Parameter Efficient Transfer Learning for Suicide Attempt and Ideation Detection

Bhanu Pratap Singh Rawat[1,*], **Hong Yu**[1,2,3,‡]
[1]CICS, UMass-Amherst, [2]U.S. Department of Veterans Affairs,
[3]Center of Biomedical and Health Research in Data Sciences
*brawat@umass.edu, ‡hong_yu@uml.edu

Abstract

Pre-trained language models (LMs) have been deployed as the state-of-the-art natural language processing (NLP) approaches for multiple clinical applications. Model generalisability is important in clinical domain due to the low available resources. In this study, we evaluated transfer learning techniques for an important clinical application: detecting suicide attempt (SA) and suicide ideation (SI) in electronic health records (EHRs). Using the annotation guideline provided by the authors of ScAN (Rawat et al., 2022), we annotated two EHR datasets from different hospitals. We then fine-tuned ScANER (Rawat et al., 2022), a publicly available SA and SI detection model, to evaluate *five* different parameter efficient transfer learning techniques, such as adapter-based learning and soft-prompt tuning, on the two datasets. Without any fine-tuning, ScANER achieve macro F1-scores of 0.85 and 0.87 for SA and SI evidence detection across the two datasets. We observed that by fine-tuning less than $\sim 2\%$ of ScANER's parameters, we were able to further improve the macro F1-score for SA-SI evidence detection by 3% and 5% for the two EHR datasets. Our results show that parameter-efficient transfer learning methods can help improve the performance of publicly available clinical models on new hospital datasets with few annotations.

1 Introduction

In the past decade, 90% of the US hospitals have adopted a certified electronic health record (EHR) system (IT, 2022). This has led to an enormous availability of EHRs with rich information about patients' health (Henry et al., 2016). With the advancement of natural language processing (NLP), there has been a significant improvement in the development of clinical models and systems to extract clinically relevant information from the EHRs for further downstream tasks (Uzuner et al., 2011; Rawat et al., 2022). Recent years have seen clinical datasets being publicly released for different NLP tasks such as named entity recognition, relation extraction, text de-identification and disease classification (Pampari et al., 2018; Sun et al., 2013; Henry et al., 2020). Medical Information Mart for Intensive Care - III (MIMIC) (Johnson et al., 2016) has enabled a large and continually growing set of de-identified EHR notes from an intensive care unit for developing other publicly available datasets such as emrQA (Pampari et al., 2018), ScAN (Rawat et al., 2022) and adverse drug reaction (ADR) extraction (Henry et al., 2020).

This increase in availability of the clinically annotated datasets has led to the improvement in performance of different NLP models. While this improvement is great, a key question is whether these improvements generalize to new datasets of the same task or not. This question is quite difficult to answer because it requires annotating multiple datasets or new datasets with the same guidelines when it is already difficult to annotate a single dataset (Laparra et al., 2021; Futoma et al., 2020). In this study, we evaluate different parameter efficient transfer learning techniques on the task of an important clinical application, namely suicide attempt (SA) and suicide ideation (SI) detection from EHRs.

Recently, a SA-SI detection dataset (ScAN) (Rawat et al., 2022) was publicly released in an effort to extract suicidal information from patients' EHRs. ScAN was released along with the annotation guidelines used by the experts and the baseline model to detect the suicidal evidences from EHR notes (ScANER). We followed the annotation guideline to annotate two new datasets: EHR notes from School of Medicine at University of Pittsburgh (hereby referred as ScAN_UP) and EHR notes from the US Veterans Health Administration (ScAN_VA). We used ScAN and ScANER as our base dataset and model for creating the two new datasets and evaluating different transfer learning techniques. In order to evaluate the transfer learning performance of ScANER, we kept the size of ScAN_UP and ScAN_VA relatively smaller than ScAN for further fine-tuning. These fine-tuned models could eventually help clinical professionals in making patient-aware clinical judgements for further treatments.

