Where do Clinical Language Models Break Down? A Critical Behavioural Exploration of the ClinicalBERT Deep Transformer Model

Alexander MacLean
Vision and Image Processing Group, University of Waterloo
Alexander Wong
Vision and Image Processing Group, University of Waterloo
Email: {alex.maclean, a28wong}@uwaterloo.ca

Abstract

The introduction of **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) was a major breakthrough for transfer learning in natural language processing (NLP), enabling state-of-the-art performance across a large variety of complex language understanding tasks. In the realm of clinical language modeling, the advent of BERT led to the creation of ClinicalBERT, a state-of-the-art deep transformer model pretrained on a wealth of patient clinical notes to facilitate for downstream predictive tasks in the clinical domain. While ClinicalBERT has been widely leveraged by the research community as the foundation for building clinical domain-specific predictive models given its overall improved performance in the Medical Natural Language inference (MedNLI) challenge compared to the seminal BERT model, the fine-grained behaviour and intricacies of this popular clinical language model has not been well-studied. Without this deeper understanding, it is very challenging to understand where ClinicalBERT does well given its additional exposure to clinical knowledge, where it doesn't, and where it can be improved in a meaningful manner. Motivated to garner a deeper understanding, this study presents a critical behaviour exploration of the ClinicalBERT deep transformer model using MedNLI challenge dataset to better understanding the following intricacies: 1) decision-making similarities between ClinicalBERT and BERT (leverage a new metric we introduce called Model Alignment), 2) where ClinicalBERT holds advantages over BERT given its clinical knowledge exposure, and 3) where ClinicalBERT struggles when compared to BERT. The hope is the insights gained about the behaviour of ClinicalBERT will help guide towards new directions for designing and training clinical language models in a way that not only addresses the remaining gaps and facilitates for further improvements in clinical language understanding performance, but also highlights the limitation and boundaries of use for such models.

1 Introduction

The introduction of **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) [1] has been seen as a recent watershed moment in transfer learning in natural language processing (NLP), and has facilitated state-of-the-art performance across a wide variety of complex natural language understanding tasks in recent years. By demonstrating the ability to learn powerful general purpose language models using large-scale unlabeled text corpus, BERT enables greatly improved supervised learning of downstream language understanding tasks with much smaller task-specific text corpus by taking advantage of the wealth of contextual relationships garnered from large general purpose language corpus. In areas where domain specific terms and language are common, however, the general-purpose BERT language models can have problems shifting to the new distribution of tokens and the effectiveness of transferring learning is reduced [2]. Clinical language modelling, such as modelling clinical notes or predicting hospital readmission, is one such area, since there is a significant amount of unique terminology, notably medical coding, which would not have been seen in a general pre-training corpus [3].

To combat this challenge, Alsentzer et al. [2] proposed a new deep transformer language model, ClinicalBERT, which was trained in the same manner of BERT but on a large corpus of clinical data, namely the MIMIC-III dataset [4, 5]. [2] shows that ClinicalBERT generally outperforms the seminal BERT model [1] on down-stream tasks using medical terminology such as the Medical Natural Language inference (MedNLI) challenge. As a result, ClinicalBERT has become a popular foundation used by the reserach community for building clinical domain-specific predictive models. However, while the authors of ClinicalBERT focused analysis solely on overall quantitative performance, there was little exploration into understanding the fine-grained behaviour and intricacies of ClinicalBERT to derive deeper contextual insights. Conducting such fine-grained behavioural explorations on clinical language models is useful in many ways. First of all, clinicians are often skeptical of releasing control in clinical decision making, so being able to explain the behaviour of models is an important step in instilling trust in relevant stakeholders [6]. Additionally, such behavioural explorations can result in better understanding of the context of success and failure of models, especially in comparison to related models. This knowledge can then provide not only direction for future development and improvements in clinical language understanding performance, but also highlight the limitation and boundaries of use for such models. Motivated to garner a deeper understanding, this study presents a critical behaviour exploration of the ClinicalBERT deep transformer model using MedNLI challenge dataset to better understanding the following intricacies: 1) decision-making similarities between ClinicalBERT and BERT, 2) where ClinicalBERT holds advantages over BERT given its clinical knowledge exposure, and 3) where ClinicalBERT struggles when compared to BERT. To the best of the authors' knowledge, such an exploration has not been previously explored in research literature and can provide a much better understanding into where such clinical language models succeed and where they break down.

