単語の階層関係に基づくデータ拡張を利用した画像キャプション生成の検討

吉田智哉 [1]　西村太一 [1]　亀甲博貴 [2]　森信介 [2]
[1] 京都大学大学院 情報学研究科　[2] 京都大学 学術情報メディアセンター
{yoshida.tomoya.25h,nishimura.taichi.43x}@st.kyoto-u.ac.jp
{kameko,forest}@i.kyoto-u.ac.jp

概要

近年の画像キャプション生成モデルは、他の画像に対しても同様に当てはまるような、一般化されたキャプションを生成する傾向がある。この低識別性の問題を解決する策の一つとして、より詳細なキャプションを持ったデータセットで学習させることが挙げられる。本研究では、識別性が単語の階層関係に依存すると仮定し、単語の階層関係に基づくデータ拡張を提案する。実験の結果、低リソースデータセット下での実験において、提案手法が生成する語彙・文を増加させるとともに、識別性を向上させることが確認できた。

1　はじめに

画像キャプション生成は、入力された画像からその画像の説明文を生成するタスクである。このタスクの主な応用先として、視覚障害者の補助 [1] や、画像検索エンジンにおけるクエリの生成などが挙げられる。これらの応用先において重要なことは、生成されるキャプションが自然な文章であるとともに、他の似た画像と区別可能であるということである。近年の画像キャプション生成モデルでは、後者の識別性の低さが問題視されている [2]。この低識別性の問題を解決するために、様々な手法が提案されている [2, 3, 4]。また、解決策の一つとして、より詳細なキャプションを持ったデータセットで学習を行うことが挙げられる。しかし、画像とそれに対応する詳細なキャプションを人手で記述することは、多くの時間的・金銭的なコストを要する。本研究では、これらの作業を自動化するために、**データ拡張**に着目する。

低識別性の問題に取り組むにあたり、キャプションの識別性について再検討を行う。キャプションにおける識別性の高さには、いくつかの観点が考えられる。一つ目は、画像中に出現する多くの物体について記述しているかという点である。先行研究の多くはこの観点に基づいている [2, 3]。二つ目は、画像中の物体について述べる単語が特徴的であるかという点である。これは、単語の階層関係に基づいている。単語の階層関係とは、単語の上位下位関係を指し、一般に下位に向かうにつれ、その単語は特徴的なものになる。((例) 哺乳類、動物、犬、プードル)。その他にも、単語の修飾関係などの観点が考えられる。本研究では、二つ目の観点により低識別性の問題が対処可能であると仮定し、**単語の階層関係に基づくデータ拡張**の提案を行う。

実験では、元データセットが高リソースである場合と、低リソースである場合に対して、提案手法を適用した。実験の結果、高リソース下での実験において、有効であることは確認できなかった。しかし、低リソース下での実験において、生成に使用する語彙・文の増加とともに識別性を向上させ、有効であることが確認できた。また、提案手法が予期しないキャプションの生成に寄与してしまっていることが確認でき、提案手法の課題についても明らかになった。

2　関連研究

2.1　識別性改善に取り組んだ手法

画像キャプション生成モデルにおける低識別性の問題を解決するために、様々な手法が提案されている。Wang らは、学習済みの画像テキスト検索モデルを利用することで識別性を評価する指標である CIDErBTW [3] を提案した。また、この指標を最適化関数に取り入れることにより、既存モデルの識別性を向上させることに成功した。その他にも、強化

