条件付き画像キャプショニングに向けた敵対的生成ネットワークの検討

Considering Generative Adversarial Network toward Conditional Image Captioning

阿部 佑樹[1*] 松森 匠哉[1] 妹尾 卓磨[1] 今井 倫太[1]
Yuki Abe[1] Shoya Matsumori[1] Takuma Seno[1] Michita Imai[1]

[1] 慶應義塾大学 理工学部
[1] Department of Science and Engineering, Keio University

Abstract: One of the challenges of image captioning is selectively generating captions using latent variables, where existing approaches have tackled by using a set of known factors and learning with an annotated dataset. On the other hand, in order to leverage potential latent variables on datasets, a mechanism that automatically learns them is required. In this research, we propose a framework based on generative adversarial networks toward conditional image captioning, a task generating captions conditioned with images and latent variables. In experiments, we demonstrated that our proposed model can learn and leverage latent variables on the image classification with several ground truth labels.

1 はじめに

画像に対する適切な説明文（キャプション）を自動で生成することは画像キャプショニングと呼ばれる．画像中の注目箇所や文章表現といったキャプションを特徴付ける要素（潜在変数）の違いにより，ひとつの画像に対するキャプションは複数通り存在し得る．

これまでに，潜在変数の値に応じて生成するキャプションを選択的に変化させる，条件付き画像キャプショニングの手法が提案されてきている [1]．既存手法は潜在変数がアノテーションされているデータセットを利用する教師あり学習を採用している．しかしながら，世の中に存在する多くの画像キャプションのデータセットは，潜在変数が未知でありアノテーションされていないことが多い．一般的な画像キャプションのデータセットを用いて，データセットに暗黙的に存在する潜在変数を利用した条件付き画像キャプショニングを実現するためには，潜在変数の教師なし学習を行う仕組みが必要である．

本研究は，潜在変数の教師なし学習を採用した条件付き画像キャプショニングに向けた，深層生成モデルのアーキテクチャの検討を行う．本提案モデルはテキスト生成モデル LaTextGAN[2] の拡張として表現される．画像キャプショニングを画像で条件づけたテキスト生成として扱うことにより，conditional GAN[3] を参考に LaTextGAN を画像で条件付けする．また，InfoGAN[4] を参考に相互情報量に関する制約項を目的関数に加えることで，キャプションの潜在変数をアノテーションのないデータセットから獲得する．

キャプションの特徴の種類や区分は曖昧であるため，キャプションの潜在変数を獲得し利用できていることの定量的評価や定性的評価は困難である．本提案モデルの評価には画像キャプショニングを直接扱わず，ひとつの画像に対して複数の正解ラベルが存在する画像分類問題を，本提案モデルの有用性を示す検証課題として扱う．

*慶應義塾大学理工学部情報工学科，〒 223-8522 神奈川県横浜市港北区日吉 3-14-1，E-mail: abe@ailab.ics.keio.ac.jp

2 関連研究

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GAN)[5] は深層生成モデルの学習フレームワークのひとつである．GAN は対立する 2 つのネットワークを利用する．生成ネットワーク G はノイズ変数 $\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})$ をデータ $G(\boldsymbol{z})$ に変換する．$p_{\boldsymbol{z}}(\boldsymbol{z})$ は一般的に一様分布や標準正規分布が用いられる．識別ネットワーク D はデータが訓練データ $\boldsymbol{x} \sim p_{data}(\boldsymbol{x})$ か G によって生成されたデータ $G(\boldsymbol{z})$ かを識別する．G と D の学習は次式に示すミニマックスゲームによって行う．

$$\min_G \max_D \mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \quad (1)$$

G は D の識別精度を下げるデータ $G(\boldsymbol{z})$ を生成するよう学習が行われる．式 1 の最適化により訓練データに似たデータを生成する生成ネットワーク G を得ることができる．

2.2 LaTextGAN

LaTextGAN は GAN をテキスト生成に応用した深層生成モデルである．LaTextGAN は Encoder-Decoder モジュールと GAN モジュールの 2 つのモジュールを利用する．Encoder-Decoder モジュールはテキストとテキストを表現する文章ベクトルの相互変換に利用される．Encoder-Decoder モジュールはオートエンコーダ [6] で構成され，学習はテキストの再構成誤差の最小化により行われる．GAN モジュールは文章ベクトルの生成に利用される．GAN モジュールの学習は式 1 で示される通常の GAN の目的関数の最適化で行われる．訓練データにはテキストを直接用いるのではなく，事前学習済みの Encoder-Decoder モジュールから得られる文章ベクトルが用いられる．最終的な LaTextGAN の出力は，GAN モジュールの出力する文章ベクトルを Encoder-Decoder モジュールによりテキストに復号することで得る．

