コンテキストアウェア敵対的生成ネットワークによる画像生成

中村　玄貴† 　馬　　強††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町 36-1
†† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町 36-1
E-mail: †nakamura-kenki@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　敵対的生成ネットワークを用いてテキストから画像を自動生成する研究が盛んに行なわれている．既存研究の多くは，一つの入力テキストからそれに対応する画像生成を行っているため，生成される画像が不自然であったり，関連性の強いテキストを入力しても生成された画像間の関連がまったくなかったりする場合がある．そのため，テキストの系列から関連性のある画像を生成することが困難である．そこで，本研究では，複数のテキストから複数の画像生成のためのコンテキストアウェア敵対的生成ネットワークモデルを提案し，入力するテキストと出力画像の関連性を考慮した画像を生成できるようにする．

キーワード　敵対的生成ネットワーク (GAN), ニューラルネットワーク, 画像生成

1. はじめに

深層学習を利用した，画像とテキストの相互生成の研究が多くなっている [1] [2] [3] [4] [5]．画像からそのタグやテキストを生成するため，画像とタグのペアを学習データとして用意してニューラルネットワークに学習させて，入力画像のタグを生成する手法が提案されている [1]．またそれを応用し，画像を複数領域に分割して，それらの領域に含まれる物体を検出し，それらを整理して人間の理解可能な文章に仕上げことで画像の説明文を自動生成する研究も行われている [2]．

画像生成の既存手法にはオートエンコーダを用いるものや物体認識のニューラルネットワークを用いるものがあるが，データが多くなるとぼやけてしまったり，原型をとどめていないものの映った画像が生成されてしまう．そのため，Ian らは画像生成のフレームワークとして GAN(敵対的生成ネットワーク) を提案している [3]．よりリアルな画像生成が可能である，GAN は現在の主要な画像生成のフレームワークとなっている．GAN は Generator と Discriminator と呼ばれる二つのニューラルネットワークから構成される相互的に学習を行うフレームワークである．Generator は Discriminator を欺ける画像を生成し，Discriminator は Generator から生成された画像を入力として，その画像が Generator により生成された偽物であるか，本物の画像であるかの判断を行う．Discriminator と Generator を交互に学習して，より本物に近い画像を生成可能としている．

Radford らは GAN に畳み込みニューラルネットワークを組み合わせ多層化することで鮮明な画像生成を行うモデルを提案している [4]．Reed らは GAN を拡張して入力テキストに対応する画像を自動生成するモデルを提案している [5]． また Huang らは複数の画像からストーリーを生成する手法を提案している [6]．しかし，我々の知る限りでは複数のテキストや画像を考慮して画像を生成する研究はまだない．

例えば，SNS へ旅行の思い出を掲載した際，適切な写真を撮っていなかった場合，生成して補いたい場面があったとする．その写真は一つのテキストの内容からのみ生成されるべきではなく，前後の写真やテキストの内容を踏まえたものとなるべきである．また，歴史的文献の写真部分が見えなくなっていた場合，検索してもそれを補える画像は見つからない．前後のテキストや画像から画像を自動生成できれば，このような欠落したコンテンツを補完できる．

Reed らの提案している従来の手法 [5] を用いて一つのテキストから画像を生成することは可能であるが，前後のテキスト・画像を考慮していないため，前後の画像やテキストと全く無関係の画像を生成してしまう可能性が高い．

そこで本研究では，この前後の画像やテキストのような関連画像・テキストをコンテキストと呼び，コンテキストに適した画像生成手法を提案する．具体的に，GAN のハイパーパラメータを自動調整する手法と，GAN の Generator と Discriminator をそれぞれ拡張して，コンテキストを扱える Context-aware GAN モデルを提案する．

- コンテキストを考慮したハイパーパラメータの自動調整:GAN のハイパーパラメータを自動調整して出力画像とコンテキスト画像との関連性を保つアプローチである．GAN の入力であるハイパーパラメータは生成される画像にランダム性を持たせる役割を果たしている．つまり，生成される画像の内容はハイパーパラメータに依存する．この値を生成される画像とコンテキストの類似などの関連性を考慮して調整し，出力画像をコンテキストに沿った内容にする手法を提案する．
- Context-aware GAN モデル:GAN を拡張して，テキストと画像のペアの系列を学習させて，テキストに対応するコンテキストに沿った画像を自動生成する Context-aware GAN を構築する．

ハイパーパラメータの調整手法では，コンテキスト画像をあらかじめ与えるものでもよいが，入力のテキストを用いて類似画像を検索してきて代用することも可能である．一方，Context-aware GAN の場合，コンテキストを含めてモデルを学習しているため，コンテキストをあらかじめ与える必要が

表 1　本研究と関連研究の位置付け

画像:テキスト	画像→テキスト	テキスト→画像
1：1	Vinyals らの手法 [2]	Reed らの手法 [5]
複数：複数	Huang らの手法 [4]	本研究

ある．

本論文の構成は以下である．2 節で関連研究について紹介する．3 節で提案手法の詳細を述べる．4 節で実験結果とその考察について述べる．5 節でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 画像の自動生成