Pre-trained language models have significantly grown in size since the inception of BERT (Devlin et al., 2018) model. BERT was introduced with 110 million parameters but recent LMs such as generative pre-trained transformer (GPT-3) (Brown et al., 2020) and Open Pretrained Transformer (OPT) (Zhang et al., 2022) have $\sim$ 175 billion parameters. Given their unprecedented performance gains over different downstream tasks, the researchers in the clinical community have also adopted these models. But all hospitals or medical organizations do not have the resources to adapt these billion parameter models in their ecosystem. Hence it is important to evaluate parameter-efficient transfer learning techniques that keep most of the model parameters frozen during fine-tuning on a newer dataset for the same task. We decided to try five different techniques: fine-tuning the classification layer, BitFit (Zaken et al., 2021), adding adapter modules (Houlsby et al., 2019), soft-prompt fine-tuning (Lester et al., 2021) and tuning the last four layers (Lee et al., 2019). Most of these techniques require fine-tuning of less than 2% of ScANER's parameters except tuning the last four layers which requires tuning of $\sim 23\%$ parameters.

In this study, we found that ScANER achieves $> 85\%$ macro F1-score for SA-SI evidence detection on two new datasets without any fine-tuning. We were able to further improve the SA-SI evidence detection by 3% for ScAN_UP and 5% for ScAN_VA by fine-tuning less than $\sim 2\%$ of ScANER's parameters. Both ScAN_UP and ScAN_VA contain less than 8% annotations when compared to the original ScAN dataset. This shows that parameter-efficient transfer learning methods can help in improving the performance of publicly available clinical models on new hospital datasets with few annotations.

2 Dataset

In order to evaluate different transfer learning techniques, we focused heavily on choosing a task that has a publicly available dataset along with the annotation guidelines and the baseline model. The annotation guidelines are very important because they would help us in keeping the annotation decisions across different datasets uniform. Hence, we chose the task of detecting suicide attempt and ideations events in EHRs because of the availability of ScAN dataset (Rawat et al., 2022). The annotations guidelines for creating ScAN are publicly available along with their proposed baseline model (ScANER).

2.1 ScAN: Suicide Attempt and Ideation Events Dataset

ScAN (Rawat et al., 2022) is a publicly available SA and SI events dataset which is a subset of the MIMIC-III (Johnson et al., 2016) dataset. The EHRs were filtered for the hospital stays that consisted of diagnostic codes associated with suicide and overdose. These EHRs were annotated at sentence-level for SA and SI events. Each hospital-stay consisting of multiple EHR notes, such as nursing note, physician note, and discharge summary, was also annotated for SA and SI. ScAN consists of $12,759$ EHR notes with $19,960$ unique evidence annotations for suicidal behavior. The publicly available annotation guidelines of ScAN allows the creation of new datasets for the same task with uniform annotations.

We decided to annotate two parallel datasets using the EHR notes of patients at School of Medicine, University of Pittsburgh and EHR notes of Veterans at Veteran Health Administration. For both datasets, we filtered the notes using the phrases related to suicidal behavior extracted from the ScAN dataset, such as *overdose*, *suicide attempt*, and *killing myself*. We were not able to map different EHRs from the same hospital-stay. Hence, we decided to focus only on extracting SA-SI evidence paragraphs from the EHRs using the *evidence retriever* module of ScANER. The *evidence retriever* module consists of a pre-trained LM (medRoBERTa) in a multi-task setting to extract all the evidence paragraphs from the EHR notes of the patients.

