単語の分散表現を用いる DAMSM と GAN との同期的学習を導入した AttnGAN モデルの提案

1　はじめに

創作は人間独自の高次な知的活動である．近年の人工知能の目覚ましい発達に伴い，その領域にまで研究分野が及んでおり，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，言語分野や画像分野などの単一の情報を対象とした研究から，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の研究も盛んになされはじめている．AttnGAN[1] は創作とマルチメディア双方の分野の特徴を持つことから，注目を集めている手法の一つである．

AttnGAN は Deep Attentional Multimodal Similarity Model (DAMSM) により単語と画像の関係を事前学習することで，テキストからの単語レベルでの画像生成を可能としている．その一方で，単語自身の持つ自然言語的な言葉の意味合いや語法・文法などは考慮されていない．自然言語処理分野の手法を基に単語の持つ文化的背景や社会的文脈を考慮することで，DAMSM による Attention の精度向上や AttnGAN 全体の性能向上につながると考えられる．

そこで本研究では DAMSM に自然言語処理分野で大規模コーパスを基に学習して得られた単語分散表現を用いて，GAN との同期的学習を可能とするモデルを提案する．提案モデルによる画像生成の性能は Inception Score に基づいて評価する．

2　AttnGAN

Generative Adversarial Networks (GAN)[2] とは，Generator (生成器) と Discriminator (識別機) という 2 つのネットワークを対立的に学習させる生成モデルである．実データとの類似性と実用上の新規性が高いデータを生成可能という点で注目を集めるモデルである．

AttnGAN は，段階的な GAN 構造と DAMSM による Attention 機構を導入し，テキストからの画像生成を目的とした GAN モデルである．

AttnGAN は (1) 式による i 段階目の Discriminator での損失関数 L_{D_i} と (2) 式による Generator での損失関数 L_G で定義される．L_{DAMSM} は 3 章で後述する DAMSM モデルによる Attention に関する損失関数である．図 1 に AttnGAN のモデル概要図を示す．

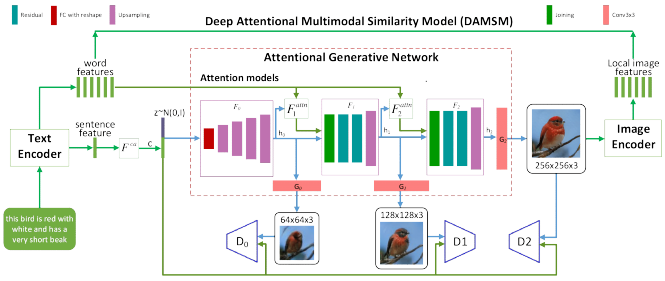

図 1: AttnGAN のモデル概要図 [1]

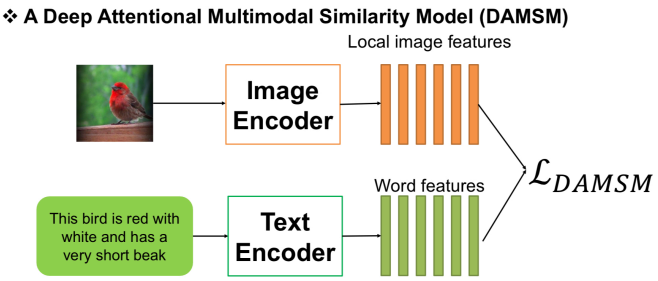

図 2: DAMSM のモデル概要図 [1]

$$
\begin{aligned}
L_{D_i} = &-\frac{1}{2}\mathbb{E}_{x_i\sim p_{\mathrm{data}}}[\log D_i(x_i)] \\
&-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log\left(1-D_i(\hat{x}_i)\right)] \\
&-\frac{1}{2}\mathbb{E}_{x_i\sim p_{\mathrm{data}}}[\log D_i(x_i, c)] \\
&-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log\left(1-D_i(\hat{x}_i, c)\right)]
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
L_G = \sum_{i=1}^{m}\langle &-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log D_i(\hat{x}_i)] \\
&-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log D_i(\hat{x}_i, c)]\rangle \\
&+\lambda L_{\mathrm{DAMSM}}
\end{aligned} \tag{2}
$$

