ETH Zürich - Deep Learning 2024 Proposal

Matteo Boglioni Francesco Rita Andrea Sgobbi Gabriel Tavernini

1. Introduction

Out-of-domain (OOD) generalization is progressively becoming one of the most limiting requirements in the deployment of LLMs in real-world applications (Ohse et al., 2024). Despite notable research effort has been dedicated to OOD Detection (Liu et al., 2024; Zhang et al., 2024), State-of-the-Art LLMs' generalization properties on OOD data still warrant further evaluation.
Yuan et al. (2023) and Mosbach et al. (2023) both investigate LLMs' robustness when faced with OOD data: although a thorough evaluation, this work does not delve into how generalization varies throughout the fine-tuning steps and with different model architectures. On top of this, the focus is only on a single OOD dataset, which implicitly assumes generalization to be a "global" property of models.

To investigate this unexplored direction, we propose to finetune multiple different models and keep track of their performance over a collection of OOD datasets for the Natural Language Inference (NLI) Task (Dagan & Glickman, 2004; Putra et al., 2024), covering the detection of textual entailment within sentence pairs.

2. Models

In accordance with Mosbach et al. (2023), we aim to evaluate the Open Pretrained Transformer (OPT) from Zhang et al. (2022) at 4 different model sizes: 125M, 350M, 1.3B and 2.7B.
Additionally, we extend the comparison by evaluating Mamba (Gu & Dao, 2024) at similar sizes (130M, 370M, 1.4B and 2.8B). This allows us to systematically compare the generalization performance of Mamba and Transformer architectures across varying parameter scales, providing deeper insights into their respective capabilities in NLI.

3. Datasets

We consider a total of 8 NLI datasets. Following the Generalization Taxonomy proposed in Hupkes et al. (2023), we consider the context of *"Across Domain"* generalization in presence of *"Covariate Shift"*, meaning that we expect a shift in the underlying data distribution but a consistent labeling rationale across domains. We use 2 *Main* large-scale datasets for finetuning the models:

- **SNLI** (Bowman et al., 2015) contains 570K crowd-sourced sentence-pairs based on image captions.
- **MNLI** (Williams et al., 2018) is a set of 433K sentence-pairs meant to cover a large range of genres of spoken and written text.

On top of this, we consider 6 other OOD datasets for evaluation. These are split into two main groups: the first group is a collection of *Standard* NLI datasets, while the second group contains *Adversarial* datasets, meant to test the robustness of the heuristic principles learned by the model. The datasets considered are the following:

- *Standard*: **SciTail** (Khot et al., 2018) is based on science multiple-choice exams, **WNLI** tackles identifying the referent of a certain pronoun and **RTE** is a general entailment dataset (the last two are from the GLUE Benchmark (Wang et al., 2019)).
- *Adversarial*: **PAWS** (Zhang et al., 2019) uses paraphrase adversaries, **HANS** (McCoy et al., 2019) tackles failure cases of 3 simple heuristics and **ANLI** (Nie et al., 2020) finds adversaries via human feedback.

4. Evaluation

During the fine-tuning of the models over each of the two *Main* datasets, we will track the test accuracy and F1 score on all 7 OOD datasets (the dataset not in use will also be considered), and compare them to the respective in-domain performance. Counting all the combinations of model sizes and architectures, we obtain a total of 16 fine-tuning runs.

5. Conclusion

We aim to expand on previous works by continuously evaluating OOD generalization across multiple datasets and model architectures. The larger pool of models and datasets will hopefully allow us to understand the generalization dynamics of LLMs, as well as test the common assumption that a single dataset is enough to prove whether a model generalizes OOD.

References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference, 2015. URL https://arxiv.org/abs/1508.05326.

Dagan, I. and Glickman, O. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004 (26-29):2–5, 2004.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R., and Jin, Z. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, October 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00729-y. URL http://dx.doi.org/10.1038/s42256-023-00729-y.

Khot, T., Sabharwal, A., and Clark, P. SciTail: A textual entailment dataset from science question answering. In *AAAI*, 2018.

Liu, B., Zhan, L., Lu, Z., Feng, Y., Xue, L., and Wu, X.-M. How good are llms at out-of-distribution detection?, 2024. URL https://arxiv.org/abs/2308.10261.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., and Elazar, Y. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation, 2023. URL https://arxiv.org/abs/2305.16938.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding, 2020. URL https://arxiv.org/abs/1910.14599.

Ohse, J., Hadžić, B., Mohammed, P., Peperkorn, N., Danner, M., Yorita, A., Kubota, N., Rätsch, M., and Shiban, Y. Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech and Language*, 88:101663, 2024. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2024.101663. URL https://www.sciencedirect.com/science/article/pii/S0885230824000469.

Putra, I. M. S., Siahaan, D., and Saikhu, A. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155, 2024. ISSN 2405-9595. doi: https://doi.org/10.1016/j.icte.2023.08.012. URL https://www.sciencedirect.com/science/article/pii/S2405959523001145.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., and Sun, M. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations, 2023. URL https://arxiv.org/abs/2306.04618.

Zhang, A., Xiao, T. Z., Liu, W., Bamler, R., and Wischik, D. Your finetuned large language model is already a powerful out-of-distribution detector, 2024. URL https://arxiv.org/abs/2404.08679.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Zhang, Y., Baldridge, J., and He, L. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*, 2019.