2.2 School of Medicine, University of Pittsburgh

There were $99,736$ EHR notes available from the School of Medicine, University of Pittsburgh. After filtering notes with the help of the selected keywords for suicidal behavior we were able to find 220 unique EHR notes with a mention of SA or SI. The dataset was annotated by two expert annotators

	ScAN_UP (220 EHRs)			**ScAN_VA** (880 EHRs)		
Evidence	*Yes*	*No*		*Yes*	*No*	
Train	302	517		1171	2171	
Validation	72	108		233	467	
Test	258	491		968	1927	
SA	*Positive*	*Neg_Unsure*	*Neutral-SA*	*Positive*	*Neg_Unsure*	*Neutral-SA*
Train	199	35	585	419	35	2888
Validation	47	11	125	77	8	615
Test	149	42	558	340	44	2511
SI	*Positive*	*Negative*	*Neutral-SI*	*Positive*	*Negative*	*Neutral-SI*
Train	80	34	702	566	440	2316
Validation	13	15	151	98	91	506
Test	60	42	638	440	364	2066

Table 1: The distribution of evidences paragraphs in ScAN_UP and ScAN_VA for train, validation and test sets. A paragraph is considered an *evidence*, labeled as *Yes*, if it has at least one sentence annotated as SA or SI. A *No* evidence paragraph is *Neutral-SA* and *Neutral-SI*.

under the supervision of a senior physician. Following the annotation guidelines provided via ScAN (Rawat et al., 2022), we created four categories for SA: *positive*, *negative*, *unsure* and *neutral-SA*. A paragraph is marked *positive* for SA if it mentions a positive suicide attempt, such as 'tried to hang myself'. A *negative* SA annotation denotes an accidental self-inflicted harm which could be misinterpreted as a suicide attempt such as a clinically diagnosed 'accidental overdose'. An annotation is marked as *unsure* for SA if it is not clear from the text whether the suicide attempt is positive or negative. Any paragraph with none of the SA annotation would be considered as *neutral-SA*. For SI, we have three categories: *positive*, *negative* and *neutral-SI*. As per ScAN (Rawat et al., 2022), we also merged our two labels *negative* and *unsure* for suicide attempt to create one label: *neg_unsure*. Similar to the original dataset, ScAN_UP is also highly imbalanced consisting of only few instances of *neg_unsure* SA labeled paragraphs.

This resulted in 853 unique annotations at sentence level where 613 were for SA and 240 for SI. Similar to ScAN (Rawat et al., 2022), we also created paragraphs from the EHR notes using an overlapping window of 5 sentences. We divided the EHRs into train, validation and test set in the ratio of 50 : 10 : 40. This resulted in total 632 *evidence* paragraphs, where an evidence paragraph is any paragraph which contains at least one annotation related to SA or SI. The annotators achieved an agreement of 97.76% at paragraph-level and 100% on document-level. The distribution of the paragraphs for SA and SI is provided in Table 1.

2.3 Veterans Healthcare Administration (VHA)

In the VHA system, we found hundreds of thousands EHR notes with keywords related to suicidal behavior. We sampled 883 notes from all the available notes to keep the size of VHA dataset roughly 4 times bigger than ScAN_UP. The dataset was again annotated by two annotators under the guidance of a senior physician. The annotators achieved an agreement of 93.97% at paragraph-level and 100% agreement on document-level. There were total of 1371 unique annotations for suicide attempt and 2270 for suicide ideation. As a preventive measure by VHA, Veterans with any form of suicidal behavior are regularly screened for suicidal ideation resulting in an inflated number of negative SI annotations in ScAN_VA dataset. Similar to ScAN_UP, we created paragraphs from the EHR notes using an overlapping window of 5 sentences. We divided the EHRs into train, validation and test set in the ratio of 50 : 10 : 40. This resulted in a total of 2372 *evidence* paragraphs. The distribution of the paragraphs for SA and SI across train, validation and test set is provided in Table 1.

These two datasets are quite different from each other as the EHR notes used for ScAN_UP are written for civilians whereas the notes for ScAN_VA are written for Veterans and contain medical linguistics specific to veteran healthcare administration. As mentioned earlier, the *negative* SI annotations are frequently observed in ScAN_VA as compared to ScAN_UP. Thus the label distribution is also quite different amongst the two datasets. These two datasets would provide a good challenge to ScANER and it's further fine-tuned versions using different transfer learning techniques.

