FEDERATED PRETRAINING AND FINE TUNING OF BERT USING CLINICAL NOTES FROM MULTIPLE SILOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large scale contextual representation models, such as BERT, have significantly advanced natural language processing (NLP) in recently years. However, in certain area like healthcare, accessing diverse large scale text data from multiple institutions is extremely challenging due to privacy and regulatory reasons. In this article, we demonstrate the possibility to both pre-train and fine tune BERT models in a federated manner using clinical texts from different silos without moving the data.

1 INTRODUCTION

In recent years, natural language processing (NLP) has been revolutionized by large contextual representation models pre-trained with large amount of data such as ELMo Peters et al. (2018) and BERT Devlin et al. (2018). Compared with traditional word level representations, such as word2vec Mikolov et al. (2013), GloVe Pennington et al. (2014) and fastText Bojanowski et al. (2017), which assign a representation vector to a word regardless of their surround context, contextual representation method models context of a word. For example, the embedings of the word "bank" are different between the context of "Bank of England" and "the bank of the Charles river" ELMo Peters et al. (2018) uses a recurrent neural network model to learn information of contexts of words from unlabelled texts. The trained contextual embedding were then used for downstream tasks. More recently, another contextual word representation model based on transformer, called BERT, was released, and has become widely used for many NLP tasks Vaswani et al. (2017); Devlin et al. (2018).

Recent studies have shown that training contextual representations in texts from specific domains improves power of the model by capturing domain specific linguistic characteristics. Texts from biomedical publications and electronic medical record have been used to pre-train BERT models for NLP task in this domain and showed considerable improvement in many downstream tasks. The BERT models trained with large amount of clinical notes are called clinical BERT Lee et al. (2019); Alsentzer et al. (2019); Si et al. (2019). Up to date, there have been several version of clinical BERT models trained on open source clinical notes Johnson et al. (2016); Si et al. (2019); Alsentzer et al. (2019).

However, sharing clinical data is difficult due to privacy and regulatory issues. Training NLP models in a federated manner is a good option to overcome these challenges. In this article, we conduct a proof-of-concept study to train BERT across clinical notes from multiple sites in a federated manner without moving notes outside of their silos. Our main contribution include:

1. We show that it is possible to conduct federated pre-training of BERT model using clinical notes from multiple silos without data transfer.

2. We show that it is possible to do federated fine tuning of BERT model for different down stream tasks such as name entity recognition(NER).

2 RELATED WORK

In 2019, Lee et al. published a BERT model pre-trained on biomedical PubMed abstracts and PubMedcal central articles Lee et al. (2019). Later in the year, Alsentzer et al. (2019) and Si et al. (2019) published almost at the same time BERT models pre-trained trained on publicly available

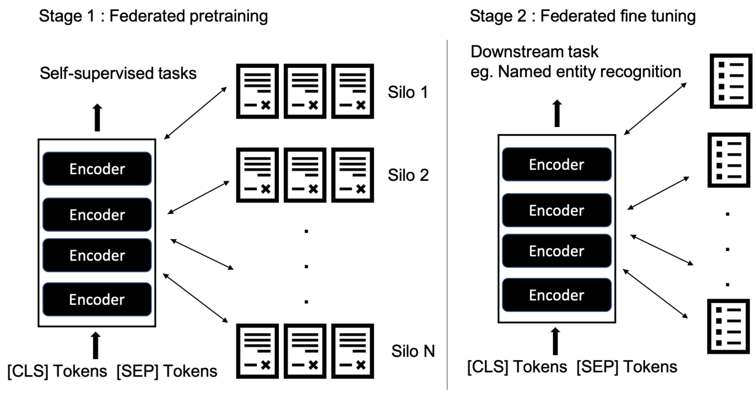

Figure 1: Federated BERT model trained can be conducted at both pre-training and fine tuning stages. In federated pre-training stage, unlabelled clinical texts from different silos, such as hospitals, are used for self-supervised training to learn domain-specific linguistic characterises in a federated manner with moving data outside their silos. In task-specific fine tuning stage, pre-trained BERT model were further trained using labelled texts from different silos in a federated manner.

clinical notes either starting from trained parameters of original BERT or BioBERT model and show improvement in clinical NLP tasks. There has been only limited work on federated NLP works in the clincal domain. In 2019, Liu et al. (2019) published a study doing federated training of machine learning models on clinical notes and used the model for patient level phenotyping based on Concept Unique Identifiers (CUIs). In comparison, in out study, we use raw clinical texts instead of CUIs and, to the best of our knowledge, we are the first one to use contextual representation methods for federated clinical NLP tasks.