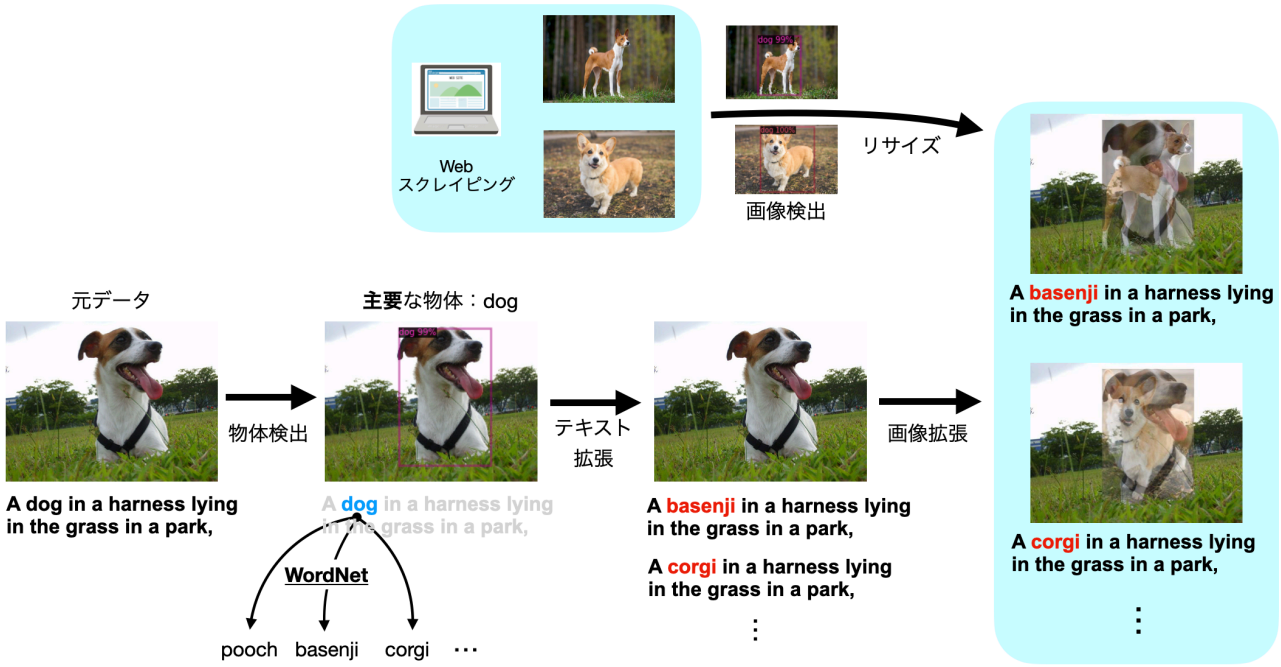

図 1　データ拡張によって作成されたデータセットの例。元データ（左）は、上記の流れにより複数のデータに拡張される（右）。元画像における主要な物体（青字）は、それぞれの拡張先において、下位語（赤字）に置き換えられる。

学習時 [5] に CLIP-Score [6] を報酬として利用することで、識別性の高いキャプションの生成に成功した手法 [2] などが挙げられる。

2.2　データ拡張を利用した手法

データ拡張は画像認識や音声認識の分類タスクにおいて、頑健性や精度の向上を図る手法 [7] として利用されている。近年の拡張手法は、単に入力情報に対してクロッピングやノイズを適用する手法のみならず、Mixup [7] や CutMix [8] と呼ばれるデータセット中のいくつかのサンプルを混ぜ合わせることにより拡張する手法なども提案されている。画像キャプション生成においても、データ拡張を利用した研究は行われている。画像に対してぼかしを施し拡張することにより、モデルの頑健性を高めた手法 [9] や、学習済みの BERT [10] を利用したテキスト拡張により精度を向上させた手法 [11] などが挙げられる。

3　提案手法

本研究では、単語の階層関係を利用することで、低識別性の問題が対処可能であると仮定し、単語の階層関係に基づくデータ拡張を提案する。提案手法により拡張されたデータセットは図 1 のようになった。

3.1　拡張手法

はじめに、データセット中の画像に対して Faster R-CNN [12] を用いて物体検出[1)]を行い、その画像中における主要な物体を定義する。主要な物体とは、人間がある画像について述べる際に着目する物体のことである。今回は、主要であることを、物体の大きさと物体認識にかけられた際の物体の信頼度の積の値が最大のものと定義することにより決定する。

テキスト拡張　テキスト拡張を行うにあたって、WordNet における、単語の階層関係を利用した名詞の置き換えを検討する。WordNet [13] は、単語が類義語ごとにグループ化されている大規模な概念辞書であり、各単語は階層関係を保持している。テキスト拡張は、正解キャプション中の主要な物体に該当する単語を、コーパス中の下位語と入れ替えることで行う。ただし、人に該当する “person” ラベルに関しては、適切でない変換候補[2)]であるため、処理を行っていない。

画像拡張　画像拡張を行うにあたって、Mixup に基づいた画像拡張を検討する。Mixup [7] とは、データセット中における 2 つの訓練サンプルを混ぜ合わせることにより、新たなデータセットを作成する

1）https://github.com/facebookresearch/detectron2 を利用。
2）“black” や “white” などを避ける、倫理観に基づいた判断。

