論文の内容の要旨

論文題目　Automatic Sentence Generation for Images
via Key-phrase Estimation using Large-Scale Captioned Images
（大規模な説明文つき画像を用いたキーフレーズ推定に基づく
画像説明文の自動生成）

氏　名　牛久　祥孝

概要

Group of people sitting at a table with a dinner.

図　生成された説明文の例.

近年の情報技術の発展により爆発的に増加している画像などのマルチメディアを効率的に活用するために，個々のデータをユーザが容易に検索し理解できる技術が求められている．視界にある物や起きている事をラベルや話し言葉で理解できる技術は，将来的にも実世界で活躍するロボットや，視覚に障害のある人々を支援するために重要な技術となる．

画像に写っている事物を認識する一般物体認識は，機械翻訳技術の発達をきっかけにこの十年で広く取り組まれてきた研究分野である．限られた種類の物体を，小さな規模のデータセットで学習し，入力画像に含まれる物体を表すラベルを一つだけ付与する画像分類は，(a)動画の分類や(b)より多様なラベルを大規模なデータセットで学習する大規模画像分類，(c)1つの入力データに複数のラベルを付与するアノテーションといった発展を見せてきた．そしてこの2，3年，入力データを複数のラベルで表現するのみならず，それらの関係を包含する自然言語の文として入力画像を説明する手法の研究が脚光を浴びつつある．

従来の研究では，それぞれの画像において「どのような物体が」「どのような光景のもと」「どのような動作を」行っているか，などの情報がついたラベルを伴う画像を学習し，新規画像に対する説明文を生成していた．さまざまな画像を説明するには大規模な画像データセットの構築が必要となるが，そのような大量の画像に同様のラベルを付与するのは極めて困難である．

本論文は，画像とそれに関連する文章だけであればWebから大規模に収集するのも容易である点に着目し，そのようなデータセットにおいて新規画像の説明文を生成する手法を提案し，実際に図に示すような説明文を生成できることを示した．以下，各章について要旨をまとめる．

第 1 章：Introduction

本章では，一般物体認識における最終目標の1つと言える画像を説明する自然文の自動生成について紹介した．既存の説明文生成手法には画像それぞれに対して「主体，動作，光景」などの属性を表すラベルなど，セマンティックな知識を手作業で与える必要があり，多様で大量なデータへの拡張が困難である点を指摘した．本論文が画像と説明文のみからなるデータセットで説明文を生成できる手法を開発することを宣言した．

第 2 章：Methods to Describe Multimedia with Natural Language

本章では，まず画像の事物を認識する一般物体認識の研究分野に対し，データセット，特徴量，認識手法の3つの観点から記述した．次に近年注目を浴びつつある画像からの説明文生成について概観し，既存手法の多くが＜物体，動作，光景＞のようにラベルの役割を明確にしたラベルを必要としている点を指摘した．これは事物間の関係性を推定するためにマルチプレットを用いているともいえる．マルチプレットを使用するものの他に，画像中の物体をその位置と共に推定する物体検出の枠組みを利用する研究もある．物体検出そのものが未だ挑戦的な課題である点と，人手で画像中の物体とその位置を記入したデータが必要になる点がやはり問題になる．このように，役割が明らかなラベルによるマルチプレットや，事物の名前と位置を各画像に対して付与するには手作業での管理が伴い，データセットを大規模に構築するのが困難である．

逆にどこから大規模なデータを収集できるか検討するとき，Web の各画像共有サービスの存在が挙げられる．こういったサイトには数十億枚以上の画像が存在し，その多くには画像と関連する文章が付随している．本章では画像とその説明文のみからなるデータセットを用いて新規画像の説明文を生成する手法を提案した．具体的には，「画像の内容はいくつかのキーフレーズで表現でき，これらを単純な文法モデルで正しく繋げば説明文を生成できる」という仮説をたて，画像と説明文のみからなるデータセットを用いて入力画像のキーフレーズを推定するマルチキーフレーズ問題を新たに提起した．

多様な事物の組み合わせからなるキーフレーズはさらに多種多様になるため，より多くのデータで学習しなければならない．キーフレーズを推定する精度はもちろん，学習時や推定時のデータ量に対するスケーラビリティが重要になる．