Ian らの提案している GAN は画像生成を行うニューラルネットワークのフレームワークである [3]．Generator と呼ばれる画像生成を行うニューラルネットワークと Discriminator と呼ばれる画像の判別を行うニューラルネットワークの二種類から構成されている．

Discriminator は入力された画像が本物画像 X であるか生成画像 G(Z) であるかを判別できるように学習させていく．出力は入力画像が本物である確率を [0,1] で出力する．Generator は Discriminetor を欺くことのできるような画像を生成できるように生成する．

Discriminator を学習させたのち，Generator によって生成された画像が Discriminator を騙せるように Generator のパラメータを学習していく．生成画像 G(Z) に対して Discriminator が本物の画像と判別する確率 D(G(Z)) が高くなるよう Generator にフィードバックさせる．また Generator を学習させたのち，それの出力画像と本物の画像を判別できるように Discriminator を学習させていく．このように交互に Generator と Discriminator を学習させていくことで本物のような画像を Generator が生成することを可能にしている．

Radford らは GAN に畳み込みニューラルネトーワークを用いて多層化することにより，それよりもさらにリアルな画像生成を可能にする DCGAN を提案している [4].

Reed らは GAN を拡張し，テキストを入力としてそれに対応する画像を生成するモデルを提案している [5]．Generator の入力ベクトルを全て乱数にするのではなく，そこにテキスト情報をベクトル化したものを用いることでその内容に沿った画像の自動生成を行う．また，Discriminator は GAN では入力される画像が生成画像か本物の画像かを見分けるものであったが，ここでは入力画像が本物の画像かどうかに加え，テキスト情報と一致しているかどうかを判断する．入力画像が本物であり，かつテキストがその内容を表すときのみ 1 と判断するように学習させていく．そのため，学習データにはテキスト (キャプション) と画像がセットになったデータセットを用いる必要がある．さらに Generator は z とテキストのベクトルを入力して画像生成を行うものであるが，テキストベクトルと画像を入力することで z を逆算することもできる．これにより求められた z は入力に用いた画像のスタイルを保持したまま，テキスト情報を踏まえた画像を出力できるものになる．

2.2 入力画像に対応するテキスト生成

Vinyals らは CNN と RNN を組み合わせたニューラルネットワークを用いて入力された一枚の画像に対応する説明文を生成する手法を提案している [2]．画像の特徴抽出を行うニューラルネットワークである CNN を用いて入力画像を解析した後，文章などの連続的な情報を利用して新たなテキスト生成を可能とするニューラルネットワークである RNN を用いて自然な文章を生成する．画像とテキストの入出力は入れ替わっているが，Reed らの提案する手法と同様に学習データにはテキスト (キャプション) と画像がセットになったデータセットを用いる必要がある．

Huang らは複数の写真 (4,5 枚程度) をアルバムとしてそれぞれの写真が一連の物語になるようにテキストを生成する手法を提案している [6]．Flickr から写真を集め，その中から数人のクラウドワーカーが，それぞれ複数枚選んでアルバムを作り，それぞれの写真に物語をつける．そして別のクラウドワーカーがこれらのアルバムを比べて，最も物語を書きやすいと思うものを選択して物語を書く．この時に選ばれなかったアルバムは破棄される．このように，クラウドワーカーの集合知を利活用して複数画像であるアルバムを説明するストーリを生成する．

2.3 動画の自動生成

Vondrick らはコマ送りになった画像の列となっているデータセットから動画を生成する GAN の拡張モデルを提案している [7]．GAN では入力された z に対して直接演算を行って，画像を生成するが，この手法では Generator 内で z を分岐させ，Foreground 部と Background 部のニューラルネットワークにそれぞれ代入して前景と背景の画像をそれぞれ生成し，合成することで動画生成を行う．Background 部は全ての動画のコマで共通部分となる背景を生成する．つまり，背景は Background 部の出力は一枚の画像である．Foreground 部はオブジェクトがコマ送りに動いている複数の画像を前景として生成する．Mask は Foreground のオブジェクトを反転させた画像となっていて，Background を合成するときに背景からオブジェクト部分を抜き取り合成できるようにするために用いられる．

3. 文脈を考慮した画像生成

本研究では，以下の二つのアプローチからコンテキストに即した画像の自動生成を試みる．

- コンテキストを考慮したハイパーパラメータの自動調整: 学習済みの Generator のハイパーパラメータを自動調整して出力画像をコンテキストとの関連を向上させる手法である．
- Context-aware GAN:テキストと画像のペアの系列を学習することでコンテキストに即した画像を自動生成する GAN の拡張モデルである．

提案手法は二つとも入力はコンテキスト画像であり，出力は入力コンテキスト画像を考慮した画像となる．提案手法 1 については入力にテキストを受け付ける Reed らのモデルに対しても用いることができ，その場合の入力はテキストとコンテキスト画像となる．

