拡散モデルを用いたキャプション生成

理学専攻・情報科学コース
2340674
平野 理子

1 はじめに

近年，大規模言語モデルが言語生成において飛躍的な進展を遂げる一方，拡散モデルは画像生成を中心に高い精度を達成し，生成モデルの新たな可能性を切り開いた．特に，拡散モデルは生成過程を通じてデータの品質と多様性を両立できる点で注目を集めている．自然言語処理タスクへの応用も進んでいるが，生成結果の制御性や流暢性に課題が残る．本研究では，この背景を踏まえ，拡散モデルを用いたキャプション生成手法の開発を目的とする．具体的には，モデル構造の改良や生成過程の制御手法を検証し，品質，制御性，多様性，および流暢性を兼ね備えたキャプション生成の実現を目指す．

2 拡散言語モデルを用いたキャプション生成

提案モデルは，拡散過程に基づく言語モデル (DLM:Diffusion Language Model)[1] と分類器の 2 つの要素で構成される (図 1)．DLM は非自己回帰型 (NAR:Non-autoregressive) 言語モデルであり，標準的な連続状態を扱う拡散モデルに対して，埋め込みと丸め込みの過程を導入することで構築される．DLM の生成過程は以下の式 (1) で表され，完全なノイズから徐々にノイズを取り除くことでデータを生成する潜在変数モデルである．

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \tag{1}$$

分類器の役割は，DLM 内の潜在変数 $\mathbf{x}_{0:T}$ を勾配更新し，最終的に生成される自然言語文が与えられた条件を満たすように制御することである [2]．

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) \propto p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \cdot p_\phi(c|\mathbf{x}_{t-1}) \tag{2}$$

2.1 手法 1 - DLM と分類器の統合

DLM の各タイムステップでのノイズ除去処理は以下の式 (3) で表される．ここで，$\tilde{\mathbf{x}_0}$ は DLM 内のノイズ除去ネットワーク (DNN:Denoising Neural Network) によって求められる．

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{3}$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{x}_0) \coloneqq \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \tilde{\mathbf{x}_0}) \tag{4}$$

ノイズの乗った状態を制御条件へ変換 (式 (2) 右辺第二項) する分類器による処理にも $\tilde{\mathbf{x}_0}$ の情報が必要となる．

$$p_\phi(c|\mathbf{x}_{t-1}) \approx p_\phi(c|\tilde{\mathbf{x}_0}) \tag{5}$$

先行研究 [1] では DLM 用と分類器用の $\tilde{\mathbf{x}_0}$ を別々のモデルで計算していた．本研究では，モデルを軽量化しつつ質と制御性を向上させるため，DLM と分類器を統合し，DLM 内の DNN が求めた $\tilde{\mathbf{x}_0}$ を分類器で活用する手法を提案した．

結果 DLM と分類器の統合による効率化は精度の向上に寄与しないことが判明した (表 1)．各タスクが必要とする特徴量の性質が異なるため，統合設計では分類器が必要とする情報が十分に保持されず，適切な特徴量を得られなかった可能性がある．一方で統合によってパラメータ数が半分に削減され，推論時間が大幅に短縮された．

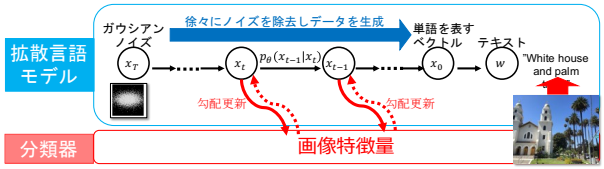

図 1: 拡散モデルを用いたキャプション生成の概要図

表 1: 拡散言語モデルと分類器の統合実験結果

統合	BLEU-1	ROUGE-L	パラメータ数
	0.627	0.524	80M
√	0.569	0.459	39M

2.2 手法 2 - ノイズ除去ネットワークの最適化

DLM 内の DNN として以下の 3 つを試行し，精度向上を目指した．

- Transformer6 層 [3]
  汎用的な構造によってノイズ除去という目的タスクに特化したモデルが構築されるか検証
- 事前学習済みでない BERT [4]
  自然言語処理タスク念頭に設計されているため，文脈を捉えるノイズ除去を期待
- 事前学習済みでない GPT-2 [5]
  サイズの大きいデコーダモデルの使用による精度への影響を調査

結果 結果 (表 2) より，Transformer6 層を DNN に採用した場合が最も良い精度を達成した．汎用的かつシンプルな構造により，ノイズ除去タスクに特化した調整が容易だったと考察される．またモデル規模が小さく，学習コストが低いことも有利に働いたと推察される．

2.3 手法 3 - 分類器の有無による性能評価

拡散モデルの制御手法には，ここまで採用してきた分類器を導入する方法と，分類器を使用しない方法がある．それぞれの精度を比較する．分類器を使用しない場合，DLM 内の DNN に制御条件である画像特徴量を渡し，ノイズの乗っていない状態を復元 (推定) させることで，入力画像に応じた制御を行う．

結果 実験結果から，分類器を使用した方が精度がわずかに高いことがわかる (表 3)．分類器を使わない場合，画像特徴量は DNN に入力されるが，ネットワーク内で十分に活用されていない可能性がある．一方使用する場合は，分類器が DNN の出力から重要な特徴を選別・強化し，その出力を補完する役割を果たすと考えられる．

