Improving prediction of visual outcomes from the electronic health record in ophthalmology using BERT and time-aware LSTMs

Stanford CS224N Custom Project

Bryan Gopal
Department of Computer Science
Stanford University
bgopal23@stanford.edu

Brian Soetikno
Department of Ophthalmology
Stanford University
briants@stanford.edu

Abstract

About 1.5 million individuals are visually impaired, which results in reduced quality of life, including increased risk of falls, depression, and anxiety. The electronic health record (EHR) contains a wealth of information about a patient, which could be used for prediction of risk of low vision. Natural language processing algorithms using deep learning could identify low-vision individuals from the EHR and could flag patients at risk for a referral to low-vision services. The purpose of this study was to investigate whether bidirectional encoder representations with transformers (BERT) models could improve the prediction of continued low-vision in the next 12 months for a patient with established low-vision (<20/40 visual acuity). We utilized a EHR database of ophthalmology clinical notes from the Stanford Byers Eye Institute. In addition, we investigated whether adding a time-aware long short-term memory layer (TLSTM) could further improve performance by capturing the sequential order of notes in the EHR. We found that the models based on BERT without the TLSTM outperformed baselines that used word-embedding-based models. Our best model achieved an AUROC of 0.844 for the prediction task of low-vision. We qualitatively analyzed our best model by performing attention visualization as measurements of intra-subject prediction variation. We conclude that BERT-based models can be a powerful method for clinical risk prediction in ophthalmology.

1 Key Information to include

- Mentor: Elaine Sui
- External Collaborators (if you have any): Sophia Wang, MD. Stanford University
- Sharing project: N/A

2 Introduction

The electronic health record (EHR) possesses a wealth of unstructured clinical data that could be unlocked for the prediction of health risk. In ophthalmology, an important measure of a patient's vision is their visual acuity, usually measured by their best line read on a standardized eye chart. Patients with low vision - defined as <20/40 visual acuity - are at increased risk of falls, depression and anxiety, and mortality. Importantly, the direct and timely intervention of providing low vision services to visually impaired patients could maximize their quality of life. However, these services are unfortunately underutilized, and patients that could benefit from low vision services receive treatment do not receive this care. Algorithms that could study a patient's EHR data and subsequently flag a patient for referral to low vision assistance could improve utilization of these services. The goal is to have algorithms help connect visually impaired patients with the services they need.

In this paper, our goal was to utilize deep learning models to predict risk of continuous low vision from a patient's EHR data. Our study builds off of previous work by Gui et al. who studied the low-vision prediction problem using a Stanford EHR database and word embedding based models [1]. In this paper, we studied two previously unexplored deep learning models for low-vision prediction: (1) Bidirectional encoder representations from transformer (BERT) and (2) BERT with a time-aware long short-term memory (LSTM).

First, we studied whether BERT-based models could improve the low-vision classification performance [2]. We implemented a classification model using SapBERT [3], a domain-specific BERT model for biomedical representations. We fine tuned the SapBERT model further on our dataset of ophthalmology notes. Due to BERT model's ability to capture long-range contextual representations, we hypothesized that this BERT-based model would outperform the baseline approaches.

Second, we studied whether using time-aware models could improve classification performance. Given that clinical notes are dispersed in time over long intervals, from days, months, and years, we reasoned that most clinical notes would provide the general context for a patient; however, the more recent notes would provide the most predictive value. Sequenced-based models, such as recurrent neural networks or long short-term memory networks (LSTM), have shown promise in improving the prediction accuracy of clinical notes. However, given that clinical notes are not evenly spaced in time, we turned to models, such as the time-aware LSTM [4][5]. We hypothesized that implementing a time-aware model would further improve predictive performance over BERT-based models.