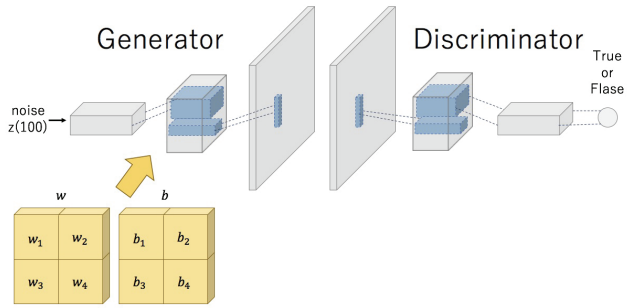

図 1　GAN の構成とパラメータ

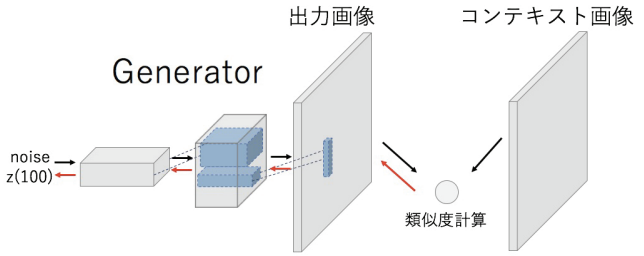

図 2　z の修正方法

3.1　文脈を考慮したハイパーパラメータの自動調整

GAN では入力にハイパーパラメータ z を入力として受け付ける．図 1 は Generator と Discriminator の略図である．Generator は正規分布に従う乱数を生成して 100 次元のベクトル z を生成し，毎回ランダムに生成された z の値に対して演算を行うことで画像を作る．この z は生成される画像にバリエーションを持たせる働きがある．一方，GAN のパラメータ (w, b) は学習後変化することはないので，生成時に毎回違う値を与えることで毎回異なった画像を出力する．

z がランダムに生成されるため，生成された画像の中身が予測できない．そのため，コンテキストに即した画像を生成できるように z を自動調整する手法を提案する．

提案手法は以下の二つのステップからなる．

（1） ランダムに生成された z に対して，GAN の構築手法を利用して w,b を学習させる．w, b はニューラルネットワークにおける重みと閾値であり，これらの値は入力された z と演算することで適切な出力が可能となるように学習させる．図 1 の重み w, バイアス b の関数である L を w, b で偏微分した値を w, b から引いて変化させていくことを繰り返すことで学習させていく．ステップ (1) の誤差関数は以下のように定義される．

$$\arg\min_{w,b} L(w,b) \tag{1}$$

（2） ステップ 1 で得られた w と b を用いて，z を学習させる．Generator によって生成された画像とコンテキスト画像の差を最小となるように z を学習する．

$$\arg\min_{z} S(G(z),C) \tag{2}$$

ただし G は Generator の関数であり，G(z) は Generator に z を入力した時の出力画像である．C はコンテキストのテキストまたは画像である．本研究では C を画像として類似度の計算を行っているが，発展的課題としてテキスト間類似度を用いるときは Jaccard 係数などを用いて，テキスト間の類似度を画像間の類似度と比例するように用いる予定である．

Algorithm 1 hyperparameter z training algorithm.

1: **for** number of GAN training iteration (Step 1) **do**
2: *Sample minibatch of m noise samples* $\{z^{(1)},...,z^{(m)}\}$
3: *Sample minibatch of m examples* $\{x^{(1)},...,x^{(m)}\}$ *from data*
4: *distribution* $p_{data}(x)$
5: *Update the discriminator by ascending its stochastic gradient*：

$$\nabla\theta_d \frac{1}{m}\sum_{i=1}^{m}[logD(x^{(i)}) + log(1 - D(G(z^{(i)})))]$$

6: *Sample minibatch of m noise samples* $\{z^{(1)},...,z^{(m)}\}$
7: *Update the generator by descending its stochastic gradient*：

$$\nabla\theta_g \frac{1}{m}\sum_{i=1}^{m}[log(1 - D(G(z^{(i)})))]$$

8: **end for**
9: **for** number of z training iteration (Step 2) **do**
10: *Generate images with sample z from noise prior* $p_g(z)$ *as input*

$$G(z)$$

11: *Update z by using the error back propagation method as follows*

$$z \leftarrow z - \alpha\frac{\partial S(G(z),C)}{\partial z}$$

12: **end for**

図 2 のように，ランダムに入力された z に対して Generator は画像を生成する．その出力画像 G(z) とコンテキスト画像 C を用いて類似度計算を行い，S(G(z),C) を求める．これを誤差関数として誤差逆伝搬法を用いることでハイパーパラメータを修正していく．

w, b は学習の終わった Generator で得られた値を用いて，定数として固定する．言い換えれば，Generator の学習時，定数としていた z と学習対象であった w, b の役割を入れ替えることになる．変数は z のみであるが，コンテキストとの比較を行うのが S であるため，実際は Generator の出力画像である G(z) とコンテキスト C を用いて演算する．

G(z) と C を比較する関数 S を適切に定めることで S の z による偏微分値を用いて z を適切に変化させていくことができる．

G(z) と C の比較を行い，誤差逆伝搬をすると z の値が変化するため，当然 G(z) の出力画像も変化する．そのため，G(z) に依存しない誤差関数 S を選ぶ必要がある．本研究において S は二画像間の全画素を RGB で総比較してその差の二乗和を計算する関数としている．

以下では z のみを入力とする DCGAN と，テキストも入力として受け付ける Reed らの提案する text-conditional GAN の場合における z の調整手法について説明する．

