

# 拡散過程を用いたキャプション生成における分類器導入の精度への影響の検証

平野理子<sup>1</sup> 小林一郎<sup>1</sup>

<sup>1</sup> お茶の水女子大学

{g1920535, koba}@is.ocha.ac.jp

## 概要

近年、拡散過程を用いた生成モデルは連続領域において最先端の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行われている。本研究では、拡散言語モデルを使って自然言語処理タスクの一つであるキャプション生成に取り組み、精度向上を目的に画像による制御を言語モデルとは別の構成要素（分類器）によって行う手法と言語モデルのみで実現する手法の性能の違いについて検証を行った。実験で測定された精度に関して提案手法は比較手法を上回ることができなかったが、提案手法によって画像の内容に応じたキャプションが一定の精度で生成できることを確認した。

## 1 はじめに

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。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（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。この拡散過程を使った制御可能な自然言語文生成を可能とした先行研究 [2] では、外部の構成要素（分類器）を導入し制御を行い良い精度を達成している一方、分類器を使用せずに良い精度を達成している拡散過程を用いた制御可能な自然言語文生成手法も提案されている [3]。これら背景を踏まえ、本研究は拡散過程を用いた画像キャプション生成手法の開発、また分類器の導入の有無による精度への影響の調査を目的とする。

## 2 関連研究

**拡散モデルによる画像生成** Stable Diffusion [4] 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。Stable Diffusion はノイズ除去ネットワーク内の U-Net 層にてテキストプロンプトを Cross-Attention を使用して Transformer [5] に条件づけることで、入力テキストに応じた画像生成を可能にしている。

**拡散過程を用いた seq2seq 自然言語文生成** DiffusionLM [2] は生成文に小さな制約をかけることに成功したが、より高難度な条件を与える seq2seq タスクにおいても拡散過程を用いたモデルは高い性能を示している。SeqDiffSeq [3] はその一例で、新しいノイズスケジュールを提案しテキストの品質と推論時間に関して他の拡散ベースのモデルと比較して多くの seq2seq 生成タスクの性能を向上させた。ノイズ除去ネットワークにはフルアテンションの BART decoder [6] を使っており、ソースシーケンスも入力として与えることで分類器を含まないモデル構造 (Classifier free) になっている。

**SSD-LM** 拡散言語モデルであり、革新的な二つの大きな特徴を持つ。一つ目は、多くの拡散言語モデルはシーケンス全体を一度に生成する非自己回帰的である一方、SSD-LM [7] では半自己回帰的にシーケンス内でトークンを左から右に生成することで出力の長さの柔軟性が向上している。また一般的な自己回帰言語モデルと同じトークナイザーを採用することで、追加の学習なしで外部の既製分類器を使った制御を可能としている。実際に条件付き条件付きでないテキスト生成どちらにおいても競合ベースラインを上回り、高いモジュール性も備えている。

### 3 拡散過程を用いたキャプション生成

#### 3.1 提案手法

本研究は拡散過程を用いたキャプション生成手法の開発と精度向上に取り組んでおり、画像による制御の方法として言語モデルとは別の構成要素となる分類器 (Classifier) を導入する場合としない場合の二つの手法を提案し、精度や結果の比較を行う。導入をする場合、提案モデルは拡散過程に基づく言語モデル (DLM: Diffusion Language Model) と分類器の二つの要素から構成され、DLM の自然言語文生成過程を分類器で制御することによって画像の内容を説明するキャプションを生成する。分類器を導入しない場合においては、DLM 内のノイズ除去ネットワーク (DNN: Denoising Neural Network) に画像の情報を渡すことで、入力画像に応じた制御を行う。

#### 3.2 拡散過程に基づく言語モデル (DLM)

DLM とは拡散過程を用いた非自己回帰の言語モデルのことである。DLM 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り、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し、丸め込み過程によって連続空間のベクトルを単語を表すベクトルに変換する。DLM の学習の対象はピュアな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、まず学習テキストデータをトークン化し、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  $t$  をノイズ除去ネットワーク  $f(\mathbf{x}_t, t)$  に与え、元のノイズの乗っていない状態のデータ  $\mathbf{x}_0$  を推測させる。推測した  $\mathbf{x}_0$  から丸め込んだトークン列  $w$  も求め、損失関数 (式 (1)) に従ってこれらと元のデータとの損失を計算する。

$$L_{\text{DLM}} = \mathbb{E}_{q_\phi(\mathbf{x}_{0:T}|\mathbf{w})} \left[ \|\tilde{\mu}_t(\mathbf{x}_t; \mathbf{x}_0)\|^2 + \sum_{t=2}^T [\|\mathbf{x}_0 - f(\mathbf{x}_t, t)\|^2] \right] + \mathbb{E}_{q_\phi(\mathbf{x}_{0:1}|\mathbf{w})} \left[ \|\mathbf{w} - f(\mathbf{x}_1, 1)\|^2 + \log p_\theta(\mathbf{w}|\mathbf{x}_0) \right] \quad (1)$$