3 Methods

3.1 Clinical data

In this study, the publicly available MIMIC-III corpus Johnson et al. (2016) was used for contextual representation learning. This corpus contains information for more than 58,000 admissions for more than 45,000 patients admitted to Beth Israel Deaconess Medical Center in Boston between 2001 and 2012.

Different type of clinical notes are available in this corpus including discharge summaries, nursing notes and so on. We included only discharge summaries in our study as previous studies have shown that performance of a model trained on only discharge summaries in this corpus is only marginally worse than model trained on all notes types Alsentzer et al. (2019).

3.2 Federated BERT training

Our goal is to develop methods for federated learning for both (1) pre-training models to capture linguistic characteristics of clinical text, and (2) fine-tuning for a specific down stream task, here named entity recognition (NER) from clinical notes. These methods will allow researchers and clinicians to utilize data from multiple health care providers to train contextual representation models like BERT, without the need to share the data directly, obviating issues related to data transfer and privacy.

In the following sections, we first describe data processing and a simple notes pre-processing step. We then discuss the method for federated pre-training of BERT model and the method for fine tuning.

In pre-training sage, clinical notes MIMIC-III corpus were randomly split into 5 groups to mimic 5 different silos by patient. The prepossessing and tokenization pipeline from Alsentzer et al. 2019 Alsentzer et al. (2019) was adapted.

To train the BERT model, we simulated sending out models with identical initial parameters to all silos. At each silo, a model was trained using only data from that site. Only model parameters of the models were then sent back to the analyzer for aggregation. An updated model is generated by averaging the parameters of models distributively trained, weighted by sample size Konečnỳ et al. (2016); McMahan et al. (2016). In this study, sample size is defined as the number of patients in the pre-training stage and number of notes in fine-tuning stage.

After model aggregation, the updated model was sent out to all sites again to repeat the global training cycle. Formally, the weight update is specified by:$Q_{ag}^t = \sum_{k=1}^{K} \frac{n_k}{N} Q_k^t$

where Q_{ag}^t is the parameter of aggregated model at global cycle t, K is the number of data silos. n_k is the number of samples at the k^{th} site, N is the total number of samples across all sites, and Q_k is the parameters learned from the k^{th} data site alone. t is the global cycle number in the range of [1,T].

In the pre-training stage, in each global cycle the BERT model was trained for one epoch through all clinical notes data at each of the silos using the default settings from the original BERT publication Devlin et al. (2018). A total of 15 global cycles were run. The downstream task performance plateaued out around cycle 10. Therefore, BERT model federately trained for 10 global cycles was used for down stream tasks. Centralized fine-tuning of the i2b2 NER tasks plateaued after 4 epochs, with the learning rate set at $2e-5$ and a batch size of 32 Alsentzer et al. (2019). When conducting federated fine tuning using the same settings as centralized fine tuning, one epoch of training was conducted in each global cycle and a total of 6 global cycles were conducted when the performance plateaued.

Each pre-training global cycle took around 4 hours on a Tesla K80 GPU which has a single precision GFLOPs of 5591–8736. A single federated fine tuning global took around 20 mins on the same device.

3.3 Downstream tasks

Clinical BERT pre-trained on MIMIC corpus has been reported to have superior performance on NER tasks in Inside-Outside-Beginning (IOB) format Ramshaw & Marcus (1999) using i2b2 2010 Uzuner et al. (2011) and 2012 Sun et al. (2013) dataAlsentzer et al. (2019).Original training/development/test splits in the challenges were used. The NER tasks classify if a token is within a span of a class, outside spans of any classes or at beginning of a span of a class. For example, the sentence "He has severe asthma" is labelled as "Null, Null,B-problem and I-problem". This means the first two words are not in any classes, the third word is the beginning of span of class "problem" and the forth word is in the class "problem".

In this study we use these two data sets for the NER tasks and for fine tuning.To conduct federated fine tuning of BERT model, training notes for downstream task were randomly split into 5 silos by note.

4 Experiments and results

4.1 Experimental design

In order to understand whether large size contextual language representation model like BERT can be pre-trained and fine tuned in a federated manner using data from different silos, we designed and conducted the following 6 experiments.

