# Self-Attention 機構を用いた文章からの画像生成手法

川會　悠生†　　青野　雅樹††

† 豊橋技術科学大学 情報・知能工学課程　〒 441–8580 愛知県豊橋市天白町字雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天白町字雲雀ヶ丘 1-1
E-mail: †kawae.yuki.ix@tut.jp, ††aono@tut.jp

**あらまし**　自然言語の文章から画像の自動生成を行う問題は近年盛んに研究されている挑戦的な研究課題のひとつである。この問題に対して Generative Adversarial Networks（GANs）に基づいた手法がいくつか提案されている。本研究では、その中で Attention 機構を導入した AttnGAN（Attentional Generative Adversarial Network）を参考にし、そのメカニズムを拡張する。拡張として他分野で提案された別な Attention 機構を導入する手法を提案し，実験と考察を行う。提案法の評価には、Inception Score を採用した。実験の結果、ベースモデルと比較して Inception Score の改善が見られた。

**キーワード**　深層学習、生成モデル、Attention 機構

## 1 はじめに

自然言語の文章から画像を生成する問題は Text-to-Image Synthesis と呼ばれ近年盛んに研究されている分野の一つである．Text-to-Image Synthesis はコンピュータ支援によるデザイン，画像検索，画像加工や芸術作品の生成など様々なアプリケーションへ応用可能な基礎技術である．

この問題に対して近年は Generative Adversarial Networks (GANs) [1] を用いた例が報告されている．一般的な手法では文章全体のみを符号化し，ノイズと連結後，GANs を用いた生成モデルへ入力し，画像を生成する．

Xu ら [2] は文章全体のみ考慮する場合，文章に含まれる単語の細かな意味が失われると指摘している．その結果生成される画像には細かな描写が抜け落ちてしまい画像の品質の低下につながる．またこの問題はより複雑な場面を説明した文章を用いて画像を生成する際，特に問題になるとも言われている．近年発表された AttnGAN(Attentional Generative Adversarial Network) [2] ではこの問題を扱うために Attention 機構が導入された．これにより単語レベルの情報を生成画像に反映させる工夫がされた．

Attention 機構は画像認識，機械翻訳など様々な分野で応用されている．Text-to-Image Synthesis と同様に自然言語と画像を扱う Visual QA と呼ばれる問題でも優れた結果を残している．そのような他分野で提案された Attention 機構は Text-to-Image Synthesis の生成モデルに対して応用可能なのではないかと考えた．

本研究では他分野で提案された Attention 機構を AttnGAN に取り入れることで文章をより考慮した精巧な画像を生成するモデルの提案と実験及び考察を行う．はじめにベースモデルとして AttnGAN の一部を拡張した AttnGAN 拡張モデルを実装する．その後，ベースモデルに対して他分野の Attention 機構を導入し，有効性を検討する．

## 2 関連研究

以下で関連研究として GAN を用いた文章からの画像生成手法と我々の研究に最も影響を与えた AttnGAN [2] について述べる．

### 2.1 GAN を用いた文章からの画像生成手法

Text-to-Image Synthesis では GANs を用いた例がいくつか報告されている．Agnese ら [3] の調査によれば大きく分けて単純な GAN 手法と発展的な GAN 手法の 2 種類があり，発展的な GAN 手法はその中で大きく 4 種類に分類できる．発展的な GAN 手法の分類としては Semantic Enhancement GANs，Resolution Enhancement GANs，Diversity Enhancement GANs，Motion Enhancement GANs の 4 種類がある．

本研究で着目する AttnGAN [2] は Resolution Enhancement GANs に分類される．Resolution Enhancement GANs に分類される手法は低解像度から徐々に高解像度の画像を生成する複数の GANs を持つ．この分類の他の代表的な手法には StackGAN [4]，StackGAN++ [5]，DM-GAN [6] などがある．

StackGAN は 2 つの GANs を持つ．Stage-I GAN では文章とノイズから物体の色や形などの原型を低解像度で生成する．Stage-II GAN では再び文章を考慮し，Stage-I GAN の低解像度画像に対する欠損の修復を行う．Stage-II GAN の出力は $256 \times 256$ サイズの高解像度画像となる．また StackGAN では Conditioning Augmentation (CA) と呼ばれる手法を提案した．CA により入力の潜在変数が不連続となる問題を軽減する．

