Meta-Learning Adaptive Knowledge Distillation for Efficient Biomedical Natural Language Processing

Abiola Obamuyide and **Blair Johnston**
Strathclyde Institute of Pharmacy and Biomedical Sciences
University of Strathclyde
Glasgow, United Kingdom
firstname.lastname@strath.ac.uk

Abstract

There has been an increase in the number of large and high-performing models made available for various biomedical natural language processing tasks. While these models have demonstrated impressive performance on various biomedical tasks, their training and runtime costs can be computationally prohibitive. This work investigates the use of knowledge distillation, a common model compression method, to reduce the size of large models for biomedical natural language processing. We further improve the performance of knowledge distillation methods for biomedical natural language by proposing a meta-learning approach which adaptively learns parameters that enable the optimal rate of knowledge exchange between the teacher and student models from the distillation data during knowledge distillation. Experiments on two biomedical natural language processing tasks demonstrate that our proposed adaptive meta-learning approach to knowledge distillation delivers improved predictive performance over previous and recent state-of-the-art knowledge distillation methods.

1 Introduction

While there has been an increase in the number of large, pre-trained language models with impressive performance on various biomedical tasks (Shin et al., 2020; Gururangan et al., 2020; Lee et al., 2020; Lewis et al., 2020; Gu et al., 2022), the training and deployment of these models can be computationally prohibitive and time-consuming, especially in resource-constrained settings. The inference latencies and storage costs of these models make their deployment for real-word biomedical applications a challenge. Knowledge distillation (Bucila et al., 2006; Ba and Caruana, 2014; Hinton et al., 2015; Romero et al., 2015), a model compression technique which aims to transfer the performance of a large and computationally inefficient teacher model to a smaller and more efficient student model, has been proposed as a way to reduce the size of large models while retaining their predictive performance.

While a variety of knowledge distillation approaches have been proposed in the literature (Hinton et al., 2015; Sun et al., 2019; Gajbhiye et al., 2021; Zhou et al., 2022), their effectiveness have largely not been evaluated on biomedical natural language processing tasks. In this work, we evaluate the effectiveness of the proposed approaches for knowledge distillation on biomedical NLP tasks. To further enhance performance, we propose an adaptive meta-learning method for distilling large and inefficient biomedical models into more efficient and smaller ones. In experiments conducted on two biomedical natural language processing tasks, we find that our proposed meta-learning approach to knowledge distillation delivers improved predictive performance over previous and recent state-of-the-art knowledge distillation methods.

2 Knowledge Distillation

Knowledge distillation is a model compression method which aims to transfer knowledge from large and accurate but computationally inefficient models to smaller and more efficient models without significant loss in task performance. This is usually achieved by training a smaller and computationally efficient student model to imitate the outputs of a larger and inefficient teacher model with a knowledge distillation objective. For instance, the knowledge distillation objective proposed in Hinton et al. (2015) uses the final output logits produced by the teacher model to transfer its hidden knowledge to the student model. Concretely, given a teacher model T parametrized by θ_T, a student model S parametrized by θ_S and a dataset $\mathcal{D}$ containing N instances $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, the knowledge transfer between teacher and student can be achieved by training the student with a knowledge distillation objective $\mathcal{L}_{KD}$ of the form:

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 131–137
November 20–23, 2022. ©2022 Association for Computational Linguistics

$$
\mathcal{L}_{KD} = \frac{1}{N} \sum_{i=1}^{N} \left[\alpha \mathcal{L}_D\big(T\left(x_i, \theta_T\right), S\left(x_i, \theta_S\right)\big) + \beta \mathcal{L}_{\mathcal{T}}\big(y_i, S\left(x_i, \theta_S\right)\big) \right] \quad (1)
$$

where $\mathcal{L}_D$ is a measure of divergence (such as the KL-divergence) between the teacher's output predictive distribution $T\left(x_i, \theta_T\right)$ and the student's output predictive distribution $S\left(x_i, \theta_S\right)$, $\mathcal{L}_{\mathcal{T}}$ is a task-specific loss function (such as the standard cross-entropy loss), x_i is an input instance with label y_i, while α and β are (scalar) hyper-parameters which determine the relative weight between the two components of the overall knowledge distillation loss function. In other words, α and β determine the rate of knowledge exchange between teacher and student during knowledge distillation. Typically, the values of α and β are manually set before knowledge distillation training, and are then kept fixed throughout. Gou et al. (2021) and Gupta and Agrawal (2022) give further overviews of various knowledge distillation methods.