#### 3.3 分類器を使用する手法

**分類器** 分類器の役割は、DLM が自然言語文をサンプルする過程の中で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により、最終的に生成される自然言語文を制御することである。このような条件  $c$  を満たすように潜在変数  $\mathbf{x}_{0:T}$  を制御するモデルは式 (2) で表せる。本研究で条件  $c$  とは画像、特に画像特徴量のことを指す。

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

右辺の確率を分解すると、

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

式 (3) 右辺第一因子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 DLM によってパラメータ化する。つまり式 (3) 右辺第二因子のデータにノイズの乗った状態  $\mathbf{x}_t$  から制御条件  $c$  への変換  $p(c|\mathbf{x}_{t-1})$  が分類器の学習対象である。

**学習過程** 分類器は DLM 内のノイズ除去ネットワーク  $f(\mathbf{x}_t, t)$  の出力を使用するため、DLM と分類器は同時に学習を行う。具体的には、分類器はノイズの乗った状態  $\mathbf{x}_t$  から  $f(\mathbf{x}_t, t)$  が推測するノイズの乗る前の状態  $\mathbf{x}_0$  を入力として受け取り、これから線形回帰で画像特徴量を予測する。予測画像特徴量  $\tilde{c}$  と正解画像特徴量  $c$  間の二乗平均誤差を DLM の損失関数の式 (1) に追加することで (式 (4))、DLM と分類器の同時学習を実現している。

$$L_{\text{DLM\&Classifier}} = L_{\text{DLM}} + L_{\text{MSE}}(\tilde{c}, c) \quad (4)$$

**生成過程** 分類器を導入する場合、DLM の生成過程を分類器が制御することで画像に応じた自然言語文が生成される (図 1)。具体的な流れとしては、まず入力としてガウシアンノイズ  $\mathbf{x}_T$  と画像をモデルに与え、 $t = T$  から 1 までの各タイムステップにおいて次の流れを繰り返す。ノイズの乗った状態  $\mathbf{x}_t$  から学習をさせた DLM 内のノイズ除去ネットワーク  $f(\mathbf{x}_t, t)$  を使って言語特徴量  $\mathbf{x}_0$  を抽出させる。抽出された言語特徴量から分類器によって予測画像特徴量  $\tilde{c}$  を求め、その予測画像特徴量と正解画像特徴量間の二乗平均誤差と  $\mathbf{x}_t$  にのっているノイズの量の和から勾配を求め  $\mathbf{x}_t$  を更新する。更新した  $\mathbf{x}_t$  と更新された  $\mathbf{x}_t$  から改めて推定した  $\mathbf{x}_0$  を使って、1 タイムステップノイズを除去した状態  $\mathbf{x}_{t-1}$  をサンプ

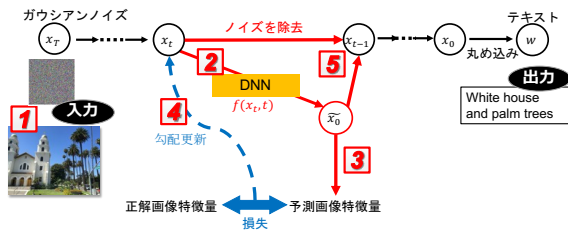


図1 分類器を使用した際のキャプション生成の流れ

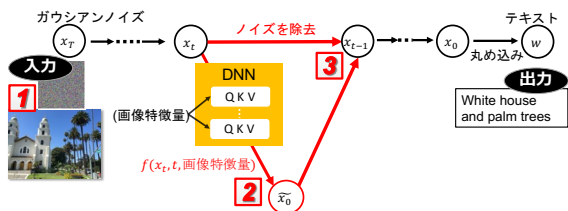


図2 分類器を使用しない場合のキャプション生成の流れ  
リングする。この流れを反復的に各タイムステップで実行し、最終的に画像に応じたキャプションの生成を実現している。

