キャプション生成時低品質データ事前検知の試み

大畑 和也 † 長澤 駿太 ‡ 北田 俊輔 ‡ 彌冨 仁 †‡
† 法政大学 理工学部 応用情報工学科
‡ 法政大学 大学院 理工学研究科 応用情報工学専攻
{kazuya.ohata.2b@stu., iyatomi@}hosei.ac.jp
{shunta.nagasawa.2u, shunsuke.kitada.8y}@stu.hosei.ac.jp

概要

視覚障がい者支援は画像キャプション生成タスクの効果的な応用の１つである。しかしながら、従来研究の多くは一般物体認識用途の対象物が明瞭に写っている高画質の画像データを学習に用いており、実用的な研究は発展途上である。近年、視覚障がい者支援を目的とし、障がい者自身の撮影による VizWiz Image Caption Dataset が発表されたが、ブレや見切れなどにより適切なキャプション生成が困難な画像が少なくない。被支援者は利用時に生成キャプションの妥当性が判断できないため、不適切なキャプションは大変不都合となる。我々は、適切なキャプション生成が困難な画像に対し利用者に再度撮影を促すことが実用上重要と考え、事前検知の可能性について検証した。本稿では、最先端のキャプション生成手法である AoANet が適切なキャプションを生成できない画像の事前検出を試みた。最先端手法を含む様々な深層学習モデルによる検証を行ったが、現時点ではこうした画像の事前検出は画像のみでは容易でないことが分かった。

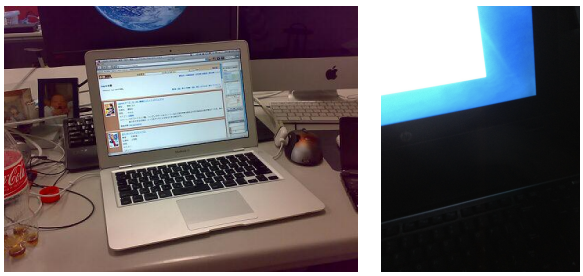
1 はじめに

画像キャプション生成は、モデルに画像内の物体の関係性を学習させ、人間が理解可能な自然言語の形で出力するタスクであり、テキストベースの画像検索や、視覚障がいを持つ人への支援など幅広い応用が考えられる。近年では深層学習を元にした画像キャプション生成の研究が盛んに取り組まれており、Vinyals ら [1] が提唱した convolutional neural network (CNN) と long short-term memory (LSTM) [2] を使用したモデルが優れた性能を示し、類似する手法が多く用いられてきた。最近では注意機構を利用したモデルが成功を収めており、Huang ら [3] は注意機構の着目箇所が適切かを測ることでより有用な情報を得る Attention on attention network (AoANet) モデルを提案した。AoANet は、一般物体検出や認識タスクの他、画像キャプション生成研究によく利用されている MSCOCO データセット [4] を元にしたキャプション生成において他の多くのモデルを上回る最先端の成果を実現している。MSCOCO データセットは、物体や人間の行動が画像中に大きく映っている上、光源等の撮影環境も適切な画像で構成されており、一般物体検出・認識タスクの他、画像キャプション生成の研究によく利用されている。

画像キャプション生成の視覚障がい者支援の応用として、深層学習を元にした歩道上での安全ナビゲーションシステムの構築が提案されている [5]。このシステムでは解析対象の画像が、影などの光源の影響や回転している場合に性能が著しく低下することが報告されており、それら撮影状況を考慮したデータセットの利用などが必要である。

視覚障がい者支援の実応用を見据えた深層学習モデルの学習のために、VizWiz データセット [6] が公開されている。これは視覚障がい者が撮影した画像をもとにキャプションがアノテーションされており、実際に近い環境の画像群が得られるが、画像中のブレやボケといった必ずしも画像の品質が良いとは言えないデータが含まれている。こうした品質の優れない画像から適切なキャプションを生成することは一般的に難しく、また、視覚障がい者ナビゲーションシステムのような応用例では、誤った推論が事故に繋がる恐れがある。

本研究では視覚障がい者の支援を目的に、解析対象画像が、キャプション生成に適しているかを画像識別により事前判断できるかについて検証する。不適切画像と事前に判断できれば、システム利用者に画像の再取得を促せるため実用上大変有意義である。本研究では最先端の性能を記録している画像キャプションモデル AoANet [3] を利用し、まず一般



(a) MSCOCO (b) VizWiz

図 1: MSCOCO と VizWiz の違い

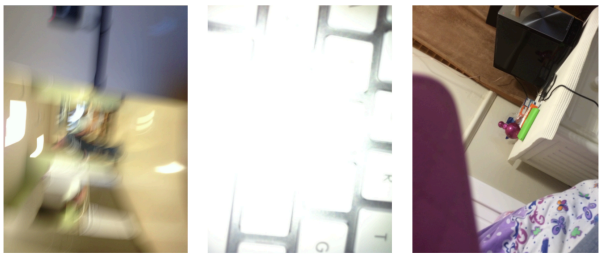


図 2: VizWiz データセットの低品質画像の例

的な画像キャプションモデルの学習に用いられている MSCOCO データセットと、実世界の障がい者支援を目的とした VizWiz データセットの差について、比較および議論する。その後、画像のみからの高精度なキャプション生成の可否の予測について、近年高いスコアを記録している複数の先端的な画像認識モデルを用いて検証する。