3 Methodology

ScANER (Rawat et al., 2022) consists of two submodules: (a) an *evidence retriever module* that extracts the evidence paragraphs related to SA and SI events and (b) a *predictor module* that predicts SA or SI label for a patient's hospital stay using all the EHR notes from the hospital admission. For our two datasets, ScAN_UP and ScAN_VA, we have sentence-level annotations in an EHR but do not have all the EHRs for patients' single admission. Hence, we only focus on the first module of ScANER which can be used to extract all the evidence paragraphs from an EHR. The *evidence retriever module* consists of a medRoBERTa model trained in a multi-task learning setting to identify the evidence paragraphs along with classifying the SA and SI event label for the paragraphs. We used the ScANER model trained on the original ScAN (Rawat et al., 2022) dataset for our experiments. We used *five* different transfer learning techniques with varying number of trainable parameters on ScAN_UP and ScAN_VA.

3.1 Fine-tuning the classifier layers

ScANER consists of three classification layers for predicting the evidence class label, SA label and SI label. We decided to only fine-tune these three final classification layers on our datasets while freezing the rest of the encoder parameters. This is the most parameter efficient transfer learning technique as it uses only $\sim$ 8 thousand parameters, out of the available 125 million, refer Table 2. This technique takes the least amount of resources for fine-tuning but provides very low capacity for the model to learn new information or patterns.

3.2 Soft prompt tuning

Soft prompt tuning (Lester et al., 2021) is a powerful technique for adapting pre-trained models for new downstream tasks. For prompt tuning, all the encoder parameters are frozen during fine-tuning except a few additional k tunable tokens for each downstream task. These tunable soft-prompts help the model in adapting to new tasks using the previously trained encoder parameters. The length of the soft-prompts (k) can be tuned as a hyper-parameter. These soft-prompts can be initialized randomly or using an existing embedding from the encoder's vocabulary (Lester et al., 2021) related to the downstream task at hand. We experimented with different length of soft prompts ranging from 10 to 40 and initializing the soft prompts with the embedding of the word 'the' and 'suicide'. This transfer learning technique uses only 0.02% of ScANER's parameters.

3.3 BitFit

BitFit (Zaken et al., 2021) is a sparse fine-tuning technique that modifies only the bias terms of the trained model. Zaken et al. (2021) showed that on small to medium sized training datasets, BitFit is competitive with fine-tuning the entire training model. BitFit is also a light fine-tuning method that only uses 0.2% of ScANER's parameters.

3.4 Adapters

Adapter modules (Houlsby et al., 2019) were proposed as another efficient transfer learning technique which requires adding a few trainable parameters for the downstream task while freezing all the original encoder parameters. Adapters require more additional tunable parameters as compared to soft prompt tuning because adapter modules are added in multiple transformer layers of the encoder. Though in comparison to training all the model parameters, it only adds $\sim 2\%$ parameters to the ScANER model.

3.5 Fine-tuning few last layers

Lee et al. (2019) studied the effect of freezing multiple early encoder layers and found that only a fourth of the final layers need to be fine-tuned to achieve 90% of the performance achieved via full model training. We experimented with fine-tuning last *two* to *five* layers for our new datasets. As compared to the earlier transfer learning methods, this technique requires the most number of parameters even with fine-tuning of only last 2 layers ($\sim 11\%$).

Evaluation Metrics As our main task is to classify a paragraph as an evidence or not, we looked at the accuracy, macro F1-score and weighted F1-score on the test sets of ScAN_UP and ScAN_VA. Since, the models are being fine-tuned in the multi-task setting we would also look at the auxiliary tasks of predicting the SA and SI labels for the paragraphs. Accuracy and weighted F1-score provides