3 Meta-Learning

Meta-learning, also known as learning to learn (Biggs, 1985; Schmidhuber, 1987; Bengio et al., 1991; Thrun and Pratt, 1998) aims to develop algorithms and models that are able to learn more efficiently with experience, by generalizing from the knowledge of related tasks. These models are able to learn how to learn, by improving their own learning process over time. Various approaches to meta-learning have been proposed and applied in various areas. These approaches include specific-architectures for learning to learn (Vinyals et al., 2016; Snell et al., 2017), learning to update model parameters from background knowledge (Andrychowicz et al., 2016; Ravi and Larochelle, 2017), and gradient-based model-agnostic meta-learning methods (Finn et al., 2017; Nichol et al., 2018; Rothfuss et al., 2021). Example natural language processing tasks to which meta-learning has been applied include machine translation (Gu et al., 2018) and quality estimation (Obamuyide et al., 2021a,b).

Gradient-based model-agnostic meta-learning algorithms such as *MAML* (Finn et al., 2017) often involve a bi-level optimization objective where feedback from the performance of an inner-learner (student model) is used to optimize a meta-learner (teacher model) with the aid of a meta-objective. In other words, in contrast with the teacher model in common knowledge distillation approaches which does not take into account feedback from the student model, the teacher model in meta-learning is able to receive and utilize feedback from the student model in order to improve itself.

Additionally, in knowledge distillation the teacher and student models are usually trained one after the other, with the teacher model trained first and then fixed during the student training. On the other hand, the student and teacher models in meta-learning are trained jointly together in order for them to improve each other.

4 Knowledge Distillation with Meta-Learning

Some works have investigated the use of the bi-level optimization framework in meta-learning to improve knowledge distillation, that is, to employ meta-learning to explicitly optimize the teacher for better knowledge transfer during the knowledge distillation process. For instance, Pan et al. (2021) trained a teacher network that can be adapted across several domains with meta-learning, and then perform standard knowledge distillation to distil the knowledge present in the teacher network into a student network. However, Pan et al. (2021) utilize meta-learning only to train a teacher model, and not throughout knowledge distillation training, thus limiting the generalizability of their approach. In order to enable the teacher model to better transfer knowledge to the student, Zhou et al. (2022) proposed the use of a meta-learning pilot update mechanism which improves the alignment between the student and the teacher in knowledge distillation. In their approach, Zhou et al. (2022) update both the teacher and student throughout the knowledge distillation training process, resulting in improved knowledge distillation performance.

5 Meta-Learning Adaptive Knowledge Distillation

An important limitation in all aforementioned knowledge distillation methods, including those that make use of meta-learning, is that they treat the rate of knowledge exchange between teacher and student (α and β in Equation 1) as fixed during training. This is not ideal, as the optimal rate and level of knowledge exchange between teacher and student should be updated during training to account for their current state.

A relevant and analogous human analogy is that school teachers teach and students learn different curricula depending on the student's educational level (e.g. nursery, primary, secondary, or university student). In most circumstances, it would not be appropriate for a human teacher to be teaching university-level knowledge to primary school students, and vice-versa. Therefore, α and β in knowledge distillation also need to be adaptive and learnable.

As a solution to the aforementioned issue, in this work we propose to treat α and β as learnable parameters which are updated during training. Our work builds on that of Zhou et al. (2022) and further enhances it with learnable α and β. This would allow the values of α and β to change to reflect the needs of the student throughout training. As we demonstrate in the experiments, this change results in improved knowledge distillation performance. We refer to our adapted approach as Meta-Learning Adaptive Knowledge Distillation (MetaAdaptiveKD), and our overall training algorithm is illustrated in Algorithm 1.

