第 27 回年次大会予稿

ヒントワードを活用した画像キャプションの自動生成 Image Caption Generation Using Hint-words

村本 尚生[†] 于 海涛^{††}

Naoki MURAMOTO Haitao YU

[†] 筑波大学大学院 図書館情報メディア研究科

Graduate School of Library, Information and Media Studies, University of Tsukuba

〒 305-8550 茨城県つくば市春日 1 丁目 2

E-mail: s1921648@s.tsukuba.ac.jp

^{††} 筑波大学 図書館情報メディア系

Faculty of Library, Information and Media Science, University of Tsukuba

〒 305-8550 茨城県つくば市春日 1 丁目 2

E-mail: yuhaitao@slis.tsukuba.ac.jp

画像キャプション生成技術により、画像の内容を説明する文を自動的に生成することが可能となった。この技術は近年注目を集めている。しかし、この技術には未だに多くの課題が残っている。本研究では、画像に関連する重要かつ有益な単語（ヒントワード）を活用し、高品質な画像キャプションを生成する手法を提案する。また、提案された画像キャプション生成手法の有用性を検証するために、有益なヒントワードを選択するためのいくつかの方法を提案し、ベンチマークデータセットで一連の実験を行う。実験結果は次のことを証明している。(1) 提案されたアプローチはより情報量豊富な画像キャプションを生成可能であり、ベースライン手法より性能が優れている。(2) ヒントワードの数と品詞情報はヒントワードを選択するための重要な指標である。

The technique of image captioning enables us to automatically generate a caption for an image, which has attracted significantly increasing attention in recent years. However, image captioning is far from being resolved and remains an open challenge. In this work, we propose to utilize the key informative words (referred to as hint words) that are associated with an image to generate a high-quality caption for this image. To validate the effectiveness of the proposed approach for image captioning, on one hand, we propose a number of methods for selecting informative hint words. On the other hand, we conduct a series of experiments on a benchmark dataset. The experimental results demonstrate that: (1) the proposed approach is able to generate captions with richer information, and outperforms the baseline method; (2) the number of hint-words and the Parts of Speech (POS) information are important indicators for selecting hint words.

キーワード: 画像キャプション生成, ヒントワード, 深層学習

Image Caption Generation, Hint-words, Deep Learning

1 序論

画像キャプション生成とは、画像の内容を理解し、説明する文を生成することである。画像キャプション生成は人間にとっては簡単な問題であるが、画像のキャプションを生成するためには、画像の内容を理解し、言語で説明する必要があるため、コンピュータにとっては、研究者が長年にわたって取り組んできた課題である。この難しい問題を解決し、コン

ピュータが画像のキャプションを生成する意義は多くある。例えば、画像にキャプションが付いているとキーワードを使って画像を探すことは簡単になり、閲覧による検索が必要なくなる。また、視覚障害者がウェブ上の画像の内容をよりよく理解できる^[1]など、画像にキャプションがついていると、我々は様々な恩恵を受ける。

また、近年の画像認識や、自然言語処理の分野の

発展により、コンピュータにも画像の内容を理解し、説明する事が可能となってきた。特に最近では、画像のキャプション生成を、画像認識と自然言語処理をつなぐ人工知能の基本的な問題としている。近年、いくつかの画像キャプションの生成手法が提案されている。その多くが再帰型ニューラルネットワーク (Recurrent Neural Network)(以下、RNN とする) に基づいており、ニューラルネットワークを活用した機械翻訳の成功からインスピレーションを受けている。画像のキャプション生成は、画像から文に“翻訳”することに類似しているため、機械翻訳のエンコーダー/デコーダーフレームワーク^[2]が適している^[3]。