2 MSCOCO と VizWiz データセット

本検証ではまず最先端のキャプション手法の 1 つである AoANet [3] を用い、画像キャプションモデルの学習に広く用いられている MSCOCO [4] データセットと実世界の障がい者支援を目的とした VizWiz [6] データセットについて、キャプション生成の難易度の差の観点から比較した。

2.1 データセットの比較

図 1 に MSCOCO と VizWiz の画像の違いを示す。

MSCOCO MSCOCO [4] は Flickr 上の画像から複数の物体がはっきりと写った画像から構成されるデータセットである。MSCOCO には一枚の画像に 1 から 5 つ程度のキャプションが付随し、MSCOCO2014 では学習用に 82,783 枚、検証用に 40,504 枚の画像が含まれている。

VizWiz VizWiz [6] は 2 に示すように画像に欠陥がないものの他にはボケ、見切れ、明暗の差が激しい、対象が不明瞭、回転のような特徴を持つ画像が含まれており、上記の MSCOCO と比べタスクの難易度が高いとされるデータセットである。VizWiz においても MSCOCO と同様に 1 枚の画像に複数キャプションが付随しており、学習用に 23,431 枚、検証用に 7,750 枚の画像が含まれている。

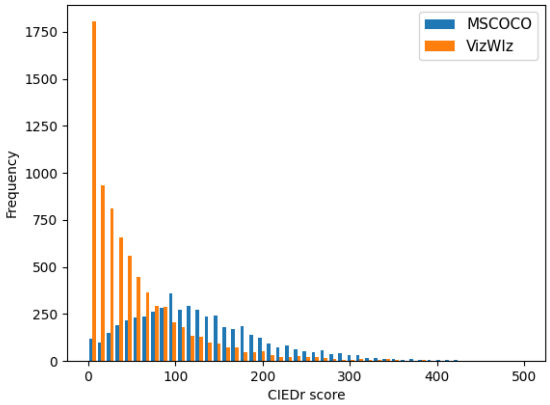


図 3: MSCOCO [4] と VizWiz [6] 検証データの CIDEr スコア分布

2.2 キャプション生成性能の比較

AoANet は、注意機構により抽出した画像中の重要な領域の情報を元に、良好なキャプションを生成可能な技術である。図 3 でこのモデルの推論により得られた文章を、画像キャプション生成の評価指標である CIDEr スコア [7] を用いて評価し、VizWiz および MSCOCO データセットによる分布の違いを比較した。

VizWiz (平均 CIDEr スコア 53.6) ではスコアの低いキャプションが生成される割合が MSCOCO (平均 CIDEr スコア 127.2) と比較して高かった。図 4 に低スコアとなる画像とキャプションの例を示す。カレンダーを PC のモニターと推論するといった見当違いの結果を予測している、0 から 10 にかけてのスコアを記録するものの割合が最も多くなった。

3 キャプション不適画像の事前検出

本研究は視覚障がい者支援のため、VizWiz データセットで学習した AoANet が生成するキャプションの妥当性を画像のみから予測可能かを検証する。本枠組みの全体像を図 5 に示す。具体的には、近年優

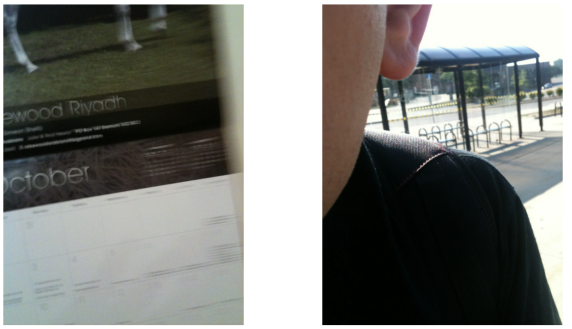


図 4: 低スコアとなる画像と正解キャプション、推論キャプションの例

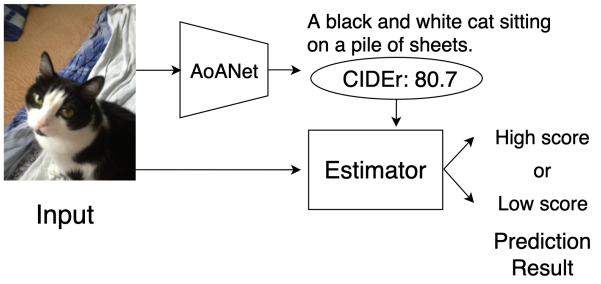


図 5: 提案するキャプション生成時低品質データ事前検出システムの全体像

れた画像認識能が報告されている複数の深層学習モデルを利用し、AoANet が生成するキャプションの CIDEr スコアが低い画像つまりキャプション不適画像を検出することを目的とする。