2.3 conditional GAN

conditional GAN は GAN を条件付き生成モデルに拡張したモデルである．ここで X を任意の種類の補足情報とする．補足情報には例えばクラスラベルやモダリティの異なる他のデータなどが挙げられる．条件付けは G および D の両方に追加の入力として X を与えることで行われる．

2.4 InfoGAN

InfoGAN は GAN を潜在変数の教師なし学習に応用した深層生成モデルである．潜在変数をマッピングさせる変数を $\boldsymbol{c} \sim p_{\boldsymbol{c}}(\boldsymbol{c})$ とする．$p_{\boldsymbol{c}}(\boldsymbol{c})$ は任意の分布を用いる．G はノイズ変数 $\boldsymbol{z}$ と変数 $\boldsymbol{c}$ をデータ $G(\boldsymbol{z}, \boldsymbol{c})$ に変換する．InfoGAN の目的関数は，式 1 で示される通常の GAN の目的関数に，変数 $\boldsymbol{c}$ を再構成するネットワーク $Q(G(\boldsymbol{z}, \boldsymbol{c}))$ を用いた相互情報量制約項 $\mathcal{L}_I$ を加えた形で表現される．

$$\min_{G,Q} \max_D \mathcal{L}_{info} = \mathcal{L}_{adv}(G, D) + \lambda_I \mathcal{L}_I(G, Q) \quad (2)$$

$$\mathcal{L}_I = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}), \boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[-\log Q(G(\boldsymbol{z}, \boldsymbol{c}))] \quad (3)$$

ここで λ_I はハイパーパラメータである．制約項 $\mathcal{L}_I$ の最小化は変数 $\boldsymbol{c}$ とデータ $G(\boldsymbol{z}, \boldsymbol{c})$ の間の相互情報量の変分下界の最大化に等しい．別の言い方をすると，制約項 $\mathcal{L}_I$ は変数 $\boldsymbol{c}$ に訓練データを特徴付ける表現を対応付けることを促進する．式 2 の最適化により，潜在変数が変数 $\boldsymbol{c}$ にマッピングされると同時に，潜在変数の値に応じてデータを生成する生成ネットワーク G を得ることができる．

3 提案

本提案モデルは LaTextGAN を元に Encoder-Decode モジュールと GAN モジュールから構成され，GAN モジュールの画像キャプショニングおよび潜在変数の教師なし学習への拡張で表現される．

本提案モデルの GAN モジュールの概要図を図 1 に示す．GAN モジュールは生成ネットワーク G，識別ネットワーク D，および復号ネットワーク Q から構成される．全てのネットワークは補足情報として画像 X を入力に受ける．G は画像 X と変数 $\boldsymbol{c}$ から文章ベクトル $\hat{\boldsymbol{s}}$ を生成する．D は画像 X に対する文章ベクトルが訓練データ $\boldsymbol{s}$ か生成データ $\hat{\boldsymbol{s}}$ かを識別する．Q は画像 X と文章ベクトル $\hat{\boldsymbol{s}}$ を入力に受け，G に入力された変数 $\boldsymbol{c}$ を予測する．目的関数を次に示す．

$$\min_{G,Q} \max_D \mathcal{L} = \mathcal{L}_{adv}(G, D) + \lambda_I \mathcal{L}_I(G, Q) \quad (4)$$

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{s}\sim p_{data}(\boldsymbol{s})}[\log D(\boldsymbol{s}, X)] + \mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[\log(1 - D(G(\boldsymbol{c}, X), X))] \quad (5)$$

$$\mathcal{L}_I = \mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[\|\boldsymbol{c} - Q(G(\boldsymbol{c}, X), X)\|^2] \quad (6)$$

