Investigating Mixture of Experts in Dense Retrieval*

Effrosyni Sokli[1,*], Pranav Kasela[1], Georgios Peikos[1] and Gabriella Pasi[1]

[1]Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Milan, Italy

Abstract

While Dense Retrieval Models (DRMs) have advanced Information Retrieval (IR), they often suffer from limited generalizability and robustness. Various studies address these limitations with representation learning techniques that leverage the Mixture-of-Experts (MoE) architecture. Unlike prior works in IR that integrate MoE within the Transformer layers of DRMs, we add a single MoE block (SB-MoE) after the output of the final Transformer layer. Our empirical evaluation investigates how SB-MoE compares, in terms of retrieval effectiveness, to standard model fine-tuning. Given MoEs sensitivity to its hyperparameters (i.e., the number of experts), we also investigate our model's performance under different expert configurations. Results show that SB-MoE is particularly effective for lightweight DRMs, consistently outperforming their fine-tuned counterparts. For larger DRMs, SB-MoE requires more training data to deliver improved retrieval performance. Our code is available online at: https://anonymous.4open.science/r/DenseRetrievalMoE.

Keywords

Mixture-of-Experts, Representation Learning, Dense Neural Retrievers

1. Introduction

Dense Retrieval Models (DRMs) can capture the semantic context of queries and documents [2] and often outperform sparse lexicon-based models such as BM25 [3] across various IR tasks. However, their dependence on large labeled datasets and limited cross-domain generalizability often requires additional fine-tuning for robust adaptation to different tasks or domains. In this paper, we investigate the effectiveness of an enhanced bi-encoder DRM architecture leveraging Mixture-of-Experts (MoE) [4] in various dense retrieval tasks. Unlike previous studies in IR that integrate MoE within each Transformer layer [5, 6], we apply a single MoE block (SB-MoE) on the final output embeddings of the underlying DRM. SB-MoE is trained in an unsupervised manner to automatically optimize each expert and dynamically aggregate their outputs, adapting predictions to the input embeddings, i.e., the query and document representations produced by the underlying DRM. We utilize two datasets of the BEIR collection [7] (i.e., Natural Questions (NQ) [8] and HotpotQA [9]), and two of the Multi-Domain Benchmark by Bassani et al. [10] (i.e., Political Science (PS) and Computer Science (CS)), to empirically evaluate SB-MoE's retrieval effectiveness for open-domain Q&A, and domain-specific academic search.

This work has the following contributions: (1) We introduce a modular MoE framework, SB-MoE, which operates on the query and document embeddings produced by an underlying bi-encoder DRM architecture; (2) We conduct an empirical evaluation using three DRMs (Contriever, BERT, and TinyBERT) investigating SB-MoE's retrieval performance and its sensitivity to hyperparameters (i.e., the number of employed experts), compared to standard model fine-tuning across four benchmarks.

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*This is an extended abstract of [1].

*Corresponding author.

effrosyni.sokli@unimib.it (E. Sokli); pranav.kasela@unimib.it (P. Kasela); georgios.peikos@unimib.it (G. Peikos); gabriella.pasi@unimib.it (G. Pasi)

0009-0003-5388-2385 (E. Sokli); 0000-0003-0972-2424 (P. Kasela); 0000-0002-2862-8209 (G. Peikos); 0000-0002-6080-8170 (G. Pasi)

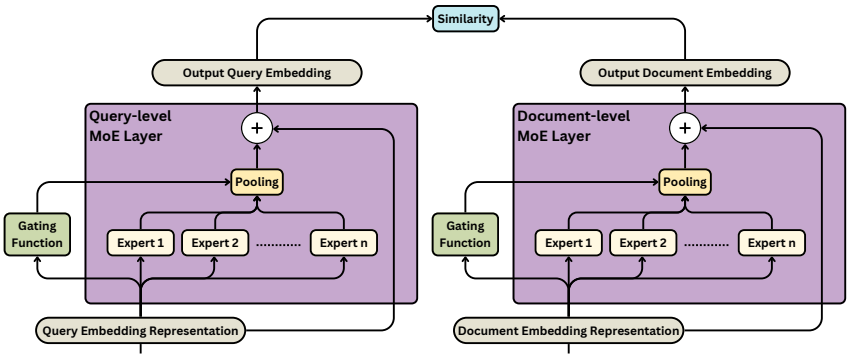

Figure 1: Overview of the SB-MoE architecture, highlighting its three main parts.