3　DAMSM

DAMSM は，テキスト中の各単語と画像の各部分領域の特徴量ベクトルの類似度から，単語と画像間の共通意味空間上における関連性を学習することで，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．

3.1 テキストエンコーダ

テキスト中の各単語の分散表現を入力として，RNN の中間層における出力を各単語の意味空間ベクトル $\boldsymbol{e} \in \mathbb{R}^{D\times T}$ として用いる．このとき，D は単語ベクトルの次元数，T は文章中の単語数である．このとき，$\boldsymbol{e}_i \in \mathbb{R}^D$ が i 番目の単語の特徴量ベクトルとなる．また，最終層における出力を文章全体の特徴量を表す $\bar{\boldsymbol{e}} \in \mathbb{R}^D$ とする．

3.2 イメージエンコーダ

画像を入力として，Google による画像識別モデルである InceptionV3[3] 内の model_6e における出力を，17×17 に分割した各部分領域の特徴量ベクトル $\boldsymbol{f} \in \mathbb{R}^{768\times 289}$ として用いる．ここで，768 は各部分領域の特徴量ベクトルの次元数である．また，最終層における出力を画像全域の特徴量ベクトル $\bar{\boldsymbol{f}} \in \mathbb{R}^{2048}$ として用いる．

全結合層 $\boldsymbol{W}$ により，画像の特徴量ベクトル $\boldsymbol{f}$ とテキストの特徴量ベクトルとの次元数を等しくし，画像の各部分領域の特徴量ベクトル $\boldsymbol{v} = \boldsymbol{W} \cdot \boldsymbol{f}$ および画像全域の特徴量ベクトル $\bar{\boldsymbol{v}} = \bar{\boldsymbol{W}} \cdot \bar{\boldsymbol{f}}$ を得る．ここで，$\boldsymbol{v} \in \mathbb{R}^{D\times 289}$，$\bar{\boldsymbol{v}} \in \mathbb{R}^D$ である．このとき，$\boldsymbol{v}_i \in \mathbb{R}^D$ を i 番目の部分領域の特徴量ベクトルとし，$\bar{\boldsymbol{v}} \in \mathbb{R}^D$ を画像全域の特徴量ベクトルとする．

3.3 マッチングスコア

画像内の各部分領域の特徴量とテキスト内の各単語の特徴量の対応を，(3) 式のようにマッチング $\boldsymbol{s}$ として

$$\boldsymbol{s} = \boldsymbol{e}^\top \boldsymbol{v} \tag{3}$$

と定義する．ここで $\boldsymbol{s} \in \mathbb{R}^{T\times 289}$ であり，$\boldsymbol{s}_{i,\,j} \in \mathbb{R}^D$ は i 番目の単語の特徴量と j 番目の部分領域の特徴量のマッチングである．(4) 式のように，マッチング $\boldsymbol{s}$ を正規化することによって，$\bar{\boldsymbol{s}}_{i,\,j}$ を得る．

$$\bar{\boldsymbol{s}}_{i,\,j} = \frac{\exp(\boldsymbol{s}_{i,\,j})}{\sum_{k=0}^{T-1} \exp(\boldsymbol{s}_{k,\,j})} \tag{4}$$

次に，i 番目の単語と画像の各部分領域の関係性を動的に表す領域コンテクストベクトル $\boldsymbol{c}_i$ を (5) 式として定義する．ここで γ_1 は Attention のかかり方の配分を決定する温度パラメータの逆数である．

$$\boldsymbol{c}_i = \sum_{j=0}^{288} \frac{\gamma_1 \exp(\bar{s}_{i,\,j})}{\sum_{k=0}^{288} \gamma_1 \exp(\bar{s}_{k,\,j})} \boldsymbol{v}_j \tag{5}$$

$\boldsymbol{c}_i$ と $\boldsymbol{e}_i$ のコサイン類似度から，i 番目の単語と画像全域の関係を $R_w(\boldsymbol{c}_i, \boldsymbol{e}_i) = \frac{\boldsymbol{c}_i^\top \boldsymbol{e}_i}{\|\boldsymbol{c}_i\|\|\boldsymbol{e}_i\|}$ と表せる．これを用いて，画像 Q とキャプション D の単語レベルでのマッチングスコア $R_w(Q, D)$ を (6) 式と定義する．

$$R_w(Q, D) = \log \left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(\boldsymbol{c}_i, \boldsymbol{e}_i)) \right)^{\frac{1}{\gamma_2}} \tag{6}$$