StackGAN++は複数の GANs を木構造のように持つ．最初の GAN で低解像度の画像を生成し，その際の潜在変数を次の GAN へ入力することで徐々に画像を高解像度にする．また損失関数として文章による条件付き項だけでなく，条件無し項を加えている．条件付き項では生成画像が文章に一致しているかを考慮し，条件無し項では生成画像が本物か偽物かを考慮する．

StackGAN++も同様に 256 × 256 サイズの高解像度画像の生成を可能とする．

最近発表された DM-GAN では Dynamic Memory 構造を取り入れることで初期に生成される画像の修正に焦点を当てている．またそのほかにもこれらを拡張した様々なモデルが提案されている [7] [8].

### 2.2 AttnGAN

Xu ら [2] は文章全体を符号化した単一のベクトルのみ条件として画像を生成すると細かな単語情報が失われると指摘する．

AttnGAN [2] はこの問題に対して Attentional Generative Network と Deep Attentional Multimodal Similarity Model (DAMSM) と呼ばれる構造を持つ．本研究では特に Attentional Generative Network のネットワークに着目した拡張を行う．Attentional Generative Network は単語と関係のある領域を描画する Attention 機構を持つ．

Attentional Generative Network は階層状に複数の Generator があり，各階層では前層の潜在変数を入力として与えることで低解像度画像から高解像度画像を生成する．以降では潜在変数を画像特徴量 (Image Feature) と呼ぶ．各階層の間には Attention 機構を持つモデル $F^{att}(e,h)$ があり，この出力を用いて単語と関係のある部分領域の描画を行う．$F^{att}(e,h)$ への入力は単語を符号化した単語特徴量 (Word Feature) $e \in \mathbf{R}^{D\times T}$ と画像特徴量 $h \in \mathbf{R}^{\hat{D}\times N}$ である．$D$ と $\hat{D}$ はそれぞれ単語特徴量と画像特徴量の次元，$T$ は単語長，$N$ は画像特徴量の部分領域を表している．部分領域 $N$ は高さ $H$，幅 $W$ の画像特徴量から $N = H \times W$ によって変形した値となる．Attention 機構 $F^{att}(e,h)$ の処理について (1)〜(4) 式に示す．(1) 式では $U \in \mathbf{R}^{\hat{D}\times D}$ により単語特徴量を画像特徴量と同じ次元に写像した $e'$ を得る．次に (2)(3) 式により単語特徴量 $e'$ と画像特徴量 $h$ の積を計算し，単語に関して Softmax で正規化した $\beta_{j,i}$ を得る．$\beta_{j,i}$ は $i$ 番目の単語が $j$ 番目の部分領域に対する重要度のような値である．その後 (4) 式で画像特徴量の $j$ 番目の部分領域に対応する word context ベクトル $c_j$ を得る．word context ベクトルを画像特徴量と連結して次の階層の入力としている．

$$e' = Ue \tag{1}$$

$$s'_{j,i} = h_j^T e'_i \tag{2}$$

$$\beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1}\exp(s'_{j,k})} \tag{3}$$

$$c_j = \sum_{i=0}^{T-1}\beta_{j,i}e'_i \tag{4}$$

Attentional Generative Network の主な損失関数を (5) 式に示す．(5) 式の $L_G$ は (6) 式のように各階層の Generator の損失関数を合計した値となる．(6) 式で示した各 Generator の損失関数 $L_{G_i}$ を (7) 式に示す．第 1 項が条件無し項で生成画像が本物か偽物であるかを考慮し，第 2 項の条件付き項は生成画像が文章全体の意味を捉えているかを考慮する．$m$ は階層の総数，$G_i$ は各階層の Generator，$D_i$ は対応する Discriminator，$\bar{e}$ は文章全体を符号化したベクトル (Sentence Feature)，$\hat{x}$ は生成画像，$\lambda$ はハイパーパラメータである．

$$L = L_G + \lambda L_{DAMSM} \tag{5}$$

$$L_G = \sum_{i=0}^{m-1} L_{G_i} \tag{6}$$

$$\begin{aligned} L_{G_i} = &-\frac{1}{2}E_{\hat{x}_i\sim p_{G_i}}[\log(D_i(\hat{x}_i))] \\ &-\frac{1}{2}E_{\hat{x}_i\sim p_{G_i}}[\log(D_i(\hat{x}_i,\bar{e})] \end{aligned} \tag{7}$$

