離散と連続の交互学習に基づく拡散言語モデルの開発

武井まりあ [1]　小林一郎 [1]
[1] お茶の水女子大学
{g2220527,koba}@is.ocha.ac.jp

概要

近年，拡散言語モデルを用いた言語生成が注目されている．拡散言語モデルには，状態を連続値で表現する連続型拡散言語モデル，及び，状態をトークンとして扱う離散型拡散言語モデルの二つがある．しかし，連続型には丸め込む際の誤差に，離散型には大域的意味構造の表現にそれぞれの課題を持つ．一方で，連続型には状態の微量かつ滑らかな変化を捉えることが可能であり，離散型には言語モデルの回帰的な性質を表現しやすいなどの利点が存在する．そこで，本研究では両者のモデルを交互に学習させることでそれぞれの長所を取り込み，短所を補えることを可能にするハイブリッド拡散言語モデルを提案する．提案手法では，拡散時間に応じてモデルと損失関数を切り替えることで，トークンレベルの正確性と文全体の意味的一貫性の両立を図る．結果として、連続型拡散言語モデル単体を上回る性能を実現できたが、離散型拡散言語モデル単体には及ばず、改善の余地がある．

1　はじめに

拡散モデルは画像生成分野での成功を背景に，言語生成への応用も活発に研究されている [1][2]．言語生成における拡散モデルは，拡散過程を連続空間で定義する手法と，離散空間で直接定義する手法に大別される．連続型拡散モデルは，トークン列を連続表現に埋め込みガウスノイズによる拡散を行うことで，理論的に安定した学習と文全体の意味構造を捉えやすいという利点を持つ．一方で，生成時には連続表現から離散トークンへの復元が必要となり，丸め込み誤差や不自然な語彙選択が生じる可能性がある．これに対し，離散拡散モデルはトークン空間上で直接拡散を行うため，生成結果のトークンレベルの正確性が高いが，長距離依存関係や文全体の大域的意味構造のモデル化が難しいという課題を持つ．このように，連続型拡散と離散拡散は補完的な性質を持つが，両者を統合的に扱う枠組みは十分に検討されていない．

2　関連研究

確率過程を生成に用いることが物理モデリングで発展し，Sohl-Dickstein ら [3] によってノイズを段階的に入れて元データを再現する生成モデルが提案された．その後，Ho ら [4] によってガウスノイズを順方向で加え，逆方向でノイズを除去する生成モデルが提案され，それまでの敵対的生成ネットワーク（GAN）における生成画像と同品質の画像生成が可能になった．また，Rombach ら [5] によって提案された潜在空間における拡散過程において画像を生成する Latent Diffusion Model である Stable Diffusion によって高解像度の画像生成が可能となり，さらに拡散過程の損失関数にそれまで使用されていた識別器を不要とする Classifier-free guidance の手法 [6] が提案されるなどによって，高品質な画像生成における制御性の向上が実現された．Li ら [1] は離散情報である単語列に拡散過程の考えを導入した初期の言語モデルを提案した．彼らの非回帰型モデルに対して，本来の言語モデルが持つ自己回帰の性質を導入した研究として，ブロック単位で自己回帰的にテキストを生成しつつ，拡散過程を実施するもの [7] やブロックの双方向自己回帰を実施するもの [8] などが提案されている．一方で，単語トークンの本来の性質をそのまま反映して拡散モデルにおける状態を連続から離散とし，ノイズをトークンのマスクとして表現する種々の研究 [9, 10, 11, 11] が進められており，連続型のモデルの精度を上回る報告がなされている．また，近年では連続型の意味表現と離散型トークン生成を同時に扱うハイブリッド型拡散言語モデルが提案されている [12, 13]．これらのハイブリッド型は連続型と離散型を同時並行して拡散過程を行うものであり，2 つのモデルを同時に利用する必要がある．そこで，本研究では離散型の拡散モデルを主としつつ，離散型トークン生成に適度に連続

値へと変換し拡散過程の状態をアニーリング可能にする，離散と連続を直列し交互に入れ替える新しいハイブリッド拡散言語モデルを提案する．

3 研究概要

3.1 連続型拡散モデル

連続型拡散モデルは，離散トークン列 $x_0 = (x_0^1, \ldots, x_0^L)$ を埋め込みにより連続表現へ写像し，$\mathbf{z}_0 \in \mathbb{R}^{L\times d}$ 上でガウスノイズによる拡散過程を定義する．前向き過程は，時刻 $t \in [0,1]$ において