3.1.1　DCGAN におけるハイパーパラメータの自動調整

Algorithm1 はハイパーパラメータ z の自動調整のアルゴリズムを示す．1 行目から 8 行目までは画像生成を行うニューラルネットワークの学習を行う．9 行目から 12 行目までは z の更

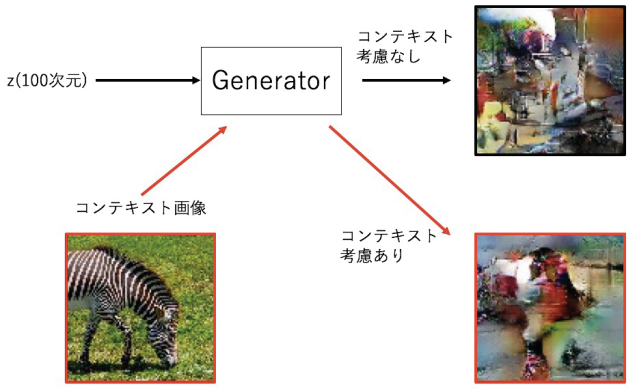

図 3　コンテキストの考慮の有無による出力画像比較 (入力テキストなし)

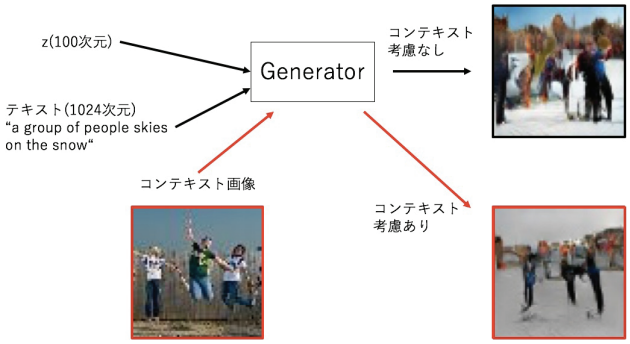

図 4　コンテキストの考慮の有無による出力画像比較 (入力テキストあり)

新手法である．

はじめは z を正規分布に従った確率分布にしたがってランダムに生成し，1 行目から 8 行目までで学習した Generator に入力して画像生成を行う．これにより生成された G(z) とコンテキスト画像 C との類似度関数を求める．その式を z で偏微分し誤差逆伝搬法を用いて z を更新する．

本研究では，Microsoft の提供しているデータセット MS COCO [8] を用いて Generator を学習する．MS COCO は画像に写るオブジェクトの発見方法などのルールを定めて，キャプションをつける過程をクラウドワーカーに分担させて行うことで，画像に適切なキャプションがつけられたデータセットである．図 3, 図 4 はコンテキストを考慮して z を調整する場合としない場合の結果を示している．どちらの場合も意味のある画像を生成できなかったが，コンテキストを考慮していないものに比べて，考慮した画像はコンテキスト画像のオブジェクトが現れているのが分かる．

3.1.2　Text-conditional GAN におけるハイパーパラメータの自動調整

Algorithm1 の 1 行目から 8 行目の間で用いるアルゴリズムを Reed らの提案しているモデルに置き換えて，テキストから画像を生成する場合のハイパーパラメータの調整手法について述べる．

図 3 では Generator の入力になっている z の値を全て，コンテキスト画像

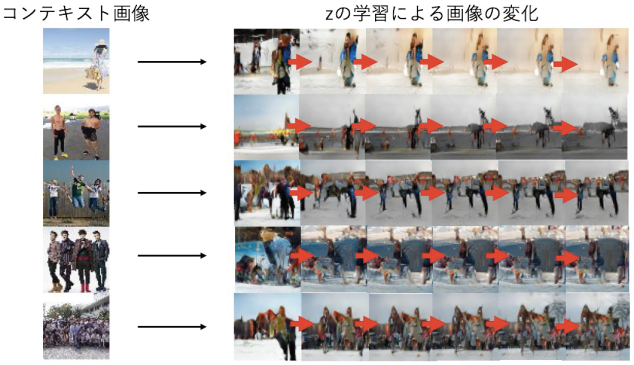

図 5　z の自動調整を行った出力画像

ンテキストに画像そのものが近づくよう変化させているので，Generator が理論的に全ての画像を表現可能であれば全く同じ画像が出力される．しかし，今回は z にテキストを合わせて入力としているので z を変化させるだけでは全く同じ画像を作ることにはならない．そこでこの手法は入力値が調整可能な変数と，出力画像の特徴を記述した変化しない定数を合わせたものである時に用いれば有効となることが分かる．

図 4 は MS COCO [8] をデータセットにして Reed らの提案したモデルを学習させた Generator に”a group of people skies on the snow” というテキストを入力して生成された画像を示している．図 3 と同様，コンテキストを考慮したハイパーパラメータの調整の有る場合と無い場合の二つの出力結果を載せている．提案手法はコンテキストと同様に 3 人グループを含む画像を生成していることが分かる．

