敵対的生成ネットワークを用いて抽出した画像の構造情報に基づく画像キャプション生成

田辺雄大 [1]　綱川隆司 [1]
[1] 静岡大学大学院 総合科学技術研究科 情報学専攻
tanabe.yudai.17@shizuoka.ac.jp　tuna@inf.shizuoka.ac.jp

概要

画像キャプション生成タスクは画像認識と自然言語処理が含まれる複雑なタスクである。本論文では画像認識の部分に生成モデルである GAN を応用する。対象の画像を GAN を用いて再構成し、再構成した画像から画像の内部表現を得る。これによりランダムで特徴量を得た場合よりも良い性能を出した。また、画像を再構成することにより、再構成する前のデータセットよりもキャプションモデルを変化させた場合の影響が大きくなった。これにより GAN での再構成によって空間的に画像を捉えるようになったことの示唆が得られた。

1　はじめに

入力された画像の説明文を生成する画像キャプション生成は画像認識とその画像に対する自然言語処理を扱う人工知能の研究分野として注目されている。画像処理においてはオブジェクトの検出と認識が必要であり、また、どのような状況なのか、そのオブジェクトの特性はどのようなものか、それらの相互作用は何かも理解する必要がある。ある言語の文として正しいテキストを生成するためにはその言語の構文的・意味的な知識が必要である。

機械が画像を理解するためには画像の特徴をどのように取得するかに大きく依存する。画像の特徴抽出には従来の機械学習ベースの手法と、深層学習ベースの手法に分けられる。

従来の機械学習ベースの手法では LBP[1]、SIFT[2]，HOG[3] などの特徴や、それらの特徴の組み合わせが広く使われている。入力データから特徴を抽出し、サポートベクターマシーン (SVM)[4] などの分類器に入力することでオブジェクトを判別する。

一方、深層学習ベースの手法では大規模な画像や動画のデータセットから自動で特徴を学習するため、多様な入力に対応できる。例として、画像からの特徴抽出には CNN(Convolutional Neural Networks)[5] が、広く用いられる。CNN での処理に続いてキャプション生成には RNN(Reccurent Neural Networks) が一般的である。

近年頻繁に用いられる生成モデルである GAN(Generative Adversarial Nets) は画像などの生成タスクで大きな成功を収めている。GAN は生成器と、生成器から生成されたものが本物か偽物かを見分ける識別器によって構成されている。GAN を画像キャプション生成に応用した手法ではキャプション生成モデルを生成器とし、生成されたキャプションを評価する識別器を交互に評価するために敵対的学習を行った [12]。この手法により従来の強化学習ベースの画像キャプション生成モデルよりも良い性能となった。さらにその過程でよく学習された識別器は画像キャプションの評価指標としても利用することができる。GAN の学習方法や識別器を画像キャプション生成タスクに応用する研究は行われている。本研究では GAN を画像処理部分に応用することによるキャプション生成への影響を調査するため、GAN を画像処理部分に応用する手法を提案する。

2　few-shot part segmentation

GAN の潜在変数を変化させると生成画像も変化するという研究もあり、GAN の各層から抽出する内部表現は生成画像と密接に関係している。Tritrong ら [10] の研究ではその内部表現を利用することによってセマンティック情報を保持することができるのではないかと仮定された。Tritrong ら [10] はその内部表現を利用することによってセマンティック情報を保持することができると仮定し、セマンティックパートセグメンテーションによる手法を提案した。セマンティックパートセグメンテーションとは画像内のオブジェクトのセグメント化を行うセマンティックセグメンテーションとは異なり、オブジェクト内のセグメント化するタスクになっている。目

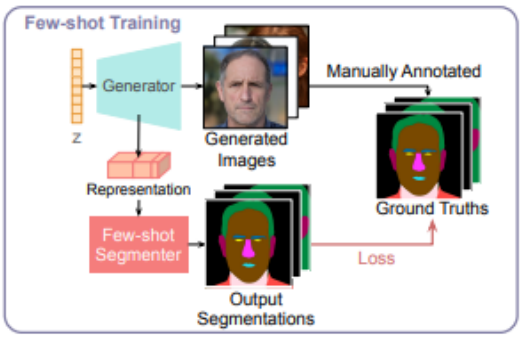

図 1　few-shot part segmentation モデル図 [10]

や鼻・顔のように 2 つの隣接するパーツ間の境界が曖昧なことがあるためセマンティックセグメンテーションよりも難しいタスクになっている。

モデルの概要図が図 1 である。

1. 対象となるドメインのデータセットで GAN を学習
2. 学習済み GAN を用いて画像の内部表現を生成
3. 2 で生成した画像に手動でアノテーション
4. 内部表現を入力として few-shot セグメンテーション
5. 4 において新にデータセットを作成し、新たにセグメンテーションマップを予測