各階層の Generator は対応する Discriminator を持つ．Discriminator の主な損失関数は (8) 式のようになる．(8) 式の $x$ は実画像 (正解データ) である．第 1 項と 2 項が条件無し項で生成画像が本物か偽物であるかを考慮し，第 3 項と 4 項が条件付き項で生成画像が文章の意味を捉えているかを考慮する．Discriminator の損失関数は階層毎に計算される．

$$\begin{aligned} L_{D_i} = &-\frac{1}{2}E_{x_i\sim p_{data_i}}[\log(D_i(x_i))] \\ &-\frac{1}{2}E_{\hat{x}_i\sim p_{G_i}}[\log(1-D_i(\hat{x}_i)] \\ &-\frac{1}{2}E_{x_i\sim p_{data_i}}[\log(D_i(x_i,\bar{e}))] \\ &-\frac{1}{2}E_{\hat{x}_i\sim p_{G_i}}[\log(1-D_i(\hat{x}_i,\bar{e})] \end{aligned} \tag{8}$$

(5) 式第二項の $L_{DAMSM}$ は DAMSM による損失関数である．DAMSM はテキストエンコーダとイメージエンコーダの 2 つのニューラルネットワークで構成されるネットワークである．DAMSM により文章全体および単語と生成画像の類似度を測る．$L_{DAMSM}$ は文章の意味を捉える画像の生成を行う損失関数として与えられる．

Xu らは 3 階層から構成される構造によって 256 × 256 サイズの画像を生成しているが，本研究では 2 階層の構造によって 128 × 128 サイズの画像を生成する．以降は 2 階層のモデルを AttnGAN と呼ぶ．AttnGAN の一部を図 1 に示す．本研究では区別のために各階層の Generator に相当する処理に対して First Stage Generator と Second Stage Generator と名前をつけた．図 1 の主に First Stage Generator と Second Stage Generator が Attentional Generative Network に相当する．

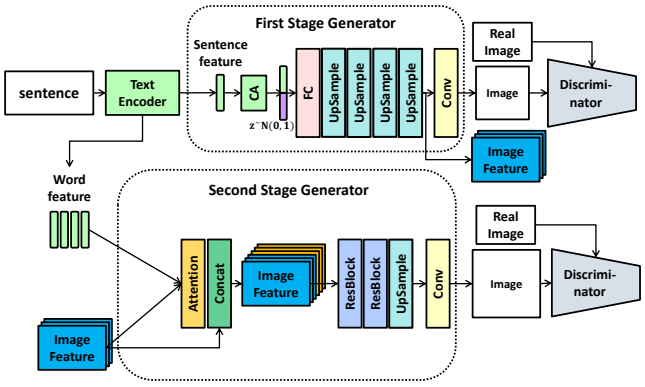



図 1　従来手法 (AttnGAN の一部)

## 3　ベースモデル

本研究におけるベースモデルである AttnGAN 拡張モデルについて述べる．AttnGAN 拡張モデルを図 2 に示す．

前節で述べた AttnGAN [2] の Attention 機構では単語特徴量と画像特徴量を Attention 後，その出力を画像特徴量と連結する．AttnGAN 拡張モデルでは連結後，畳み込み層，Batch Normalization と Gated Linear Unit (GLU) [9] を追加した．ベースモデルの連結後に導入した GLU は入力をチャネル方向に 2 分割し，後半のデータを Sigmoid で処理後に前半のデータとの要素積を取るようにして用いる．GLU は Xu ら [2] の AttnGAN の実装においてすでに用いられているが連結後に導入することで，以降の処理で扱う入力サイズを削減する．AttnGAN 拡張モデルの [Attention，連結 (concat)，畳み込み (conv)，GLU] をまとめて Attention Block と呼ぶことにする．畳み込みはカーネルサイズ $3 \times 3$、ストライド 1、ゼロパディング 1 としている。