2. Related Work

DRMs often outperform lexicon-based models (e.g., BM25 [3]), since they can capture the semantic context of queries and documents. They project both queries and documents in a common dense vector space and score documents through similarity functions for a given query [11, 12, 13]. In this work, we leverage three DRMs. Contriever [14] is a state-of-the-art BERT-based model that exploits contrastive learning, a Machine Learning technique that uses pairs of positive and negative examples to learn meaningful and distinctive representations of queries and documents. Besides BERT [15], we also use TinyBERT [16], which leverages knowledge distillation [17] to transfer knowledge from its larger counterpart (BERT) to a tinier version, reducing training times and computational expenses. DRMs often showcase continuous adaptation needs, which can lead to low generalizability and robustness [18, 19]. The MoE [4] framework has been employed in various approaches to mitigate these issues. MoE can handle multiple types of data and tasks [20, 21] and has been used in tasks such as classification [22], and multi-lingual machine translation [23]. MoE has been employed for IR tasks such as passage retrieval [5, 24], and Q&A [6, 25, 26]. These approaches either integrate MoE blocks into every layer of the Transformer model (substantially increasing the number of parameters) or only partially leverage MoE by applying it solely to the query representation. In our work, we apply a single MoE block to both query and document representations and train the obtained architecture end-to-end for retrieval.

3. Methodology

SB-MoE builds upon a bi-encoder DRM architecture [27], which allows for independent encoding of documents and queries to enhance scalability and to enable the computation of relevance scores through a similarity function (e.g., cosine similarity). The proposed model's architecture consists of three parts (Figure 1): (1) the experts, operating on the query and document representations produced by the underlying DRM; (2) the gating function, trained in an unsupervised manner to indicate the most appropriate expert(s) for a given input; and (3) the pooling module, used in the final stage to aggregate the experts' representations and produce the final embedding to be used for similarity estimation between the query and documents.

The experts receive as input the query or document embedding as produced by the underlying DRM. The output is n modified representations, where n is the number of employed experts. The gating function receives the same input and produces an n-dimensional vector, which indicates the importance of each experts contribution to the final query or document embedding. We rely on noisy Top-1 gating, as proposed by Shazeer et al. [23], for training the gating function. This approach ensures that SB-MoE can explore every expert during training, enhancing the robustness of the model. During inference, the pooling module uses two different strategies. The

Table 1
Results on all four datasets. Symbol * indicates a statistically significant difference over Fine-tuned. The best results for each model are underlined.

		NQ		HotpotQA		PS		CS	
Retriever	Variant	NDCG@10	R@100	NDCG@10	R@100	NDCG@10	R@100	NDCG@10	R@100
TinyBERT	Fine-tuned	.216	.689	.158	.394	.125	.262	.150	.308
	SB-MoE$_{\text{TOP-1}}$	.219	.693	.162*	.399*	.130*	.271*	.153*	.313*
	SB-MoE$_{\text{ALL}}$	.217	.697*	.171*	.411*	.129*	.270*	.153*	.315*
BERT	Fine-tuned	.265	.846	.372	.660	.183	.374	.172	.362
	SB-MoE$_{\text{TOP-1}}$	.261	.842	.349*	.642*	.183	.377*	.175*	.364
	SB-MoE$_{\text{ALL}}$	.258*	.840	.362*	.649*	.184	.378*	.167*	.355*
Contriever	Fine-tuned	.426	.934	.672	.862	.251	.483	.224	.437
	SB-MoE$_{\text{TOP-1}}$	.416*	.930*	.653*	.853*	.250	.479*	.222*	.435
	SB-MoE$_{\text{ALL}}$	.416*	.932	.667*	.861	.251	.483	.223	.438

first one is Top-1 gating [28] (SB-MoE$_{\text{TOP-1}}$), which selects solely the output of the expert that the gating function assigned the highest score to. The second strategy (SB-MoE$_{\text{ALL}}$) calculates probability scores from the gating function's output vector through a softmax normalization [29], and produces the final embedding, which is the weighted sum of all experts' outputs.

