拡散過程を用いたキャプション生成への取り組み

平野理子　　小林一郎
お茶の水女子大学 理学部 情報科学科
{g1920535,koba}@is.ocha.ac.jp

概要

近年，拡散過程を用いたモデルが連続データ生成において非常に良い性能を達成しており，離散データ生成においても盛んに研究が進められている．本研究は，拡散過程を用いた画像キャプショニング手法の開発を目的とする．拡散過程による自然言語文生成を画像を対象に制御するために，推測されたノイズ除去状態の潜在変数から分類器によって求められた画像特徴量と実際の特徴量の損失に基づき勾配の更新を繰り返すことにより，最終的に画像に適したキャプションを生成する．生成したキャプションは，SOTA との比較において優位な結果ではなかったが，単純な手続きでも拡散過程を用いた画像キャプショニングが可能であることを確認した．

1　はじめに

これまで人工知能が発展しない理由とされていた「創造する」という課題が深層学習を使った画像生成や自然言語文生成によって可能になりつつある．自然言語文生成においては，汎用言語モデルの出現により言語モデル中心の文生成が主流となっている．そのような背景において，自然言語文生成における課題としては大量のコーパスから学習した汎用言語モデルを再学習を必要とせずに言語モデルの振る舞いを制御することが挙げられる．一方，近年，画像生成においては，拡散過程（Diffusion Process, DP）を採用した手法が，敵対的生成ネットワーク（Generative Adversarial Networks, GAN）による従来の最高性能を超える画像の生成を可能にした [1]．また，Li ら [2] によって，本来，連続的な情報を扱う DP に対して，離散情報である自然言語を扱えるようにした Diffusion Language Model (DLM) が提案されており，従来の最高性能を超えるような制御可能な自然言語文生成の可能性が示されている．これらの背景から，本研究では拡散過程を用いた画像キャプショニング手法を提案する．手法の開発においては，拡散過程に基づく非自己回帰の言語モデルとその外部に拡散過程を制御する識別器を導入することで，画像に対応したキャプション生成を可能にする．

2　関連研究

拡散過程を用いた画像生成　Stable Diffusion [3] や DALL・E2 [1] は拡散過程を用いて画像を生成するモデルである．これらは，与えられたテキストからその内容に従った画像を生成するタスクや画像から画像への変換タスクなどにおいて，非常に高い精度を達成している．特に DALL・E2 は，入力テキストから画像埋め込みを生成するモデルと，画像埋め込みから画像を生成する二つのモデルから構成され，どちらのモデルでも拡散過程を用いることで質の高いサンプルを生成している．また，外部の基盤モデルで、CLIP [4] という，大量の画像とテキストのペアデータで学習したニューラルネットワークを使用することで，テキストの意味的な内容を忠実に画像内で表現することを可能にしている．

Bit Diffusion　Bit Diffusion [5] は，連続状態を扱う拡散モデルを用いて，離散データである自然言語文を生成している．具体的には，まず離散データを 2 値ビットで表し，これを実数に写像することで拡散モデルの扱える連続データに変換する．これにより，ピュアなノイズからノイズを徐々に除去することで，テキストをサンプリングすることを可能とした．また，Self Conditioning や Asymmetric Time Intervals といったサンプルの質を向上させる技術もその枠組みに取り入れている．本研究では，Bit Diffusion とは異なり，離散データを埋め込み表現に写像することで，拡散モデルの扱える連続状態にテキストを変換している．

離散拡散過程を用いた画像キャプショニング　Zhu ら [6] は，本研究と同じく拡散過程を用いた画像キャプショニング手法を提案している．画像からキャプションの内容や長さを推測し生成過程の制