図 5 はコンテキスト画像との全画素比較を行った時に出力画像がどのような変化を行うかを示したものである．出力となっている画像はコンテキストに近づいたものになっている．コンテキスト画像が 1 人，3 人のものなどはコンテキストに近づいているのがはっきり分かる．GAN はデータセットと同じではないがデータセットに近い画像を出力する仕組みであるため，z を修正してもデータセットに存在しない画像は生成できないので，z の自動調整をするにはできるだけ多くの画像を学習させる必要がある．また全体として画像が暗くなる傾向があり，今後，これについて検討する予定である．またこの手法ではテキストやコンテキストごとに適切な z を学習するため時間がかかってしまう．そのため，z を効率よく高速に調整できる手法が今後の課題となる．

3.2　Context-aware GAN

図 6 のように Reed らのモデルを拡張して，コンテキストを入力に含められるような Generator と Discriminator を学習して画像を生成する context-aware GAN モデルを提案する．Generator は Reed らのモデルと比べて入力にコンテキストも受け付けるようにする．Discriminator は入力にコンテキストも加えている．そして，出力が真（1）である条件には，出力が本物かつテキスト情報に即しているほかに，コンテキストとの関連を加えている．

context-aware GAN は以下のような式で定式化できる．

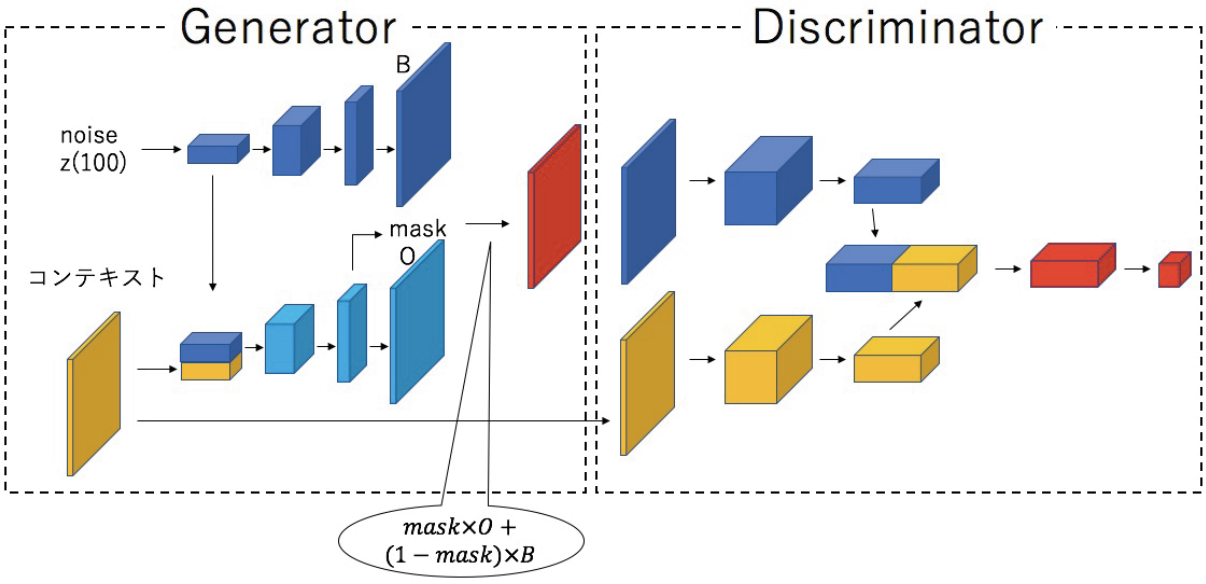

図 6　Context-aware GAN

$$G : \mathbb{R}^Z \times \mathbb{R}^C \to \mathbb{R}^O \tag{3}$$

$$D : \mathbb{R}^O \times \mathbb{R}^C \to 0, 1 \tag{4}$$

ここで C はコンテキスト画像の次元数であり，O は出力画像の次元数である．

$$z \in \mathbb{R}^Z \sim N(0, 1) \tag{5}$$

また z は上式のように定義されたハイパーパラメータである．

Generator 内では画像生成を行うことに加え，コンテキスト情報を画像に反映させることが必要である．Reed らの提案しているモデルの Generator の出力画像と同等のものが出力でき，かつ Generator の内部ではオブジェクト部分と背景部分に分かれるようにする．そのオブジェクト部分の出力画像に対し，コンテキストのテキスト情報を加えて演算することでオブジェクトがよりコンテキストに沿ったものになるようにする．

オブジェクトと背景を分離する手法は Vondrick らの提案しているモデル [7] を利用する．本研究で提案している手法ではオブジェクトと背景情報を分離する必要はないが，発展的課題として，Context-aware GAN に対してハイパーパラメータの自動調整を行い，オブジェクトを適切なものへ変化させようとした場合，オブジェクトと背景を分離しなければ，図??のように，背景が暗くなってしまい，鮮明な画像を得られない可能性が高い．将来の拡張のため，オブジェクトと背景の分離モデルを導入している．

Discriminator も同じくコンテキスト情報を加えて，Generator の出力画像がコンテキストに沿っているかを判定する必要がある．ここでは正しくコンテキストになっているデータセットを用いて学習させなければならない．

学習のデータセットには Huang らが提案した手法 [6] で収集された複数のアルバムを用いる．これは写真からストーリーを作ることができるとクラウドワーカーが判定したもので作成されてはいるが，それぞれのアルバムには類似性がないと思われる画像のペアも含まれているので，それぞれにつけられたテキストを Jacarrd 法で比較し，テキスト間類似度が高いと思われるものを用いてデータセットとする．その他にアルバムを自作しモデル学習とテストのデータセットとしている．

Algorithm 2 Context-aware GAN training algorithm.