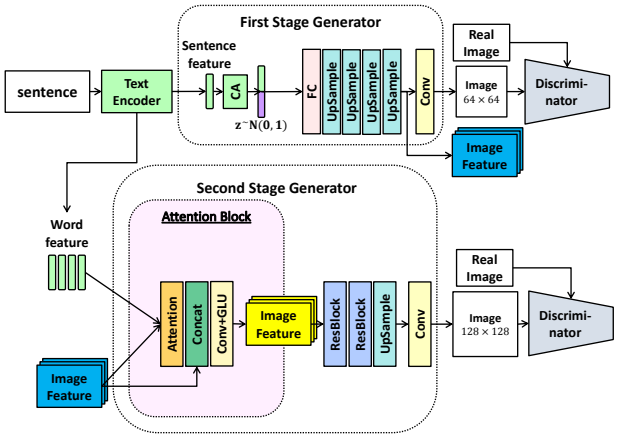
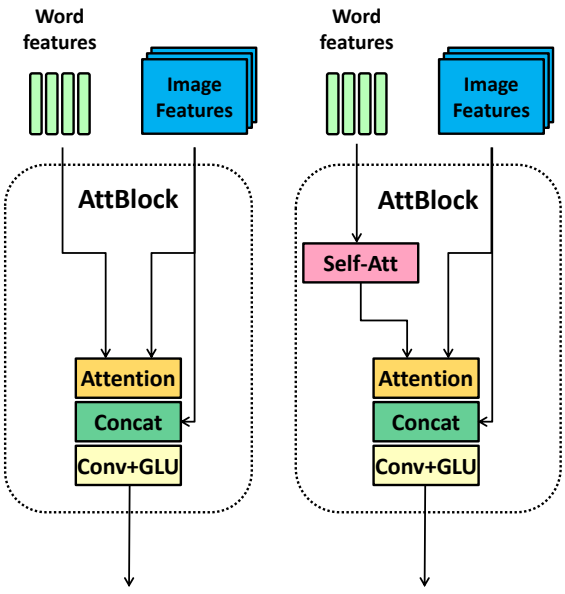


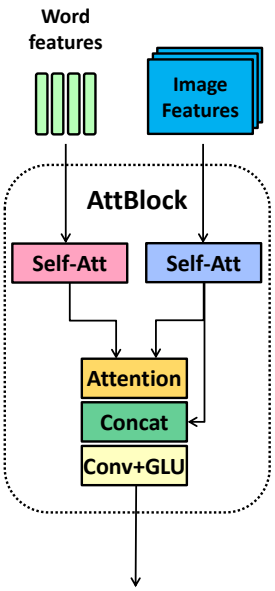

図 2　ベースモデル (AttnGAN 拡張モデル)

## 4　提案モデル

提案モデルはベースモデルに加えて 2 種類の Attention 機構を追加した．ベースモデルと提案モデルの Attention Block を図 3 に示す．4.1 節と 4.2 節ではベースモデルの Attention Block に追加した 2 種類の Attention 機構について述べる．4.3 節 Att1 提案モデルでは単語特徴量に対して Attention 機構を導入した提案について述べる．4.4 節 Att2 提案モデルでは単語特徴量だけでなく画像特徴量に対しても Attention 機構を導入した提案について述べる．




(a) ベースモデル　　(b) Att1 提案モデル




(c) Att2 提案モデル

図 3　ベースモデル及び提案モデルの Attention Block

### 4.1　Modular Co-Attention Networks

Modular Co-Attention Networks (MCAN) は Yu ら [10] によって提案された．Yu らは Self-Attention (SA) と Guided-Attention (GA) で構成された 3 種類の Modular Co-Attention Layer (MCA Layer) を挙げ，MCA Layer を用いた MCAN を提案している．図 3(b)(c) に示す提案モデルはこの 3 種類の MCA Layer の形を参考に実装した．

図 3(b) で示す Att1 提案モデルでは単語特徴量に対する Attention として MCAN で用いられる Self-Attention 機構 (図 4) を用いた．Yu らは Self-Attention と Guided-Attention について Vaswani ら [11] の提案から着想を得たと述べているが，

本研究では Yu らの実装を参考とする．Self-Attention 機構では入力に対して Multi-Head Attention を行い，その出力と入力を加算し，Layer Normalization で正規化を行う．さらにその出力をフィードフォワードなネットワークにより変換し，加算後，再び Layer Normalization で正規化を行うという処理の流れである．本研究で Self-Attention 機構に適用した Dropout 率は全て 0.1 である．また Multi-Head 数は 8 とした．

