DSCI 691 - NLP with Deep Learning, Spring 2024

Contributors:

* Samson Adeniyi (saa384@drexel.edu)

* Alec Peterson (ap3842@drexel.edu)

* Morgan Purcell (mrp366@drexel.edu)

Abstract

The goal of this project was to fine-tune sentence-transformer models to produce semantic-rich embeddings for clinical text. The Clinical-BERT model was used as a base-model for our work. The experimental conditions fine-tuned the Clinical-BERT base-model with different sets of NLI sentence pairs. The models were evaluated by testing their performance when used to create sentence embeddings for three tasks including: calculating the MRR of matched sentence pairs (information retrieval), classifying sentence-pairs based on relatedness, and finding the cosine similarity of embedded sentences with different relatedness scores. Overall, we found Clinical-BERT models fine-tuned on clinical NLI data outperformed those fine-tuned solely on general NLI data on our evaluation tasks.

Introduction

This repo contains files to develop sentence-transformer models for producing semantic rich embeddings for clinical texts, exploring different control and experimental conditions to examine the effects of fine-tuning transformer models on general natural language inference (NLI) and medical NLI sentence pairs.

Files:

1) Jupyter notebook files

    * `DSCI_691_Model_Finetuning.ipynb`

* `DSCI_691_Mean_Reciprocal_Rank.ipynb`

* `DSCI_691_MedNLI_Classification.ipynb`

* `DSCI_691_BIOSESS_Comparison.ipynb`

2) `data` subdirectory – contains data as described in Background -> Datasets

3) `images` subdirectory – contains select images of results plots/figures

Background

Models

Clinical-BERT (https://huggingface.co/medicalai/ClinicalBERT) is a fine-tuned version of the popular BERT encoder transformer architecture, trained on MIMIC-III (Johnson et al., 2016), a large corpus of de-identified clinical records. Clinical-BERT, specifically a version that used the knowledge-distilled DistilBERT model, was used as a base model for our work. For control comparisons to examine the effect of fine-tuning alone, the distilbert-base-uncased model was used.

Datasets

Two datasets were used for fine-tuning:

* Stanford NLI (SNLI, https://nlp.stanford.edu/projects/snli/) + Multiple-Genre NLI (MultiNLI, https://cims.nyu.edu/~sbowman/multinli/) combined, also referred to as “AllNLI” - ~1,000,000 samples.

* MedNLI (https://jgc128.github.io/mednli/) - ~15,000 samples. This dataset consisted of a development, training, and test split. The training split (~11,000 samples) was used for fine-tuning, while the test split (~1400 samples) was used for evaluation

These datasets consist of sets of sentence pairs consisting of a “premise” sentence, and a “hypothesis” sentence. Each pair has an associated label of “entailment”, “neutral”, or “contradiction” referring to how semantically similar the sentences are.

Another dataset, BIOSESS (https://huggingface.co/datasets/tabilab/biosses) was used for evaluation. This dataset consisted of 100 biomedical sentence pairs and a score for each pair ranging from 0 (no relation) to 4 (equivalent) indicating the similarity of the sentences, as assessed by five human annotators.

Methods

Hypothesis

Because labeled clinical textual datasets are resource-intensive and challenging to obtain compared to general texts, we were interested in examining the embedding quality of a transformer model with clinical knowledge and pre-training, such as Clinical-BERT, when fine-tuned on *general* NLI data compared to *clinical-specific* NLI data.

Conditions

To this end, several control and experimental conditions were devised to assess the impact of NLI fine-tuning:

* Control 0 (CTRL0): Negative control, the base ClinicalBERT model with no fine-tuning. Indicates embedding quality from clinical pre-trained knowledge alone.

* Positive Controls: BERT models (i.e. no clinical pre-trained knowledge) to examine the effect of fine-tuning along
    * Control 1 (CTRL1) - SNLI+MultiNLI
    * Control 2 (CTRL2) - MedNLI

    * Control 3 (CTRL3) - SNLI+MultiNLI followed by MedNLI

* Experimental Conditions: Clinical-BERT models to examine the impact of fine-tuning a model with clinical pre-trained knowledge

    * Experimental 1 (EXP1) - SNLI+MultiNLI

    * Experimental 2 (EXP2) - MedNLI

    * Experimental 3 (EXP3) - SNLI+MultiNLI followed by MedNLI

Training

Sentence Transformer Modules

The `sentence-transformers` library was leveraged to add the required modules to the Hugging Face transformer models previously mentioned. These modules consisted of:

* Mean Pooling Layer - a layer which averages the token embeddings produced from the base transformer model

* Multiple Negative Ranking (MNR) Loss - a loss function that, within a batch of sample pair embeddings for a given sentence a_i and other sentences b_j:

    1) calculates the similarity between all combinations of a_i and b_j

    2) calculates the cross-entropy loss between these similarities and assigned labels, where the semantically similar positive pair a_i and b_i within a given batch are assigned the "ideal" label

    3) Backpropagation proceeds to optimize cross-entropy loss and thus make the semantically similar a_i and b_i close together in the embedding space, and all other b_j sentences farther away

    To align with the intent of MNR loss and standard practice, in effect only the "entailment" and "contradiction" samples were used.

