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　(指導教員：小林 一郎)

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 [3]．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．

2　DLM を用いたキャプション生成

2.1　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（DLM）と外部の分類器（Classifier）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，DLM 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で Classifier を学習させる．ノイズ除去を繰り返し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，DLM の自然言語文生成過程を Classifier 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．

2.2　DLM

DLM(Diffusion Language Model) とは，拡散過程を用いた言語モデルのことを指す．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る．図 1 のピンクの枠にある，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

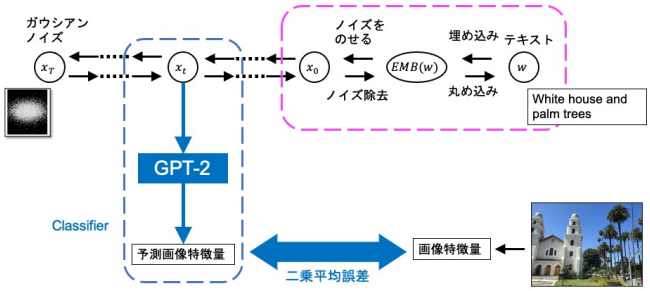

図 1: Diffusion-LM を用いたキャプション生成

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．

2.3　Classifier

言語モデルとは別の外部のモデル Classifier の役割は，DLM 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を用いて勾配更新を行うことで，最終的に生成される自然言語文を制御することである．

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．

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

分解すると，

$$\begin{aligned} p(\mathbf{x}_{t-1}|\mathbf{x}_t, c) &\propto p(\mathbf{x}_{t-1}|\mathbf{x}_t) \cdot p(c|\mathbf{x}_{t-1}, \mathbf{x}_t) \\ &\propto p(\mathbf{x}_{t-1}|\mathbf{x}_t) \cdot p(c|\mathbf{x}_{t-1}) \end{aligned} \tag{2}$$

式 2 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 2 右辺第二因子のデータにノイズの乗った状態から制御対象への変換 $p(c|\mathbf{x}_{t-1})$ が Classifier の学習対象である．具体的には図 1 の青い枠にあるように，まずノイズの乗っている状態の潜在変数 $\mathbf{x}_t$ から自己回帰の汎用言語モデル（GPT-2）を用いて言語特徴量を抽出する．続いて，抽出した言語特徴量から画像特徴量を予測し，正解画像特徴量との二乗平均誤差をとり，これが小さくなるよう学習を行う．

キャプションを生成する時は，学習をさせた二つのモデル DLM と Classifier を組み合わせて使用し，生成される自然言語文を制御する．具体的な流れとしては，まずガウシアンノイズを DLM 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．DLM によって推測された $\mathbf{x}_{t-1}$ に付加されているノイズの量（式 3 の第 1 項）と，Classifier が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 3 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

表 1: キャプション生成の実験結果

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGEL	CIDEr
OFA	0.837	0.692	0.548	0.424	0.312	0.613	1.451
LaBERT	0.727	0.568	0.424	0.306	0.254	0.548	1.028
提案手法	0.189	0.126	0.078	0.048	0.144	0.269	0.138

表 2: キャプション生成例

生成文: A man flying a kite on the beach.
正解文: A man on a beach holding a kite.

生成文: A red fire hydrant on the side of a street.
正解文: A red fire hydrant in grassy area next to street.

生成文: A pizza is served on a pizza pan
正解文: a pizza on a pan sitting on a table.

生成文: A man on a motorcycle on a street.
正解文: A man is riding a motorcycle on the street.

生成文: A man is riding on the back of a horse.
正解文: A man riding on the back of a white horse.

生成文: A man sitting on a park bench with a dog.
正解文: a woman is sitting with a statue on a bench

する．

$$\nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$$
$$= \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1}|\mathbf{x}_t) + \nabla_{\mathbf{x}_{t-1}} \log p(c|\mathbf{x}_{t-1}) \quad (3)$$

以上の流れを繰り返すことで，最終的に画像からのキャプション生成を実現する．

3 実験

この節では，言語モデル（DLM）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実際に行い，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．

3.1 実験設定

データセット データセットには，Microsoft COCO[1]を使用する．広く使われている Karpathy 分割方法に従って，113,287 枚を学習データ，5,000 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．標準的な評価方法に従って，他の手法と提案手法の性能の比較を 4 つの評価指標 BLEU，METEOR，ROUGE-L，CIDEr を用いて行う．

比較手法 比較手法として OFA [4] と LaBERT [1] を用いる.OFA は現在の SOTA な画像キャプショニング手法の一つであり，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 は文の長さの制御が可能な，DLM と同じく非自己回帰型のキャプション生成モデルである．

3.2 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 や非自己回帰モデリングを採用している LaBERT には及ばない結果となった．表 2 は，実際に生成されたキャプションの一部である．

3.3 考察

評価指標は画像キャプショニングの他の SOTA な手法などと比べても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例をみると様々な入力画像から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ことがわかる．一方で，ベンチに座る人と銅像の画像から生成されたキャプションでは銅像を犬と認識しており，モデルの学習には改善が必要である．

4 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，分類器による潜在変数への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．実際に画像からキャプションを生成する実験を通して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．今後は，言語モデルや分類器を改良し，提案手法の精度の向上に取り組みたい．

参考文献

[1] Deng, C., Ding, N., Tan, M. and Wu, Q.: Length-Controllable Image Captioning, in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, p. 712–729, Berlin, Heidelberg (2020), Springer-Verlag.

[2] Li, X. L., Thickstun, J., Gulrajani, I., Liang, P. and Hashimoto, T. B.: Diffusion-LM Improves Controllable Text Generation, *CoRR*, Vol. abs/2205.14217, (2022).

[3] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents, *CoRR*, Vol. abs/2204.06125, (2022).

[4] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J. and Yang, H.: OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, *CoRR*, Vol. abs/2202.03052, (2022).

[1] https://cocodataset.org/#home