding a bag watches a baseball game unfold • there is a baseball game going on • a baseball player holding a bat on top of a field • the baseball player is up to bat at the stadium full of players • a man holding a baseball bat while waiting on a field

MS COCO [60]: Microsoft Common Objects in Context (MS COCO) has become a de facto standard dataset for image captioning task. MS COCO consists of training, validation, and test set, which contains roughly 80k, 40k, and 40k images respectively. Each image contains 5 human-written ground truth captions. **Flickr 30k** [102] and **Flickr 8k** [35] are also frequently used datasets for image captioning task.

MS SIND [37]: Microsoft Sequential Images Narrative Dataset (SIND) was collected to tackle the visual storytelling task, and contains roughly 81k images that constitute 20k sequences. In most cases, a sequence consists of 5 images that loosely compose a story. Each image in a sequence is accompanied by a caption, but unlike captions in other datasets, each caption takes part in the overall story of the sequence. In other words, each caption's primary objective is not to describe the respective image, but to assign the image a certain role in the scheme of storytelling for the overall sequence. As such, it stylistically deviates from usual image captions that describe a single image. Also, since each caption is part of a story, it frequently contains contextual wordings, such as pronouns. As we will see later on in this thesis, this results in a very low correlation between each image and its caption.

Visual Genome [51]: Visual Genome dataset presents a highly dense collection of annotations in various types including attributes, relationships, descriptions, and question answer pairs. Some of the works [43, 107] on image captioning and VQA have relied on this dataset for training.

VQA [2]: Some of the other datasets on VQA task are worth mentioning. DAQUAR [63]. COCOQA [77] was built by applying rule-based transformations to captions in MS COCO. Visual7W [107] consists of crowd-sourced questions and answers, but questions for each image are designed to observe the *7w*-form, arguably restricting the spectrum of contents. FSVQA [84] attempted to eliminate the ambiguity of the answer by making the answers full sentences. VQA 2.0 [29] has also been released. It not only increased the size of the dataset, but attempted to balance and diversify the answers depending on the

questions, so that question alone cannot be the substantial clue for answering.

VQG [68]: While very few datasets were designed solely for VQG task [68], any dataset for VQA task can be used for VQG task as well, since they inevitably contain images and the associated visual questions. In this thesis, we utilize both [68] and [2] for visual question generation.

3.3.2 Evaluation Metrics

Evaluation metrics for text data can be classified into two categories; automatic evaluation and human evaluation. Popular automatic evaluation metrics have proven to be correlated to human judgment to a moderate degree, and are considered to be indicative of the overall quality of the text data to a plausible degree. Since most of them refer to provided ground truth answers for comparison, however, they display highly task-specific characteristics, and have low portability to novel tasks. On the other hand, human evaluation can be applied to nearly any type of text evaluation task, thus highly flexible to novel tasks. Yet, participating subjects may have widely varying standards for evaluation, which requires highly attentive setting for the evaluation requesters. It also necessitates financial cost, which can easily skyrocket depending on the difficulty and size of the task.

We first review some of the most frequently used automatic evaluation metrics. **BLEU** [71] was originally introduced as an evaluation metric for machine translation task, and has since become one of the most popular evaluation metrics not only for machine translation task, but also for image captioning task. Its principal focus is on the precision of n-gram, hence there can be multiple versions of BLEU depending on the order of n, usually from BLEU-1 to BLEU-4. Discriminative BLEU [23] has been proposed to incorporate qualitative weights into BLEU, where each reference sentence $r_{i,j}$ is assigned a weight $w_{i,j}$ in [-1,+1].

ROGUE [59] is an evaluation metric originally designed for evaluating document summarization tasks. As opposed to BLEU's emphasis on precision, ROGUE is more shifted towards recall.

METOER [18] is another n-gram based metric designed for evaluating machine translation task. Unlike BLEU or ROGUE, it mainly focuses on unigram matching, but also takes recall into consideration. It also attempts to account for semantic matching by looking up synonyms from WordNet.

While the evaluation metrics described above were originally designed for natural language processing tasks, **CIDEr** [94] was designed specifically for evaluating image captioning task. Given an image I_i, Its evaluates a candidate caption c_i by referring to the consensus of a set of ground truth captions $S_i = s_{i1},...,s_{im}$. Sentences are mapped to sets of n-grams up to n=4, with each word converted to its root form. In order to discern which n-grams are informative, it also employs *tf-idf* calculated from the entire dataset.

Evaluation metrics described above are all automatic evaluation metrics that produce an overall score calculated from pre-defined equations or rules. On the other hand, **human evaluation** may be more appropriate for certain types of tasks. A number of crowdsourcing services exist, but Amazon Mechanical Turk is by far the most popular. The overall procedures is as follows: a requester sets up a human intelligence task (HIT), such as providing image descriptions, or evaluating them, for instance. The requester also sets

up the financial reward to be given to the workers upon completion and approval of the HIT. The requester also sets up the upper bound for the number of workers to participate in the HIT. The workers complete the HIT by following the directions provided by the requester, and the requester may approve or reject the completed work. It is also possible to filter out the workers for quality insurance; for example, only the native speakers of a certain language may be permitted to participate, or only the workers with approval rate over a given threshold may participate.

Evaluation metric for VQA task has been solidified as the one provided by [2], which is available on its evaluation server. It is basically a precision matching with the generated answer and human-provided ground truth answers, as defined by the following equation:

$$\min(\frac{\text{\#humans that provided that answer}}{3}, 1) \times 100 \tag{3.3}$$

Thus, the generated answer is deemed correct if 3 or more of the 10 ground truth answers match the generated answer.

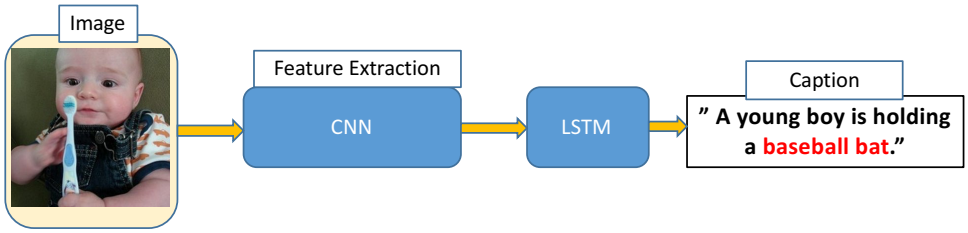
Chapter 4

Dense Image Representation for Locally Robust Captioning

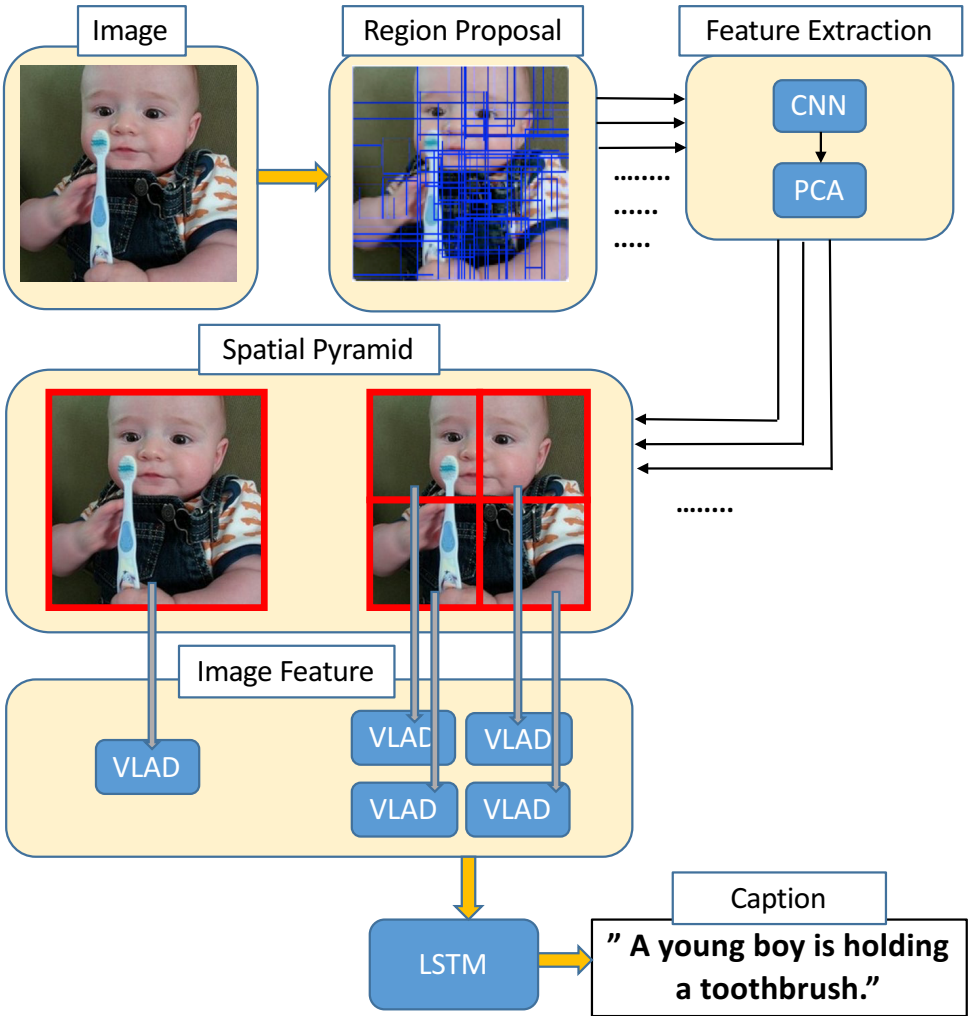
The workflow of extracting features from images using convolutional neural networks (CNN) and generating captions with recurrent neural networks (RNN) has become a de-facto standard for image captioning task. However, since CNN features are originally designed for classification task, it is mostly concerned with the main conspicuous element of the image, and often fails to correctly convey information on local, secondary elements. We propose to incorporate coding with vector of locally aggregated descriptors (VLAD) on spatial pyramid for CNN features of sub-regions in order to generate image representations that better-reflect the local information of the images. Our results show that our method of compact VLAD coding can achieve comparable performance to raw CNN features with much lower dimensionality, and, when combined with spatial pyramid, it results in image captions that more accurately take local elements into account.

4.1 Accounting for Local Details in Captioning

Image captioning task has gained unprecedented attention with successful application of convolutional neural networks (CNN) and recurrent neural networks, especially long short-term memory (LSTM) units [47, 95, 100]. Such pipeline of extracting features from images using CNN, and mapping the representation to ground truth captions using RNN or LSTM has become a de-facto standard, employed by most recent works on image captioning task. With the current standard of CNN-LSTM pipeline, the novelty can come from either representation part (CNN), or learning and generation part (LSTM). We tackle the former part in this chapter. While CNN provides a powerful yet relatively compact representation of the image, it is noteworthy that CNNs are originally trained for classification of objects, with the goal of correctly identifying mostly a single, main object in the image. In image captioning task, however, it is frequently necessary to account not only for main objects in the image, but also for local, secondary objects. Although CNN mostly results in correct captioning with regards to the main object, it frequently results in incorrect captioning for local, secondary objects, as shown in Figure 1(a). This is natural in a sense that CNNs were originally trained for classification of main objects in the image.



(a)



(b)

Figure 4-1: a) Example of incorrectly captioned local objects using conventional approach, b) Overall workflow of our model.

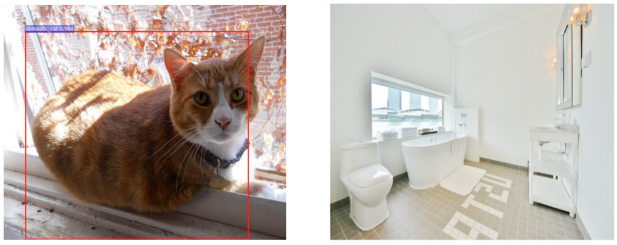


Figure 4-2: Examples of object detections using RCNN. Although it works well for objects classes present in the dataset (e.g. *cat* in the image on the left-hand side), it frequently fails to detect objects that are not present in the dataset (e.g. *toilet*, *bathtub*, *window* in the image on the right-hand side not detected).

In this chapter, we introduce a novel application of spatial pyramid VLAD coding to CNN features at different sub-region levels, in order to generate more locally robust representation, and more accurate captioning. VLAD has been popular coding method for compactly representing images from a large-scale dataset. However, its drawback of discarding spatial information has also been pointed out. In order to compensate for this drawback while preserving compact representation of VLAD, spatial pyramid VLAD has been suggested, and we apply it to CNN features.

In the conventional approach, CNN features are extracted from the image in its entirety without explicitly dealing with local objects. On the other hand, in our model, CNN features are extracted from a large number of bounding boxes from sub-regions proposed by selective search, which are mostly oriented towards local objects. This way, features are extracted not only from the entire image, but from each object or region whose importance is likely to be neglected in the conventional way. We then cluster the CNN features into a number of codewords, and perform VLAD coding using the codewords. Such coding results in very compact representation of images, as little as 3% of the CNN features at its minimum, and yet shows comparable performances. We then implement VLAD coding at different regions of different levels, thus implementing spatial pyramid VLAD so that the spatial information of the features can be preserved. By doing so, we generate captions that more accurately and frequently account for local elements of the image that have been overlooked. Figure 1(b) illustrates the overall workflow of our approach.

We optimize our method with various settings to investigate the influence of parameters and to find the best-performing combination. We also compare our method to previous works, as well as combining our method with conventional approach. Experimental results show that our method can more accurately and frequently account for local objects than the conventional approach, frequently providing details at the level of human-written ground truth captions.

Our main contributions comprise 1) showing that VLAD coding of CNN features from sub-regions can represent the images more compactly, 2) combining it with spatial pyramid to account for spatial information, and 3) applying it to image captioning task to generate more locally robust captions.

4.2 Related Works for Dense Image Captioning

Most previous works have represented images with CNN features extracted from the whole image, usually without paying explicit attention to local objects. In that regard, Johnson et al. [43] show similar motivation to our work. They introduced DenseCap, which attempts to localize the regions in the image and incorporate it into captioning. DenseCap generates multiple captions from multiple regions, but is not fundamentally different from previous works when it comes to the single caption from the entire image. Also, their dense localization layer is trained on a dataset whose construction process is highly costly, with manual box-setting and labelling on crowdsourcing. Our method does not involve any manual labelling, and can work with any existing dataset.

Karpathy et al. [47] exploited multimodal RNN to generate descriptions of image regions, using alignment model of CNN over image regions and bidirectional RNN over sentences, which are intermingled via a multimodal embedding. This model relies on region convolutional neural network (RCNN) [27] to detect objects. However, since RCNN model is fine-tuned on the limited number of classes (only 20 classes of objects on PASCAL dataset [20], 80 classes on MS COCO [60], or 200 classes from ILSVRC 2012), it frequently fails to detect objects not included in the object classes of the dataset. Figure 2 shows examples of success and failure in object detection using RCNN. Since our model relies on selective search for object detection and region proposal, it is not limited by the number of object classes in the dataset.

Vector of locally aggregated descriptors (VLAD) was introduced by Jegou et al. [40], as a model to compactly represent images in a large-scale dataset, and has been a popular coding method for images. They used the simple L2 normalization method for normalizing VLAD descriptors. Arandjelovic et al. [4] demonstrated that intra-normalization and recording multiple VLADs for an image, along with vocabulary adaptation, can further enhance the performance of VLAD. As a method to approximate global non-invariant geometric statistics, Lazebnik et al. [57] introduced spatial pyramid matching technique, a simple extension of bag-of-features representation, in which histograms for local features are aggregated in each sub-region. Although spatial pyramid can find useful global features from each level, it has been reported to be weak at high geometric variability, necessitating a combination with invariant features. On the other hand, VLAD coding is usually performed on locally invariable descriptors, such as SIFT, yet it does not preserve spatial information. In order to compensate for these mutual weaknesses, Zhou et al. [106] introduced spatial pyramid VLAD, which plays a central role in the method introduced in our model.

Sanchez et al. [80] showed a far simpler approach for taking spatial information into account, by simply incorporating the coordinate information into the feature vector and augmenting it. We will also examine this approach and compare it to our model in Section 4.

Some previous works [28, 15] have applied similar methods to ours by extracting deep activation features from local patches at multiple scales, and coding them with VLAD or Fisher Vector. However, previous works mainly dealt with scene classification or object classification, in which the necessity for explicitly dealing with local objects and spatial information is less pronounced. On the contrary, image captioning task requires that local

objects be very clearly reflected in the captions. To our knowledge, our work is the first to apply such workflow to image captioning.

4.3 Proposed Model: SPVLAD of R-CNN Features

4.3.1 Region-Based Feature Extraction

We first obtain a set of region proposals from images with selective search [92]. Selective search starts by superpixel segmentation, and proceeds with hierarchical grouping of regions in a bottom-up manner, in which neighbouring regions are combined iteratively. We used the *fast* mode of selective search, in which HSV and Lab colorspaces are employed, and complementary similarity metrics, such as color, texture, and fittingness between neighbouring regions, are also taken into consideration. Although fast mode lacks some of the features present in *quality* mode, such as intensity, it is roughly 5 times faster than quality mode, while sacrifice in performance is relatively small (98% recall compared to 99% on Pascal 2007 test set). As discussed in Section 2, the benefit of using selective search is that it is not limited by the number of object classes, which was the case for object detection using RCNN and other object detection methods based on datasets.

We then extract CNN features from *all* regions proposed by selective search. The rationale behind extracting CNN features from region proposals is that, since regions now tightly encompass particular objects, CNN features from the regions will be highly representative of that particular object. The motivation for feature extraction from *all* regions, instead of running non-maximum suppression to reduce the number of regions, consists of two reasons. First, CNN features will go through spatial coding, in which an insufficient amount of region samples can cause data sparsity problem. Second, it is intuitive that conspicuous objects will have multiple proposals of different sizes, so that the influence of such objects will remain strong even after coding, and are likely to be reflected in captions.

Region-based variations of CNN, such as Fast-RCNN [26], have made the idea of extracting CNN featrues from multiple regions feasible. Although Fast-RCNN was originally designed for detection task, we took advantage of it to perform high-speed feature extraction. Instead of provided network models trained for detection task, we used VGG network [86] trained for classification on ImageNet [17], and extracted 4096-dimensional features from the second fully-connected (fc7) layer.

4.3.2 VLAD Coding

Since we eventually have to perform VLAD coding of the features separately on each grid of spatial pyramid, 4096-dimensions will be too large and redundant. We thus performed dimensionality reduction with principal component analysis (PCA) on CNN features extracted from the regions. We trained the PCA with features of 250k randomly sampled regions, and separately performed reduction to 128, 256, 512, and 1024-dimensions.

We then performed codeword learning with K-means. The number of clusters included 1, 2, 4, 8, 64. Centroids were initialized with K-means++ [5], as random Gaussian initialization resulted in skewed clustering. Based on the codewords learned using K-means, we

Figure 4-3: Dividing the images into multiple grids using spatial pyramid enables us to focus on various local objects.

perform the VLAD coding on CNN features obtained after PCA, with signed square rooting normalization as in [73, 3, 39, 41]. Thus, given a set of d-dimensional PCA-applied CNN features $X = (x_1,...,x_n)$ from n region proposals, and a set of d-dimensional k codewords $C = (c_1,...,c_k)$ obtained using K-means++, CNN feature x_i is mapped to its closest codeword as follows:

$$x_i \mapsto c_j \text{ where } c_j = \arg\min sqn(x_i - c) \text{ for } c \in C \tag{4.1}$$

and $sqn()$ indicates signed-square rooting normalization. Then, VLAD coding for each centroid is performed by summing up the differences between the centroid c_i and all CNN features x_j assigned to that centroid as follows:

$$V_i = \sum x_j - c_i \forall x_j \mapsto c_i \tag{4.2}$$

The final VLAD vector is obtained by concatenating the VLAD codings for each centroid:

$$V = \sum V_i \text{ where } |C| = i \tag{4.3}$$

Dimensionality of the resulting vector at this point is the dimensionality of CNN features after PCA times the number of clusters from K-means.

4.3.3 Spatial Pyramid

Although VLAD encoding is known to perform well on preserving locally invariant features, it is at the cost of discarding spatial information. Previous works have thus proposed Spatial Pyramid VLAD [106], in which VLAD coding is performed at multiple levels of different sizes, going from coarse to fine sub-regions, inspired by spatial pyramid matching. Spatial pyramid allows us to focus on local objects that would have been neglected in the entire image. In Figure 4-3, for instance, local objects such as chair, coffee, cell phone, and newspaper are placed in a relatively inconspicuous position at a small scale, and are

likely to be neglected in the whole frame. By placing spatial pyramid, however, many of them are now closer to the center of each grid at a larger scale, and are more likely to be reflected in the feature extraction process.

In order to determine which grid a particular region belongs to, we simply locate the center of the region. An alternative would be to assign a region to all overlapping grids. However, our examination of the alternative resulted in grids not being much different from each other, and thus not discriminative enough. This may be attributed to the fact that many conspicuous elements in the images are frequently large in size, thus occupying multiple grids, and consequently making the grids all similar. We thus resorted to the center of the region for its grid assignment, as it is also more concurrent with our initial motivation of preserving local information.

Since previous works [57] have reported that levels beyond three obtain only an insignificant amount of improvement at the cost of enlarged dimensions, we also set our spatial pyramid at three levels; 1×1, 2×2, and 3×1 (left, middle, right). Thus, using up to second level of the pyramid will result in representations of the image at 5 different sizes or locations, and up to third level will have representations at 8 different sizes or locations per image. Because the number of grids is fairly small in our case, we did not perform any normalization for different levels or different grids.

