Query Encoder Distillation via Embedding Alignment is a Strong Baseline Method to Boost Dense Retriever Online Efficiency

Yuxuan Wang and **Hong Lyu**[*]
University of Pennsylvania
{wangy49, hlyu}@seas.upenn.edu

Abstract

The information retrieval community has made significant progress in improving the efficiency of Dual Encoder (DE) dense passage retrieval systems, making them suitable for latency-sensitive settings. However, many proposed procedures are often too complex or resource-intensive, which makes it difficult for practitioners to adopt them or identify sources of empirical gains. Therefore, in this work, we propose a trivially simple recipe to serve as a baseline method for boosting the efficiency of DE retrievers leveraging an asymmetric architecture. Our results demonstrate that even a 2-layer, BERT-based query encoder can still retain 92.5% of the full DE performance on the BEIR benchmark via unsupervised distillation and **proper student initialization**. We hope that our findings will encourage the community to re-evaluate the trade-offs between method complexity and performance improvements.

1 Introduction

Recent advances in neural-based NLP techniques have led to powerful neural encoders that can generate high-quality, semantic-rich, dense vector text representations (Reimers and Gurevych, 2019; Cer et al., 2018; Conneau et al., 2018; Schick et al., 2023), making it possible to calculate the text relevancy with simple vector operations like dot product. Thus, the Dual Encoder (DE) neural Information Retrieval (IR) architectures, combined with optimized semantic search implementations (Andoni et al., 2018; Johnson et al., 2019; Boytsov and Nyberg, 2020), have achieved comparable or even superior performances to their Cross Encoder (CE) based predecessors (Thakur et al., 2021; Menon et al., 2022; Ni et al., 2022; Yu et al., 2022) while being significantly more efficient (Reimers and Gurevych, 2019).

Despite the numerous proposed efficiency enhancements for making DE-based IR models suitable for production settings, they may pose challenges for practitioners with limited resources in terms of adoption and replication (Hooker, 2020). However, by leveraging two key facts, we can simplify model development while achieving higher efficiency. Firstly, documents, in contrast to queries, are typically longer and more complex, necessitating specialized architectures (Zhang et al., 2019; Dai et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020). Secondly, document embeddings remain mostly static after indexing, allowing for a high-quality and computationally expensive document encoder without online overhead. Based on these insights, we propose an asymmetric IR architecture that pairs a lightweight query encoder with a robust document encoder.

In this study, we present a minimalistic baseline approach for constructing the aforementioned asymmetric retriever using any existing query encoder. As depicted in Figure 1, by employing suitable initialization and simply minimizing the Euclidean distance between student and teacher query embeddings, even a 2-layer BERT-based query encoder (Devlin et al., 2018) can retain 92.5% of the full DE performance on the BEIR benchmark (Thakur et al., 2021). Similarly, the 4-layer encoder preserves 96.2% of the full performance, which aligns with the supervised outcome (96.6%) achieved by a 6-layer encoder (Kim et al., 2023). We hope that these findings will motivate the research community to reassess the trade-offs between method complexity and performance enhancements. Our code is publicly available in our GitHub repository.

2 The Trivially Simple Distillation Recipe

2.1 Student Initialization

The initialization of student model weights is frequently not given enough attention in the knowledge distillation literature for IR. We find that a

[*] Both authors contributed equally to this research.

Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 290–298
July 13, 2023 ©2023 Association for Computational Linguistics