画像のキャプションを生成する人工知能のモデルの多くは、畳み込みニューラルネットワーク (Convolutional Neural Network)(以下、CNN とする) により画像から特徴を抽出し、抽出された特徴をキャプションに翻訳している。本研究では、画像内容を CNN による画像特徴抽出に加え、画像内容をよく表す単語をヒントワードとして活用する画像キャプション生成手法の提案を目的とする。一方で、提案された画像キャプション生成手法の有用性を検証するために、有益なヒントワードを選択するためのいくつかの方法を提案する。また、有益なヒントワードの選び方についても検証する。

2 関連研究

この章では、本研究に関連する先行研究について紹介する。

近年、様々な画像キャプション生成手法が提案されている。その多くが RNN に基づいており、ニューラルネットワークを用いた機械翻訳の成功からインスピレーションを受けている^[3]。

Kiros ら (2014) が初めて、ニューラルネットワークを画像キャプション生成に活用した。これは、2つのマルチモーダル言語モデルである。彼らのモデルを畳み込みネットワークと共に訓練することによって、単語表現と画像特徴を共同して学習することができることが示されている^[4]。また、Vinyals ら (2014) では RNN として長・短期記憶 (Long Short-Term Memory)(以下、LSTM とする)^[5]を彼らのモデルに活用した^[6]。Vinyals ら (2015) が提案したモデル Neural Image Caption (NIC) では、訓練済

みの CNN から画像の特徴を抽出し、それに続く言語生成 RNN が、画像キャプションを生成する^[1]。

3 提案フレームワーク

この章では、提案するフレームワークについて紹介する。まず 3.1 節で、ヒントワードのコンセプトについて紹介し、次に 3.2 節で、ヒントワードの選び方について紹介する。そして、最後に 3.3 節で、提案モデルを紹介する。

3.1 ヒントワード

本研究では画像の特徴を表している単語を、ヒントワードとして画像キャプション生成に活用する。

3.2 ヒントワードの選び方

本論文では、適切なヒントワードを与えると適切なキャプションを生成するモデルを訓練するために、画像とキャプションのペアデータからなるデータセット、Flickr8k^[7]を用いる。

ヒントワードの品詞がキャプション生成の精度に寄与するという仮説から、名詞、動詞をそれぞれ選びヒントワードとする。また、それらと比較するために、ランダムにヒントワードを選ぶ、名詞をヒントワードとする手法を Hint-NN、動詞をヒントワードとする手法を Hint-VB、ランダムに選んだ単語をヒントワードとする手法を Hint-Random とする。

Hint-NN、Hint-VB の場合、Flickr8k のデータセットの画像に対するキャプションの文に対応する品詞の単語を k 個ランダムサンプリングした単語をヒントワードとする。

Hint-Random の場合、Flickr8k のデータセットの画像に対するキャプションの文から k 個ランダムサンプリングした単語をヒントワードとする。

3.3 提案モデル

画像のキャプションを生成する際に、画像キャプション生成モデルは、画像の特徴量をキャプションへ翻訳している。画像の特徴量として、先行研究では、多くの場合、CNN が抽出した、特徴ベクトルを使用している。しかし、画像には、複数の特徴があ

り、画像キャプション生成の際に、どの特徴に焦点を当てるべきか判断することは、大きな課題となっている。

図 1: 提案する手法のフレームワーク

本研究では、CNN が抽出した特徴ベクトルに加えて、焦点を当てるべき特徴を、ヒントワードとして与えるモデルを提案する。図 1 は、本研究が提案する画像キャプション生成手法のフレームワークである。画像について、CNN を用いて特徴ベクトルを抽出する。また、画像の特徴を表す単語をヒントワードベクトルとする。2 つ合わせた特徴量から、RNN はキャプションを生成する。

図 2 に提案モデルの概要図を示した。画像は訓練済みの VGG CNN モデル [8] から 4096 要素の特徴ベクトルを抽出し、Dense レイヤで 256 要素の画像特徴ベクトルとなる。k 個の選ばれたヒントワードは、Embed レイヤで 256 要素のヒントベクトルとなる。画像特徴ベクトルとヒントベクトルは、Concatenate レイヤで 512 要素の画像表現ベクトルとなる。n 個の単語からなるキャプションは、Embed レイヤで 512 要素のベクトルとなり、LSTM 層に続く。LSTM 層の最後の状態に画像表現ベクトルがマージされる。そして、Softmax レイヤで、キャプションの次の単語をボキャブラリ全体に渡り Softmax 予測する。

