CSCI-GA.2565-001 Machine Learning: Project Report
Unsupervised Deep Learning for Clinical Natural Language Processing

Vasudev Awatramani (va2134) va2134@nyu.edu
Computer Science Department
New York University Courant Institute of Mathematical Sciences
New York, NY 10012, USA

Alyssa Liu (xl2500) xl2500@stern.nyu.edu
Department of Technology, Operations, and Statistics
New York University Stern School of Business
New York, NY 10012, USA

Instructor: Rajesh Ranganath
TA: Mark Goldstein, Aahlad Manas Puli, Raghav Singhal

Code: https://github.com/vasudev13/HealthLearning along with Appendix A

Abstract

The application of self-supervised language models in a clinical setting has proven to have significant performance gains. Though these models do have some understanding of language, they are often perplexed by the unusual linguistic structure and heterogeneity of clinical terminologies. In this work, we aim at utilizing Knowledge bases as a means of enhancing understanding of language models. Focusing on this direction, we propose a novel transformation technique that is unique to the clinical domain and train a language model Discharge Summary Albert that achieves 78.5% accuracy on the MedNLI(Romanov and Shivade) dataset.

Keywords: Self-supervised Learning, Clinical Natural Language Processing

1. Introduction

Unsupervised Learning may be described as *capturing rich patterns in raw data with models (like,deep networks) in a label-free way.* For this project, we focus on its specific aspect: **Self-Supervised Learning**, a form of unsupervised learning where the data provides the supervision. The idea of learning good representations of the data through a *pre-training* task (in an unsupervised fashion) has become dominant in Natural Language Processing (NLP) in the last few years, essentially due to the significant performance gains that are observed with respect to the fine-tuning on downstream task.

Pre-trained transformer models such as Bert(Devlin et al., 2019) have shown to work well on NLP benchmarks and consequently have been adapted to domains like Legal,Biomedical and even in Multilingual setting. Although BERTology models perform primarily well on web data, clinical narratives (e.g. physician notes) have known differences in linguistic

characteristics from both general text and non-clinical biomedical text, motivating the need for specialized clinical domain language models. Moreover, the sheer diversity in vocabulary and representation of identical concepts through various alternatives makes the domain more challenging.

In particular, through this project we make the following contributions:

1. Develop a Clinical-domain specific transformation that employs vast knowledge bases such as UMLS. To some extent the transformation does preserve the meaning of text.

2. Train a language model using text perturbed by the transformation and evaluate it over a downstream task: Natural Language Inference for Clinical Domain.

2. Related Work

Due to the excellent performance of Bert on NLP benchmarks such as Glue(Wang et al., 2018), several works have adapted the utility of it in the various domains. In Bert model, apart from final output layers (which are specific to task), the same architectures are used in both pre-training and fine-tuning procedures. The same pre-trained model parameters are used to initialize models for different downstream tasks. This idea of pre-training on unsupervised features is not unique to Bert and has been actively explored in Computer Vision for image representation (Doersch et al., 2015; Zhang et al., 2017; He et al., 2019; Chen et al., 2020). Bert can be viewed as a *Denoising Autoencoder*, such that during pre-training, the input embeddings (sum of the token embeddings, the segmentation embeddings and the position embeddings) have a small subset of tokens randomly masked (corrupted). The model is then trained to recovery the original input (denoising), so it can have a deeper sense of language context and flow.

However, Bert may work well on general domain, but not for specialty corpora, such as for the clinical domain, where unusual grammar structure, abbreviations are common. Therefore, attempts have been made to pre-train over clinical corpora as a means to achieve a deeper understanding of clinical and biomedical knowledge. In BLUE Benchmark (Peng et al., 2019), they facilitate research in the development of pre-training language representations in the biomedicine domain. The benchmark consists of five tasks with ten datasets that cover both biomedical and clinical texts with different dataset sizes and difficulties. They find that the BERT model pre-trained on PubMed abstracts and Mimic-III clinical notes achieves the best results.

Similar to models mention before, Clinical BERT(Alsentzer et al., 2019) models, demonstrated that continual learning yields performance improvements on three common clinical NLP tasks such that they initialize their language model with Bio Bert(Lee et al., 2019) (Bert trained over PubMed abstracts and full articles) and train over Mimic-III notes. In contrast, PubMed Bert(Gu et al., 2021) pre-trained model on biomedical domain from scratch, showing that for domains with abundant unlabeled text, such as biomedicine, pretraining language models from scratch results in substantial gains over continual pre-training of general-domain language models.

In our study, we draw comparison with Clinical Bert as it follows similar training method (including data) as ours.