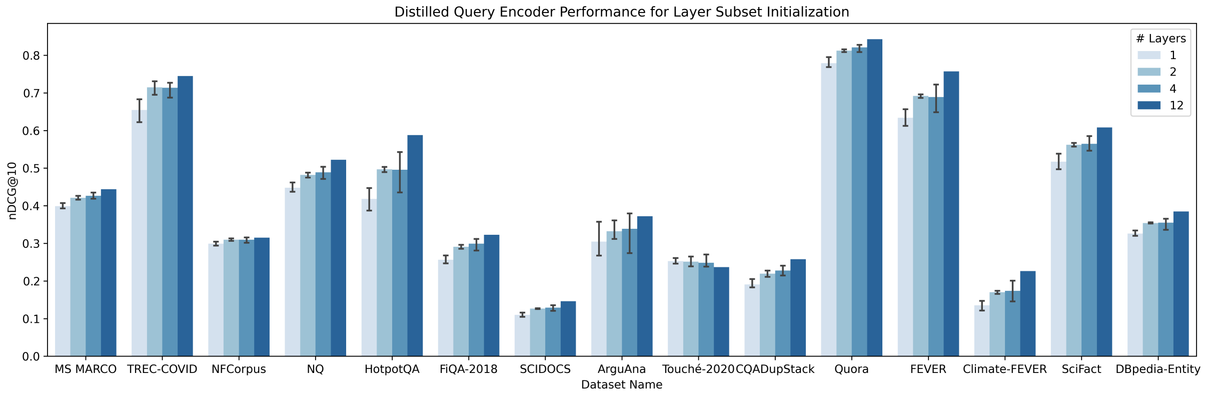

Figure 1: The dual encoder retriever performance with distilled query encoders of a varying number of layers. The student models are initialized by extracting subsets of the teacher model (msmarco-bert-base-dot-v5) layers. Variances in performances come from different layer subsets chosen, discussed in section 4.

"well-prepared" student model can considerably alleviate distillation challenges. In this study, we investigate two classes of initialization approaches.

Extract a Subset of Teacher Layers In this initialization method, we establish the student model by taking a subset of the teacher model's transformer layers while keeping the embedding and pooling layers. By inheriting some of the teacher model's structural properties and knowledge, the student model is intuitively better prepared for efficient distillation in comparison to a randomly initialized student model. We conduct experiments using various combinations of teacher model layers to assess their impact on both performance and efficiency. Details would be discussed in subsection 3.2, section 4, and subsection A.1.

Adopt Other Pretrained Models We also explore initializing the student model from other efficient pretrained models. Simultaneously, we investigate the influence of multiple factors, e.g., fine-tuning tasks and distance functions, on achieving a "well-prepared" initialization for the student model. We take DistilBERT (Sanh et al., 2019) as our student model candidate and experiment with different DistilBERT models fine-tuned on diverse tasks. Examples include the distilbert-base-uncased model and DistilBERT models fine-tuned on the MS MARCO dataset (Bajaj et al., 2018) with distinct objectives from sentence-transformers on HuggingFace (Wolf et al., 2019). This approach enables us to evaluate the efficacy of using alternative pretrained models as starting points for student model initialization. Student model cards are listed in subsection A.2.

2.2 Embedding Alignment

DE-based IR systems often use vector similarity for searching (Andoni et al., 2018), making it logical to match student and teacher embedding spaces.

Contextualized Embedding Pooling Strategies BERT-based encoders produce contextualized representations for all tokens from the input text. Common ways to aggregate token embeddings are selecting [CLS] embedding, computing mean values across all token embeddings, concatenating multiple pooled embeddings together, etc. We stick with the average pooling strategy for all presented experiment results in this paper, as in Reimers and Gurevych (2019).

Alignment Objective Let $\text{Enc}^s_\theta(\cdot)$ denote the student query encoder parameterized by θ and $\text{Enc}^t(\cdot)$ denote the teacher query encoder, we minimize the expected Euclidean distance between the student and teacher embeddings,

$$\mathcal{L}(\theta) = \mathbb{E}_{q\sim\mathcal{D}_q}\left[\|\text{Enc}^s_\theta(q) - \text{Enc}^t(q)\|_2\right]$$

Thus, θ is found by minimizing the empirical loss,

$$\theta = \underset{\theta}{\text{argmin}} \frac{1}{|\mathcal{Q}|}\sum_{q_i\in\mathcal{Q}} \|\text{Enc}^s_\theta(q_i) - \text{Enc}^t(q_i)\|_2$$

where $\mathcal{Q}$ denotes a set of queries sampled from the distillation domain. In our experiment, we set $\mathcal{Q}$ to be the queries of the IR datasets used by teacher query encoders. This simple optimization objective yields surprisingly performant student models **when paired with proper initialization.**

3 Experiments

3.1 Evaluation Datasets and Metrics

Retrieval Performance For in-domain evaluation, we keep the dataset consistent with our teacher models' training corpus MS MARCO (Bajaj et al., 2018). As for the out-of-domain (zero-shot) evaluation, we use the BEIR benchmark (Thakur et al., 2021) to evaluate our distillation method. It is a diverse collection of seven categories[1] of IR-related tasks. We report normalized Discounted Cumulative Gain (nDCG@10) as the performance metric and average the **relative** performance drops to compare the distillation results.