GAN には StyleGAN2 が用いられ、few-shot ネットワークには評価実験の結果から CNN が採用されている。内部表現の抽出は 1 のように、生成器の全ての層から活性化マップを抽出し、それらを最大次元へアップサンプリングし連結することによって得る。この内部表現は生成された各画像にのみ有効であり、テスト画像には使用することができないが、任意のテスト画像を生成する潜在変数を最適化することで同様に内部表現を得ることができる。

これらの内部表現を利用し学習した結果、非常に少ない数のラベル付きデータでパートセグメンテーションを可能にし、10〜50 倍のラベルを必要とする完全教師有りの手法と同程度の性能を出した。本研究により、GAN が画像を生成する際に画像の構造情報を学習していることの示唆を与えた。筆者はこの技術に着目し、GAN の生成器が保持している画像の構造情報をキャプション生成モデルに入力することで、よりキャプションに画像の構造情報が反映された結果となることに期待した。

3　提案手法

提案手法のモデル図が図 2 である。提案手法は GAN を用いて画像から内部表現を抽出するタスクとキャプション生成モデルを用いて内部表現からキャプションを生成するタスクの大きく 2 つに分かれる。

3.1　画像投影器

データセットの画像を GAN で生成するための潜在変数を得る。StyleGAN2[8] では与えられた画像に対応する潜在変数を得る際に、拡張などを行うことなく元の潜在空間から潜在変数を得る。最適化の際に潜在変数にランダムノイズを加え、そのノイズの最適化も行う。

元画像と再構成画像の LPIPS[13] 距離を $D_{\mathrm{LPIPS}}[\boldsymbol{x}, g(\tilde{g}^{-1}(\boldsymbol{x}))]$ として計算し、画像の類似度とする。ここで x は対象の画像、$\tilde{g}^{-1}$ は近似投影演算を表す。

3.2　画像からの内部表現抽出

ベースライン: Rennie ら [11] の研究では 101 層の Resnet-101 で特徴量を抽出している。Attention モデルでは最後の畳み込み層で画像を符号化し、出力が $14 \times 14 \times 2048$ のサイズになるように最大プーリングを行う。後述するキャプション生成モデルに FC モデルを用いる場合は出力層を空間的に平均化することで 2048 次元の特徴量を得る。

提案手法: Tritrong ら [10] の研究では、生成器の各層からから $a_1, a_2, ..., a_n$ のように活性化マップを抽出する。さらにこれらを以下の式で結合する:

$$F = \mathbb{U}(a_1) \oplus_c \mathbb{U}(a_2) \oplus_c ... \mathbb{U}(a_n)$$

ここで、$\mathbb{U}()$ は画像サイズに空間的にアップサンプリングする関数であり、$\oplus_c$ はチャンネル次元に連結する処理である。

提案手法ではキャプションモデルの入力やベースラインの次元数に合わせるためアップサンプリングではなく、14×14 に次元数を整え各層を連結する。

提案手法では学習済み GAN の生成器を用いて画像から特徴量を抽出する。まず初めに対象となるドメインの画像群で GAN を学習し、そのドメインの画像を生成するモデルを作成する。次にキャプション付きデータセットの画像を画像投影器に入力し、学習済み GAN に入力するための潜在変数を取得するとともに、学習済み GAN の生成器を用いて画像の再構成を行う。そしてその潜在変数を同様の学習済み GAN の生成器に入力し、内部表現を得る。このようにして画像から対応する内部表現を抽出する。

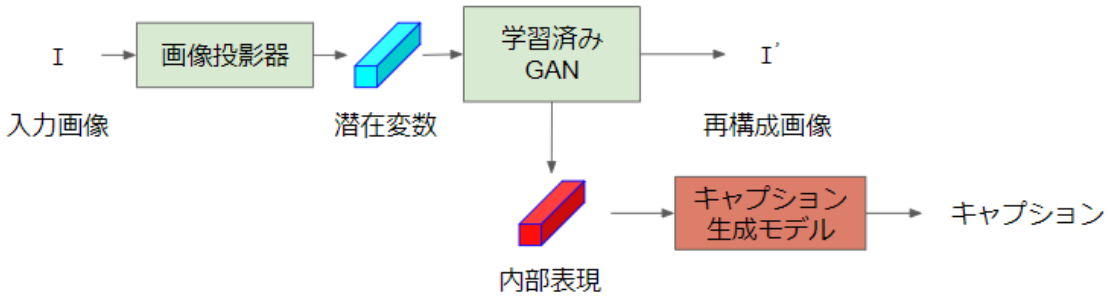

図 2　提案手法のモデル図

3.3　キャプション生成モデル

FC モデル: 入力画像を CNN などで特徴量を取得し、線形射影で埋め込む。単語はワンホットベクトルで表現され、同様の次元数の線形埋め込みで埋め込まれる。このモデルではまず単語が生成され、LSTM にフィードバックされる。モデルのパラメータを θ とすると教師データが与えられたときのクロスエントロピー誤差を損失関数 $L(\theta)$ として最小化する:

$$L(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^*|w_1^*, ...w_{t-1}^*)) \quad (1)$$