ここで γ_2 は $R_w(\boldsymbol{c}_i, \boldsymbol{e}_i) = \frac{\boldsymbol{c}_i^\top \boldsymbol{e}_i}{\|\boldsymbol{c}_i\|\|\boldsymbol{e}_i\|}$ の重要度の拡張性を決定づける温度パラメータの逆数である．

また文章レベルでのマッチングスコア $R_s(Q, D)$ は (7) 式として定義する．

$$R_s(Q, D) = \frac{\bar{\boldsymbol{v}}^\top \bar{\boldsymbol{e}}}{\|\bar{\boldsymbol{v}}\|\|\bar{\boldsymbol{e}}\|} \tag{7}$$

3.4 L_{DAMSM}

バッチサイズ が M の時，$R_w(Q, D)$ と $R_s(Q, D)$ を用いて M 組の画像と単語のペア $\{(Q_i, D_i)\}_{i=1}^M$ において，キャプション D_i が画像 Q_i と正解マッチングとなる条件付き確率はそれぞれ (8) 式，(9) 式として求められる．ここで γ_3 はスムーズ化係数である．

$$P_w(D_i|Q_i) = \frac{\exp(\gamma_3 R_w(Q_i, D_i))}{\sum_{j=1}^M \gamma_3 R_w(Q_j, D_j)} \tag{8}$$

$$P_s(D_i|Q_i) = \frac{\exp(\gamma_3 R_s(Q_i, D_i))}{\sum_{j=1}^M \gamma_3 R_s(Q_j, D_j)} \tag{9}$$

これらの条件付き確率の負の対数尤度を取ることによって，(10) 式，(11) のように損失関数 L_1^w，L_1^s を得る．

$$L_1^w = -\sum_{i=1}^M \log P_w(D_i|Q_i) \tag{10}$$

$$L_1^s = -\sum_{i=1}^M \log P_s(D_i|Q_i) \tag{11}$$

また，ベイズ推定により，画像 Q_i がキャプション D_i と正解マッチングとなる条件付き確率は，それぞれ (12) 式，(13) 式として求められる．

$$P_w(Q_i|D_i) = \frac{\exp(\gamma_3 R_w(D_i, Q_i))}{\sum_{j=1}^M \gamma_3 R_w(D_i, Q_i)} \tag{12}$$

$$P_s(Q_i|D_i) = \frac{\exp(\gamma_3 R_s(D_i, Q_i))}{\sum_{j=1}^M \gamma_3 R_s(D_i, Q_i)} \tag{13}$$

これらの条件付き確率の負の対数尤度を取ることによって，(14) 式，(15) 式のように損失関数 L_2^w，L_2^s を得る．

$$L_2^w = -\sum_{i=1}^{M} \log P_w(Q_i|D_i) \tag{14}$$

$$L_2^s = -\sum_{i=1}^{M} \log P_s(Q_i|D_i) \tag{15}$$

これらの総和により，$L_{\mathrm{DAMSM}} = L_1^w + L_2^w + L_1^s + L_2^s$ として定義する．

4 自然言語処理における単語の分散表現獲得手法

4.1 Word2Vec

予測的な手法で実用上現在最も広く知られている Word2Vec[4] は，大量のテキストデータを解析し，各単語の意味をベクトル表現化する手法である．周辺単語から対象後を推測する Continuous Bag of Words (CBOW) と，ある語から周辺単語を対象として推測する skip-gram の 2 つの方法が存在する．

4.2 GloVe

Global Vectors for Word Representation (GloVe)[5] とは、コーパスを集約して得られたグローバルな単語の共起行列から統計的に学習し、単語の分散表現を獲得するという教師なし学習モデルである．

4.3 fastText

fastText[6] とは，Word2Vec の延長線上にあるモデルであり，単語の構成要素をサブワードとして分解していることが特徴である．単語をサブワードに分解し，活用による変化のない基幹部分のベクトルを扱うため，使用メモリを削減でき，学習速度も早く，未知語にも強いという長所があるとされる．