ScAN_UP		**Evidence**			**SA**			**SI**		
Transfer Learning	# Tunable Params ↑	Acc	F1	Wt-F1	Acc	F1	Wt-F1	Acc	F1	Wt-F1
ScANER	-	0.88	0.87	0.88	0.81	0.57	0.82	0.89	0.58	0.88
Classifier	8 Thousand	0.88	0.87	0.88	0.85	0.54	0.82	0.89	0.56	0.88
Soft Prompt-tuning	23 Thousand	0.91	0.90	0.91	0.86	0.56	0.84	0.88	0.49	0.86
BitFit	130 Thousand	0.88	0.87	0.88	0.85	0.54	0.83	0.89	0.54	0.88
Adapter	2 Million	0.91	0.90	0.91	0.87	0.56	0.84	0.89	0.50	0.87
Last 4 layers	28 Million	0.89	0.88	0.89	0.85	0.54	0.83	0.89	0.52	0.87
All layers	125 Million	0.91	0.90	0.91	0.87	0.56	0.84	0.89	0.52	0.87

ScAN_VA		**Evidence**			**SA**			**SI**		
Transfer Learning	# Tunable Params ↑	Acc	F1	Wt-F1	Acc	F1	Wt-F1	Acc	F1	Wt-F1
ScANER	-	0.86	0.85	0.86	0.81	0.49	0.84	0.79	0.63	0.79
Classifier	8 Thousand	0.90	0.88	0.89	0.90	0.50	0.89	0.81	0.63	0.81
Prompt-tuning	23 Thousand	0.90	0.89	0.90	0.91	0.52	0.89	0.81	0.63	0.80
BitFit	130 Thousand	0.91	0.89	0.90	0.91	0.52	0.90	0.82	0.64	0.81
Adapter	2 Million	0.91	0.90	0.91	0.92	0.53	0.91	0.82	0.65	0.82
Last 4 layers	28 Million	0.90	0.89	0.90	0.90	0.51	0.89	0.82	0.64	0.81
All layers	125 Million	0.92	0.91	0.92	0.93	0.58	0.92	0.84	0.70	0.84

Acc: Accuracy, F1: Macro F1-score, Wt-F1: Weighted F1-score

Table 2: Evidence, SA and SI classification performance of all the transfer learning techniques on ScAN_UP and ScAN_VA datasets. The transfer learning techniques are *Classifier* layers tuning, Soft *prompt-tuning* (Lester et al., 2021), *BitFit* (Zaken et al., 2021), *Adapter* modules fine-tuning (Houlsby et al., 2019) and fine-tuning *last 4 layers* (Lee et al., 2019). *ScANER* (Rawat et al., 2022) refers to the original model without any fine-tuning on ScAN_UP and ScAN_UP and *all layers* refers to the fine-tuning of all the parameters of ScANER model.

overall model performance whereas the macro F1-scores provides class level model performance and is quite important in our cases as our dataset is highly imbalanced (refer Table1). All the final hyper-parameter settings for the transfer learning techniques are provided in Appendix A.

4 Results and Discussion

For evidence retrieval, even without fine-tuning the original ScANER model is able to achieve a macro-F1 score of 0.87 and 0.85 for ScAN_UP and ScAN_VA datasets respectively, refer Table 2. When all the parameters of ScANER are fine-tuned, the macro F1-score of the *evidence retriever* module improved by 3% and 6% for evidence retrieval for ScAN_UP and ScAN_VA. For ScAN_VA, the macro F1-score for SA and SI also improved by 9% and 7% respectively. But for ScAN_UP, the performance for both SA and SI classification dropped when all the layers of the encoder are fine-tuned. This is mainly because of the extreme imbalance for both SA and SI in the ScAN_UP dataset. The accuracy and weighted F1-score performance for SA classification improved by 6% and 2% respectively because the fine-tuned ScANER model performed well for the *positive* and *neutral-SA* class but performed poorly for the under-represented *neg_unsure* class. We tried multiple techniques to counter the imbalance, such as up-sampling and weighted log-loss learning as described in Rawat et al. (2022), but none of the techniques helped in improving the performance of fully-trained model on ScAN_UP. One thing to notice is that the performance for the main task of *evidence retrieval* improved for both datasets with transfer learning. We also observed that the performance improvement is not strictly correlated with the number of tunable parameters available for transfer learning.