3 Related Work

Gui et al. explored whether deep learning models using structured and free text could predict visual prognosis [1]. In our work, we utilized the same dataset, which was courteously provided by our clinical collaborator Dr. Sophia Wang, MD. The criteria for inclusion in the study were: (1) *low vision*, defined as a documented visual acuity of <20/40, and (2) at least one year of follow up. The prediction label was whether the patient continued to have a *low vision* at 1 year of followup. The index date was defined as the first date the patient had low vision. Structured data and unstructured free text notes prior to the index date were extracted from the Stanford Clinical Data Warehouse from 2009 to 2018. This structured and unstructured data served as inputs into the prediction model. Eight models were compared in the paper, using different combinations of inputs (e.g. structured inputs alone vs. unstructured free text alone) and different word embedding approaches (PubMed word embeddings vs. ophthalmology embeddings). We selected the top 3 performing neural networks (based on their AUROC values) for implementation and to serve as our baseline models. The three baseline models are a fully connected neural network (FNN), a convolutional neural network (CNN), and a FNN+CNN fusion.

Clinical notes are long documents with important information scattered throughout. Information that aids in prediction likely lies in the long range interaction between words. The embedding models, which serve as our baseline, cannot capture long-range dependencies. BERT models are suitable to capture the long-range interactions between words in notes. Applications of BERT to prediction from EHR data have been demonstrated. One such example is ClinicalBERT, which outperformed word embedding models in the task of predicting probability of 30-day readmission from EHR notes in a large open-source critical care dataset called MIMIC-III [6]. The success of ClinicalBERT in prediction tasks inspired us to using a BERT-based model for the prediction task of low-vision.

Another aspect of clinical notes that is not well captured by word-embedding models is temporal information. Patients may see the ophthalmologist once a year if their condition is chronic and less serious, or they may have a visit every month for treatment and monitoring of disease. We hypothesize that integrating time information into our model would improve predictive performance.

The time-aware LSTM (TLSTM) proposed by Baytas et al. incorporates time information into the traditional LSTM. [4]. TLSTM has forget, input, and output gates of the standard LSTM. However, short term memory content is adjusted, such that if there has been a large time gap between two notes, the previous memory has a smaller effect on the current cell's output. TLSTM performs this adjustment by weighting the short-term memory content with a function $g(\Delta t)$. The weighting function was defined as:

$$g(\Delta t) = \frac{1}{\log(e + \Delta t)} \tag{1}$$

Zhang et al. further studied the TLSTM and the application to the MIMIC-III dataset (the same dataset used for ClinicalBERT). The authors implemented a hierachical model that used ClinicalBERT to encode chunks of notes. The chunks were then fed into the TLSTM model. [5]. To make predictions, the last hidden state of the TLSTM was fed into a single layer perceptron and sigmoid layer. The authors reported improved performance over BERT-based methods alone. This work inspired us to pursue using the TLSTM to see if low vision prediction could be further improved with temporal information.

4 Approach

4.1 Baseline models

4.1.1 FNN: Fully connected neural network

Structured data alone was fed into a fully connected network with two dense layers. No clinical notes were used for this model.

4.1.2 CNN: Convolutional neural network

Using word embeddings pre-trained on abstracts related to ophthalmology [7], words from each sentence were converted into vectors. Following Gui et al., a convolutional neural network (CNN) from the paper "Convolutional neural networks for sentence classification" by [8]. The word vectors were passed to a fully connected layer before several parallel convolutional layers with varying kernels sizes. For each convolutional layer the output was passed through a MaxPool to obtain the features with maximum value (most important feature). These features were then passed through two fully connected layers, before reaching the final output single unit with sigmoid activation. Only unstructured data (clinical notes) was used to train this model.

4.1.3 FNN+CNN: Fusion

This model combined the last output layers from the FNN and CNN models (concatenation). This was then fed into two fully connected layers before reaching the final output single unit with sigmoid activation. This model thus incorporates both structured and unstructured data during training.

