STAIR Captions: 大規模日本語画像キャプションデータセット

吉川 友也 † 重藤 優太郎 ‡* 竹内 彰一 †
† 千葉工業大学 人工知能・ソフトウェア技術研究センター ‡ 奈良先端科学技術大学院大学
{yoshikawa,takeuchi}@stair.center yutaro-s@is.naist.jp

1 はじめに

1.1 背景

自然言語と画像の統合的な処理は, 近年注目を集めている. 実際に, 2011 年に自然言語言語と画像処理に関するワークショップ (Workshop on Vision and Language) が開催されて以来, このワークショップは毎年開催されている.[1] この研究分野の中でも, 画像に対して説明文 (キャプション) を自動で生成させる試み (image captioning) [2, 6, 8, 11] が大きな注目を集めている.

画像キャプション生成とは, 1 枚の画像を入力してその画像のキャプションを自動的に生成することである. 画像キャプション生成が精度良く可能になると, 画像に対する自然文での検索や, キャプションを音声で出力することで視覚障害者の画像認識支援に役立つ. 画像キャプション生成においては, 多様な画像を認識し, それに対して適切な表現のキャプションを出力するために, 画像とキャプションのペアを大量に用意することが重要である.

1.2 研究目的と貢献

本研究では, 画像キャプションを日本語で生成することを考える. これまでのデータセットのほとんどは英語キャプションで, 日本語キャプションが付与されたデータセットは少ない. これに対する解決策として, 機械翻訳により英語キャプションを日本語に翻訳することが考えられる. しかしながら, 機械翻訳を介することで, 直訳的で不自然なキャプションが生成される可能性がある. したがって, 本論文では, 日本語キャプションデータセットの構築を行い, 自然な日本語キャプションを生成することを目指す.

具体的な本研究の貢献は, 以下のとおりである.

- 日本語キャプションの生成を行うために, 日本語キャプション付きデータセットを構築した (3 節).
- 構築したデータセットに対して, ニューラルネットワークによるキャプション生成法を使用することで, 日本語キャプションが生成できることを確認した (5 節).

本研究において, 構築した日本語キャプションデータセットは http://captions.stair.center/ でダウンロードできる.

2 関連研究

これまで, 画像に対して英語キャプションが付与されたデータセットは, いくつか構築されている. 代表的なものとして, PASCAL [10], Flickr3k [5, 10], Flickr3k を拡張した Flickr30k [12], MS-COCO (Microsoft Common Objects in Context) [1] が存在する.

3 節で詳しく述べるが, 本研究で構築した日本語キャプションデータセットは, MS-COCO[2] が提供している画像に対し日本語キャプションを独自に付与している.

MS-COCO [7] は画像分類, 物体認識や英語キャプション生成の研究用に構築されたデータセットである. データが公開されて以来, 単純な画像分類やキャプション生成などのベンチマークデータセットとして使われるだけでなく, 多くの研究によって, データセット自体の拡張がなされている.[3]

近年では, 英語以外の言語で書かれたキャプションが付与されたデータセットも構築されている [3, 4, 9]. 中でも, 本研究と最も関連のある研究に Miyazaki and Shimizu [9] がある. 彼らは, 本研究とは独立に MS-COCO に対して日本語のキャプションを付与し, そのデータセットを用いて, 日本語キャプション生成の実験を行っている. 3 節において, 本論文で提案するデータセットと彼らが構築したデータセットの比較を行う.

*本研究の一部は, 千葉工業大学 人工知能・ソフトウェア技術研究センターで行ったインターンジップ期間中に行った.

[1] 近年では EMNLP や ACL などの併設ワークショップとして開催されている; https://vision.cs.hacettepe.edu.tr/vl2017/

[2] http://mscoco.org

[3] http://mscoco.org/external/

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

図 1: キャプション付与の作業画面イメージ

3 STAIR Captions

3.1 構築方法

この節では, STAIR Captions をどのように構築したのかを説明する. STAIR Captions では, MS-COCO 2014 年版の訓練･開発･テスト用データの全ての画像 164,062 枚を取得し, これらをキャプション付与の対象とした. 各画像に対して, 5 つのキャプションが付与した. したがって, キャプションの総数は 820,310 個である. ただし, MS-COCO を用いて作成したデータの公開ルールに基づき, 公開データセットではテスト画像に付けられたキャプションデータを除いている.

効率的にキャプション付与作業を行うために, まず最初にキャプション付与専用の Web アプリケーションを開発した. 図 1 は, その Web アプリケーションの作業画面イメージを示す. 作業者は, 表示されている画像を見て, 説明文を画像の下のテキストボックスに記入する. そして, 送信ボタンを押すことで一つの作業が完了し, 次の画像が表示される.

上記アプリケーションを用いてキャプション付与を並行的かつ安価に行なうために, アルバイトやクラウドソーシングのワーカーに作業を依頼した. キャプション付与に際して, 作業者には以下のガイドラインを伝えた. (1) 15 文字以上で説明すること, (2) 「である」調で書くこと, (3) 画像に映っていないことを想像して書かないこと, (4) 単文で書くこと, (5) 感情, 意見を書かないこと. また, キャプションの質を担保するために, 作成されたキャプションに対しては抜き取り検査を行い, ガイドラインに沿わないキャプションは排除した. この作業は, 約 2,100 人のワーカーによって約半年間で行った.

