キーワード付与による画像キャプション生成

木村文飛 [1]　新納浩幸 [2]
[1] 茨城大学工学部情報工学科　[2] 茨城大学大学院理工学研究科 情報科学領域
18t4030r@vc.ibaraki.ac.jp　hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

画像の内容を自動的に記述する画像キャプション生成という技術は、自然言語処理と画像処理という二つの分野を組み合わせて行う分野である。本論文では、キーワード付与というデコーダの入力ベクトルにキーワードのベクトルを付与する手法を提案し、それを行うことで画像キャプション生成における問題点の解決を図った。提案手法を行わなかったモデルと比較することで評価を行い、結果として提案手法が CIDEr のスコアを 0.7 更新したことが確認できた。

1　はじめに

本論文では、画像とキャプションのペアを用いて行うキャプション生成について、キーワード付与を用いることで、生成されるキャプションの精度を向上させる手法を提案する。

画像キャプション生成は、画像からその画像の簡単な説明文（キャプション）を生成するタスクである。人力で作成された、画像とキャプション文のセットを用いて学習する手法によって、近年目覚ましい発展を遂げている分野である。

画像キャプション生成における従来の手法の問題点として、十分なデータセットの用意が困難という点があげられる。人力で作成された画像とキャプション文のセットをさらに拡張することは、非常にアノテーションコストがかかる作業なのである。これの解決策として、教師なし学習 [1, 2] やデータ拡張 [3] などがある。また画像キャプション生成の大きな問題点の一つとして、不適当なキャプションが生成された場合、原因が何であるかを特定することが容易ではないという点があげられる。画像キャプション生成をする際には、画像畳み込みといった画像側の処理と、画像キャプション文の学習という言語側の処理を一連の流れの中で行わなければならない。つまり、原因が言語側にあるか画像側にあるか不明確なのである。本論文ではこれらの問題点に着目し、画像を畳み込み作成されたベクトルに、学習に使われるキャプション内に出現するキーワードに対応したベクトルを付与することで、生成するキャプションの精度を向上させることを目指す。また、キーワード付与をすることで物体検出が正常に行われた状況下となるかということを実験する。

実験では、訓練データの画像に MSCOCO のデータセット [4]、キャプションに STAIR Captions の日本語キャプション [5] を用いた。テストデータは MSCOCO のデータセットから一部を抜き出して使用した。そして、キーワード付与を行ったデータで学習を行ったシステムで得られたキャプションと、キーワード付与を行っていないデータで学習を行ったシステムで得られたキャプションを比較し評価を行った。

2　関連研究

本研究では、[6] で提唱された画像キャプション生成モデル、NIC の構造を参考にした。

このモデルは畳み込みニューラルネットワーク（CNN）をエンコーダ、長短期記憶ニューラルネットワーク（LSTM）をデコーダとして、それらを組み合わせることで画像キャプションを生成するモデルである。入力画像から畳み込みニューラルネットワーク（CNN）によって特徴量を取り出したものと、画像の説明文を単語埋め込みによりベクトル化したものを、長短期記憶ニューラルネットワーク（LSTM）に入力することで文章を生成するものである。このモデルの特徴として、画像キャプション生成に利用可能なデータセットのサイズが大きくなればなるほど、出力するキャプションの精度が上がることは明らかであると述べられている。しかし、実際には利用可能なデータセットを増やすことは非常にコストがかかり容易ではない。

3　提案手法

画像キャプション生成の問題点は、データセット拡張のコストが高いことと、不適切なキャプション生成の原因が言語側にあるか画像側にあるか不明確であることの 2 点である。これらを解決するため、学習に用いられるデータセットに、キャプションから取り出したキーワードをもとに作成したベクトルを付与する手法を提案する。