4.2 BERT-based models

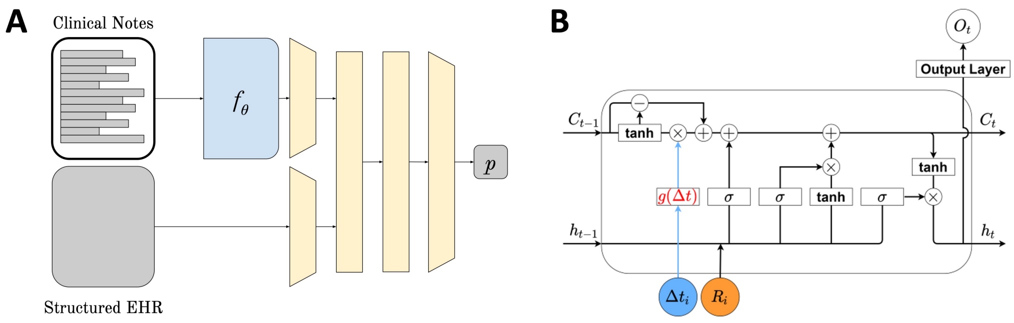

Figure 1: Structure Fused BERT Fine-tuning architecture proposed in this study. (A) Structure Fused Chunked Pretrained BERT Embeddings FNN (B) TLSTM unit (schematic adapted from [5])

We decided to improve upon our baseline models by leveraging state-of-the-art pretrained transformers as our free text encoder. We used a BERT model pretrained on a large selection of PubMED abstracts and notes as our initialization for our transformer architecture. For our first BERT pipeline, we chunked patient notes, duplicated labels across the chunk, and created a training set full of the pooler embeddings of each chunk component. We also experimented with fusion of structured data with our model at both early and late stages. Structured labels were also duplicated for fusion with the chunked BERT approach.

4.3 BERT with LSTM and TLSTM

We implemented the TLSTM by Baytas et al [4]. We utilized open-source code implemented by Zhang et al. provided in the github repository (https://github.com/zdy93/FTL-Trans) [5]. In the bottom layer of the model, the content of notes are chunked and encoded using a pre-trained SapBERT model [3]. The TLSTM layer utilizes the chunk representation and the clinical note's time information to learn a single representation for the sequence of clinical notes for a patient. We tested two versions of the model. In the first model BERT+LSTM, we set the weighting function $g(\Delta)$ equal to 1. This essentially removes any temporal weighting. In the second model, we used equation 1 as $g(\Delta t)$. Figure 1B shows the TLSTM unit and the placement of the weighting function as a multiplication of the short term memory. The final hidden layers of the model were fed into a single layer perceptron and sigmoid layer for classification.

5 Experiments

5.1 Data

Data were extracted the private ophthalmology EHR dataset provided by our collaborator Dr. Sophia Wang, MD. This is a non-deidentified dataset containing both structured data and free-text clinical notes from Stanford Hospital between 2009 and 2018 organized in a SQL database. We extracted the relevant data to our project and stored them in .csv file for training. The structure data contained 556 features, including visual acuity measurements, age and gender. Free-text clinical notes included physician notes, surgical operative reports, and imaging interpretation notes. The target prediction outcome is a binary prediction task, where 1 represents persistent low vision (<20/40 measured visual acuity) in the next 12 months and 0 represents return to normal visual acuity in the next 12 months. The dataset contained 15,409 notes related to 2258 patients that have low vision, and 52,267 notes where 3290 patients have normal visual acuity.

5.2 Evaluation method

We evaluated our models by measuring the area under the receiving operating curve (AUROC) for the binary classification task described in section 5.1.

5.3 Experimental details

5.3.1 Baseline models

The training of the FNN model only consisted of structured data. Therefore, no clinical notes were used for training. Structured data consisted of 556 features from the EHR for each patient. Our hyperparameters and optimization procedure matched that of Gui et al, and our AUROC (0.80) result was in line with that of the original paper as well.

The training of the CNN model only consisted of free-text note data. No structured data was used for training. Our hyperparameters and optimization procedure matched that of Gui et al, and our AUROC (0.82) result was in line with that of the original paper as well.

5.4 Chunked pretrained BERT FNN

For the chunked pretrained BERT-embeddings approach, instead of truncating our data to a fixed width, we chunk our data, duplicate labels across each chunk, and generate pretrained BERT pooler embeddings for each member of the chunk. These embeddings were then used to train an FNN on our binary classification task. We fine-tuned our model using an Adam optimizer, a learning rate of 1×10^{-2}, a weight decay of 1×10^{-7}, a batch size of 4096, and a Cosine Annealing with Warm Restarts learning rate scheduler.

