敵対的生成ネットワークによる文からの画像生成の改良手法の提案

第 1 グループ　　置名 一元

1. はじめに

創作とは人間ならではの高度な知的活動である．近年，人工知能（Artificial Inteligence： AI）の目覚ましい発展に伴い，情報工学における研究分野は AI による創造の分野にまで拡大しており，計算機による創作物の理解や自動生成への試みは工学的にも興味深く大きな意義を持つようになっている．また，言語や画像などの単一の情報だけでなく，多分野の情報を複合的に取り扱うマルチモーダルな自動生成に対する研究も盛んになされはじめている．Attentional Generative Adversarial Networks（AttnGAN）[1] は，創作とマルチメディアと双方の分野の特徴を持つことから，高い注目を集めている自動生成手法の一つである．

AttnGAN では，自然言語から成る説明文中の関連する単語に Attention を向けることにより，テキスト入力から高精細な画像の生成が可能となっている． 単語と画像の関係を Attention として事前学習するために，Deep Attentional Multimodal Similarity Model（DAMSM）が提案されている．DAMSM により，入力テキスト中の単語のニュアンスが反映された画像生成を可能としている．

しかし AttnGAN では，単語の意味，構文，および文法はまったく考慮されていないという問題点がある．自然言語処理の方法に基づいて単語の文化的背景と社会的文脈を考慮することは，DAMSM による Attention と AttnGAN 全体の性能向上につながる可能性が高いと考えられる．そこで本論文では，単語の分散表現によって得られた特徴ベクトルをテキストエンコーダの入力として使用する AttnGAN モデルを提案する．

また，DAMSM の訓練画像データの不足を補う目的で，事前学習だけでなく GAN との同期学習を可能にすることで，生成画像の情報も活用できるモデルについても提案する．

さらに，テキスト入力による画像生成の不安定性を低減するため，実画像と生成画像との距離として Fréchet Inception Distance（FID）[2] と Sliced Wasserstein Distance（SWD）[3] を測定し損失関数に組み込む．これにより，入力テキスト中の強い影響力を持つ単語が原因となり，意図せず実画像からかけ離れた画像が生成されてしまう現象の抑制を図る．

提案モデルによる画像生成の性能は従来研究と同じく Inception Score [4] に基づいて評価する．

2. 提案手法

初めに AttnGAN と DAMSM について説明する．

AttnGAN は Generative Adversarial Networks (GAN) [5] をベースとした生成モデルである．GAN の最大の特徴は，Generator と Discriminator の 2 つのニューラルネットワークに対立的な教師なし学習をさせるアルゴリズムである．学習がやや不安定で膨大な訓練データが必要という課題はあるものの，実データとの類似性と実用上としての新規性がともに高いデータを生成可能なため，期待と関心を集めている将来性の高いモデルである．AttnGAN では，段階的な GAN 構造と Attention 機構を導入することで，テキストからの高精細な画像生成を可能としている．図 1 に AttnGAN のモデル概要図を示す．

DAMSM は画像の各部分領域とテキスト中の各単語をそれぞれ分散表現化し，得られた特徴量ベクトルによるマッチングスコアから L_{DAMSM} を算出するモデルである．図 2 に DAMSM のモデル概要図を示す．

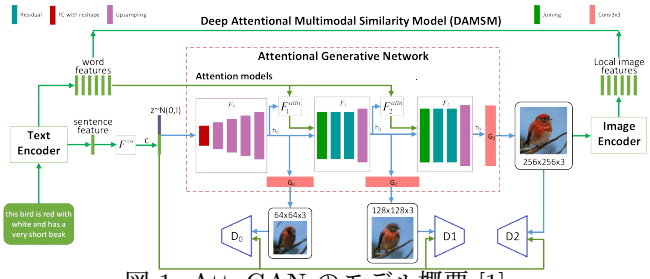

図 1: AttnGAN のモデル概要 [1]

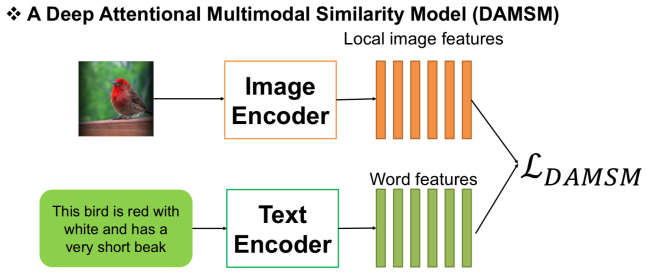

図 2: DAMSM の概要図 [1]

本研究では，DAMSM に単語の持つ意味を考慮させることで AttnGAN の性能向上につながると考え，自然言語処理における Word Embedding 手法により学習された単語の分散表現を DAMSM のテキストエンコーダの入力とするモデルを構築した．

さらに，DAMSM の訓練データ不足の補填を目的として，事前学習だけでなく GAN の学習時にも同期的に学習可能にすることで，Generator によって得られる生成画像の情報も活用できるモデルを提案する．

また，GAN によるテキストからの画像生成の不安定さの緩和のため，AttnGAN の学習時に実画像と生成画像との FID 及び SWD を Generator の損失関数に加える．これにより，入力テキストによっては実画像とかけ離れた画像が意図せず生成されてしまう現象の抑制を図る．