4.3.4 Caption Generation

Since the motivation of this chapter is to tackle the representation part of the image captioning process, we generally follow the conventional approach for caption generation part, applying LSTM to our representation of the images and ground truth annotations. We employ the "vanilla" architecture for LSTM as used in [95], which is a modified version of the Equation (3.1). Gate functions are defined as following:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \end{aligned} \tag{4.4}$$

Word vectors were trained with random initialization, and sigmoid function is used for non-linearity throughout all gates except along with hyperbolic tangent for memory cell update. Training was performed for 100 epochs in all experiments, and beam size of 1 was used.

4.4 Experiment

4.4.1 Setting

We apply our proposed model to MS COCO [60]. Train and validation split add up to roughly 120,000 images, and running selective search on the entire dataset resulted in 45.9M region proposals, approximately 385 regions per image. From these regions, we follow the procedures described in Section 3 to generate captions.

Table 4.1: Performances on BLEU with varying dimensionalities of CNN features after PCA (1 cluster, no spatial pyramid)

Dimension	BLEU-1	BLEU-2	BLEU-3	BLEU-4
128	**59.3**	39.3	25.4	16.9
256	**59.3**	**39.4**	**25.5**	**17.0**
512	51.2	30.8	17.9	10.9
1024	48.0	25.4	13.2	7.8

Table 4.2: Performances on BLEU with varying number of clusters (256 dimension, no spatial pyramid)

Clusters	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	59.3	39.4	25.5	16.9
2	58.7	38.8	24.9	16.3
4	**60.7**	**41.0**	**27.3**	**18.4**
8	57.7	36.4	22.5	14.6
64	46.4	25.0	10.4	4.4

We compare the performance of our proposed model with baseline, in which 4096-dimensional CNN features are extracted from the entire image, and inserted to LSTM as input with no further preprocessing. Note that, since the evaluation server for test split of MS COCO allows only a limited number of submissions, we split the validation split into two splits for validation and test respectively. Parameter validation was performed on the former split of the validation split, and the remaining experiments with comparison to baseline models were carried out on the latter split from the validation split.

4.4.2 Parameter Validation

A number of factors can potentially contribute to a large increase in dimensionality of the final representation; dimensionality of reduced CNN features, number of clusters, and level of pyramid. Even if performance increases, excessively high dimensionality would be impractical. We would thus like to consider an appropriate trade-off between performance and dimensionality, and performed a number of validations to set up appropriate parameters.

We first examined the influence of dimensionality of the CNN features after PCA. The number of codewords was fixed to 1, and spatial pyramid was not employed. We varied the dimensionality of CNN features as 128, 256, 512, and 1024, to which VLAD coding was applied. The 4096-dimensionality CNN features prior to PCA were not employed since it will cause the final vector to be impractically large, when combined with multiple clusters and spatial pyramid.

Table 1 shows the performances of our model with various dimensionality of CNN features after PCA, using BLEU [71] as evaluation metric. Surprisingly, lower dimensionality outperforms higher ones by a considerable margin. This indicates that 4096-dimensional CNN features are not optimal and contain redundant information. Inspection of our pipeline provides another explanation. CNN features in our model are extracted from small regions suggested by selective search, which mostly contain objects at a large proportion, as opposed to the "whole" images containing various objects and components at varying scales.

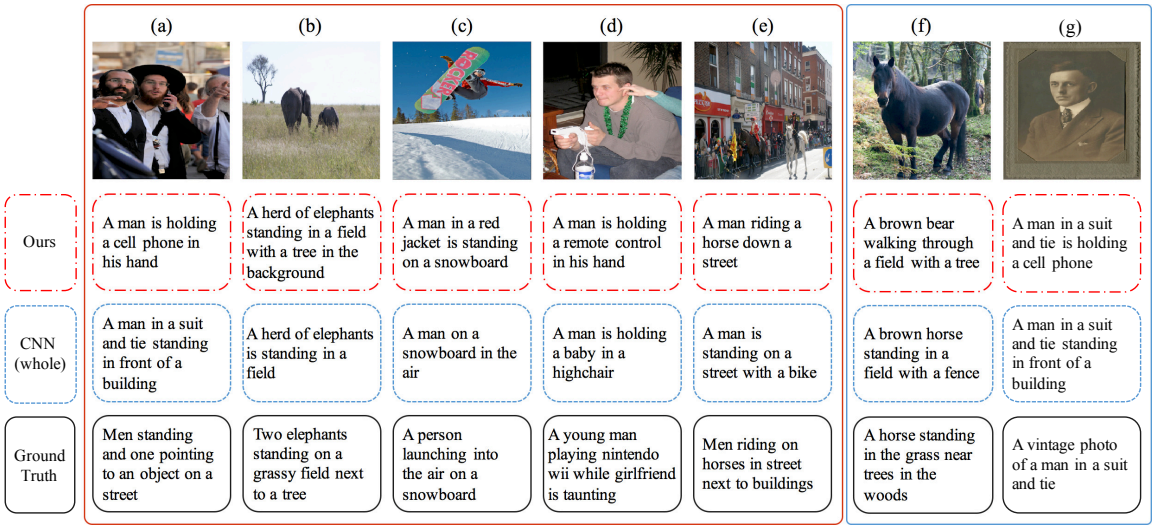


Figure 4-4: Qualitative analysis of our model and baseline approach. Examples in the red solid box demonstrate that our model of VLAD coding of CNN feature generates more accurate captions with regards to local objects. Blue solid box corresponds to failure cases.

Thus, much fewer dimensionality is needed to correctly classify and represent the objects. Furthermore, since clusters are obtained from these features, less compact representation with high dimensionality is likely to result in noisy clusters, leading to noisy VLAD coding, which negatively affects the accuracy of captions.

Notably, 256-dimensional coded representation, even with only one cluster and no spatial pyramid, resulted in best performance, almost equal to the performance of 4096-dimensional CNN features used in conventional image captioning task. Further-reducing the dimensionality to 128 resulted in very slight decrease, but still comparable to 256-dimension features and CNN features, despite being only 1/32 of its size. This demonstrates that VLAD coding of CNN features from region proposals contains highly discriminative ability, while being very compact.

Secondly, we examined the influence of the number of clusters. Dimensionality of the CNN features was fixed to 256, which achieved the best performance in the first validation, and spatial pyramid was not employed. We varied the number of clusters as 1, 2, 4, 8, and 64, with which VLAD coding was performed upon 256-dimensional CNN features. Table 2 shows the performances of our model with various numbers of clusters. The differences in performance between low number of clusters are relatively small, while larger number with 64 clusters noticeably degrades the performance. Similarly to the dimensionality case, a large number of clusters results in sparse clustering, where many clusters end up with no vector assigned to it. Increasing the number of iterations is likely to improve the performance to similar levels as lower numbers of clusters, but it indicates that its convergence is much slower. Although using only one or two clusters resulted in comparable performances, it would not be much different from average pooling, and thus would not fully utilize the benefit of VLAD coding. In the following experiments for comparison to previous works, we mostly proceed with the combination of 256-dimensional CNN features, 4 or 8 clusters, and spatial pyramid of level 2 or 3.

Table 4.3: Performances of each model on BLEU.

Model	Cluster	SP Lev.	Total Dim.	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN (whole)	N/A	N/A	4096	62.0	42.4	28.0	18.7
Ours	4	1	1024	60.7	41.0	27.3	18.4
		2	5120	61.3	41.4	27.5	**18.9**
		3	8192	61.1	40.8	26.9	**18.8**
	8	1	2048	57.7	36.4	22.5	14.6
		2	10240	58.5	37.4	23.7	15.5
		3	16384	58.9	38.8	24.9	16.5
[80]	4	N/A	1036	56.5	35.9	21.9	13.9
Ours+CNN	N/A	N/A	9216	60.5	41.0	27.1	18.4

4.4.3 Additional Setup 1: Feature Augmentation

In object detection and classification literature, some alternatives to spatial pyramids have been proposed. We examine one of such proposals in order to examine whether, and to what extent, local information can be preserved without using spatial pyramids.

We examine a simple feature augmentation method proposed by Sanchez et al. [80], where coordinate information is concatenated to the vector of descriptor. Specifically, given 2D-coordinates of a region patch $m_t = [m_{x,t}, m_{y,t}]^T$ with a descriptor x_t of size D, which in our case corresponds to CNN features, and the patch scale σ_t, where the image is of size H and W, we augment the dimension of the descriptor by 3, resulting in a new vector $\hat{x}_t \in \mathbb{R}^{D+3}$ as follows:

$$\hat{x}_t = \begin{pmatrix} x_t \\ m_{x,t}/W - 0.5 \\ m_{y,t}/H - 0.5 \\ \log \sigma_t - \log \sqrt{WH} \end{pmatrix} \tag{4.5}$$

Thus, it accounts for location information of each region implicitly in the feature vector, rather than explicitly dividing regions and generating separate representations. The benefit of this approach is the simplicity of its implementation. We implement this method to the best-performing set of parameters in our model, except without spatial pyramid; 256-dimension CNN features with 4 clusters. Features are thus augmented to 259-dimension, and the resulting final representation becomes 259×4=1036 dimensions, roughly a quarter of the dimensionality of conventional CNN features.

4.4.4 Additional Setup 2: Ours + CNN (Whole)

We examine a combination of our model with the conventional one, where CNN features are extracted from the image in its entirety. We extracted activations from second fully-connected (fc7) layer of 4096-dimension with VGG-19 network. For our model, 256-dimensional features with 4 clusters and spatial pyramid up to level 2 were employed. Combining the two adds up to 9216-dimensional vector per image.

4.4.5 Evaluation

Table 3 summarizes the results from our model and previous works, with various combinations of parameters, and BLEU as the primary evaluation metric. The scores for whole CNN features are from our own experiment under the same condition for fair comparison. Other papers have reported varying, sometimes higher results with their own methods, although mostly in a close range. While enhancing the performance of LSTM is out of scope of this chapter, it is an active research area, and replacing our vanilla LSTM with more recent versions is likely to boost the overall performance of all models.

Most variations of our model achieve performances very close to CNN features, especially with the best-performing combination outperforming CNN features at BLEU-4. Since BLEU-4 is computed from higher-order of n-grams than others, it is frequently employed as the primary source of evaluation metric [95, 81], as it better-indicates the overall semantic similarity. It thus shows that accounting for local objects as in our model enhances the overall semantic accuracy.

Overall, 256-dimensional CNN features with 4 clusters up to level 2 and 3 resulted in the best performance. In all cases, models with spatial pyramids outperform those without, which demonstrates that paying attention to local elements by dividing the images into sub-regions is able to reflect more detailed aspects of the images. Patterns observed in Section 4.2 mostly hold true, with compact dimensionality and a small number of clusters performing better. Feature augmentation achieved reasonably high performance, but fell short of our model. If there are multiple regions covering the same objects, those regions have close feature vectors as well as close coordinates. Thus, when they are assigned to clusters, the explicitness of coordinate information is likely to become subdued to an insignificant extent. This again shows the importance of explicitly accounting for local objects in image representation.

In order to examine how different our captions are from original CNN features, we calculated BLEU score of our model with captions generated from CNN features as references, which resulted in 59.4/47.9/40.2/35.3. The score indicates that both captions feature similar contents, but also that they have considerable differences in their wordings and in their dealing of details.

We also performed human evaluation to compensate for the limitations of automatic evaluation metrics. We asked on Amazon Mechanical Turk which caption reflects the images in more details for 5,000 images. In addition, we also asked the workers to determine whether each image contains local elements or not. Overall, our captions had a marginal lead with roughly 37.5% of the responses preferring our captions as opposed to 34.2% of CNN features, while both captions were considered to be at the same level in 28.2%, as shown in Table 4. Our model shows clearer advantage when the image contains local elements. Of 3,022 images classified as containing local elements, ours model was preferred in 39.0% of the responses as opposed to 28.2% of CNN features, while both models were considered to be at the same level in 34.1%. On the other hand, it showed comparative weakness when it comes to the images not containing any local elements. Our model was preferred in 35.2% of the responses, while CNN features were preferred in 42.7%. Such result is concurrent with the failure examples shown above, where the images cannot benefit from spatial pyramid. Yet, the overall results demonstrate that our model did achieve

Table 4.4: Number of votes for each model on human evaluation.

	# images	Ours	Both	CNN
All	5000	**1876**	1411	1713
with local elements	3022	**1180**	1032	868
without local elements	1978	696	379	845

our objective of better reflecting local elements.

4.4.6 Discussion

Figure 3 shows examples of images and captions generated by our model, CNN features, and the combination of two, along with ground truth. Resulting captions show that our model does capture local, secondary objects more correctly and frequently than CNN features, frequently providing details at the level of human-written ground truth captions. It verifies that our model has successfully learned to apply mapping between local objects and their linguistic correspondences to unseen images, and that our motivation of capturing spatial information with spatial pyramid has succeeded to a plausible extent.

There were indeed cases where our models performed more poorly than CNN features. Such cases were mostly the ones in which it was hard to find any component other than main object in the image. Our model often talks about non-existent secondary object, or, in worse case, incorrectly describes the main object. As much as it can deal better with particular local objects when they are present, it turns out to be less efficient when there are no secondary objects so that segmenting the image into spatial pyramid becomes unnecessary. Figure 4 shows examples of such failure cases.

Since our model better-deals with local objects, while original CNN features can handle main objects well, it seems intuitive to combine the two and expect balanced results. However, their performances were lower than respective models, presumably due to their large dimensionality, which requires more training time to fully converge. The resulting captions seem to display somewhat mixed characteristics, slightly leaning more towards captions from CNN features in terms of contents.

Since applying PCA to our VLAD-coded CNN features from sub-regions not only reduced the dimensionality but also enhanced the performance, one possible alternative would be to apply dimensionality reduction to original CNN features as well, and see whether it retains its discriminative strength. If successful, it will make the combination of two models more compact and thus more practical.

We would also like to discuss why our proposed model is better at reflecting local variations than using global CNN features only. In CNN, global pooling is performed at each local patch. Whether it is maximum pooling or average pooling, local variations are either discarded completely, or subdued. In our model, however, calculating vector distance between the centroids and the extracted region features is performed evenly regardless of the size of the region. Thus, local variations, even from small regions, are more likely to be preserved and reflected in the final representation.

4.5 Summary & Discussion

We introduced a method for image representation incorporating spatial pyramid VLAD to CNN features of sub-regions suggested by selective search, in order to generate more locally robust image captions. Our VLAD coding of CNN features, both with and without spatial pyramid, was able to achieve performance comparable to CNN features, while having much lower dimensionality, as little as 3% of its size at its minimum. We optimized our model via parameter validations, and learned that combination of low dimensionality after PCA with appropriate number of clusters yields the best results. Combining spatial pyramid turned out to enhance performance not only on evaluation metrics, but in resulting captions dealing well with local objects.

Our model more accurately and frequently accounts for local objects than previous methods, such as feature augmentation or conventional CNN representation for the whole image. It frequently dealt with local objects at the level of human-written ground truth captions. Our model did show weaknesses when there are no local elements so that spatial pyramid is hardly necessary, but a more carefully crafted combination with conventional CNN features is likely to complement mutual weaknesses of respective models.

A number of potential improvements can be made from each stage of our model. For example, other region proposal methods, such as DeepProposal [25] may as well be employed, as long as they are not bound by a pre-determined set of object classes. Employing intra-normalization for VLAD coding as in [4] can also potentially improve the performance. Furthermore, since spatial pyramid has pre-determined division of cells that may not always correspond to the ideal localization of objects in the image, it may be helpful to build a spatial pyramid in which the size and location of the cells are determined by the results of region proposals followed by non-maximum suppression. Discriminative spatial pyramid [31] in this sense is a potential candidate for improving the performance, and will be our immediate future work. Finally, since we exclusively dealt with the representation part of the image captioning task, a novel approach to tackle the generation part of it would naturally be of interest.

Figure 4-4 shows more examples of the generated captions from our model, conventional CNN features from the entire image, ground truth, and additional setup where we combined our model with conventional CNN features from the entire image.

Figure 4-5: More examples.

Chapter 5

Image Captioning with Sentiment Terms

Image captioning task has become a highly competitive research area with successful application of convolutional and recurrent neural networks, especially with the advent of long short-term memory (LSTM) architecture. However, its primary focus has been a factual description of the images, including the objects, movements, and their relations. While such focus has demonstrated competence, describing the images along with non-factual elements, namely sentiments of the images expressed via adjectives, has mostly been neglected. We attempt to address this issue by fine-tuning an additional convolutional neural network solely devoted to sentiments, where dataset on sentiment is built from a data-driven, multi-label approach. Our experimental results show that our method can generate image captions with sentiment terms that are more compatible with the images than solely relying on features devoted to object classification, while capable of preserving the semantics.

5.1 Accounting for Sentiments in Captioning

Image captioning task bridges the gap between two of the most fundamental artificial intelligence domains, namely language and vision. Recent surge of deep learning approaches has escalated the task to an unprecedented stage, where generated captions can nearly rival those by humans [12][19][21][47][95][100]. However, the objective of image captioning task has revolved around the factual description of the images, such as the objects, their motions, and their relations. On the contrary, non-factual components subject to viewers' interpretation of the images, mostly appearing in a form of adjective or adverb, have been missing. We define such subjective elements as the *sentiment* of the image, and modifying terms describing it as *sentiment terms*. Such non-factual sentiment terms broaden the expressibility, enrich the aesthetics of the language, and are more human-like.

The reason that research on image captioning with sentiment terms has stagnated is partly due to lack of dataset specialized in sentiments, and the difficulty of building such dataset, which inevitably poses several conundrums. First, there is no clear boundary between classes. An image labeled as 'happy' may also be labeled as 'cute,''beautiful,' etc., and the same holds true in the opposite sentiment polarity. One way to deal with this issue may be to have a highly limited number of inclusive classes, as is often done in facial

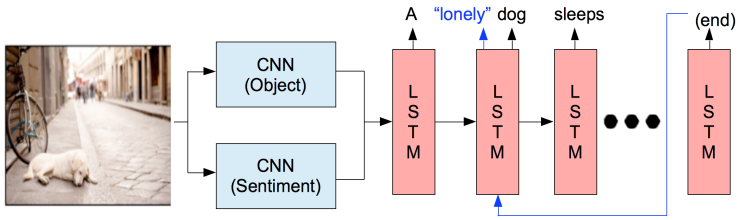


Figure 5-1: Overall workflow of our model. Our model extracts features from two convolutional neural networks, one for object classification, and the other for sentiment classification. Two sets of features are combined and input to LSTM, which generates the caption. LSTM returns to the most likely state for sentiment term, reloads the output from previous unit, and determines the term based on probability distribution from separate vocabulary for sentiment terms.

expression classification task [7]. While this has an advantage that the distinction between classes is comparatively clear, it is at the cost of losing the subtle nuances apparent within the inclusive classes. For example, a non-negligible discrepancy lies in between 'hilarious' and 'peaceful,' both of which belong to the inclusive positive sentiment polarity. Also, it is difficult to port such limited number of classes to the images of a broader domain, in which the range of possible subject matters is extremely wide and humans are frequently not present. Unlike facial expressions, sentiments from the images of general domain can be interpreted with a great variety, often accompanying disagreements among the viewers. Furthermore, certain images may permit labels from opposite polarities to be attached (*e.g.*, 'friendly' and 'eerie' for a smiling pierrot). In fact, the results from our human evaluation in Section 2 testify that humans indeed find it very difficult to agree on a single label for given images, even when the number of classes is relatively few. We thus conclude that the sentiments should be represented with multiple labels, as there is no single 'correct' label, but only an indefinite set of acceptable, appropriate labels.

Another practical issue has to do with the financial cost of building such dataset. If we were to rely on crowd sourcing services to have 1 million images manually labeled, as was the case for ImageNet [17], the cost would easily skyrocket up to tens of thousands of dollars. Even so, due to the subjective nature of sentiments, it is not guaranteed that the results will be reliable. As an alternative to manual labelling, we note that the viewers' comments towards the images on social network frequently reflect the sentiments of the images. Figure 2 shows examples of comments reflecting the sentiments associated with the images. We exploit this characteristic of the comments in order to inexpensively label the images. As we will see in Section 5.4, it requires attentive filtering processes and is only weakly supervised, but is capable of building a fairly agreeable dataset at virtually zero financial cost.

In this chapter, we tackle a novel problem of image captioning with sentiment terms. We build sentiment dataset in a data-driven, multi-label setting, from which an additional convolutional neural network (CNN) learns sentiment features. Since our work is fundamentally an incremental work built on top of conventional image captioning task, we generally follow the approach of CNN-LSTM pipeline for the most part, except features for object classification and sentiment classification are obtained separately, and the LSTM unit with highest probability is revisited after sentence is complete, in order to produce the

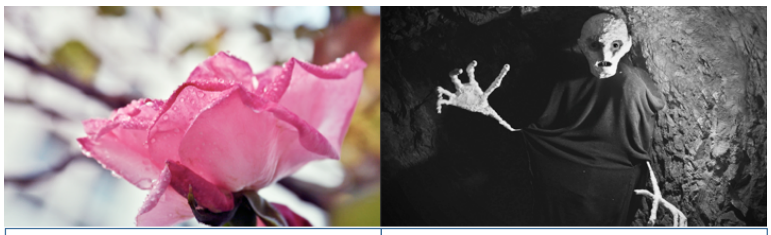


Figure 5-2: Actual examples of comments on social network services describing the sentiments of the images. Underlined words are the sentiment terms to be used as labels.

sentiment term. Figure 5-1 presents a diagram of the workflow of our model. Throughout the chapter, we refer to 'sentiment terms' as the words whose positive or negative score on SentiWordNet [6] is 0.5 or higher. Features for object classification are combined with sentiment features and are fed into LSTM [34], which generates the caption in its conventional way, except it returns to the state with highest probability, and reloads the output from previous unit to determine the sentiment term.