4. Experimental Analysis

This section presents the empirical evaluation conducted to answer the following research questions (RQs):

RQ1 How does SB-MoE compare, in terms of effectiveness, to standard model fine-tuning?
RQ2 How does the number of experts impact the retrieval effectiveness of SB-MoE?

4.1. Experimental Setup

For RQ1, we employ 6 distinct experts across all models and datasets. For RQ2, we vary the number of experts from 3 to 12 with a step of 3. This setup is based on previous studies [30, 31], which suggest that a high number of experts does not always yield performance improvements [32], and experiment with expert counts ranging from 2 to 8 [25, 24, 33]. We follow the architecture proposed by Houlsby et al. [34], where each expert consists of a feed-forward network (FFN) with a down-projection layer that reduces the input dimension by half, followed by an up-projection FFN layer, which restores the vector dimension to the original embedding size. The gating function includes a single hidden layer that reduces the input dimension by half, and an output FFN layer with dimensionality equal to the number of experts. During training, we use a batch size of 64. The learning rate is set to 10^{-6} for the underlying DRM and 10^{-4} for the experts. TinyBERT is trained for 30 epochs across all datasets, while BERT and Contriever are trained for 20 epochs due to resource constraints and longer training times, on all datasets except CS, where they are trained for 10 epochs since the collection's training queries are ~3.5 times more than the second largest collection used (PS). We reserve 5% of each training set for validation and keep the checkpoint with the lowest validation loss. We set the random seed to 42 and use contrastive loss [14] with a temperature of 0.05. For our evaluation, we use NDCG@10 and R@100, two metrics commonly used on BEIR, for comparability. Statistical significance is assessed using two-sided paired Student's t-tests with Bonferroni multiple testing correction, at a significance level of 0.05. We integrate SB-MoE into three different DRMs and compare its retrieval effectiveness to that achieved by the underlying DRM, fine-tuned on the same training data and hyperparameters. We refer to these baseline experiments as Fine-tuned.

4.2. Results and Discussion

RQ1. As shown in Table 1, SB-MoE consistently improves NDCG@10 and Recall@100, especially for lightweight models. For example, on TinyBERT, SB-MoE leads to noticeable performance

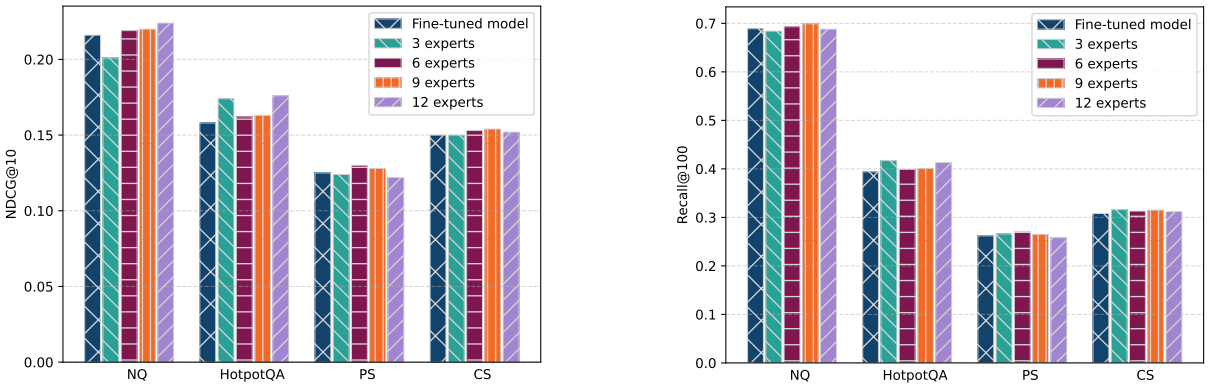

Figure 2: SB-MoE$_{\text{TOP-1}}$ on TinyBERT with 3, 6, 9, and 12 experts.