### 3.4 分類器を使用しない手法 (Classifier Free)

続いて分類器を導入しない手法について説明をする。制御を行う外部要素がない場合、画像による制御を行うために単純に DLM 内のノイズ除去ネットワークに画像特徴量  $c$  を条件として与える。

$$\tilde{x}_0 \approx f(x_t, t, c)$$

ノイズ除去ネットワークに画像特徴量  $c$  を条件付けする方法として、ネットワーク内の Multi-Head Attention 層において Cross-Attention を導入している。つまりノイズの乗った状態  $x_t$  と画像特徴量  $c$  から、 $Q = x_t \cdot W_Q, K = c \cdot W_K, V = c \cdot W_V$  として  $Q, K, V$  を用意する。このようにすることで DLM 内の潜在変数に画像の情報が注入されるようになっている (図 2)。

## 4 実験

本実験の目的は提案手法を用いて入力画像から実際に生成されるキャプションの精度を求め評価を行うことで、分類器の導入の有無による精度の違いを検証することである。

### 4.1 実験設定

**データセット** データセットには、Microsoft COCO<sup>1)</sup>を使用する。広く使われている Karpathy 分割方法 [11] に従って 113,287 枚を学習データ、5,000

1) <https://cocodataset.org/#home>

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。DLM の学習時語彙数は 13,461、埋め込み次元は 256 に設定されている。また DLM 内の DNN (ノイズ除去ネットワーク) には BERT Encoder [5] を使用しており、画像から特徴量を抽出する手法には BLIP [12] を採用している。

**評価指標** 他の手法と提案手法の性能比較として、生成されたキャプションを品質と多様性の 2 つの側面から評価する。品質の評価としては 5 つの評価指標を導入している。まず画像キャプションの標準的な評価方法に従って 4 つの評価指標 BLEU [13], METEOR [14], ROUGE-L [15], CIDEr [16] を用いる。BLEU や METEOR は機械翻訳、ROUGE-L は要約、CIDEr は画像キャプションタスクによく使われる評価指標である。これらは生成文と正解文の間の文字列類似性を図るため、意味的類似性の評価を目的に BERTScore [17] を導入する。続いて多様性の指標として 3 つ導入をしている。一つ目の dist-1 は各生成文内の多様性を測定するものであり、値が低いと生成文に繰り返しの単語がより含まれていることを示唆する。続いて文レベルの自己 BLEU を使って文レベルの多様性を評価する。これは 1 つの画像からの出力セットにおける n-gram 重複を測定するもので、値が低いことは多様性が高いと考えられる。最後に出力セット内の異なる 4-gram の比率を表す多様な 4-gram (div-4) を計算する。提案手法についてひとつの画像ごとに 5 つのサンプルを生成し多様性の評価指標を計算している。

**比較手法** 比較手法として 3 つの手法、OFA [8], LaBERT [9] と DDCap [10] を使って性能を比較する。OFA は現在の SOTA な画像キャプション手法の一つであり、モダリティ (画像や言語) とタスクを統合して扱う seq2seq モデルである。LaBERT は文長制御が可能な DLM と同じく非自己回帰的にキャプションを生成する。DDCap は拡散過程を用いてキャプション生成に取り組んでいる Classifier free 手法である。提案手法の DLM では内部のデータは連続状態であるのに対し、DDCap はデータを離散状態のまま DLM 内で取り扱う。

### 4.2 結果









評価指標において、提案手法は分類器の有無に関わらず比較手法と比べて良い精度を達成することは



表 1 実験結果

分類器	Bleu-4 ↑	METEOR ↑	ROUGEL ↑	CIDEr ↑	BERTScore ↑	dist1 ↑	selfBleu ↓	div4 ↑	length
OFA [8]	0.424	0.312	0.613	1.451	0.782	0.861	-	-	9.71
LaBERT [9]	0.306	0.254	0.548	1.028	0.786	0.894	-	-	8.89
DDCap [10]	0.350	0.282	0.574	1.178	-	-	-	-	-
提案手法 有	0.191	0.239	0.514	0.688	0.711	0.933	0.0117	0.980	12.46
提案手法 無	0.192	0.206	0.498	0.594	0.696	0.893	0.0128	0.989	11.04

表 2 分類器の有無による生成キャプション例の比較

	有:A man holding a plate of food standing in a kitchen. 無:A man that is sitting on in a kitchen.		有:A piece of pizza sitting on top of a green plate. 無:A cup of on a plate near a.
	有:a man sitting on a bench reading a book 無:A bench on the edge of a park bench.		有:A man riding a motorcycle down a drive. 無:A man riding a motorcycle down the street.
	有:Two young people sitting on a court tire. 無:A group of people sitting around a bench or UNK.		有:A bird perched on a branches of a tree forest in Michael. 無:A bird sitting on top of a tree.
	有:A classroom vintage of luggage two pieces of various suitcases 無:A luggage suitcase is out in the middle of the luggage.		有:A papers of buses that are parked on the ground. 無:A city is down the street.