Attention モデル (Att2in2): Attention モデルは各タイムステップで画像の特定の領域に焦点を当てるために入力された空間的特徴を動的に再重み付けする。Xu ら [9] によるキャプションのための Attention モデルのアーキテクチャを修正し、LSTM のセルにのみ Attention によって得られた特徴量を入力する。式 (1) と同様の損失を最適化する。

4　実験

4.1　データセット

画像の内部表現を抽出するための GAN の学習には BIRDS450[7] という 450 種類の鳥が含まれるデータセットを使用した。BIRDS450 のうち、GAN の学習には学習用の 70626 枚を使用した。

本論文でのキャプションモデルの学習には全て CUB-200(Caltech-UCSD Birds-200-2011) を使用した。このデータセットは 200 のサブカテゴリからなる学習用 5994 枚、テスト用 5794 枚の合計 11788 枚となっている。各画像には 10 個のキャプションが付けられており、10 単語以上の文となっている。語彙は 5 個以下の単語を削除した。

提案手法では全てのデータセットの GAN による再構成を行う。CUB-200 を GAN で再構成した画像を用いてベースラインモデルの学習も行い、提案手法と比較する。

評価指標は BLEU、METEOR、ROUGE-L、CIDEr-D を用いる。

4.2　画像の内部表現

画像から特徴量を抽出する方法は 3.2 節の通りである。ベースラインである ResNet からは FC モデルでは 2048、Attention モデルでは 14 × 14 × 2048 次元である。提案手法からは FC モデルでは各層の次元数が (512,512,512,512,512,512,512,512,512,256,256,128,128) となっているので全て合わせて 5376 次元、Attention モデルでは 14 × 14 × 5376 次元となる。

また、提案手法と比較するためにランダムで特徴量を生成し比較実験を行う。正規分布から FC モデル、Attention モデルそれぞれ 5376 次元、14 × 14 × 5376 次元の特徴量を得る。

4.3　実験結果

4.3.1　画像の再構成

StyleGAN2 によって再構成された画像と元の画像の LPIPS 距離のヒストグラムが図 3 である。画像のリサイズを行っているため LPIPS の値が高くなってしまう画像がある。画像から潜在変数、潜在変数から画像へのマッピングが複雑なため再構成が失敗する例も存在する。

画像の再構成によって元の画像とは異なる画像となってしまう場合がある。そのような画像にも同様のキャプションを付けた場合、結果にどのような影響があるかの検証を行う。画像の特徴抽出はベースラインモデルで行う。

表 1 の 2 行目はテストデータの画像だけを再構成画像に変えた場合、3 行目は学習・テストデータ両方の画像を再構成画像に変えた場合の結果となって

表 1　モデルと特徴量による実験結果

特徴量	モデル	学習データ	テストデータ	Bleu4	METEOR	ROUGE-L	CIDEr-D
ResNet	FC	CUB-200	CUB-200	0.486	0.320	0.620	0.423
ResNet	FC	CUB-200	CUB-200 再構成	0.482	0.314	0.625	0.381
ResNet	FC	CUB-200 再構成	CUB-200 再構成	0.450	0.304	0.591	0.298
GAN	FC	CUB-200 再構成	CUB-200 再構成	0.437	0.302	0.594	0.231
random	FC	CUB-200	CUB-200	0.463	0.302	0.632	0.251
ResNet	Att2in2	CUB-200	CUB-200	0.500	0.317	0.622	0.432
ResNet	Att2in2	CUB-200 再構成	CUB-200 再構成	0.493	0.314	0.626	0.400
GAN	Att2in2	CUB-200 再構成	CUB-200 再構成	0.461	0.302	0.603	0.297
random	Att2in2	CUB-200	CUB-200	0.422	0.286	0.588	0.193

