Low-cost Matryoshka Embeddings for NLP

Mateusz Boruń
mb417908@students.mimuw.edu.pl

Paweł Wojciechowski
pw431929@students.mimuw.edu.pl

Piotr Zalewski
pz430869@students.mimuw.edu.pl

Supervisior:
Mateusz Doliński

June 14, 2024

1 Introduction

The widespread adoption of machine learning systems in real-world applications has underscored the essential role of learned representations. These representations, once trained and subsequently leveraged across various downstream tasks, form the cornerstone of many modern ML systems. Traditional deep learning models produce high-dimensional embeddings that, while information-dense, incur substantial computational and storage overheads, especially at scale [1]. The inflexibility of these representations often cause the need to maintain high-dimensional vectors across diverse tasks, irrespective of varying resource and accuracy requirements. This rigidity also leads to other inefficiencies, particularly when it necessitates training multiple models to accommodate different embedding sizes for specific tasks, compounding the computational and storage overhead in large-scale deployments.

Human perception naturally processes information in a coarse-to-fine manner [2], which contrasts with the flat and fixed nature of conventional deep learning embeddings. This discrepancy has driven research into more flexible and adaptive representation learning techniques. Existing methods, such as training multiple low-dimensional models, or post-hoc compression, often fall short due to training overheads, multiple expensive forward passes, storage costs, and accuracy drops [3].

Matryoshka Representation Learning (MRL) [3], inspired by the nested structure of Matryoshka dolls, addresses these challenges by learning coarse-to-fine representations within a single high-dimensional vector. MRL optimizes lower-dimensional vectors nested within the

high-dimensional space, enabling adaptive deployment at no additional inference cost. This flexibility is particularly beneficial for large-scale classification and retrieval tasks, where it offers significant computational efficiency without compromising accuracy.

Therefore, in this study, we aim to:

1. **Expand Application Scope:** While the original MRL research provided some limited insights into Matryoshka token representations, this study extends the investigation to snippet embeddings, and the curve of performance penalty resulting from down-scaling of the full-size Matryoshka representations.

2. **Study MRL-style Fine-Tuning of Pretrained Models:** Evaluating the potential of fine-tuning existing pretrained NLP models to output high-quality Matryoshka embeddings will be crucial in determining the adaptability of MRL in the current NLP landscape.

The successful adaptation of Matryoshka representations to NLP could revolutionize how embeddings are generated and utilized, offering a flexible, efficient, and scalable solution for diverse NLP applications, especially in resource-constrained environments.

Despite the promising potential of MRL, its application in the natural language processing (NLP) domain and specifically in the context of existing models remains underexplored. The original research predominantly focused on the visual domain, leaving a gap in understanding how MRL influences downstream task performance of the representations of a given size, especially when using the nested representations with reduced fidelity, which are crucial for various NLP tasks such as classification, retrieval, and for low-resource or large scale applications. This study aims to bridge this gap by investigating the application of Matryoshka embeddings in the NLP domain, particularly in low-resource settings, where the original method of full training with MRL objective might not be feasible.

2 Related work

Matryoshka embeddings, which encode representations of varying capacities within the same high-dimensional vector, are already frequently used in both computer vision (CV) and natural language processing (NLP), as they allow for a flexible and adaptive approach to model deployment.

In computer vision, the performance of Matryoshka embeddings has been well documented [3]. This technique has shown significant improvements in tasks like large-scale classification and retrieval. For example, models like ResNet and Vision Transformers (ViTs) have leveraged MRL to achieve state-of-the-art results on datasets such as ImageNet, demonstrating negligible accuracy differences while maintaining computational efficiency [3].

While Matryoshka embeddings are also utilized in NLP, their performance has not been as rigorously measured as in the vision domain. There is a notable gap in the literature regarding the evaluation of these embeddings for NLP tasks, and the exact quantitative nature of the trade-off between computational efficiency and accuracy when using the scaled-down versions of Matryoshka embeddings.

Another critical aspect that remains unclear is the feasibility of fine-tuning existing embedding models to output Matryoshka embeddings. Although fine-tuning pretrained models has become a standard practice in NLP, it is uncertain whether this approach can effectively produce high-quality Matryoshka embeddings. Exploring this possibility is essential for understanding how these embeddings can be adapted and optimized for various NLP applications.

3 Experimental setup

To evaluate the effectiveness of Matryoshka Representation Learning in NLP tasks, we will employ parts of the MTEB benchmark, specifically focusing on classification and retrieval tasks. The MTEB benchmark provides a comprehensive suite of tasks that are particularly relevant for assessing the performance of embedding models in real-world scenarios. Classification and retrieval are critical applications where the flexibility and efficiency of Matryoshka embeddings can be effectively demonstrated. By leveraging these benchmark tasks, we aim to highlight the potential advantages of MRL in delivering high-quality embeddings that can adapt to varying resource constraints without significant losses in performance.

For our base model, we have selected the pretrained DistilRoBERTa base [4] model trained for masked language modeling, which is particularly suitable for our study due to its fairly compact size (82M parameters) and respectable performance on downstream tasks [5], making it an ideal candidate for low-resource scenarios where computational and storage efficiency are critical. We intend to finetune that base model for embeddings in two versions: standard fixed feature embeddings and Matryoshka embeddings. These two models, trained on the same dataset, will serve as our baselines. By evaluating both the standard fixed feature embeddings and the Matryoshka embeddings, we will also provide accurate results on the applicability and effectiveness of Matryoshka embeddings in NLP. Furthermore, we will explore the feasibility of fine-tuning to Matryoshka embeddings, using the first model (trained with standard fixed feature embeddings) as the initial state for this fine-tuning process. This approach will help us understand the potential benefits and challenges of adapting existing pretrained models to output Matryoshka-style embeddings for various downstream tasks.

For our experiments, we will use the Multitask Embeddings Data with Instructions (MEDI) dataset [6]. MEDI, consisting of 330 datasets from sources such as Super-NI, sentence-transformer embedding training data, KILT, and MedMCQA, covers a broad spectrum of domains and tasks. This dataset is ideal for our study due to its relatively small yet comprehensive nature, making it suitable for low-resource settings.

References

[1] Jeffrey Dean. Challenges in building large-scale information retrieval systems: invited talk. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, New York, NY, USA, 2009. Association for Computing Machinery.

[2] Mike G Harris and Christos D Giachritsis. Coarse-grained information dominates fine-grained information in judgments of time-to-contact from retinal flow. *Vision Research*, 40(6):601–611, 2000.

[3] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc., 2022.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[5] https://sbert.net/docs/sentence_transformer/pretrained_models.html#original-models.

[6] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada, July 2023. Association for Computational Linguistics.