The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

画像キャプションに対する表現学習に向けた敵対的生成ネットワーク

Generative Adversarial Networks
toward Representation Learning for Image Captions

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

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

Captions generated from a single image are possibly different from each others as for representations (e.g. attention points or sentence expressions). However, a vast amount of image captioning datasets in the world have few or no annotations of latent variables. Learning latent variables of captions with no supervision is an important from perspectives of scalability and interpretability of conditional image captioning models. In this research, we propose a deep generative model to learn and leverage latent variables of image captions. In experiments, we used the task of image classification with several MNIST images and ground truth labels as down-scaled setting of image captioning, and we show that our proposed model acquired latent variables which represent sub-groups of 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．ひとつの画像に対するキャプションは，注目箇所の偏りや文章表現の違いにより複数通り存在することが考えられる．別の言い方をすると，キャプションは，画像および画像と独立したキャプションを特徴付ける要素（潜在変数）によって生成されると考えることができる．潜在変数の利用や獲得は，コンテキストの考慮や分析といった実用上の有用性から，画像キャプショニングにおける重要な課題のひとつとして位置付けられる．

これまでに，画像，キャプション，および潜在変数を用いた教師あり学習により，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手法が提案されてきている [1]．しかしながら，世の中に存在する画像キャプションのデータセットは，潜在変数がアノテーションされておらず未知であることが多い．

アノテーションのない訓練データから潜在変数を学習により獲得することは一般的に表現学習と呼ばれ，深層生成モデルを用いた手法が提案されてきている [2]．深層生成モデルを用いた画像キャプションに対する表現学習は，潜在変数による選択的な画像キャプショニングの拡張性の向上に繋がる．また，画像のみからキャプションを生成する方法 [3] と比較して，意味的な潜在変数を用いて生成するキャプションを制御できることは，モデルの解釈性の観点から重要である．

本研究は，画像キャプションに対する表現学習に向けた，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[4] 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捉えることにより，conditional GAN[5] を参考に LaTextGAN を画像で条件付けする．また，InfoGAN[2] 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，キャプションに対する潜在変数をアノテーションのない訓練データから獲得する．

キャプションの潜在変数の種類や区分は曖昧であるため，キャプションの特徴を捉えた潜在変数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，ある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．

連絡先: 阿部佑樹，慶應義塾大学理工学部，〒 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) は Goodfellow *et al.* [6] によって提案された深層生成モデルの学習フレームワークである．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．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．

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

2.2 LaTextGAN

LaTextGAN[4] 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 [7] 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 モジュールは通常の GAN の枠組みで学習が行われる．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により得る．

2.3 conditional GAN

conditional GAN[5] 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．

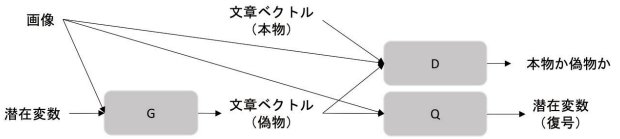


図 1: 本提案モデルの GAN モジュールの概要図.

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[2] は GAN を表現学習に応用した深層生成モデルである．潜在変数を $\boldsymbol{c} \sim p_{\boldsymbol{c}}(\boldsymbol{c})$ 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 $\boldsymbol{z}$ と潜在変数 $\boldsymbol{c}$ 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，潜在変数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．

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

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

ここで λ_I はハイパーパラメータである．制約項 $\mathcal{L}_I$ の最小化は潜在変数 $\boldsymbol{c}$ 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ため，訓練データを特徴付ける表現を潜在変数 $\boldsymbol{c}$ として獲得することが促される．

2.5 Variational AutoEncoder

Variational AutoEncoder (VAE)[8] はオートエンコーダの拡張モデルである．VAE のエンコーダは入力 $\boldsymbol{x}$ で条件づけられた事後分布 $q(\boldsymbol{z}|\boldsymbol{x})$ で表される．すなわち VAE は入力 $\boldsymbol{x}$ の表現 $\boldsymbol{z}$ を分布で表現するモデルである．VAE では $q(\boldsymbol{z}|\boldsymbol{x})$ が事前分布 $p(\boldsymbol{z})$ に近づくような制約が加えられる．VAE の目的関数は次式で表される．

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) \quad (4)$$

ここで θ は VAE のパラメータ，D_{KL} はカルバックライブラー情報量である．第 1 項が再構成誤差に対応し，第 2 項が事後分布に対する制約に対応する．

3. 提案

本提案モデルは LaTextGAN を元に，Encoder-Decoder モジュールでの VAE の利用，および GAN モジュールの画像キャプショニング及び表現学習への拡張で表現される．

3.1 Encoder-Decoder モジュール

本提案モデルはテキストと文章ベクトルの相互変換に VAE を用いる．しかしながら，式 4 で表させる通常の VAE の目的関数は事後分布に対する制約が強すぎるため，テキストの再構成誤差が大きくなり相互変換に適さない．そこで制約係数 $0 < \lambda_{KL} < 1$ を用いて VAE の学習を行う．目的関数を次に示す．

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q_\theta(\boldsymbol{s}|\boldsymbol{S})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{s})] + \lambda_{KL} D_{KL}(q_\theta(\boldsymbol{s}|\boldsymbol{S}) \| p(\boldsymbol{s})) \quad (5)$$

[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.