表 1　高リソース下における提案手法とベースラインの比較

		Uniq-l	Uniq-S	B-4	M	R	C	CLIP-S	R@1	R@5	R@10
MS COCO	Att2in	742	3662	**31.6**	**25.9**	**54.4**	**102.3**	**74.1**	**13.1**	**32.9**	**45.4**
	+MixAug(Ours)	**758**	**3706**	30.3	25.6	53.8	99.3	74.0	12.8	31.9	44.3
Flickr30k	Att2in	**397**	**877**	**17.0**	**16.8**	**41.5**	**34.1**	**67.6**	**16.5**	**36.0**	**48.6**
	+MixAug(Ours)	374	861	16.7	16.5	41.3	33.2	67.1	15.0	35.2	47.0

表 2　低リソース下における提案手法とベースラインの比較

		Uniq-l	Uniq-S	B-4	M	R	C	CLIP-S	R@1	R@5	R@10
MS COCO	Att2in	543	3529	**26.5**	**23.2**	**51.1**	**83.0**	71.4	6.6	20.0	30.4
	+MixAug(Ours)	**889**	**4055**	24.6	22.7	50.0	78.1	**71.5**	**7.6**	**21.9**	**31.9**
Flickr30k	Att2in	321	857	**14.0**	**15.2**	**39.4**	**26.0**	**64.3**	10.1	24.1	**34.8**
	+MixAug(Ours)	**364**	**895**	13.8	15.0	39.1	24.4	64.2	**10.4**	**25.4**	34.2

データ拡張であり、画像や音声領域の分類タスクにおいて有効であることが確認されている。結合方法は以下のように計算される。

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda) x_j \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

x_i, x_j は、データセットにおけるベクトル（画像）であり、y_i, y_j は、正解ラベル（one-hot ベクトル）である。また、$\lambda \in [0, 1]$。

本研究中では $\lambda = 0.5$ とし、画像中における主要な物体の領域と、テキスト拡張時に得られた下位語の画像[3]に対して、この手法を適用することで画像拡張を行う。なお、正解ラベルである $\hat{y}$ の処理は行なっていない。

4　実験

今回は、ベースラインのモデルとして、Rennie らのモデル [5] を利用した。このモデルは、エンコーダに CNN、デコーダに LSTM [14] を利用しており、それらは注意機構 [15] により結合されている。

4.1　データセット及び前処理

データセット　提案手法は、MS COCO[16] データセットの Karpathy らの分割 [17]（MS COCO）における訓練セットに対して適用した。MS COCO は、113,287 画像からなる訓練セット、各 5,000 画像からなる検証セット、テストセットから構成されている。実験では、元データセットが高リソースである場合と、低リソースである場合に対して、提案手法を適用した。高リソース時では元データセットの 100% を利用し、低リソース時では、元データセットの 10%を利用した。

前処理の詳細　エンコーダには、学習済みの Resnet101 [18] を利用した。先行研究に従い、画像は、クロッピングやリサイズ処理を行わずに利用した。キャプションは、Stanza [19] をトークナイザとして利用し、データセットを通して出現頻度が 5 回未満の単語を $\langle UNK \rangle$ に置き換えて学習に利用した。

4.2　評価データ及び指標

評価の際は、MS COCO テストセットを利用した。また、モデルの頑健性を確認するために他データセットのテストセットでも評価を行った。他データセットは、キャプションの形式が MS COCO と似ている、Flickr30k のテストセット（1,000 画像）を利用した。

評価指標は先行研究に従い、標準的な評価指標として、BLEU-4 (B-4) [20]、METEOR (M) [21]、ROUGE-L (R) [22] 及び、CIDEr (C) [23] を用いた。また、識別性の評価指標として、Uniq-l・Uniq-S [4]、CLIP-Score (CLIP-S) [6] 及び、R@K を用いた。Uniq-l・Uniq-S [4] は、生成されたキャプションのうちの異なり語・文の数を表す。R@K は、生成されたキャプションを学習済みの画像テキスト検索モデルに入力した際に、他の画像の中から対応する画像が上位 K 位以内に得られた割合を表す指標である。学習済みの画像テキスト検索モデルについては、CLIP [24] を利用した。

4.3　実験結果

高リソース下での実験　学習時間の観点から、訓練セットを**約 2 倍**に拡張を行った。各テストセット

3)　下位語の画像は Web スクレイピングにより複数枚取得。

表 3　生成されたキャプションの例

			CLIP-S	生成されたキャプション
(a)		ベースライン	65.0	A green **plant** with a green plant and a tree.
		提案手法	**84.6**	A green **dwarf banana** is hanging from a tree.
		正解キャプション	92.9	A strange plant hanging off a banana tree.
			81.1	A bunch of bananas are hanging from the banana tree.
(b)		ベースライン	**80.2**	A **dog** is standing in front of a window.
		提案手法	78.5	A **black and white springer** is outside of a window.
		正解キャプション	80.0	A dog that is standing near an open window.
			89.8	A dog peers out of the bottom of a window in a multi-story building.
(c)		ベースライン	**71.5**	A man holding a surfboard walking in the ocean.
		提案手法	41.7	A man and a woman walking in the air.
		正解キャプション	78.9	A man in black wet suit holding a surfboard under his arm.
			73.2	A person standing on rocks holding a surfboard.