3. 数値実験

本研究では CUB 鳥画像データセット [6] を用いて実験する．このデータセットの画像の総数は 11788 枚であり，画像 1 枚につきキャプションが 10 文与えられている．この内 200 枚をテストデータとして選出し，残りを 8:2 の割合で訓練データと評価データに分割する．

本実験では AttnGAN の DAMSM のテキストエンコーダに対する入力は次の 5 パターンを比較する．

- 従来手法：単語ごとにランダムなベクトルの入力
- 提案手法：自然言語処理における Word Embedding 手法である Word2Vec, GloVe, fastText, ElMo によりそれぞれ学習された単語の分散表現の入力

訓練データを基に学習を進めた AttnGAN の各手法に対し，テストデータのキャプションを入力テキストとして画像をサンプリング生成させ，各手法による生成画像群の Inception Score を比較した．

表 1: 各手法による生成画像の Inception Score

手法	Inception Score	
テキストエンコーダの RNN	LSTM	GRU
同期的学習なし／ FID・SWD なし		
先行研究論文掲載値	4.36	-
従来手法	4.31	4.33
Word2Vec	4.26	4.24
GloVe	4.21	4.15
fastText	4.33	3.80
ElMo	4.45	4.93
同期的学習あり／ FID・SWD なし		
従来手法	4.48	4.28
Word2Vec	4.22	4.37
GloVe	4.00	4.05
fastText	3.84	4.04
ElMo	4.76	4.80
同期的学習あり／ FID・SWD あり		
従来手法	4.24	4.38
Word2Vec	4.12	3.96
GloVe	4.30	4.50
fastText	4.08	4.07
ElMo	4.58	4.78

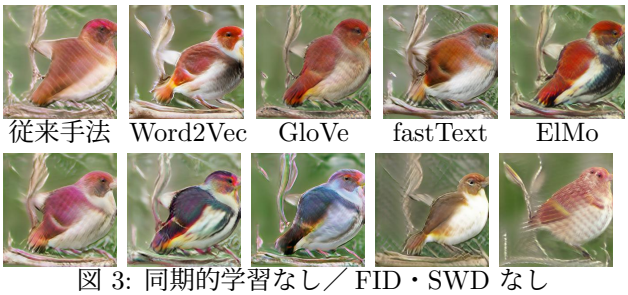

図 3: 同期的学習なし／ FID・SWD なし

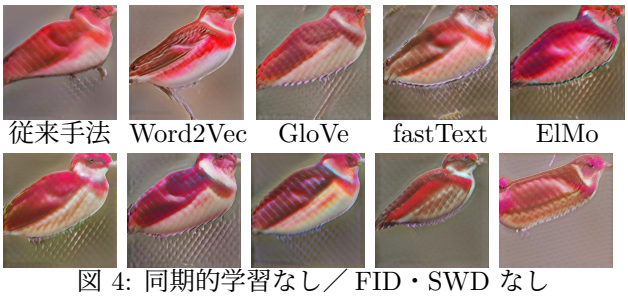

図 4: 同期的学習なし／ FID・SWD なし

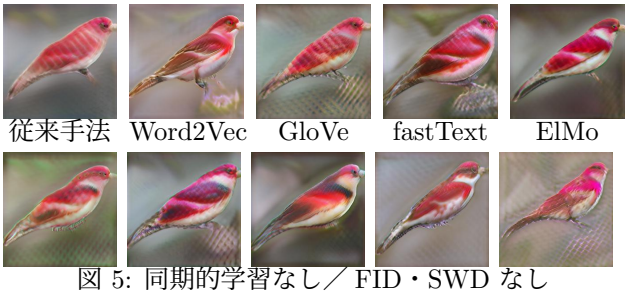

図 5: 同期的学習なし／ FID・SWD なし

4. まとめと今後の課題

表 1 に各手法により得られた Inception Score を示す. 図 3 から図 5 に, 各手法によりテストデータのキャプション “this is a red bird with a white belly and a brown wing.” を入力としてサンプリング生成された画像 (上段：LSTM, 下段：GRU) を示す.

生成画像の結果から, DAMSM と GAN との同期的学習を導入することにより, キャプション中の “red” という単語の特色がより濃く反映されていることが確認できる. これは, 従来手法では GAN しか学習しないため, 単語と画像の齟齬よりも Discriminator による識別が優先されるために生じる現象だと考えられる. 具体的には, 赤い鳥の画像が訓練データセット内に少なく, 従来手法では Discriminator によって「赤い鳥は鳥ではない」と誤識別されてしまうため, 赤い鳥は生成されにくい傾向があると考えられる.

今後の課題として, 提案モデルのパラメータ調整や, BERT の導入の検討などがあげられる.

参考文献

[1] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks, 2018.

[2] Martin Heusel and Hubert Ramsauer and Thomas Unterthiner and Bernhard Nessler and Günter Klambauer and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium, 2017.

[3] G. Peyré and Marco Cuturi. Computational optimal transport, 2019.

[4] Christian Szegedy and Vincent Vanhoucke and Sergey Ioffe and Jonathon Shlens and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[5] Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua. Generative adversarial nets, 2014.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset, 2011.