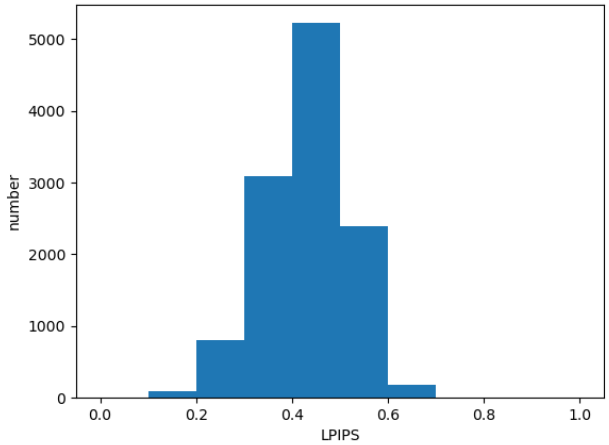

図 3　LPIPS 距離のヒストグラム

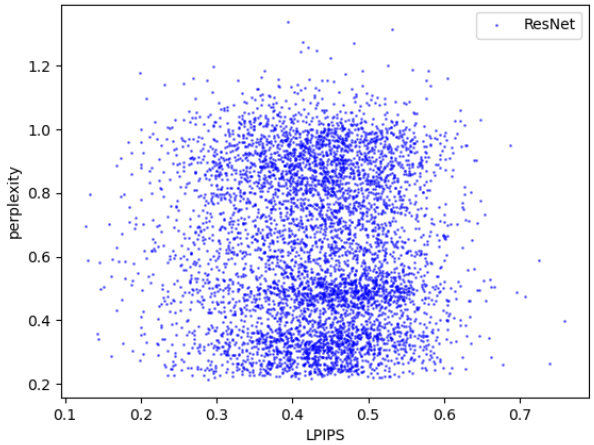

図 4　ベースラインモデルでの LPIPS と perplexity

いる。テストデータのみ変化させた場合は、どちらも変化させた場合よりも評価指標に大きな差がないことがわかる。図 4 はベースラインモデルで学習・テストデータどちらも再構成画像を用いた場合でのテストデータの LPIPS と perplexity の散布図である。これより、再構成画像の LPIPS と perplexity には相関がない、すなわち元画像と再構成画像の差と生成されたキャプションの精度には相関がない。

4.3.2　画像の内部表現抽出方法による実験

画像の内部表現抽出による比較を行う。結果が表 1 である。FC モデル、Attention モデルどちらも ResNet から得た特徴量の方が良い性能となっている。データセットに CUB-200 再構成を使ったケースでは Attention モデルにすることで性能が大幅に向上するが、CUB-200 では向上が見られない。その結果 CUB-200 で学習した場合と CUB-200 再構成のデータセットで学習した Attention モデルの精度が同程度となった。一方、ランダムの特徴量では性能が劣化した。提案手法はランダムで特徴量を得た場合よりも良い性能となっているため提案手法は画像の意味を持つ特徴量を抽出できていることがわかる。

5　結論

本論文では画像キャプション生成タスクの画像から特徴量を取得するモジュールに GAN を応用する手法を提案した。GAN が画像生成時に画像の構造も学習しているという性質を利用し、その画像キャプション生成への適用について調査した。その結果、本論文では GAN から得た特徴量がランダムで得た特徴量よりも優れていることが示された。さらに画像を再構成することで画像の特徴量を線形に取得する FC モデルに対して空間的に画像の特徴量を取得する Attention モデルを適用した場合の精度の向上幅が上昇し、Attention モデルでは精度が FC モデルでもとの特徴量を用いた場合と同程度となった。これにより画像を再構成することで、空間的に画像を捉えるようになったことの示唆が得られた。

参考文献

[1] Ojala, Timo, Matti Pietikäinen, and Topi Mäenpää. ”Gray scale and rotation invariant texture classification with local binary patterns.” European conference on computer vision. Springer, Berlin, Heidelberg, 2000.

[2] Lowe, David G. ”Distinctive image features from scale-invariant keypoints.” International journal of computer vision 60.2 (2004): 91-110.

[3] Dalal, Navneet, and Bill Triggs. ”Histograms of oriented gradients for human detection.” 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05). Vol. 1. Ieee, 2005.

[4] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. ”A training algorithm for optimal margin classifiers.” Proceedings of the fifth annual workshop on Computational learning theory. 1992.

[5] LeCun, Yann, et al. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[6] Hossain, MD Zakir, et al. ”A comprehensive survey of deep learning for image captioning.” ACM Computing Surveys (CsUR) 51.6 (2019): 1-36.

[7] https://www.kaggle.com/datasets/gpiosenka/100-bird-species

[8] Karras, Tero, et al. ”Analyzing and improving the image quality of stylegan.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[9] Xu, Kelvin, et al. ”Show, attend and tell: Neural image caption generation with visual attention.” International conference on machine learning. PMLR, 2015.

[10] Tritrong, Nontawat, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. ”Repurposing gans for one-shot semantic part segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[11] Rennie, Steven J., et al. ”Self-critical sequence training for image captioning.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Chen, Chen, et al. ”Improving image captioning with conditional generative adversarial nets.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

[13] Zhang, Richard, et al. ”The unreasonable effectiveness of deep features as a perceptual metric.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.