Fine-tuning BERT for Medical Natural Language Inference

W266 Summer 2019 Project Report

Ted Pham (Tues 630pm), Subha Vadakkumkoor (Thurs 630pm)

Part of final project for NLP with Deep Learning Class, UC Berkeley

# Abstract

Deep learning methods for natural language inference have reached accuracy over 90% on benchmark datasets such as SNLI and MultiNLI. However, these models don’t generalize well outside the datasets or domains they were trained on and require further refinement steps to become effective for tasks on different domains. In this project, we explored applications of Bi-LSTM and BERT along with the use of word embeddings and fine-tuning on the clinical notes (MIMIC-III) derived Medical Natural Language Inference dataset. With an aim to improve the originally reported accuracy of 73.5% , we achieved: 1. 76.6% accuracy with Bi-LSTM and PubMed+MIMIC-III word embedding; 2. Comparable state-of-the-art result of 82.4% by fine-tuning BioBERT v1.1 (BERT base + 1M PubMed article abstracts). Our experiments show that the clinical data corpora (MIMIC-III) or a PubMed full-text corpus (as opposed PubMed abstracts alone) might not be necessary for achieving state-of-the-art scores on the MedNLI dataset.

# Introduction

Recognizing entailment, contradiction, and neutral relationships between two sentences has been a common task in Natural Language Inference (NLI) research. The task is crucial to developing computational systems that can perform semantic parsing, commonsense reasoning, information extraction, and question answering. However, entailment task is challenging because the developed algorithms have to account for both lexical and compositional semantics. The state-of-the art neural network based methods for this include ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and most recently XLNet (Yang et al. 2019) pushing the accuracy on large-scale benchmark datasets such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) to above 90%. Talman and Chatzikyriakidis (2019) reported that ELMo and BERT do not generalize well to datasets outside their respective trained and tested datasets. They attributed the discrepancy in out of domain accuracy to the differences in domains and sampling methods. However, two strategies exist for applying pre-trained language representations to downstream NLI tasks. The first one is to include additional features to the downstream task, while the other applies fine-tuning to the pre-trained parameters with the downstream task’s training data. In this project, we focus on fine-tuning BioBERT v1.1 for the Medical Natural Language Inference Dataset (Romanov and Shivade 2018) in the clinical domain, with the main goal of improving the original classification accuracy of 73.5%.

# Background

## MIMIC-III and MedNLI Dataset

MIMIC-III is a freely accessible critical care database, containing sanitized records of 52,423 hospital admissions. The data includes clinical notes, outcomes, patients lab results, vital signs, medications, etc associated with these admissions. The only barrier to obtain access to MIMIC-III is to take a research ethics course and providing a reference. The MedNLI dataset was derived from the Past Medical History in MIMIC-III, comprising of 11,232 training pairs, 1,395 development paris and 1,422 test pairs. Unlike SNLI and MultiNLI, which were crowd-sourced, the MedNLI dataset was annotated by four clinicians using 4,683 premises. Three classes: entailment, contradiction and Neutral are balanced in train, dev, and test sets along with the variance in medical topics. Some examples are shown in Table 1.

|  |  |  |
| --- | --- | --- |
| **Label** | **Premise** | **Hypothesis** |
| Entailment | Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4. | Patient has elevated Cr |
| Contradiction | Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4. | Patient has normal Cr |
| Neutral | Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4. | Patient has elevated BUN |

**Table 1:** Examples of MedNLI dataset

## State-of-the-art Results on MedNLI Dataset

Lee et al. introduced BioBERT, essentially BERT pre-trained on general domain corpora then pre-trained on biomedical domain corpus. They demonstrated that BioBERT significantly outperforms previous state-of-the-art models on all of biomedical data. They released two variants of BioBERT pre-trained weights: 1. BioBERT v1.0 ( 200K PubMed abstracts and 270K PubMed Central full-text articles), 2. BioBERT v1.1 (1M PubMed abstracts).

