アンカー表現の適応的選択に基づく拡散言語モデル

佐藤真 [1]　江口浩二 [2]
広島大学 [1] 広島大学大学院 [2]
{b220083, kxeguchi}@hiroshima-u.ac.jp

概要

拡散言語モデルは，ノイズを除去する過程を学習するノイズ除去拡散確率モデル（DDPM: Denoising Diffusion Probabilistic Models）に基づく大規模言語モデル（LLM: Large Language Models）を指し，現在主流となっている自己回帰型の LLM における累積誤差などの課題を解決しうる手段として注目されつつある．とりわけ，ノイズ除去ネットワークに加えて，生成条件を与えるためのアンカーネットワークをともなった，アンカー拡散言語モデル（ADLM: Anchor Diffusion Language Model）が最近提案され，その有効性が報告されている．そこではアンカーネットワークの構成要素であるアンカー表現の選択基準が固定的かつアドホック的であるため改善の余地が認められる．そこで，本研究では，アンカー表現の適応的選択をともなう拡散言語モデルについて検討する．

1　はじめに

近年，大規模言語モデル（LLM: Large Language Models）の成長が著しく，とりわけ自己回帰型の LLM （autoregressive (AR) models）は高品質のテキスト生成や推論能力を達成している．GPT-3[1] や Gemini[2]，LLaMA[3]，Claude[4] に代表される AR モデルは，テキスト生成において逐次的に左から右への一方向のトークン生成を行っている．これはそれ以前のトークンをもとに次のトークンを予測するものであり，このモデルはシーケンス全体の情報を利用することができない．また逐次生成により遅延や累積誤差が生じる問題がある．

AR モデルの代替モデルとして，拡散言語モデル（DLMs: Diffusion Language Models）が注目されている．これはノイズを除去する過程を学習するノイズ除去拡散確率モデル（DDPM: Denoising Diffusion Probabilistic Models）[5] に基づく大規模言語モデルである．拡散言語モデルは，マスクされたトークンを予測する過程を繰り返し学習することにより，AR モデルにおける前述の課題の解決に資することが期待される．

しかしながら，拡散言語モデルはノイズを付与する拡散過程において，早い段階で重要なトークン（出現回数が低頻度のトークンや意味的に重要なトークン）がマスクされてしまうと，文脈の意味が失われ，もとの文章を正確に再現することが難しくなる．これを解決する手段として，ノイズ除去ネットワークに生成条件を与えるためのアンカーネットワークをともなった，アンカー拡散言語モデル（ADLM: Anchor Diffusion Language Model）[6] が最近提案され，その有用性が報告されている．ADLM では，重要語であるアンカー表現の選択基準として，一定長のシーケンスにおける単語の出現頻度が一定数以下であることが仮定されていた．この基準は固定的でアドホック的であるため，改善の余地が認められる．局所的な低頻度語が必ずしも重要な単語とは限らず，大域的に低頻度かつ局所的に高頻度であるような語が重要な単語であることが経験的に知られている [7, 8]．

本研究では，TF-IDF を用いた重要語スコアと学習可能な閾値を利用した，より適応的かつ効果的なアンカー選択を行う手法を提案する．以下，本提案手法を AdaADLM（Adaptive ADLM）と呼ぶ．

2　関連研究

2.1　アンカー拡散言語モデル

ADLM[6] は，アンカーネットワークとデノイジングネットワークの２つから構成される．アンカーネットワークは，逆拡散過程においてマスクされたトークンから，重要なトークンの尤度を予測するネットワークである．ADLM の概念図を図 1 に示す．