ScAN_UP The adapter module and soft-prompt fine-tuning performed the best for the evidence retrieval task. Both techniques were able to achieve the same performance as the fully-trained *evidence retrieval* module while using less than 2% of the module parameters. They also achieved similar SA prediction performance in terms of F1-score but under-performed for the SI prediction task. Amongst the two, adapter modules based fine-tuning performed better for SI prediction by 1%. These results are encouraging as they suggest that with the help of only 132 annotated EHRs and fine-tuning of less than 2% of the parameters, we can significantly improve the performance of the *evidence retrieval* module. BitFit performed almost similar to only classifier fine-tuning even when it has 16 times more tunable parameters. For last few layers technique, we found that tuning last 4 layers yield the best results. It was also able to improve over the baseline ScANER performance but under-performed as compared to adapter and soft-prompt tuning.

For adapter modules, we found that 64 dimensional adapters work the best for our dataset. For soft-prompt fine-tuning, we tried initializing the soft prompt randomly, using the existing vocabulary embedding of the token 'the', and the vocabulary embedding of the token 'suicide'. For our dataset, the model with soft-prompt initialized using the embedding of the token 'suicide' performed the best. It outperformed the model with the soft-prompts using the emebeddings of the token 'the' by $\sim 1\%$. The results also show that even without any fine-tuning ScANER can retrieve evidences with a strong performance of macro F1 of 0.87.

ScAN_VA This dataset is almost twice the size of ScAN_UP which allows the ScANER model to improve even more. This is evident as the fully-trained *evidence retrieval* module outperformed the original ScANER module by 6%, 9% and 7% for evidence, SA and SI classification respectively. The adapter based model is able to achieve the best macro F1-score of 0.90 amongst all the transfer learning fine-tuning techniques. It outperformed all the other models for SA and SI classification as well while improving the performance of the original ScANER model by 5% for evidence retrieval, 4% for SA classification and 2% for SI classification. Even the classifier only fine-tuning technique is able to improve the performance of *ScANER* by 3% for evidence detection. The rest of the fine-tuning techniques improved the macro F1-score for evidence retrieval by atleast 4%.

Recommendations We observed that for both datasets, adapter based fine-tuning performed the best for evidence retrieval and SA classification. It also outperformed the other transfer learning techniques for SI classification on ScAN_VA but under-performed on ScAN_UP. As a result, for improving any publicly available clinical model using transfer learning we would recommend the use of adapter modules. If the availability of computational resources is still a problem, we would recommend using soft-prompt based fine-tuning as it uses ~ 86 times lesser parameters as compared to adapter modules while consistently performing very well across both datasets. BitFit performed well on ScAN_VA dataset but under-performed when compared with most of the fine-tuning techniques on ScAN_UP dataset.

Overall, ScANER generalizes well on new datasets and achieved a macro F1-score of 0.87 and 0.85 on two new datasets without any further fine-tuning. With the help of parameter efficient transfer learning techniques, such as adapter and soft-prompt fine-tuning, we can significantly improve the performance of ScANER on new datasets. We observed that the SA-SI label distribution and the size of the dataset can also significantly affect the SA-SI classification performance of the fine-tuned models.

5 Related Works

Laparra et al. (2021) performed an extensive review to study the recent work on building more adaptable and generalizable NLP models for clinical domain using adaptive and transfer learning techniques. They reviewed the most recent relevant work to characterize different type of methods and tasks that are being used and studied in the clinical domain. They showed that most of the work is using pre-trained language models such as BioBERT (Lee et al., 2020) and clinicalBERT (Alsentzer et al., 2019). Laparra et al. (2021) also discussed work that uses multi-task learning, sequential transfer learning and cross-lingual adaptation but did not review any recently developed parameter efficient transfer learning techniques such as adapter modules (Houlsby et al., 2019), soft-prompt tuning (Lester et al., 2021) and BitFit (Zaken et al., 2021). They also mentioned that the high costs of creating and distributing new clinical datasets favor creating a new dataset for a new task rather than creating another dataset for an existing task. In order to mitigate such imbalance, we study the effectiveness of transfer learning techniques by creating *two* new datasets for an existing task with a publicly available dataset (ScAN) and evaluating newly introduced transfer learning techniques.