4.4 ElMo

Embeddings from Language Models (ElMo)[7] とは，双方向言語モデル (bidirectional language model) を大規模コーパスを用いて学習させ，文脈を考慮した単語の分散表現の獲得を目的とするモデルである．従来の単語の分散表現獲得手法では多義語であっても 1 単語ごとに 1 つの特徴量ベクトルしか得られないが，ElMo では文脈に応じて異なる単語特徴量ベクトルを獲得できるとされる．

5 生成画像への評価指標

5.1 Inception Score

Inception Score[8] は特に GAN による生成画像に対するスコアとして広く用いられてきた指標であり， N 枚の生成画像群に対し InceptionV3 を用いて以下の 2 つの観点から算出される．

1. 各画像に対して画像識別器による識別のしやすさ

2. 画像郡内の物体クラスのバリエーションの豊富さ

x_i を画像， y を識別ラベルとすると， 1. は条件付き分布 $p(y|x_i)$ のエントロピーが小さくなると向上する．

一方で 2. は $p(y|x_i)$ を i について積分した周辺分布 $p(y)$ のエントロピーが大きくなると向上する．

$p(y|x_i)$ と $p(y)$ の両分布間の差異を測るために KL ダイバージェンスを用い，これを i について平均化し指数を取ったものが Inception Score である．

$$\text{Inception Score} = \exp(\frac{1}{N}\sum_i \mathrm{KL}(p(y|x_i)||p(y))) \tag{16}$$

生成画像の Inception Score の値が大きいほど，新規性が高いとみなすことができる．

実画像と生成画像との比較がなされないため，類似性を測定することはできないが，テキスト入力による画像生成モデルでは比較対象とすべき実画像が存在しない場合も多く，AttnGAN も含めてこのスコアにより評価される．

5.2 FID

Fréchet Inception Distance (FID)[9] は， IS の実画像の分布が考慮されていないという欠点を改善するため，InceptionV3 から得られる画像の特徴量ベクトルは普遍的に多変量正規分布に従っているという仮定を基に，実画像と生成画像の各特徴量ベクトルの分布間の距離を Fréchet 距離により算出しスコアとしている．

特徴量ベクトルの分布の平均と共分散行列が実画像について μ_r，Σ_r，生成画像について μ_g，Σ_g と得られているとすると，FID は次式で定義される．

$$\mathrm{FID} = ||\mu_r - \mu_g||_2^2 + \mathrm{tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \tag{17}$$

生成画像の FID の値が小さいほど，InceptionV3 から見た特徴量が実画像に近いとみなすことができる．

5.3 SWD

Sliced Wasserstein Distance (SWD)[10] は，Inception Score や FID と異なり InceptionV3 の物体クラスや識別能に依存しないという利点があるスコアである．特に InceptionV3 は識別時に画像サイズを 299×299 に変換しなければならないという欠点があり，近年の GAN 研究における高解像度の生成画像の測定に対しては SWD が用いられることが多い．

Wasserstein Distance は分布間の最適輸送コストから算出される距離関数である．SWD では実画像と生成画像の各ラプラシアンピラミッドの各レベルから 7×7 のパッチを取り出したものをそれぞれ 1 次元に変換して Wasserstein Distance を計算し，画像間の特徴量分布の距離として用いる．

生成画像の SWD の値が小さいほど，エッジなどの構造上の特徴が実画像に近いとみなすことができる．

6 提案手法

本研究では，単語の持つ意味を考慮することで DAMSM の性能向上につながると考え，自然言語処理分野の単語分散表現獲得手法により得られた特徴量ベクトルを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，テキストからの GAN による画像生成の不安定さの緩和のため，AttnGAN の Generator 学習時に実画像と生成画像との FID 及び SWD を損失関数 L_G 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．

7 実験

本研究では CUB 鳥画像データセット [11] 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 64 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の RNN に対する入力として次の 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を入力

表 1: AttnGAN の学習パラメータ

訓練データ画像数	9379
検証データ画像数	2345
テストデータ画像数	64
DAMSM 事前学習時のバッチサイズ	64
AttnGAN 学習時のバッチサイズ	8
最大エポック数	50 エポック
学習データの記録	毎 10 エポック
学習率	0.0002
各特徴量ベクトルの次元数	256
γ_1	4.0
γ_2	5.0
γ_3	10.0
λ	5.0