As of April 2019, the state-of-the-art accuracy on MedNLI was 73.5% from the original Romanov and Shivade paper. The model for this was Bi-LSTM trained on MIMIC-III corpora. After we proposed our project idea and obtained access to the dataset, several papers came out in June 2019 with significant improvement in accuracy. The top two that we are aware of are: 1. Peng et al. from the National Institutes of Health published NCBI BERT, base BERT trained on MIMIC-III and 200K PubMed abstracts corpus, achieving 84.0% accuracy; 2. Alsentzer et al. at MIT reported 82.7% accuracy with Clinical-BioBERT, which was BioBERT v.1.0 (Lee 2019) pre-trained on MIMIC-III. The two reports used similar approach by training BERT on MIMIC-III and another large biomedical related corpus. However, they also covered other biomedical related language tasks without emphasis on error analysis on MedNLI. With this project, we explored fine-tuning BioBERT v1.1 for MedNLI without using MIMIC-III clinical notes corpora. To our knowledge, there hasn’t been a report on MedNLI score with BioBERT v1.1. Our main aim is to improve from the original accuracy of 73.5%. We also experimented with Clinical-BioBERT and MT-DNN for comparison.

# Implementation

Our code is available [here](https://github.com/tedapham/w266_finproj_summer19_tp_sv.git)

<https://github.com/tedapham/w266_finproj_summer19_tp_sv.git>

MedNLI data can be obtained [here](https://physionet.org/physiotools/mimic-code/mednli/)

|  |  |  |
| --- | --- | --- |
| **Platform** | **Specs** | **GPU with CUDA** |
| Macbook | 2018 Macbook Pro 15’’ | None |
| Gcloud1 | 2 vCPUs, 7.5 GB memory | 1 x NVIDIA Tesla V100 |
| Gcloud2 | 8 vCPUs, 30 GB memory | 1 x NVIDIA Tesla T4 |

**Table 2:** Computation Infrastructure

# Experiments

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment Description** | **Method** | **Implementation on** | **Framework** |
| Baseline | -Feedforward Network  - LSTM | Macbook | Command Line Interface, pytorch |
| Finetuning BioBERT | BioBERT | Gcloud2 | Jupyter notebook, Mxnet, GluonNLP |
| Fine-tuning ClinicalBERT | Clinical BERT | Gcloud1 | Command line interface, pytorch |
| Fine-tuning MT-DNN | MT-DNN | Gcloud1 | Command line interface, pytorch |

**Table 3:** List of experiments

# Results and Discussion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Model Base** | **embeddings** | **Dev score** | **Test**  **Score** |
| 1 | FeedForward | gLove | 60.5% | 59.8% |
| 2 | LSTM | glove | 74.6% | 73.3% |
| 3 | LSTM | mimic | 77.4% | 75.9% |
| 4 | LSTM | BioWordVec | 77.8% | 76.6% |
| 5 | BioBERT | BioBERTv1.1 | 82.7% | 82.4% |
| 6 | Clinica-BioBERT | (BioBERT v.1.0 + MIMIC-III) | 84.1% | 81.3% |
| 7 | MT-DNN | Glove | - | 0.693 |
| 8 | MT-DNN | BioWordVec | - | 0.694 |

**Table 4:** Results of on MedNLI using different combination of models and embeddings

## Baseline

We implemented the original MedNLI codebase (Romanov and Shivade 2018) and modified the config.py and train.py scripts to obtain models 1-4, serving as our baseline, in table 1. We trained these models iteratively through 20 epochs with the Macbook and only updated the model parameters when a higher dev score was obtained. Model 1 is a simple feedforward network with ReLU activation with GloVe embedding. Unsurprisingly, model 1 only resulted in 0.605 and 0.598 for dev and test scores. For models 2-4 , we used a bidirectional LSTM network and ReLU activation.

While models 1-3 were from the original paper, for model 4 we converted (from binary format to pickle format) and applied the BioWordVec embedding, which was trained on the MIMIC III dataset and PubMed abstracts with fastText (Zhang et al. 2019). We observed marginal gains on dev (+.4%) and on test (+0.7%) with model 4 compared to model 3 which resulted in the best baseline from using the MIMIC embedding in Romanov and Shivade 2018. By comparing the embedding vocabulary size and the number of unknown token, we found that BioWordVec has fewer unknown token compared to MIMIC and GloVe. This might explain why BioWordVec (model 4) works better than model 3 with MIMIC. In addition, MIMIC embedding is better than GloVE despite having a smaller vocabulary size might be because MIMIC contains more relevant tokens. We also hypothesize that if we retrain BioWordVec to 300 dimensions, we might obtain better performance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Embedding** | **Vocab Size** | **Unknown Token Count** | **Dimension** |
| GloVE | 2,196,019 | 1,149 | 300 |
| MIMIC | 973,187 | 1,114 | 300 |
| BioWordVec | 16,545,452 | 321 | 200 |

**Table 5:** Comparing Embeddings

## BioBERT v1.1

We used BioBERT v1.1 from GlounNLP Model Zoo. This BioBERT version reflects a language representation model whose pretrained corpus includes both general and biomedical domains, specifically the original BERT and 1M PubMed abstracts. We utilized a NVIDIA Tesla T4 for this experiment. Training data was loaded to the GPU with a data-loader set to shuffling mode for batch size of 15. To compute dev and test accuracy we copy the prediction to CPU and export the prediction on test along with the original sentence pairs for error analysis. We achieved the best accuracy of 82.4% on the MedNLI testset with BioBERT v1.1. One limitation was that we didn’t save the model parameters and tested the model first on Dev. This is a feature for future improvement.