1.1 Related Work

Alsentzer et al. [2] were influenced by the BioBERT model, which attempted to solve a similar problem in the context of biomedical science research rather than clinical medicine [7]. BioBERT was initialized using weights from BERT-base, the BERT model with 12 attention heads, 12 attention layers, and 110 million parameters in total [7]. [2] experimented with ClinicalBERT initialized both from BioBERT and BERT, thus each model in the experiments has the exact same architecture, and found that the BioBERT-ClinicalBERT generally outperformed BERT and ClinicalBERT without the biomedical text corpus training, likely due to the at least partial similarities of biomedical text to clinical text. It is that former model which was selected to be examined in this study, and to which ClinicalBERT will refer in the rest of this analysis.

2 Methodology

2.1 Task

The Medical Natural Language Inference (MedNLI) challenge was selected [5, 8] in this study as the basis for critical behavioural exploration in order to evaluate the performance of ClinicalBERT by fine-tuning the model on a specific task. The MedNLI challenge consists of two statements, with the goal for the model to determine how the second statement relates to the first. There are three possibilities: Entailment, meaning that the second statement can be logically deduced from the first; Contradiction, the second statement cannot logically be true based on the first; and Neutral, where there is no relation between the veracity of the first and second statements. Examples of some of these statement pairs can be found in Tables 2 and 3 along with further analysis. The MedNLI dataset is split into 11,232 training samples, 1,395 validation samples, and 1,422 testing samples.

2.2 Research Questions

Motivated by the desire for a better qualitative understanding of ClinicalBERT, three research questions were explored in this critical behavioural exploration: 1) What are the decision making similarities between ClinicalBERT and BERT? 2) In which contexts is ClinicalBERT improving upon on BERT given its clinical knowledge exposure?, and 3) Where does ClinicalBERT struggle in comparison to BERT?

2.2.1 Decision Making Similarities

It has been shown, in many contexts including with clinical text, that pre-training on relevant corpora improved absolute quantitative performance on fine-tuned downstream tasks when looking at accuracy and related metrics [2]. However, these values do not differentiate between cases where improved models build on the successes of the ones to which they are compared, or if they are just simply correct on a different subset of "difficult" samples. To this end, in addition to studying confusion matrices, we introduce a new metric called **Model Agreement** to better evaluate decision making similarities. Model Agreement is calculated in the same manner as model accuracy, but using the predictions of the two models on the test set rather than comparing the predictions to the baseline labels. This is shown in Eq. 1:

$$\textit{Model Agreement} = \frac{\textit{Samples classified the same by both models}}{\textit{Total number of test samples}} \tag{1}$$

In this way, a value higher than the accuracy of an individual model and close to 1 suggests that the two models are consistent in their decision making, with the variation coming from a few misclassified samples by one model being corrected in the output of the other models. As the value decreases, more and more of the errors are different across the two models, meaning that even though one model has higher overall performance, there are a significant amount of its errors which are correct in the output of the other model. Investigating this behaviour will lead to insights into the effectiveness of the models.

2.2.2 Sample-level Model Disagreement Analysis

To obtain a much finer-grain understanding into the behavioural intricacies of ClinicalBERT, we further conduct a sample-level analysis of specific model disagreement scenarios where ClinicalBERT exhibits differing predictive behaviour when compared to the seminal BERT model. By studying the areas of model disagreement, one can gain much deeper behavioural insights into: 1) the strengths of ClinicalBERT when dealing with clinical language understanding tasks (where ClinicalBERT leads to a correct prediction while BERT does not), and 2) the limitations of ClinicalBERT (where ClinicalBERT leads to an incorrect prediction while BERT provides a correct one).

3 Results and Discussion

The BERT-Base and ClinicalBERT models were trained on the MedNLI challenge dataset for 50 epochs using the Adam optimizer within the Keras deep learning environment. Confusion matrices for the two models for the testing set are shown in Figure 1 and Figure 2 respectively. Additionally, both standard performance statistics (e.g., accuracy, precision, recall, and F1-score) along with the proposed Model Agreement scores are shown in Table 1.

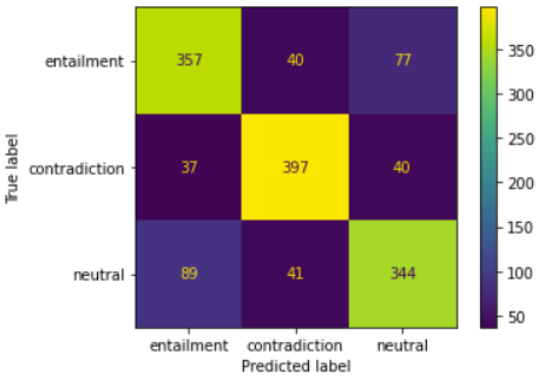

Fig. 1: BERT Confusion Matrix.

