一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IMQ2019-48,IE2019-130,MVE2019-69(2020-03)

心像性に基づく画像キャプショニングの検討

梅村　和紀[†]　カストナーマークアウレル[†]　井手　一郎[††,†]　川西　康友[†]　平山　高嗣[†††]
道満　恵介[††††,†]出口　大輔[†]　村瀬　洋[†]

† 名古屋大学 大学院情報学研究科 〒464-8601 愛知県名古屋市千種区不老町
†† 名古屋大学 数理・データ科学教育研究センター 〒464-8601 愛知県名古屋市千種区不老町
††† 名古屋大学 未来社会創造機構 〒464-8601 愛知県名古屋市千種区不老町
†††† 中京大学 工学部 〒470-0393 愛知県豊田市貝津町床立 101
E-mail: † {umemurak, kastnerm}@murase.is.i.nagoya-u.ac.jp,
†† ide@i.nagoya-u.jp, † {kawanishi, ddeguchi, murase}@i.nagoya-u.ac.jp
††† takatsugu.hirayama@nagoya-u.jp, †††† kdoman@sist.chukyo-u.ac.jp

あらまし　様々な場面や用途に応じた画像キャプショニングを目指し，文の心像性に基づく画像キャプショニング手法を提案する．まず，既存の画像キャプションデータセットを拡張して，様々な抽象度のキャプションをもつ画像キャプションデータセットを構築する．そして，そのデータセット中のキャプションに対して心像性を算出することで，心像性スコア付きキャプションデータセットを構築する．それを用いてキャプショニングモデルを学習し，指定した心像性をもつ画像キャプションを生成する枠組みを構築する．評価実験では，画像キャプションデータセットの拡張の有効性と提案手法による画像キャプショニング実現の可能性を確認した．
キーワード　マルチメディア処理，画像キャプショニング，心理言語学，セマンティックギャップ

A study on image captioning considering its imageability

Kazuki UMEMURA[†], Marc A. KASTNER[†], Ichiro IDE[††,†], Yasutomo KAWANISHI[†], Takatsugu HIRAYAMA[†††], Keisuke DOMAN[††††,†], Daisuke DEGUCHI[†], and Hiroshi MURASE[†]

† Graduate School of Informatics, Nagoya University
†† Mathematical and Data Science Center, Nagoya University
††† Institutes of Inovation for Future Society, Nagoya University
†††† School of Engineering, Chukyo University
E-mail: † {umemurak, kastnerm}@murase.is.i.nagoya-u.ac.jp,
†† ide@i.nagoya-u.jp, † {kawanishi, ddeguchi, murase}@i.nagoya-u.ac.jp
††† takatsugu.hirayama@nagoya-u.jp, †††† kdoman@sist.chukyo-u.ac.jp

Abstract　We propose an imageability-aware image captioning method tailoring generated captions to various applications. In this study, we first extend an existing image captioning dataset by augmenting its captions. Then, an imageability score for each augmented caption is calculated. A modified image captioning model is trained using this extended dataset to generate captions tailored to a specified imageability score. The evaluation shows the possibility that the extended dataset and the proposed method can generate imageability-aware captions.
Key words　Multimedia proccesing, image captioning, psycholinguistics, semantic gap

1. はじめに

近年，画像処理技術ならびに自然言語処理技術は急速に発達している．そして，それらの技術を組み合わせ，画像内容を描写したキャプションを生成する「画像キャプショニング」の技術も，目ざましい発展を遂げている．一方，画像キャプションは様々な場面で利用されるが，その性質や特徴は状況によって異なる．現在の画像キャプショニング技術では，このような状

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2020 by IEICE

表 1　生成したキャプションの例.

画像	心像性	生成したキャプション.
	0.1	A brown and white dog sitting on a bench.
	0.2	A brown and white dog standing next to a bike.
	0.3	A dog laying on the ground next to a bike.
	0.4	A dog sitting on a leash on a bike.
	0.5	A dog sitting in front of a red door.
	0.6	A brown and white dog standing next to a red container.
	0.7	A brown and white dog sitting on a leash.
	0.8	A brown and white dog sitting on a leash.
	0.9	A brown dog standing next to a red container.

について述べる.

画像キャプションの自動評価指標を用いて生成したキャプションを評価する．具体的には，BiLingual Evaluation Understudy（BLEU）[11] を採用する．BLEU は機械翻訳における翻訳文の自動評価指標であり，生成文と参照文との N-gram 適合率で行なうが，ここでは $N = 4$ とする.