gains in both metrics across all datasets, with a marked increase in HotpotQA, where SB-MoE$_{\text{ALL}}$ achieved an NDCG@10 score of .171 compared to .158 of the fine-tuned version. However, for larger models like BERT and Contriever, the integration of SB-MoE had a marginal impact, with similar or slightly worse retrieval performance compared to Fine-tuned. These results suggest that in models already equipped with a substantial number of parameters, SB-MoE's advantages may not be so prominent, potentially due to redundancy when additional experts are employed. Therefore, the integration of SB-MoE particularly benefits lightweight models.

RQ2. As SB-MoE seems to benefit significantly lightweight models, we leverage TinyBERT to understand the impact of the number of experts, by configuring SB-MoE with 3, 6, 9, and 12 experts and evaluating across all datasets (Figure 2). Our findings show variations in performance for different expert counts across datasets, which can also lead to the maximization of different performance measures, as observed in the case of NQ, where the employment of 12 experts maximizes NDCG@10, but Recall@100 is maximized with 9 experts. Therefore, the number of employed experts is a hyperparameter that requires tuning with respect to the domain and the addressed retrieval task.

5. Conclusions

In this work, we integrate a single Mixture-of-Experts block (SB-MoE) into Dense Retrieval Models (DRMs) and conduct an experimental investigation on its effectiveness in different dense retrieval tasks. Results show that SB-MoE significantly enhances the retrieval performance of lightweight DRMs, consistently improving NDCG@10 and R@100 across datasets. However, larger DRMs only marginally benefit from SB-MoE, indicating that models with a higher parameter count need dataset-specific optimization to see measurable gains. Our analysis reveals that the number of employed experts is a key hyperparameter, which influences SB-MoE's performance and requires task and domain-specific calibration.

Acknowledgments

This work has received funding from the European Unions Horizon Europe research and innovation programme under the Marie Skodowska-Curie grant agreement No 101073307.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. Sokli, P. Kasela, G. Peikos, G. Pasi, Investigating mixture of experts in dense retrieval, CoRR abs/2412.11864 (2024). URL: https://doi.org/10.48550/arXiv.2412.11864. doi:10.48550/ARXIV.2412.11864. arXiv:2412.11864.

[2] B. Mitra, N. Craswell, An introduction to neural information retrieval, Foundations and Trends® in Information Retrieval 13 (2018) 1–126.

[3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, Okapi at trec-3, Nist Special Publication Sp 109 (1995) 109.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive Mixtures of Local Experts, Neural Computation 3 (1991) 79–87. doi:10.1162/neco.1991.3.1.79.

[5] J. Guo, Y. Cai, K. Bi, Y. Fan, W. Chen, R. Zhang, X. Cheng, Came: Competitively learning a mixture-of-experts model for first-stage retrieval, ACM Trans. Inf. Syst. (2024). doi:10.1145/3678880, just Accepted.

[6] S. Shen, L. Hou, Y. Zhou, N. Du, S. Longpre, J. Wei, H. W. Chung, B. Zoph, W. Fedus, X. Chen, T. Vu, Y. Wu, W. Chen, A. Webson, Y. Li, V. Y. Zhao, H. Yu, K. Keutzer, T. Darrell, D. Zhou, Mixture-of-experts meets instruction tuning: A winning combination for large language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=6mLjDwYte5.

[7] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[8] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural Questions: A Benchmark for Question Answering Research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466. URL: https://doi.org/10.1162/tacl_a_00276. doi:10.1162/tacl_a_00276.

[9] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. doi:10.18653/v1/D18-1259.

[10] E. Bassani, P. Kasela, A. Raganato, G. Pasi, A multi-domain benchmark for personalized search evaluation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 38223827. doi:10.1145/3511808.3557536.

[11] L. Gao, J. Callan, Unsupervised corpus aware language model pre-training for dense passage retrieval, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2843–2853. doi:10.18653/v1/2022.acl-long.203.