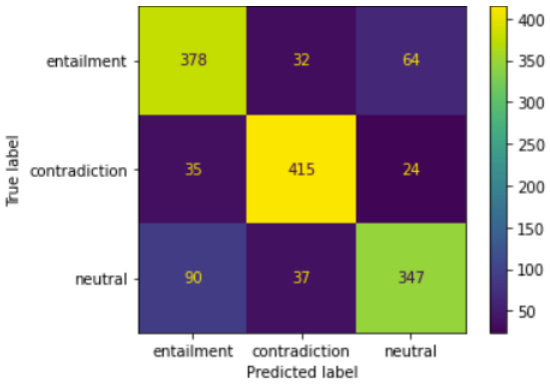

Fig. 2: ClinicalBERT Confusion Matrix.

Model	**BERT**	**ClinicalBERT**
Class	Entailment	
Precision	0.7391	**0.7515**
Recall	0.7532	**0.7975**
F1-score	0.7461	**0.7738**
Class	Contradiction	
Precision	0.8305	**0.8574**
Recall	0.8376	**0.8755**
F1-score	0.8340	**0.8663**
Class	Neutral	
Precision	0.7462	**0.7977**
Recall	0.7258	**0.7321**
F1-score	0.7358	**0.7635**
Overall Accuracy	0.7721	**0.8017**
Model Agreement	0.8143	

Table 1: Results on Test Dataset.

3.1 Decision Making Similarities

As seen in both the figures and Table 1, much of ClinicalBERT's improvements over BERT come in the form of fewer Entailment and Contradiction samples being classified as Neutral. The largest decreases in off-diagonal elements in ClinicalBERT's confusion matrix are seen where those classes were predicted as Neutral. Similarly, the individual class statistic which increased the most was precision for the Neutral case. Precision is defined as $\frac{TP}{TP+FP}$, so ClinicalBERT led to fewer false Neutral predictions more than it improved on any other case.

Additionally, Table 1 contains the Model Agreement score. This value of 0.8143 is slightly higher than the accuracy of either model, but still far from 1, suggesting that a significant amount of cases where ClinicalBERT is incorrect are cases where BERT was successful, and vice versa. If that were not the case, and every sample that BERT predicted accurately were successfully classified as well by ClinicalBERT, then Model Agreement would instead be $1-(acc_{ClinicalBERT}-acc_{BERT}) = 1-(0.8017-0.7721) = 0.9704$, with the only differences coming from cases where ClinicalBERT improved over BERT. At the other extreme, where every error in ClinicalBERT was predicted successfully by BERT, Model Agreement would then be $1-((1-acc_{ClinicalBERT})+(1-acc_{BERT})) = 1-((1-0.8017)+(1-0.7721)) = 0.5738$. In this case, ClinicalBERT fixed every error made by BERT, but for almost every one made a new error; the times where a new error was not made account for the increased overall accuracy. 0.8143 falls partway between these two extremes, showing that some errors were consistent across the two models, while others differed. This result suggests that the third research question, in which cases did ClinicalBERT struggle in comparison to BERT, is in fact relevant and should be studied closely.

The following tables show selected test samples to illustrate model behaviour. Table 2 displays samples whose label is Entailment, but which were misclassified by one or both of ClinicalBERT and BERT. Table 3 does the same, but for test samples whose label is Contradiction.

Sample	Sentence1	Sentence2	BERT	ClinicalBERT
E1	The patient stopped intravenous fluids and got 10 mg of intravenous Lasix times three and put out 500 cc of urine output.	The patient received too much fluid.	CONTRADICTION	NEUTRAL
E2	DM x 20yrs 4.	Patient has elevated blood glucose	NEUTRAL	NEUTRAL
E3	Labs notable for WBC 10.5 without bands, Hct 32.2 (prior baseline mid to upper 20s), Cr 0.9, CE neg X 1, and lactate 1.5.	The patient does not have an infection.	CONTRADICTION	**ENTAILMENT**
E4	She was found to have new onset a.fib w/ rate in the 120's to 130's and lateral ST depressions c/w demand ischemia.	The patient has coronary artery disease.	NEUTRAL	**ENTAILMENT**
E5	En route to the Emergency Department, she developed worsening substernal chest pain without any radiation.	patient has an acute MI	**ENTAILMENT**	NEUTRAL
E6	She was evaluated by neurosurgery, deemed to be intact neurologically.	No history of cerebrovascular accidents	**ENTAILMENT**	CONTRADICTION

Table 2: Examples of failed test samples where the label is Entailment, with the predicted labels from each of the models also provided.