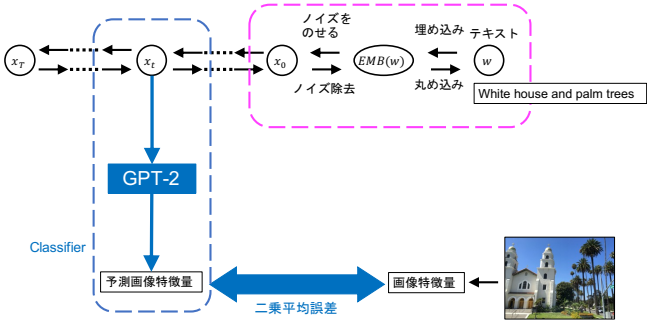


図 1　Diffusion-LM を用いたキャプション生成

御を行うことで，質の高い画像キャプショニングに成功している．本研究では拡散過程を用いた言語モデルの学習の際，データにノイズが乗った状態とタイムステップの情報の二つから，ノイズの乗っていない状態のデータを予測しているが，DDCap ではそれらに加えアテンションマスクや，画像を入力として与えた CLIP [4] の出力なども用いて，ノイズの乗っていない状態のデータを予測している．また，DALL・E2 [1] と同様，CLIP という外部の基盤モデルを使用しているところも特徴である．

3　DLM を用いたキャプション生成

3.1　提案手法

本研究で提案する，拡散過程を用いた言語モデルと外部の分類器の二つのモデルを用いたシンプルな画像キャプショニング手法の概要を図 1 に示す．学習の大まかな流れとしては，拡散過程を用いた言語モデルを大量のテキストデータで学習させたのち，画像とキャプションのペアデータを用いて分類機を学習させる．ノイズ除去の過程を辿りながらデータをサンプリングする際は二つのモデルを組み合わせ，学習させた言語モデルで文を生成する過程を分類器で制御することで，画像の内容を説明する自然言語文，キャプションの生成を行う．

3.2　Diffusion LM

Diffusion LM (Diffusion Language Model) とは，拡散過程を用いた言語モデルのことを指す．Diffusion LM を構築するには，標準的な連続状態を扱う拡散モデルに幾つかの修正を加える必要がある．図 1 のピンクの枠にあるような，埋め込みと丸め込みの過程の導入がその一つである．埋め込み関数を定義することで，離散データであるテキストを連続空間に写像する．丸め込み過程を導入することで，埋め込み空間のベクトルを単語を表すベクトルに写像し返す．言語モデルの学習の対象は，ガウシアンノイズからノイズを徐々に除去し，最終的に流暢性のある自然言語文を生成する過程である．つまり，各タイムステップにおけるノイズの除去 $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ を実現する際，必要となるパラメータを学習する．

具体的な学習の流れとしては，まずキャプションをトークン化し，各トークンをベクトル空間に埋め込む．サンプリングされたタイムステップ t によって決められる量のノイズを埋め込み表現に乗せ，ノイズの乗った状態 $\mathbf{x}_t$ にする．ノイズの乗った状態 $\mathbf{x}_t$ とタイムステップ t をニューラルネットワーク $f(\mathbf{x}_t, t)$ に入力として与え，ピュアなデータを推測させる．損失関数は以下の式に従う．

$$
\begin{aligned}
&L^{\text{e2e}}_{\mathbf{x}_0\text{-simple}}(\mathrm{w}) \\
&= \mathbb{E}_{q_\phi(\mathbf{x}_{0:T}|\mathrm{w})}\left[\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t;\mathbf{x}_0)\|^2 + \sum_{t=2}^{T}[\|\mathbf{x}_0 - f_\theta(\mathbf{x}_t,t)\|^2]\right] \\
&+ \mathbb{E}_{q_\phi(\mathbf{x}_{0:1}|\mathrm{w})}\left[\|\mathrm{EMB}(\mathrm{w}) - f_\theta(\mathbf{x}_1,1)\|^2 + \log p_\theta(\mathrm{w}|\mathbf{x}_0)\right]
\end{aligned}
\tag{1}
$$