First of all, we looked at the scenarios no domain specific BERT model pre-training was conducted. In those cases, the parameters (checkpoint) from original BERT base model trained on Books Corpus and English Wikipedia Zhu et al. (2015); Devlin et al. (2018). We also looked at the scenarios where BERTbase model was pre-trained by MIMIC3 discharge summaries in a centralized manner, also called clinical BERT Alsentzer et al. (2019). Lastly, we look at the scenarios where BERTbase model was pre-trained by MIMIC3 discharge summaries in a federated manner, called federated clinical BERT. For each of these conditions, we fine tuned pre-trained BERT models for downstream tasks using a centralized or federated learning strategy.

Table 1: Performance on i2b2 NER tasks

Task	Pretraining	fine tuning	Prec	Rec	F1
i2b2_2010 NER	BERTbase	Centralized	0.775	0.794	0.784
i2b2_2010 NER	ClinicalBERT	Centralized	0.844	0.873	0.858
i2b2_2010 NER	Fed_ClinicalBERT	Centralized	0.81	0.831	0.820
i2b2_2010 NER	BERTbase	Federated	0.73	0.703	0.716
i2b2_2010 NER	ClinicalBERT	Federated	0.819	0.868	0.843
i2b2_2010 NER	Fed_ClinicalBERT	Federated	0.811	0.806	0.808
i2b2_2012 NER	BERTbase	Centralized	0.704	0.754	0.728
i2b2_2012 NER	ClinicalBERT	Centralized	0.711	0.774	0.741
i2b2_2012 NER	Fed_ClinicalBERT	Centralized	0.708	0.764	0.735
i2b2_2012 NER	BERTbase	Federated	0.667	0.707	0.686
i2b2_2012 NER	ClinicalBERT	Federated	0.697	0.769	0.731
i2b2_2012 NER	Fed_ClinicalBERT	Federated	0.687	0.745	0.715

To summarize, six experiments were conducted:

1. Original BERT + centralized fine tuning
2. Clinical BERT + centralized fine tuning
3. Federated clinical BERT + centralized fine tuning
4. Original BERT + federated fine tuning
5. Clinical BERT + federated fine tuning
6. Federated clinical BERT + federated fine tuning.

4.2 Experimental results

The results of our experiments are shown in Table 1. In experiment 1, when original BERT was used without domain-specific pre-training on clinical notes and downstream task specific fine tuning was conducted in a centralized manner, a F1 score of 0.784 for NER task was achieved for i2b2 2010 and 0.728 for i2b2 2012. In experiment 2, where BERT was pre-training using clinical notes in a centralized manner and fine tuning for NER tasks were also conducted in a centralized manner, the F1 scores for NER tasks were 0.858 for i2b2 2010 and 0.741 for i2b2 2012. In experiment 3, where BERT was pre-trained in a federated manner and fine tuned using centralized data, the F1 scores for NER tasks were 0.820 for i2b2 2010 and 0.735 for i2b2 2012.

In Experiment 4 where original BERT was not pre-trained using clinical notes and fine tuning was conducted in a federated manner, the F1 scores for NER tasks were 0.716 for i2b2 2010 and 0.686 for i2b2 2012. In comparison, if the BERT is trained using centralized clinical notes before federated fine tuning (Experiment 5), F1 scores of i2b2 2010 NER task improved to 0.843 and F1 scores of i2b2 2012 NER task improved to 0.731. While the both the pre-training and fine tuning were conducted in a federated, as in Experiment 6, the F1 score were 0.808 and 0.715 for i2b2 2010 and i2b2 2012 respectively, which is superior to BERT model without pre-training. We made our BERT model federatedly trained with clinical notes publicly available at XXXX and all the codes at XXXX.

5 Discussion and conclusion

There are several limitations in this study. Firstly of all, due to limit of data access, we used clinical notes from a single healthcare system to simulate different silos. Secondly, due to limitation of resources, we did not compare effects of factors such as number of silos and non-IID data distribution in this proof-of-concept study. In a future step, we would like to conduct evaluation on more downstream tasks using data from different healthcare systems, and explore possible medical insights by analyzing model trained form different silos.

REFERENCES

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

Dianbo Liu, Dmitriy Dligach, and Timothy Miller. Two-stage federated phenotyping and patient representation learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 283–291, 2019.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pp. 157–176. Springer, 1999.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*, 2019.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.