Inference Efficiency We evaluate the efficiency of our distilled query encoder by measuring the wall-clock time required to process queries from the NQ dataset (Kwiatkowski et al., 2019). We simulate various scenarios, ranging from nearly online settings to batched processing, by selecting batch sizes of 4, 8, 16, 32, and 64. For each batch size, we record the elapsed time to process approximately 4×10^3 queries on a single Nvidia Tesla T4 GPU, repeating the process three times and taking the median time to calculate the number of queries processed per second as the evaluation result.

3.2 Teacher and Student Models

The Teacher Model In this work, we use a siamese DE model `msmarco-bert-base-dot-v5` hosted on the HuggingFace hub for its competitive performance (Figure 3). The model was fine-tuned on MS MARCO using the dot score as the relevancy measurement and Margin Mean Squared Error (MarginMSE) as the objective function.

Extractive Initialization We select a total of thirteen combinations, comprising five combinations of 4-layer models, four combinations of 2-layer models, and four combinations of 1-layer models. The full combinations are listed in subsection A.1.

DistilBERT Initialization We explore six DistilBERT checkpoints. The students are initialized from the full model without extracting subsets of layers. Please refer to subsection A.2 for the HuggingFace model cards. We discuss the potential relationship between distillation performance and model characteristics in section 4.

3.3 Implementation Details

We use the first 80% of over eight million queries from the MS MARCO training set as our training data and the rest 20% for validation. We train the student models using the AdamW optimizer (Loshchilov and Hutter, 2017) for one epoch with Mean Squared Error (MSE) loss, applying a batch size of 128, a learning rate of 10^{-4} and 10^3 warm-up steps.

4 Results and Discussions

Initializing from Subsets of Teacher Layers Figure 1 illustrates the performance of the distilled query encoders. We observe that different initialization strategies can lead to up to 6% variability in performance, even with the same number of layers. However, **we find that initializing the students with the first and last few layers consistently yields preferable results**, which aligns with previous findings (Fan et al., 2019; Sajjad et al., 2020; Dong et al., 2022). For instance, considering the 1-layer student encoder (Figure 5), initializing from the last layer yields the best outcomes across all datasets except for ArguAna (Wachsmuth et al., 2018) and Touché-2020 (Bondarenko et al., 2021), preserving an average relative performance of 86.1%. This observation applies similarly to the 2-layer (retaining the first and last layers) and 4-layer (retaining the first and last two layers) students, which exhibit performance preservation rates of 92.5% and 96.2% respectively, aligning closely with the performance of the supervised distilled 6-layer encoder at 96.6% (Kim et al., 2023).

Initializing from DistilBERTs The results in Table 1 reveal the within-group performance comparison. Since all student models undergo the same embedding-alignment distillation process, the final performance preservation rate can serve as a proxy for the "well-preparedness" of students. `msmarco-dot` performs the best. Its tuning configuration is the same as its teacher's, i.e., the same dataset, distance function, and objective function. `msmarco-tas-b`, tuned with the balanced topic-aware sampling technique (Hofstätter et al., 2021b), closely follows. Such a variation poses a slightly greater challenge in embedding alignment. On the other side of the spectrum, changing a distance measurement alone makes alignment drastically harder, as shown from `msmarco-cos`. Interestingly, using a different objective function (`msmarco-base`

[1] The original publication presents nine categories, but the news and tweet retrieval datasets are not publicly available.