Sample	Sentence1	Sentence2	BERT	ClinicalBERT
C1	Diastolic CHF, LVEF >70% 2/06 10.	Patient has angina	ENTAILMENT	ENTAILMENT
C2	(Lactate only 1.3 and pt afebrile).	Elevated temperature	NEUTRAL	ENTAILMENT
C3	Pt saw PCP next day and atenolol was stopped but no further w/u done (ie scans/xray) for fall on [** Location **]us day.	The patient is kept on a beta blocker.	NEUTRAL	**CONTRADICTION**
C4	Because neonatology was not present at delivery, resuscitation was initiated by the labor and delivery nurses.	The patient did not need any additional help after birth.	ENTAILMENT	**CONTRADICTION**
C5	She was discharged on bed rest and treated with terbutaline.	Patient is no longer taking medications	**CONTRADICTION**	NEUTRAL
C6	He returned to [**Hospital 8682**] clinic three weeks later and was prescribed antibiotics	the patient is not infected	**CONTRADICTION**	ENTAILMENT

Table 3: Examples of failed test samples where the label is Contradiction, with the predicted labels from each of the models also provided.

3.2 Sample-level Model Disagreement Analysis

While it is not possible to capture all possible patterns with only a few samples, some patterns do seem to be emerging. The first case to be considered is when both models have incorrect predictions. In each case shown, samples E1, E2, C1, and C2, there is not only unique clinical terminology, but also clinical diagnostic knowledge that is relevant to the logical connection between the two statements. Thus, even if the model is able to interpret correctly what "Lasix" and "cc" (E1), "DM" (E2), "CHF" and "LVEF" (C1), and "pt afebrile" (C2) all mean, the meaning of the following sentences are additionally dependent on the quantities contained in the initial ones. There could have been training samples with the same terminology, but with different quantities, and the model would need to be able to interpret those relationships. In a sense, the model would need to see enough samples with differing quantities to build an internal "classifier" to know that "LVEF >70%" does not indicate that the patient has angina (C1) or that "500 cc of urine output" is related to too much input of fluid (E1).

Regarding cases where ClinicalBERT is correct while BERT is not, which are samples E3, E4, C3, and C4, they tended to be long, complex statements with significant amounts of clinical terminology, with the extra pre-training for ClinicalBERT perhaps allowing it to better follow the connections between the various terms within a sentence as well as between sentences. Sample E3 contains plenty of terms (WBC, Hct, Cr, CE) which are unique to clinical text, and MedNLI training alone may not have allowed BERT to recognize the meaning of those terms in the context of infection. In sample C3, BERT may not have connected the term "atenolol" to beta blockers, and defaulted to the sentences being neutral, while ClinicalBERT was able to identify that relationship properly. Similar discussion can be had for the terms "ischemia" and "coronary artery disease" (E4). In sample C4, BERT's error in labelling the sample as Entailment shows that it understood the birth-related terms ("neonatology", "delivery", "labor") perhaps could not determine that the focus of the logic was in fact on "resuscitation" and "additional help" which ClinicalBERT successfully identified.

In examining these cases, a fitting analogy can be found. When BERT undergoes pre-training, it is essentially a student moving through education and life, learning grammatical structures and meanings of words, especially how they vary based on context. Once graduated high school, an individual can make sense of most sources of text in their native language, at least when the distribution of terms is familiar to them. However, when they are thrust into a situation where terms are new to them, or used in new ways, they have a much harder time understanding what they are reading or hearing. If a recent high school graduate were shown the MedNLI dataset, they would likely have a hard time succeeding - not because they cannot read English, but instead because they often cannot make the proper connections between the unique terms that are important to the meanings of the various statements, even if they had taken a high school biology course or watched Grey's Anatomy. On the other hand, ClinicalBERT has essentially been "sent" to medical school - by training it on the MIMIC-III data, it was introduced to domain specific terminology that help it to perform its future function. It has not lost the knowledge gained during "high school", which is the general understanding of the English language, and has learned how clinical terminology fits into its understanding of language. Since [2] used BioBERT as the initialization for ClinicalBERT, we can go one step further and say that BioBERT was analogous to a high school student who went to university and completed an undergraduate degree in biomedical science. It learned terminology required for understanding relevant texts, in addition to its previous general education. By initializing