Our main contributions can be summarized as following: 1) proposal of a novel task of image captioning with sentiment terms, 2) utilization of multi-label learning to deal with subjective nature of sentiments, and 3) introduction of a data-driven approach to inexpensively build a dataset on sentiments and its public release.

5.2 Related Works for Image Captioning with Sentiment Terms

Traditionally, sentiment classification of images has been carried out mostly with hand-crafted features. For example, Siersdorfer et al. [61] suggested that SIFT combined with global color histogram can be a good indicator of the sentiments of the images, although dealing only with positive/negative binary classification. Borth et al. [9] represented images with adjective-noun pairs collected from web mining and analyzing tags associated with the images. More recently, Katayev et al. [46] demonstrated that neural networks can be fine-tuned to distinguish between different styles and atmospheres, and that it outperforms other hand-crafted features, such as GIST or color histogram. This led to an idea that we may also be able to fine-tune neural networks to determine the appropriate sentiment of a given image.

However, these works have mostly overlooked the inclusion of sentiment terms in their captions. In this regard, most intimate to the nature of our work is by Mathews et al [66]. They proposed a switching RNN model, consisting of two parallel RNNs for factual and sentiment description respectively. However, they built separate models for positive and negative terms and applied it to the same set of images under premise that any image can be interpreted from either sentiment polarity. While it is true for certain images as was discussed in Section 5.1, there are a substantial amount of images that hardly permit an interpretation from both sentiment polarities (for example, it is rare to see a description with a negative term given a close-up of a toddler's smiling face or a blooming flower). We

thus believe that the polarity of the sentiment terms in the description should be determined automatically, unaided by manual choice of polarity. It consequently follows that a single RNN suffices for us, although we still need two CNNs for separate feature extractions.

5.3 Proposed Model: Fine-tuned Sentiment CNN

5.3.1 Multi-Label Learning

Note that, although we utilize multi-label setting, our objective deviates from that of traditional multi-label learning in that we do not necessarily aim to predict identical set of labels as ground truths, as there exists no definitive set of labels. In fact, prediction of a single appropriate label suffices since there is usually only one modifying term for an object at a time. Thus, multi-label setting in our case is for representing the images and projecting them in a sensible space, rather than replicating identical set of labels.

Multi-label classification itself is an active research area with a variety of approaches. The bottom-line for us is that the approach should be implementable with ease in standard deep learning frameworks, Caffe [42] in our case. One possibility is to utilize the approach known as *Binary Relevance* [10][104] which decomposes the multi-label learning into a set of independent binary classification problems. Thus, m training examples x_i whose associated labels form a set Y are viewed as following:

$$D_j = \{(x_i, \phi(Y_i, y_j))|1 \leq i \leq m\}$$
$$\text{where } \phi(Y_i, y_j) = \begin{cases} 1, & \text{if } y_j \in Y_i \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

In our case, x_i corresponds to CNN features, extracted from 2nd fully-connected layer (fc7) of VGG [86]. Then, the set of labels for unseen example is determined by the obtained binary classifiers g_j for q classes:

$$Y = \{y_j|g_j(x) > 0, 1 \leq j \leq q\} \tag{5.2}$$

While simplistic, it has proven to generalize well in various domains, and has become a foundation for more sophisticated multi-label learning techniques [104]. Another feasible approach is *Random k-Labelsets* [91], in which every unique set of labels is considered a distinct class. It has two obvious downsides that the number of classes exponentially grows, and that there may be classes in test set that are unseen in training set. We thus opt to proceed with the mechanism of binary relevance.

While multi-label setting can be implemented with slice layers in deep learning frameworks, its setup can be highly tricky. A much simpler method that essentially performs the same task is to simply duplicate the images and assign them different labels. The benefit is its simplicity, while the downside is that the size of dataset grows, which in our case approximately doubled. Note that, since the predicted label for a given image will always be the same, we limit the images in the test set to those containing only one label. Otherwise, the accuracy will never be able to go beyond 100/(average number of labels)% at best.

5.3.2 Caption Generation

Encouraged by its recent successes in image captioning task [19][47][66][95][100], we employ LSTM [34] as our caption generator, and follow its conventional setting for the most part. The input to LSTM are the features extracted from the second fully-connected layer (fc7) of CNN, although our model necessitates additional CNN features as will be discussed in Section 5.5. Word vectors are trained with random initialization, and sigmoid function is used for non-linearity throughout all gates except along with hyperbolic tangent for memory cell update as defined in equation (4.4) reprinted below for convenience:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \end{aligned} \tag{5.3}$$

The unique part of our LSTM is that we force it to contain at least one sentiment term in its prediction. We hypothesize that sentiment terms are most likely to modify the nouns most characteristic of the sentence, hence most characteristic of the image. We thus keep track of the probability every time a noun is predicted, and once the prediction of the entire sentence is complete, return to the LSTM unit which predicted the noun with highest probability, and feed it again with the output from the previous LSTM unit. We keep a separate vocabulary V_{senti} consisting of sentiment terms only, which is a subset of V consisting of all terms, and predict a word again at the LSTM unit we return to, but this time only from V_{senti}. Thus, we are essentially forcing an insertion of a modifying sentiment term that may have been skipped in favour of the characteristic noun due to smaller likelihood. In summary, sentiment term w_{senti} is the term in V_{senti} which maximizes

$$p_{t_{\text{return}}+1}(w_{\text{senti}}) \; = \text{LSTM}(x_{t_{\text{return}}})(w_{\text{senti}}), \tag{5.4}$$

where $x_{t_{\text{return}}}$ is the input at $t = t_{\text{return}}$ determined by the learned parameters and word vectors up to that state. Also, originally predicted word $w_{t_{\text{return}}+1}$ at this state, which is part of the generated caption of length N, satisfies the following:

$$\begin{aligned} w_{t_{\text{return}}+1} &= \arg\max_{w} p_{t_{\text{return}}+1}(w) \; , w_{t_{\text{return}}+1} \in V_{\text{noun}}, \\ t_{\text{return}} &= \arg\max_{t} \max \text{LSTM}(x_t)(w) \; , 0 \leq t \leq N-1 \end{aligned} \tag{5.5}$$

5.4 Weakly-Supervised Sentiment Dataset

5.4.1 Construction

We first collected 2.5M images and 28M comments associated with those images from image hosting services, namely Flickr and DeviantArt. Although comments are of different nature from captions, they have been reported to be highly indicative of the sentiment of the images [13], and thus fit our purpose of representing visual sentiments. Sentiment terms that frequently appear on ground truth descriptions of existing standard datasets were manually chosen as queries to retrieve the images. From the collected comments, we count the occurrences of sentiment terms, accompanied by a series of filtering processes as following:

- negation: sentiment terms that are negated are filtered out (*e.g.*, "not very funny")

- spam: suspicious comments are ignored, and comments with URL are also ignored, regardless of the contents
- color and motion terms: sentiment terms describing specific colors are filtered out. Also, sentiment terms describing motions in the appearance of a gerund (*e.g.*, "jumping") are filtered out with a few exceptions (*e.g.*,"smiling").
- first-person subject: sentiment terms used to modify the first-person subject are filtered out (*e.g.*, "I'm serious")
- inflection: adverbs and comparative forms of adjectives are inflected to their respective original adjective forms (*e.g.*, "happily" or "happier" to "happy") except they are filtered when followed by an adjective (*e.g.*, "simply" as in "simply beautiful")
- dual part-of-speech: sentiment terms that have high frequency as a different part-of-speech and require more sophisticated usage of parser are filtered (*e.g.*, "mean","pretty")
- general, non-visual terms: sentiment terms with unclear description criteria that provide no visual clue are manually filtered out from the final counts (*e.g.*, "good,""bad")

After filtering and counting of the sentiment terms, we need to determine the appropriate number of classes. We experimented with three different number of classes (20, 50, and 100) determined by the frequency of terms in the comments. According to the number of classes, images without any comment that contains at least one label from the classes are filtered out, and most frequent labels up to maximum of five that appear in the comments for each image and exist in the classes are selected as the labels for the image. Sizes of the resulting datasets and their performances on various evaluation methods are summarized in Table 5.1. We refer to this dataset as *Sentiment Dataset* in the rest of the chapter.[1]

5.4.2 Validation

In order to approximate the reliability of the comment-generated labels, we performed a human validation over a subset of our dataset, consisting of 10k images and 19,975 labels. Two workers were assigned per image, and each worker was asked to select the labels that do not seem appropriate given the image. We marked the labels inappropriate if two workers agreed that the label is inappropriate, and 572 labels were agreed to be inappropriate, which amount to 2.9% of the tested labels. While not entirely satisfactory, this yields a fair bound for the reliability of a comment-generated label. Frequent sources of biases were viewers commenting on the overall quality of the images rather than sentiments, or compliments to the uploaders. More refined filtering process considering these biases will further enhance the reliability. Figure 5-3 shows examples of comment-generated labels and the labels that turned out to be inappropriate.

In order to comprehend the proximity of classification models' performances to human capability, we performed classifications by humans for each number of classes on Mechanical Turk with a subset of dataset, consisting of 1,000 images respectively. Table 5.1 shows the performances of neural networks and human classification on each dataset with varying number of classes. It is noteworthy that it is difficult even for humans to achieve high

[1] http://www.mi.t.u-tokyo.ac.jp/static/projects/sentidata/

Table 5.1: Top-1 accuracy of classification by various models. Apart from human evaluation carried out on 1,000 sampled images, all other tests are performed on the entire dataset.

Dataset	Class	Size	SIFT+RGB [61]	VGG [86]	Human
ImageNet [17]	1000	1M	-	71	85
FlickrStyle [46]	20	80k	-	40.7	75.1
Sentiment	100	1.1M	6.8	11.3	16.1
	50	.7M	14.2	20.5	25.8
	20	.5M	19.4	28.7	40.1

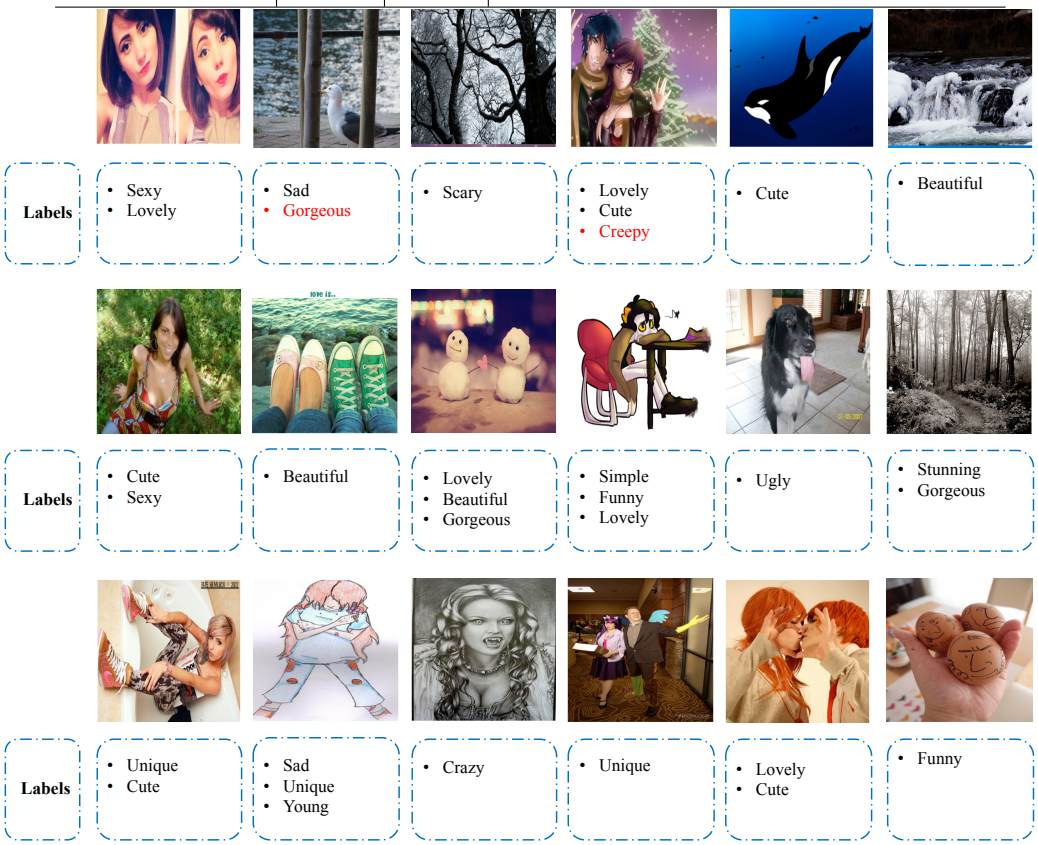


Figure 5-3: Examples of comment-generated labels. Labels in red color indicate the labels agreed to be inappropriate on Mechanical Turk.

accuracy due to the unique nature of sentiments in which subjectivity prevails. However, since we are assuming a single label per image in the test instead of multi-label, the actual accuracy is supposedly higher. Table 5.3 shows the final 20 classes, their sentiment scores as defined on SentiWordNet [6], and the number of images in the dataset belonging to the class.

Some of the traditional hand-crafted features have been known to correlate well in sentiment classification. We applied SIFT and RGB color histogram features followed by a linear SVM in a similar manner as [61] to our dataset, and compared the performances. The results were not as competent as in binary classification, again confirming the complication of multi-class sentiment classification of images and that hand-crafted features may not be adequate for more elaborate classification tasks.

Table 5.2: Size of vocabulary, parameters, dimensions for each dataset

Dataset	V	V_{senti}	Param	$\text{Param}_{\text{senti}}$	$\text{Dim}_{\text{senti}}$
Flickr8k	4,448	104	3,855,968	4,186,208	1,290
Flickr30k	11,794	399	7,624,466	7,824,658	782
MSCOCO	14,141	512	8,828,990	9,260,350	1,685

Table 5.3: Sentiment score and number of images for each class.

Class	POS	NEG	OBJ	Images	Class	POS	NEG	OBJ	Images
angry	0	.875	.125	25,824	lovely	.625	0	.375	123,004
beautiful	.750	0	.250	254,905	sad	.125	.750	.125	75,263
crazy	.625	.500	-	37,810	scary	0	.750	.250	30,773
creepy	0	.875	.125	28,830	sexy	.625	0	.375	72,186
cute	.625	0	.375	325,606	simple	.875	.500	-	46,874
dirty	0	.750	.250	16,417	stunning	.750	.625	-	24,049
funny	.500	.500	-	85,590	ugly	0	.750	.250	21,840
gorgeous	.750	0	.250	71,712	unique	.500	0	.500	24,981
handsome	.625	0	.375	28,404	weird	0	.250	.750	51,072
hot	.625	0	.375	48,486	young	.625	.250	.125	39,612

5.5 Experiment

5.5.1 Setting

We chose VGG with 19 layers [86] as our network model. Presumably, the characteristics of our sentiment dataset substantially deviate from datasets devoted to object classification task, and we thus aim to adjust the network parameters slightly more aggressively. We fine-tune the layers from the first fully-connected layer (fc6) and on, as opposed to the conventional approach in which only the last fully-connected layer (fc8) is fine-tuned. The initial hyper-parameter setting for fine-tuning is as follows; gaussian weights, initial learning rate of 0.001, step decay of 0.1 at every 20k iterations, etc. Features are extracted from the images via fine-tuned network above using Caffe framework [42].

Since we are aiming to generate only one sentiment term in the caption, it is likely redundant to have the same number of dimensions for sentiment features as the ImageNet features, and we thus reduce the dimensionality of sentiment features using PCA, after which they are concatenated to ImageNet features. The size of reduced dimensionality varied depending on the dataset for experiment (Table 5.2) and was determined by Minka et al.

We compare the performance of our proposed method with those of four baselines. Note that sentiment terms are force-inserted in all models except for the first baseline. :

- ImageNet: ImageNet features with conventional LSTM caption generation without sentiment term force-inserted
- ImageNet+: ImageNet features with conventional LSTM caption generation with sentiment terms force-inserted by LSTM
- Bigram: ImageNet features with sentiment terms chosen by an external bigram corpus, namely Google Web Trillion Word Corpus [11]. No additional features were

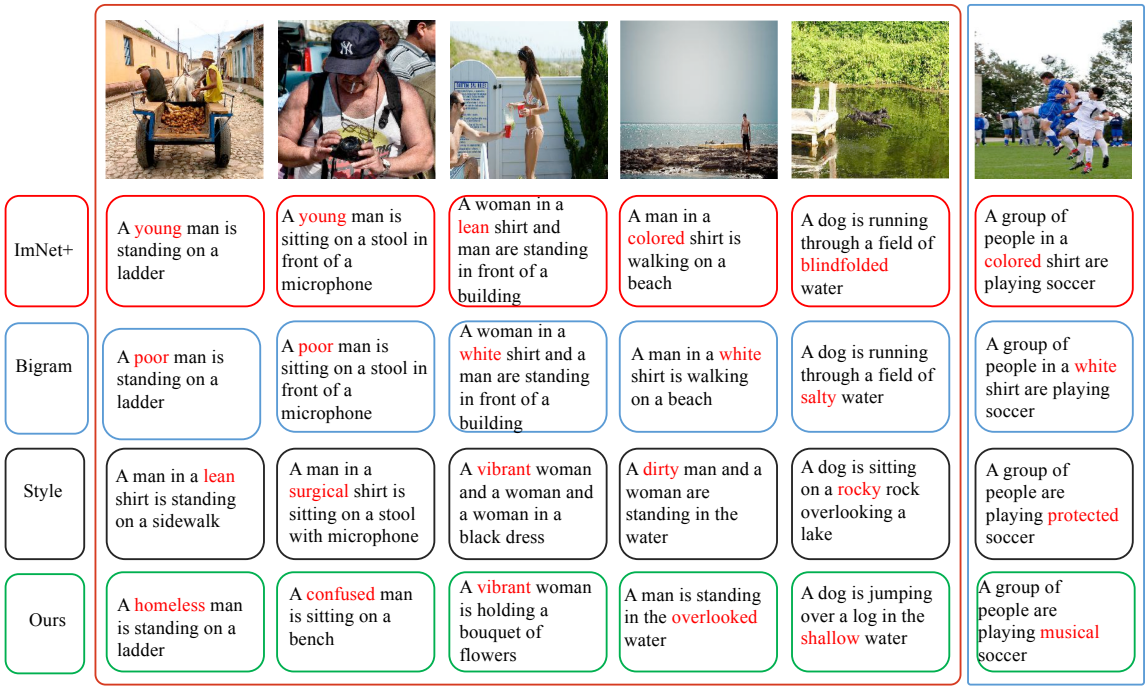


Figure 5-4: Examples of captions generated by each model with sentiment term. Words in red color indicate the inserted sentiment term by each model. Failure case is also shown in the right-hand side. Note that removing sentiment terms from captions by ImNet+ or bigram models will be identical as the captions generated by original ImageNet model without sentiment terms.

added, but most likely sentiment term was inserted according to the external bigram corpus, regardless of the probability distribution determined by LSTM, and thus regardless of the features from corresponding images.

- Flickr Style: ImageNet features combined with features from Flickr Style dataset [46] that contains forced sentiment terms from LSTM in the same way as our model. Caffe provides the CaffeNet model fine-tuned on Flickr Style dataset, which achieves about 39.2% accuracy on its own test data. Using VGG 19-layers and fine-tuning from fc6 as we did in our model slightly boosts up the accuracy to 40.7%, and we refer to Flickr Style features as those extracted by this network.

Experiments are carried out on three standard datasets for image captioning task; Flickr 8k [35], Flickr 30k [102], and Microsoft COCO [60]. In each dataset, we build vocabulary V consisting of the words that appear twice or more in the ground truth captions. Vocabulary of sentiment terms V_{senti} are also built in the same way for each dataset. Sizes of the vocabularies and dimensionality of sentiment features for each dataset, along with the number of learnable parameters, are shown in Table 5.2.

5.5.2 Evaluation & Discussion

Figure 5-5 shows some of the figures and captions with and without sentiment terms, along with all sentiment terms generated by the models. For ImageNet+ and bigram models,

sentiment terms are frequently distant from dominant sentiment of the image. Bigram model results in terms that are commonplace, *e.g.*, 'hot dog','punk rock,'smart phone,' yet frequently irrelevant to the image. Inspecting the captions generated by our model demonstrates that, while the captions are frequently composed in different wording, the contents of the descriptions are mostly identical to the captions from ImageNet model, with only a few exceptions. This testifies to the modificative linguistic effect of our newly added features, and yet shows that they can preserve the main semantics for most images. Most of the incorrectly described concepts (for example, color turns out to be frequently inaccurate) also prevailed in other models. Another noticeable phenomenon is that our model occasionally has highest confidence on different subjects from ImageNet+ model, inserting sentiment terms at different part of the caption, which is another intriguing effect of the new features. See Figure 5-5 for more examples of the generated captions from the baseline models and our model.

Table 5.4 shows the performances of our model and baselines on a number of automatic evaluation metrics. First, note that BLEU scores [71] are seemingly impaired for all models in which additional terms are inserted. This is inevitable since there are a plethora of ground truth captions that do not contain any sentiment term, and insertion of sentiment terms will inevitably lower the overall resemblance to those captions, especially as the size of n-gram grows. Since the captions from ImageNet+ and bigram models are exact replica of the original ImageNet model except for the sentiment term, their BLEU scores decrease less than two other models. Since new features are added in the remaining two models including our proposed model, their scores deviate slightly more from ImageNet model, but our model's scores are comparable to those of two fore-mentioned models, especially as the size of dataset grows, and consistently outperform the scores by Flickr Style model.