Fine-tune Dataset	MS MARCO		MS MARCO		-		MS MARCO		NLI + STS		MS MARCO	
Fine-tune Objective	MarginMSE		MarginMSE		-		MultiNegRanking		CosineSimilarity		MarginMSE	
Similarity Function	Dot		Dot		-		Cosine		Cosine		Cosine	
Dataset (↓) **Ckpt** (→)	msmarco-dot		msmarco-tas-b		base-uncased		msmarco-base		nli-stb		msmarco-cos	
MS MARCO (In-domain)	**.415**	**(- 6.58%)**	.412	(- 7.17%)	.406	(- 8.60%)	.389	(-12.40%)	.389	(-12.39%)	.346	(-22.04%)
TREC-COVID	**.689**	**(- 7.58%)**	.676	(- 9.30%)	.634	(-14.86%)	.575	(-22.78%)	.573	(-23.03%)	.512	(-31.31%)
NFCorpus	.290	(- 7.84%)	**.297**	**(- 5.87%)**	.283	(-10.04%)	.271	(-14.06%)	.269	(-14.56%)	.230	(-27.04%)
NQ	**.481**	**(- 7.94%)**	.480	(- 8.04%)	.463	(-11.42%)	.433	(-17.09%)	.440	(-15.75%)	.404	(-22.67%)
HotpotQA	**.441**	**(-25.06%)**	.421	(-28.38%)	.394	(-33.06%)	.352	(-40.12%)	.348	(-40.87%)	.287	(-51.15%)
FiQA-2018	**.291**	**(- 9.84%)**	.289	(-10.61%)	.282	(-12.81%)	.269	(-16.59%)	.271	(-16.15%)	.232	(-28.18%)
ArguAna	**.426**	**(14.33%)**	.417	(12.12%)	.392	(5.40%)	.408	(9.68%)	.368	(- 1.07%)	.402	(8.09%)
Touché-2020	.232	(- 2.22%)	.238	(0.59%)	.245	(3.36%)	**.247**	**(4.22%)**	.244	(3.17%)	.234	(- 1.43%)
SCIDOCS	**.121**	**(-17.02%)**	.119	(-18.93%)	.110	(-25.10%)	.099	(-32.29%)	.100	(-31.29%)	.086	(-41.18%)
CQADupStack	**.218**	**(-15.54%)**	.212	(-17.63%)	.204	(-21.08%)	.180	(-30.10%)	.185	(-28.26%)	.148	(-42.58%)
Quora	**.812**	**(- 3.59%)**	.809	(- 3.98%)	.806	(- 4.31%)	.799	(- 5.23%)	.792	(- 6.04%)	.751	(-10.88%)
FEVER	**.620**	**(-18.06%)**	.616	(-18.64%)	.568	(-25.00%)	.535	(-29.39%)	.519	(-31.40%)	.446	(-41.05%)
Climate-FEVER	**.182**	**(-19.77%)**	.175	(-22.64%)	.163	(-28.12%)	.156	(-30.89%)	.150	(-33.70%)	.145	(-35.77%)
SciFact	.541	(-11.09%)	**.546**	**(-10.29%)**	.522	(-14.23%)	.492	(-19.14%)	.493	(-19.02%)	.439	(-27.76%)
DBpedia-Entity	**.337**	**(-12.41%)**	.329	(-14.50%)	.314	(-18.45%)	.287	(-25.54%)	.302	(-21.46%)	.265	(-31.08%)
Avg. Δ Performance	**-10.01%**		-10.88%		-14.55%		-18.78%		-19.45%		-27.07%	

Table 1: The DistilBERT-based students' nDCG@10 and percentage change compared to the teacher model across BEIR evaluation datasets. The models are ordered (left to right) according to their average performance degradation.

and nli-stsb) appears to alleviate misalignment, suggesting the potential interaction between objective and distance functions. Additionally, a clean, pretrained-only student (base-uncased) performs better when a perfect replication of the teacher's fine-tuning setting is not present. Notably, all DistilBERT-based students perform worse than the top-performing extractive students. The 2-layer extractive student outperforms the 6-layer msmarco-dot with a performance gap of 2.5%.

Where are the Well-prepared Students? Student pretraining has been demonstrated to be crucial for knowledge distillation in language understanding tasks (Turc et al., 2019). However, in our asymmetric DE system, the student encoder operates in conjunction with the document encoder of the teacher system, deviating from the conventional distillation procedure. In this case, **the effectiveness of student models lies not in their sheer capability but rather in their compatibility.** Dong et al. (2022) employed t-SNE (van der Maaten and Hinton, 2008) to visualize the embedding spaces of DE encoders in the context of QA tasks. They observed that the two encoders of an asymmetric system tend to map questions and answers onto distinct parameter spaces, even when trained jointly. This observation elucidates the reason why extractive initialization significantly reduces the difficulty of knowledge distillation in our scenario. Furthermore, we extend these findings by demonstrating that aligning the training objectives, similarity measures, and fine-tuning datasets with those of the teacher model can enhance embedding space compatibility. Note that fine-tuning on similar tasks without aligning other elements, e.g., the distance function, may undermine compatibility. Our findings, in conjunction with the results from Kim et al. (2023), suggest that supervision signals play a crucial role in alignment while parameter-sharing inherently addresses this issue.

