LoRA-Adapted Diffusion

Ruurd J.A. Kuiper[1], Lars de Groot[1], Bram van Es[2], Maarten van Smeden[1], Ayoub Bagheri[2]

[1] University Medical Center Utrecht, Utrecht
[2] Utrecht University
r.j.a.kuiper@umcutrecht.nl

1 Introduction

Large language models (LLMs) predominantly use an autoregressive (AR) paradigm for text generation, which, despite its success, is limited by its left-to-right decoding constraints. This can impact efficiency and prevents bidirectional reasoning during generation reasoning [1–4]. Diffusion models, which have shown state-of-the-art performance in continuous domains like image synthesis [5], offer a promising non-autoregressive alternative. However, adapting them for discrete text data presents significant challenges, such as decreased inference and training efficiency [1, 6]. Our work, LoRA-Adapted Diffusion (LAD), addresses some of these challenges by proposing a novel, parameter-efficient framework that adapts existing AR LLMs for non-autoregressive generation using Low-Rank Adaptation (LoRA) [7]. We introduce a structural noising objective that combines masking with text perturbations (swaps, duplications, and span shifts), enabling full-sequence editing during generation and unifying diffusion adaptation and instruction-tuning. The main objective of this work is to demonstrate that LAD can be a viable and efficient alternative to training diffusion models from scratch, with minimal computational cost. The full paper version of this abstract is available from https://openreview.net/pdf?id=h8Zo1K6zpY, and two interactive demos are available online through: https://ruurdkuiper.github.io/tini-lad/

2. Methods

We developed LAD by fine-tuning pre-trained Llama 3.2 1B/3B and Llama 3.1 8B models [8]. To enable bidirectional generation, we replaced their causal attention mechanisms with bidirectional masks. We applied LoRA to the query and value projections, freezing all other weights to ensure parameter-efficient fine-tuning. We trained six variants of LAD with varying model sizes and LoRA ranks to study the effects of scaling. The models were trained on a diverse instruction-finetuning dataset of 951k examples from various domains, including question-answering, multiple-choice, math and coding.

Our unique training strategy incorporates a hybrid noising scheme. Instead of relying solely on masking as in traditional Masked Diffusion Models (MDMs) [4, 9–11], we introduced a structured corruption process involving token swaps, duplications, and span shifts. This trains the model to correct common structural errors observed during iterative refinement. This combined approach allows the model to denoise from a fully masked input, but makes it also able to refine partially coherent sequences without masking. For inference, we developed two modes: a scheduled denoising approach with intermediate re-masking, and a self-refining approach that omits re-masking, allowing the model to continuously refine its output without losing information.

3. Results

The training loss curves showed that models with larger parameter sizes and higher LoRA ranks achieved lower final cross-entropy loss, indicating better convergence. The training process was highly efficient, with the 8B model being trained on 200 million tokens in just 16.5 hours on a single NVIDIA A100 GPU. For comparison, the base model LLaMA 3.1-8B was trained on 15 trillion tokens, and comparable diffusion models such as LlaDa [4], used 2.3 trillion tokens. This means LAD used just 0.001% and 0.008% of their respective token counts.

Benchmark evaluations on ARC-Easy, MMLU, ARC-Challenge, and HellaSwag demonstrated that LAD models performed comparably to their autoregressive base models. However, they showed significantly lower accuracy on the complex reasoning tasks in GSM8K. We also observed that increasing the number of iterations at inference led to a decrease in perplexity and an improvement in fluency, without significant changes in lexical diversity as measured by the distinct 2-gram fraction. Finally, our qualitative analysis showed that the self-refining, no-remasking inference method is faster but can sometimes result in less polished outputs compared to the remasking approach.

4. Discussion

This study showed that LAD successfully adapts pretrained AR LLMs into diffusion models with efficient training and competitive performance. The structural noising strategy enabled text refinement without remasking. The ability to decide between test time compute and quality, controlled by the number of iterations, is a key feature of diffusion networks. While LAD showed competitive performance on some tasks, it has as of yet limited capabilities on complex reasoning and coding benchmarks. Future work should also focus on exploring the bidirectional reasoning capabilities of the model and conducting broader comparative evaluations. In conclusion, LAD could be an efficient alternative for scalable, controllable, and efficient diffusion models for text generation.

References

1. Zhu, Y., Zhao, Y.: Diffusion Models in NLP: A Survey, https://www.semanticscholar.org/paper/376e86cc85555041ccab0d51b7bd07d452cf0cd8, (2023). https://doi.org/10.48550/arxiv.2303.07576.
2. Yi, Q., Chen, X., Zhang, C., Zhou, Z., Zhu, L., Kong, X.: Diffusion models in text generation: a survey. PeerJ Comput. Sci. 10, e1905 (2024).
3. Zou, H., Kim, Z.M., Kang, D.: A survey of diffusion models in natural language processing, https://www.semanticscholar.org/paper/4ca824e792f3a2c777ffa5896a2e7cdf11b9518d, (2023).
4. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., Li, C.: Large Language Diffusion Models, http://arxiv.org/abs/2502.09992, (2025).
5. Sohl-Dickstein, J.N., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. ICML. abs/1503.03585, 2256–2265 (2015).

6. Li, Y., Zhou, K., Zhao, W.X., Wen, J.-R.: Diffusion models for non-autoregressive text generation: A survey, https://paperswithcode.com/paper/diffusion-models-for-non-autoregressive-text, (2023). https://doi.org/10.48550/arxiv.2303.06574.
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of large language models, http://arxiv.org/abs/2106.09685, (2021).
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., De Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C.,

Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The Llama 3 herd of models, http://arxiv.org/abs/2407.21783, (2024).

9. Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., Sutton, C.: Program synthesis with large language models, http://arxiv.org/abs/2108.07732, (2021).
10. Lou, A., Meng, C., Ermon, S.: Discrete diffusion modeling by estimating the ratios of the data distribution, http://arxiv.org/abs/2310.16834, (2023).
11. Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., Li, C.: Your absorbing discrete diffusion secretly models the conditional distributions of clean data, http://arxiv.org/abs/2406.03736, (2024).