The same can be argued for different metrics. For instance, since perplexity is computed from the inverse probability of each predicted word, sentiment terms, which in general have lower probability than non-sentiment terms in ground truth captions, are prone to increase the perplexity. This reveals one of the unsettling aspects of currently popular automatic evaluation metrics, particularly for rating image captions; while convenient to use and indicative of the overall quality to a plausible extent, what most of them measure is the resemblance of the generated captions to the ground truth captions (via n-grams or alignment) or the probabilistic likeliness of the predicted caption, not the overall appropriateness with regard to the attributes of the images. As such, higher scores on automatic evaluation metrics occasionally do not necessarily imply better quality of the captions, and more so when sentiment terms are concerned. Hence,

In order to compensate for limitations of evaluation metrics for dealing with sentiment terms, we also resort to human evaluation, and interpret it as a complementary criterion of evaluation. We performed two types of human evaluation tasks on Mechanical Turk. In the first task, workers were given an image and one of the captions from four models with sentiment terms, and were asked to determine whether the sentiment term is appropriate. In the second task, workers were given an image and all four captions with sentiment terms, and were asked to rank the captions in consideration of both semantic accuracy and appropriateness of the sentiment terms. Two workers were assigned per image in the second task. In both tasks, the same set of 2,000 images from MS COCO was used.

Our model was able to receive the highest appropriateness rating in the first task, which

Table 5.4: Performances of the captions generated by each model on MS COCO determined by automatic evaluation metrics. Note that no additional features were added in first three models. ImNet refers to original ImageNet features from VGG with no sentiment term inserted. ImNet+ indicates that sentiment terms are inserted to captions from ImNet model. Our model is referred to as Sentiment. Results reflect that sentiment terms have become noise and were put at disadvantage in evaluation.

Dataset	Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR [18]	Cider [94]
MSCOCO	ImNet	62.0	42.4	28.0	18.7	12.1	62.3
	ImNet+	56.6	36.2	21.9	13.1	11.8	44.1
	Bigram	56.7	36.2	22.0	13.3	11.7	44.1
	Style [46]	55.5	34.8	20.7	12.4	11.3	38.2
	Ours	56.5	35.9	21.7	13.0	11.6	43.0
Flickr8k	ImNet	51.1	33.0	20.1	12.6	11.4	39.4
	ImNet+	47.8	29.3	16.6	9.6	10.9	30.6
	Bigram	48.2	29.4	16.7	9.7	10.9	30.5
	Style	44.6	26.3	14.6	8.3	10.4	29.6
	Ours	46.6	28.0	15.6	9.0	10.5	30.3
Flickr30k	ImNet	55.0	35.7	22.9	14.8	10.5	33.0
	ImNet+	51.6	31.9	19.2	11.6	10.4	25.9
	Bigram	51.8	31.9	19.1	11.5	10.4	27.1
	Style	49.3	30.1	17.9	10.8	9.9	22.0
	Ours	51.1	31.0	18.5	11.1	10.1	23.0

demonstrates that our model was more frequently able to capture the dominant sentiment in the image and generate appropriate terms. In other words, newly added features in our model were more compatible with sentiment terms in the ground truth captions, and the prevalent sentiment of the images. On the other hand, our model was below ImageNet+ model in the ranking task, although only by a close margin. A possible cause is that some of the sentiment terms in ImageNet+ model's captions were considered *compatible* with the image, even when it is not a dominant sentiment in the image (*e.g.* "lean shirt,""young man"). It may also be a reason for its agreement score being lower. Table 5.5 summarizes the results from human evaluation. Inter-rater agreement was calculated based on S metric given by [8]:

$$S = \frac{QP_a - 1}{Q - 1} \tag{5.6}$$

where Q is the number of possible choices, and P_a is the probability of raters assigning the same decision. All agreements fall into a range of 'fair' agreement according to [55].

Our final remark is that many sentiment terms in the generated captions, regardless of the model, turned out to be irrelevant to emotions or aesthetics, which was dominant in our choice of classes, reflecting different natures of online comments and ground truth captions. In fact, sentiment terms in our classes appeared rather infrequently. If we were to shift our focus more strictly towards emotions or aesthetics, building V_{senti} more restrictively, for example, by setting a higher threshold on SentiWordNet, may have done the job.

Table 5.5: Performances of each model on human evaluation

Dataset	% Appropriateness	Avg. Rank	Agreement
ImNet+	.404	2.17	.209
Bigram	.383	2.66	.260
Style [46]	.367	2.98	.301
Sentiment	**.448**	2.25	**.307**

5.6 Summary

We tackled a novel problem of image captioning with sentiment terms. We introduced a method to inexpensively build a dataset on sentiments, trained in a form of multi-label learning, and exploited the learned features on long short-term memory to generate image captions with sentiment terms. It was comparable on automatic evaluation metrics to conventional models, and human evaluators found the captions from our model to be more appropriate with regards to the sentiment of the image.

Figure 5-5: More examples of captions with sentiment terms generated by each model.

Chapter 6

Visual Question Generation (VQG) and Answering (VQA)

Visual question answering (VQA) task not only bridges the gap between images and language, but also requires that specific contents within the image are understood as indicated by linguistic context of the question, in order to generate the accurate answers. Thus, it is critical to build an efficient embedding of images and texts.

6.1 Motivation for Visual Question Generation and Answering

Recent rise of deep learning methods including convolutional neural networks (CNN) and recurrent neural networks (RNN) has escalated a large number of artificial intelligence tasks to an unprecedented stage, where the performance frequently rivals that of humans. Tasks such as object classification, scene classification, and object detection demonstrated the ability to correctly recognize and locate the images both holistically and regionally, whereas tasks such as caption generation or object retrieval demonstrated that deep learning methods can successfully bridge the gap between images and language. Visual question answering (VQA) task further promotes the boundary of deep learning applicability and complicates the problem by necessitating multiple prerequisites, potentially encompassing all of the above-mentioned capabilities; as it needs to understand the question, locate or classify the objects/scenes mentioned in the question, and generate appropriate answers.

6.2 Related Works for Visual Question Generation and Answering

Visual Question Generation (VQG): While question answering task has been a classic NLP task, visual question generation (VQG) task has hardly been tackled. [68] introduced VQG dataset with crowd-sourced questions, but their goal was to generate questions that humans might naturally ask, rather than questions designed for AI. In fact, it explicitly

excluded questions that can be answered by visual cues. Since the paper's goal was to generate the questions only, their dataset does not contain the answers to them. In order to examine and exploit the unique design characteristic of this dataset, we collected the answers to the questions in this dataset, and used them as a part of our training data. Details will be shown in Section 7.3.

The paper also shows that multimodal recurrent neural network outperforms other generative models including maximum entropy language model (MELM) [21] and machine translation (MT) model [14]. We also generate questions with multimodal recurrent neural network, except we replace the gated recurrent neural network (GRNN) with LSTM.

Visual Question Answering (VQA): Visual question answering (VQA) task itself has been popularized with the advent of dataset provided by [2], consisting of 0.25M images, 0.76M questions, and 10M answers. They also report baseline results from methods with multi-layer perceptron and LSTM [34].

VQA: Real Images Real images category is currently by far the more popular and competitive task in VQA. [64] introduced Ask Your Neurons. Unlike the baseline provided by [2], in which image features and question features are embedded to common space at the last stage prior to classification, they built a system where image features are shared at each LSTM unit for processing question features. They also performed comparison of different operations for fusing input features, and concluded that summation performs better than multiplication. In our work, however, both summation and multiplication are performed, which demonstrates significant improvements.

Many recent papers reporting competitive results have relied heavily on various types of attention mechanism. [101] introduced stacked attention networks (SANs), which relies on semantic representation of each question to search for relavant regions in the image. More specifically, they built multiple-layer attention mechanism, which locates the relevant region multiple times so that more accurate region of interest can be retrieved. In a similar manner, [82] attempts to locate relevant regions in the image. They map the textual queries to features from different regions by embedding them to a common space and comparing their relevance via inner product.

[98] proposed a number of improvements to dynamic memory network (DMN). Their proposed DMN+ model introduced a novel input module based on a two-level encoder with sentence reader and input fusion layer, and implemented memory based on gated recurrent units (GRU). [38] proposed focused dynamic attention (FDA) model, which exploits an object detector to determine regions of interest. LSTM is used to embed the region features and global features into common space. [99] proposed spatial memory network in which neuron activations of different spatial regions are stored in memory, and regions with high relevance are chosen depending on the question. The latter step was made possible by their novel spatial attention architecture designed to align words with patches.

Unlike most of the works mentioned above, our work does not employ any attention mechanism, yet demonstrates superior performance by fully exploiting features provided to the network.

VQA: Abstract Scenes Since drawings or illustrations possess fundamentally different characteristics from real images, simply extracting CNN features for abstract scenes does not result in performance as reliable as in real images. As such, relatively few results have been reported on abstract scene categories compared to real images.

[105] converted the questions to a tuple containing essential clues to the visual concept of the images. Each tuple (P, R, S) consists of a primary object (P), secondary object (S), and their relation (R). Mutual information was employed to determine which object corresponds to primary object and secondary object. They also augmented the dataset using crowd-sourcing in order to balance the biases in the dataset. Their visual features included histogram-like vectors for primary and secondary objects, as well as absolute and relative locations of the objects modeled by GMMs. We show that this model's performance is enhanced by addition of deep features, both holistically and regionally, and applying our DualNet further improves the performance.

[90] proposed to represent the image as a scene graph with nodes corresponding to the objects and edges representing their spatial relationship. Question is also parsed into words with their syntactic relationships. This yields the state-of-the-art results on abstract scenes category at the time of this writing.

6.3 Proposed Model: Question-Dependent Region Features

We first describe the overall workflow of our model, which won the first place in VQA Challenge 2016 in the abstract scenes category.

On top of the features described in [105], we added features from the uppermost fully-connected layer from ResNet with 152 layers, and fc7 layer of VGG with 19 layers for holistic features. We alternate between two different setups for regional features as following:

6.3.1 Average Softmax of Top Regions

We first extract 10 regions from each image using Deep Proposal [25], which proposes regions based on objectness measure and applies non-maximum suppression to filter out overlaps. We then extract the 201-dimensional features from softmax layer for each region, where softmax function is defined as below:

$$f_i(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{6.1}$$

where f_j refers to the *j*-th element in the probability distribution for all classes, and $z \in \mathbb{R}$ is some real-valued probability distribution for all classes. Dimensionality of the features correspond to 201 classes used in ILSVRC object detection task. In other words, we end up having a probability distribution for all classes. We used Fast-RCNN [26] and VGG-16 trained for the task. Finally, we average the softmax probabilities of all 10 regions to obtain a single 201-dimensional vector per image.

6.3.2 VLAD Coding of CNN with Coordinates

The general procedure is similar to the procedure employed in Section 4, except we do not employ spatial pyramid. We run selective search for each image, which returns approximately 1,000 region proposals for each image. Using Fast-RCNN, we extract fc7 features

from all regions. Dimensionality of fc7 features is reduced to 256 using PCA. We then concatenate 8-dimensional coordinate vector:

$$f_{coordinate} = (x_{min}, y_{min}, x_{max}, y_{max}, x_{center}, y_{center}, W, H) \tag{6.2}$$

as in [36] so that each region yields a 264-dimenisonal vector. Finally, we apply VLAD coding to all regions of an image with one cluster to obtain the final one 264-dimensional vector for each image.

6.3.3 Discussion

It turns out that average softmax of top regions performs better on yes/no and number categories, while VLAD coding of CNN with coordinates performs better on others category. Average softmax of top regions revolves around the most conspicuous regions only, and the features are extracted from softmax layer instead of fully-connected layer. Thus, it has more strong correspondence to classification of objects, and can yield better performances for the questions on the existence of or the number of certain objects, such as "*is there a dog?*" or "*how many dogs are there?*" which belong to yes/no and number categories. On the other hand, VLAD coding of CNN features extracts features from fully-connected layer, especially from hundreds of regions. It thus reflects more of the overall characteristics of the image, with slightly less emphasis on particular conspicuous objects. Thus, while it may be less efficient in yes/no and number categories, it turns out to be more efficient in more general questions about the image.

6.4 Experiment for VQA

6.4.1 Setting

We apply our proposed VQA model to VQA dataset [2], with emphasis on abstract scenes category. Each category consists of two sub-categories, namely multiple-choice category and open-ended choice, depending on whether the choices for the answers are provided or not. Taking advantage of each method's strength as discussed in Section 6.3.3 , we alternate between average softmax and VLAD coding, depending on the type of question, which was predicted by key phrase extraction; e.g., 'how many' indicating number category, etc. We had batch size of 400, and 500 possible answers, and set number of word embeddings for questions as 1,000. LSTM with one hidden layer of 256 hidden units was employed. Hyperbolic tangent was employed for non-linearity, as sigmoid and rectified linear unit (ReLU) resulted in slightly inferior performances, and training was performed for 100 epochs.

6.4.2 Evaluation

Evaluation was performed on the evaluation server provided by [2]. As discussed in 3.3.2, evaluation metric is a precision matching to human-provided answers. Our model's performances in both categories are shown in Table 6.1 and 6.2, along with performances from

Table 6.1: Performances of each method on open-ended category

	All	Y/N	Num	Others
all “yes”	29.15	64.9	0.22	1.67
question alone	57.19	76.88	49.55	38.79
[48]	62.56	79.06	51.57	48.94
[2]	65.02	77.45	52.54	56.41
Ours	**67.39**	**79.59**	**57.06**	**58.20**

Table 6.2: Performances of each method on multiple-choice category.

	All	Y/N	Num	Others
all “yes”	29.15	64.9	0.22	1.67
question Alone	61.41	76.9	49.65	49.19
[48]	67.99	79.08	52.57	61.99
[2]	69.21	77.46	52.90	66.65
Ours	**71.18**	**79.59**	**56.19**	**67.93**

other models. Our model demonstrates a clear superiority over other models in both categories, and in all types of questions. Indeed, using our proposed model, we took the first place in abstract scenes category, in both multiple-choice and open-ended categories, at VQA Challenge held as CVPR 2016 workshop.

(a) **Q**: Does this plane appear to be leaving or arriving? **A**: leaving

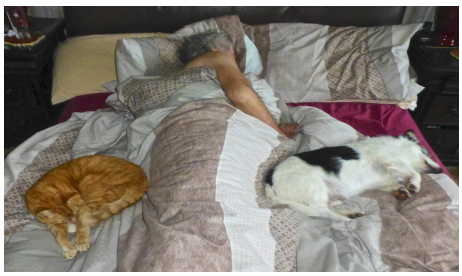
(b) **Q**: What is she wearing on her eyes? **A**: goggles

(c) **Q**: Are the animals of the same breed? **A**: yes

(d) **Q**: Does the bear look happy or sad? **A**: sad

(e) **Q**: What color is the chair in the corner? **A**: yellow

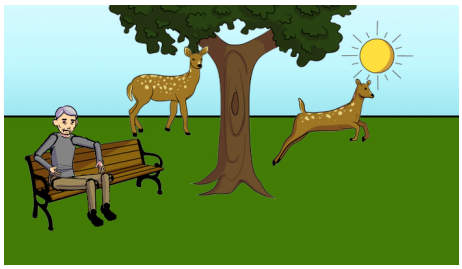
(f) **Q**: What fruit is on top? **A**: cherry



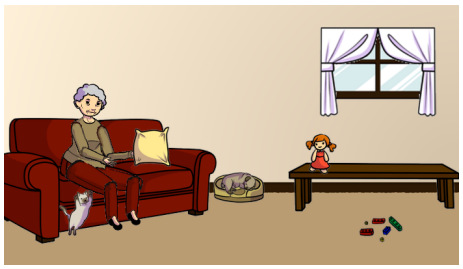
(g) **Q**: How many pets are in the bed? **A**: 2

(h) **Q**: Where is he looking? **A**: camera

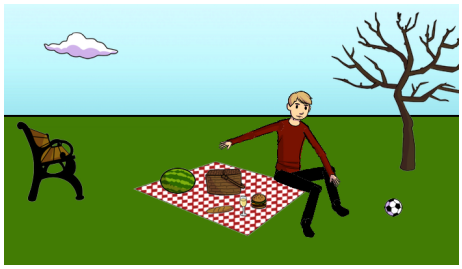
Figure 6-1: Examples of questions and generated answers in real images



(a) **Q**: How many deer are there? **A**: 2

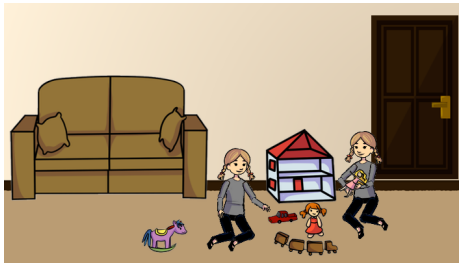


(b) **Q**: Where are the legos? **A**: floor



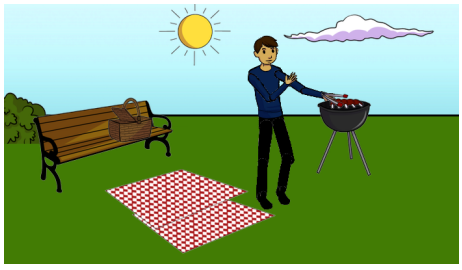
(c) **Q**: Is the man standing? **A**: no

(d) **Q**: What is the sidewalk made of? **A**: concrete



(e) **Q**: Are the girls twins? **A**: yes

(f) **Q**: What type of animal is sitting next to the chair that the girl is in? **A**: dog



(g) **Q**: What color is the man's shirt? **A**: blue

(h) **Q**: Is he playing video games? **A**: yes

Figure 6-2: Examples of questions and generated answers in abstract scenes

Chapter 7

Single Image Narrative Generation

It has been taken for granted that a single sentence of factual description suffices for single image. Yet, images frequently provide more contents than what can be described in a sentence, whether it is visual (further details or sentiments), or non-visual (inference about the image). Incorporating such elements can make image description more human-like. While visual storytelling task aims to generate story-like text from the images, it requires a sequence of images. Likewise, dense captioning generates multiple captions for single image at region-level, but it is restricted to factual description of each region and involves very expensive human annotation. We introduce a novel task of **single image narrative generation**, in which we attempt to generate multiple-sentence description from a single image that consists of both visual and non-visual elements. We note that visual question answering (VQA) datasets cover a wider range of topics than caption datasets, and exploit them by generating multiple questions about the image and collecting answers. Experimental results demonstrate that our proposed model can generate image narratives that are richer in contents and more human-like than the baseline models.

7.1 Motivation for Image Narrative Generation

Image captioning task has enticed a remarkable amount of research in recent years, thereby continuously setting up new milestones, and has now achieved a comparable performance to that of humans. However, the objective of image captioning task has almost invariably been limited to the generation of factual description of the images in a single sentence. Yet, images frequently encompass a wide variety of elements that may be difficult to capture in a single sentence. Some of those elements are *visual*, as in further details, or sentiment of the image. On the other hand, images may further stimulate the viewers to envision beyond what is visually present; for example, one may wonder why the event in the image is happening, or conjecture what happens after the event described. One may even creatively imagine and assign story-like elements to the image. Incorporating such *non-visual* elements can make the image description more human-like, and enables us to further examine the creative aspects of visual language beyond conventional captioning.

In this chapter, we propose a novel task of *single image narrative generation*, where we define *image narrative* as an image description 1) consisting of multiple sentences, and

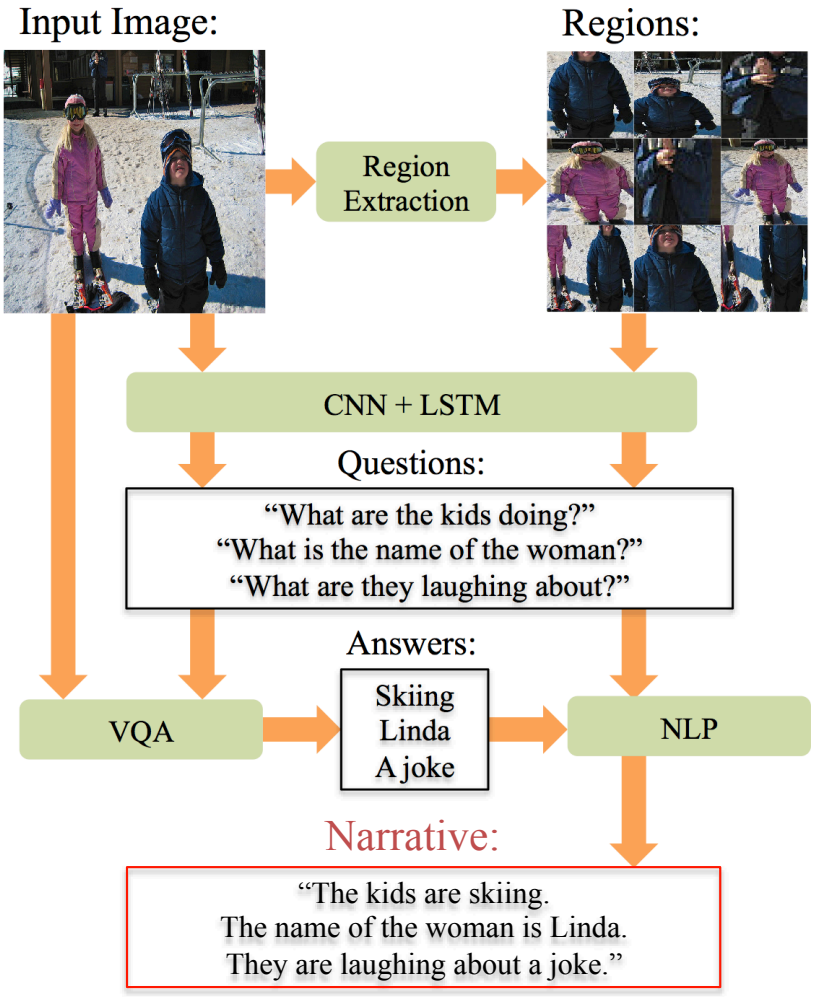


Figure 7-1: Overall workflow of our model for single image narrative generation. We generate region proposals, from which we generate questions. We then answer those questions with VQA and apply elementary NLP techniques to obtain the image narrative.