これは，既存の拡散モデルにおいて早い段階で重要なトークンがマスクされる問題に関して，ADLM

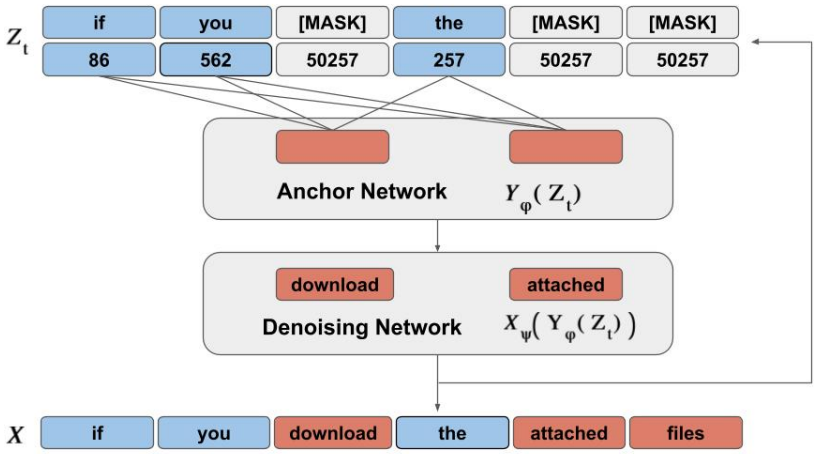

図 1　ADLM の概念図．サンプリングステップ t において，アンカーネットワークが潜在変数 Z_t のマスクトークンからアンカーを予測する．予測されたアンカー情報をヒントに，デノイジングネットワークがマスクトークンを予測する．

では 2 層のネットワークを用いることにより抑制できる．すなわち，まずアンカーネットワークがアンカー（重要語）を予測し，次にそのアンカー予測結果をデノイジングネットワークが受け取ってマスクされたトークンを予測する．このような構造を持つ ADLM により，既存の AR モデルや拡散言語モデルに比して，より高品質なテキストを生成できることが報告されている [6].

しかし ADLM におけるアンカーネットワークでは，一定長のシーケンスでの出現頻度が 5 回以下の単語をアンカーとして選択しており，固定的でアドホック的である．データセットによりシーケンスのテキスト長が異なるため，この固定的な指標はデータセットの性質を無視している．Rout ら [6] による ADLM の実験では，OWT（OpenWebText）[9] と LM1B[10]（One Billion Words）の 2 つのデータセットを利用していた．これらのデータセットのテキスト長はそれぞれ経験的に 1024 と 128 と設定されており，アンカーであると選択されるトークンは，データセットのテキスト長に大きく依存すると考えられる．

2.2　テキストにおける重要語抽出

テキストにおける重要語抽出手法として，TF-IDF（Term Frequency-Inverse Document Frequency）がよく利用される [7, 8]．TF-IDF は，TF(Term Frequency) と IDF(Inverse Document Frequency) の積であり，それぞれ単語出現頻度，逆文書頻度（目的の単語の文書頻度の逆数を対数化したもの）である．TF-IDF は，データセット全体における単語の出現状況と各文書における単語の出現状況を両方考慮して算出されるスコアであり，単語の重要語抽出において広く用いられてきた手法である．

3　提案手法

ADLM において，アンカー選択基準は単語の出現頻度ベースで固定的であった．ここでは，TF-IDF を用いた重要語スコアと学習可能な閾値を導入した手法を提案する．

アンカー選択基準を式 (1) で定義する．

$$\text{tf-idf}\,(\text{token}) \le \tau \tag{1}$$

ここで，tf-idf (token) は一定長の各シーケンスを文書とみなし，各シーケンス内の特定の token の TF-IDF[7, 8] を求める関数であり，τ はアンカー（重要語）選択の閾値とする．tf-idf (token) は通常の TF-IDF スコアを，事前にデータセットから算出した平均・分散によって標準化したスコアを指す．