Inference Efficiency Figure 2 shows that student models initialized from a subset of teacher layers have significantly improved inference speed compared to the teacher model, even with small batch sizes. Considering the marginal performance loss, query encoder distillation provides substantial benefits over the siamese DE encoder.

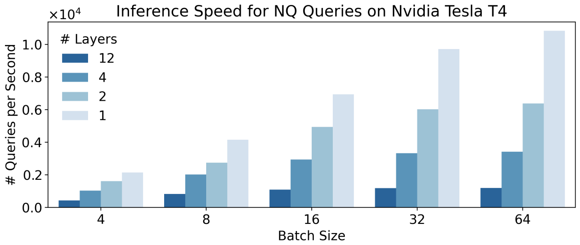

Figure 2: Inference speeds of the distilled query encoders compared to that of the full teacher model. The improvements in inference efficiencies become more drastic as batch size increases.

5 Related Work

Efficient Methods for DE-based IR Systems Various techniques have been proposed to enhance encoder performance in IR systems, including knowledge distillation (Hofstätter et al., 2021a;

Zeng et al., 2022; Lin et al., 2023b; Kim et al., 2023), improved pretraining objectives (Lee et al., 2019; Chang et al., 2020; Gao and Callan, 2021; Izacard et al., 2021), data augmentation (Oguz et al., 2021), better sampling techniques (Lin et al., 2021; Zhang et al., 2021), ensembles (Hofstätter et al., 2021b; Lin et al., 2023a; Ren et al., 2021). However, most of these methods focus on siamese architectures, as asymmetric DE pairs are prone to representation collapse (Leonhardt et al., 2022) or misalignment of embedding spaces (Dong et al., 2022), making them challenging to train. Due to the shared parameters between query and document encoders, practitioners often need to constrain model size for practicality in production settings, despite the significance of larger models for better retrieval and generalization performance (Ni et al., 2022; Yu et al., 2022). Consequently, this constraint often leads to complex training procedures. In contrast, our simple recipe adopts the train-large-distill-small paradigm, offering a straightforward and effective approach to model development and can be adopted out of the box for existing systems.

Embedding Alignment for IR Concurrently, Kim et al. (2023) propose incorporating embedding alignment loss into the supervised distillation pipeline. However, they initialized models from other checkpoints without recognizing the importance of using teacher weights as initialization. Additionally, Campos et al. (2023) suggest minimizing the KL divergence between student and teacher embeddings in an unsupervised manner. Yet, to the best of our knowledge, they do not explore the impact of different layer subsets, whereas our work demonstrates the significant variance caused by such choices.

6 Conclusion

In this work, we leverage the characteristics of typical production DE-based IR systems to propose a minimalistic baseline method for improving online efficiency through embedding-alignment distillation. We explore the significance of student initialization for asymmetric DEs and demonstrate that a "well-prepared" student can achieve over five times improvement in efficiency with only 7.5% average performance degradation. We also observe that "well-prepared" students generally have aligned embedding spaces with their teachers, and a simple approach to construct such students is by extracting the first and last few layers from the teacher models. Our findings aim to enhance the accessibility of neural IR systems and encourage the research community to reassess the trade-offs between method complexity and performance improvements.

Limitations

Limited Experimental Scope Our study's experimental scope was limited to testing distilled student models against a single teacher model. A more comprehensive evaluation would involve multiple teacher models of varying sizes, fine-tuning tasks, and datasets. Additionally, in our experiment with DistilBERT-based student models, incorporating more checkpoints would enable a more thorough comparison across different factors.

Unexplored Embedding Size Variations We kept the embedding size (768) consistent across student models to maintain variable consistency. Future research could investigate student models with different embedding sizes to determine if the observed trends hold true across models of varying widths.