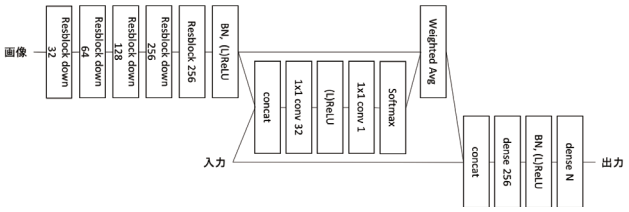


図 3: ネットワーク構成の概要図．BN はバッチ正規化，LReLU は LeakyReLU を示す.

ここで $\boldsymbol{S}$ はテキスト，$\boldsymbol{s}$ は文章ベクトルである．事後分布 $q_\theta(\boldsymbol{s}|\boldsymbol{S})$ には対角ガウス分布を用い，事前分布 $p(\boldsymbol{s})$ には標準正規分布を用いる．$\lambda_{KL} = 0$ のときオートエンコーダの目的関数と一致し，$\lambda_{KL} = 1$ のとき式 4 で示す VAE の目的関数と一致する．小さい λ_{KL} では再構成誤差の最小化が重要視される．

VAE の学習した文章ベクトル $\boldsymbol{s}$ を利用すると GAN モジュールの学習が安定する．文章ベクトル $\boldsymbol{s}$ は対角ガウス分布に従うノイズを含む．識別ネットワーク D の入力にノイズを加えることで GAN の学習を安定化する手法 [9] と同様に，訓練データの分布とモデル分布のサポートの重なりが増加する．また，生成ネットワーク G に対する近似誤差の緩和の効果が得られる．

3.2 GAN モジュール

図 1 に本提案モデルの GAN モジュールの概要図を示す．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，及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画像 X を入力に受ける．G は画像 X と潜在変数 $\boldsymbol{c}$ 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 に入力された潜在変数 $\boldsymbol{c}$ を予測する．目的関数を次に示す．

$$\min_{G,Q} \max_{D} \mathcal{L}(G, D, Q) = \mathcal{L}_{adv}(G, D) + \lambda_I \mathcal{L}_I(G, Q) \quad (6)$$

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{s}\sim q_\theta(\boldsymbol{s})}[\log D(\boldsymbol{s}, X)] + \mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[\log(1 - D(G(\boldsymbol{c}, X), X))] \quad (7)$$

$$\mathcal{L}_I = \mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[\|\boldsymbol{c} - Q(G(\boldsymbol{c}, X), X)\|^2] \quad (8)$$

ここで，式 8 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に対する潜在変数 $\boldsymbol{c}$ の獲得が促される．

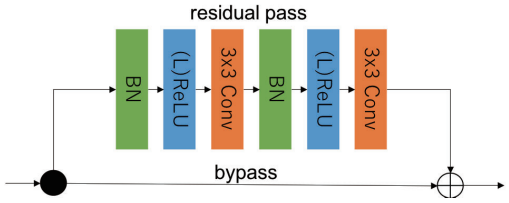


図 4: ResBlock の概要図．BN はバッチ正規化，LReLU は LeakyReLU を示す．ResBlock を用いてサイズを半分にする場合 (ResBlock down) は，bypass および residual pass の最後に平均プーリング層を挿入する．ResBlock を用いてチャネル数を変化させる場合は，bypass の最初に 1×1 の畳み込み層を追加し，この畳み込み層と residual pass の最初の畳み込み層でチャネル数を変化させる．画像に対して直接 ResBlock を適用する場合は，residual pass の最初の BN と (L)ReLU を取り除く．

4. 実験

潜在変数の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．画像キャプショニングにおいて，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対応する．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する．

4.1 データセット

複数の正解ラベルが存在する画像分類問題として，MNIST を元に GridMNIST を作成した．GridMNIST のサンプルの一例を図 2 に示す．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 画像であり，正解ラベル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 実装

Encoder-Decoder モジュールはクラスラベル L とクラスラベルを表現する潜在表現 $\boldsymbol{l}$ の相互変換に利用され，VAE を用いて構成される．クラスラベル L は 10 次元の Onehot ベクトルで表現し，潜在表現 $\boldsymbol{l}$ の次元数は 2 次元とした．VAE のエンコーダおよびデコーダは 1 層の全結合層を用いた．制約係数 $\lambda_{KL}=0.2$，バッチサイズ 64，エポック数 100 とし，最適化手法に Adam[10] を用いた．