Algorithm 1 Meta-Learning Adaptive Knowledge Distillation (MetaAdaptiveKD)

```
Require: Training data $\mathcal{D}^{train}$, holdout data $\mathcal{D}^{hold}$
Require: Teacher $\theta_T$ and student $\theta_S$ models
Require: Teacher $\mu$ and student $\epsilon$ learning rates
Require: Learnable $\alpha$ and $\beta$
 1: Initialize $\theta_T$, $\theta_S$, $\alpha$, $\beta$
 2: while not done do
 3:     Create a copy of student parameter $\theta_S$ to $\theta'_S$
 4:     Sample mini-batches of train data $\boldsymbol{x}_{train} \sim \mathcal{D}^{train}$
 5:     for each $\boldsymbol{x}_{train}$ do
 6:         $\theta'_S \leftarrow \theta'_S - \epsilon \nabla_{\theta'_S} \mathcal{L}_{KD}\left(\boldsymbol{x}_{train}, \theta'_S, \theta_T, \alpha, \beta\right)$
 7:     end for
 8:     Sample mini-batches of holdout data $\boldsymbol{x}_{hold} \sim \mathcal{D}^{hold}$
 9:     for each $\boldsymbol{x}_{hold}$ do
10:         $\alpha \leftarrow \alpha - \mu \nabla_{\alpha} \mathcal{L}_{\mathcal{T}}\big(\boldsymbol{x}_{hold}, \theta'_S\left(\theta_T, \alpha, \beta\right)\big)$
11:         $\beta \leftarrow \beta - \mu \nabla_{\beta} \mathcal{L}_{\mathcal{T}}\big(\boldsymbol{x}_{hold}, \theta'_S\left(\theta_T, \alpha, \beta\right)\big)$
12:         $\theta_T \leftarrow \theta_T - \mu \nabla_{\theta_T} \mathcal{L}_{\mathcal{T}}\big(\boldsymbol{x}_{hold}, \theta'_S\left(\theta_T, \alpha, \beta\right)\big)$
13:     end for
14:     Update $\theta_S \leftarrow \theta_S - \epsilon \nabla_{\theta_S} \mathcal{L}_{KD}\left(\boldsymbol{x}_{train}, \theta_S, \theta_T, \alpha, \beta\right)$
15: end while
```

Our approach described in Algorithm 1 assumes access to both training and holdout datasets[1]. We start by initializing parameters of the teacher and student models, and α and β (line 1). At each training step, we first create a copy of the student parameters (line 3) and sample a number of mini-batches from the training data (line 4). Then for each mini-batch of training data, we update the copy of the student model (lines 5-7). Because the updated student model θ'_S as well as its loss on the holdout set $\mathcal{L}_{\mathcal{T}}\big(\boldsymbol{x}_{hold}, \theta'_S\left(\theta_T, \alpha, \beta\right)\big)$ is now a function of α, β and θ_T, we can use the holdout loss to optimize α, β and θ_T. Thus, we sample mini-batches of data from the holdout set (line 8), and for each mini-batch of holdout data, we update α, β and θ_T (lines 9-13). Finally, we update parameters of the original student model θ_S (line 14). At the end of training, the final student model θ_S can be evaluated and deployed.

6 Experimental Setup and Details

6.1 Datasets

Given our interest in improving the efficiency of biomedical models with knowledge distillation, we conduct experiments on the following two (2) biomedical datasets:

ChemProt: The Chemical Protein Interaction corpus (ChemProt) (Krallinger et al., 2017) is a dataset of PubMed [2] abstracts annotated with interactions between chemical and protein entities. Following common practice, we evaluate on five(5) classes from this dataset.

GAD: The Genetic Association Database (GAD) (Bravo et al., 2014) is a binary relation classification corpus containing a list of gene-disease associations, with the corresponding sentences reporting the association.

Table 1 provides a breakdown of the instances in both datasets.

Dataset	Train	Dev	Test
ChemProt	18035	11268	15745
GAD	4261	535	534
Total	22296	11803	16279

Table 1: Number of instances in the train/dev/test splits of the ChemProt and GAD datasets.