また，閾値 τ を学習可能なパラメータに設定し，ADLM において固定的でアドホック的な閾値に依拠する問題に対処する．

学習時の目的関数は，ADLM と同じアンカー付き負の変分下界（ANELBO: Anchored Negative Evidence Lower Bound）とし，式 (2) で定義する．

$$\begin{aligned} &L_{\text{ANELBO}}(\mathbf{x}, \mathbf{y}; \varphi, \psi) \\ &= \mathbb{E}_{Z_0 \sim q(\cdot|\mathbf{x})}\left[-\log p_\psi(\mathbf{x} \mid \mathbf{y}_\varphi(Z_0))\right] \\ &\quad + \sum_{i=1}^{T} \mathbb{E}_{Z_{t(i)} \sim q(\cdot|\mathbf{x})}\left[\frac{(1-\sigma_{t(i)})\alpha_{t(i)} - \alpha_{s(i)}}{1-\alpha_{t(i)}}\right. \\ &\quad \left. \times \sum_{l=1}^{L} \log\langle \mathbf{x}_\psi^l(\mathbf{y}_\varphi(Z_{t(i)})), \mathbf{x}^l\rangle + \gamma \log\langle \mathbf{y}_\varphi^l(Z_{t(i)}), \mathbf{y}^l\rangle\right] \end{aligned} \tag{2}$$

ここで，$\mathbf{x}$ は入力シーケンス，$\mathbf{y}$ はアンカートークンの混合分布，φ はアンカーネットワークのパラメータ，ψ はデノイジングネットワークのパラメータである．T はタイムステップの総数であり，$i \in \{1, 2, \ldots, T\}$ において，$t(i) = \frac{i}{T}, s(i) = \frac{i-1}{T}$ と定義する．$Z_{t(i)}$ はタイムステップ $t(i)$ における潜在変数，$q(\cdot|\mathbf{x})$ は拡散過程の分布，$p_\psi(\mathbf{x} \mid \mathbf{y}_\varphi(Z_0))$ は逆拡散過程の分布，α_t はマスキングスケジュール，σ_t はリマスキング確率，γ はアンカーの強さを示す指標である．

目的関数の $\log\langle \mathbf{x}^l_\psi(\mathbf{y}_\varphi(Z_{t(i)})), \mathbf{x}^l \rangle$ は，アンカーネットワークの出力をもとにトークンの尤度を評価するデノイジングネットワークの項であり，$\log\langle \mathbf{y}^l_\varphi(Z_{t(i)}), \mathbf{y}^l \rangle$ は早期段階で重要トークンを予測するアンカーネットワークを監視する項である．

4 実験設定

データセット 本実験ではベンチマークとして One Billion Words（LM1B）を使用する．評価においては，LM1B の標準的な test split を用いる．

モデル設定 ADLM と AdaADLM の 2 つのモデルを使用する．ADLM の実験設定に合わせてバッチサイズは 512，モデルのテキスト長は 128，トークナイザーは BERT-base-uncased[11] を用いる．また，モデルの学習の終了基準は Perplexity の減少率が 0.5% 以下の状態が 3 回連続で続いたときとし，アンカーの強さを示す指標は $\gamma = 3e - 3$，アンカー選択の閾値 τ の初期値は 1.0 とする．

評価指標 次の 3 つの評価指標を用いる．

- **Test Perplexity (PPL)**：モデルが未知のテキストをどれだけ正確に予測できるかを示す指標．低いほどよい．
- **Generative Perplexity (Gen PPL)**：外部の学習済みモデル（ここでは GPT-2 Large を使用）に，モデルが生成したテキストを入力した際の Perplexity．低いほど良い．生成テキストの品質を評価する指標である．
- **Zero-shot Perplexity**：学習していないドメインのデータに対する Perplexity．これにより汎用性を測る．

表 1 Test perplexity

model	PPL (↓)
ADLM	75.81
AdaADLM	**48.65**

5 実験結果・考察

5.1 尤度推定における評価

AdaADLM の Test Perplexity を評価した．表 1 に，LM1B における Test Perplexity の結果を示す．これより，AdaADLM が ADLM よりも高い性能を示した．AdaADLM は 35.8% の精度向上を達成しており，アンカーの適応的選択により精度が向上した．

5.2 生成テキストの品質評価

事前学習した AdaADLM のサンプリングステップ数 N を変化させ，Generative Perplexity を評価した．サンプリング数は，$N = \{128, 256, 512, 1024, 2048, 4096\}$ とする．また，同時に生成テキストのエントロピーも評価する．表 2 に，Generative Perplexity とエントロピーの結果を示す．