図 2: 提案モデルの概要

(1) の式を用いることにより、画像からキャプションを生成する際に、正解となるキャプションとなる

確率を最大化する。

$$\theta^* = \arg \max_{\theta} \sum_{(V,S)} \log p(S|V;\theta) \quad (1)$$

(1) の数式は、 θ はモデルのパラメータ、 V は画像の特徴ベクトルとヒントワードの表現ベクトルを連結したベクトル。 S は正解データのキャプション。 S は様々なキャプションを表現し、それらはそれぞれ長さを持つ。従って、 $\log p(S|V;\theta)$ は、 N を特定の例の長さとする、連鎖律 (Chain rule) が適用され、 $S_0, \dots, S_N$ の連鎖確率となり、

$$\log p(S|V) = \sum_{t=0}^N \log p(S_t|V, S_0, \dots, S_{t-1}) \quad (2)$$

と表される。 $(S|V)$ はペアで訓練される。また、(2) の対数確率の合計は Adam 最適化アルゴリズム [9] で最適化される。

4 実験

この章では 4.1 節で、実験で使用するデータセット、Flickr8k を紹介し、4.2 節で評価手法、BLEU について、4.3 節で実験のフレームワークについて、4.4 節でベースライン手法について紹介する。

4.1 Flickr8k

Flickr8k は画像とキャプションのペアデータからなるデータセットである。8000 枚の画像とキャプションからなるデータセットであり、3 つの部分から構成されている。6000 枚は訓練データ、1000 枚は検証データ、1000 枚はテストデータである。また、1 枚の画像につき、5 つのキャプションが付いている。Flickr8k は何かしらの動きをしている人々、もしくは動物 (多くは犬) に焦点を当てている。画像は Flickr の 6 つの異なったグループから選ばれている。また、有名人や有名な場所は含まない傾向にあるが、シーンや状況はバラエティ豊かになるように選ばれている [7]。

4.2 BLEU

実験の評価指標として BLEU^[10] を使用する。BLEU とは、機械翻訳の自動評価手法である。BLEU は、

機械翻訳結果と参照訳との類似度を、両者の n-gram 一致数を基に、以下の式で算出する。

$$BLEU = BP_{BLEU} \times \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (3)$$

ただし、

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-gram}' \in C'} Count_{clip}(n\text{-gram}')}$$

p_n は、評価コーパス全体について、翻訳文と参照訳を比較し、n-gram の一致率を算出しているものである。また、 $w_n = \frac{1}{N}$ である。まず、文ごとに n-gram の一致を計算する。次に、すべての候補文章について切り取られた n-gram 数を加算し、テストコーパス内の候補 n-gram 数で除算し、テストコーパス全体に対する修正精度スコア p_n を計算する。これを 1-gram から n-gram について幾何平均を求めることにより、スコアを算出する。1-gram は、単語訳の正しさを表す指標となっており、高次の n-gram は、翻訳の流暢さを表す指標であるので、BLEU スコアは両者を組み合わせた指標となっている。また、 BP_{BLEU} は、翻訳文が参照訳より短い場合に与えるペナルティ (brevity penalty) であり、翻訳文が参照訳より長い場合は 1 である。本研究では、単一のコーパスレベルの BLEU スコアを計算する。

このように、BLEU スコアは 0 から 1 の間の実数で表現され、値が高いほど良好な翻訳文であると判断される^[11]。

4.3 実験のフレームワーク

適切なヒントワードを与えると、適切なキャプションを生成するモデルを訓練するために、Flickr8k を用いる。Flickr8k のキャプションからヒントワードを与えると、そのキャプションを生成するように図 2 のモデルを訓練する。モデルを訓練する際に、本研究では TensorFlow¹ のライブラリを用いる。Flickr8k からヒントワードを与える際にヒントワードの数、選び方、また、モデルが使用する語彙群全体の語彙数をどのように設定するとパフォーマンスが向上するか検証する。

