Language-Specific Pruning for Efficient Reduction of Large Language Models

The Third Ukrainian Natural Language Processing Workshop 2024

Maksym Shamrai

PhD Student
Institute of Mathematics of NAS of Ukraine

Table of contents

Introduction

Pruning

- Pruning, a technique involving the selective removal of model weights, has also shown prominent results in general contexts [5, 10, 4].
- However, their application to different languages and the implications for model performance remain unexplored.

Objective and Significance

Hypothesis

LLMs, which are trained using data from various languages, exhibit unique weight distributions that are specific to each language.

- It means that each language has its own distinct set of weights, which express and reflect the linguistic characteristics and patterns present in the training data specific to that language.
- The aim of this paper is to empirically validate the hypothesis through experimentation with both Ukrainian and English languages and also explore language-specific considerations of model pruning.

Related Work

Methods

- In the context of low-resource languages like Ukrainian, training-free approaches play a crucial role.
- Wanda [7] and SparseGPT [2] are training-free layer-wise pruning methods that require a small calibration dataset for efficient pruning.
- While they share the same framework, they differ in their weight importance metrics:

Wanda

$$S_{ij} = |W_{ij}| \cdot ||X_j||_2,$$

SparseGPT

$$S_{ij} = \left[|W|^2 / \mathrm{diag}\left((X^T X + \lambda I)^{-1} \right) \right]_{ij},$$

where W denotes the weights, X represents the inputs.

Experimental Methodology and Setup

Data

- The UberText 2.0 corpus [1] was utilized since it offers diverse language contexts for the Ukrainian language.
- We sample 4000 records for calibration and 200 for evaluation.
- To emphasize the significance of the language of the calibration data, the English dataset c4 [6] was utilized.

Setup

- Models were pruned to 50% sparsity with unstructured and 2:4 semi-structured configurations.
- LLaMA 7B [8], LLaMA 2 7B [9] and Mistral v0.1 7B [3] in 16-bit floating point precision were chosen for the experiments.
- Perplexity metric, which measures the effectiveness of a language model in predicting a sequence, was used for the evaluation:

$$\text{PPL}(X) = exp\{-\frac{1}{t}\sum_{i=0}^{t}\log p_\theta(x_i|x_{<i})\}.$$

Objective

The objective of the experiments is to empirically and statistically investigate several key aspects:

1. The impact of the size of the calibration dataset on the performance of pruned models.
2. Comparison of the language-specific pruning efficiency of Wanda and SparseGPT.
3. Assessment of the significance of the language of the calibration data for pruning effectiveness.

Results

Calibration Dataset Size Significance

CS	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
64	12.162 ± 0.025	11.283 ± 0.007	**9.314 ± 0.098**
128	12.161 ± 0.012	11.278 ± 0.007	9.726 ± 0.125
256	**12.148 ± 0.008**	11.275 ± 0.009	10.385 ± 0.038
512	12.152 ± 0.007	**11.254 ± 0.012**	12.262 ± 0.424

Table 1: Perplexity values of different models after pruning using unstructured configuration of Wanda and various number of calibration samples[1].

[1] **CS** denotes **Calibration Samples**

Calibration Dataset Size Significance

CS	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
64	31.533 ± 0.169	**30.101 ± 0.406**	**29.822 ± 0.381**
128	31.438 ± 0.348	30.177 ± 0.361	30.741 ± 0.231
256	31.496 ± 0.327	30.651 ± 0.353	32.709 ± 0.328
512	**31.198 ± 0.446**	30.883 ± 0.271	34.471 ± 0.704

Table 2: Perplexity values of different models after pruning using 2:4 semi-structured configuration of Wanda and various number of calibration samples.

Calibration Dataset Size Significance

CS	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
64	10.632 ± 0.027	9.703 ± 0.013	7.109 ± 0.003
128	10.559 ± 0.011	9.683 ± 0.028	7.095 ± 0.011
256	10.531 ± 0.006	9.671 ± 0.015	7.085 ± 0.003
512	**10.529 ± 0.020**	**9.652 ± 0.012**	**7.074 ± 0.004**

Table 3: Perplexity values of different models after pruning using unstructured configuration of SparseGPT and various number of calibration samples.

Calibration Dataset Size Significance

CS	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
64	13.319 ± 0.092	11.559 ± 0.082	8.582 ± 0.036
128	13.148 ± 0.192	11.515 ± 0.072	8.551 ± 0.041
256	13.093 ± 0.054	11.457 ± 0.035	8.497 ± 0.006
512	**12.994 ± 0.047**	**11.379 ± 0.008**	**8.476 ± 0.031**

Table 4: Perplexity values of different models after pruning using 2:4 semi-structured configuration of SparseGPT and various number of calibration samples.