2) describing both *visual* and *non-visual* aspects. Each of these two components in our definition has been tackled individually. Recently introduced visual storytelling task [37] is parallel to the objective of our task in that it strives to generate a story-like text, but it is applied upon a sequence of images. Humans on the other hand can contemplate upon, or even imagine things with regards to a single image, and linguistically express it. Likewise, dense captioning [43] has provided a highly reliable model for obtaining descriptive details from a single image. Yet, it is restricted to factual, visually observable components, and as such, deviates from our objective of generating a narrative.

We develop a simple, yet highly competitive model to address both of the issues raised above, generating multiple captions per image, involving both visual and non-visual contents. We exploit a number of existing datasets and techniques, which are put together in a novel way to tackle the task. Namely, we employ off-the-shelf image captioning

Table 7.1: Examples of captions and questions for the same image. While captions essentially describe the same contents, questions widely vary in terms of the topics.

Image	
	
COCO Captions	**VQA Questions**
• A group of people sitting on the back • Several people are taking a ride on elephants • Some people are riding elephants in the jungle • The people are riding on the two elephants • People riding on elephants in the jungle	• Is this standard transportation in the United States? • Are they on a paved roadway? • How many people are riding elephants?

techniques, visual question answering (VQA), and elementary natural language processing (NLP) techniques as integral components of the model, along with region proposal algorithm.

In caption datasets, annotations for a single image are semantically identical in essence since they aim to describe the objective components of the image. Thus, simply generating multiple captions from single image will not lead to significant increase in contents diversity, not to mention its inability to obtain non-visual contents. On the other hand, VQA task and the relevant datasets were designed to deal not only with visually descriptive aspects, but also with aspects that require common-sense, inference, or even imagination, since their goal is to build an answering system that is robust to any type of image-relevant questions. As such, VQA task permits a much wider range of topics than captions, as long as they are image-relevant. An example of such discrepancy between caption dataset and VQA dataset is illustrated in Table 9.1. We exploit this beneficial characteristics of VQA to retrieve both visual and non-visual contents about the image.

Figure 7-1 illustrates the overall workflow of our model; we extract multiple region proposals from an image, generate questions about the regions using image captioning techniques, answer those questions with VQA, and put together the obtained contents by rule-based conversions using elementary NLP techniques. Further implementation details will be presented in Section 7.3.

Our main contributions can be summarized as following:

- introduction of a novel task of single image narrative generation,
- development of a model to efficiently tackle the task, exploiting a wide range of techniques and datasets,
- construction of auxiliary dataset and evaluation metrics that further help tackle the task.

7.2 Related Works for Image Narrative Generation

Image Captioning: The workflow of extracting image features with convolutional neural network (CNN) and generating captions with long short-term memory (LSTM) [34] has been consolidated as a standard for image captioning task. [96] presented a model inspired by statistical machine translation, which maximizes the probability of target image's description. [21] took a more linguistically oriented approach, in which they train visual detectors for words using multiple instance learning. It has also been shown that implementing attention mechanism helps boost the performance [100].

Dense Captioning: [47] generated region-level descriptions by implementing alignment model of region-level CNN and bidirectional recurrent neural network (RNN). [43] proposed DenseCap that generates multiple captions from an image at region-level. However, both works generate factual, descriptive captions at region-level, whose appearance is hardly different from captioning from the entire image. Our model covers much wider range of topics because we generate questions from datasets that were not necessarily intended to be descriptive, and as a result, frequently deviate from merely descriptive contents.

Stylistic Captioning: There have been a series of attempts at incorporating non-factual elements into image captioning. [66] and [85] inserted adjectives into the caption by training a network to predict the appropriate sentiment of the image. Apart from the adjective, however, the rest of the caption is identical as the usual image captioning, and does not incorporate any non-visual element. [37] built SIND dataset whose image descriptions display a more casual and natural tone, involving aspects that are not factual and visually apparent. While this work resembles the motivation of our research, it is also restricted by the premise of single-description-per-image, and require a sequence of images to fully construct a narrative. We on the contrary build a multiple-sentence narrative from a single image.

Visual Question Answering (VQA): Visual question answering (VQA) has escalated the interaction of language and vision to a new stage, by enabling a machine to answer a variety of questions about the image, not just describe certain aspects of the image. It is noteworthy that the questions consist not only of objective, visually verifiable questions, but also of questions that require common-sense, inference, even imagination. This characteristic enables us to obtain both visual and non-visual contents about the image.

A number of VQA datasets are available; DAQUAR [63] was the first VQA dataset to be introduced, followed by VQA dataset [2] and COCOQA [77]. VQA dataset constructed the dataset by crowdsourcing the questions and answers, while COCOQA applied rule transformations to MS COCO [60] captions. Visual7w [107], which is a subset of Visual Genome project [51], also collected the questions and answers by crowdsourcing, enforcing

questions to observe *7w*-form.

A number of different approaches have been proposed to tackle VQA task, but so far, classification approach has been shown to outperform generative approach [1, 45]. [22] proposed multimodal compact bilinear pooling to compactly combine the visual and textual features. This model is currently the state-of-the-art method at the time of this writing. [82] proposed an attention-based model to select a region from the image based on text query. It is yet arguable as of now whether visual attention is a must-have prerequisite for higher performance [45]. [62] introduced co-attention model, which not only employs visual attention, but also question attention. Finally, [97] introduced a model to extract information from general knowledge base to answer image-based questions.

7.3 Proposed Model: Region-Oriented Self Q&A

7.3.1 Region Extraction

As described in Section 7.1, we obtain contents about the image first by generating questions about the image. Since our goal is eventually to generate narrative consisting of multiple contents, it is insufficient to generate a single question, which necessitates generation of multiple questions. However, unlike in classification tasks where retrieving top k results is straightforward, it is tricky to retrieve multiple best outputs in generative models that employ recurrent networks. We deal with this problem by generating questions from multiple regions of the image. A straightforward approach to obtain regions would be to employ sliding window or spatial pyramid. However, they inherently exhibit a drawback that they generate regions without accounting for contents of the image. In order to generate regions while also accounting for the contents of the image, we employ recently proposed region extraction method.

Following [25], we first extract region candidates from the feature map of an image, by applying linear SVM trained on annotation bounding boxes at multiple scales, and applying non-maximal suppression. The region candidates then go through inverse cascade from upper, fine layer to lower, coarser layers of CNN, in order to better-localize the detected objects. This results in region proposals that are more contents-oriented than selective search [92] or Edge Boxes [56]. We first extracted top 10 regions per image. Figure 7-2 shows an example of the regions extracted in this way. See Supplemental Material for comparison to other basic region extraction models. In the experiments to follow, we set the number of region proposals K as 5, since the region proposals beyond top 5 tended to be less congruent, thus generating less relevant questions. It also reflects the consideration of trade-off between legibility and the amount of contents to provide; while we may indefinitely add up new questions by increasing the number of region proposals, the incremental informativeness of additional contents decreases, as do the legibility and accuracy of the narrative.

7.3.2 Image Feature Generation

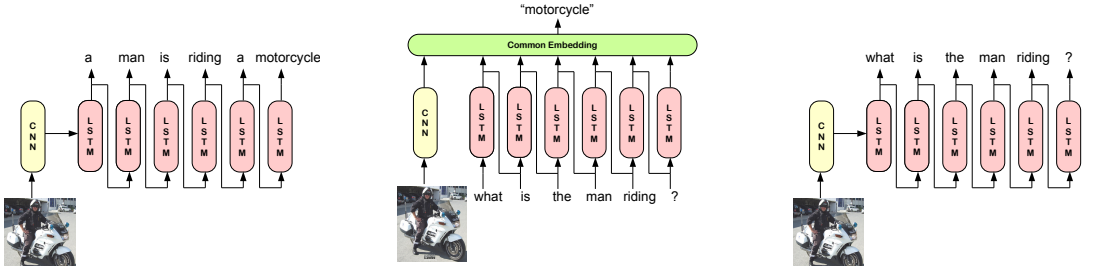
Conventional approach has been to extract the image features from the uppermost fully-connected layer or pooling layer to train an image captioning module or VQA module.

Figure 7-2: Example of regions extracted from the image, and the questions generated from each region. The color of the question corresponds to that of boundary around each region, from which the question was generated.

However, our efforts so far have been for including the elements that are difficult to capture such conventional CNN features, and we have demonstrated our proposed models' superior performances in previous chapters. We thus substitute CNN features for our own proposed image features by combining the models introduced in Chapter 4 and Chapter 5.

For object part, procedure in Chapter 4 with the best-performing combination of hyper-parameters is chosen. More specifically, fc7 features from VGG [86] are extracted for all region proposals suggested by selective search [92], and PCA is applied to reduce the dimensionality from 4,096 to 256. Then, k-means++ [5] is applied to learn the codewords with 4 as the number of clusters, from which VLAD coding [40] is applied. The same procedure is applied to one 1×1 and 4 2×2 grids, summing up to 5,120-dimensional vector per image.

For sentiment part, procedure in Chapter 5 is replicated. We extract fc7 features from fine-tuned sentiment CNN, and apply PCA. Instead of 1,685-dimensional vector as described in Chapter 5, we further reduce it to 1,024-dimensional vector to prevent it from interfering with object features. We do not apply selective search, VLAD coding, or spatial pyramid for sentiment features, since sentiment is generally present throughout the image,



(a) Image captioning. (b) Visual question answering. (c) Visual question generation.

Figure 7-3: Illustration of the overall workflow for each task.

rather than being locally variant. Combining the object features and sentiment features, we end up having 6,144-dimensional vector per image. For the rest of this thesis, we will refer this to image features as our proposed image representation, unless otherwise noted.

7.3.3 Visual Question Generation

Key point in visual question generation is the ability to generate image-relevant questions. The question should directly address the components of the image, whether the answer to it is visually verifiable or not. We achieve this by utilizing existing techniques and datasets in a novel way.

In image captioning task, it is conventional to train an LSTM with human-written captions as ground truth annotations. On the other hand, in VQA task, questions are frequently inserted to LSTM in series with fixed image features, and the answers to the questions become the ground truth labels to be classified. Instead, we replace the human-written captions with human-written questions, so that the LSTM is trained to predict the question, rather than caption. See Figure 7-3 for the illustration of the workflow for each task.

Note that, while questions are *generated* from both the image in its entirety and the region proposals from the previous step, the training procedure will proceed only with the image in its entirety, under a premise that even the region-level questions pertain to the whole image as long as they can relate. Thus, given an image I, a question $Q = (q_0,...q_N)$, the training proceeds as [96]:

$$\begin{aligned} x_{-1} &= CNN(I) \\ x_t &= W_e q_t \\ p_{t+1} &= LSTM(x_t) \end{aligned} \tag{7.1}$$

where W_e is a word embedding, x_t is the input features to LSTM at t, and p_{t+1} is the resulting probability distribution for the entire dictionary at t. In the actual generation of captions, it will be performed over all region proposals $r_0,...,r_N \in I$:

$$\begin{aligned} x_{-1} &= CNN(r_i) \\ x_t &= W_e q_{t-1} \\ q_t &= \text{argmax}_{q \in p} p_{t+1} \\ &= \text{argmax} LSTM(x_t) \end{aligned} \tag{7.2}$$

for q_0,...$q_N \in Q_{\mathrm{r_i}}$. We implement an off-the-shelf LSTM in which each gate is computed as in equation (4.4) reprinted below for convenience:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \end{aligned} \tag{7.3}$$

We trained LSTM for 50 epochs with batch size of 128. Our proposed image representation as described above was employed as image features. Beam size of 2 was used to generate questions. Figure 7-2 shows examples of questions generated from each region including the question generated from the entire image. As shown in the figure, by focusing on different regions and extracting different image features, we can generate multiple image-relevant questions from single image. Generated questions, starting from the one generated from the entire image and then in the order of their score, proceed to the next stage, except duplicate questions are eliminated.

A question may arise as to why not to directly generate captions from the regions. The answer to this question has to do with the range of contents generated. As discussed in Section 7.1, most ground truth annotations in caption datasets were written with the objective of depicting the factual components of the image, and it is thus hard to generate anything beyond such characteristics. On the other hand, the human-written questions in VQA datasets are not restricted by such rigid objective, as long as they are relevant to the image. They greatly vary in terms of the topic even for the same image, whereas in captions datasets all ground truth annotations per image essentially describe the identical contents. Thus, a larger amount of flexibility in terms of contents can be expected. In fact, as we will see in Chapter 9, generating image narrative by directly generating multiple captions from region proposals results not only in less interesting narratives, but also in less accuracy.

Table 7.2: Statistics from the crowd-sourcing task.

# of answers collected	48,090
# of unique answers	15,469
# of workers participated	187
max. # assignments by worker	1609
avg. # assignments per worker	51.43
rewards per assignment	$.10
10 most common answers	'*yes*','*no*','*tom*', '*london*','*mine*', '*downtown*','*john*','*me*' '*halloween*','*new york*'

So far, we were concerned with generating "visual" questions. We now seek to generate "non-visual" questions. As was described in Section 7.2, [68] generated questions that a human may naturally ask and require common-sense and inference. We examine whether we can train a network to ask multiple questions of such type by visual cues. If such is possible, it would further enrich the contents of generated narrative. We thus replicated the

Table 7.3: Examples of answers collected on VQG.

Question	Answer
'*What is the name of the man?*'	'*Tom*'
'*What is the score in the game?*'	'*0-0*'
'*What kind of record is being played?*'	'*rap records*'
'*How long until the bathroom is fixed?*'	'*1 week*'
'*Why is he making weird face?*'	'*he's drunk*'
'*What's the cat's name?*'	'*Moni*'
'*How much did that cost?*'	'*10 dollars*'
'*What destroyed this town?*'	'*bomb*'
'*Why are the trees lit up?*'	'*It's Christmas time*'
'*What are the ingredients?*'	'*fish,bread,broccoli*'

image captioning process described above, with 10,000 images of MS COCO and Flickr segments of VQG dataset, with 5 questions per image as the annotations. Examples of questions generated by training the network solely with non-visual questions are shown in Table 7.4. In Chapter 9, these non-visual questions will be put into effect in the form of data augmentation.

7.3.4 Visual Question Answering

Since we generated questions from images, we should now answer those questions. We train the question answering with VQA dataset [2]. So far, classification approach has been shown to outperform generative approach in VQA task [45], and we also implement our VQA network as a classification task. Question words are sequentially encoded by LSTM as one-hot vector. Our proposed image representation has been employed for image features. Hyperbolic tangent non-linearity activation was employed, and element-wise multiplication was used to fuse the image and word features, from which softmax classifies the final "label" as the answer for visual question. We set the number of answers to 1,250. Alternatively, Visual7w dataset [107] can be used to train the network, but due to its rigid design rule of enforcing 7-w questions, training with this dataset resulted in less diverse and interesting questions of the form "*what is X?*" which is present for every image in the dataset.

While some papers have reported performance boost with implementation of attention mechanism, it is arguable to what extent it contributes [45], and some of the previous works [79] have reported superior performance without using attention mechanism. Indeed, the result from our model demonstrates that it can answer the questions dealing with regions successfully without the help of attention mechanism. Yet, it is still a factor for potential future improvement of our model's overall performance, which would remain as our research interest.

As we augmented the training data to generate "visual" and "non-visual" questions, we now also need to train the network to "answer" those non-visual answers. Since *et al.*[68] was concerned with generating questions only, however, the dataset provides the questions only. We thus collected the answers to these questions on Amazon Mechanical Turk. Since

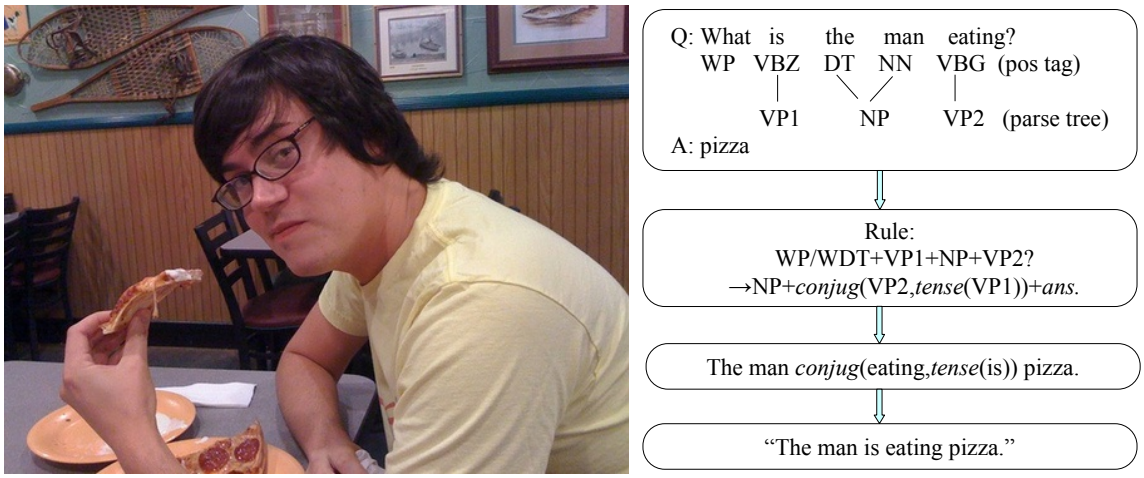


Figure 7-4: Example of question and answer converted to a declarative sentence by conversion rule.

many of these questions cannot be answered without specific knowledge beyond what is seen in the image (*e.g.* "*what is the name of the dog?*"), we encouraged the workers to use their imagination, but required them to come up with answers that an average person might also think of. For example, people answered to the question "*what is the name of the man?*" with "*John*" or "*Tom*." Such non-visual elements add vividness and story-like characteristics to the narrative as long as they are compatible with the image, even if not entirely verifiable. Table 7.2 shows the statistics for the crowd-sourcing task, and Table 7.3 shows examples from our collected answers. Dataset with the answers we collected is publicly available.

7.3.5 Natural Language Processing

At this stage, we are given multiple pairs of questions and answers about the image. By design of the VQA datasets, most of which comprise simple questions regarding only one aspect and the answers mostly being single words, the grammatical structure of most questions and answers can be reduced to a manageable pool of patterns. Exploiting these design characteristics, we can combine the obtained pairs of questions and answers to a declarative sentence by application of rule-based transformations, as was performed in [77, 84].

To briefly illustrate the conversion process, we first rephrase the question to a declarative sentence by switching the word positions, and then insert the answers to its appropriate position, mostly replacing *wh*-words. For example, a question "*What is the man holding?*" is first converted to a declarative statement "*The man is holding what*" and the corresponding answer "*frisbee*" replaces "*what*" to make "*The man is holding frisbee*." Part-of-speech tags with limited usage of parse tree were used to guide the process, particularly conjugation according to tense and plurality. Figure 7-4 illustrates the workflow of converting question and answer to a declarative sentence. See Table 7.5 for specific conversion rules. Part-of-speech tag notation is as used in PennTree I Tags [65].

Alternatively, we may be able to perform preprocessing in the same way by converting the training data of questions and answers into declarative sentences beforehand and gen-

erating captions from regions, but it would require us to process the entire training dataset beforehand. Another alternative would be to use full-sentence VQA datasets, such as [84], but the overall performance is reported to decline, due to exponentially increased number of labels.

Another advantage of performing sentence transformation later in the stage is that we can utilize the beneficial property of VQA task. It is generally treated as a classification problem, which turns out to be more reliable than generative approach [45]. Thus, although the questions are *generated*, VQA can retrieve more accurate answers by performing classification, as long as the questions are relevant to the image.

7.4 Conclusion & Future Work

We introduced a novel task of single image narrative generation, and proposed a model to effectively tackle the task, by utilizing region proposal, image captioning, visual question answering, and natural language processing techniques, with mostly existing datasets and some auxiliary dataset collected anew. Our model demonstrated superior performance to baseline models. Since our model's performance is contingent on the performance of image captioning and VQA techniques, replacing each module with more advanced ones, instead of off-the-shelf models as in our work, is highly likely to boost the overall performance of our model.

We would finally like to outline three aspects that must be accompanied for further improvements in single image narrative generation task; namely, dataset, evaluation metric, and causality. It must be admitted that applying natural language processing techniques is a heuristic approach to overcome the current insufficiency of dataset comprising narratives for single images. Building such dataset will be a part of our subsequent research. Although we relied on human evaluation for its reliability, development of an automatic evaluation metric to deal with the plausibility of the generated image narrative must also follow. Lastly, our current narrative consists of contents that are not structurally organized. Eventually, more structure-oriented narrative involving causality would be desirable, and will be our subsequent research topic.

Table 7.4: Examples of questions generated using non-visual questions in VQG dataset.

Image	Generated Questions
	• What is the name of the player? • What is he speaking about? • What is the score of the match? • Is this costume for a race? • Has he worked there?
	• Do you think the boy can win the prize? • Was this for a charity event? • Did they have a child? • What is she looking at? • What are they waiting for? • Who is that guy? • What is he looking at?
	• Is this a hotel room? • Is that a picture of your house? • Where did you get those pillows? • Is this new tile? • Was that clean there? • How big is that room?
	• Is the woman drunk? • Is this a church? • Is this structure in a museum? • What city was this in? • Are they protesting?
	• What kind of pizza is that? • Is it for dinner? • What kind of topping is this on the pizza? • What does the plate say?
	• What is this bird staring at? • How long will it be there? • Is that a real bird? • What sort of bird is that? • What kind of flower is that?