- 提案手法：自然言語処理分野の単語分散表現獲得手法である Word2Vec，GloVe，fastText，ElMo のそれぞれから学習済みの特徴量ベクトルを入力

上記の入力に対し，それぞれ DAMSM のテキストエンコーダの RNN として LSTM と GRU の 2 種類を取り扱い，計 10 種類の手法で比較実験する．実験 1，実験 2 において，各手法による生成画像群の Inception Score を比較した．表 1 に本研究で設定した AttnGAN の学習パラメータを示す．

7.1 実験 1：未知語を含まないテキストからの画像生成

テストデータのキャプション 64 文を入力テキストとした．テキストの内容は訓練データと似通っており，キャプション内に未知語は含まれない．訓練データに近いテキストの入力から各手法により出力される生成画像を Inception Score で比較する．

7.2 実験 2：未知語を含むテキストからの画像生成

新規に作成した短文 “this bird has [unlearned color] [parts]” を入力テキストとした．[unlearned color] には matplotlib で扱える 148 色のうち，データセットに含まれない未学習の 102 色から 1 色の色名が入る．[parts] には [’crown’ ,’head’, ’eyering’, ’bill’, ’wing’, ’wings’, ’breast’, ’belly’, ’tail’, ’leg’] の 10 部位から 1 つの部位名が入る．未知語が含まれるテキストの入力から各手法により出力される生成画像を観察する．従来手法では入力時に未知語が取り扱われないため，生成画像の Inception Score による比較はしない．

表 2: 各手法により得られた Inception Score

手法	Inception Score	
テキストエンコーダの RNN	LSTM	GRU
先行研究論文掲載値	4.36	-
従来手法	4.36	4.58
Word2Vec	4.56	4.48
GloVe	4.55	4.45
fastText	4.38	4.90
ElMo	4.67	4.91

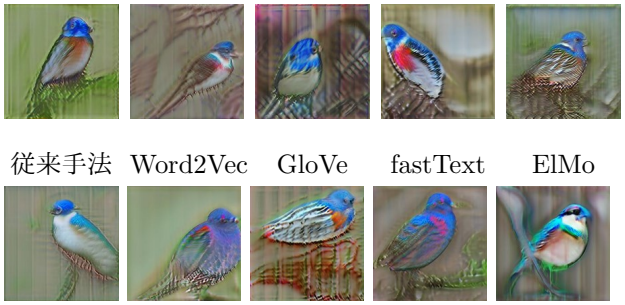

図 3: “the colorful bird has a white belly and is otherwise mostly blue, and it has a short stubby beak.” を入力テキストとして各手法により生成された画像 (上段：LSTM，下段：GRU)

8 結果

8.1 実験 1 の結果

表 2 に実験 1 で各手法により得られた生成画像群の Inception Score を示す．

図 3〜図 5 に実験 1 で各手法により生成された画像の一例を，テキストエンコーダの RNN の種類別にそれぞれ示す．

8.2 実験 2 の結果

図 6〜図 9 に実験 2 で各手法により生成された画像の一例を，テキストエンコーダの RNN の種類別にそれぞれ示す．この時，”deeppink”, ”deepskyblue”, ”orangered”, ”navajowhite” はいずれも学習データに含まれない未知語である．

9 まとめと今後の課題

本研究では，単語の分散表現を用いる DAMSM と GAN との同期的学習を導入した AttnGAN モデルを提案し，従来研究より高い Inception Score を得た．

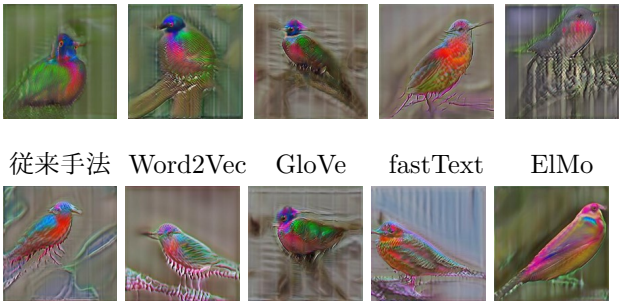

図 4: “a distinct red bird with a blue head, green wings and a red eyering.” を入力テキストとして各手法により生成された画像 (上段：LSTM，下段：GRU)