ここで，式 6 の最小化は，Q の出力がガウス分布に従うと仮定した場合の式 3 の最小化と同一である．画像キャプショニングを画像で条件づけられたテキスト生成として扱うことにより，G は画像に対する適切な文章を生成するモデル，すなわち画像キャプションを生成するモデルとして学習が行われる．また，相互情報量制約 $\mathcal{L}_I$ により，文章ベクトルの潜在変数の獲得が促される．

本提案モデルの最終的な出力は，学習済みの生成ネットワーク G を用い画像 X と変数 $\boldsymbol{c}$ から文章ベクトル $\hat{\boldsymbol{s}}$ を生成し，学習済みの Encoder-Decoder モジュールを用いて生成された文章ベクトル $\hat{\boldsymbol{s}}$ を有効なテキストに復号することによって得る．

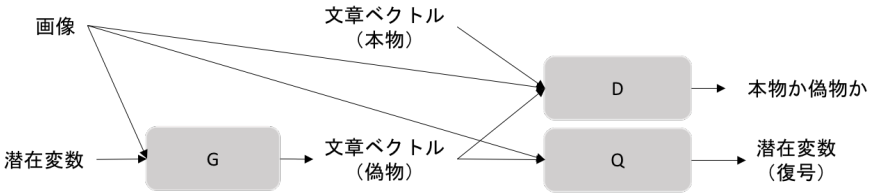
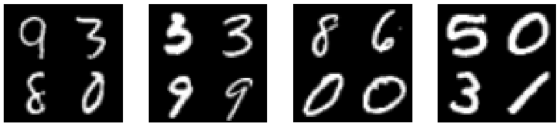


図 1: 本提案モデルの GAN モジュールの概要図.



[9,3,8,0] [3,3,9,9] [8,6,0,0] [5,0,3,1]

図 2: GridMNIST のサンプルの一例.

4 実験

本提案モデルは条件付き画像キャプショニングに向けたものであるが，評価の明確化と簡単化のため，検証課題として複数の正解ラベルが存在する画像分類問題を用いる．潜在変数の値に応じて生成するキャプションの特徴を変化させることは，画像分類問題において潜在変数の値に応じた特定のラベルを出力することと考える．ひとつの潜在変数の値に対して複数の画像を用いたときの本提案モデルの出力を評価することで，潜在変数として獲得された表現を明らかにし，同時に潜在変数を利用した条件付き画像キャプショニングが可能であることを示す.

4.1 データセット

ひとつの画像に対して複数の正解ラベルをもつ画像分類問題として，MNIST を元にした GridMNIST を作成した．図 2 は GridMNIST のサンプルの一例である．GridMNIST は 64×64 の白黒画像と 4 つの正解ラベルを持った正解ラベル集合のペアから構成される．画像は 2×2 の格子状に配置された MNIST であり，正解ラベルの集合の要素は画像中に存在する MNIST の正解ラベルである．MNIST の学習データとテストデータを元に，GridMNIST の学習データを 10000 件，テストデータを 1000 件作成した.

4.2 実験方法

ランダムに決定されたひとつの潜在変数の値 $\boldsymbol{c}_n \sim p_{\boldsymbol{c}}(\boldsymbol{c})$ に対して，入力画像として GridMNIST のテストデータ 1000 件全てを用い，本提案モデルの分類精度および出力ラベルの出現頻度比を評価する．変数 $\boldsymbol{c}$ の分布 $p_{\boldsymbol{c}}(\boldsymbol{c})$ には 3 次元の標準正規分布を用いる．分類精度は，ある画像に対して出力したラベルが正解ラベル集合に含まれる場合を正解として計算される．分類精度は，画像に対して適切な出力を得ることができるかを評価する指標であり，本提案モデルのキャプショニングの性能を調べるための代替として有用である．出力ラベルの出現頻度比は，ひとつの潜在変数の値に対する出力ラベルの傾向を評価する指標であり，潜在変数として獲得された表現を明らかにするために有用である．出力ラベルの出現頻度比の比較手法には，潜在変数を画像ごとに全てランダムに決定する場合を用いる.