with BioBERT, [2] sent a biomedical science graduate to medical school, and this contextualizes why Bio-ClinicalBERT provided better results for those authors; there is a reason why so many medical students come from related educational backgrounds, as the relevant background can make it easier to acquire medicine specific knowledge.
In the final cases, E5, E6, C5, and C6, the statements tended to be simpler and had fewer terms specific to clinical terminology. This may show why BERT was able to be successful, but it is less clear why ClinicalBERT had difficulties. Of note, most of these failures where the true label was Entailment were misclassified as Neutral, while most of the Contradictions were wrongly labelled Entailment. In the former cases, ClinicalBERT made errors by believing that the statements were not related, possibly because the model did not find a strong enough connection between them. In any case, it rarely falsely believed the statements to be contradictory, which would have been a greater error. In the latter case, ClinicalBERT instead understood that the statements were related, but failed in understanding the manner of the connection by saying that the second was true based on the first. Many of these examples contained negation in one of the statements ("not", "no symptoms", "no recent history", etc.) suggesting that failure was due to the model incorrectly understanding the purpose of that specific negation. Unfortunately, these insights do not provide an explanation as to why BERT was successful in these cases despite the same challenges being present.

3.3 Recommendations

The results of this study emphasize the importance of understanding the context of what a model is being asked to do. As shown in examples E1, E2, C1, and C2, ClinicalBERT's improvements come from an improved understanding of clinical language; in essence, its time at medical school focused on the meanings of clinical terminology and while some decision making processes could be embedded in that understanding, there likely are not enough training examples in MIMIC-III or the MedNLI datasets to trust its output in those situations. Recognizing these limitations, what may be more effective is integrating such a language model into larger systems, whereby ClinicalBERT can interpret incoming information and forward it to other subsystems which are designed for such a task, either via clinician-defined rulesets or possibly other machine learning models.
Another area of future study would be to investigate the MedNLI error cases with clinicians to understand if there are any further patterns emerging. Recent work has gone into examining the behaviour of BERT-type models on inference tasks, namely the Stanford NLI task, and noticed that some of the samples in the data set are somewhat ambiguous [9, 10]. In their experiments, the authors surveyed 100 individuals to acquire a distribution of human opinions for a subset of test samples, and noticed that samples which have higher disagreement across human participants generally are predicted less accurately by their BERT model [10]. In the case of MedNLI, it would be much harder to find 100 individuals with the required medical knowledge to reliably provide annotations for a subset of test samples, but performing similarly designed experiments would provide an interesting comparison to this recent work. Discussing said results with clinicians may lead to more actionable insights which can either validate ClinicalBERT's performance despite errors according to the structure of MedNLI, or guide development of architecture and data curation to further improve performance.

4 Conclusion

In conclusion, this study examined the qualitative performance of a publicly available ClinicalBERT model trained on the MIMIC-III dataset applied to the MedNLI challenge. During investigation, its advantages over a model trained from BERT-Base were identified to occur when samples contain higher amounts of clincal language-specific terminology, as expected. Additionally, by analyzing its failures the limitations of the model were explored, which is especially critical for any technology used in clinical contexts. Applying BERT to new language domains has consistently shown to improve upon state-of-the-art results, but blind application runs the risk of overlooking unacceptable errors or routes to improvement. Hopefully, studies such as this one will lead to improvements in design and better trust in application to the clinical domain, both of which will have positive impacts on health outcomes.

Acknowledgments

The authors would like to acknowledge Dr. Helen Chen from the Waterloo Health Information Systems and Technology Lab at the University of Waterloo for providing initial motivation for the exploration of this problem area.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[2] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," 2019.

[3] R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *Journal of Biomedical Informatics*, vol. 57, pp. 28–37, Oct. 2015. [Online]. Available: https://doi.org/10.1016/j.jbi.2015.07.010

[4] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[5] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[6] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, p. m689, Mar. 2020. [Online]. Available: https://doi.org/10.1136/bmj.m689

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," 2019.

[8] C. Shivade, "Mednli - a natural language inference dataset for the clinical domain," Oct 2019. [Online]. Available: https://physionet.org/content/mednli/1.0.0/

[9] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. [Online]. Available: https://www.aclweb.org/anthology/D15-1075

[10] Y. Nie, X. Zhou, and M. Bansal, "What can we learn from collective human opinions on natural language inference data?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9131–9143. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.734