語彙群の大きさを変えることによるモデルのパ

フォーマンスの変化を検証するために、訓練データにおいて、閾値より出現頻度の低い単語を “unknown” に置き換える。単語の出現頻度の閾値を 3 から 5 の間で変更する。3, 4, 5 の閾値の場合 Flickr8k の語彙数は、3478, 2918, 2539 となる。単語の出現頻度の閾値、ヒントワードの選び方を、3.2 節で説明した 3 つの方法、ヒントワードの数、1, 2, 3, 4 を組み合わせた 36 パターンの実験を行う。

訓練では、Adam 最適化アルゴリズム^[9] を使用する。損失関数としては、多クラスの交差エントロピーを使用する。36 パターンのモデルは訓練時、検証データのパフォーマンスが低下し始めるとアーリーストッピングする。

各モデルは 3 回に分けてトレーニングを行う。本章の結果は、これら 3 回の別々の実行の平均値である。訓練されたモデルを評価するために、ビーム幅 3, 最大クリップ長 20 単語のビームサーチを使用し、テストデータセット内の画像のキャプションを生成する。生成されたキャプションを評価するために、BLEU スコアを計算し、比較する。

4.4 ベースライン

Marc Tanti らによって発表された論文^[12, 13] にある “merge-model” に基づいたモデルをベースラインとする。

5 結果

5.1 節ではベースラインの結果を、5.2 節では提案手法の結果を紹介する。

5.1 ベースラインの結果

出現頻度の閾値	BLEU-1	BLEU-2	BLEU-3	BLEU-4
3	0.592	0.403	0.266	0.175
4	0.593	0.405	0.268	0.177
5	0.596	0.407	0.269	0.178

表 1: ベースラインの BLEU スコア結果

表 1 は、ベースライン手法の BLEU スコアを示している。左端の列に訓練データにおける単語の出現頻度の閾値を示した。訓練データにおいて、この閾値以上出現している単語をそれぞれのモデルの語

¹<https://www.tensorflow.org/>

図 3: 提案モデルの BLEU スコアの結果

彙群としている。BLEU スコア全てにおいて、単語の出現頻度の閾値が3から5に大きくなるにつれてスコアが高くなった。表1にあるスコアをベースラインのスコアとする。

5.2 提案手法の結果

図3に、3つのヒントワードの種類、1から4のヒントワードの数、3から5の訓練データにおける単語の出現頻度の閾値をそれぞれ組み合わせた36パターンのBLEUスコアを示した。36パターン全てのBLEUスコア結果それぞれがベースラインのBLEUスコア結果を上回った。

3種類のヒントワードの選び方について、それぞれの選び方のBLEUスコアの平均値を比較すると、Hint-VB>Hint-NN>Hint-Randomの順で高い結果となった。ヒントワードの数については、1から4まで多くなるにつれBLEUスコアは高くなった。また、訓練データにおける単語の出現頻度の閾値について、それぞれの閾値のBLEUスコアを比較すると、閾値が4>5>3の順でBLEUスコアが高い結果となり、ベースラインとは異なる結果となった。