Lack of Error Analysis A common distillation limitation, as noted by Hooker et al. (2020), is the considerable performance decline for certain data subsets. In our study, we couldn't conduct a thorough error analysis due to the lack of appropriate tools for comparing individual data points in retrieval tasks.

Ethics Statement

Although our method improves accessibility for IR systems, it is essential to evaluate whether the proposed approach might introduce biases or unfairness in the retrieval results. As our work lacks extensive error analysis, we cannot entirely rule out the possibility that distilled query encoders may discard certain hard-to-process cases critical for ensuring fairness across various query topics and user groups. A comprehensive error analysis would be beneficial in future research to identify and address potential biases in the distilled query encoders, ultimately fostering fair and unbiased retrieval results for all users.

Acknowledgements

We thank the anonymous reviewers for their time and constructive feedback. Also, we would like to thank Prof. Mark Yatskar for his precious suggestions during the development of this work.

References

Alexandr Andoni, Piotr Indyk, and Ilya P. Razenshteyn. 2018. Approximate nearest neighbor search in high dimensions. *CoRR*, abs/1806.09823.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. ArXiv:1611.09268 [cs].

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2021. Overview of touché 2021: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 450–467. Springer.

Leonid Boytsov and Eric Nyberg. 2020. Flexible retrieval with NMSLIB and flexneuart. *CoRR*, abs/2010.14848.

Daniel Campos, Alessandro Magnani, and ChengXiang Zhai. 2023. Quick Dense Retrievers Consume KALE: Post Training Kullback Leibler Alignment of Embeddings for Asymmetrical dual encoders. ArXiv:2304.01016 [cs].

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. ArXiv:1803.11175 [cs].

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *CoRR*, abs/2002.03932.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. ArXiv:1705.02364 [cs].

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zhe Dong, Jianmo Ni, Dan Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. 2022. Exploring dual encoder architectures for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9419, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Luyu Gao and Jamie Callan. 2021. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. ArXiv:2108.05540 [cs].

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2021a. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. ArXiv:2010.02666 [cs].

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021b. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. ArXiv:2104.06967 [cs].

Sara Hooker. 2020. The hardware lottery. *CoRR*, abs/2009.06489.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. 2023. EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval. ArXiv:2301.12005 [cs].

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *CoRR*, abs/1906.00300.

Jurek Leonhardt, Marcel Jahnke, and Avishek Anand. 2022. Distribution-aligned fine-tuning for efficient neural retrieval. *ArXiv*, abs/2211.04942.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023a. How to Train Your DRAGON: Diverse Augmentation Towards Generalizable Dense Retrieval. ArXiv:2302.07452 [cs].

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, and Nan Duan. 2023b. PROD: Progressive Distillation for Dense Retrieval. ArXiv:2209.13335 [cs].

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Aditya Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank Reddi, and Sanjiv Kumar. 2022. In defense of dual-encoders for neural ranking. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15376–15400. PMLR. ISSN: 2640-3498.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barlas Oguz, Kushal Lakhotia, Anchit Gupta, Patrick S. H. Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, and Yashar Mehdad. 2021. Domain-matched pre-training tasks for dense retrieval. *CoRR*, abs/2107.13602.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv:1908.10084 [cs].

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. ArXiv:2110.07367 [cs].

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor man's bert: Smaller and faster transformer models. *ArXiv*, abs/2004.03844.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. ArXiv:2104.08663 [cs].

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *arXiv preprint arXiv:2210.15212*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.

Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum Learning for Dense Retrieval Distillation. ArXiv:2204.13679 [cs].

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. *CoRR*, abs/2110.03611.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A Technical Details

A.1 Layer-subtraction Schemes

The 4-layer models were initialized using the following schemes: [1,4,7,10], [0,1,10,11], [0,1,2,3], [4,5,6,7], and [8,9,10,11]. The first two schemes were inspired by the results from Fan et al. (2019), which suggested that the input and output layers are often more influential in embedding representations than the middle layers. The latter three schemes were used to validate this intuition and guide our selection schemes for 2-layer and 1-layer initialization. Combinations of 2-layer include [0, 10], [0, 11], [1, 10], and [1, 11]. Layers extracted to make 1-layer models are [0], [1], [10], and [11].