Narayanan et al. (2020) studied different transfer learning techniques for adverse drug event (ADE) and medication entity extraction. They mainly focused on evaluating different biomedical contextual embeddings and using these pretrained embeddings for improved performance on their tasks. Similarly, Sun and Yang (2019) also studied the effectiveness of multilingual BERT and BioBERT for a named entity recognition (NER) task of extracting chemical and protein entities from Spanish biomedical texts. Zhou et al. (2019) adapted a CRF trained on general medical domain for NER on nursing handover data to achieve improved performance. A participant at MediQA 2019 challenge (Abacha et al., 2019) combined multiple classification tasks such as sentence classification, pairwise text classification and relevance ranking for improved performance in the shared task of the challenge. All the studies, either used a pre-trained LM or multi-task learning to improve the performance of their model on a task. Whereas in our study, we use an openly available trained LM-based classification model and further fine-tune it using recently developed parameter efficient transfer learning techniques (Houlsby et al., 2019; Lee et al., 2019; Lester et al., 2021; Zaken et al., 2021) to improve it's performance on two new datasets of the same downstream tasks.

6 Conclusion

In this paper, we evaluated different parameter efficient transfer learning techniques on the task of suicide attempt (SA) and suicide ideation (SI) events detection in the EHR notes. According to the publicly available annotation guidelines of ScAN (Rawat et al., 2022) dataset, we created two new datasets: ScAN_UP and ScAN_VA. We tested the baseline model ScANER on these two datasets and achieved macro F1-scores of 0.87 and 0.85 for SA-SI evidence detection. We were able to further improve the performance of ScANER by at least 3% after fine-tuning only 2% of ScANER's parameters. We show that parameter efficient transfer learning can help improve the performance of publicly available clinical models on new hospital datasets with few annotations. We would recommend the use of adapter modules for further transfer learning of clinical models as they consistently performed well for SA-SI detection while tuning only 2% of the parameters. If the computational resources are still a constraint, we would recommend using soft-prompt tuning as they only tune 0.02% of the parameters while achieving a performance quite close to adapter module tuning.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492.

J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. 2016. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35(35):2008–2015.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Health IT. 2022. Non-federal acute care hospital electronic health record adoption.

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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Egoitz Laparra, Aurelie Mascio, Sumithra Velupillai, and Timothy Miller. 2021. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of Medical Informatics*, 30(01):239–244.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Sankaran Narayanan, Kaivalya Mannam, Sreeranga P Rajan, and P Venkat Rangan. 2020. Evaluation of transfer learning for adverse drug event (ade) and medication entity extraction. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 55–64.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Wilfred Pigeon, and Hong Yu. 2022. ScAN: Suicide attempt and ideation events dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: an evaluation of bert in the pharmaconer task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12.

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

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Liyuan Zhou, Hanna Suominen, Tom Gedeon, et al. 2019. Adapting state-of-the-art deep language models to clinical information extraction systems: Potentials, challenges, and solutions. *JMIR medical informatics*, 7(2):e11499.

A Hyper-parameters for transfer learning techniques

Transfer Learning	# Prompts	LR	Epochs	Size
Classifier	-	1e-3	5	-
Soft prompts	20	1e-3	5	-
BitFit	-	1e-3	5	-
Adapter	-	1e-3	5	64
Last 4 layers	-	1e-4	5	-

Table 3: Best hyperparameters for classifier only, BitFit, soft promp, adapters, and last 4 layers fine-tuning