6.2 Teacher and Student Models

Both the teacher and student models are based on the transformer architecture (Vaswani et al., 2017). Specifically, the teacher model is a transformer model with 12 layers and 110M parameters. It is initialized with weights from

[1] The holdout dataset can, for instance, be obtained by splitting from the training set.

[2] https://pubmed.ncbi.nlm.nih.gov

BioLinkBERT$_{base}$ (Yasunaga et al., 2022), a state-of-the-art biomedical transformer model with same architecture as BERT (Devlin et al., 2019), but pre-trained using citation links between PubMed articles. In contrast, the student model is a 6-layer transfomer with 66M parameters. It is initialized with weights from the first six(6) layers of BioLinkBERT$_{base}$.

6.3 Baselines

We compare our approach with the following baselines:

Finetune This is the conventional finetuning approach, where a pre-trained transformer student model is finetuned on each dataset without any knowledge distillation loss. This student model has the same number of parameters as the student model used by our approach and the other baseline knowledge distillation approaches. It is initialized with weights from the first six(6) layers of BioLinkBERT$_{base}$.

KD This is the original knowledge distillation approach proposed in (Hinton et al., 2015). This approach first trains a teacher model, which is then kept fixed while the student is trained with the standard knowledge distillation objective in Equation 1.

PatientKD This approach to knowledge distillation was proposed by Sun et al. (2019). It works by aligning intermediate layer feature representations from the teacher and the student.

MetaDistil This is a recent, state-of-the-art meta-learning approach to knowledge distillation proposed by Zhou et al. (2022). Different from our approach, *MetaDistil* uses fixed values for α and β.

6.4 Experimental Details

Hyper-parameter	**Value**
Learning rate	5e-5
Mini-batch size	8
Max. sequence length	128
Distillation temperature	2
Number of training epochs	20

Table 2: Hyper-parameter values for all compared approaches

Our implementation makes use of Pytorch (Paszke et al., 2019), transformers (Wolf et al., 2020) and higher (Grefenstette et al., 2019) libraries. All compared knowledge distillation approaches, including ours, make use of the same values for hyperparameters such as the number of training epochs, learning rate and batch size. These values were selected by manual search in initial experiments, and are provided in Table 2. Each experiment is repeated across five (5) different random seeds, and we report the average.

6.5 Evaluation

We make use of the F1 measure as performance metric. We repeat each distillation experiment five(5) times and report the average F1 performance of the distilled student on the test set of each dataset.

7 Results and Discussion

The results obtained by our approach and the other knowledge distillation methods on the two biomedical datasets are as shown in Table 3. All student models have nearly twice (x1.94) the inference speed of the teacher model and only about 60% (66M) of the teacher's parameters.

Method	#	Speed↑	F1 (%)	
			ChemProt	GAD
BioLinkBERT (Teacher)	110M	x1.00	77.57	84.39
Finetune	66M	x1.94	72.17	78.53
KD	66M	x1.94	72.49	78.84
PatientKD	66M	x1.94	72.10	78.89
MetaDistil	66M	x1.94	72.73	79.08
MetaAdaptiveKD	66M	x1.94	**73.03**	**79.62**

Table 3: Experimental results on the ChemProt and GAD datasets. The # column represents the number of parameters in each model, while the Speed↑ column represents the speedup of each approach when compared to the teacher model. F1 results of the teacher model are obtained from Yasunaga et al. (2022). The F1 results for all student models including ours are the average of five(5) runs with different random seeds.

In terms of F1 performance of the student models, we find that just finetuning the student model (*Finetune*) without any knowledge distillation objective underperforms all other distillation methods on the GAD dataset and also underperforms all other methods except *PatientKD* on the Chemprot dataset, which demonstrates the effectiveness of knowledge distillation in general. *PatientKD* outperformed *KD* on the GAD dataset but not on the

ChemProt dataset, while *MetaDistil* outperforms *KD* and *PatientKD* on both datasets.