Table 7.5: Conversion rules for transforming question and answer pairs to declarative sentences.

Type	Rule (Q→A)	Question	Ans.	Converted Ans.
yes/no	VB1+NP+VB2/JJ?	–	–	–
	→NP+*conjug*	-	-	
	(VB2/JJ,*tense*(VB1))	*Did he get hurt?*	*yes*	*He got hurt.*
	or, NP	-		
	+*negate*(*conjug*	-	-	
	(VB2/JJ,*tense*(VB1)))	*Is she happy?*	*no*	*She is not happy.*
	MD+ NP+VB?	–	–	–
	→NP+MD+VB **or**,	*Will the boy fall asleep?*	*yes*	*The boy will fall asleep.*
	NP+*negate*(MD)+VB	*May he cross the road?*	*no*	*He may not cross the road.*
number	"*How many*"+NP+	-	-	
	/is/are+EX?	-	-	
	→EX+*is/are*+*ans*+NP	*How many pens are there?*	*2*	*There are 2 pens.*
	"*How many*"+NP1(+MD)	-	-	
	+VB(+NP2)?	–	–	–
	→*ans*(+MD)+VB(+NP2)	*How many people are walking?*	*3*	*3 people are walking.*
	"*How many*"+NP1+	-	-	
	VB1/MD+NP2+VB2?	–	–	–
	→NP2	-	-	
	+(MD+VB2)/*conjug*	-	-	
	(VB2,*tense*(VB1))	-	-	
	+*ans*+NP1	*How many pens does he have?*	*4*	*He has 4 pens.*
others	WP/WRB/WDT+	-	-	
	"*is/are*"+NP?	-	-	
	→ NP+"*is/are*"+*ans.*	*Who are they?*	*students*	*They are students.*
	WP+NP+VP? → *ans.*+VP	*What food is on the table?*	*apple*	*Apple is on the table.*
	WDT+NP+VP(+NP2)?	-	-	
	→*ans.*(+NP)+VP(+NP2)	*Which hand is holding it?*	*left*	*Left hand is holding it.*
	WP/WDT+MD+VB?	-		
	→*ans.*+MD+VB	*Who would like this?*	*dog*	*Dog would like this.*
	WP/WDT+MD+NP+VB?	-	-	
	→NP+MD+VB+*ans.*	*What would the man eat?*	*apple*	*The man would eat apple.*
	WP/WDT+VP(+NP)?	-	-	
	→*ans.*+VP(+NP)	*Who threw the ball?*	*pitcher*	*Pitcher threw the ball.*
	WP/WDT+VB1+NP+VB2?	–	–	–
	→NP+*conjug*	-	-	
	(VB2,*tense*(VB1))+*ans.*	*What is the man eating?*	*apple*	*The man is eating apple.*

Chapter 8

Interactive Image Narrative Generation

We now extend the image narrative generation task described in Chapter 7 to interactive environment, in which users participate in the narrative generation process by answering questions about the image, and the generated narrative varies depending on the provided user input. We hypothesize that implementing such environment enables us to learn the user's respective interest, which can be utilized later for automatic reflection of specific interests. We first discuss the motivation for interactive narrative generation, and proceed to describe how to integrate it into our current pipeline of image narrative generation, particularly with focus on generating visual questions to be answered by the user.

8.1 Motivation for Interactive Narrative Generation

So far, we have taken a broad assumption that it suffices that an identical description be derived from a single image. Such assumption would indeed be safe if holistic perspective upon seeing the images is assumed [74]. That is, when we consider the image as a whole in its entirety, there hardly remains any room for the viewer's subjectivity or interpretation. For example, annotations for classification of the image would hardly vary among different viewers. As such, there would be little overall disagreement as to the contents of the image description, except for minor details. On the other hand, humans can frequently pay attention to different parts of the image depending on their interest or personal tendency. For example, given an image of various types of fruits, one may be more interested in apples whereas another one may pay attention to watermelons. In order for such diversity to exist, there should be a certain context as a catalyst that allows the user to pay attention to different parts rather than the image as a whole. Figure 8-1 shows an example of how attention varies depending on the context provided by the questions. Such context needs not only to have a possibility for various attentions or interpretations, but also to be able to reflect the user's distinct interest. If such context is provided, it would be possible to learn the user's interest, or to which aspect the user tends to attend to. Learning of such personal tendencies would further enable us to customize and optimize a variety of tasks, including narrative generation, to meet the user's specific interest. Throughout this chapter, we refer to the user's interest as a certain tendency that each user exhibits when confronted with the image and the questions (that provide the context) that allow for multiple interpretations or

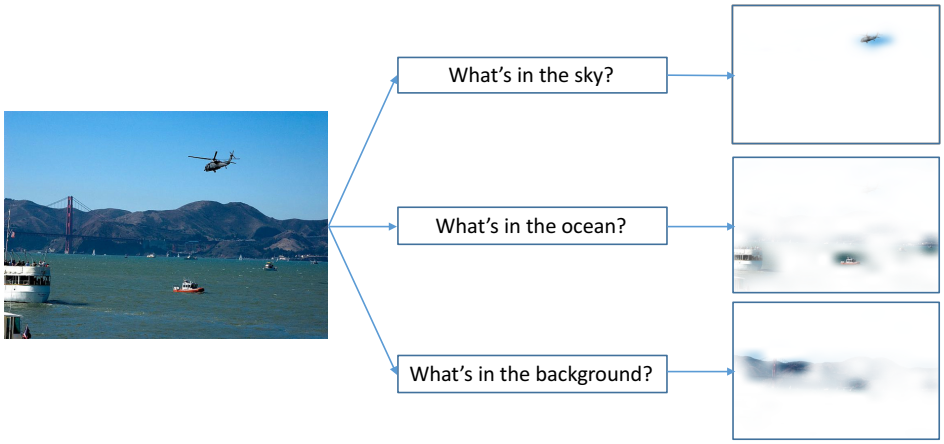


Figure 8-1: Viewer's attention varies depending on the context provided.

(a) before learning user's interest (b) after learning user's interest

Figure 8-2: Example of how learning of user's interest can be applied..

answers, mostly by enabling attention to multiple regions within the image.

Under our current proposed pipeline of generating image narrative from single image as described in Chapter 7, one of the most critical modules is to generate multiple visual questions about the image, and to answer them. In the fully automated setting, the generation of visual question and answers was performed over multiple regions of the image, so that contents diversity could be obtained. Selection of the region to discuss was automatically determined based on objectness measure using region proposal algorithm. In an interactive setting, however, we assign a portion of answering part to the users, so that the region selection itself reflects the user's choice, from which content diversity is derived. Since different regions can be selected from different users, instead of automatic selection of regions, we can now generate multiple image narratives from single images.

A question may arise as to why not to simply ask the users to select the region or part of the image that stands out the most to them. In such case, there would be no need to *generate* the questions for each image, as the question '*what stands out the most?*' would suffice for all images. This, however, would be equivalent to a simple saliency annotation task, and would not allow for any meaningful customization or optimization per user. Thus,

as discussed above, generating a question for each image is intended to provide a context in which each user can apply their own specific interest.

Apart from simply generating diverse image narratives based on the user input, many potential applications can be conceived of. For example, in cases where thorough description of an entire scene results in a redundant amount of information both quality and quantity-wise, application of our model can be applied to describe just the aspect that meets the user's interest that was learned. In its overall mechanism and purpose, it can be said to be roughly analogous to the automatic recommendation system in an online shopping environment or automatic sub-topic extraction and summary from large text corpora. See Figure 8-2 for an illustration of a potential application of our model.

8.2 Related Works for Interactive Narrative Generation

Most previous works on image description generation tasks focus on fully automated generation, and it is only recently at the point of writing this that tasks involving user interaction started to appear. Most representative, and the closest to our work in its spirit, is Visual Dialog [16], in that it actively involves user interaction, which in turn affects the responses generated by the system. Its core mechanism, however, is technically an inversion of our model, where the users ask the questions about the image, and the system answers them. Thus, the focus is on extending the VQA system to a more context-dependent, robust, and interactive direction. On the other hand, our model's focus is on generating customized image descriptions, and user interaction is employed to learn the user's interest, whereas Visual Dialog is not concerned about the users themselves, and the user interaction is purely for the sake of providing questions to the system.

8.3 Proposed Model: Visual Question Generation for User Interaction

8.3.1 Applying User Interaction within the Same Images

As discussed in the earlier part of this chapter, the key component of image narrative generation in an interactive setting is to generate a visual question that provides a context for analytic perception of the image, and allows for attentions to various parts that reflect the user's interest. The foremost prerequisite for the interactive questions to perform that functionality is the possibility of various answers or interpretations. In other words, a question whose answer is so obvious that it can be answered in an identical way unanimously would not be valid as an interactive question. In order to make sure that each generated question allows for multiple possible answers, we internally utilize the VQA module. More specifically, the question generated by the VQG module is passed on to VQA module, where the probability distribution p_{ans} for all candidate answers C is determined. If the most likely candidate $c_i = \arg\max p_{ans}$, where $c_i \in C$, has a probability of being answer over a certain threshold α, then the question is considered to have a single obvious answer, and is thus considered ineligible. The next question generated by VQG is passed on to VQA to repeat

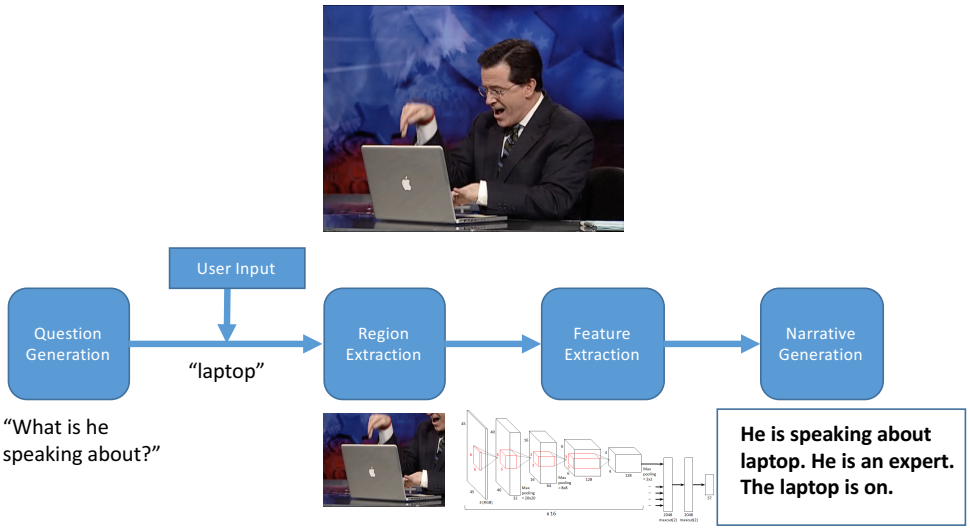


Figure 8-3: Overall workflow of interactive image narrative generation.

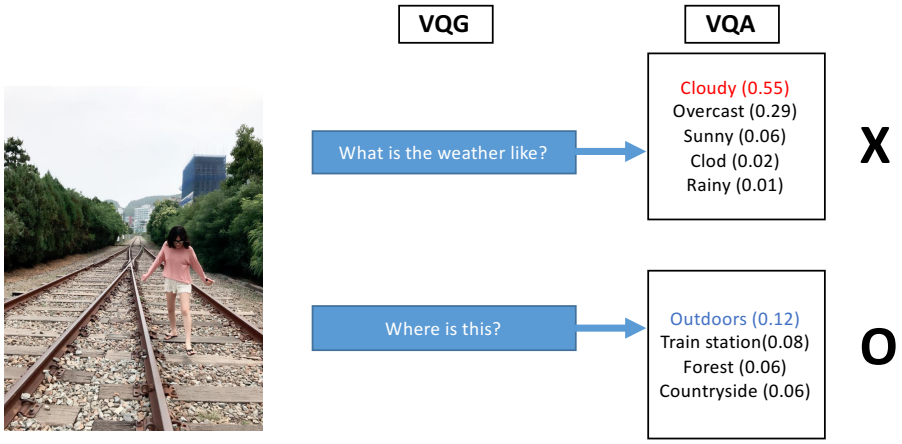


Figure 8-4: Examples of valid and invalid visual questions for interaction.

the same process until the the following requirement is met:

$$c_i < \alpha, c_i = \arg\max p_{ans} \tag{8.1}$$

In our experiments, we set α as 0.33. Figure 8-4 illustrates an example of a question where the most likely answer had a probability distribution over the threshold (and is thus ineligible), and another question whose probability distribution over the candidate answers was more evenly distributed (and is thus eligible).

Once the visual question that allows for multiple responses is generated, a user inputs his answer to the question, which is assumed to reflect his interest. We now need to extract a region within the image that corresponds to the user's response. We slightly modify the attention networks introduced in [101] in order to obtain the coordinates of the region that correspond to the user response. In [101], the question itself was fed into the network, so that the region necessary to answer that question is "*attended to*." On the other hand, our task requires that we attend to the region indicated by the user's response instead of

the question. We make a very simple yet efficient modification, in which we replace the *wh-* question terms (also including other frequent inquisitive terms such as "*how many?*") with the response provided by the user. For example, a question "what is on the table?" with a user response "pizza" will be converted to a phrase "pizza is on the table," which is fed into attention network just as a question would. This is similar to the rule-based NLP conversion as employed in Chapter 7, but much simpler. We obtain the coordinates of the region from the second attention layer, by obtaining minimum and maximum values for *x*-axis and *y*-axis in which the attention layer reacts to the input phrase.

Since the regions tightly contain objects at a large scale, they have a risk of being misclassified. We thus extract the regions allowing slightly larger space than indicated by the coordinates. Specifically, a region $r_{i,j,k}$ of size (w_r, h_r) with coordinates $x_0, y_0, x_{max}, y_{max}$ for image *i* of size (W, H) with a question *j* answered by the user *k* is to be extracted as r' with a magnifying factor $0 < \alpha < 1$, resulting in the following coordinates:

$$\begin{aligned} r'_{i,j,k} = (&\max(0, x_0 - w_r\alpha), \max(0, y_0 - h_r\alpha), \\ &\min(W, x_{max} + w_r\alpha), \min(H, y_{max} + h_r\alpha)) \end{aligned} \tag{8.2}$$

where we set α as 0.25. After the region corresponding to the user input is extracted, we now need to extract features from that region. Note that, since we are now given the *local* regions that hardly encompass further sub-local elements, we follow the conventional feature extraction of generic CNN features, instead of local features as described in Chapter 4. We also bypass the extraction of sentiment features for similar reason. We extracted 4096-dimensional fc7 layer features from VGG-19 [86] for each region extracted.

Given the region and its features, we can now apply the single image narrative generation process described in Chapter 7 with minor modifications in setting. Regions are extracted, visual questions are generated and answered, and rule-based natural language processing techniques are applied to organize them. Since now there are few local elements within the image, we reduce the number of regions *K* to be extracted to 3. Figure 8-3 shows an overall workflow of our model for interactive image narrative generation.

8.3.2 Applying User Interaction to New Images

Once we implemented interactive environment, and collected data of user choices, we can now represent each instance of image, question, and user choice as a triplet consisting of image feature, question feature, and the label vector for the user's answer. In addition, collecting multiple choices from identical users enables us to represent any two instances by the same user as an order pair of triplets, assuming source-target relation. With these pairs of triplets, we can train the system to predict a user's choice on a new image and a new question, given the same user's choice on the previous image and its associated question. Starting with the user's previous choice, we obtain image feature x_{img_i} via spatial pyramid VLAD coding of CNN features (Chapter 4) and sentiment network (Chapter 5), as described in Chapter 7. Question feature x_{q_i} for question q_i is obtained via LSTM,

$$x_{q_i} = LSTM(w_0, w_1, w_2, ..., w_{|q_i|}) \tag{8.3}$$

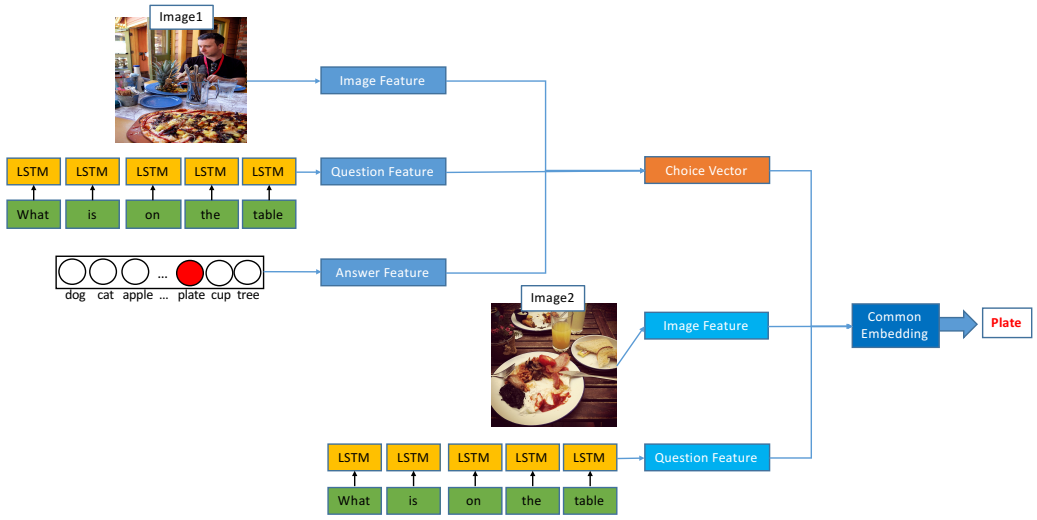


Figure 8-5: Training with pair of choices made by the same user upon being asked specific questions about the images. In the figure, given the choice vector for image 1 and new image feature and question feature for image 2, it is trained to predict the answer for the question on image 2.

where w_i is a one-hot vector corresponding to each word in the question, and the user's choice x_{ans_i} is also represented as one-hot vector where the size of the vector is equal to the number of possible choices. We refer to the fused feature representation of this triplet consisting of image, question, and the user's choice as **choice vector**.

Now, we can represent the image feature x_{img_j} and question feature x_{q_j} for the second triplet in the same way as above, and project it onto the same embedding space as the choice vector. We can now train a softmax classification task in which the feature from the common embedding space predicts the user's choice x_{ans_j}. We set the batch size as 400 and hyperbolic tangent was employed as non-linearity. Figure 8-5 shows the overall workflow for training.

Prediction can now be made with a new user's choice. Suppose we are given a new triplet $x_k = (x_{img_k}, x_{q_k}, x_{ans_k})$ of image, question, and a choice from a new user, which corresponds to our understanding of the user's preference. Now, by projecting the choice vector along with the image feature x_{img_l} for a new image, and the question feature x_{q_l} for the new question for which we would like to predict an answer, we can predict what choice the user (who made the choice as given in choice vector) would make, with our prediction being the outcome label of the classification. In short, we postulate that the answer with index u, which maximizes the probability calculated by LSTM, is to be chosen as x_{ans_l} by the user who chose x_{ans_k}, upon seeing a tuple (x_{img_l}, x_{q_l}) of new image and new question:

$$u = \arg\max_v P(v; c_k, x_{img_l}, x_{q_l}) \tag{8.4}$$

where P is a probability distribution determined by softmax over the space of possible choices, and c_k is the choice vector corresponding to $(x_{img_k}, x_{q_k}, x_{ans_k})$. This overall procedure and structure are essentially identical as in VQA task, except we augment the feature space to include choice vector, on top of image feature and question feature.

8.4 Conclusion & Future Work

We introduced a novel task of interactive image narrative generation task, where image narrative generation is not fully automated, but works with the aid from use interaction, leading to variations in outcomes. We proposed a parallel architecture of VQG and VQA that generates visual questions allowing for multiple responses. Such questions allow diversity in user responses, and we extract regions corresponding to the user response by modifying query vector to feed into attention network. Since we end up with different regions per response, even with the same image and same question, we end up having different image narratives from single images. As we will see in the following chapter, such interactive environment not only leads to diversity in generated descriptions, but also enables us to learn and customize to the users involved in the generation.

Chapter 9

Experiments

9.1 Experiments for Single Image Narrative Generation

In this section, we first perform and evaluate our model for single image narrative generation as described in Chapter 7. Note that user interaction is not involved in the generation process at this stage yet.

9.1.1 Setting

We applied the model described in Section 7.3 to 40,775 images in test 2014 split of MS COCO [60]. We compare our proposed model to three baselines as following:

Baseline 1 (COCO): general captioning trained on MS COCO applied to both images in their entireties and the region proposals

Baseline 2 (SIND): captions with model trained on MS SIND dataset [37], applied to both images in their entireties and the region proposals

Baseline 3 (DenseCap): captions generated by DenseCap [43] at both the whole images and regions with top 5 scores using their own region extraction implementation.

For all baseline models, we added a constraint to prevent duplicate captions for the same image for fair comparison. Note that, while baseline models used region proposals to generate region-level captions, our model used region proposals only to generate questions, and the answering stage was performed with the entire image, as is conventional in VQA task. This again shows the comparative advantage of applying VQA to generate multiple sentences; once presented with questions, it no longer has to look at regions to generate multiple contents about the image. It generates multiple contents looking only at the entire image, lessening the prospect of obtaining irrelevant descriptions, which frequently happens with region-level captioning, as we will see in this section.