Calibration Dataset Size Significance

- We can conclude that dependency of the calibration dataset size and pruning efficiency depends on the pruning method and the pruned model.
- Nevertheless, models pruned using SparseGPT demonstrated negative correlation between number of calibration samples and perplexity.

Language Significance

Model	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
Dense	8.950	8.269	6.460
UWc4	13.953 ± 0.060	13.829 ± 0.087	41.466 ± 6.314
USc4	15.797 ± 0.761	15.011 ± 0.283	9.208 ± 0.086
UWUT	12.148 ± 0.008	11.254 ± 0.012	9.314 ± 0.098
USUT	**10.529 ± 0.020**	**9.652 ± 0.012**	**7.074 ± 0.004**
2:4Wc4	52.346 ± 1.628	79.801 ± 7.338	433.940 ± 282.154
2:4Sc4	89.772 ± 28.306	57.460 ± 5.379	165.516 ± 90.769
2:4WUT	31.198 ± 0.446	30.101 ± 0.406	29.822 ± 0.381
2:4SUT	**12.994 ± 0.047**	**11.379 ± 0.008**	**8.476 ± 0.031**

Table 5: Perplexity values of different models and different pruning configurations[2].

[2] **U** denotes **Unstructured**, **W** denotes **Wanda**, **c4** denotes **c4 dataset**, **S** denotes **SparseGPT**, **UT** denotes **UberText 2.0 dataset**, **2:4** denotes **2:4 semi-structured**.

Language Significance

- Among both unstructured and especially 2:4 semi-structured configurations, the most effective pruning method is SparseGPT.
- Also, the extreme variances observed in models pruned with c4 data indicate a significant dependency on randomness in the pruning process, suggesting that the outcome is less influenced by the dataset itself.

Conclusion and Discussion

Conclusion

- We observed a dependency on the calibration dataset size only when using SparseGPT, in both unstructured and 2:4 semi-structured configurations.
- The SparseGPT is a better choice in the context of language-specific pruning.
- There is a clear dependency of the effectiveness of the pruned model on the language of the calibration dataset.

Discussion and Future Work

- Given that the accuracy of pruned models depends on the language of the calibration dataset, we can conclude that the hypothesis may be valid because the pruning methods remove only the less significant weights.
- In future work, this pruning technique can serve as a foundational framework for linguistic comparisons by introducing new metric space for languages.
- For instance, a further exploration could involve comparing the languages of Polish and Ukrainian, given their Slavic roots and linguistic proximity.
- Demonstrating their linguistic closeness in the LLM context suggests that fine-tuning the LLM on data from both languages could potentially enhance overall performance.

Questions?

Languages Metric Space

- Want to define metric space where elements of this space are languages.
- Let M_W be the pruning mask of the weight W obtained after pruning it using e.g. SparseGPT.
- Let's stack the pruning masks from all weights of the model into a single vector m_{l_1}, where l_1 is a language we pruned the model.
- Then the distance between language l_1 and language l_2 will be:

$$d(l_1, l_2) = ||m_{l_1} - m_{l_2}||_1$$

- D. Chaplynskyi.
  Introducing UberText 2.0: A corpus of modern Ukrainian at scale.
  In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- E. Frantar and D. Alistarh.
  SparseGPT: Massive language models can be accurately pruned in one-shot.
  arXiv preprint arXiv:2301.00774, 2023.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al.
  Mistral 7b.
  arXiv preprint arXiv:2310.06825, 2023.

References ii

X. Ma, G. Fang, and X. Wang.
Llm-pruner: On the structural pruning of large language models.
Advances in neural information processing systems, 36, 2024.

P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz.
Importance estimation for neural network pruning.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu.
Exploring the limits of transfer learning with a unified text-to-text transformer.
arXiv e-prints, 2019.

M. Sun, Z. Liu, A. Bair, and J. Z. Kolter.
A simple and effective pruning approach for large language models.
arXiv preprint arXiv:2306.11695, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al.
Llama: Open and efficient foundation language models.
arXiv preprint arXiv:2302.13971, 2023.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al.
Llama 2: Open foundation and fine-tuned chat models.
arXiv preprint arXiv:2307.09288, 2023.

N. Yang, Y. Jang, H. Lee, S. Jung, and K. Jung.
Attribution-based task-specific pruning for multi-task language models.
arXiv preprint arXiv:2205.04157, 2022.