## Clinical-BioBERT

For our analysis purposes, the clinical-BioBERT model fine tuned from BioBERTv1.0 on all MIMIC-III notes. Hyper parameter settings include attention dropout probability of 0.1, gelu activation, learning rates of 2e-5, 3e-5 and 4e-5, 12 hidden layers and attention heads and batch sizes of 16 and 32. Adam optimizer was used. Further training MedNLI resulted in an eval accuracy of 84.1% and a test accuracy of 81.3%. While multiple hyper parameters were attempted, there is still further scope to fine tune the parameters.

## MT-DNN

For comparison purposes, we also used pretrained MT-DNN models on MedNLI data. While further fine tuning was not attempted, we used Glove as well as BioWordVec. Without fine tuning the models on domain specific data, the accuracies were 0.63 and 0.69 respectively. This is an opportunity for further research and analysis as to why the accuracies were only slightly higher and also how much the accuracy can be improved by further fine tuning.

# Error Analysis

Unlike SNLI and MultiNLI, each example in the MedNLI dataset was single annotated. In Gururangan et al. (2018) discovered annotation artifacts in NLI datasets that could be present in MedNLI as well.

## 

## 

## Confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted Class** |  |  |
| **True Class** | Entailment | Contradiction | Neutral |
| Entailment | 79.96% | 5.70% | 14.35% |
| Contradiction | 7.17% | 88.61% | 4.22% |
| Neutral | 15.40% | 5.91% | 78.69% |

**Table 6:** Confusion Matrix

Highest accuracy of prediction was for contradiction and highest misclassification seem to be between neutral and entailment.

Below are some examples of differences. A high level analysis of errors indicate that the possibility of handling negations could be improved for better accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted Class** | **True Class** | **Premise** | **Hypothesis** |
| Neutral | Entailment | He denied headache or nausea or vomiting . | He has no head pain |
| Entailment | Neutral | He had no EKG changes and first set of enzymes were negative . | the patient has negative enzymes |

**Table 7:** Samples of Error

The fine tuning on clinical data are typically limited to selected hospitals that release their records for research purposes due to HIPAA and PHI. This can also be limiting and better accuracy could possibly be achieved by using more data sources.

# 

# Conclusions

Overall, our results demonstrate that even without using the original clinical notes (MIMIC-III) on which the MedNLI was derived, we still achieved accuracy comparable to the state-of-the-art methods which did include MIMIC-III in their BERT training. We hypothesized that deidentification of the MIMIC-III dataset might be the reason for this observation. Our approach of fine-tuning BioBERT v1.1 further showed that biomedical abstracts might be enough to provide biomedical context to BERT instead of both abstracts and full-text articles. However, our best accuracy on MedNLI was 82.4 hence further work needs to be done to reach 90% like BERT did on other general language tasks.

# References

Peters M, Neumann M, Lyyer ,Gardner N, Clark C,Lee K, and Zettlemoyer L. Deep contextualized word representations. Proceedings of NAACL. 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. Proceedings of NAACL. 2019.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. Proceedings of EMNLP.

Talman A and Chatzikyriakidis S. Testing the generalization power of neural network models across NLI benchmarks. CoRR, abs/1810.09774. 2018.

Romanov A and Shivade C. 2018. "Lessons from Natural Language Inference in the Clinical Domain". Proceedings of EMLP. 2018.

Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. [BioWordVec, improving biomedical word embeddings with subword information and MeSH](https://www.nature.com/articles/s41597-019-0055-0). Scientific Data. 2019.

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. (Under Review). 2019.

Lee J, Yoon W, Kim S, Kim D., Kim S, So C., Kang J. BioBERT: a pretrained biomedical language representation model for biomedical text mining. arXiv Preprint. 2019.

Alsentzer E., Murphe J, Boag W, Weng W, Jin D, Naumman T, McDermott M. Publicly Available Clinical BERT Embeddings. NAACL Clinical NLP Workshop. 2019.