4　実験

4.1　実験設定

本実験で付与するベクトルとは、キーワードとなる単語 150 個を用いて、学習に用いられるキャプションから作成したキーワードの bag of words である。これを各キャプションごとに作成し、キャプションに対応する画像を畳み込むことで得られるベクトルにアペンドした。尚、キーワードとなる単語 150 個は、データセット全体における名詞の出現回数をもとに決定しており、物体検出で出力されないような名詞（色、数字等）は手動で除外した。

本実験では、MSCOCO の画像データセットと STAIR Captions の日本語キャプションを組み合わせて学習を行った。キーワード付与を行った場合におけるキャプションの品質の変化を調べるために、デフォルトのデータセットとキーワード付与を行ったデータセットの合わせて 2 パターンの訓練データを用意した。これらを用いて学習を行い、デフォルトのデータセットを用いて学習を行ったモデルと、提案手法を用いて作成されたデータセットを用いて生成されたモデルとを比べることによって品質が向上したかを調査する。

今回の実験では CNN は学習させずに、学習対象をエンコーダの最後に行う 1 層の全結合層と、デコーダとなる LSTM のみを学習させた。また、本実験では CNN に学習済みネットワークである VGG-16[7] を用いて画像畳み込みを行った。

学習に用いたデータセットは、MSCOCO の訓練データ画像と、STAIR Captions の 5 つのキャプション文のセットを組み合わせた 82732 個である。図 1 に訓練データ例を示す。

これに対してバッチサイズ 100、エポック数 10 として学習を行った。そして、学習する際の各エポック終了時にモデルを保存し、各モデルに対してテスト

stair caption
長い 机 と 椅子 が 、 並ん で 置い て ある
テーブル の 上 に 、 植物 が 置い て ある
背もたれ が 丸い 椅子 が たくさん 並べ られ て いる テーブル の 真ん中 に 大きな ガラス の 花瓶 に 活け られ た 花 が 飾ら れ て いる
長い テーブル に お 揃い の イス が 置か れ て いる
10 人 掛け の テーブル の 上 に は 分厚い 本 や フラワー アレンジメント が 置か れ て いる

図 1　訓練データ例

トデータとなる MSCOCO のデータセットから抜き出した画像 10000 枚からキャプションを生成させ評価を行った。

実験後、キーワードとして埋め込んだ単語がどれほどの出現率となるかを調査するために、提案手法を用いたモデルで最高精度だったモデルと提案手法を用いないモデルで最高精度だったモデルを用いて生成されたキャプション 10000 文に対して、ベクトルを付与する際に用いたキーワードの bag of words を作成し分析した。

分析内容は以下の 3 通りである。

- キーワードとして付与した単語の出現率。
- bag of words の値で場合分けした、 キーワードとして付与した単語の出現率。
- 不適切なキーワードの出現率。

これらについて調査を行い、分析を行った。

4.2　評価手法

生成キャプションの評価には CIDEr[8] を用いた。CIDEr とは、動画や画像のキャプション生成で評価に用いるメソッドであり、n-gram 形式を用いた TF-IDF による平均コサイン類似度を表すものである。キャプション画像の枚数やアノテーション数を考慮しているという点から本実験の評価手法に用いた。

4.3　実験結果

実験結果として、キーワード付与前と付与後のモデルで行ったテスト結果をまとめたものを図 3 に示

す。キーワードの付与前のモデルで最高精度となったのはエポック 3 のものであり、値は 42.1 であった。キーワードの付与後のモデルで最高精度となったのはエポック 4 のものであり、値は 42.8 であった。この結果により、CIDEr という評価手法では提案手法がベースラインを上回る最高精度を記録し、提案手法が精度向上に起因することが確認できた。

キーワードとして付与した単語の出現率についての分析結果をまとめたものを表 1 に示す。この結果から、キーワードとして付与した単語の出現率は、上昇したが大きな変化がなかったことが確認できた。

キーワードを付与する際に埋め込まれるベクトルは、正解キャプション 5 つから作成したキーワードの bag of words であるので、各キーワードごとに出現回数によって重みが違うという問題がある。表 1 はその問題点を考慮していないため、画像ごとではなく単語単位で、正解キャプション 5 つにおけるキーワードの出現回数によって場合分けし、分析した。その結果を表 2 に示す。

