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

心像性スコア
0.7 → A boy is riding a snowboard
0.3 → A person is standing on the ground

図 1　本手法により生成したいキャプションの例.

況に応じたキャプション生成はできないため，依然実用化には程遠い．例えば，視覚障害者が画像内容を認識する際には，少しでも的確かつ詳細に画像内容を描写したキャプションが必要とされる．一方，ニュース記事では画像とキャプションが同時に提示されるため，キャプションが画像内容を詳細に描写する必要はなく，むしろニュース記事の内容をふまえた簡潔なものほど適切である．

我々は，このような様々な用途の状況に応じた画像キャプションの生成を目指している．その中で，心理言語学における概念である「心像性」に着目した．心像性とは，単語概念の想像しやすさを指す [10]．我々はこれまで，この単語の心像性に基づいて，文の心像性を算出する方法について検討してきた [15]．本研究では，このような文の心像性に基づく画像キャプション生成手法を検討する．

近年，画像キャプショニングに関して様々な研究が行なわれている．一般には，入力した画像に対して，畳み込みニューラルネットワーク（CNN）を用いて画像特徴量を抽出し，再帰型ニューラルネットワーク（RNN）によりキャプションを生成する手法 [13] が主流である．また，画像中のアテンション情報を考慮した手法 [14] も提案されている．

一方，様々な性質のキャプション生成に関する研究が行なわれている．Mathews ら [8] は，画像を的確に描写したうえで，“positive” や “negative” といった印象をふまえたキャプショニング手法を提案した．Gan ら [2] は “humorous” や “romantic” といった人間の感情を含めた魅力的なキャプショニング手法を提案した．これらのように，印象や感情を加味した画像キャプショニング手法が提案されている．

これらに対して，我々は心像性を考慮した画像キャプショニング手法を提案する．本手法により生成したいキャプションの例を図 1 に示す．既存のキャプショニングモデルはラベルの有無という離散値を入力とするが，本研究では心像性という連続値を入力とし，それに対応したキャプショニングを実現する．

以降，2. で提案手法の詳細について述べる．次に，3. では提案手法により生成された画像キャプションに関する評価実験を行なう．最後に 4. で本報告をむすび，今後の課題について述べる．

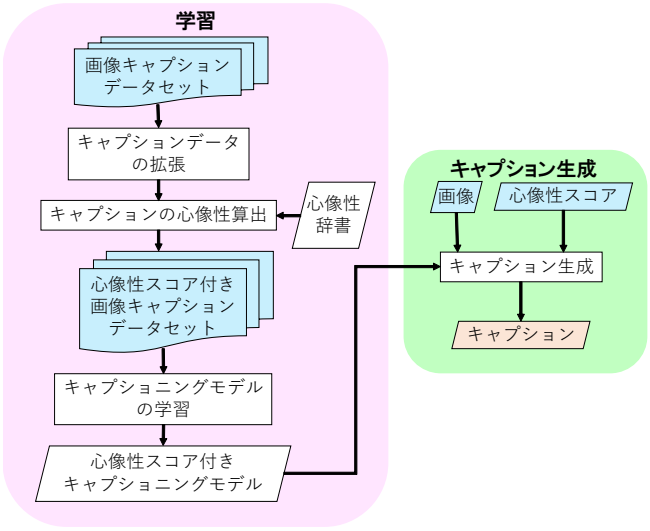


図 2　提案手法の処理手順.

## 2. 文の心像性に基づく画像キャプショニング手法

本節では，文の心像性に基づく画像キャプショニング手法について説明する．まず，キャプショニングモデルの学習に必要な心像性スコア付き画像キャプションデータセットを構築する．そのためにまず，既存の画像キャプションデータセット中の各キャプションの単語を置換することでキャプションデータを拡張し，それらの心像性を算出する．そして，心像性スコアを特徴ベクトルとみなし，既存の画像キャプショニングモデルを拡張することで，文の心像性に基づいた画像キャプショニングモデルを学習する．キャプション生成時には，このキャプショニングモデルに対して画像と心像性スコアを入力することで，必要なレベルの心像性をもったキャプションを生成する．提案手法の処理手順を図 2 に示す．以降，各処理について順に説明する．

### 2.1 画像キャプションデータセットの拡張

一般に画像キャプションデータセットの構築は人手により行なわれるが，それには多大なコストを要する．そこで同一の画像に対する様々な心像性スコアのキャプションを生成するために，既存の画像キャプションデータセットを拡張する．

既存の画像キャプショニングデータセットには，1 枚の画像に対して複数のキャプションが付与されているものもあるが，より多様な心像性をもったキャプションを生成するために，データセット中のキャプションに対し，名詞をその上位語に置換することで，元のキャプションよりも抽象的なキャプションを生成する．

ここでは，画像 1 枚あたり 5 つのキャプションが付与された MS COCO データセット [6] を用いる．まず，各キャプション中の名詞について，WordNet [9] の木構造を用いて，その上位語を探索する．そして，各名詞についてそれぞれ最大 5 階層上まで置換し，キャプションを生成する．その際，キャプション中の各名詞を同時に置換するのではなく，1 つずつ順に上位

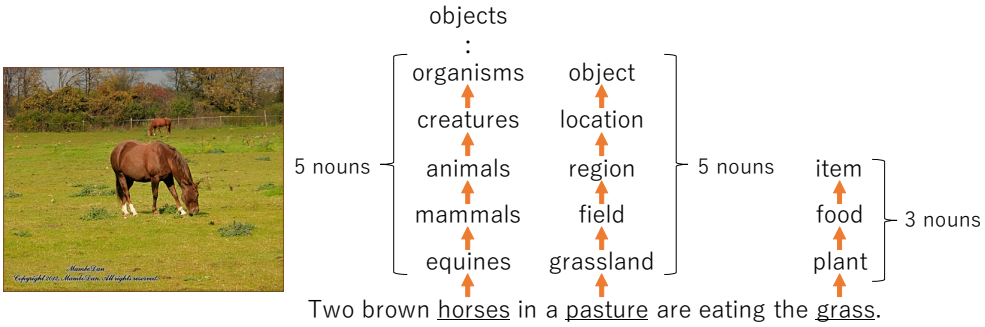


図 3　上位語への置き換えによる学習用キャプションデータ拡張の例.