表 1: データセットの統計量比較. () 内は公開データセットのみの統計量を示す.

	STAIR Captions	YJ! Captions
画像数	164,062 (123,287)	26,500
キャプション数	820,310 (616,435)	131,740
語彙数	35,642 (31,938)	13,274
平均文字数	23.79 (23.80)	23.23

3.2 統計量

この節では, STAIR Captions の定量的な性質を紹介する. その際, STAIR Captions と同様に MS-COCO の画像に対して日本語キャプションを付与した YJ! Captions [9] との比較を行なう.

表 1 は, データセットの統計量を示す. データセット全体では, STAIR Captions は YJ! Captions の 6.19 倍の画像に対して, 6.23 倍のキャプションが付けられている. また, 公開データセットでは, 画像数は 4.65 倍, キャプション数は 4.67 倍である. 画像数やキャプション数が多いことは, テストにおいて未知のシーンやオブジェクトが現れる可能性を減らすことができるため, キャプションの自動生成において重要な点である. 語彙数に関しては, STAIR Captions は YJ! Captions の 2.69 倍である. 語彙が多いことで, キャプション生成モデルが多様な説明文を学習・生成できるようになることが期待される. 平均文字数は, STAIR Captions は YJ! Captions はほとんど同じである.

4 ニューラルネットワークに基づくキャプション生成法

ここでは, 日本語キャプション生成の評価実験 (5 節) で用いる, ニューラルネットワークに基づくキャプション生成法を説明する. 本論文では, Karpathy and Fei-Fei [6] が提案したキャプション生成法を用いる.

Karpathy and Fei-Fei が提案したキャプション生成法は, convolutional neural network (CNN) と long short term memory (LSTM)[4] により構成される. 具体的には, まず, CNN を用いて画像の特徴を抽出し, その後, 抽出

[4]彼らの原論文では LSTM ではなく, RNN を用いているものの, appendix において, 投稿後に RNN と LSTM の性能比較を行った結果, LSTM の方が良い結果を得ることを報告している. そのため, 本論文では RNN ではなく LSTM を用いる.

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

された特徴を LSTM に入力することでキャプションを生成する.

実際には以下の式により, 画像 I からキャプション $Y = (y_1, y_2, \cdots, y_n)$ を生成する.

$$\begin{aligned}
\mathbf{x}_{im} &= \mathrm{CNN}(I) \\
\mathbf{h}_0 &= \tanh\left(\mathbf{W}_{im}\mathbf{x}_{im} + \mathbf{b}_{im}\right) \\
\mathbf{c}_0 &= \mathbf{0} \\
\mathbf{h}_t, \mathbf{c}_t &= \mathrm{LSTM}\left(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}\right) \qquad (t \geq 1) \\
y_t &= \mathrm{softmax}\left(\mathbf{W}_o\mathbf{h}_t + \mathbf{b}_o\right)
\end{aligned}$$

式中の CNN(·) は, CNN によって抽出された特徴ベクトル (CNN の最終層) を出力する関数である. また, y_t は t 番目の単語であり, 時刻 t における LSTM の入力 $\mathbf{x}_t$ には, 時刻 $t-1$ で予測された単語 y_{t-1} に対応する単語ベクトルを入力する. この y_t の出力は, 最終的に文の終わりを表す終端記号が出力されるまで繰り返す.

訓練時には, 画像 I_i とキャプション Y_i のペアの集合 $\{(I_i, Y_i)\}_{i=1}^{N}$ が与えられ, これを用いて, $\mathbf{W}_*, \mathbf{b}_*$[5] に加えて, CNN と LSTM のパラメータを学習する.

5 実験

この節では, 構築した日本語キャプションデータセットを用いて, 実際にキャプション生成を行う. 本実験の目的は, 日本語キャプションデータセットの必要性を示すことである. また同時に, 既存のニューラルネットワークによるキャプション生成法により, 日本語のキャプション生成がどの程度できるかを検証する.

5.1 実験設定

評価指標. 先行研究 [1, 6] の設定に従い, 本実験では, BLEU, ROUGE, CIDEr を評価指標として用いる. 本来, BLEU は機械翻訳, ROUGE は要約のための評価指標であるが, キャプション生成の評価指標としてよく用いられているため, 本論文でも評価指標として採用する.

比較手法. 本実験の目的は, 日本語キャプションデータセットの必要性を示すことなので, 訓練に日本語キャプションデータを使用しない方法をベースラインとする. 具体的には, 機械翻訳を用いて英語キャプションを日本語に翻訳する.

本実験では, 以下の 2 種類のキャプション生成法を比較する.

- MS-COCO + MT: 英語キャプション生成と機械翻訳のパイプライン. はじめに英語キャプションデータ (MS-COCO) を用いて, 英語キャプションを生成するニューラルネットワークを学習する. 評価時には, 学習したニューラルネットワークを用いて英語キャプションを生成し, それを機械翻訳で日本語に翻訳する. 英語から日本語への翻訳には, Google 翻訳を用いた. この方法が本実験のベースラインである.
- STAIR Captions: 3 節で述べた日本語キャプションデータセットを用いてニューラルネットワークを学習したもの. 上述した手法とは異なり, 機械翻訳は用いない.