A.2 DistilBERT-based Student Checkpoints

The HuggingFace model cards of the DistilBERT checkpoints adopted in the experiments are listed below. Except for `distilbert-base-uncased`, all other models have `sentence-transformers/` prefix. The same order also maps to Table 1.

1. `msmarco-distilbert-dot-v5`
2. `msmarco-distilbert-base-tas-b`
3. `distilbert-base-uncased`
4. `msmarco-distilbert-base-v3`
5. `distilbert-base-nli-stsb-mean-tokens`
6. `msmarco-distilbert-cos-v5`

B Additional Results

B.1 Other Visualizations

Figure 3 shows the performances of various teachers provided by SentenceTransformers on a subset of BEIR benchmarks. We select the teacher with the highest retrieval performance `msmarco-bert-base-dot-v5`. Figure 4 shows the performances of the student models initialized from other DistilBERT checkpoints. In general, students initialized from pretrained models perform worse than direct layer extraction. Figure 5 and Figure 6 demonstrate that The first few layers and last few layers are more preferable in terms of initialization strategy.

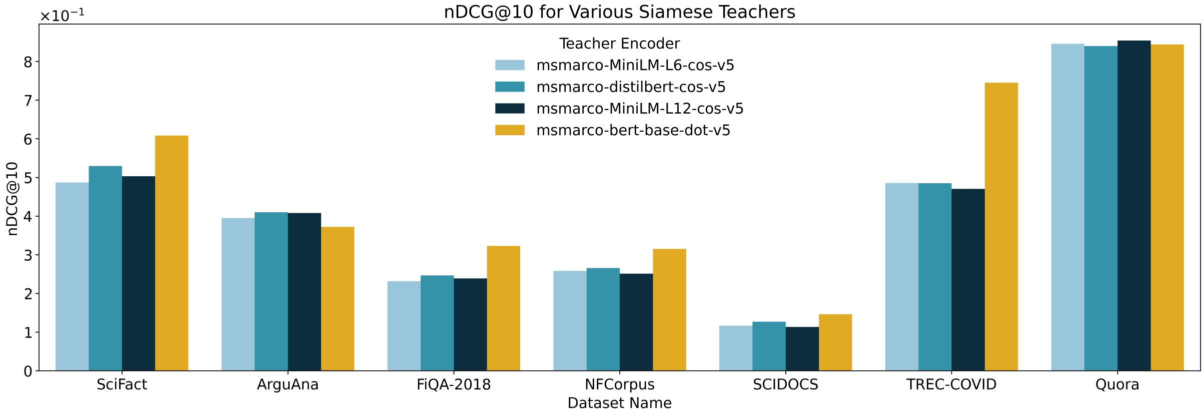

Figure 3: The performances of various teachers provided by SentenceTransformers on a subset of BEIR benchmarks. We select the teacher with the highest retrieval performance.

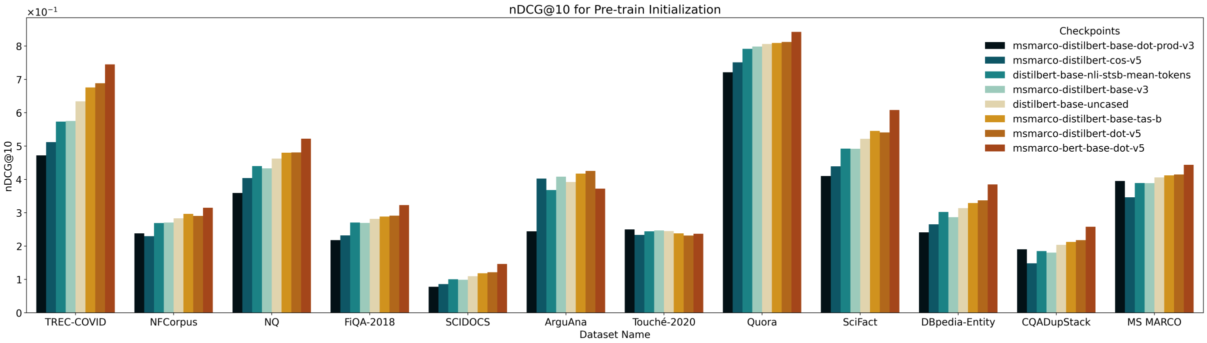

Figure 4: In general, students initialized from pretrained models perform worse than direct layer extraction.