結果としてすべてのサンプリングステップにおいて，AdaADLM が ADLM よりも高い性能を示した．どちらの手法もサンプリングステップの増加に伴い，Perplexity は減少しており，とりわけサンプリングステップ数が最も大きい $N = 4096$ では，AdaADLM は非常に低いスコアを達成した．

またエントロピーについては，サンプリングステップ数が大きくなるほど，ADLM よりも高いエントロピーを維持しており，性能を維持していることが示された．

5.3 ゼロショット汎化性能

他ドメインのデータセットを利用した，学習済み AdaADLM のゼロショット生成における Perplexity を評価した．ここでは，Lambada[12]，PTB[13]，Wikitext[14]，AG News[15]，PubMed[16]，ArXiv[17] の 6 つのデータセットを用いる．表 3 に，ゼロショット汎化性能の結果を示す．

結果として，すべてのデータセットにおいて，AdaADLM が ADLM よりも高い性能を示した．これにより，AdaADLM は学習した分布が変動してもロバスト性を示しており，高い汎用性を示している．

表 2 Generative Perplexity and Entropy

	Gen PPL. (↓)			Entropy (↑)		
	N=1024	N=2048	N=4096	N=1024	N=2048	N=4096
ADLM	204.9	148.7	98.3	4.148	4.039	3.856
AdaADLM	**131.3**	**106.6**	**85.2**	**4.176**	**4.136**	**4.078**
	N=128	N=256	N=512	N=128	N=256	N=512
ADLM	364.0	312.7	258.7	**4.284**	**4.258**	**4.215**
AdaADLM	**218.9**	**158.7**	**157.7**	4.248	4.224	4.202

表 3 Zero-shot Perplexity (↓)

	Lambada	PTB	Wikitext	AG News	PubMed	ArXiv
ADLM	261.38	383.58	350.84	415.55	521.63	657.76
AdaADLM	**185.61**	**266.45**	**227.19**	**264.32**	**261.32**	**438.53**

特に，学術的な分野の理解を測る PubMed と ArXiv において，ADLM は他のデータセットと比較して高い Perplexity を示すのに対して，AdaADLM はそれぞれ 49.9%，33.3%の精度向上が達成された．

6 おわりに

本研究では，ADLM の固定的かつアドホック的なアンカー選択基準を，重要語抽出手法である TF-IDF を用い，学習可能なパラメータとして学習することで，AdaADLM の有用性を示した．AdaADLM は，モデルの生成テキスト品質や他ドメインのデータセットに対しても高い性能を発揮することが示された．

今後は，アンカー選択基準の改善やより大きなデータセットでの性能評価などを行い，AdaADLM の有用性をさらに高めることにより，拡散言語モデルの性能向上に寄与することが期待される．

謝辞

本研究の一部は科学研究費補助金基盤研究 (C)（23K11231）の援助による．

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[4] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. **Claude-3 Model Card**, Vol. 1, p. 1, 2024.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. **NeurIPS**, 2020.

[6] Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Anchored Diffusion Language Model. **NeurIPS**, 2025.

[7] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. **Introduction to Information Retrieval**. Cambridge University Press, 2008.

[8] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. **Modern Information Retrieval**. Addison Wesley, 1999.

[9] Aaron Gokaslan and Vanya Cohen. OpenWebText Corpus. 2019.

[10] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. **arXiv preprint arXiv:1312.3005**, 2013.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina

Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[12] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1525–1534, 2016.

[13] Mitch Marcus, Beatrice Santorini, and Mary-Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. **Computational linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[14] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. **International Conference on Learning Representations**, 2017.

[15] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. **Advances in neural information processing systems**, Vol. 28, , 2015.

[16] Arman Cohan, Franck Dernoncourt, Doo-Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 615–621, 2018.

[17] Arman Cohan, Franck Dernoncourt, Doo-Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 615–621, 2018.