間違ったキーワードが出現した確率も同時に調査した。具体的には、追加したベクトルに含まれていないキーワードが出現する確率を計算した。計算結果は追加前が 52.75 %、追加後が 47.74 %であった。

4.4 出力例

図 2 にモデル別のキャプション出力例を示す。

「パソコン」、「サーフ」はキーワードであり、ベクトルに埋め込まれているため、提案手法を用いることで生成されるキャプションがより正確になっている場合があることが確認できる。

5 考察

実験結果より、本実験の提案手法は CIDEr の評価値こそ微増させたものの、埋め込んだキーワードをキャプションに出現させるという観点においてはあまり効果がないということが分かった。特に、正解キャプション内で出現回数が少ないキーワードに関してはほとんど改善が見られなかった。埋め込んだキーワードをよりキャプションに出現させることができなかった要因の一つに、画像に様々な物体が移っているという要因が考えられる。

提案手法を追加する前から、生成キャプション 1 つにつき平均 1 つのキーワードを出現させることが出来ている。よってこの提案手法によってキャプションの精度を向上させるには、すでに生成キャプションに含まれている物体に加えて他の物体も表現しなければならないのである。

生成キャプション

前	男性 が ベッド の 上 で 本 を 読ん で いる <EOS>
後	男性 が パソコン を 見 て いる <EOS>

生成キャプション

前	砂浜 で 男性 が フリスビー を し て いる <EOS>
後	砂浜 で サーフ ボード を 持っ て いる 人 が いる <EOS>

図 2 テスト画像例

埋め込んだキーワード数の平均から、多くの画像には 3 つ以上の物体が写っていることが確認でき、これらは提案手法によって、学習に用いるベクトルへと確実に付与されている。故に、正確に物体検出出来ている状況に近い状態にあると言える。しかし、それでもあまり改善が見られないことから、本実験によって、単純に画像の物体すべてを物体検出することが出来ても、それらを複数用いて正しく画像キャプション生成をすることは困難であるということが考えられる。

また、提案手法は表記ゆれを吸収できないという欠点がある。例えば、「男性」、「男」、「青年」などといった意味上は似通った単語については、本研究では特に操作せずに別の単語として扱っている。加えて、「男」や「青年」等の単語は埋め込むキーワードにバリエーションを持たせるために除去した。そのため、意味上は似通った単語であっても追加するベクトルには反映されない。これが影響し、埋め込んだキーワードをキャプションに出現させることが出来なかった可能性が考えられる。