不適画像の推定に用いた画像認識モデルは以下の 6 種類及び、比較のためにランダム推定を用いた。

1. ImageNet [8] 事前学習済みの ResNeXt50 [9]
2. Instagram 事前学習済み ResNeXt101 [10]
3. ImageNet21k 事前学習済み ResNet50
4. ImageNet21k 事前学習済み ResNet50 (高解像度版)
5. ImageNet21k 事前学習済み VisionTransformer (ViT) base [11]
6. ImageNet21k 事前学習済み VisionTransformer (ViT) large [11]
7. ランダムに検出対象ラベルとする推定

4 実験

4.1 実験詳細

評価用データセット キャプション不適画像を判断する深層学習モデルを学習させるために、VizWiz の検証用データセット (7,750 枚) に対して AoANet が推定したキャプションの CIDEr スコアに基づき、高スコアクラス (high score 群) と低スコアクラス (low score 群) のラベルを生成した。これらの生成結果に対して、学習用 : 検証用 = 8 : 2 に分割した。

前処理とデータ拡張 前処理として画像を事前学習済みモデルの入力サイズと一致するようにリサイズした。その後 ImageNet の RGB の平均値と標準偏差を用いて画像の正規化を施した。データ拡張としてランダムに画像を左右反転させる水平反転、および画像を $-30°$ ~ $30°$ の間で回転させた。

学習と評価 学習時には、誤差関数には交差エントロピー誤差を利用し、学習回数は 100 エポックとした。評価指標には、precision、recall、F1 を用いた。

4.2 モデル比較による実験

本実験では、図 3 のスコア分布より、CIDEr スコアが下位約 35%を占める 20 以下の画像を対象とし、低品質クラスとしてそれらの検出を試みた。表 1 に低スコアクラス画像の検出能結果を示す。今回の実験では ResNeXt50 が一番高い結果を示したが、最先端の ViT モデルを含めいずれのモデルも適切な検出が行えていない結果となった。なお、モデルの特性の違いを明らかにするために結果は precision, recall のバランス調整のための閾値調整などは行わず、モデルの出力そのままに基づく結果を記載した。また ViT large 以外のモデルにおいて、学習データに対しては推定精度が十分に高いが、検証データの場合に大幅に下がってしまう過学習が見られた。

5 考察

本実験においては、先端的な画像識別手法を使用してもキャプション生成が難しい低スコア画像の検出能は低く、単純なランダム推定と同様かそれ以下の結果しか得られなかった。入力される画像が光源の影響を強く受けていたり、ボケていたりする場合などの根本的な悪条件画像の場合、適切なキャプション生成は望めず、CIDEr スコアは大幅に低下する傾向にある。本実験ではこうした画像の検出が期待されたが、実験結果に掲載していない、より浅いネットワークも含めて成果は得られなかった。その

表 1: 異なるモデルを用いた入力画像に対する品質判断結果の比較

使用モデル	解像度	Precision	Recall	F1
ResNeXt50	224 × 224	0.449	0.389	0.417
ResNeXt101	224 × 224	0.393	0.379	0.386
ResNet50	224 × 224	0.432	0.395	0.397
ResNet50	512 × 512	0.422	0.384	0.402
ViT base	384 × 384	0.447	0.348	0.388
ViT large	384 × 384	0.650	0.024	0.047
Random†	—	0.354	0.500	0.414

† 期待値

理由として光源やボケの影響で本来認識すべき物体を別のものと認識してしまったことが原因として挙げられる。また、物体が見切れている画像では全体が写っていないために正しく認識できなかったと考えられる。

6 結論

本稿では VizWiz Image Caption Dataset を用い推論時に低スコアとなってしまう入力画像事前検出の提案と調査を行なった。様々な先端モデルの適用やデータ拡張手法を適用したが、本実験の範囲においては、効果的な結果は得られなかった。学習時に多くのモデルにおいて過学習が確認されたため、より多くの学習画像の導入を検討するとともに今後も有効なデータ拡張の手法を探すなど、検出精度を高くする研究を続けていきたい。

参考文献

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

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

[3] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proc. of the IEEE International Conference on Computer Vision*, pp. 4634–4643, 2019.

[4] 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. of the European conference on computer vision*, pp. 740–755. Springer, 2014.

[5] Faruk Ahmed, Md Sultan Mahmud, Rakib Al-Fahad, Shahinur Alam, and Mohammed Yeasin. Image captioning for ambient awareness on a sidewalk. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pp. 85–91. IEEE, 2018.

[6] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Proc. of the European conference on computer vision*, pp. 417–434. Springer, 2020.

[7] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[9] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[10] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR preprint arXiv:2010.11929*, 2020.