SqueezeBioBERT: BioBERT Distillation for Healthcare Natural Language Processing

Hongbin George Du[1] and Yanke Hu[2(✉)]

[1] University of Texas at Austin, Austin, TX 78712, USA
dugeorge21@utexas.edu
[2] Humana, Irving, TX 75063, USA
yhu@humana.com

Abstract. Healthcare text mining attracts increasing research interest as electronic health record (EHR) and healthcare claim data have skyrocketed over the past decade. Recently, deep pre-trained language models have improved many natural language processing tasks significantly. However, directly applying them to healthcare text mining won't generate satisfactory results, because those models are trained from generic domain corpora, which contains a word distribution shift from healthcare corpora. Moreover, deep pre-trained language models are generally computationally expensive and memory intensive, which makes them very difficult to use on resource-restricted devices. In this work, we designed a novel knowledge distillation method, which is very effective for Transformer-based models. We applied this knowledge distillation method to BioBERT [5], and experiments show that knowledge encoded in the large BioBERT can be effectively transferred to a compressed version of SqueezeBioBERT. We evaluated SqueezeBioBERT on three healthcare text mining tasks: named entity recognition, relation extraction and question answering. The result shows that SqueezeBioBERT achieves more than 95% of the performance of teacher BioBERT on these three tasks, while being 4.2X smaller.

Keywords: Natural language processing · Transformer · Deep learning · Knowledge distillation · Healthcare

1 Introduction

Healthcare text mining attracts increasing research interest as electronic health record (EHR) and healthcare claim data have skyrocketed over the past decade. Recently, deep pre-trained language models, such as BERT [2] and GPT [3], have improved many natural language processing tasks significantly. However, it won't give satisfactory results by directly applying those deep pre-trained language models to healthcare text mining. One important reason is that those models are trained from generic domain corpora, which contains a word distribution shift from healthcare corpora. Moreover, deep pre-trained language models

S. Chellappan et al. (Eds.): CSoNet 2020, LNCS 12575, pp. 193–201, 2020.
https://doi.org/10.1007/978-3-030-66046-8_16

are difficult to use on resource-restricted devices due to their huge computation complexity and memory consumption. It's very important to have embedded models that can directly inference on mobile for healthcare related apps in the US because: 1) it can provide better user experience at poor cell phone signal locations, and 2) it doesn't require users to upload their health sensitive information onto the cloud. In the US, health related data are only allowed to upload to the cloud by mobile apps being developed by certified institutes, which greatly suppresses the enthusiasm of developing healthcare mobile apps from individual developers. There are some model compression techniques developed recently for generic BERT [6–8], but there doesn't exist a small and efficient enough pre-trained language model in healthcare domain. In this work, we developed SqueezeBioBERT. SqueezeBioBERT has 3 transformer layers, and inference much faster while being accurate on healthcare natural language processing tasks. Our contributions are summarized as below:

- We designed a novel knowledge distillation method, which is very effective for compressing Transformer-based models without losing accuracy.
- We applied this knowledge distillation method to BioBERT [5], and experiments show that knowledge encoded in the large BioBERT can be effectively transferred to a compressed version of SqueezeBioBERT.
- We evaluated SqueezeBioBERT on three healthcare text mining tasks: name entity recognition, relation extraction and question answering. The result shows that SqueezeBioBERT achieves more than 95% of the performance of teacher BioBERT on these three tasks, while being 4.2X smaller.

2 Transformer Layer

As the foundation of modern pre-trained language models [2–4], transformer layer [1] can capture long-term dependencies of the input tokens with attention mechanism. A typical transformer layer contains two major components: *multi-head attention* (MHA) and *feed-forward network* (FFN).

2.1 Multi-head Attention

Practically, we calculate the attention function on a query set $\mathbf{Q}$, with key set $\mathbf{K}$ and value set $\mathbf{V}$. The attention function can be defined as below:

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \tag{1}$$

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{A})\mathbf{V} \tag{2}$$

where d_k denotes the dimension of $\mathbf{K}$.