Figure 2: Quantitative results

Model Type	AUROC
FNN	0.784
CNN	0.805
FNN + CNN	0.826
Chunked Pretrained BERT Embeddings FNN	0.702
Fine-tuned BERT	**0.843**
Structure Fused Chunked Pretrained BERT Embeddings FNN (Late Fusion)	0.712
Structure Fused Fine-tuned BERT	**0.844**
BERT w/ LSTM	0.553
BERT w/ TLSTM	0.538

5.5 Fine-tuned BERT

For the fine tuning of our BERT model, we truncated our text to an input length of 384 tokens in order to meet compute constraints. We fine-tuned our model using an AdamW optimizer, a learning rate of 1×10^{-3}, a weight decay of 1×10^{-7}, a batch size of 256, and a Cosine Annealing with Warm Restarts learning rate scheduler.

5.6 Structure Fused Models

All structure fused models shared the same optimization and hyperparameter choices as their notes-only complements.

5.7 BERT with LSTM and TLSTM Models

Because these models require a sequence of notes for training. We had to limit our training dataset to patients who had multiple encounters or visits (>=3) to the ophthalmology department. This meant the patients had multiple notes with different timestamps. The sequence of these notes could then be captured by our TLSTM model. This limited our dataset sample size to 400 patients for training.

We trained the models with learning rates from 1×10^{-4}. We inspected the gradients during training and verified that gradients were not exploding or vanishing, which can be a common problem during training of sequence based models.

5.8 Results

Figure 2 shows a table of all the measured AUROCs for the models that we experimented with in this study.

We found that BERT based models outperformed the baseline approaches only after finetuning the BERT model on ophthalmology notes (see second section of Figure 2). The fine-tuned BERT model outperformed the best baseline model FNN+CNN.

We found that adding structured data fusion to our BERT-based models only slightly improved performance (third section of Figure 2). After inspecting examples of the notes in our visualization of attention heads (see next section), we realized that this is likely because the most important structured information is likely already contained in the clinical note. For example, key structured features such as age, visual acuity, and gender are all included in the clinical notes.

We found that our BERT-based models with LSTM and TLSTM underperformed the baseline models significantly. There are a few reasons why these models under performed. First, the number of training samples was reduced for these models because they required patients that had multiple documented visits. Second, it is possible that the ordered temporal sequence of notes in the ophthalmology setting is not an important predictor of low vision. In the critical care setting, where these models have been shown to outperform BERT models without sequence structure, the timing of events likely has stronger influence on the prediction of tasks such as mortality and readmission [5][9]. As a critical care example, it is important to know that a patient had hypotension, was diagnosed with a heart attack, and received cardiac catherization as treatment, in that specific order, for determining mortality or readmission. However, in ophthalmology, the temporal order may not be as important as many eye conditions have a chronic nature and interventions given to patients may be given over long periods of time. Therefore, we do not think it is surprisingly that the BERT based LSTM and TLSTM models underperformed our baseline models.

6 Analysis

6.1 Attention Head Visualization

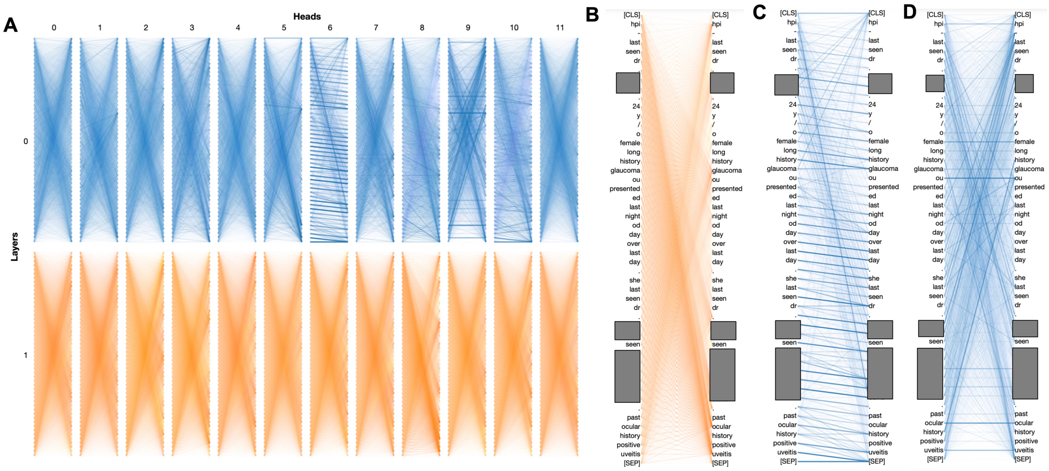

Figure 3: Examples of attention visualization. (A) Attention maps of the 0th and 1st layers of the BERT model. (B-D) selected examples of attention head weights for a representative chunk of a patient note. Grey boxes are censored areas of the note in order to protect patient privacy.