3 トレースベースの意図反映キャプション生成

近年，画像認識と自然言語処理の発展により，高品質なキャプション生成が可能となっているが，生成さ

表 2: DNN 最適化の実験結果

DNN	BLEU-1	ROUGE-L	BERTScore
Transformer6 層	0.670	0.525	0.721
BERT	0.653	0.514	0.711
GPT-2	0.654	0.504	0.690

表 3: 分類器の有無による実験結果

分類器	BLEU-1	ROUGE-L	BERTScore
√	0.653	0.514	0.711
	0.650	0.498	0.696

れるキャプションは画像全体の大まかな説明にとどまることが多い．本研究ではユーザが画像上をなぞった軌跡であるトレースを制御信号として活用し，座標や滞在時間などの多様な情報をもとに，個人の嗜好や視点を反映した柔軟なキャプション生成の実現を目指す．特にトレースの密集度からユーザの関心領域を特定，座標変化から説明の順序を決定，さらに滞在時間から各領域への関心の高さを評価して，生成キャプション内に表現する．

実験　まずは，マウスで画像をなぞりながら口頭で説明を加えた画像アノテーションデータセットである Localized Narratives(LN) [6] を基にデータセットを構築する．具体的にはペン先の座標変化量を解析し，トレースを画像内の特定領域を説明している区間と，次の説明領域への移動区間に分割し，画像から一連のバウンディングボックス (BBox:Bounding Box) を抽出する．時間情報をもとに発語キャプションを区切り，各 BBox に対応付けることで，新たなデータセットを構築する．

3.1　非自己回帰型生成 (NAR)

DNN の出力である言語特徴量 $\tilde{\mathbf{x}}_0$ の配列全体から画像全体の特徴量の予測を行っていたのを，言語特徴量の各 BBox に対応する配列の部分から各 BBox の画像特徴量を線形回帰で予測するように変更を加えることで，各ユーザ固有の意図を生成文内に表現することを目指す．

結果　表 4 よりトレース情報を活用して生成を制御する提案手法が自己回帰モデル (AR) を超える高精度を達成することが示された．生成例においても (図 2)，ユーザが説明した物体がトレース順序に従い，ペン先の滞在時間が長い領域についてより詳細に記述されていることが確認できる．一方，例 2 では NAR 言語モデル特有の文法誤りが一部見受けられ，さらなる改良の余地があることが示唆される．

3.2　半自己回帰型生成 (SAR)

NAR 型生成手法における流暢性不足を改善するため，自己回帰と非自己回帰の両者の利点を活かす半自己回帰型 (SAR:Semi-Autoregressive) 手法を提案する [7]．トレースから指定された領域ごとに順次生成処理を行い，文脈を考慮した流暢なテキスト生成を目指す．

結果　表 4 の最下段より SAR 型生成にすることで発語キャプションとの類似度を測る指標において高い精度向上を達成したことがわかる．一方でトレースの順番通りに各物体が説明されているか評価する F 値に関しては NAR 型手法に及ばない結果となった．流暢性と制御性のトレードオフの発生の可能性が考えられる．

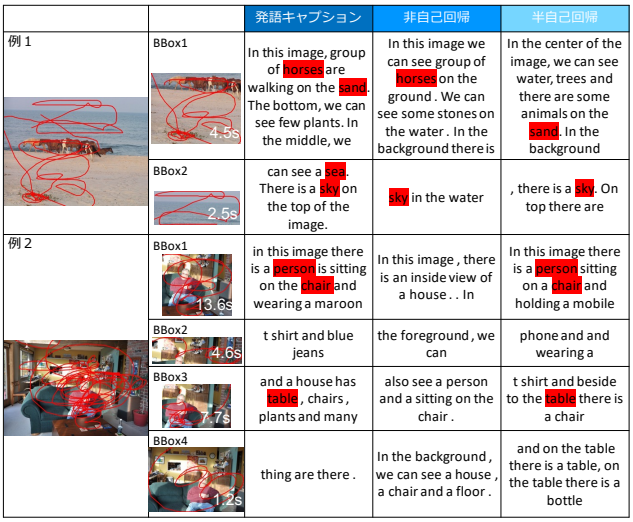

図 2: 生成キャプション例

表 4: トレースベースの意図反映キャプション生成の実験結果

手法	BLEU-1	ROUGE -L	BERT Score	CLIP Score	F 値
AR	0.0950	0.1135	0.4991	23.64	0.0141
NAR	0.1866	0.2109	0.5578	23.41	**0.2568**
SAR	**0.2165**	**0.2138**	**0.5706**	**23.74**	0.2199

4　まとめ

本研究では拡散モデルを自然言語処理タスクの一つであるキャプション生成に適用し，高い質・制御性・多様性を実現することを目指した．拡散言語モデルの改良やさまざまな制御手法を検討し，どのような条件下で拡散モデルが精度高く条件付きテキスト生成を可能にするか検証した．また，画像を見ているユーザに特化したキャプション生成手法の開発にも取り組み，拡散言語モデルの用途拡張や多様な応用可能性を示した．今後はさらなる用途拡大や，自己回帰型生成への改良による精度向上に取り組みたい．

参考文献

[1] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. *ArXiv*, 2022.

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. Association for Computational Linguistics, 2019.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[6] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.

[7] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. Toronto, Canada, 2023. Association for Computational Linguistics.