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}\Big(\alpha(t)\mathbf{z}_0,\ \sigma^2(t)\mathbf{I}\Big) \tag{1}$$

で与えられる（$\alpha(t), \sigma(t)$ はノイズスケジュール）．このとき，標準的な学習はノイズ予測として式 (2) の損失関数 $\mathcal{L}_{\text{cd}}$ を用いて行われる．

$$\mathcal{L}_{\text{cd}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\big\|_2^2\Big], \tag{2}$$

$$\mathbf{z}_t = \alpha(t)\mathbf{z}_0 + \sigma(t)\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

生成時は $\mathbf{z}_t$ を逆拡散により $t : 1 \to 0$ で逐次復元し，最終的な連続表現 $\mathbf{z}_0$ を最近傍探索等により離散トークン列へ射影することで文を得る．

3.2 離散型拡散モデル

離散型拡散モデルは，語彙集合 $\mathcal{V}$ 上のトークン列 $x_0 = (x_0^1, \ldots, x_0^L)$ に対して，離散状態空間上で直接拡散過程を定義する．各時刻 $t \in [0,1]$ における前向き過程は，連続時間マルコフ過程として

$$\frac{dp_t}{dt} = Q_t\, p_t, \qquad p_0 \approx p_{\text{data}}. \tag{3}$$

で与えられる．ここで $\mathbf{Q}$ はトークン置換を定義する拡散行列である．

Lou ら [14] が提案した Score Entropy Discrete Diffusion models (SEDD) では，逆拡散に必要なスコアとして，周辺分布の比率

$$s_\theta(x_t, t)_y \approx \log \frac{p_t(y)}{p_t(x_t)} \tag{4}$$

をニューラルネットワークにより近似する．学習は，スコアエントロピー損失を用いて

$$\mathcal{L}_{\text{SEDD}} = \mathbb{E}_{x\sim p}\left[\sum_{y\neq x}\left(w_{xy}\, s_\theta(x)_y - \frac{p(y)}{p(x)}\log s_\theta(x)_y + K\frac{p(y)}{p(x)}\right)\right] \tag{5}$$

として定式化される．生成時には，推定されたスコアに基づき，τ-leaping 等の手法を用いて複数トークンを同時に更新することで効率的に逆拡散を行う．

3.3 提案手法：ハイブリッド拡散言語モデル

本研究では，連続型拡散モデルと離散型拡散モデルの利点を統合するハイブリッド拡散言語モデルを提案する．提案手法では，拡散過程を単一のグローバル時間 $u \in [0,1]$ で管理し，これを三つの区間に分割する：

$$[0,a), \quad [a,b), \quad [b,1], \qquad 0 \le a < b \le 1. \tag{6}$$

$u \in [0,a)$ の初期区間では，離散型拡散モデル（SEDD1）を用い，クリーンなトークン列 x_0 を直接予測するように学習を行う．中間区間 $u \in [a,b)$ では，トークン列を連続表現に写像し，連続型拡散モデル（Diffusion-LM） [1] を用いて，境界時刻に対応する連続表現 z_a の予測を行う．最後の区間 $u \in [b,1]$ では，再び離散型拡散モデル（SEDD2）を用い，連続型拡散から得られた表現に整合するトークン列 x_b を予測する．

学習時には，サンプリングされた拡散時間 u に応じて使用するモデルおよび損失関数を切り替える．

このように，連続・離散の拡散過程を時間方向に明示的に接続することで，文全体の意味的一貫性とトークンレベルの正確性の両立を図る．図 1 に，提案手法の全体構成を示す．拡散時間の初期および終盤に離散拡散を配置することでトークンレベルの正確性を確保し，中間区間に連続型拡散を挿入することで文全体の意味構造を効率的に学習する．

4 実験

4.1 実験設定

本節では，提案手法の有効性を検証するためのデータセット，学習設定，および評価方法を述べる．

データセット　学習には OpenWebText[1] を用い，評価には WikiText[2] を用いた．OpenWebText は Web 由来の大規模テキストからなるデータセットであり，WikiText は Wikipedia を整形した言語モデリング用データセットである．学習と評価で異なるコーパスを用いることで，提案手法の汎化性能を確認する．