1: **for** number of training iteration **do**
2: *Sample minibatch of m noise samples* $\{z^{(1)}, ..., z^{(m)}\}$
3: *Sample minibatch of m examples* $\{x^{(1)}, ..., x^{(m)}\}$ *from data*
4: *distribution* $p_{data}(x)$
5: *Sample minibatch of m examples* $\{c^{(1)}, ..., c^{(m)}\}$ *from data*
6: *to distribution* $p_{data}(x)$
7: *Update the discriminator by ascending its stochastic gradient* :

$$\nabla\theta_d \frac{1}{m}\sum_{i=1}^{m}[logD(x^{(i)}, c^{(i)}) + \frac{1}{2}(log(1 - D(G(z^{(i)}, c^{(i)}), c^{(i)}))$$

$$+log(1 - D(x^{(i)}, c^{(j)})))](i \neq j)$$

8: *Sample minibatch of m noise samples* $\{z^{(1)}, ..., z^{(m)}\}$
9: *Sample minibatch of m examples* $\{c^{(1)}, ..., c^{(m)}\}$ *from data*
10: *Update the generator by descending its stochastic gradient* :

$$\nabla\theta_g \frac{1}{m}\sum_{i=1}^{m}[log(1 - D(G(z^{(i)}, c^{(i)}), c^{(i)}))](i \neq j)$$

11: **end for**

今回はアルバム内でテキスト間類似度の高かったものを選んで学習データを作成している．つまりそれぞれのペアには二つの画像と，それぞれに対応するテキストが含まれている．これをさらに一方の画像が生成すべき画像でもう一方がコンテキストとなる二パターンに分けている．これを用いて学習を行う．

Algorithm2 は提案している Context-aware GAN のアルゴリズムを示している．Generator と Discriminator を交互学習させる時，まず Discriminator の学習から行う．

2 行目から 7 行目までが Discriminator の学習部分である．7 行目の式が大きくなるように Discriminator の w, b を学習させていく．$logD(x^{(i)}, c^{(i)})$ は Discriminator に学習用のデータ

図 7　context-aware GAN 学習過程に生成された画像

セットと適切なコンテキストを入力とした時の Discrminator の判定値なので大きくなるように学習する．$D(G(z^{(i)}, c^{(i)}), c^{(i)})$ は Generator から生成された画像とそれに対応するコンテキスト画像が入力とされた Discrminator の判定値，$D(x^{(i)}, c^{(j)})$ は Discriminator への入力画像は本物であるがコンテキスト画像が適切でなかった場合の Discrminator の判定値であるため小さくなるように学習する．アルゴリズムでは，1 から引いた数を大きくするように学習を行う．よって Discriminator の学習式は 7 行目のようになる．

8 行目から 10 行目が Generator の学習方法である．10 行目の式が小さくなるように Generator の w, b を学習させていく．$D(G(z^{(i)}, c^{(i)}), c^{(i)})$ は Generator の出力画像を Discriminator へ入力した時の値であり，Disciminator を騙せるようにこの値を大きくする必要があるので 1 から引いた値を大きくするように学習させていく．

Reed らの提案している手法では 100 次元の z，入力テキストから生成された 128 次元のベクトルから構成される 228 次元のベクトルを，Generator に代入して画像を生成している．これに対して，本研究で提案している手法ではテキストでなくコンテキスト画像を入力している．画像を CNN を用いて，[4,4,1568] の行列へ変換したものに z を加える．この時 z が 100 次元で [4,4,100] の行列とすると，入力画像に比べ z の影響が小さくなり，出力画像がコンテキスト画像に依存して，バリエーションが少なくなってしまう可能性が高い．

図 7 はスキーをしている人の画像を生成するため，学習データセットを用意し，コンテキストにはスキーをしている人数と同じ人が映った画像を用いて学習させたものの，学習途中で生成された画像である．スキー画像の学習はされているが，同じ画像が多く生成されているのがわかる．これは入力になっているコンテキスト画像の値が z に比べて大きな割合を占めているために z の影響が小さくなり，コンテキストの入力に対して出力が一つに収束してしまっているからである．

入力されたコンテキスト画像に対して適切な画像生成を行うことはできているが，画像に変化をもたせられない場合は，学習データの一つに解が近づいているだけなので，生成できる画像の種類が限られてしまう．多様な画像を生成するため，入力となる z のベクトルの次元数を増やして，たとえば，1000 次元