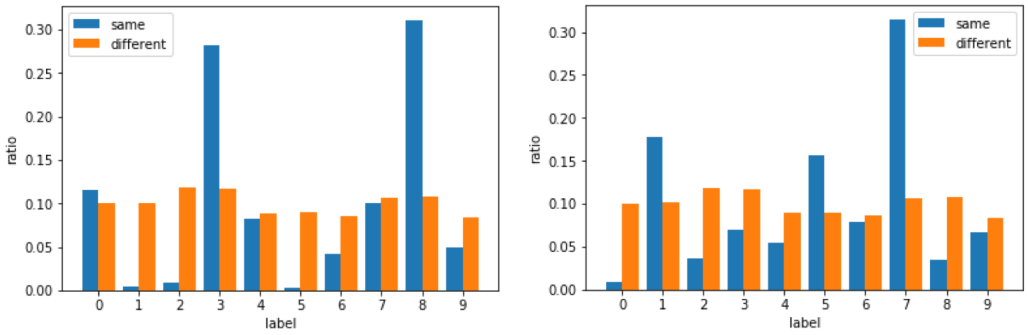
4.3 実験結果

図 3 に実験結果を示す．分類精度はいずれも 98%を超えており，本提案モデルが画像に対して適切なラベルを出力していることを示している．潜在変数を固定した場合にはラベルの出現頻度に偏りが見られる．例えば，潜在変数の値 $\boldsymbol{c}_1$ を用いた場合はラベル 3，8 を多く出力しており，潜在変数の値 $\boldsymbol{c}_2$ を用いた場合はラベル 1，7 を多く出力している．一方で，潜在変数をランダムに決定した場合は全てのラベルを概ね均等に出力している.

分類精度を維持しながら特定のラベルを多く出力できるため，潜在変数はラベルのサブグループに対応すると考えられる．例えば，潜在変数の値 $\boldsymbol{c}_1$ はラベル 3，8 を要素としたサブグループを表現し，入力画像中に数字 3 か数字 8 があれば優先的に対応するラベルを生成ネットワーク G から出力させる．潜在変数の値 $\boldsymbol{c}_1$ はラベル 3 と 8 のサブグループ，潜在変数の値 $\boldsymbol{c}_2$ はラベル 1 と 7 のサブグループと，数字の形が似ているものはサブグループ化されやすいと考えられる.

5 将来研究

本研究では本提案モデルの有用性の検証に画像分類問題を用いた．画像キャプショニングは自然画像を用



(1) $\boldsymbol{c}_1$, acc: 98.7%　　(2) $\boldsymbol{c}_2$, acc: 98.0%

図 3: 入力画像 1000 件に対する出力ラベルの出現頻度比．横軸はラベルであり，縦軸は出現頻度比である．潜在変数の値を固定したものが same，潜在変数の値がランダムなものが different である．図 (1)(2) では same は異なる潜在変数の値を用いている．acc は same に対する本提案モデルの分類精度を示す．

いた問題であるのに対し，今回扱った画像は自然画像ではない．画像キャプショニングに向け，自然画像を用いた画像分類問題において本提案モデルの有用性を検証することは直近の課題のひとつである．

潜在変数の教師なし学習のため，InfoGAN と同様に相互情報量制約項を目的関数に加える方法を用いた．本提案モデルの挙動の分析により潜在変数の表現を検証したが，潜在変数の各次元や値が表現とどう対応付いているかは不明である．モデルの挙動の解釈性の観点から，人にとって解釈しやすい潜在変数を獲得することは重要である．潜在変数の応用についても今後の課題として挙げられる．本提案モデルは画像とキャプションから潜在変数を予測する復号ネットワーク Q の学習も行なっているため，画像を用いた対話システムで潜在変数の予測と利用を応用し，一貫したコンテキストで対話を行うシステムを構築することが考えられる．

6　まとめ

本研究では，潜在変数の教師なし学習を採用した条件付き画像キャプショニングに向けた，深層生成モデルのアーキテクチャの検討を行った．画像キャプショニングを画像で条件づけたテキスト生成として扱うことで，本提案モデルをテキスト生成モデルの拡張として構築した．本提案モデルの有用性を示す検証課題として画像分類問題を扱い，本提案モデルが潜在変数としてクラスラベルのサブグループを獲得し，生成するラベルを選択的に変化させることが可能であることを示した．将来研究として，画像キャプショニングへ向けた本提案モデルの拡張や，潜在変数の解釈性の向上および応用が挙げられる．

参考文献

[1] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pp. 3574–3580, 2016.

[2] David Donahue and Anna Rumshisky. Adversarial text generation without reinforcement learning. *arXiv preprint arXiv:1810.06640*, 2018.

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

[5] 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.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 8, pp. 1798–1828, 2013.