語に変換して，キャプションを生成する．その理由として，同時に複数の名詞を上位語に置換すると，元のキャプションとは意味的に大きく異なったキャプションが生成されてしまうのを避けるためである．なお，WordNet の上位語を探索する際に，複数の選択肢がある場合には，1 つ目に見つかったノードを採用することにした．また，根ノードに達した場合には，そこで置換を打ち切る．以上の手法により，キャプション 1 つにつきキャプション内名詞数の最大で 5 倍の数のキャプションが生成される．また，入れ替えの際，単数形複数形は元の形に対応させる．このようなキャプションデータの拡張例を図 3 に示す．このようにして，既存の画像キャプションデータセットを拡張することで，様々な心像性をもったキャプションからなるキャプションデータセットを構築する．

### 2.2　文の心像性推定

我々は，構文木を用いて，ルールベースで既存の単語心像性を組み合わせることで，文の心像性を算出する手法を提案した [15]．ここでは，この手法を改良した手法を使用して文の心像性を推定する．まず，構文解析により構文木を生成する．そして，葉ノードからボトムアップに心像性スコアを組み合わせて算出する．その際兄弟ノード間の関係性により異なる算出方法をとる．そこで，兄弟ノード間が修飾関係にある場合，修飾語が被修飾語の心像性を増加させるという仮定に基づき，式 1 を用いて心像性スコア $I$ を算出し，それら兄弟ノードの親ノードのスコアとする．なお，$x_i (i = 1, \ldots, n \,|\, i \neq m)$ は兄弟ノードの心像性スコア，$x_m$ は被修飾語の心像性スコアとする．ここで，$n$ は兄弟ノード数である．

$$I = x_m \prod_{i=1(\neq m)}^{n} f(x_i) \tag{1}$$

$$f(x) = 2 - e^{-x} \tag{2}$$

また，兄弟ノードに等位接続詞を含み，兄弟ノード間が並列関係にある場合，全ノードのスコアの和を親ノードのスコアとする．兄弟ノードがない場合，そのノードのスコアをそのまま親ノードのスコアとする．

最終的に，根ノードに達した際のスコアに対し，式 3 を適用することで [0,1] の範囲に正規化し，それを文の心像性とする．

$$g(x) = 1 - e^{-x} \tag{3}$$

以上の手法により，2.1 節で構築した拡張画像キャプションデータセット中の全キャプションに対して心像性を算出する．このようにして構築した心像性付き画像キャプションデータセットを用いて，画像キャプショニングモデルを学習する．

### 2.3　画像キャプショニング

既存のアテンションを考慮した画像キャプショニング手法 [14] を利用する．この手法において，アテンション付き画像特徴，単語特徴と心像性特徴を結合し学習する．そして最終的に，画像と心像性を入力し，それらに基づくキャプショニングを実現する．本手法は式 4 のように定式化することができる．ここで，キャプション $c_i = \{w_0, w_1, ..., w_N\}$，$w_i$ は $i$ 番目の単語ベクトルとし，$I_t$ はアテンション付き画像特徴ベクトル，$\mathrm{IA}_t$ は心像性特徴ベクトルである．$W_e, W_i$ は学習パラメータである．

$$\begin{aligned} x_t &= W_e * w_t \\ o_t &= \mathrm{LSTM}(\mathrm{concat}(x_t, I_t, \mathrm{IA}_t)) \\ w_{t+1} &= \mathrm{softmax}(W_i O_t) \end{aligned} \tag{4}$$

アテンション付き画像特徴ベクトル $I_t$ の構成方法を以下の式 5 に表す．

$$\begin{aligned} I_f &= \mathrm{CNN}(I) \\ I_t &= \mathrm{Att}(h_{t-1}, I_f) \end{aligned} \tag{5}$$

$I$ を入力画像とし，$I_f$ は ImageNet [1] で学習済みの Resnet [3] で抽出された画像特徴，$h_{t-1}$ は LSTM [4] における 1 ステップ前の隠れ状態を表す．

次に，心像性特徴ベクトル $\mathrm{IA}_t$ の構成手法について述べる．まず，2.2 節で算出した文の心像性について，$[0, 1]$ の範囲に正規化されたスコアを $[-1, 1]$ の範囲に変換する．次に，画像特徴とキャプション中の単語特徴と次元数を揃えるため，心像性スコアを 512 次元に次元拡張する．この 512 次元の心像性スコアを心像性特徴ベクトルとし，使用する．

## 3.　評 価 実 験

本節では，拡張した画像キャプションデータセットを用いた画像キャプショニング手法に関する評価実験を行なう．本実験では，一般的な画像キャプションの評価指標を用いて評価する．

### 3.1　実 験 方 法

本節では，生成した画像キャプションの評価実験の実施方法

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.