3.3　Classifier

言語モデルとは別のモデルである，外部の分類器（Classifier）の役割は，Diffusion LM が最終的に自然言語文をサンプリングする過程中に反復的に生成する潜在変数に対して勾配更新を行うことで，最終的に生成されるテキストを制御することである．潜在変数 $\mathbf{x}_{0:T}$ を制御するモデルは以下のように書くことができる．

$$
p(\mathbf{x}_{0:T}|c) = \prod_{t=1}^{T} p(\mathbf{x}_{t-1}|\mathbf{x}_t, c) \tag{2}
$$

分解すると，

$$
\begin{aligned}
p(\mathbf{x}_{t-1}|\mathbf{x}_t, c) &\propto p(\mathbf{x}_{t-1}|\mathbf{x}_t) \cdot p(c|\mathbf{x}_{t-1}, \mathbf{x}_t) \\
&\propto p(\mathbf{x}_{t-1}|\mathbf{x}_t) \cdot p(c|\mathbf{x}_{t-1})
\end{aligned}
\tag{3}
$$

第一項の各タイムステップでのノイズの除去 $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ は，Diffusion LM によってパラメータ化された．つまり，第二項のデータにノイズの乗った状態から制御対象への変換 $p(c|\mathbf{x}_{t-1})$ を，ニューラルネットワークを用いた classifier に学習させ，パラメータ化する．具体的には図 1 の青い枠にあるように，自己回帰の言語モデル（GPT-2 [7]）を用いて，ノイズの乗っている各タイムステップの潜在変数か

ら画像特徴量の予測を行い，正解画像特徴量との二乗平均誤差をとることで機械学習を行う．正解画像特徴量は，画像を Resnet50 [8] に通すことで取得している．

サンプリング時は，学習をさせた Diffusion LM と Classifier を組み合わせ，画像に応じた自然言語文生成過程の制御を行う．具体的な流れとしては，まずガウシアンノイズを Diffusion LM に与え，反復的に 1 タイムステップノイズを除去した状態の潜在変数 $\mathbf{x}_{t-1}$ を推測させる．各タイムステップにおいて，$\mathbf{x}_{t-1}$ から Classifier が画像特徴量を推測する．Diffusion LM によって $\mathbf{x}_t$ を用いて推測された $\mathbf{x}_{t-1}$ に付加されているノイズの量（式 4 の第 1 項）と，Classifier が $\mathbf{x}_{t-1}$ から推測した画像特徴量と正解画像特徴量間の二乗平均誤差（式 4 の第 2 項）の和から、誤差逆伝播法を用いて勾配を求めパラメータを更新する. 勾配更新を行うことで，$p(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$ を最大化し，流暢性があり画像に応じた適切なテキストを生成する．

$$\begin{aligned}&\nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1}|\mathbf{x}_t, c)\\&= \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1}|\mathbf{x}_t) + \nabla_{\mathbf{x}_{t-1}} \log p(c|\mathbf{x}_{t-1})\end{aligned} \quad (4)$$

以上の流れを各タイムステップで繰り返し，最終的に画像に応じた適切なキャプション生成を目指す．

4 実験

この節では，言語モデル（Diffusion LM）と分類器 (Classifier) を用いて，実際に画像からのキャプション生成を行う．本実験の目的は，提案手法を用いて画像キャプショニングを行い，生成されたキャプションから精度を求め評価を行うことである．

4.1 実験設定

評価指標 本実験では，BLEU [9]，ROUGE-L [10] を評価指標として用いる．BLEU は機械翻訳，ROUGE はテキストの自動要約の評価指標として知られている．また，これら二つはキャプション生成の評価指標として用いられることも多いため，本論文でも画像キャプショニングの評価指標として使用する．

データセット データセットには，Microsoft COCO[1)] を使用する．Microsoft COCO が定めた訓練画像のうち，108,302 枚を本実験の訓練データ，4,985 枚を評価データとしこれを用いてパラメータ調整を行った．残りの 5,000 枚をテストデータとし，評価を行う．Diffusion LM の学習時，語彙数は 13,461，埋め込み次元は 256 に設定している．