[12] E. Kamalloo, N. Thakur, C. Lassance, X. Ma, J.-H. Yang, J. Lin, Resources for brewing beir: Reproducible reference models and statistical analyses, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 14311440. doi:10.1145/3626772.3657862.

[13] Y. Yu, C. Xiong, S. Sun, C. Zhang, A. Overwijk, COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1462–1479. doi:10.18653/v1/2022.emnlp-main.95.

[14] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsu-

pervised dense information retrieval with contrastive learning, Transactions on Machine Learning Research (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[16] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, TinyBERT: Distilling BERT for natural language understanding, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4163–4174. doi:10.18653/v1/2020.findings-emnlp.372.

[17] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, arXiv (2014).

[18] Y. Liu, R. Zhang, J. Guo, M. de Rijke, Y. Fan, X. Cheng, Robust neural information retrieval: An adversarial and out-of-distribution perspective, CoRR abs/2407.06992 (2024). URL: https://doi.org/10.48550/arXiv.2407.06992. doi:10.48550/ARXIV.2407.06992. arXiv:2407.06992.

[19] G. Sidiropoulos, E. Kanoulas, Analysing the robustness of dual encoders for dense retrieval against misspellings, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2132–2136. URL: https://doi.org/10.1145/3477495.3531818. doi:10.1145/3477495.3531818.

[20] R. Collobert, S. Bengio, Y. Bengio, A parallel mixture of svms for very large scale problems, Advances in Neural Information Processing Systems 14 (2001).

[21] M. Li, M. Li, K. Xiong, J. Lin, Multi-task dense retrieval via model uncertainty fusion for open-domain question answering, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 274–287. doi:10.18653/v1/2021.findings-emnlp.26.

[22] D. Eigen, M. Ranzato, I. Sutskever, Learning factored representations in a deep mixture of experts, arXiv preprint arXiv:1312.4314 (2013).

[23] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, in: International Conference on Learning Representations, 2017. URL: https://openreview.net/forum?id=B1ckMDqlg.

[24] G. Ma, X. Wu, P. Wang, S. Hu, Cot-mote: exploring contextual masked auto-encoder pre-training with mixture-of-textual-experts for passage retrieval, arXiv preprint arXiv:2304.10195 (2023).

[25] D. Dai, W.-J. Jiang, J. Zhang, W. Peng, Y. Lyu, Z. Sui, B. Chang, Y. Zhu, Mixture of experts for biomedical question answering, in: Natural Language Processing and Chinese Computing, 2022. URL: https://api.semanticscholar.org/CorpusID:248218762.

[26] P. Kasela, G. Pasi, R. Perego, N. Tonellotto, Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 111–125.

[27] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[28] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. M. Dai, Z. Chen, Q. V. Le, J. Laudon, Mixture-of-experts with expert choice routing, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728e

b-Abstract-Conference.html.

[29] M. I. Jordan, R. A. Jacobs, Hierarchical Mixtures of Experts and the EM Algorithm, Neural Computation 6 (1994) 181–214. URL: https://doi.org/10.1162/neco.1994.6.2.181. doi:10.1162/neco.1994.6.2.181.

[30] X. Li, S. He, J. Wu, Z. Yang, Y. Xu, Y. jun Jun, H. Liu, K. Liu, J. Zhao, Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 11475–11485. URL: https://aclanthology.org/2024.lrec-main.1003.

[31] T. Zadouri, A. Üstün, A. Ahmadian, B. Ermis, A. Locatelli, S. Hooker, Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024.

[32] T. Chen, Z. Zhang, A. K. Jaiswal, S. Liu, Z. Wang, Sparse moe as the new dropout: Scaling dense and self-slimmable transformers, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=w1hwFUb_81.

[33] Y. Wang, S. Agarwal, S. Mukherjee, X. Liu, J. Gao, A. H. Awadallah, J. Gao, AdaMix: Mixture-of-adaptations for parameter-efficient model tuning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5744–5760. doi:10.18653/v1/2022.emnlp-main.388.

[34] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.