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-world 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 $\theta_T$, a student model $S$ parametrized by $\theta_S$ and a dataset $D$ containing $N$ instances $D = \{(x_i, y_i)\}_{i=1}^N$, the knowledge transfer between teacher and student can be achieved by training the student with a knowledge distillation objective $L_{KD}$ of the form:
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 Laroche, 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

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

where \( \mathcal{L}_D \) is a measure of divergence (such as the KL-divergence) between the teacher’s output predictive distribution \( T(x_i, \theta_T) \) and the student’s output predictive distribution \( S(x_i, \theta_S) \). \( \mathcal{L}_T \) is a task-specific loss function (such as the standard cross-entropy loss), \( x_i \) is an input instance with label \( y_i \), while \( \alpha \) and \( \beta \) are (scalar) hyper-parameters which determine the relative weight between the two components of the overall knowledge distillation loss function. In other words, \( \alpha \) and \( \beta \) determine the rate of knowledge exchange between teacher and student during knowledge distillation. Typically, the values of \( \alpha \) and \( \beta \) 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.
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 $D^{train}$, holdout data $D^{hold}$
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 $x_{train} \sim D^{train}$
5: for each $x_{train}$ do
6: $\theta'_S \leftarrow \theta'_S - \epsilon \nabla_{\theta'_S} L_{KD} (x_{train}, \theta'_S, \theta_T, \alpha, \beta)$
7: end for
8: Sample mini-batches of holdout data $x_{hold} \sim D^{hold}$
9: for each $x_{hold}$ do
10: $\alpha \leftarrow \alpha - \mu \nabla_{\alpha} L_T(x_{hold}, \theta'_S (\theta_T, \alpha, \beta))$
11: $\beta \leftarrow \beta - \mu \nabla_{\beta} L_T(x_{hold}, \theta'_S (\theta_T, \alpha, \beta))$
12: $\theta_T \leftarrow \theta_T - \mu \nabla_{\theta_T} L_T(x_{hold}, \theta'_S (\theta_T, \alpha, \beta))$
13: end for
14: Update $\theta_S \leftarrow \theta_S - \epsilon \nabla_{\theta_S} L_{KD} (x_{train}, \theta_S, \theta_T, \alpha, \beta)$
15: end while

Our approach described in Algorithm 1 assumes access to both training and holdout datasets. 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 $\theta'_S$ as well as its loss on the holdout set $L_T(x_{hold}, \theta'_S(\theta_T, \alpha, \beta))$ is now a function of $\alpha, \beta$ and $\theta_T$, we can use the holdout loss to optimize $\alpha, \beta$ and $\theta_T$. Thus, we sample mini-batches of data from the holdout set (line 8), and for each mini-batch of holdout data, we update $\alpha, \beta$ and $\theta_T$ (lines 9-13). Finally, we update parameters of the original student model $\theta_S$ (line 14). At the end of training, the final student model $\theta_S$ 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 abstracts annotated with interactions between chemical and protein entities. Following common practice, we evaluate on five 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemProt</td>
<td>18035</td>
<td>11268</td>
<td>15745</td>
</tr>
<tr>
<td>GAD</td>
<td>4261</td>
<td>535</td>
<td>534</td>
</tr>
<tr>
<td>Total</td>
<td>22296</td>
<td>11803</td>
<td>16279</td>
</tr>
</tbody>
</table>

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

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

BioLinkBERT\textsubscript{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 transformer with 66M parameters. It is initialized with weights from the first six(6) layers of BioLinkBERT\textsubscript{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\textsubscript{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 $\alpha$ and $\beta$.

6.4 Experimental Details

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>5e-5</td>
</tr>
<tr>
<td>Mini-batch size</td>
<td>8</td>
</tr>
<tr>
<td>Max. sequence length</td>
<td>128</td>
</tr>
<tr>
<td>Distillation temperature</td>
<td>2</td>
</tr>
<tr>
<td>Number of training epochs</td>
<td>20</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>#</th>
<th>Speed↑</th>
<th>ChemProt</th>
<th>GAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioLinkBERT (Teacher)</td>
<td>110M</td>
<td>x1.00</td>
<td>77.57</td>
<td>84.39</td>
</tr>
<tr>
<td>Finetune</td>
<td>66M</td>
<td>x1.94</td>
<td>72.17</td>
<td>78.53</td>
</tr>
<tr>
<td>KD</td>
<td>66M</td>
<td>x1.94</td>
<td>72.49</td>
<td>78.84</td>
</tr>
<tr>
<td>PatientKD</td>
<td>66M</td>
<td>x1.94</td>
<td>72.10</td>
<td>78.89</td>
</tr>
<tr>
<td>MetaDistil</td>
<td>66M</td>
<td>x1.94</td>
<td>72.73</td>
<td>79.08</td>
</tr>
<tr>
<td>MetaAdaptiveKD</td>
<td>66M</td>
<td>x1.94</td>
<td>73.03</td>
<td>79.62</td>
</tr>
</tbody>
</table>

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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$) demonstrates the importance of not keeping $\alpha$ and $\beta$ 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


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.