前処理　トークン化には GPT-2 系トークナイザを用い，系列長を $L = 64$ に固定した．各文書はトーク

1） https://huggingface.co/datasets/Skylion007/openwebtext
2） https://huggingface.co/datasets/Salesforce/wikitext

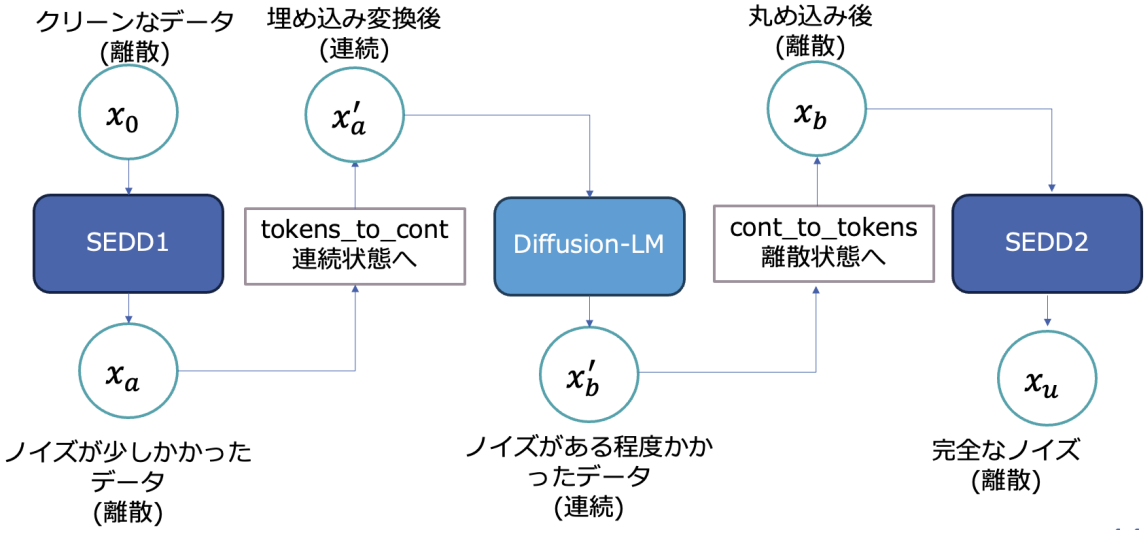

図 1　提案するハイブリッド拡散言語モデルの概念図．グローバル時間 u を三つの区間に分割し，初期および終盤では離散型拡散モデル（SEDD）[14]，中間区間では連続型拡散モデル（Diffusion-LM）[1] を用いる．

ン列へ変換後，長さ L のチャンクに分割して学習に供した（短い列はパディングし，損失計算ではマスクする）．

学習設定　提案手法は，グローバル時間 $u \in [0, 1]$ をサンプリングし，u が属する区間に応じて (i) 離散拡散（SEDD1），(ii) 連続型拡散（Diffusion-LM），(iii) 離散拡散（SEDD2）のいずれかを選択して更新する．区間境界 a, b は $0 \le a < b \le 1$ を満たすハイパーパラメータであり，$u < a$ ではスコアエントロピー損失，$a \le u < b$ では連続型拡散のノイズ予測損失，$u \ge b$ では境界状態を予測対象とする離散拡散損失を用いる．最適化には AdamW を用い，学習率，バッチサイズ，学習ステップ数などの詳細は表 1 にまとめる．

ベースライン　提案手法の効果を明確にするため，以下のベースラインと比較する：(i) 連続型拡散のみ（Diffusion-LM），(ii) 離散拡散のみ（SEDD）．

評価指標　言語モデルの生成品質評価として，Perplexity（PPL）を用いる．PPL は評価コーパス上の負の対数尤度から算出する．

表 1　実験に用いた主なハイパーパラメータ．

項目	設定
学習データ	OpenWebText
評価データ	WikiText
トークナイザ	GPT-2 tokenizer
系列長	$L = 64$
最適化手法	AdamW
学習率	1×10^{-4}
バッチサイズ	32
学習ステップ数	2000
時間分割	$[0, a), [a, b), [b, 1]$
a, b	$a = 0.35,\ b = 0.75$