第 3 章：Investigation of Online Learning Methods for Multiclass Classification

本章では，まず大量のデータで学習する大規模画像分類問題に注目した．大規模データで学習するには，データを1つずつ読み込んで行うオンライン学習が有用である．しかし，従来のオンライン学習手法は，人工データや自然言語処理に関するデータによって評価されてきた．本論文では，大規模画像分類で現在主流となっている画像特徴抽出手法と主たるオンライン学習手法を横断的に組み合わせ，画像認識におけるオンライン学習の重要な知見を得た．

- 古典的なパーセプトロンでも現在の手法に匹敵する性能が得られる．
- 過去の識別器から現在の識別器までの Averaging は全てのオンライン学習に必須である．
- 2値分類器を多クラス分類に応用させる one-vs.-the-rest 法よりも，分類器の学習手法としてマルチクラスに対応させた方が短い時間でより安定して収束できる．

第 4 章：Multi-Keyphrase Problem and Sentence Generation

本章では，入力画像に対して複数のキーフレーズを推定し，文法モデルでキーフレーズを繋ぐ手法について提案した．実際の文生成技術への要求機能は(1)所望の文長に近づけること，(2)推定されたキーフレーズはなるべく多く用いること，(3)キーフレーズ同士をつなげる際には文法的に確からしい単語列を用いること，の3つである．本章ではそれぞれの要求機能に基づく複数のコスト関数を設計し，候補文の合計コストが最小化されるものを入力画像への説明文とするアプローチを取った．こ

れはグラフ探索の一種である Multi-stack beam search を改良することで近似的に解ける. 実際に画像と説明文からなるデータセットを用い, 収集コストの高いセマンティックな知識が無くとも画像の説明文が生成可能であることを確認した. あわせて, 大規模画像分類に対する実験で得られた知見をもとに従来のオンライン学習手法を改良し, 各事物に固有の分類器と照合するだけで高精度にアノテーションできる手法を提案した. 従来のアノテーションでは入力画像に類似した画像のラベルを入力画像のラベルとして出力するノンパラメトリックな手法が主流であったが, 類似度の学習や画像検索におけるスケーラビリティに問題があった. 実際に複数のデータセットを用い, スケーラビリティに優れた提案手法が精度よく画像をアノテーションできることを示した.

第 5 章：CoSMoS: Common Subspace for Model and Similarity

本章では, 部分空間の学習に着目し, 効率的かつ高精度なキーフレーズ推定手法である Common Subspace for Model and Similarity (CoSMoS) を提案した. 多くのキーフレーズを学習する際には, 各事物の分類器を学習するためのパラメータが膨大になってしまい, 安定した高精度なキーフレーズ推定が困難になる. 同じキーフレーズを持つ画像同士が高い類似度, すなわちお互い近傍に位置するような部分空間を学習する手法は従来も複数存在した. CoSMoS は部分空間内で実際に各事物の分類器を同時学習することで, 分離すべきキーフレーズ同士を判別しやすいような空間を効率的に学習できる. 関連研究でもよく用いられるデータセットに提案手法を適用し, 画像アノテーション精度やキーフレーズ推定精度がより改善されることを確かめた. さらに, キーフレーズ推定精度が改善されることで, 説明文の精度も改善されることを確認した.

第 6 章：Evaluation of Sentential Description for Images

本章では, 数千から数万の画像それぞれに対して人手で複数の説明文が付与されたデータセット2つと, Web から収集された百万枚の画像と関連分からなるデータセット1つに提案手法を適用した. 人手で付与しなければならないマルチプレットやバウンディングボックスを用いず, 画像と文章のみのデータセットで関連研究より高精度に説明文を生成できた. また, Web から収集されたデータセットの規模を変えながら説明文生成を行った結果, データセットの規模が増えるに従って説明文の精度も向上した.

第 7 章：Conclusion and Future Work

本論文の学術的貢献は, 以下のようにまとめられる.

- 画像と文章のみからなるデータセットで説明文を生成するために, キーフレーズ推定アプローチを提案した.
- 大規模画像データからキーフレーズ推定を行うにあたり, 現在主流となっているオンライン分類学習手法と画像特徴量を組み合わせた横断的な実験を行い, 新たな知見を得た.
- 大量なキーフレーズを効率的に学習するために, 部分空間を用いた新たな学習手法を提案した.

提案手法はデータ数に対するスケーラビリティを持ち, 実験結果からもより大規模なデータセットによってより精度よく説明文を生成できることが示されている. 今後, さらに精度を向上させるためにはキーフレーズの設計が重要になると予想される. たとえば本論文では連続単語列からキーフレーズを抽出したが, 非連続単語列をも含めればより多様な事物間関係を画像から推定できるようになる.