7T-05

ViT エンコーダを活用した画像キャプション生成モデルの構築と評価

岡本翔汰 †　　黒田 久泰 ‡
愛媛大学工学部工学科 †　　愛媛大学大学院理工学研究科 ‡

1. はじめに

現代社会において画像は，ウェブ，SNS，広告，教育，医療などさまざまな分野で広く使用され，情報の共有や理解に欠かせないツールとなっている．自然な言葉による説明やキャプションは，画像に対する理解を深めるために重要である．

2014 年に，Kiros らが初めてニューラルネットワークを画像キャプション生成に利用した [1]．画像キャプション生成は，コンピュータビジョンと自然言語処理（NLP）の融合として位置付けられコンピュータが画像を認識し，その内容を自然な言葉で説明する技術である．本研究では，ViT エンコーダと Transformer デコーダから成る画像キャプション生成モデルを構築し，評価する．

2. Transformer

Transformer は，発表されて以来，自然言語処理タスクにおいて基盤となるネットワークとして幅広く活用されている．アーキテクチャに使われる重要なメカニズムの 1 つにセルフアテンションがある．系列データ内の各要素が他の要素とどの程度関連しているかを計算するための方法である．異なる重みを持った複数のヘッドでセルフアテンションを並列化して行い，最終的な出力を得る．系列データにおいて長い文脈や依存関係を捉えるのに効果的な手法となっている．本研究の提案モデルはエンコーダ，デコーダともに Transformer アーキテクチャで構成されている．

3. ViT

Vision Transformer(ViT) は，Dosovitskiy らが開発した Transformer アーキテクチャを画像認識に有効化した画像処理モデルである [2]．入力画像は，固定サイズのパッチに分割され，これらのパッチが Transformer の入力として使用される．分割されたパッチには位置情報がないため，Transformer への入力前に位置エンコーディングが導入される．Transformer ではマルチヘッドアテンション層や FNN 層を介して画像の重要な情報を回収する．情報は全結合層に送られ，クラス分類が行われる．ViT は，CIFAR-100 や ImageNet といった画像分類データセットでのテストにおいて CNN と同等かそれ以上の結果であると報告されている [3]．本研究では，全結合層を削除し，情報の抽出までの処理を行う．

4. 実験

4.1. モデルの概要

提案モデルでは画像の特徴抽出を行うエンコーダ部分に ImageNet で事前学習された学習済み ViT を設定し，クラス分類を想定した全結合層を取り除いている．既存モデルでは学習済み CNN として Resnet152 を用いている．提案モデルと同様に，全結合層を取り除いている．デコーダ部分は Transformer を用いた文生成デコーダとなっており，両者とも同じ構成となっている．

4.2. 実験の概要

本実験では，MSCOCO データセットから 40,505 枚の画像を用いて学習と検証を行った．1 枚の画像につき約 5 つのキャプションが付与されている．このうち 70 %を学習データ，30 %を検証データとして設定した．学習時は画像サイズは 224×224 にリサイズし，モデルの汎化性能を向上させるためランダムな水平フリップと画像の標準化を施した．1 エポックごとに検証を行い，交差エントロピー誤差を計算した．学習は 500 回とし，損失がもっとも小さいモデルを保存した．その後，MSCOCO データセットから新たに 1,000 枚の画像を用いて保存したモデルの性能を評価した．画像に対して，生成したキャプションと正解のキャプションを評価関数に入力した．それぞれの評価指標に関して取得した 1,000 個のスコアで平均値を計算し，最終的な結果を得た．

4.3. 評価指標

本研究の実験では，モデルを比較するために以下の 4 つの評価指標を用いた．値の範囲はどれも 0〜1 である (高いほど良い)．

(1) BLEU：n-gram を用いて，短いフレーズや長いフレーズの一致度を総合的に判断し，評価を行う．

Construction and evaluation of image captioning model using ViT encoder
† Okamoto Shota, Faculty of Engineering, Applied Information Engineering, Ehime University
‡ Kuroda Hisayasu, Graduate School of Science and Engineering, Ehime University

(2) CIDEr：n-gram 形式を用いた TF-IDF による平均コサイン類似度を表す．
(3) ROUGE：数種類ある中で ROUGE-L という手法を用いる．2 つのキャプションの最長共通部分を用いて適合率を計算する．
(4) SPICE：文章を意味的に解析したあとに関連のある言葉でタプルを複数生成し，タプル間の重複を見て評価を行う．

4.4. 実験の結果

提案モデルにおいて損失がもっとも小さくなったのは 213 エポックのとき，既存モデルでは 166 エポックのときであった．これらの学習データをロードして画像キャプション生成を行い，出力されたキャプションと正解キャプションをもとに評価を行った．

それぞれのモデルの評価は表 1 のようになった．すべての評価指標において ViT は CNN のスコアを上回った．もっとも差が大きかったのは CIDEr で，その差は約 0.04 ポイントであった．評価スコアで既存モデルを超えていることから提案モデルがより優れていることが示唆される．自然なキャプションが生成された例を図 1 に示す．一部では，同じ文字列を繰り返している不自然なキャプションが生成された結果もみられた．例を図 2 に示す．

表 1　評価スコアの比較

	ViT	CNN	2 つの差
BLEU	0.6374	0.6172	0.0202
CIDEr	0.7088	0.6641	0.0447
ROUGE	0.4217	0.4109	0.0108
SPICE	0.1647	0.1529	0.0118

生成文：A dog is catching a frisbee in a park
正解文：A white dog is holding a frisbee in its mouth waiting to play

BLEU Score：0.6544
CIDEr Score：0.7645
ROUGE Score：0.4317
SPICE Score：0.1670

図 1　出力キャプションの例 1

生成文：A cup of coffee and a cup of coffee
正解文：A dessert and a cup of coffee sit next to a book and a purse

BLEU Score：0.6492
CIDEr Score：0.7222
ROUGE Score：0.4281
SPICE Score: 0.1828

図 2　出力キャプションの例 2

5. まとめ

実験結果より，すべての評価指標において ViT は CNN のスコアを上回ったため，ViT を用いた画像キャプション生成は CNN モデルよりも精度の高い出力結果を得るために有効であると考える．しかし，不自然なキャプション生成結果もみられた．こういったキャプションにも，自然なキャプションの例と近いスコアが与えられてしまっている．不自然なキャプション生成を減らすこと，人間による評価との相関が高い評価指数を検討することについては今後の課題である．

参考文献

[1] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel, “*Multimodal Neural Language Models*”, Proceedings of the 31st International Conference on Machine Learning, Vol.32, No.2, pp.595-603, 2014.
[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*”, arXiv preprint arXiv:2010.11929, 2021.
[3] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao, “*A Survey on Visual Transformer*”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.45, No.1, pp.87-110, 2023.