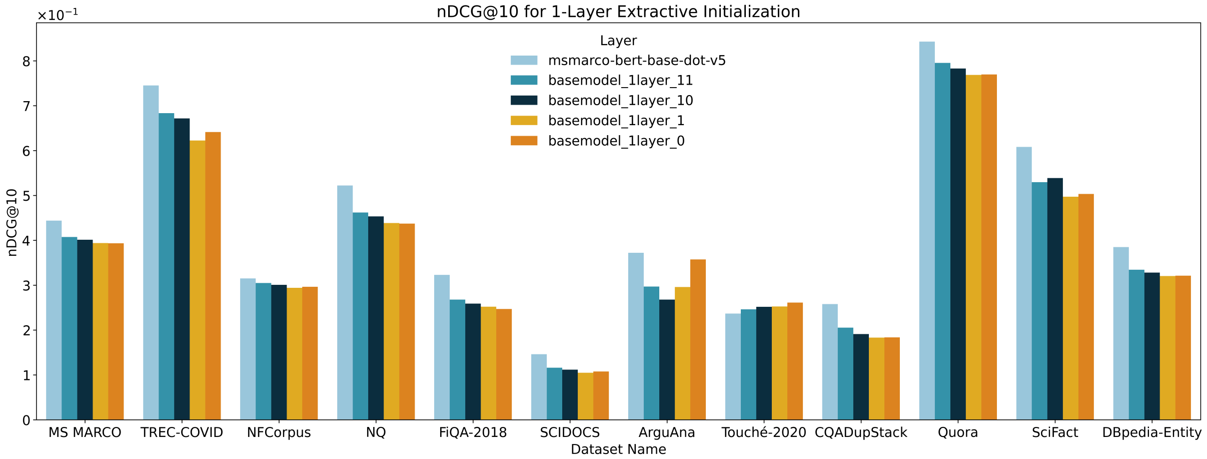

Figure 5: Deeper layers are more preferable than the shallower layers in terms of initialization strategy.

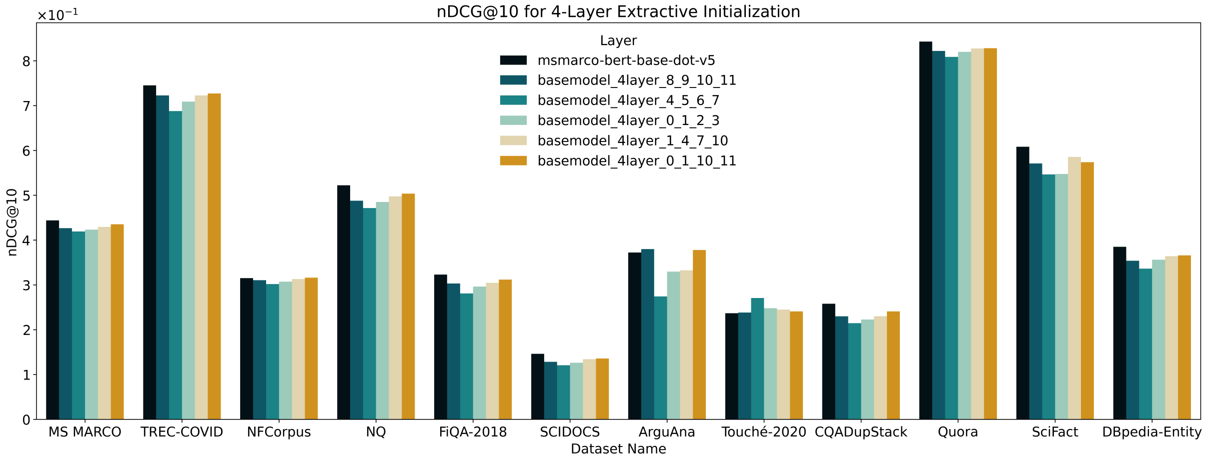

Figure 6: The first few layers and last few layers are more preferable in terms of initialization strategy.