36パターン全てのモデルの中で、ヒントワードの選び方がHint-VB、単語の出現頻度が3、ヒントワードの数が4のモデルのBLEUスコアが最も優れていた。

6 考察

表1、図3にあるように提案モデルのBLEUスコアは全てベースラインスコアを上回った。従って、画像キャプション生成におけるヒントワード活用の有用性が示された。ヒントワードの選び方、3種類のBLEUスコアの平均値を比較すると、Hint-VB>Hint-NN>Hint-Randomとなったことから、ヒントワードを選ぶ際に品詞情報が有用であり、中でも動詞が最もキャプション生成の精度に寄与すると示された。また、ヒントワードの数が多くなるにつれ、BLEUスコアが高くなることから、ヒントワードの数のパラメータの影響が確認された。ベースライン手法では、訓練データにおける単語の出現頻度の閾値が3から5に増えるにつれてBLEUスコアが高くなった事に対し、提案手法ではそれぞれの閾値のBLEUスコアを比較すると、閾値が4>5>3の順でBLEUスコアが高い結果となった。これは閾値が大きくなるにつれて語彙群の数が減り、モデルがヒントワードのベクトル表現を正確に理解できないことが増加する事が原因であると考えられる。

7 結論

本研究は、画像キャプション生成において画像内容をよく表す単語をヒントワードとして活用する手法の提案、また、ヒントワードの選び方の検証を目的とした。

実験の結果、最も優れていたヒントワードの選び

方が Hint-VB, 単語の出現頻度が 3, ヒントワードの数が 4 の手法を, 画像内容をよく表す単語をヒントワードとして活用する画像キャプション生成手法として提案する.

ヒントワードを選ぶ際に, 品詞情報が有用であり, 中でも動詞が最もキャプション生成の精度に寄与し, また, ヒントワードの数が 1 から 4 まで増えるにつれ, キャプション生成の精度が向上すると示された.

7.1 今後の展開

本研究では, 画像内容をよく表す単語をヒントワードとして活用する画像キャプション生成手法を提案した. ヒントワードを与える際に本実験では Flickr8k から抽出した. 従って, Flickr8k の画像以外についてはヒントワードを抽出する事が出来ない. 今後は, 画像検索ログデータの画像についても, 対応するクエリからヒントワードを抽出し, キャプション生成を可能にしたい.

参考文献

- [1] Oriol, Vinyals; Alexander, Toshev; Samy, Bengio. “Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge”. IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1-12, 2016.
- [2] K, Cho et al. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation”. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1724-1734, 2014.
- [3] Kelvin, Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. Proceedings of the 32nd International Conference on Machine Learning. pp. 2048-2057, 2015.
- [4] Ryan, Kiros; Ruslan, Salakhutdinov; Richard, Zemel. “Multimodal Neural Language Models”. Proceedings of the 31st International Conference on Machine Learning. pp. 595-603, 2014.
- [5] Sepp, Hochreiter; Jurgen, Schmidhuber. “Long Short-Term Memory”. Neural Computation. 1997, pp. 1735-1780, 1997.
- [6] Ilya, Sutskever; Oriol, Vinyals; Quoc, V. Le. “Sequence to Sequence Learning with Neural Networks”. Proceedings of the 27th International Conference on Neural Information Processing Systems. pp. 3104-3112, 2014.
- [7] Micah, Hodosh; Peter, Young; Julia, Hockenmaier; “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”. Journal of Artificial Intelligence Research 47. pp. 853-899, 2013.
- [8] Karen, Simonyan; Andrew, Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. CoRR. 2014.
- [9] Diederik, P. Kingma; Jimmy, Ba. “Adam: A Method for Stochastic Optimization”. CoRR. 2014.
- [10] Kishore, Papineni; Salim, Roukos; Todd, Ward; WeiJing, Zhu; “BLEU: A Method for Automatic Evaluation of Machine Translation”. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311-318, 2002.
- [11] 今村 賢治; 隅田 英一郎; 松本 裕治. 「機械翻訳自動評価指標の比較」. 第 10 回言語処理学会年次大会. pp. 452-455, 2004.
- [12] Marc, Tanti; Albert, Gatt; Kenneth, P. Camilleri. “What is the Role of Recurrent Neural Networks (RNN) in an Image Caption Generator?”. Proceedings of the 10th International Conference on Natural Language Generation. pp. 51-60, 2017.
- [13] Marc, Tanti; Albert, Gatt; Kenneth, P. Camilleri. “Where to Put the Image in an Image Caption Generator”. Natural Language Engineering. pp. 467-489, 2018.