Multi-head attention will jointly train the model from different representation subspaces. It is denoted as below:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, ..., head_h)\mathbf{W} \tag{3}$$

where h denotes attention head number, $head_i$ is computed by Eq. (2), and W is the linear parameter weight.

2.2 Feed-Forward Network

After multi-head attention, a fully connected feed-forward network will follow, which is denoted as below:

$$FFN(x) = max(0, x\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2 \tag{4}$$

3 Knowledge Distillation

A very common way to boost the performance of a machine learning algorithm is to train several models, and then ensemble. Deep learning models are generally heavy neural networks, so it's normally considered too computationally expensive and inefficient to deploy the ensemble of deep neural networks in the production environment. [9] first proposed *Knowledge Distillation* and showed the possibility of compressing the function learned from a large complex model into a much smaller and faster model without significant accuracy loss [10]. As deep learning models are becoming more and more complex, knowledge distillation has shown its power of transferring the knowledge from a group of specialist networks to a single model [10–12].

Formally, *Knowledge Distillation* process can be defined as the process of minimizing the loss function between the a large teacher network **T** and a small student network **S** as below:

$$\mathcal{L}_{KD} = \sum_{x \in X} L(f^T(x), f^S(x)) \tag{5}$$

where L denotes the loss function to evaluate the difference between **T** and **S**, x is the token input, X is the training set, f^T denotes the output of the teacher network **T** and f^S denotes the output of the student network **S**.

4 BioBERT

BioBERT [5], with almost the same structure as BERT and pre-trained on biomedical domain corpora such as PubMed Abstracts and PMC full-text articles, can significantly outperform BERT on biomedical text mining tasks.

BioBERT has been fine-tuned on the following three tasks: Named Entity Recognition (NER), Relation Extraction (RE) and Question Answering (QA). NER is to recognize domain-specific nouns in a corpus, and precision, recall and F1 score are used for evaluation on the datasets listed in Table 1. RE is to

classify the relationships of named entities, and precision, recall and F1 score are used for evaluation on the datasets listed in Table 2. QA is to answer a specific question in a given text passage, and strict accuracy, lenient accuracy and mean reciprocal rank are used for evaluation on BioASQ factoid dataset [24].

Table 1. BioBERT Named Entity Recognition evaluation datasets

Dataset	Entity type
NCBI Disease [13]	Disease
2010 i2b2/VA [14]	Disease
BC5CDR [15]	Disease/Drug
BC4CHEMD [16]	Drug
BC2GM [17]	Gene
JNLPBA [18]	Gene
LINNAEUS [19]	Species
Species-800 [20]	Species

Table 2. BioBERT Relation Extraction evaluation datasets

Dataset	Entity type
GAD [21]	Gene/Disease
EU-ADR [22]	Gene/Disease
CHEMPROT [23]	Protein

5 BioBERT Distillation

In this section, we developed a novel distillation method for BioBERT. Experiments show that knowledge encoded in the large BioBERT can be effectively transferred to the compressed version of SqueezeBioBERT.

Figure 1 shows the overview of the proposed knowledge distillation method. Supposing that the teacher BioBERT has M transformer layers and the student SqueezeBioBERT has N transformer layers, we distillated BioBERT both on transformer layers and task-specific layers.

Transformer layer distillation consists of multi-head attention distillation and feed forward network distillation. For multi-head attention distillation, we combine Eqs. (2), (3) and (5), and use the mean squared error (MSE) as the loss function since it's more suitable for regression tasks. Thus, the multi-head attention distillation process is denoted as below:

$$\mathcal{L}_{MHA} = \frac{1}{h}\sum_{i=1}^{h} MSE(M_i^T, M_i^S) \tag{6}$$

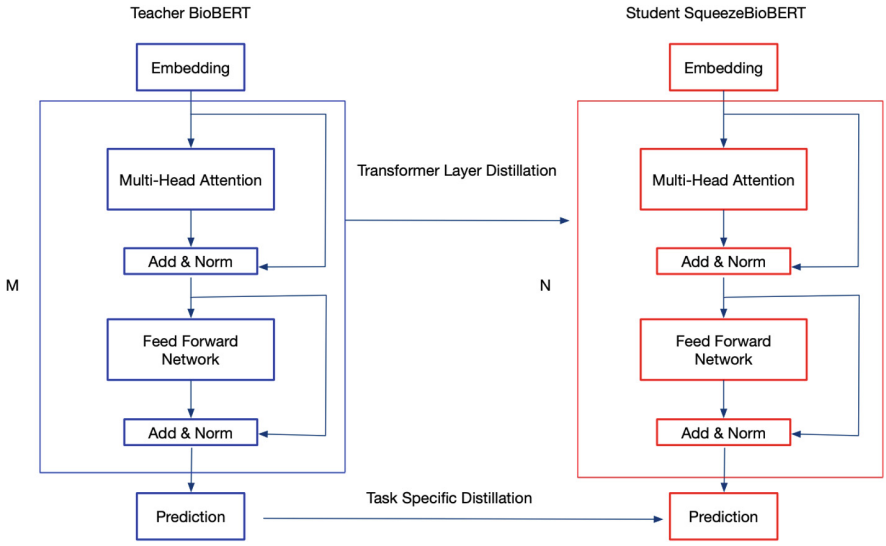

Fig. 1. The overview of distillation from BioBERT to SqueezeBioBERT

where h denotes the number of attention heads, M_i^S denotes the output of i-th student attention head, and M_i^T denotes the output of i-th teacher attention head.

For feed forward network distillation, we can use a single linear transformation W_{FFN} to transform the output of the teacher network into the student network. Thus, the feed forward network distillation process is denoted as below:

$$\mathcal{L}_{FFN} = MSE(O_{MHA}^T W_{FFN}, O_{MHA}^S) \tag{7}$$

For task-specific prediction layer distillation, we use *softmax cross-entropy* as the loss function, since it's more suitable for classification tasks. Thus, the task-specific prediction layer distillation is denoted as below:

$$\mathcal{L}_{pred} = -softmax(O_{FFN}^T) log(softmax(O_{FFN}^S)) \tag{8}$$

In summary, Eqs. (6), (7) and (8) describes the overall procedure of the BioBERT distillation process.

6 Experiments

We use BioBERT-Base v1.1 [25] as our source model, and distillated it to SqueezeBioBERT on the same three healthcare NLP tasks. BioBERT-Base v1.1

Table 3. Named Entity Recognition metrics comparison

Dataset	Metrics	BioBERT-Base v1.1	SqueezeBioBERT v1.0
NCBI Disease [13]	Precision	88.22	86.19
	Recall	91.25	88.42
	F1	89.71	87.74
2010 i2b2/VA [14]	Precision	86.93	83.97
	Recall	86.53	83.85
	F1	86.73	85.26
BC5CDR [15]	Precision	86.47	82.84
	Recall	87.84	84.94
	F1	87.15	85.23
BC4CHEMD [16]	Precision	92.80	89.83
	Recall	91.92	87.78
	F1	92.36	90.33
BC2GM [17]	Precision	84.32	82.46
	Recall	85.12	83.16
	F1	84.72	82.11
JNLPBA [18]	Precision	72.24	69.93
	Recall	83.56	81.67
	F1	77.49	75.09
LINNAEUS [19]	Precision	90.77	89.41
	Recall	85.83	84.29
	F1	88.24	85.15
Species-800 [20]	Precision	72.80	70.47
	Recall	75.36	74.38
	F1	74.06	72.73

Table 4. Relation extraction metrics comparison

Dataset	Metrics	BioBERT-Base v1.1	SqueezeBioBERT v1.0
GAD [21]	Precision	77.32	74.69
	Recall	82.68	81.61
	F1	79.83	77.04
EU-ADR [22]	Precision	77.86	75.37
	Recall	83.55	80.54
	F1	79.74	77.19
CHEMPROT [23]	Precision	77.02	75.79
	Recall	75.90	72.41
	F1	76.46	74.01