Finally, we find that our approach *MetaAdaptiveKD*, which adaptively learns α and β with meta-learning, outperforms all previous distillation methods on both datasets. The fact that our approach outperforms *MetaDistil* (a meta-learning method which uses fixed α and β) demonstrates the importance of not keeping α and β fixed during knowledge distillation, but instead learning their optimal values from the distillation data during training, as done in our approach.

8 Conclusion

In this work, we proposed a new meta-learning approach to knowledge distillation. In contrast to previous methods which manually set the rate of knowledge exchange between student and teacher and keep them fixed throughout training, our approach learns their optimal values adaptively from the distillation data during training. In experiments conducted on two biomedical datasets, we demonstrated that our approach outperforms previous knowledge distillation methods.

Limitations, Risks and Ethical Considerations

Meta-learning methods for knowledge distillation in general require additional computational resources compared to traditional distillation methods. The *MetaAdaptiveKD* algorithm for knowledge distillation introduced in this work is a meta-learning based approach with similar computational requirements as previous meta-learning methods.

Although this computational cost can be high, it is a one-time investment with long-term returns since it would result in an efficient and more accurate compressed model with reduced run-time costs. In addition, while we have conducted experiments on two english biomedical datasets, *MetaAdaptiveKD* is a generic distillation technique that can be applied to data from other languages and domains.

In terms of risks and ethical considerations, *MetaAdaptiveKD* improves on the performance of previous knowledge distillation methods and does not introduce additional risks and ethical concerns in comparison with these previous methods. Nevertheless, as has been noted in previous work (Hooker et al., 2020), the introduction or amplification of algorithmic biases is a common risk of model compression methods in general, and devising ways of addressing these concerns is an important line of future work.

Acknowledgment

This work is co-funded by the Made Smarter Innovation challenge at UK Research and Innovation (Grant Reference: EP/V062077/1).

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662.

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. 1991. Learning a synaptic learning rule. *IJCNN-91-Seattle International Joint Conference on Neural Networks*, ii:969 vol.2–.

John B. Biggs. 1985. The role of metalearning in study processes. *British Journal of Educational Psychology*, 55:185–212.

Àlex Bravo, Janet Piñero, Núria Queralt, Michael Rautschka, and Laura Inés Furlong. 2014. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 535–541.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia.

Amit Gajbhiye, Marina Fomicheva, Fernando Alva-Manchego, Frédéric Blain, Abiola Obamuyide, Nikolaos Aletras, and Lucia Specia. 2021. Knowledge distillation for quality estimation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5091–5099, Online. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819.

Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. 2019. Generalized inner loop meta-learning. *CoRR*, abs/1910.01727.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

Manish Gupta and Puneet Agrawal. 2022. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data*, 16(4):61:1–61:55.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily L. Denton. 2020. Characterising bias in compressed models. *ArXiv*, abs/2010.03058.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, J. A. Lopez, Umesh K. Nandal, Erin M. van Buel, Ambika Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Patrick S. H. Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 146–157. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*.

Abiola Obamuyide, Marina Fomicheva, and Lucia Specia. 2021a. Bayesian model-agnostic meta-learning with matrix-valued kernels for quality estimation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 223–230, Online. Association for Computational Linguistics.

Abiola Obamuyide, Marina Fomicheva, and Lucia Specia. 2021b. Continual quality estimation with online Bayesian meta-learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 190–197, Online. Association for Computational Linguistics.

Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2021. Meta-KD: A meta knowledge distillation framework for language model compression across domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3026–3036, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Sachin Ravi and Hugo Larochelle. 2017. Optimization As a Model for Few-Shot Learning. In *International Conference on Learning Representations 2017*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua

Bengio. 2015. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. 2021. PACOH: bayes-optimal meta-learning with pac-guarantees. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9116–9126. PMLR.

Jurgen Schmidhuber. 1987. Evolutionary principles in self-referential learning. *On learning how to learn: The meta-meta-... hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4700–4706. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4322–4331. Association for Computational Linguistics.

Sebastian Thrun and Lorien Pratt. 1998. Learning to Learn: Introduction and Overview. In *Learning to Learn*, pages 3–17. Springer US, Boston, MA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049, Dublin, Ireland. Association for Computational Linguistics.