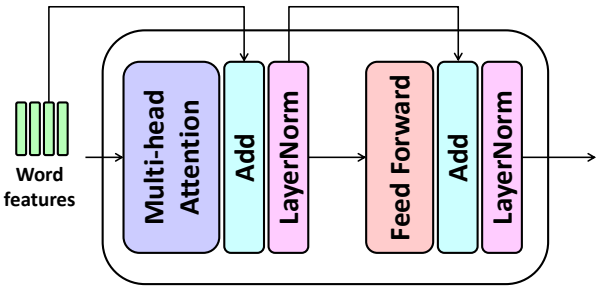



図 4　単語特徴量に対する Self-Attention 機構

### 4.2　Convolutional Block Attention Module

画像特徴量に対する Attention は画像特徴量サイズが大きく，計算時間がかかることが予想されるため，軽量な Attention 機構が望ましい．そこで画像特徴量に対する Attention として比較的軽量な機構である Convolutional Block Attention Module (CBAM) [12] を用いる．

CBAM は Channel Attention Module と Spatial Attention Module と呼ばれる 2 つの構造を持つ Attention 機構である．CBAM の概要を図 5 に示す．Channel Attention Module はチャネル方向に対する Attention で意味のあるものに焦点を当てる働きがある．空間方向に MaxPool と AveragePool を行い，共通の MLP に通す．その出力を加算し，Sigmoid した重みを元のデータに要素積を行うことで Attention により強調されるパターンを反映させる．Spatial Attention Module は空間方向に対する Attention で何がどこにあるかに焦点を当てる働きがある．チャネル方向に MaxPool と AveragePool を行い，連結後，畳み込み層に通す．その出力を Sigmoid した重みを元のデータに要素積を行うことで反映させる．最後にその出力を元の入力と加算する．本研究では Spatial Attention Module の畳み込み層のカーネルサイズを $5 \times 5$ とした．本研究では CBAM による Attention を画像に対する Self-Attention と呼ぶ．

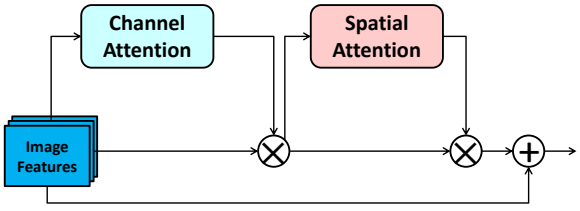



図 5　Convolutional Block Attention Module

### 4.3　Att1 提案モデル

Att1 提案モデルはベースモデルの Attention Block に比べて単語特徴量に対して Attention 機構を導入している．Att1 提案モデルの Attention Block を図 3(b) に示す．Att1 提案モデルは 4.1 節で述べた Self-Attention 機構を単語特徴量に対して適用する．これにより単語特徴量の重要なパターンを強調し，画像特徴量の Attention において重要単語が示す部分領域がより強調されることを狙う．

### 4.4　Att2 提案モデル

Att2 提案モデルでは単語特徴量だけでなく画像特徴量に対しても Attention 機構を導入する．Att2 提案モデルの Attention Block を図 3(c) に示す．Att2 提案モデルは 4.2 節で述べた Attention 機構を Att1 提案モデルの Attention Block に追加した．これにより画像特徴量に対しても，視覚的な物体位置などの領域をあらかじめ示すことで，単語特徴量との Attention を有効に作用させることを狙う．

## 5　実　　験

本研究で実施した実験について述べる．5.1 節では実験設定として使用したデータセットやその他の実験環境について述べる．5.2 節では評価手法である Inception Score [13] と生成画像比較について述べる．5.3 節では Inception Score を用いた定量的な評価の結果を示す．またベースモデルと提案モデルの生成画像に対して定性的な評価を行う．

### 5.1　実 験 設 定

実験設定を表 1 に示す．各種実験設定及び実装は Xu ら [2] の実装を参考にした．また損失関数は Xu らの実装と同様にし，DAMSM は学習済みのモデルを用いる．本研究ではデータセットとして The Caltech-UCSD Birds-200-2011 Dataset(CUB) [14] を用いた．CUB データセットは 200 クラスで全 11,788 枚 (訓練:150 クラス 8,855 枚，テスト:50 クラス 2,933 枚) の画像が有る．また各画像に対して 10 個のキャプションを持つ [4], [15].