Table 5. Question answering metrics comparison

Dataset	Metrics	BioBERT-Base v1.1	SqueezeBioBERT v1.0
BioASQ 4b [24]	Strict Accuracy	27.95	27.31
	Lenient Accuracy	44.10	42.12
	Mean Reciprocal Rank	34.72	33.26
BioASQ 5b [24]	Strict Accuracy	46.00	43.58
	Lenient Accuracy	60.00	58.08
	Mean Reciprocal Rank	51.64	49.94
BioASQ 6b [24]	Strict Accuracy	42.86	41.83
	Lenient Accuracy	57.77	56.48
	Mean Reciprocal Rank	48.43	46.83

has 12 transformer layers and 109M weights. SqueezeBioBERT has 3 transformer layers and 26M weights.

NER results are show in Table 3, RE results are show in Table 4, and QA results are show in Table 5. From the results, we can see that SqueezeBioBERT is 4.2X smaller than BioBERT, but still achieves more than 95% accuracy performance of the teacher BioBERT on the three NLP tasks. This proves the efficiency of the proposed method of transferring knowledge encoded in the large BioBERT to the compressed version of SqueezeBioBERT.

7 Conclusion

Although recent deep pre-trained language models have greatly improved many natural language processing tasks, they are generally computationally expensive and memory intensive, which makes them very difficult to use on resource-restricted mobile or IoT devices. Embedded models that can directly inference on mobile is important for healthcare related apps in the US because: 1) it can provide better user experience at poor cell phone signal locations, and 2) it doesn't require users to upload their health sensitive information onto the cloud. In this paper, we designed a novel knowledge distillation method, which is very effective for compressing Transformer-based models without losing accuracy. We applied this knowledge distillation method to BioBERT, and experiments show that knowledge encoded in the large BioBERT can be effectively transferred to a compressed version of SqueezeBioBERT. We evaluated SqueezeBioBERT on three healthcare text mining tasks: name entity recognition, relation extraction and question answering. The result shows that SqueezeBioBERT achieves more than 95% of the performance of teacher BioBERT on these three tasks, while being 4.2X smaller.

Acknowledgement. This work was supported by Humana.

References

1. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
4. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)
5. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**, 1234–1240 (2019)
6. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
7. Jiao, X., et al.: TinyBERT: distilling BERT for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
8. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D.: MobileBERT: a compact task-agnostic BERT for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
9. Bucilua, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, New York, NY, USA, pp. 535–541. ACM (2006)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Urban, G., et al.: Do deep convolutional nets really need to be deep (or even convolutional)? In: Proceedings of the International Conference on Learning Representations (2016)
12. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2654–2662 (2014)
13. Dogan, R.I., et al.: NCBI disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform. **47**, 1–10 (2014)
14. Uzuner, O., et al.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J. Am. Med. Inform. Assoc. **18**, 552–556 (2011)
15. Li, J., et al.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database **2016**, 1–10 (2016)
16. Krallinger, M., et al.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. J. Cheminform. **7**, 1–17 (2015)
17. Smith, L., et al.: Overview of BioCreative II gene mention recognition. Genome Biol. **9**, 1–19 (2008). https://doi.org/10.1186/gb-2008-9-s2-s2
18. Kim, J.-D., et al.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), Geneva, Switzerland, pp. 73–78 (2004). COLING. https://www.aclweb.org/anthology/W04-1213
19. Gerner, M., et al.: LINNAEUS: a species name identification system for biomedical literature. BMC Bioinform. **11**, 85 (2010). https://doi.org/10.1186/1471-2105-11-85
20. Pafilis, E., et al.: The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. PLoS One **8**, e65390 (2013)

21. Bravo, A., et al.: Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinform. **16**, 55 (2015). https://doi.org/10.1186/s12859-015-0472-9
22. Van Mulligen, E.M., et al.: The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. J. Biomed. Inform. **45**, 879–884 (2012)
23. Krallinger, M., et al.: Overview of the BioCreative VI chemical-protein interaction track. In: Proceedings of the BioCreative VI Workshop, Bethesda, MD, USA, pp. 141–146. https://doi.org/10.1093/database/bay073/5055578 (2017)
24. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
25. https://github.com/naver/biobert-pretrained