9.1.2 Evaluation

Automatic Evaluation

Although the dataset employed in this experiment [60] comes with ground truth captions and ground truth answers for the questions, it does not contain ground truth annotations

for image narratives. Even if it did, the difficulty of automatically evaluating image narratives would still persist due to the unique characteristics of image narrative generation task. In image captioning task, upon which automatic evaluation is usually performed, both the length and semantics of the "correct" answers are generally in a highly fixed range, mostly single sentence describing the main event in an objective way, thus making the automatic evaluation convenient. Likewise, in machine translation task, the length of the translation is generally proportional to the length of the source language, and the permissible range of semantics is strictly limited. On the other hand, image narrative consists of multiple sentences of much wider range of possible contents whose order of appearance mostly does not have to be fixed. Thus, resembling human-written image narrative does not necessarily imply itself that such image narrative is better than the one that does not resemble human-written image narrative. In other words, reliability of the automatic evaluation is diminished.

Even so, however, being able to take a look at human-written image narratives can still provide us with important insights. Not only can we perform automatic evaluation for reference, but we can also have a comprehension of what characteristics would be shown in actual human-written image narratives. It also enables us to examine image narrative generation via training with human-written image narratives, which can be a meaningful comparison to our proposed model, as we will see later in this chapter. As such, we collected image narratives for a subset of MS COCO dataset. We asked the workers to write a 5-sentence narrative about the image in a story-like way. We made it clear that the description can involve not only factual description of the main event, but also local elements, sentiments, inference, imagination, etc., provided that it can relate to the visual elements shown in the image. Table 8.1 shows examples of actual human-written image narratives collected. Table 8.2 shows statistics for collected human-written image narratives. Examining the actual human-written image narratives displays a number of intriguing remarks. On top of the elements and styles we asked for, the participants actively employed many other elements encompassing humor, question, suggestion, etc. in a highly creative way. It is also clear by looking at the human-written image narratives that conventional captioning alone will not be able to capture or mimic the semantic diversity present in them.

We performed automatic evaluation with popular metrics [71, 18, 94] on a subset of MS COCO with collected image narratives as ground truth annotations. Table **??** shows the results. While resemblance to human-written image narratives may not necessarily guarantee better qualities, our model, along with DenseCap, showed highest resemblance to human-written image narratives. As we will see in human evaluation, these two models turn out to be most reliable models, and such result is consistent here, suggesting that resemblance to human-written image narratives may indeed provide a meaningful reference.

Human Evaluation

Although we collected human-written image narratives for a subset of the dataset and performed automatic evaluations above, it is questionable whether more resemblance to the human-written image narratives has direct correlation to higher integrity of the generated image narrative, since image narratives deal with much longer texts with a much wider range of possible contents as we discussed. Few previous works have attempted to tackle

the task of narrative evaluation. [69] proposed Story Clonze Evaluator, which evaluates vector representation beyond textual similarity and attempts to reflect context to predict what should happen next. However, it works under a premise that the narrative is highly structural with clear causality, which is yet to be present in our task. More importantly, it does not take visual information into consideration. As such, we resort to crowd-sourcing for evaluation of the models.

We asked the workers to rate each model's narrative with 5 metrics that we find essential in evaluating narratives; *Diversity*, *Interestingness*, *Accuracy*, *Naturalness*, and *Expressivity* (DIANE). *Diversity* deals with the coverage of diction and contents in the narrative, roughly corresponding to *recall*. *Interestingness* measures the extent to which the contents of the narrative grasp the user's attention. *Accuracy* measures the degree to which the description is relevant to the image, corresponding to *precision*. Contents that are not visually verifiable are considered accurate only if they are compatible with salient parts of the image. For example, a description with a name of a man is considered compatible if there is a salient man in the image, but incompatible if there are no humans in the image, or the salient agent is female, or there is a crowd of people with none of them standing out. *Naturalness* refers to the narrative's overall resemblance to human-written text or human-spoken dialogue. *Expressivity* deals with the range of syntax and tones in the narrative. Evaluation was performed for 5,000 images with 2 workers per image, and all metrics were rated in the scale of 1 to 5 with 5 being the best performance in each metric. We asked each worker to rate all 4 models for the image on all metrics.

Table 9.6 shows example narratives from each model. See Supplemental Material for more examples. Table 9.4 shows the performance of each model on the evaluation metrics with mean and standard deviation, along with the percentage of each model receiving the highest score for a given image, including par with other models. Our model obtained the highest score on *Diversity*, *Interestingness* and *Expressivity*, along with the highest overall score and the highest percentage of receiving best scores. In all other metrics, our model was the second highest, closely trailing the models with highest scores. Note that standard deviation is high due to each worker rating with their own range. We thus performed additional test to confirm the significance of the results. Table 9.5 shows our model's performance against each baseline model, in terms of the counts of wins, losses, and pars. χ^2 values on 2 degrees of freedom are evaluated against the null hypothesis that all models are equally preferred. The rightmost column in Table 9.5 corresponds to the one-sided p-values obtained from binomial probability against the same null hypothesis. Both significance tests provide an evidence that our model is clearly preferred over others.

Closely inspecting the results on the evaluation reveals interesting characteristics of each model. General image captioning trained on MS COCO shows weaknesses in accuracy and expressivity. Lower score in accuracy is presumably due to quick diversion from the image contents as it generates captions directly from regions. In fact, examples in Table 9.6 show that the contents of the captions start to deviate in the later part of the narrative. This reveals one of the limitations of general image captioning described in Section 7.1. Since it is restricted by an objective of describing the entire image, it frequently generates irrelevant description on images whose characteristics differ from typical COCO images, such as regions within an image as in our case. On the contrary, VQA questions provide a much wider range of acceptable topic contents, as long as they can be related to

the image. For example, given a region of a sky with nothing else, our model finds it sufficient to ask "*what is the weather like*?" whereas the captioning models struggle to generate a full description with concrete subjects, frequently resulting in irrelevant captions as in "*a bathroom with a sink and a mirror*."

Story-like captioning trained on MS SIND obtained the lowest scores in all metrics. In fact, examples in Table 9.6 also display that the narratives from this model are almost completely irrelevant to the corresponding images, since the correlation between single particular image and assigned caption is very low. DenseCap turns out to be the most competitive among the baseline models. It demonstrates the highest accuracy among all models, but shows weaknesses in interestingness and expressivity, due to their invariant tone and design objective of factual description. Our model, highly ranked in all metrics, demonstrates superiority in many indispensable aspects of narrative, while not sacrificing the descriptive accuracy.

9.1.3 Additional Experiment

We also performed an experiment in which we generate image narratives by following conventional image captioning procedure with human-written image narratives collected on Amazon Mechanical Turk. In other words, we trained LSTM with CNN features of images and human-written image narratives as ground truth captions. If such setting turns out to be successful, our model would not have much comparative merit.

We trained an LSTM with collected image-narratives for training split of MS COCO. We retained the experimental conditions identically as previous experiments, and trained for 50 epochs. Table 9.10 shows example narratives generated. Not only does It utterly fail to learn the structure of image narratives, but it hardly generates text over one sentence, and even so, its descriptive accuracy is very poor. Since LSTM now has to adjust its memory cells' dependency on much longer text, it struggles to even form a complete sentence, not to mention inaccurate description. This tells us that simply training with human-written image narratives does not result in reliable outcomes.

9.2 Experiments for Interactive Image Narrative Generation

We now perform and evaluate the experiments for image narrative generation with user interaction involved.

9.2.1 Experiment Setting

We first need to obtain data that reflect personal tendencies of different users. Thus, we not only need to collect data from multiple users so that individual differences exist, but also to collect multiple responses from each user so that individual tendency of each user can be learned.

We generated 10,000 questions that allow for multiple responses following the procedure described in Chapter 8. We grouped every 10 questions into one task, and allowed

3 workers per task so that up to 3,000 workers can participate. Since multiple people are participating for the same group of images, we end up obtaining different sets of responses that reflect each individual's tendency.

We have permutation of 10 choose 2, or $P(10, 2) = 90$ for each group. Note that we are assuming a source-to-target relation within the pair, so the order within the pair does matter. Since we have 3 workers assigned per group and there are 1,000 groups of 10 questions, we end up having 270,000 pairs of training data. We randomly split these data into 250,000 and 20,000 for training and validation splits, and performed 5-fold validation with training procedure described in Chapter 8. With 705 labels as possible choices, we had an average of 68.72 accuracy in predicting the the choice on new image, given the previous choice by the same user. Randomly matching the pairs with choices from different users seemingly drops the average score down to 45.17, confirming that the consistency in user choices is a key point in learning preference. As we will see later in this chapter, this result is consistent with the results from human evaluation.

9.2.2 Evaluation

We performed 3 experiments, each of which is supposed to examine different aspects of our proposed model; question generation, user-dependent image narrative generation, and the learning of user's interest.

Question Generation

For question generation, our interest is whether our model can generate questions that allow for various responses, rather than single fixed response. We showed the workers on Amazon Mechanical Turker whether the question can be answered in various ways or has multiple answers, given an image. 1,000 questions generated using both VQG and VQA modules following our proposed model, and another 1,000 questions generated using only the VQG technique without following our model, were presented to the workers. The workers' task was simply to answer with yes or no with regards to whether the visual question can be answered in multiple ways.

Table 9.11 shows the number of votes for each model. It is very clear that the questions generated from our proposed model of parallel VQG and VQA outperformed by far the questions generated from VQG only. This is inevitable in a sense that VQG module was trained with human-written questions that were intended to train the VQA module to answer, i.e. with questions that mostly have clear answers. On the other hand, our model deliberately chose the questions from VQG that have evenly distributed probabilities for answer labels, thus permitting multiple possible responses. Table 9.12 shows examples of visual questions generated from our model and VQG only respectively. In questions generated from our model, different responses are possible, whereas the questions generated from VQG only are restricted to single obvious answer.

Reflection of User's Choice on the Same Image

Our next experiment is on the user-dependent image narrative generation. We presented the workers with 3,000 images and associated questions, with 3 possible choices as a response to each question. Each worker freely chooses one of the choices, and is asked to rate the image narrative that corresponds to the answer they chose (i.e., the image narrative was generated based on that choice), in consideration of how well it fits and reflects their answer choices. Thus, even with the same image and the same question, workers may be asked to rate different image narratives depending on their answer. As a baseline model to compare, we also examine a model where the question is absent in the learning and representation stages, so that only the image and the user input are provided. Rating was performed over scale of 1 to 5, with 5 indicating that the image narrative is highly reflective of their choice. Table 9.13 shows the result. Our model clearly has an advantage over using image features only with a margin considerably over standard deviation. Agreement score among the workers was calculated based on [8]. Agreement score for our model falls into the range of 'moderate' agreement, whereas, for baseline model, it is at the lower range of 'fair' agreement, as defined by [55], demonstrating that the users more frequently agreed upon the reliability of the image narratives for our model. Figure 9-1 shows examples of images, generated questions, and image narratives generated depending on the choice made for the question respectively.

Reflection of User's Choice on New Images

Finally, we experiment with the learning of user's interest. As in the previous experiment, each worker is presented with an image and a question, with 3 possible choices as an answer to the question. After they choose an answer, they are presented with a new image and a new image narrative. Their task now is to determine whether the newly presented image narrative reflects their choice and interest. As a baseline model to compare, we again examined a model where the question is absent in the learning and representation stages, as in the previous experiment on the reflection of user's choice on the same image. In addition, we performed an experiment in which we trained preference learning module with randomly matched choices. In other words, each pair of choices in the training data did not consist of the choices from the same user, but the choices that were randomly matched. This is to examine whether there exists a consistency in user choices that enables us to apply the learned preferences to new image narratives. While the number of training samples by random matching can be very large, we restricted it to 270k as in our proposed model. Rating was again performed over scale of 1 to 5, with 5 indicating that the image narrative is highly reflective of their general preference or interest.

Table 9.14 shows the result. As in reflection of user's choice on the same images, our model clearly has an advantage over using image features only. Inter-rater agreement score is also more stable for our model. Training preference learning module with randomly matched pairs of choices resulted in a score below our proposed model, but above using the image features only. This may imply that, even with randomly matched pairs, it is better to train with actual choices made by the users with regards to specific questions, rather than just with most conspicuous objects as in using image features only. However,

its performance is still below that of our proposed model, confirming the importance of training with choices made by the same user that have consistency. Overall, the result again confirms that it is highly important to provide a context, in our case by generating visual questions, for the system to learn and reflect the user's specific preferences. It also shows that it is important to train with consistent choices. Table 9.15 shows examples of image narratives generated for new images, depending on the choice the users made for the original image, given the respective questions.

Discussion

It was shown via the experiments above that there exists a certain consistency over the choices made by the same user, and that it is thus beneficial to train with the choices made by the same users. Yet, we also need to investigate whether such consistency exists throughout different categories of images. We ran Fast-RCNN on the images used in our experiment, and assigned the classes with probability over 0.7 as the labels for each image. We then define any two images to be in the same category if any of the assigned labels overlaps. Of 3,000 pairs of images used in the experiment, 952 pairs had images with at least one label overlapping. Our proposed model had average score of 4.35 for pairs with overlapping labels and 2.98 for pairs without overlapping labels. Baseline model with image features only had 2.57 for pairs with overlapping labels and 2.10 for pairs without overlapping labels. Thus, it is shown that a large portion of the superior performance of our model comes from the user's consistency for the images of the same category, which is an intuitively correct conclusion.

However, our model also has superiority over baseline model for pairs without overlapping labels. This may seem more difficult to explain intuitively, as it is hard to see any explicit correlation between, for example, a car and an apple, other than saying that it is somebody's preference. We manually examined a set of such examples, and frequently found a pattern in which the color of the objects of choices was identical; for example, a red car and an apple. It is difficult to conclude whether such pattern was detected due to our feature representation reflecting local variations, or by biases in the dataset, but in any case, it is likely that there exists some degree of consistency in user choices over different categories, although to a lesser extent than for images in the same category. Also, it is once again confirmed that it is better to train with actual user choices made on specific questions, rather than simply with most conspicuous objects.

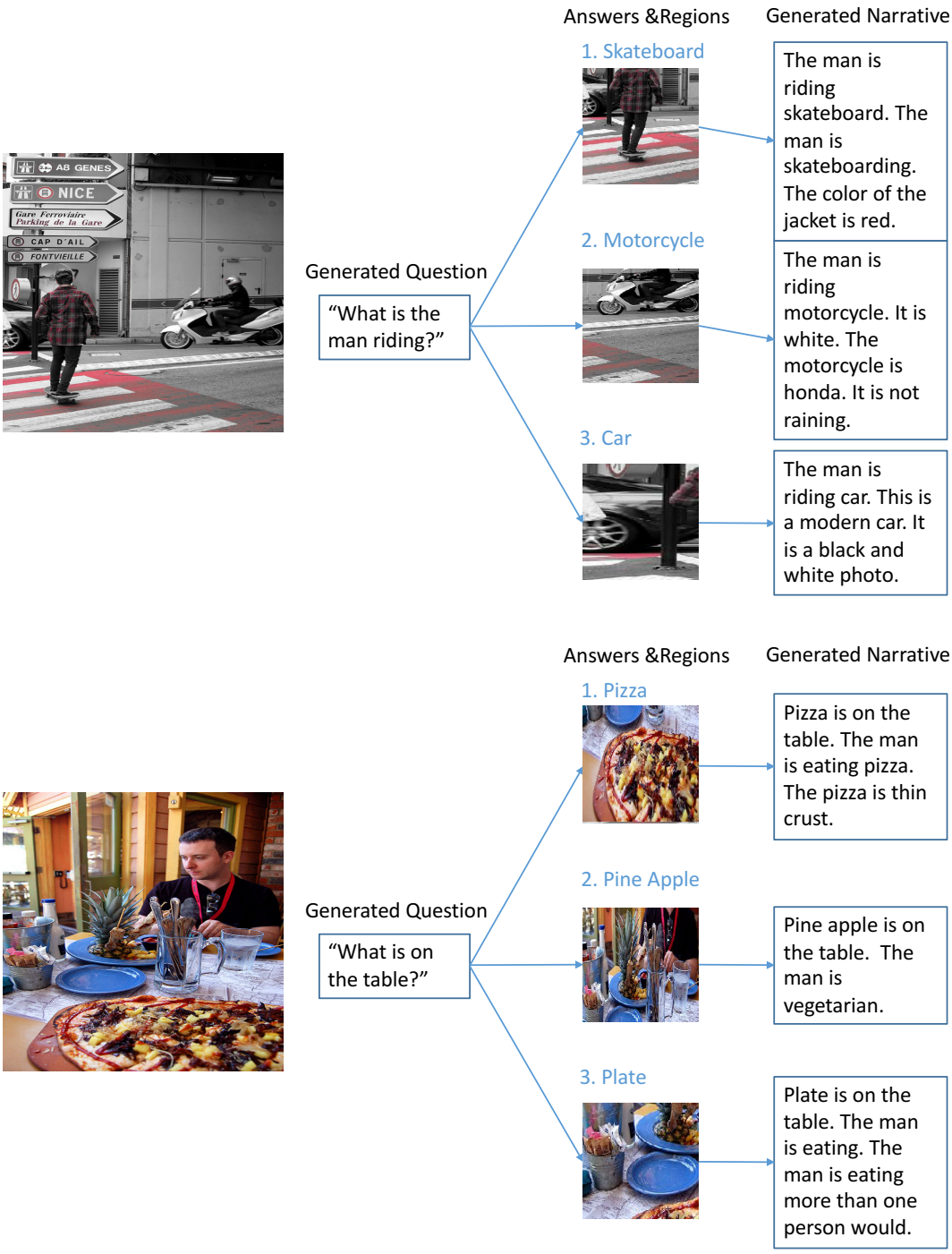


Figure 9-1: Examples of region extracted and image narratives generated depending on the answer to the question.

Table 9.1: Examples of human-written image narratives collected on Amazon Mechanical Turk.

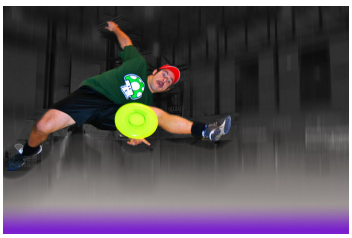
Image	Human-written Narrative
	This cat is having fun. She is very confused about the change in the carpet. It is funny that this has interested her so much. Cats are very picky and they do not like changes. She is probably mad about this.
	The pizza cook makes the pizza. The couple looks forward to pizza. The oven is very hot. He is a master at making pizza. He was born in italy.
	The food truck looks good. I bet they have good food. Does everyone in a food truck have a beard? I am so done with the beard thing. Hope his beard does not get into the food.
	Car is in very good shape for the age. This is a prefect car for California. I think i see that this is in Huntington beach. This would attract a lot of attention. Great way to pick up girls or guys.
	A dad and his daughter are sitting on the couch. They have just woken up. They each have a cup of juice. They use cups with lids so they don't spill on the couch.
	Tom is playing with a frisbee. He is practicing new moves. He jumped up in the air. He is trying to catch it between his legs. He was successful in his attempt.

Table 9.2: Statistics for human-written image narratives collected on Amazon Mechanical Turk.

# of answers collected	13,221
rewards per assignment	$.20
minimum length of image narrative	10
maximum length of image narrative	83
average length of image narrative	31.629

Table 9.3: Performances of the image narratives generated by each model on a subset of MS COCO determined by automatic evaluation metrics, with human-written image narratives as ground truth references.

Model	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
COCO	13.97	6.13	2.85	1.39
SIND	13.39	2.99	0.82	0.18
DenseCap	20.77	9.26	4.15	1.90
Ours	**20.87**	8.71	3.58	1.41

Table 9.4: Each model's performance on DIANE.

Metric	**COCO**	**SIND**	**DenseCap**	**Ours**
Diversity	2.972	2.060	3.102	**3.580**
Interesting	2.875	2.100	3.336	**3.489**
Accuracy	2.812	2.105	**3.188**	3.132
Naturalness	2.754	2.059	3.146	**3.374**
Expressivity	2.819	2.141	3.257	**3.381**
Overall	2.846	2.093	3.201	**3.391**
Std. Dev	$\pm$0.93	$\pm$ 0.86	$\pm$0.93	$\pm$0.89
% of Win.	.300	.195	.357	**.400**

Table 9.5: Our model's performance against each model on χ^2 with 2 degrees of freedom, and one-sided p-value obtained from binomial probability (rightmost column). $>$ refers to the cases where our model was rated higher than vs. Model, and so on.

vs. Model	$>$	**=**	$<$	χ^2	***p*-value**
COCO	2,208	1,222	1,570	133.37	1.4e-25
SIND	2,970	538	1,492	812.93	1.1e-11
DenseCap	1,890	1,454	1,656	271.33	4.5e-05

Table 9.6: Examples of narratives generated by each model (K=5). Each baseline is referred to as COCO, SIND, and DenseCap, respectively.

Image	COCO	SIND	DenseCap	Ours
	An elephant standing in a field of grass. A large elephant standing in a field of grass. A bathroom with a sink and a mirror. A large building with a clock on it.	The dog was very happy to see the animals. We had a great time. I went to the museum today. We went to the city to see the sights. We saw a lot of old buildings. The view from the top was amazing.	An elephant standing in a field of grass. A gray elephant. Elephant trunk is curled. Elephant in the photo. Trunk of an elephant. Elephants walking on the road.	This is a baby elephant. The elephants are standing on grass. They are bored. The elephant is sitting. This is not a zoo.
	A teddy bear sitting on top of a wooden table. A teddy bear sitting on top of a bed. A stuffed bear is sitting on a bed.	The cake was delicious. I had a great time. The food was delicious.	A teddy bear sitting on top of a wooden table. Teddy bear on a table. A brown teddy bear. A teddy bear. A teddy bear on a table. The head of a teddy bear.	Bear is on the cake. That stuffed animal has a funny face. It is mine. The bear is wearing hat. The cake is white.
	A giraffe standing in a field of grass. A giraffe standing in front of a building. A man and a woman standing next to a giraffe. A giraffe stan ding next to a wooden fence. A giraffe is standing in the middle of a forest.	We went to the city to see the sights. We had a great time. I went to the fair today. We saw a lot of old buildings.	A giraffe standing in a field of grass. A cloudy blue sky. Two giraffes in a zoo. Trees in the background. A giraffe behind the fence. Two peop le sitting in a chair.	There are 2 giraffes. The giraffes love each other. The giraffe is eating. This is a zoo.
	A man is eating a hot dog in a restaurant. A man holding a hot dog in his hand. A man holding a hot dog in a bun. A man in a suit and tie standing in front of a building. A man in a hat is holding a hot dog.	We had a great time.	A man is eating a hot dog in a restaurant. Woman holding a sandwich. Woman has brown hair. Woman in black jacket. A sandwich on a white plate. A brown wooden wall.	The girl is eating sandwich. Her name is mary. She is hungry. She eats a lot. She is smiling.