表 1　実 験 設 定

| | |
|---|---|
| 最適化関数 | Adam |
| Discriminator 学習率 | 0.0002 |
| Generator 学習率 | 0.0002 |
| バッチサイズ | 20 |
| 最大エポック数 | 600 |
| 評価手法 | Inception Score |

### 5.2　評 価 手 法

各モデルに対する評価は Xu ら [2] を参考に Inception Score [13] を用いる．Inception Score の計算式を (9) 式に示す．Inception Score は画像の多様性と識別可能性を測る尺度として使用される評価手法である．$\mathbf{x}$ は生成画像，$y$ は inception モデルによるラベルの予測結果である．条件付き確率 $p(y|\mathbf{x})$ は生成画像に対する識別が行いやすいほどエントロピーが小さくなる．

一方周辺確率 $p(y)$ は生成画像が多様なラベルに分類されるほどエントロピーが大きくなる．その結果 $KL$ ダイバージェンス ($\mathrm{D_{KL}}$) は周辺確率 $p(y)$ と条件付き確率 $p(y|\mathbf{x})$ の分布の差が大きいほど大きくなる．Inception モデル [16] には StackGAN [4] を参考に CUB データセット用に fine-tune されたモデルを使用した．Inception Score の計測を各モデルで生成した 30,000 枚の画像に対して行う．ただし全ての生成画像に対して一度に Inception Score の計測を行うのではなく，10 分割しその平均と標準偏差を結果として示す．

$$\mathrm{InceptionScore} = \exp(\mathbf{E}_{\mathbf{x}}\mathrm{D_{KL}}(p(y|\mathbf{x})||p(y))) \tag{9}$$

Inception Score では文章の描写を適切に反映しているかを調べることができない．そこで Inception Score が高いモデルで生成した一部の画像に対して比較を行い，文章の描写を捉えることができているか調べる．またその結果から InceptionScore との関係を考察する．

### 5.3 Inception Score による評価

5.3.1 節では Inception Score [13] による結果を示す．5.3.2 節および 5.3.3 節では生成画像を示し，定性的に評価する．

#### 5.3.1 各モデルの Inception Score の結果

各モデルにおける inception Score [13] の計測結果を表 2 に示す．表 2 に示すベースモデル以下の結果は 300epoch から 600epoch まで 10epoch 毎に Inception Score を計測した最大値である．表 2 の 1 行目にある AttnGAN(論文) は Xu ら [2] の計測による結果である．なお本研究における生成画像の画像サイズが 128 × 128 サイズであるのに対して，Xu らの結果は 256 × 256 サイズである．

表 2 よりベースモデルに比べて提案モデルの Inception Score が高いことがわかる．

表 2　Inception Score

| モデル | Inception Score |
|---|---|
| AttnGAN(論文) | $4.36 \pm 0.03$ |
| ベースモデル | $4.43 \pm 0.05$ |
| Att1 提案モデル | $\mathbf{4.59 \pm 0.06}$ |
| Att2 提案モデル | $4.50 \pm 0.05$ |

#### 5.3.2 各モデルの生成画像

図 6 に示す生成画像はそれぞれキャプション 1 から各モデル 10 枚生成した結果である．

図 6 より配色，ポーズや背景がモデル毎に違うことがわかる．例えば図 6(a) は比較的黒色が強く，図 6(b)(c) は色の境界がはっきりしているように見える．また同じモデルで生成した 10 枚を比べると，物体の配色は似ているが，背景やポーズなどが画像毎に異なることがわかる．

---

キャプション 1
“this small bird has a bright fiery red breast tapering down to a yellow tail that is contrasted by a completely black head.”

---

(a) ベースモデル

(b) Att1 提案モデル

(c) Att2 提案モデル

図 6　キャプション 1 に対する生成画像

#### 5.3.3 キャプションの一部を変更した生成画像

図 7，8，9 は各モデルで，キャプション 2 の “red” を “black”，“yellow”，“blue” に変えた時に生成画像がどのように変化するかを示す．

ベースモデルによる図 7(c) は全体的に色が広がっているように見えるが，提案モデルによる図 8，9 は “crown” の位置が特に色の情報を反映しているように見えるなど文章の内容と合っている．しかし “blue” に関してはどのモデルも全体的に色の広がりが見える．また “speckles” に対して色の情報が反映されていない．

キャプション 2
“this smaller bird has a gray belly and breast with **red** speckles, a **red** crown, and a short pointy bill.”