比較手法 現在の SOTA な画像キャプショニング手法の一つである OFA [11] の本実験のテストデータにおける精度を求め，提案手法との比較を行う．

4.2 実験結果

表 1 に実験結果を示す．すべての指標において，画像キャプショニングの SOTA な手法の一つである OFA には及ばない結果となった．しかし，BLEU-1 のスコアは質の高いサンプルが生成されていると一般的に言われる 0.40 の値を超えることができた．

表 2 は，実際に生成されたキャプションの一部である．トイレやテニス，鳥の写真では正解キャプションと似たような，画像に応じた適切な文を生成できている．しかし，一方で，電車の画像での生成キャプションを見ると，本来，赤一色である電車の色を赤と白と認識していることから，色に対しての学習が足りていないことがわかる．また花瓶の画像では文法のミスがあり，生成文の流暢性に問題があることから，言語モデルの学習にも改善が必要である．サーフボードを持った人の画像から生成されたキャプションには，画像に存在しない物体についての記述があり，ベンチに座る人の画像からは画像に関係はあるが対応はしていないキャプションが生成されるなどしている．これより生成過程の流暢性と制御の充足度のトレードオフにも問題があることがわかる．

4.3 考察

精度は画像キャプショニングの SOTA な手法を比べると良い結果を出すことができなかったが，実際に生成されたキャプションから，非常にシンプルなモデルでも拡散過程を用いた画像キャプショニングを行えることを示した．

5 まとめ

拡散過程を用いた生成モデルは，ピュアなノイズからノイズを除去した潜在変数を反復的に生成することで，最終的にデータをサンプルする．本研究では，分類器を用いてこれら潜在変数に対して勾配更新を行うことで，言語モデルの再学習を行わずに生

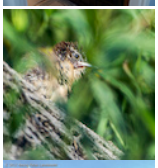
1） https://cocodataset.org/#home

表 1 キャプション生成の実験結果

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Bleu-4 com	ROUGELpre	ROUGELrec
OFA	0.8013	0.5510	0.334	0.1952	0.3067	0.4836	0.4524
提案手法	0.6338	0.3253	0.1742	0.0906	0.1557	0.3264	0.2534

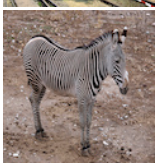
表 2 キャプション生成例

Gen.: A bathroom with a toilet , a shower and a sink .
GT: Small bathroom with a toilet, sink, shower, and mirror.



Gen.: A small bird perched on a tree branch .
GT: A bird perched on top of a tree branch.

Gen.: A red and white train on a track .
GT: A train parked in a train depot loading passengers.



Gen.: A zebra standing in the grass in a field .
GT: A zebra standing on a rocky dirt field with no grass.

Gen.: There is a man sitting on a bench with his dog on the beach .
GT: a person sitting on a bench on a city street

Gen.: A man holding a tennis racquet on a court .
GT: A man holding a tennis racquet on a tennis court.

Gen.: A small cheese pizza is sitting on a plate .
GT: A plate topped with a cheesy meaty pizza on a table.

Gen.: A vase of flowers in a vase on a wooden table .
GT: A vase of flowers sits on a table.

Gen.: A large building with a clock on the top .
GT: A very tall clock mounted to the side of a building.

Gen.: A man riding a wave while riding a surfboard .
GT: A man on a bike carrying a surf board.

成過程を制御し，画像に応じた自然言語文を生成する手法を提案した．

今後は，言語モデルの改良や外部の分類器を用いずに制御する方法などに取り組み，提案手法の改良を進めたい．

参考文献

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. **CoRR**, abs/2204.06125, 2022.

[2] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. **CoRR**, abs/2205.14217, 2022.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[5] Ting Chen, Ruixiang Zhang, and Geoffrey E. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. **ArXiv**, abs/2208.04202, 2022.

[6] Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. Exploring discrete diffusion models for image captioning. **arXiv preprint arXiv:2211.11694**, 2022.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational**

Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[10] Unnat Jain, Svetlana Lazebnik, and Alexander G. Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2018.

[11] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. **CoRR**, abs/2202.03052, 2022.