In order to qualitatively evaluate our best performing models, we turned to attention head visualization to attempt to gain insight into our model, which is depicted in Figure 3.

We carefully studied these attention maps for several representative samples of patient notes. We show one example in Figure 3 for illustration. We found three patterns in the attention heads worth mentioning.

First, we noticed that some attention heads had strong attention patterns on key clinical words. For example in Figure 3(B), attention was drawn mostly to words such as ’glaucoma’ and ’uveitis’, which are key clinical disease entities.

Second, we noticed that some attention heads developed the ’next word’ attention pattern, where attention is focused on the next word of a sequence (Figure 3(C)). It is reasonable that attention would focus on the next word because the next word often helps understand a word’s context.

Third, we found attention patterns where attention is divided evenly across all the words in a chunk (Figure 3(D)). These type of patterns illustrate how attention can allow long range contextual interactions within a clinical note.

6.2 Structure Fused Fine-tuned BERT Longitudinal Analysis

We would like to analyze the behavior of our best performer (structure fused fine-tuned BERT) at predicting low-vision outcomes of patients who had multiple visits recorded in our dataset. To do so, we analyzed the variation in prediction scores for this model for patients that had a change in low vision prognosis over time (either from 0 to 1 or 1 to 0) and compared this variation to that of patients that had consistently positive or negative labels. Our results are shown in Figure 4. Although the model fails to model changes in label status across visits (predicting the same label uniformly for all visits for a given patient), the variance in predicted scores is much higher than the categories where the ground truth label remained the same. This suggests that adding some time-aware component to this pipeline may be beneficial as a future experiment.

Figure 4: Prediction Variation based on Label Type

Patient Label Progression	Average Prediction Score Variation
0 only	1×10^{-20}
1 only	1×10^{-20}
Label Change	1×10^{-8}

7 Conclusion

We successfully implemented BERT-based models to predict visual outcomes in a single-center ophthalmology EHR dataset. We found that BERT based models can outperform word embedding based models especially after fine tuning on the corpus. We also experimented with time-aware LSTM's to capture the temporal sequence of notes; however, we found that this underperformed baseline models. We used attention visualization to qualitatively understand the attention patterns in our BERT-based models.

The primary limitation of this work is that it was limited to a single center, thus results of these EHR models may not generalize to other populations or hospitals as the structure of clinical notes varies between healthcare systems.

The incorporation of time into prediction modelling is still an important area for future research. While the temporal sequence of notes in the ophthalmology setting may not be important for prediction, as evident by the poor performance of our LSTM-based models, incorporation of time into the model in other ways may still provide additional value. One possibility would be to encode the temporal order of a sequence using positional encodings rather than a recurrent structure such as LSTM. This has been demonstrated by the "Attend and Diagnose" model by Song et al [10]. Future work may explore this representation of time in clinical notes.

References

[1] Haiwen Gui, Benjamin Tseng, Wendeng Hu, and Sophia Y. Wang. Looking for low vision: Predicting visual prognosis by fusing structured and free-text data from electronic health records. *International Journal of Medical Informatics*, 159:104678, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online, June 2021. Association for Computational Linguistics.

[4] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 65–74, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pages 566–588. PMLR, 2020.

[6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

[7] Sophia Wang, Benjamin Tseng, and Tina Hernandez-Boussard. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. *International Journal of Medical Informatics*, 150:104464, 2021.

[8] Y Kim. Convolutional neural networks for sentence classification. arxiv 2014. *arXiv preprint arXiv:1408.5882*, 2019.

[9] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[10] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*, 2018.