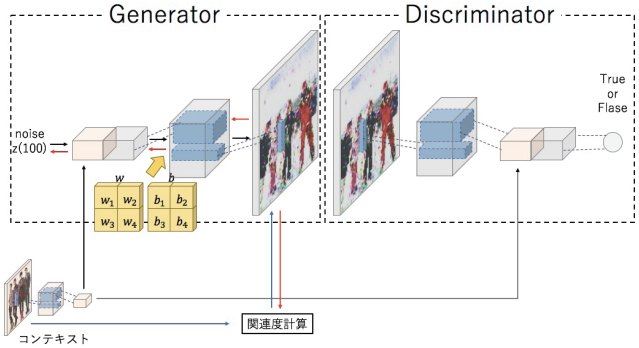

図 8　Context-aware GAN の構成

図 9　ベースラインとなる Reed らの手法でのスキーの出力画像

とし，入力されたコンテキスト画像のベクトルに加える時に占める割合が半分近くになるようにすれば，より出力画像をランダムにして学習することができて，変化を持った画像を生成することができる．

図 8 は Context-aware GAN の全体図である．GAN を拡張してコンテキスト画像を入力できるようにする．またハイパーパラメータの調整過程も書いてある．Discriminator には入力画像とコンテキスト画像の関係性があるかどうかを判断できるよう学習していくため，Generator は入力されたコンテキスト画像に対して関係性がある画像を生成できればいいため，z が小さかった場合，z を完全に無視して Discriminator に通る一枚の画像を生成し続けるように学習してしまう．そのため z を大きくすることが必要である．またそれにより，本研究で提案されている z の自動調整手法を適用することができるようになると考えられる．今後，この点について検討していく予定である．

4.　実験と考察

4.1　実 験 概 要

提案手法を評価するため，Reed らの提案しているモデルと，それに対して z の自動調整を行ったもの，そして Context-aware GAN を用いたものの三つで学習を行い，その出力結果を比較する．今回はスキーの画像生成を行う．コンテキストを考慮した出力結果がどのように変わるかを評価するため，ベースラインとして図 9 のような Reed らの提案するモデルによる出力画像を用いる．ベースライン画像と五種類のコンテキスト画像を用いて，z の自動調整を行った画像と Context-aware GAN の出力画像を比較することでコンテキストを考慮することの有用性を評価する．

4.2　データセット

モデルの学習には公開データセット MS COCO [8] やデータ

図 10　比 較 画 像

セットには Huang らの提案している手法 [6] により生成された画像とテキストのペアの系列となっているアルバムなどを用いる．Context-aware GAN の学習について，Huang らの提供しているアルバムのペアには画像間に明らかに関連がないと思われるものがあり，ニューラルネットワークが収束しなかったために，一時的に自ら集めたスキーの画像と，それに対して関連があると思われる画像の組み合わせをペアとし，データセットとして用いた．

4.3　実 験 結 果

図 10 は比較画像を列挙したものである．B1, B2, B3 は Reed らのモデルに MS COCO を学習させたものにテキスト : “a group of people skies on the snow.“を入力した時のランダムな出力画像 3 つである．C1, C2, C3, C4, C5 はコンテキスト画像として提案している二つの手法の入力に用いる画像である．この五つの画像を用いて Reed らのモデルのハイパーパラメータ z を自動調整した際の出力画像が Z1〜Z5 であり，それぞれコンテキストに用いた画像が Z1 は C1, Z2 は C2, Z3 は C3, Z4 は C4, Z5 は C5 である．また Context-aware GAN の出力画像は G1〜G5 であり．それぞれコンテキストに用いた画像が G1 は C1, G2 は C2, G3 は C3, G4 は C4, G5 は C5 である．

4.3.1　文脈を考慮したハイパーパラメータの自動調整

Z1 や Z3 はコンテキスト画像と同じ人数の画像になっている．しかし，全体的に出力されている画像は白っぽいためにスキーと見ることができるが鮮明さに欠けている．コンテキスト画像に全画素比較でハイパーパラメータを修正していくと人のいる位置が全体的に近づいていく．そのため C5 のような，写っている人の多い画像をコンテキストとして用いると全体的な形は似るものの，人間には分かりにくい画像が生成された．

4.3.2　Context-aware GAN

Context-aware GAN を学習させた際にはスキーをしている画像と，全く無関係な場面の人の映っている画像を，映っている人の人数が同じになっていることを条件にペアを作って入力データセットとして学習させた．そのため，コンテキストの条件は人数が一致することとなっている．出力画像は z の自動調整に比べて鮮明な画像となっている．

表 2　図 10 の比較結果

文脈	出力画像	類似度	文脈	出力画像	類似度
C1	baseline1	9.55172e+07	C3	G3	1.10899e+08
C1	baseline2	1.15396e+08	C3	Z3	7.02971e+07
C1	baseline3	4.69348e+07	C4	baseline1	1.46336e+08
C1	G1	8.37656e+07	C4	baseline2	1.29559e+08
C1	Z1	2.41218e+07	C4	baseline3	1.65401e+08
C2	baseline1	1.28037e+08	C4	G4	1.47822e+08
C2	baseline2	9.24582e+07	C4	Z4	7.63125e+07
C2	baseline3	1.22453e+08	C5	baseline1	1.12958e+08
C2	G2	1.43344e+08	C5	baseline2	8.79883e+07
C2	Z2	2.25361e+07	C5	baseline3	1.19363e+08
C3	baseline1	1.24424e+08	C5	G5	7.53389e+07
C3	baseline2	1.08328e+08	C5	Z5	3.38725e+07
C3	baseline3	1.3485e+08			