Image	COCO	SIND	DenseCap	Ours
	A herd of zebras grazing in a field. A herd of zebra standing on top of a lush green field. A bird flying over a building with a clock. A man standing on a sidewalk next to a street sign. A group of zebras are standing in a field.	We had a great time. I went to the museum today. We saw a lot of interesting things. We went to the city to see the sights. We saw many different types of animals. We went to the museum.	A herd of zebras grazing in a field. A field of grass. Two zebras in a field. The photo was taken in the daytime. White clouds in blue sky. The grass is tall.	The zebras like each other. These animals are related. The zebras are not in a zoo. The animal is grazing.
	A close up of a pizza on a plate. A close up of a sandwich on a plate. A cat sitting on top of a window sill. A bathroom with a toilet and a sink. A plate of food with a sandwich and french fries. A person holding a hot dog in a bun.	The food was delicious. We had a great time. I went to the museum today.	A close up of a pizza on a plate. Pizza on a plate. Pizza on a table. The hand of a person. A cup of coffee. The pizza has red sauce.	500 calories are in the meal. This is a pizza. This is not a healthy meal. This is not for vegetarian.
	A street with cars parked on the side of it. A car parked in front of a parking meter. A street sign on a pole on a street. A car parked on the side of a road. A street sign that is on a pole.	We went to the city to see the sights. The car was covered in snow. I went to the museum today. We went to the museum. We had a great time. We went to the location.	A street with cars parked on the side of it. A silver car parked on the street. A black car parked on the street. A white truck. Blue sky with no clouds. A black truck.	The car is gray. The car is parked illegally. Where the car is is inappropriate. That is pine tree behind.
	A teddy bear sitting on a wooden bench. A teddy bear sitting on top of a tree. A teddy bear is sitting on the ground. A train traveling down tracks next to a forest. A teddy bear is sitting on a tree branch.	The kids had a great time. The dog was very happy to see me. I had a great time. The view was amazing. We had a great time.	A teddy bear sitting on a wooden bench. A teddy bear in a red hat. Red teddy bear. A white teddy bear. The nose of a sheep. The teddy bear is s itting on the ground.	These are stuffed animals. That teddy bear can be scary when you see it at night. The animals are there for fun. The bear is sleeping. This is not a real bear.

Table 9.7: More examples of image narratives.

Image	COCO	SIND	DenseCap	Ours
	A red and white train traveling down train tracks. A red and white train on a train track. A white and red train on a train track. A street sig n that is on a pole. A train station with a train on the tracks.	We had a great time. The car was covered in snow. I went to the museum today.	A red and white train traveling down train tracks. Front of train is yellow. Yellow door on train. Trees in the background. Train tracks on the ground. A train in the photo.	The train is yellow. The train is moving. These locomotive are likely expensive. The man is driving.
	A plate of food with a fork and knife. A pizza with a lot of toppings on it. A plate with a sandwich and a salad. A close up of a plate of food with broccoli.	The food was delicious.	A plate of food with a fork and knife. Pizza on a table. A pizza on a plate. A slice of pizza. The pizza has red sauce. A slice of tomato.	This is a vegetarian pizza. This is not a cheese pizza. The green vegetable is spinach. This is a healthy meal.
	A man and a woman playing a game with nintendo wii controllers. A man and a woman sitting on a couch. A man and a woman are playing a video game. A vase with flowers in it sitting on a table.	The kids were having a good time. We had a great time. The dog was very happy to see me. I had a great time.	A man and a woman playing a game with nintendo wii controllers. Man wearing a gray shirt. Man sitting on couch. Dog with black and white fur. A lamp on the table. A black fireplace.	The dogs are eating. They are celebrating christmas. They are talking about politics. The dog is sitting. The man is drinking beer.
	A table topped with a plate of food and a glass of wine. A table topped with a plate of food. A glass of wine sitting on a table. A hot dog on a bun with ketchup and mustard. A kitchen counter with a bunch of food items.	We had a great time. The food was delicious. I had a great time.	A table topped with a plate of food and a glass of wine. A bottle of beer. Two people sitting on a table. Wine glasses on the table. Wine glass with wine. Red and green apples.	The name of this restaurant is jonn's bakery. This would be a delicious gift. The table is made of wood. There are 3 glasses. This is a healthy meal.

Table 9.8: More examples of image narratives.

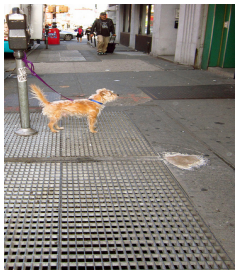
Image	COCO	SIND	DenseCap	Ours
	A cat laying on top of a bed next to a remote control. A cat laying on top of a bed next to a laptop. A cat laying on top of a bed next to a window. A person is holding a piece of broccoli. A large building with a clock on it.	The dog was very happy to see me. I had a great time. We saw a lot of old buildings. The view from the top was amazing.	A cat laying on top of a bed next to a remote control. A cat laying on a bed. The head of a cat. Ear of a cat. The cat is brown. The ear of a cat.	There is 1 cat. This cat looks lonely. The cat is not sleeping. The weather is sunny.
	A dog that is sitting on a bench. A street sign on a pole in front of a building. A man riding a skateboard down a street. A dog is running with a frisbee in its mouth. A large building with a clock on it.	The dog was very happy to see me. We had a great time. The house was very nice. I went to the museum today.	A dog that is sitting on a bench. A brown dog. A brick sidewalk. Man walking on sidewalk. Dog walking on sidewalk. A white line on the ground.	There is a dog. The dog is sad. The color of the wall is white. The color of the fire hydrant is gray.
	A man riding skis down a snow covered slope. A man in a red jacket is snowboarding. A man wearing a hat and a tie. A street light with a building in the background. A group of people standing on a beach with a kite.	We had a great time. I was so happy to see me. We went to the city to see the sights. I was so excited to see my friends. I went to the city last weekend.	A man riding skis down a snow covered slope. Man in red jacket. Snow covered mountain. The woman is wearing a helmet. The man is wearing black pants. Black and white jacket.	The person is skiing. This is alps. This person is having fun. He won the competition. The person is holding ski poles.
	A plate with a piece of cake on it. A close up of a pair of scissors on a table. A plate with a sandwich and a salad on it. A piece of cake on a plate with a fork. A close up of a plate of food on a table.	The food was delicious. The flowers were so beautiful. I had a great time.	A plate with a piece of cake on it. A yellow piece of donut. A red basket on the table. A box of donuts. A table with a wooden table. A donut with sprinkles.	There are 3 different types of food. This is not a healthy breakfast. This cake looks so fun. The orange color is bread.

Table 9.9: More examples of image narratives.

Table 9.10: Examples of image narratives generated by training with human-written image narratives. It shows that simply training with human-written image narratives utterly fails to generate reliable outcomes.

Image	Generated Narrative
	a man is sitting on a chair he is wearing a white shirt he seems to
	a man is holding a hot dog he is wearing a white shirt he seems to

Table 9.11: Results from evaluation on Mechanical Turk on whether the generated questions allow for multiple responses.

Model	# Overall	# Yes	# No
Ours	1,000	**664**	336
VQG	1,000	217	783
Overall	2,000	881	1119

Table 9.12: Examples of generated questions using our proposed model and VQG respectively. Questions in bold fonts are from our proposed model.

Image	Generated Question	Image	Generated Question
	**What is the color of the shirt?** How many children are there?		**What is on the table?** What is the table made of?
	**What is the dog doing?** What is the color of the couch?		**What is the color of the car?** What is the weather like?

Table 9.13: Results from evaluation on Mechanical Turk on how well the generated image narrative reflects the choices they made for the questions.

Model	Avg. Score	Agreement
Ours	**3.851±1.12**	**.601**
image only	2.636±1.01	.432

Table 9.14: Results from evaluation on Mechanical Turk on how well the generated image narrative for the new image reflects their interest or attention.

Model	Avg. Score	Agreement
Ours	**3.455±0.93**	**.527**
random match	2.772±0.79	.489
image only	2.238±1.24	.428

Table 9.15: Examples of image narratives generated on new images, depending on the choices made for the original input image.

Image & Question	Choice	New Image	Narrative
 What animal is this?	giraffe		The giraffe is standing. The weather is sunny.
	zebra		Zebra is thinking. It is not in a zoo.
	rhino		2 animals are in the picture. The sky is blue.
 What is the man riding?	skateboard		No one is riding bicycle. The man is standing.
	motorcycle		The motorcycle is red. No one is riding motorcycle.
	car		This is not a modern building. The image is not in black and white.
 What kind of animal is that?	dog		The horse is running. The car is white.
	sheep		The boy is wearing red shirt. Tree is in the background.
	person		The man is riding horse. The man is wearing hat.
 What color is the car?	white		The white object is bus. The car is white.
	green		The bus is green. The train is headed to Washington.
	yellow		The train is yellow. The image is not in black and white.

Chapter 10

Conclusion & Future Works

10.1 Conclusion

Generating image narrative requires not only factual description of the main components in the image, but also an understanding of how to account for local elements, sentiments, and non-visual elements present in the image. It also requires an efficient model to organize such elements into a well-defined structure consisting of multiple sentences. Most previous works on image description dealt with only a part of these many requirements, and thus fell short of generating an image narrative. In this thesis, we developed a model in which we generate an image narrative encompassing a variety of image-relevant components. In order to accomplish our objective, we proposed and implemented a number of modules, each of which plays an indispensable role within the model. We further developed our model to reflect the users' distinct preferences by implementing an interactive setting.

Dense Image Representation for Locally Robust Captioning

Previous works on image captioning have almost invariably dealt with main events occurring in the image. This is inevitable in a sense that CNN features are originally designed for classification of single object. In this chapter, we designed a model to generate dense image features that are robust to local elements and can reflect them in image captioning. First, we generated a large number of region proposals using selective search [92], extracted their CNN features, and applied PCA to make the features more optimal and practical. We then learned the codewords with k-means++, and coded the features with VLAD. This process was performed on a number of grids using spatial pyramid. Quantitative and qualitative evaluations demonstrate that our model can more accurately and frequently reflect local elements that frequently may not be captured by CNN features alone.

Image Captioning with Sentiment Terms

Another aspect that has been overlooked in previous works is the sentiment of the image. The difficulty lies in that sentiment can be ambiguous and is subject to multiple labels. For this reason, a dataset specializing in sentiments has been considered very difficult to build. We proposed to alleviate the ambiguity of sentiments by treating it as a multi-label problem.

In addition, based on our remark that comments on the images from social photo-sharing services frequently reflect the associated sentiments, we built a weakly-supervised sentiment dataset that is publicly available, with which we fine-tuned a separate CNN designed to classify the sentiments of the image. While automatic evaluation is difficult to carry out due to lack of ground truth captions involving sentiment terms, we performed human evaluation, which indeed confirmed that our captions better-reflected the sentiments of the image than compared models.

Visual Question Generation and Answering

Discussing a variety of elements from an image necessitates a decision of *what to discuss*, unlike image captioning task where the goal is settled from the beginning. Visual question generation is an efficient way to model the mapping between an image, or regions of the image, to relevant topics that can be discussed. Subsequently, the generated questions must now be answered in order to constitute an image narrative. Visual question answering task is currently one of the most competitive techniques to achieve that goal of answering visual questions, and we proposed a model to boost the performance of VQA. We alternated between average softmax probabilities from top regions and VLAD coding of CNN features for all region proposals, depending on the type of questions. Our proposed model took the first place in abstract scenes category at an international competition on the task, ascertaining its efficiency.

Single Image Narrative Generation

Each of the modules described so far is now organized to tackle the novel task of single image narrative generation, which in essence is the principal task this thesis is concerned with. Image features were generated by combining the spatial pyramid VLAD coding and fine-tuned sentiment CNN described in previous chapters. Top regions were extracted based on their *objectness*, from which visual questions were generated, and VQA technique was applied to answer the generated visual questions. Finally, simple natural language processing techniques were applied to connect the questions and answers into declarative formats, completing an image narrative. We performed both human evaluation and automatic evaluation, with human-written image narratives as references.

Interactive Image Narrative Generation

Considering that different viewers may attend to different parts of the image given a context, we tackled the task of interactive image narrative generation, under a premise that user's interest or attention can be learned to generate customized image narratives. In order for the user interaction to reflect each user's distinct tendency, we generated visual questions that can be answered in multiple ways, by making use of the probability distribution over answer labels from VQA module. We employed word-based attention heatmap to extract appropriate regions corresponding to the user input, and applied our image narrative generation model to the extracted regions to create user-dependent image narratives. We

also showed that by collecting inputs from the users interactively, we can learn their interest and attention tendency, and apply it to unseen images to customize the image narrative even without explicitly involving the users.

10.2 Remaining Problems & Future Work

While our model for image narrative generation task proved superior to other existing models, many aspects still remain to be improved or implemented. Although we leave them as future work in this thesis, it is worthwhile discussing those remaining problems, as it is likely to be of immediate research interest for ourselves as well as the research community.

First of all, while our image narratives deal with multiple aspects about the image, they are yet to have a strong structural orientation, including causality. Foremost difficulty would be the lack of such dataset, but constructing image narrative with clear causality may require a number of additional modules, such as correlation analysis between sentences, on top of just building a dataset.

Second, our model as a whole is highly dependent on the performance of each component module. For example, the state-of-the-art VQA module at the time of this writing still achieves less than 70% accuracy in real images category of VQA dataset [2]. Classifying and representing sentiments of the image with higher accuracy is also imperative for better performance. Likewise, improvements in sub-modules within the component, such as region extraction, are certain to improve the performance of the entire model as well. More advanced learning of user's interest in the interactive environment also remains an important future work.

Finally, applying similar motivation to other visual domains, videos in particular, would be another important challenge. Some of the previous works have attempted to generate video captions of multiple sentences [83, 103]. Yet, as in image captioning, they are yet to examine descriptions other than mere factual components of primary event, whose accuracy still has a long way to go. Likewise, VQA task and dataset for videos have been introduced [89], but have not been able to ignite an active research interest as yet. We briefly examined illustrations with VQA module in Chapter 6, but generating image narratives from illustrations will also be an intriguing as well as significant challenge.

Bibliography

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, 2016.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[3] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] Relja Arandjelović and Andrew Zisserman. All about VLAD. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[6] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[7] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] E. Bennett, R. Alpert, and A. Goldstien. Communications through limited-response questioning. *Public Opinion Quarterly*, 18:303–308, 1954.

[9] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of ACM International Conference on Multimedia (MM)*, 2013.

[10] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37:1757 – 1771, 2004.

[11] Thorsten Brants and Alex Franz. Web 1t 5-gram version 1. In *Proceedings of Philadelphia: Linguistic Data Consortium*, 2006.

[12] Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] Yan-Ying Chen, Tao Chen, Winston H. Hsu, Hong-Yuan Mark Liao, and Shih-Fu Chang. Predicting viewer affective comments based on image content in social media. In *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

[14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, 2014.

[15] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Abhishek Das, Satwik Kottur, Avi Singh, Deshraj Yadav, Jose M.F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[18] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *EACL Workshop on Statistical Machine Translation*, 2014.

[19] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010.

[21] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, 2016.

[23] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of Association for Computational Linguistics (ACL)*, 2015.

[24] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000.

[25] Amir Ghodrati, Ali Diba, Marco Pedersoli, Tinne Tuytelaars, and Luc J. Van Gool. Deepproposal: Hunting objects by cascading deep convolutional layers. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[26] Ross Girshick. Fast r-cnn. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proceedings of European Conference on Computer Vision*, 2014.

[29] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(1):307–361, 2012.

[31] Tatsuya Harada, Yoshitaka Ushiku, Yuya Yamashita, and Yasuo Kuniyoshi. Discriminative spatial pyramid. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[35] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47:853–899, 2013.

[36] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.

[38] Ilija Ilievski, Shuicheng Yan, and Jiashi Feng. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*, 2016.

[39] Hervé Jégou and Ondrej Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[40] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating local descriptors into a compact image representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[41] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions of Pattern Analysis and Machine Intelligence (TPAMI)*, 34(9):1704–1716, 2012.

[42] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of ACM International Conference on Multimedia (MM)*, 2014.

[43] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Bela Julesz. A theory of preattentive texture discrimination based on first-order statistics of textons. *Biological Cybernetics*, 41(2):31–138, 1981.

[45] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *https://arxiv.org/abs/1610.1465*, 2016.

[46] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of British Machine Vision Conference (BMVC)*, 2014.

[47] Andrej Karpathy and Fei-Fei Li. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[48] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

[49] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[50] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*, 2017.

[51] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *https://arxiv.org/abs/1602.07332*, 2016.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS*, 2012.

[53] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[54] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[55] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 1977.

[56] Piotr Dollar Larry Zitnick. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[57] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[58] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[59] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. Proceedings of Association for Computational Linguistics (ACL), 2004.

[60] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[61] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004.

[62] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

[63] Mateusz Malinowski and Mario Fritz. Towards A Visual Turing Challenge. In *NIPS Workshop on Learning Semantics*, 2014.

[64] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A deep learning approach to visual question answering. *CoRR*, abs/1605.02697, 2016.

[65] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate rgument structure. In *HLT*, 1994.

[66] Alexander Mathews, Lexing Xie, and Xuming He. SentiCap: generating image descriptions with sentiments. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[67] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2013.

[68] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of Association for Computational Linguistics (ACL)*, 2016.

[69] Nasrin Mostafazadeh, Lucy Vanderwende, Wen tau Yih, Pushmeet Kohli, and James Allen. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *ACL Workshop on Evaluating Vector Space Representations for NLP*, 2016.

[70] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[71] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of Association for Computational Linguistics (ACL)*, 2002.

[72] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[73] Florent Perronnin, Yan Liu, Jorge Sanchez, and Herve Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[74] Mary A. Peterson and Gillian Rhodes. *Perception of faces, objects, and scenes: Analytic and Holistic Processes*. Oxford University Press, 2003.

[75] Ehud Reiter. NLG vs. Templates. In *Proceedings of the 5th European Workshop on Natural Language Generation (EWNLG)*, 1995.

[76] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[77] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring Models and Data for Image Question Answering. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

[78] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

[79] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[80] Jorge Sanchez, Florent Perronnin, and Teofilo de Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33:2216–2223, 2012.

[81] Rakshith Shetty and Jorma Laaksonen. Video captioning with recurrent networks based on frame- and video-level features and visual content classification. In *ICCV Workshop on Describing and Understanding Video and The Large Scale Movie Description Challenge*, 2015.

[82] Kevin Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[83] Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. Beyond caption to narrative: Video captioning with multiple sentences. In *Proceedings of International Conference on Image Processing (ICIP)*, 2016.

[84] Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa). *arXiv:1609.6657*, 2016.

[85] Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset. In *Proceedings of British Machine Vision Conference (BMVC)*, 2016.

[86] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[87] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.

[88] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[89] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[90] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *CoRR*, abs/1609.05600, 2016.

[91] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 99, 2010.

[92] Jasper R.R. Uijlings, Koen E.A. van de Sande, Theo Gevers, and Arnold W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104:154–171, 2013.

[93] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Efficient image annotation for automatic sentence generation. In *Proceedings of ACM Conference on Multimedia (MM)*, 2012.

[94] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[95] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[96] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[97] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[98] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*, 2016.

[99] Huijuan Xu and Kate Saenko. Ask, attend, and answer: Exploring question-guided spatial attention for visual question answering. In *arXiv:1511.05234*, 2015.

[100] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

[101] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[102] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[103] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[104] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions of the Knowledge Data Engineering*, 26(8):1819–1837, 2014.

[105] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[106] Renhao Zhou, Qingsheng Yuan, Xiaoguang Gu, and Dongming Zhang. Spatial pyramid vlad. In *Proceedings of Visual Communications and Image Processing*, 2014.

[107] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Publications

International Conference

1. Kuniaki Satio, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada, "DualNet: Domain-Invariant Network for Visual Question Answering," Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2017), accepted, Hong Kong, July, 2017.

2. Atsushi Kanehira, Andrew Shin, and Tatsuya Harada, "True-Negative Label Selection for Large-Scale Multi-Label Learning," Proceedings of International Conference on Pattern Recognition (ICPR 2016), pp.3662-3667, Cancun, Mexico, Dec., 2016.

3. Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada, "Beyond Caption To Narrative: Video Captioning with Multiple Sentences," Proceedings of IEEE International Conference on Image Processing (ICIP 2016), pp.3364-3368, Phoenix, Arizona, Sep., 2016.

4. Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada, "Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset," Proceedings of British Machine Vision Conference (BMVC 2016), York, United Kingdom, Sep., 2016.

5. Andrew Shin, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura, "Context-Dependent Automatic Response Generation Using Statistical Machine Translation Techniques," Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL 2015), pp.1345-1350, Denver, Colorado, June, 2015.

Awards

1. 1st Place, Visual Question Answering (VQA) Challenge on Abstract Scenes at CVPR, Las Vegas, Nevada, 2016.