Training Hyperparameters

Training was performed in DSCI_691_Model_Finetuning.ipynb and used the following hyperparameters:

* `batch_size` = 16, mostly due to GPU memory limitations

* `warmup_steps` = 10% of training data (a good rule-of-thumb from sentence-transformers authors at sbert.net)

* `num_epochs`:

    * = 1 for the SNLI+MultiNLI dataset, primarily due to GPU resource limitations

    * = 10 for the MedNLI training dataset. Since the MedNLI train set was ~1/10 the size of the SNLI+MultiNLI dataset, 10x the epochs were used so that the model was exposed to the same number of iterations/samples

Evaluation

1) Mean Reciprocal Rank (MRR)

MRR is a common information retrieval measure. While the MedNLI dataset is not of the ideal format for this assessment, the "entailment" positive pairs from the MedNLI validation split were used. Cosine similarity is calculated between premise sentence `a_i` and all other entailment hypothesis sentences `b_j`. Assuming uniqueness between all other pairs except `a_i` and `b_i`, the cosine similarity between `a_i` and `b_i` should be of rank 1 (or at least close to it). This evaluation. is performed in `DSCI_691_MedNLI_Classification.ipynb`

2) Classification on MedNLI Test Set

One way of assessing the quality of the produced embeddings is to use them as features for a multi-class classifier. A multi-class classifier neural network was trained on the MedNLI train set by concatenating the embeddings of each premise and hypothesis sentence within a given sentence pair. This evaluation, as well as specifics

for the classifier neural network, are shown in `DSCI_691_MedNLI_Classification.ipynb`

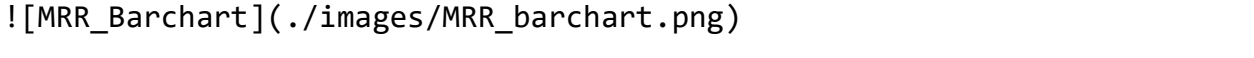
3) BIOSESS Correlation to Gold Labels

A similarity score is calculated between each of BIOSESS sentences. After normalization to 0 - 4 like the target labels, a Pearson correlation is calculated. A correlation coefficient close to 1 would indicate agreement with human assessment.

Results

1) Mean Reciprocal Rank (MRR)

Model	MRR
CTRL0	0.015862
CTRL1	0.012445
EXP1	0.017658
CTRL2	0.047981
EXP2	0.058847
CTRL3	0.018088
EXP3	0.064095



2) Classification on MedNLI Test Set

Model	Accuracy	Precision	Recall	F1-score

| CTRL0 | 0.682841 | 0.687191 | 0.682841 | 0.684240 |

| CTRL1 | 0.668073 | 0.673463 | 0.668073 | 0.669328 |

| EXP1 | 0.718003 | 0.719465 | 0.718003 | 0.718617 |

| CTRL2 | 0.661041 | 0.662777 | 0.661041 | 0.658784 |

| EXP2 | 0.720816 | 0.738972 | 0.720816 | 0.715694 |

| CTRL3 | 0.667370 | 0.674723 | 0.667370 | 0.668987 |

| EXP3 | 0.729255 | 0.741342 | 0.729255 | 0.724303 |

3) BIOSESS Correlation to Gold Labels

| Model | Pearson Correlation | P-value |

|-------|---------------------|-----------------|

| CTRL0 | 0.574784 | 4.008899e-10 |

| CTRL1 | 0.498372 | 1.316569e-07 |

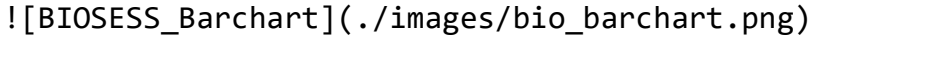
| EXP1 | 0.669632 | 2.591172e-14 |

| CTRL2 | 0.586494 | 1.439203e-10 |

| EXP2 | 0.517410 | 3.544764e-08 |

| CTRL3 | 0.513374 | 4.713201e-08 |

| EXP3 | 0.590170 | 1.034596e-10 |



Discussion

Information Retrieval (MRR)

Our study revealed that Clinical-BERT models fine-tuned on clinical NLI data (MedNLI) outperformed those fine-tuned solely on general NLI data. The Mean Reciprocal Rank (MRR) results indicated that positive pairs (entailment) generally had higher ranks, though overall MRR values were low, suggesting room for improvement in distinguishing entailment pairs. The low MRR results could be partially attributed to the lack of richness or detail in the “hypothesis” second sentence relative to the “premise” first sentence in a pair. The hypothesis was often a terse summary compared to the description in the premise.