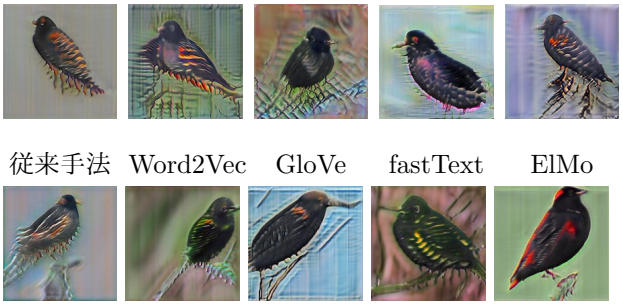

図 5: “medium sized black bird with one orange stripe on the wing, medium tarsus and medium beak.” を入力テキストとして各手法により生成された画像 (上段：LSTM，下段：GRU)

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

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

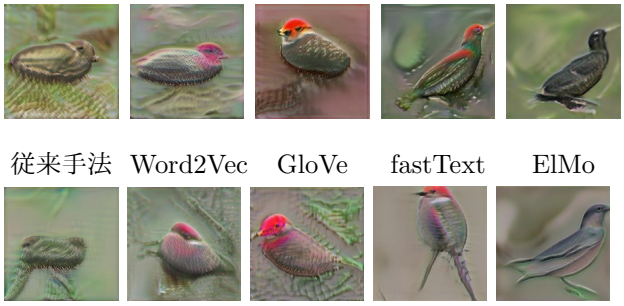

図 6: “this bird has deeppink head”
を入力テキストとして各手法により生成された画像
(上段：LSTM，下段：GRU)

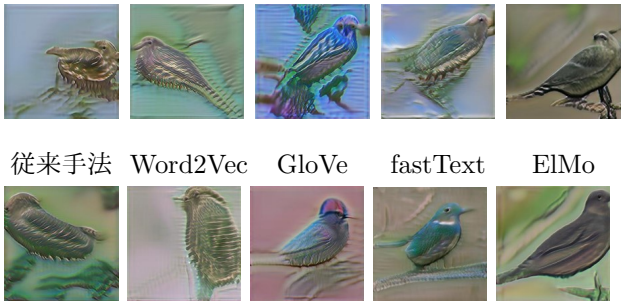

図 7: “this bird has deepskyblue wings”
を入力テキストとして各手法により生成された画像
(上段：LSTM，下段：GRU)

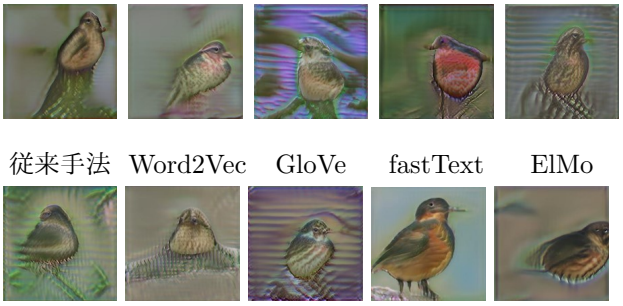

図 8: “this bird has orangered breast”
を入力テキストとして各手法により生成された画像
(上段：LSTM，下段：GRU)

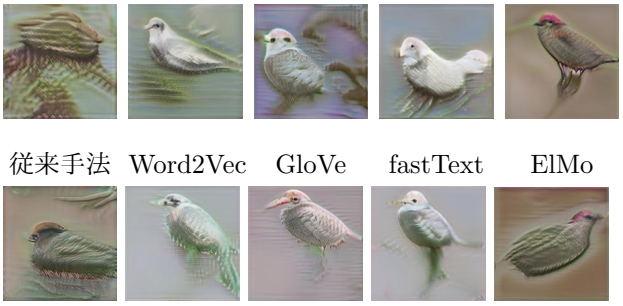

図 9: “this bird has navajowhite crown”
を入力テキストとして各手法により生成された画像
(上段：LSTM，下段：GRU)

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, Vol. abs/1512.00567, , 2015.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[8] Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. Vol. 29, pp. 4844–4852, 2016.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, Vol. abs/1706.08500, , 2017.

[10] G. Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, Vol. 11, pp. 355–607, 2019.

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