論 文 の 内 容 の 要 旨

論文題目　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）

氏　名　シン・アンドリュー

Image captioning task has enticed an unprecedented amount of attention with the advent of deep learning techniques. However, its objective has invariably been limited to the generation of factual description of the image, mostly through single sentence. Yet, images frequently provide an ample amount of contents that cannot be fully reflected by a single sentence of factual description. We examine a novel task of ***image narrative*** generation in which we attempt to overcome such 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 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. In particular, we actively engage the users with the learning and generation process,

thereby providing an interactive model to guide, correct, and enhance the learning process. We experimentally demonstrate that our proposed model for image narrative generation can generate a highly expressive and interesting image description with much wider range of topic contents and without sacrificing the descriptive accuracy, which turns out to be difficult to realize via conventional models.