(a) “red”

(b) “black”

(c) “yellow”

(d) “blue”

図 7　ベースモデルによるキャプション 2 に対する生成画像

(a) “red”

(b) “black”

(c) “yellow”

(d) “blue”

図 8　Att1 提案モデルによるキャプション 2 に対する生成画像

(a) “red”

(b) “black”

(c) “yellow”

(d) “blue”

図 9　Att2 提案モデルによるキャプション 2 に対する生成画像

## 6　考　　察

はじめに表 2 の Inception Score [13] の結果をもとに各モデルの評価を行う．結果より提案モデルは生成画像の多様性や識別可能性に向上の可能性を示した．特に Att1 提案モデルが最も高い値を記録した．一方で Att2 提案モデルは Att1 提案モデルに比べて Inception Score が低い結果となった．原因として CBAM [12] はチャネル方向と空間方向ともに MaxPool と AveragePool を用いる Attention 機構であることが挙げられる．MaxPool と AveragePool を用いる Attention 機構では Attention による強調箇所が大きくなるため，詳細な描写が失われているのではないかと考えられる．

次に図 6 に示す生成画像を比較する。ベースモデル (図 6(a)) では全体的に黒色が目立つ．これは “black”が画像全体に影響を及ぼしているためであると考えられる．一方提案モデル (図 6(b)(c)) では色の境界がある．そのため頭や胴体の区別をつけやすい．また提案手法間の比較では “red breast”に注目すると Att2 提案モデル (図 6(c)) の方が Att1 提案モデル (図 6(b)) よりも特徴を捉えているように見える．一方図 6(c) に比べて図 6(b) の方が物体の形が崩れていないため頭や胴体などを区別しやすい．

また 5.3.3 節で述べたように図 7，8，9 に関しても一部の画像で配色によってパーツの区別に違いが見られる．このように生成画像の一部の例でパーツの区別や配色に違いが見られる．Xu ら [2] が指摘しているように，単語レベルの描写ができることにより，頭や胴体などのパーツの区別や配色などの面で画像品質が向上し，その結果識別が行いやすくなり，Inception

Score の向上につながったのではないかと考えられる．

以上から Att1 提案モデルと Att2 提案モデルの両方に導入した単語特徴量に対する Attention 機構が有効である可能性が高い．また画像特徴量に対する Attention 機構は Inception Score の結果で単語特徴量のみ Attention 機構を導入したモデルに劣るが，生成画像の一部において色の情報が反映できている画像があることから，改善の余地があると考えられる．

## 7 終わりに

本研究では MCAN で実装された 3 種類の MCA Layer の構造を参考に，他分野で提案された Attention 機構を導入したモデルを提案し，実験と考察を行った．Inception Score を用いた定量的な評価の結果，提案モデルがベースモデルより高いスコアを記録した．特に単語特徴量に対する Attention を行う Att1 提案モデルが Inception Score において有効である可能性が高い結果を示した．さらにベースモデルと提案モデルの生成画像を比較した結果，一部の画像においてベースモデルよりも文章に沿った細かな描写が可能であること確認した．今後の課題として Inception Score 以外の評価手法による各モデルの評価，他のデータセットでの実験，ベースモデル及び提案モデルの改良と $256 \times 256$ サイズの画像生成が考えられる．また試行回数を増やすことで有効性や再現性についてさらに検討する必要がある．

## 謝　辞

本研究の一部は，科研費基盤 (B) (課題番号 17H01746) の支援を受けて遂行した．

## 文　献

[1] 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*, pages 2672–2680. Curran Associates, Inc., 2014.

[2] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. 10 2019.

[4] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. 10 2017.

[6] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dmgan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] Y. Cai, X. Wang, Z. Yu, F. Li, P. Xu, Y. Li, and L. Li. Dualattn-gan: Text to image synthesis with dual attentional generative adversarial network. *IEEE Access*, 7:183706–183716, 2019.

[8] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2063–2073. Curran Associates, Inc., 2019.

[9] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 933–941. JMLR.org, 2017.

[10] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[13] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

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

[15] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.