表 2　各モデルのパープレキシティ．

Model	PPL
SEDD	49.60
Continuous Diffusion	209.93
Hybrid Diffusion	203.20

4.2　実験結果

表 2 は，各モデルの生成性能をパープレキシティ（PPL）により比較した結果を示している．離散型拡散モデルである SEDD は PPL = 49.60 と最も低い値を示し，言語モデリングにおいて高いトークン予測精度を達成していることが分かる．一方，連続型拡散モデルは PPL = 209.93 と高く，連続空間での生成に起因する量子化誤差やトークン単位での整合性の難しさが反映されていると考えられる．提案するハイブリッド拡散モデルは PPL = 203.20 と，連続型拡散モデルと比較してわずかながら改善が見られたものの，SEDD には及ばなかった．この結果は，ハイブリッドモデルが連続型拡散の欠点を部分的に緩和している一方で，トークンレベルの尤度最適化という点では離散型拡散モデルが依然として有利であることを示唆している．

4.3 考察

表 2 の結果から，ハイブリッド拡散モデルは連続型拡散モデルと比較して一定の改善を示すものの，SEDD には及ばない性能であることが確認された．この差は，ハイブリッドモデルにおいて連続表現から離散トークンへ切り替える際の処理に起因している可能性がある．特に，埋め込み空間から離散トークン空間への射影は近似的に行われることが多く，この過程で情報の損失が生じ得る．今後の改善としては，連続表現と離散トークンの整合性を高めるより精緻な変換手法の導入や，変換誤差を明示的に考慮した学習目標の設計が有効であると考えられる．

5 まとめ

本研究では，連続型拡散モデルと離散型拡散モデルの利点を統合するハイブリッド拡散言語モデルを提案した．提案手法は，拡散過程を単一の時間軸で管理し，初期および終盤に離散拡散，中間に連続型拡散を配置することで，文全体の意味的一貫性とトークンレベルの正確性の両立を目指す．今後の課題として，連続型拡散から離散型拡散へ切り替わる際の表現変換に伴う誤差の低減が挙げられる．本研究で得られた結果から，埋め込み空間から離散トークン空間への写像がモデル性能に大きく影響する可能性が示唆された．このため，連続表現と離散トークンの対応関係をより厳密に学習過程へ組み込む手法や，変換誤差を明示的に考慮した学習目標の設計が重要であると考えられる．また，離散・連続拡散の切り替えタイミングや時間配分についても，タスクやデータ特性に応じた最適化が必要であり，これらの点を検討することで，ハイブリッド拡散言語モデルのさらなる性能向上が期待される．

参考文献

[1] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In **Proceedings of the 36th International Conference on Neural Information Processing Systems**, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[2] Tong Wu, et al. Ar-diffusion: auto-regressive diffusion model for text generation. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[3] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In **Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37**, ICML'15, p. 2256–2265. JMLR.org, 2015.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 10684–10695, 2022.

[6] Jonathan Ho. Classifier-free diffusion guidance. **ArXiv**, Vol. abs/2207.12598, , 2022.

[7] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11575–11596, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Yonggan Fu, Lexington Whalen, Zhifan Ye, Xin Dong, Shizhe Diao, Jingyu Liu, Chengyue Wu, Hao Zhang, Enze Xie, Song Han, Maksim Khadkevich, Jan Kautz, Yingyan Celine Lin, and Pavlo Molchanov. Efficient-dlm: From autoregressive to diffusion language models, and beyond in speed, 2025.

[9] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024.

[10] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025.

[11] Michael Hersche, Samuel Moor-Smith, Thomas Hofmann, and Abbas Rahimi. Soft-masked diffusion language models, 2025.

[12] Bocheng Li, Zhujin Gao, and Linli Xu. Unifying continuous and discrete text diffusion with non-simultaneous diffusion processes. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11530–11551, Vienna, Austria, July 2025. Association for Computational Linguistics.

[13] Cai Zhou, Chenxiao Yang, Yi Hu, Chenyu Wang, Chubin Zhang, Muhan Zhang, Lester Mackey, Tommi Jaakkola, Stephen Bates, and Dinghuai Zhang. Coevolutionary continuous discrete diffusion: Make your diffusion language model a latent reasoner, 2025.

[14] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In **Proceedings of the 41st International Conference on Machine Learning**, ICML'24. JMLR.org, 2024.