3. Methods

3.1 Clinical Synonym Replacement Transformation

To improve the representation capability of the language model, we use synonym replacement transformation such that the model captures representation invariant to the variability of similar clinical concepts. For instance, in clinical notes, abbreviated mentions account for a substantial proportion of the failure cases, so substituting those for their long form definitions is worth to explore for improving the downstream tasks' performance. For example, if the input is "patient has elevated BUN", after the transformation, it becomes "patient has elevated blood urea nitrogen measurement" (Clinical Bert predicts on the transformed sample incorrectly (`entailment`) while the original input is predicted correctly (`neutral`) during the downstream task). This transformation can be implemented by ScispaCy(Neumann et al., 2019), a Python library for biomedical/scientific text processing. We utilize its *Entity Linking* framework designed to link to a subset of the Unified Medical Language System (UMLS; Bodenreider (2004)) (2.78M unique concepts).

Though, ScispaCy's entity linking approach does provide a reasonably accurate method to recognize a given clinical term and map it to its corresponding UMLS concept, the approach fails in case of ambiguity among possible candidates. The present method involves approximating nearest neighbour search over UMLS concepts (and aliases) with respect to given term. The concepts are encoded using vectors of TF-IDF scores over character tri-grams with document frequency ≥ 10. The search can be regularized by K nearest neighbours, and their proximity distance can be used as a means to filter out the closest candidate concept. Apart from UMLS being a noisy knowledge base, the major flaw with this approach is that it neglects context dependent information. This is becomes apparent when the candidate concepts are *equi-distant* from the mentioned term. Then the appropriate concept can only be identified from the context information in the input sequence of the clinical note and definitions of the candidate concepts. (Figure 1)

Another variation to the above method is to utilize specialized entity recognition models trained over datasets such as Bc5cdr(for entities relating to chemicals and diseases) and BioNLP13CG(for entities relating to cancer and genetics) . This offers more control to which category of mentions to perturb, depending on the downstream task. Apart from lack of context information, a short-coming of using such transformations is significant compute requirement. The transformations scale poorly to large number of sequences as the search for a candidate synonym involves finding nearest neighbours among 2.78M UMLS concepts and associated aliases.

3.2 Self-Supervised Pre-training

Using the perturbations from Synonym Replacement, we train a language model over the discharge summaries present in the Noteevents collection from Mimic-III. Though there are various categories (15) of clinical notes in Mimic-III, the authors of Clinical BERT observed similar performance on downstream tasks when pre-trained on the entire Mimic notes or solely on discharge summaries. This is because most of the downstream tasks largely use discharge summaries for their labelled datasets. Therefore, to make the pre-training computationally feasible, we employ 10K discharge summaries.

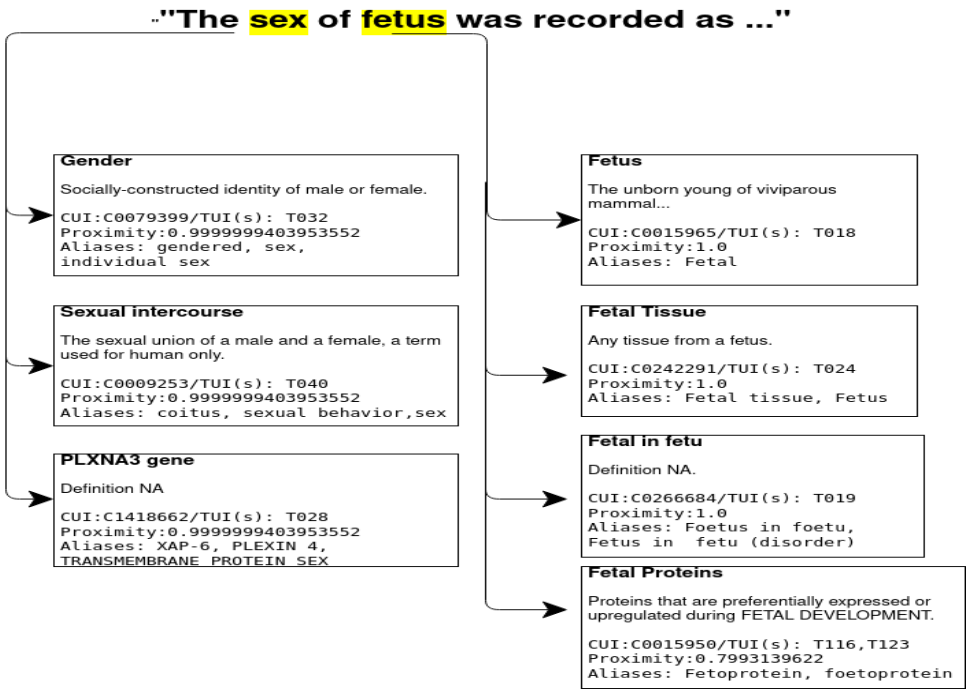

Figure 1: Ambiguous candidate concepts for clinical terms. It can be seen that there is ambiguity between the candidate for 'sex' which requires context dependent information to resolve to correct UMLS entity.