表 3　図 10 の比較平均

用いた手法	正規化を行った平均値
既存手法	0.71
提案手法 1	0.88
提案手法 2	0.72

4.3.3　画像の類似度比較

それぞれをコンテキスト画像との全画素の比較を行った．

表 2 はコンテキスト画像とそれぞれの出力画像の全画素比較の出力結果である．コンテキスト画像と出力画像は図 10 の画像をそれぞれ参照している．全画素比較の類似度は次のように算出している．出力画像とコンテキスト画像を [64,64,3] の行列に RGB で読み込んで変換する．形の同じ行列なので，それぞれの値の差分を取り，その全ての要素の平方和を取る．出力画像を O, コンテキスト画像を C として式 (7) のように求めている．

$$\sum_{i=1}^{64}\sum_{j=1}^{64}\sum_{k=1}^{3}(O[i][j][k]-C[i][j][k])^2 \qquad (6)$$

表 3 は値を正規化して平均を出した値を表示している．関連度が大きくなるように学習させている提案手法 1 の値は他に比べて関連度が高くなっている．しかし提案手法 2 は既存手法に比べて値に大きな違いはない．これは関連度の計算方法として用いている手法が最適ではないためであり，関連度の計算手法について改良を行うのは今後の課題である．

4.4　考　　察

出力された画像はスキーをしているように見える画像となっているが，多様な画像となっていて，それぞれ実用的に用いられる場面は異なることがわかる．これらの画像の違いを生み出しているのは z の違いであり，z を適切な値に変化させることでコンテキストに沿った画像を出力することができる．

表 2 を見ると直接画素値を近づけている z の比較は値が画素値が，baseline の画像と比較して近いものになっている．Context-aware GAN を用いて作られた画像はこの比較方法では baseline の画像との差はあまりない．しかし，画像を直感的

に比較すると，Context-aware GAN により生成された画像は baseline のものと比べるとコンテキストを踏まえていることがはっきり分かる．これより全画素を単純に全比較するだけではコンテキストを踏まえているかどうかの判定は完全にはできないことが分かる．今後，これらのコンテキストを踏まえているかどうかをより正確に判定するニューラルネットワークを構築して比較手法に取り入れることができれば，コンテキストを考慮した z の自動調整の精度を高くすることができる．またその比較方法を用いて Context-aware GAN がコンテキストを踏まえたものになっているかの定量評価を行う実験も今後の課題としておく．

今回の例の場合は，学習データにないコンテキスト画像を入力しても出力画像はコンテキスト画像に写っている人と近い人数の人が写っている画像を出力している．画像が Reed らの提案しているモデルに比べて鮮明になっているのは，ニューラルネットワークの画像データの占める割合が z の占める割合に比べ大きいためである．z の影響が小さくなると，生成される画像にランダム性が生じなくなるため，出力画像が特定のものに収束していくため早い段階で綺麗な画像が出力されることになる．しかし z の占める割合を大きくすると，使用メモリが多くなることが分かった．また，Context-aware GAN のメモリ使用量が大きいため，さらにデータ次元の圧縮などの対策が必要と考える．今後，これらの課題について取り込みたい．

5. ま と め

コンテキストを考慮した画像生成を行うために GAN の入力となっているハイパーパラメータの自動調整と，GAN を拡張した Context-aware GAN モデルを提案した．MS COCO や Huang らの提供しているデータセットなどを用いて出力結果について従来手法と比較を行った．本研究における提案手法は従来手法よりもコンテキストとの関連の強い画像を生成することが可能であることが分かった．

しかし，ハイパーパラメータを自動調整することで，画像全体が暗くなる可能性が高い．また，Context-aware GAN のメモリ使用量が大きいため，効率的な実現手法が必要である．実験で分かった課題は以下に列挙し，今後これらについて検討する予定である．

- z の自動調整によって画像が暗くなるのを防ぐこと．
- 現在，コンテキストの要素としたのはコンテキスト写真の人数だけであり，これでは不十分であるので，より様々な要素 (天候や季節，性別など) を反映できるような Contect-aware GAN を作ること．
- 今は画像とコンテキストのペアを入力として与えているので，これを時系列を考慮した画像生成が行えるように改良すること．
- Contect-aware GAN の出力結果がメモリ不足のためにバリエーションを持っていないので，z の与える影響を大きくし出力画像に変化を持たせること．
- Contect-aware GAN の出力結果がメモリ不足のためにコンテキストを考慮しないパターンを反映できなかったので，学習を行えるようにすること．
- Contect-aware GAN の出力結果に対してハイパーパラメータの自動調整を行うことでよりコンテキストに沿った画像が出力できるようにすること．

今後，さらに大規模な実験を行って，提案手法の定量評価を行う予定である．また，ハイパーパラメータの調整手法を Context-aware GAN に適用してモデルを進化させる予定である．さらに，Context-aware GAN の改善として，コンテキストデータに存在している時空間連続性を考慮する点について検討していきたい．

6. 謝　　辞

本研究の一部は，科研費（課題番号 16K12532) による．

文　　献

[1] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pp. 487–495, 2014.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[5] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[6] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. *CoRR*, Vol. abs/1604.03968, , 2016.

[7] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pp. 613–621, 2016.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.