Classification

Metrics

Classifier accuracy for the MedNLI validation split indicated that training on NLI datasets for ClinicalBERT, even the general SNLI+MultNLI corpus, improved performance relative to the base model. Accuracies of ~71% were observed for EXP1 compared to ~67% for CTRL1 or ~68%, indicating limited benefit without pre-training domain knowledge. Slightly higher accuracies of 72% and 73% were observed in EXP2 and EXP3, respectively while CTRL2 and CTRL3 had similarly lower accuracies of ~66%. The F1 scores were similar in magnitude to the accuracies for a given condition.

Confusion Matrix

The confusion matrix for each condition gives more insight into potential weaknesses in the different classifier models. While there was generally strong performance in identifying the correct class across all the conditions, CTRL0/CTRL1/EXP1/CTRL3 predicted “entailment” (label 0) more often for the “contradiction” (label 2) class relative to CTRL2/EXP2/EXP3, which were either fine-tuned on MedNLI only (CTRL2, EXP2), or in the case of EXP3 had clinical domain knowledge (differentiating it from CTRL3).

Notably, the classifiers often predicted the inverse case of “contradiction” (label 2) for the “entailment” class (label 0). These were highest for EXP2 and EXP3 indicating potential room for improvement in differentiating those in the embedding space.

While these classification metrics give an indication of embedding quality, the sentence-transformer models are not intended for developing features for classifier models. Sentence-transformer models

are more suited to applications of information retrieval and semantic search, and a labeled corpus of documents should be used.

BIOSESS

The Pearson correlation for the BIOSSES dataset for each condition has some mixed interpretations compared to the other datasets, with all results being statistically significant. EXP1 surprisingly had the highest correlation at 0.67, with the next highest being EXP3 at 0.59, CTRL2 at 0.59, and even the base model CTRL0 at 0.57. This could indicate a potential benefit to fine-tuning on general NLI.

These correlation coefficients are not the strongest however, and the BIOSESS dataset is quite small at 100 samples. The sentences could be considered more biomedical in nature as opposed to clinical and so the domain knowledge and fine-tuned knowledge may not have translated.

Nevertheless, the BIOSESS evaluation gives some indication of alignment with human similarity scoring.

Overall

Combining general NLI fine-tuning with subsequent clinical NLI fine-tuning appeared to yield the best performance overall, highlighting the potential complementary benefits of both datasets. Still, there seems to be more value in fine-tuning on domain-specific datasets like MedNLI alone where possible. The challenge remains for curating a comprehensive corpus of labeled datasets in the clinical domain for creating a performant model for clinical embeddings.

Conclusion

Fine-tuning transformer models on domain-specific datasets enhances their performance in specialized fields like clinical text processing. Clinical-BERT, when fine-tuned with clinical NLI data, produced superior embeddings and classification results compared to models trained on general NLI data alone. Future work should explore larger models with increased context lengths, additional domain-specific datasets of clinical documents, and (parameter-efficient) fine-tuning techniques to further improve model performance and utility in an application requiring information retrieval or semantic search. Additional benefit may be gained from experimenting with fine-tuning

hyperparameters like batch size – Reimers & Gurevych (2019) suggest increased performance from larger batch sizes. Experimenting with this parameter was limited due to resource constraints (GPU availability and cost).

References

1. Reimers & Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084.
    a. https://sbert.net/https://sbert.net/ -
2. Devlin et al. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805.
3. Huang et al. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. https://doi.org/10.48550/arXiv.1904.05342
4. Sanh et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter. https://arxiv.org/abs/1910.01108.
5. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).
6. Williams, Adina, Nangia, Nikita, Bowman, Samuel (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference" In: Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana (1112-1122). Retrieved from http://aclweb.org/anthology/N18-1101
7. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research

Resource for Complex Physiologic Signals," Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
8. Romanov, Alexey and Shivade, Chaitanya. "Lessons from Natural Language Inference in the Clinical Domain" Proceedings of EMLP (2018).
9. Soğancıoğlu, G., Öztürk, H., & Özgür, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics, 33(14), i49-i58. [Oxford University Press].