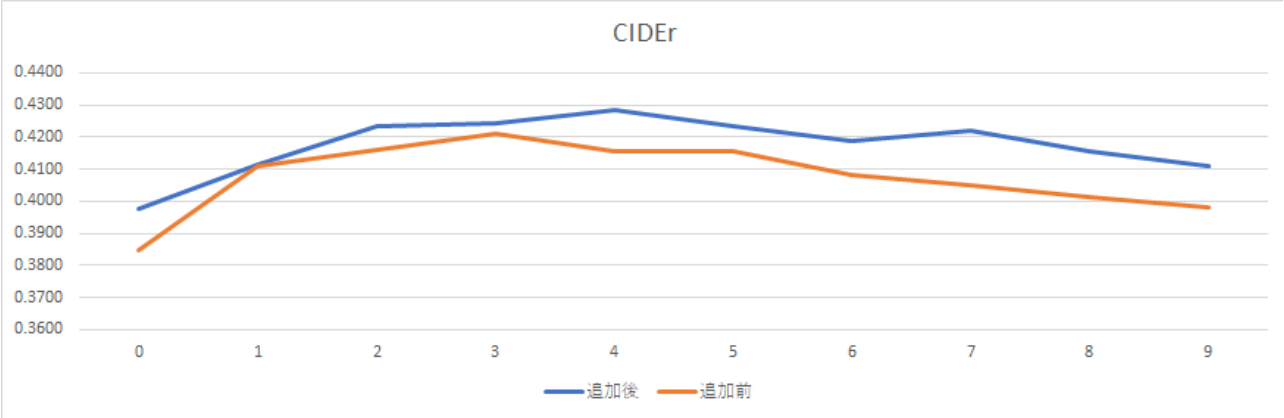

図 3　CIDEr の評価

表 1　埋め込んだキーワード数（平均）に対する、出力されたキーワード数（平均）

kw 数	出力できた kw 数（追加前）	出力できた kw 数（追加後）
3.618	1.039	1.062

表 2　正解文内の各キーワードの出現回数（bag of words の数値）をもとにした、キーワードの出力率

正解文の各 kw 出現回数	追加前	追加後	単語総数
1 回	12.01 %	11.31 %	15564
2 回	24.68 %	24.35 %	6941
3 回	37.21 %	39.18 %	4934
4 回	49.51 %	52.20 %	4272
5 回	61.58 %	64.88 %	3964
6 回	69.30 %	73.24 %	355
7 回	75.00 %	77.00 %	100
8 回	80.00 %	86.67 %	30
9 回	73.33 %	66.67 %	15
10 回	100.00 %	75.00 %	4

しかし、提案手法が全く機能していなかったというわけではない。表 2 から、正解キャプション内で出現回数が 5 回の単語に関しては、出現確率が約 3.305 %、4 回の単語に関しては、出現確率が 2.692 %ほど上昇している。この結果から、出現回数が多い、つまり重要度の高いキーワードの出現確率は向上していることが確認出来る。逆に、出現回数が 1 回や 2 回だけの、出現回数が少ないキーワードに関しては出現確率が下降している。これらは単純に bag of words の数値に起因するものであると考察できる。bag of words が高いキーワードのみが優先されてしまい、低いキーワードの生成に至らなかったと考えられる。

また、誤ったキーワードを含めたキャプションが 52.75 %から 47.74 %へと減少したことから、図 2 の例のように、当初の目的である物体検出の安定、キャプションの正確性を増すという点では提案手法は効果的であるということが確認できた。

6　おわりに

本論文では、キーワード付与を行うことで画像キャプション生成における品質の向上を行った。具体的には、画像を畳み込んだベクトルに、厳選したキーワードを用いた正解キャプションの bag of words をアペンドした。提案手法を用いて生成したキャプションは評価値については微増することが確認できた。しかし、複数のキーワードの生成という点においてあまり大きな効果がないことが実験結果の分析からわかった。

本実験を行っていくうえで Chive のような日本語単語分散表現を追加するベクトルに使ってみるというアイデアは出ていた。本実験ではよりシンプルになる用に bag of words を採用したが、これらのベクトルを利用して実験を行うことも検討する。また、本提案手法の弱点として述べた表記ゆれについての対抗策も検討する予定である。

キーワードを付与することで物体検出は出来ていると考えられることから、問題点は言語処理側にあるといえるはずである。今後は言語処理側に注視し、追加するベクトルについて改良に関する研究を重ねてく。

謝辞

本研究は JSPS 科研費 JP19K12093 および 2021 年度国立情報学研究所公募型共同研究（2021-FC05）の助成を受けています.

参考文献

[1] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[2] Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3692–3702, Online, April 2021. Association for Computational Linguistics.

[3] 岩村紀与彦, ルイ笠原純ユネス, モロアレッサンドロ, 山下淳, 淺間一. アテンション機構を用いたクロップとマスクによるキャプション生成のためのデータ拡張. 精密工学会誌, Vol. 86, No. 11, pp. 904–910, 2020.

[4] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pp. 740–755, Cham, 2014. Springer International Publishing.

[5] Stair captions: Constructing a large-scale japanese image caption dataset. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics.

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

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, 2015.

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