繰り返しになるが, ベースライン (MS-COCO + MT) は, 訓練時に, 新たに構築した日本語キャプションデータセットを使用しない.

どちらの方法においても, Karpathy and Fei-Fei [6] に従い, CNN を固定し, LSTM のみを学習させた. CNN は, 公開されているパラメータ (VGG[6]) を使用した. LSTM の学習は, ミニバッチ, RMSProp を用いた. この時, バッチサイズを 20 とした.

データセット分割. MS-COCO を用いたキャプション生成に関する研究 [1, 6] の実験設定に従い, MS-COCO が定めた訓練画像のうち 111,091 枚 (とそれに対応する日本語キャプション) を本実験の訓練データとし, 残りを二分割し, それぞれ開発データ (5,000 枚) と評価データ (5,000 枚) とした. 開発データを用いてパラメータ調整を行った. その際, 最良のパラメータは CIDEr を基準に決定した. 日本語キャプションは, 形態素に分解し使用した. 形態素解析には, MeCab[7] を用いた.

5.2 実験結果

表 2 に実験結果を示す. 全ての指標において, 日本語キャプションを用いない場合 (MS-COCO + MT) より, 日本語キャプションを訓練データとして用いる方法 (STAIR Captions) が良い結果となった.

表 3 は, MS-COCO + MT では不自然なキャプションが生成されたが, STAIR Captions では尤もらしいキャプションが生成された例を示す. 表 3 上部の例では, MS-COCO + MT は, “A double decker bus” を直訳し “二重デッカーバス” として, 不自然なキャプションを生

[5] ∗ はワイルドカードを表す.

[6] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

[7] http://taku910.github.io/mecab/

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 日本語キャプション生成の実験結果. 各評価指標で最も良い数値を, 太字で表している.

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE_L	CIDEr
MS-COCO + MT	0.565	0.330	0.204	0.127	0.449	0.324
STAIR Captions	**0.763**	**0.614**	**0.492**	**0.385**	**0.553**	**0.833**

表 3: 生成されたキャプションの具体例.

MS-COCO:
A double decker bus driving down a street.
MS-COCO + MT:
ストリートを運転する二重デッカーバス。
STAIR Captions:
二階建てのバスが道路を走っている。

MS-COCO:
A bunch of food that are on a table.
MS-COCO + MT:
テーブルの上にある食べ物の束。
STAIR Captions:
ドーナツがたくさん並んでいる。

成している. これは, 機械翻訳を用いると直訳的で不自然なキャプションになる典型的な例である. 一方で, STAIR Captions では, 適切に “二階建てのバス” と表現し, 自然なキャプションが生成できた. 表 3 下部の例では, MS-COCO + MT は大量のドーナツを “食べ物の束” と表現し, 不適切なキャプションとなった. 一方で, STAIR Captions では, 画像に写る食べ物をドーナツと正しく表現できた.

構築したデータセットにおいて, 画像に存在しない物体を想像したり, 人物に対してセリフをつけるなど, 客観的な画像の説明になっていないキャプションがいくつか存在した. 訓練データにこのようなキャプションが存在することは, キャプションの自動生成に対して悪影響を与えることが懸念されたが, それらの数が少ないせいか, 本実験の評価データに含まれていた画像に対して, そのようなキャプションは生成されなかった.

6 おわりに

本論文では, 新たに日本語キャプションデータセット STAIR Captions を構築した. このデータセットは, MS-COCO が提供している画像に対して, 日本語キャプションが付与されており, 合計 820,310 文ある.

本実験では, 日本語キャプションの必要性を示すため, 日本語キャプションを用いない場合と用いる場合でキャプション生成の比較を行った. その結果, 構築したデータセットの必要性を示すことができた. また, 既存のキャプション生成手法を単純に適応するだけで日本語キャプションが生成できることを確認できた.

今度は, 付与した日本語キャプションと英語キャプションを用いて, 二言語の情報を利用したキャプション生成法を開発する予定である.

謝辞 データセット構築に協力していただいた千葉工業大学の学生有志, 株式会社 mokha 蒲地氏, スケールアウト株式会社 荻野氏に感謝致します.

参考文献

[1] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint*, 1504.00325, 2015.

[2] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[3] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. In *Workshop on Vision and Language*, pages 70–74, 2016.

[4] M. Grubinger, P. D. Clough, H. MÃijller, and T. Deselaers. The IAPR Benchmark: A new evaluation resource for visual information systems. In *LREC*, 2006.

[5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[7] T. Lin, M. Maire, S. Belongie, J. Hays, and P. Perona. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (M-RNN). In *ICLR*, 2015.

[9] T. Miyazaki and N. Shimizu. Cross-lingual image caption generation. In *ACL*, pages 1780–1790, 2016.

[10] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[12] P. Young, A. Lai, M. Hodosh, and J. 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.

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.