文の心像性推定には，既存の単語心像性辞書 [7], [12] を用いる．ここで，いずれの心像性辞書にも含まれる語に関しては，Scott ら [12] の心像性を優先的に用いることとする．なお，いずれの心像性辞書にも含まれない語（未知語）を含むキャプションは画像キャプションデータセットから除去する．また，拡張後のキャプションデータ数が 10 文に満たない画像も画像キャプションデータセットから除去する．キャプショニングモデルの学習には画像 1 枚あたり 10 文のキャプションデータを用いるが，それ以上ある場合は 10 文を選択して使用する．MS COCO データセット中の学習・評価・テストデータの参照には，Karpathy splits [5] を用いる．具体的には，学習データには画像 113,287 枚，評価データには 5,000 枚，テストデータには 5,000 枚を用いる．上記データセット中から以上の基準で画像データを除去した後，学習データは画像 113,235 枚，評価データは 4,999 枚，テストデータは 4,998 枚となった．以降，このデータセットを用いる.

比較手法として，データ拡張しないデータセットを用いて，心像性を考慮しないキャプショニング手法，データ拡張したデータセットを用いて，心像性を考慮しないキャプショニング手法，及びデータ拡張しないデータセットを用いて，心像性を考慮したキャプショニング手法の 3 つの手法を用意した．なお，データ拡張しないデータセットを用いる際には，画像 1 枚あたり 5 文のキャプションを用いる.

3.2　実験結果

まず，画像キャプションの自動評価指標による評価結果を表 2 に示す．心像性を考慮する場合としない場合いずれの場合においても，拡張したデータセットによるキャプショニングの方が高精度であった.

最後に，生成した画像キャプションの例を表 1 に示す．キャプション生成の際に指定した心像性，それぞれの心像性において生成したキャプションを表している.

表 2　自動評価指標による実験結果.

データセット	BLEU-4
元データセット（心像性なし）	23.3
拡張データセット（心像性なし）	25.9
元データセット（心像性あり）	24.5
拡張データセット（心像性あり）	25.2

3.3　考　　察

画像キャプションの自動評価指標による評価実験について考察する．表 2 より，データ拡張により精度が向上したと考えられる．そのため，提案手法によるデータ拡張は有効であったと考えられる.

なお，本実験ではキャプションの自動評価指標として BLEU を用いたが，単に生成したキャプション中の語彙と評価データ中のキャプションの語彙の一致割合を評価しているに過ぎない．そのため，本研究のように様々な語彙を含んだキャプションを生成するタスクに適した評価指標の使用を検討する.

4.　ま　と　め

本報告では，文の心像性に基づく画像キャプショニング手法を提案した．具体的には，既存の画像キャプションデータセットを拡張して，様々な抽象度のキャプションをもつ画像キャプションデータセットを構築した．そして，その画像キャプションデータセット中のキャプションに対して心像性を算出することで，心像性スコア付きキャプションデータセットを構築した．更に，それを用いてキャプショニングモデルを学習し，指定した心像性をもつ画像キャプションを生成する枠組みを構築した.

評価実験により，提案手法による心像性に基づく画像キャプショニングの可能性を示した．また，画像キャプションデータセットにおけるデータ拡張の有効性を確認した.

今後は，他の評価指標による評価実験と，被験者による定性的な評価実験を行なう必要がある．また，精度の向上を目指し，手法の改善を検討していく.

謝　　辞

本研究の一部は，科学研究費補助金による.

文　　献

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

[2] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3137–3146, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[5] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[6] 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 *Proc. European Conference on Computer Vision*, pp. 740–755, 2014.

[7] Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proc. 3rd Workshop on Representation Learning for NLP*, pp. 217–222, 2018.

[8] Alexander P. Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proc. 30th AAAI Conf. on Artificial Intelligence*, pp. 3574–3580, 2016.

[9] George A. Miller. WordNet: A lexical database for English. *Comm. the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[10] Allan Paivio, John C Yuille, and Stephen A Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psycho.*, Vol. 76, No. 1, pp. 1–25, 1968.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.

[12] Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. *The Glasgow Norms: Ratings of 5,500 Words on Nine Scales*. Springer, 2018.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. 32nd Int. Conf. on Machine Learning*, pp. 2048–2057, 2015.

[15] **梅村和紀**, **カストナーマークアウレル**, **井手一郎**, **川西康友**, **平山高嗣**, **道満恵介**, **出口大輔**, **村瀬洋**. **画像キャプションの質的評価に向けた文の心像性推定手法の検討**. **言語処理学第** 25 **回年大**, A4-9, 2019.