GAN モジュールの G，D，および Q のネットワーク構成を図 3 に示す．各ネットワークの計算処理は画像 X と追加の入力からある値を出力するという点で同一であり，Kazemi *et al.* [11] の提案するモデルを参考にした．学習の安定化のため，ネットワークの構成要素には，図 4 に示す ResBlock[12]，バッチ正則化 [13]，G の活性化関数は ReLU[14]，D および Q の活性化関数は負側の傾きが 0.2 の LeakyReLU[15] を用いた．また式 7 を負の対数尤度から最小二乗誤差に置き換えた [16]．すなわち G は $\mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[(D(G(\boldsymbol{c},X),X)-1)^2]$ を，D は $\mathbb{E}_{\boldsymbol{s}\sim q_\theta(\boldsymbol{s})}[(D(\boldsymbol{s},X)-1)^2]+\mathbb{E}_{\boldsymbol{c}\sim p_{\boldsymbol{c}}(\boldsymbol{c})}[D(G(\boldsymbol{c},X),X)^2]$ を最小化するように学習する．潜在変数 $\boldsymbol{c}$ の分布 $p_{\boldsymbol{c}}(\boldsymbol{c})$ は 3 次元の標準正規分布 $\mathcal{N}(0,I)$ とした．相互情報量制約の係数は $\lambda_I=1.0$ とした．バッチサイズ 64，イテレーション数 400000 とし，最適化手法に Adam を用いた．Adam のパラメータは，$lr=0.0002$，$\beta_1=0.5$，$\beta_2=0.999$ とした．パラメータの更新は式 6 に示す最適化を 1 イテレーションごとに行った．

4.3 実験方法

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する．ランダムに決定されたひとつの潜在変数の値に対して，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．ここで分類精度は，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，いかなる潜在変数の値に対しても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，潜在変数とラベルが直接対応していないことを調べるために有用である．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調べ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，潜在変数を全てランダムに決定する方法を用いる．

4.4 実験結果

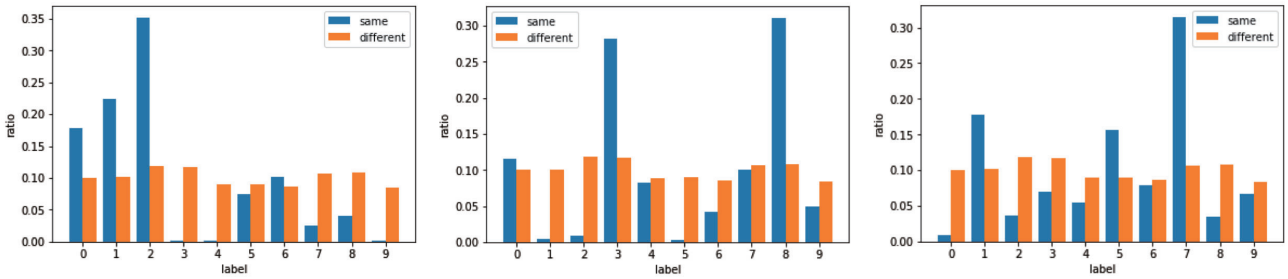
図 5 に実験結果を示す．異なるグラフでは異なる潜在変数の値を用いている．分類精度はいずれも 98%を超えており，潜在変数とラベルが直接対応していないことがわかる．潜在変数を固定した場合にはラベルの出現頻度に偏りが見られる．例えば，潜在変数の値 $\boldsymbol{c}_1$ を用いた場合はラベル 0，1，2 を多く出力している．潜在変数の値 $\boldsymbol{c}_2$ を用いた場合はラベル 3，8 を多く出力しており，潜在変数の値 $\boldsymbol{c}_3$ を用いた場合はラベル 1，7 を多く出力している．一方で，潜在変数をランダムに決定した場合は全てのラベルを概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 0，1，2 を要素としたサブグループを表現し，入力画像中の数字 0，1，2 を優先的に扱う．潜在変数の値 $\boldsymbol{c}_2$ はラベル 3 と 8 のサブグループ，潜在変数の値 $\boldsymbol{c}_3$ 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. おわりに

本研究では，画像キャプションに対する表現学習に向けた，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，本提案モデルを既存の GAN のアーキテクチャの組み合わせとして構築した．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，生成するクラス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．

将来研究として以下が挙げられる．画像キャプショニングは自然画像と自然言語を用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．また，本提案モデルの出力の分析により潜在変数の表現を分析したが，潜在変数の各次元や値が表現とどのように結びついているかは不明である．画像キャプションに対する表現学習の有効性の観点から，潜在変数に対するより詳細な分析は必要である．モデルの解釈性の観点から，人にとって解釈しやすいディスエンタングルな潜在変数を獲得することも重要である．



(1)$\boldsymbol{c}_1$, acc: 98.9%　(2)$\boldsymbol{c}_2$, acc: 98.7%　(3)$\boldsymbol{c}_3$, acc: 98.0%

図 5: 入力画像 1000 件に対する出力ラベルの出現頻度比．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)(3) 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．

参考文献

[1] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pp. 3137–3146, 2017.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[15] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30, p. 3, 2013.

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2813–2821. IEEE, 2017.