A measure to reduce compute requirements is to employ smaller models. We came across two alternatives: Albert(Lan et al., 2020) and Electra(Clark et al., 2020), that achieve performance close to Bert-Base with lesser number of parameters. Electra(Clark et al., 2020) employs a different pre-training task than masked language modelling: *replaced token detection* and may not fit the appropriate apples to apples comparison to study effectiveness of the transformation. Hence, we train a Albert-Base-V2 model that uses masked language modelling as pre-training and has 11M parameters in comparison to Bert-Base with 109M parameters. Albert employs factorization of the embedding parametrization and parameter sharing among transformer layers this achieve parameter reduction (89%).

Using the Clinical Synonym Replacement transform to perturb sequences in discharge summaries, we train a Albert-Base-V2 and refer it as Discharge Summary Albert.

4. Experiments & Discussion

To evaluate the language model, we fine-tune over a sequence classification downstream task: MedNLI(Romanov and Shivade) dataset, where the objective is determine if given hypothesis and premise are in agreement (`entailment`), disagreement (`contraction`) or no relation (`neutral`). The results are showing in Table 1. Though our model does not

perform better than the state-of-the-art model, our model does attempt to associate clinical concepts even when pre-trained on comparatively lesser amount of data.

Model	Accuracy (%)	Parameters (million)	Corpora Size (No. of Mimic Notes)
Clinical Bert(Alsentzer et al., 2019)	80.8	109	2.3 M
Discharge Summary Bert (Alsentzer et al., 2019)	80.6	109	50 K
Blue Bert(Peng et al., 2019)	**84**	109	2.3M + PubMed Abstracts
Discharge Summary Albert (ours)	78.5	11	10 K

Table 1: Comparison of existing models with Discharge Summary Albert.

We examine some of the samples that our model predicts correctly and models like ClinicalBert gets wrong (Table 2). We observe that the model tries to associate terms like "febrile" and its synonyms "fever", as indication of relation between the premise and hypothesis; other models fail to identify such relation (predict `neutral`).

Premise	Hypothesis	Model	Prediction
He denies recent fevers, chills or rigors.	The patient is not febrile./ He is afebrile.	ClinicalBert	Neutral
		Discharge Summary Albert (ours)	Entailment
		Ground Truth	Entailment
He denied headache or nausea or vomiting.	He has no head pain.	ClinicalBert	Neutral
		Discharge Summary Albert (ours)	Entailment
		Ground Truth	Entailment

Table 2: Samples where the transformation helps model prediction.

5. Conclusion & Future Work

In this work, we attempt to employ knowledge bases as means to achieve better understanding of clinical terms for a language model. However,our transformation technique has significant pitfalls. In that direction, our future work shall focus around:

1. Generating unambiguous synonym candidates that take account context-dependent information, such that the meaning of the text is always preserved.
2. Make the transformation scalable to large number of sequences.
3. Train on entire Mimic-III clinical notes corpora with larger model architectures.

Apart from language modelling, the transformations can be applied with variety of tasks such as *data augmentation* technique, *replaced token detection*, and *sentence-level pre-training*. The Clinical synonym Replacement transformation can act as basis to other transformations. For instance, the identified entity in the text can be replaced with an associated `IsA` relationship entity from Snomed-CT concept relationship knowledge base.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.

Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. URL https://openreview.net/pdf?id=r1xMH1BtvB.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1422–1430, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.167. URL https://doi.org/10.1109/ICCV.2015.167.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, 2019.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. URL http://arxiv.org/abs/1808.06752.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://www.aclweb.org/anthology/W18-5446.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 645–654, 2017.

Appendix A. Implementation Details

We pre-train our model Discharge Summary Albert over 1.3 million sentences as a masked language model. The model is trained over 436,332 steps with batch size of 8 and learning rate of 5×10^{-5}.

We implement the Clinical Synonym Replacement Transformation in PyTorch(Paszke et al., 2019) `torchvision.transforms` style. This framework allows us to employ composite transforms as used by (Chen et al., 2020). The transforms are regulated by a probability `p` which dictates whether to perturb given sentence and a `synonym replacement probability` whether to replace the given term in a perturb sentence. We use `en_core_sci_sm` (p=0.4,synonym replacement probability=0.6),`en_ner_bionlp13cg_md` (p=0.7,synonym replacement probability=0.75) and `en_ner_bc5cdr_md` (p=0.7,synonym replacement probability=0.75).

More details about implementation of our model can be found at `https://github.com/vasudev13/HealthLearning` along with model weights at `https://huggingface.co/Vasudev/discharge_albert`.