できなかった。特に質を評価する指標の数値に関しては差が大きくあり、実際に提案手法によって生成されたキャプションは文法の誤りが多く見受けられたり流暢性に欠けていたりすることが確認できる。分類器を使用する手法では言語モデルと分類器は同時に学習を行なったため言語モデルの学習へ悪影響があった可能性がある。同様に分類器を使用しない手法においては、DLM 内のノイズ除去ネットワークに画像特徴量も追加して与えることは生成されるテキストの流暢性の低下につながることを示唆される。提案手法については分類器を導入し制御を行う方法がより良い性能を持つことが分かった。ノイズ除去ネットワークに画像の情報を渡す分類器を使用しない手法より、生成時に複数回画像特徴量を使って DLM 内の潜在変数を更新する分類器を用いた手法の方が生成されるテキストに画像の情報がより反映されると予測する。また生成キャプションの平均長から分類器を使用した場合により長いキャプションが生成される傾向があることがわかり、これが文内の多様性を図る指標における良い値に寄与したことが考えられる。分類器を使用しない手法によって生成されたキャプションのうち特に精度が悪いものは画像の内容を全く捉えてられていない場合が見受けられる (左一番下)。DLM 内での画像特徴量の処

理方法について修正が必要であると考えている。

## 5 まとめ

拡散過程を用いた生成モデルは、プリアなノイズから反復的に少しずつノイズを除去することで、最終的にデータをサンプルする。本研究では拡散過程を用いた画像キャプション生成手法を開発し、また分類器の導入の有無による精度の比較を行った。実際に画像からキャプションを生成する実験を通して、分類器を導入しない手法よりも分類器を使用して制御を行う手法の方が評価指標において高い精度を達成することを確認した。また分類器の導入の有無に関わらず、提案手法は拡散過程を用いて画像から画像の内容を適切に捉えたキャプションを一定の精度で生成できることを示している。一方提案手法は他の拡散言語モデルを使った画像キャプション生成手法を含め比較手法と比べ良い精度を達成できていない。今後は、言語モデル DLM 内のノイズ除去ネットワーク (DNN) について様々な構造を試したり、分類器を使わない場合において画像の特徴量の DNN への入力仕方を改良したり、提案手法の更なる精度の向上に取り組みたい。

## 謝辞

本研究は JSPS 科研費 18H05521 の助成を受けて行ったものです。

## 参考文献

- [1] Ramesh Aditya, et al. Hierarchical text-conditional image generation with CLIP latents. **CoRR**, 2022.
- [2] Xiang Lisa Li, et al. Diffusion-lm improves controllable text generation. **ArXiv**, Vol. abs/2205.14217, , 2022.
- [3] Yuan Hongyi, et al. Seqdiffuseq: Text diffusion with encoder-decoder transformers. **ArXiv**, 2022.
- [4] Rombach Robin, et al. High-resolution image synthesis with latent diffusion models. **CoRR**, 2021.
- [5] Vaswani Ashish, et al. Attention is all you need. **CoRR**, Vol. abs/1706.03762, , 2017.
- [6] Mike Lewis, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2020.
- [7] Xiaochuang Han, et al. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. **ACL: Annual Meeting of the Association for Computational Linguistics**.
- [8] Wang Peng, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. **CoRR**, 2022.
- [9] Chaorui Deng, et al. Length-controllable image captioning. **arXiv preprint arXiv:2007.09580**, 2020.
- [10] Zixin Zhu, et al. Exploring discrete diffusion models for image captioning. **arXiv preprint arXiv:2211.11694**, 2022.
- [11] Karpathy Andrej, et al. Deep visual-semantic alignments for generating image descriptions. **CoRR**, Vol. abs/1412.2306, , 2014.
- [12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In **ICML**, 2022.
- [13] Kishore Papineni, et al. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, 2002.
- [14] Satanjeev Banerjee, et al. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005.
- [15] Unnat Jain, et al. Two can play this game: Visual dialog with discriminative question generation and answering. 2018.
- [16] Ramakrishna Vedantam, et al. Cider: Consensus-based image description evaluation. 2015.
- [17] Zhang Tianyi, et al. Bertscore: Evaluating text generation with bert. 2020.