におけるベースライン及び提案手法の結果を表 1 に示す。

ベースラインと比べて、Uniq-l・Uniq-S が増加していることが確認できる。しかし、他データセットである Flickr30k においては、増加が確認できない。また、識別性を評価する CLIP-S や R@K が向上していないことから、増加した語彙・文が識別性に貢献できていないことがわかる。

低リソース下での実験　学習時間の観点から、訓練セットを**約 5 倍**に拡張を行った。各テストセットにおけるベースライン及び提案手法の結果を表 2 に示す。

ベースラインと比べて、Uniq-l・Uniq-S ともに大きく増加している。CLIP-S 及び R@K が向上していることから、増加した語彙・文が識別性に貢献していることが示唆される。また、ベースラインと比較して標準的な評価指標を大きく損なっていないことから、ベースラインの出力するキャプションの形式と似た、より識別性の高いキャプションが生成できていることが示唆される。他データセットである Flickr30k においても、Uniq-l・Uniq-S 及びいくつかの R@K に向上が確認できる。

4.4　生成されたキャプションの例

低リソース下の実験における、生成されたキャプションの例を表 3 に示す。

(a) と (b) の結果から、ベースラインと比較して、提案手法によって生成されたキャプションが、より特徴的な単語を用いて記述されていることが確認できる。これらから、提案手法が、生成に使用する単語をより特徴的な単語になるように促していることが示唆される。この結果、(a) については CLIP-S が向上しており、特徴的な単語で記述することが識別性を向上させていることが確認できる。しかし、(b) については、CLIP-S が低下しており、必ずしも特徴的な単語が CLIP-S を向上させるとは言えないことがわかる。

(c) では、ベースラインでは妥当なキャプションの生成が可能であるのに対し、提案手法では誤ったキャプションが生成されている。これは提案手法が、予期しない、誤った生成に寄与してしまっていることが示唆される。このような例は他の生成結果にも確認され、この誤った生成が CLIP-S や R@K を大きく向上させることができなかった原因と考えられる。

5　終わりに

本研究では、識別性が単語の階層関係に依存すると仮定し、既存のデータセットに対して、単語の階層関係に基づくデータ拡張を施すことで、低識別性の問題に対処する手法の提案を行った。実験の結果、低リソース下の実験において、出現語彙・文の増加とともに識別性を向上させ、有効であることが確認できた。しかし、高リソース下の実験では、出現語彙・文は増加したものの、識別性の評価指標である CLIP-S や R@K においてベースラインを上回ることができなかった。今後の課題として、4.4 で述べた、提案手法が誤った生成に寄与する原因について調べていくとともに、改良に努めていきたいと考える。

参考文献

[1] Burak Makav and Volkan Kılıç. A new image captioning approach for visually impaired people. In **IEEE**, 2019.

[2] Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distincive image captioning via clip guided group optimization. In **ECCVW**, 2022.

[3] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, Chan, and Antoni B. Compare and reweight: Distinctive image captioning using similar images sets. In **ECCV**, 2020.

[4] Ukyo Honda, Taro Watanabe, and Yuji Matsumoto. Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning. In **WACV**, 2023.

[5] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In **CVPR**, 2017.

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In **ACL**, 2021.

[7] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In **ICLR**, 2018.

[8] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In **ICCV**, 2019.

[9] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. Data augmentation to improve robustness of image captioning solutions. In **CVPR**, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **ACL**, 2019.

[11] Viktar Atliha and Dmitrij Šešok. Text augmentation using bert for image captioning. **Applied Sciences**, 10(17):5978, 2020.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In **NeurIPS**, 2015.

[13] George A. Miller. Wordnet: A lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

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

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NeurIPS**, 2017.

[16] 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, **ECCV**, 2014.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In **CVPR**, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, 2016.

[19] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In **ACL**, 2020.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL**, 2002.

[21] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **ACL**, 2005.

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **ACL**, 2004.

[23] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **CVPR**, 2015.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **ICML**, 2021.