UNIVERSITY OF CAPE TOWN

MASTERS THESIS

Hospital readmission prediction with long clinical notes

Author:
Yassin NURMAHOMED

Supervisor:
Dr. Jan BUYS

Minor Dissertation presented in
partial fulfilment of the requirements for
the degree of Master of Science

in the

Department of Computer Science

October 20, 2022

Declaration of Authorship

I, Yassin NURMAHOMED, declare that this thesis titled, "Hospital readmission prediction with long clinical notes" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Signed by candidate

Date: 20/10/2022

UNIVERSITY OF CAPE TOWN

Abstract

Faculty of Science

Department of Computer Science

Master of Science

Hospital readmission prediction with long clinical notes

by Yassin NURMAHOMED

Electronic health records (EHR) data is captured across many healthcare institutions, resulting in large amounts of diverse information that can be analysed for diagnosis, prognosis, treatment and prevention of disease. One type of data captured by EHRs are clinical notes, which are unstructured data written in natural language. We can leverage Natural Language Processing (NLP) to build machine learning (ML) models to gain understanding from clinical notes that will enable us to predict clinical outcomes. ClinicalBERT is a pre-trained Transformer based model which is trained on clinical text and is able to predict 30-day hospital readmission from clinical notes. Although the performance is good, it suffers from a limitation on the size of the text sequence that is fed as input to the model. Models using longer sequences have been shown to perform better on different ML tasks, even with clinical text. In this work, a ML model called Longformer which pre-trained then fine-tuned on clinical text and is able to learn from longer sequences than previous models is evaluated. Performance is evaluated against the Deep Averaging Network (DAN) and Long short-term memory (LSTM) baselines and previous state-of-the-art models in terms of Area under the receiver operating characteristic curve (AUROC), Area under the precision-recall curve (AUPRC) and Recall at precision of 70% (RP70). Longformer is able to best ClinicalBERT on two performance metrics, however it is not able to surpass one of the baselines in any of the metrics. Training the model on early notes did not result in substantial difference when compared to training on discharge summaries. Our analysis shows that the model suffers from out-of-vocabulary words, as many biomedical concepts are missing from the original pre-training corpus.

Acknowledgements

I would like to thank my supervisor Dr. Jan Buys for his indispensable guidance and insights.

The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za.

Contents

List of Figures

List of Tables

List of Abbreviations

BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CNN	**C**convolutional **N**eural **N**etwork
DAN	**D**eep **A**veraging **N**etwork
EHR	**E**lectronic **H**ealth **R**ecord
ICU	**I**ntensive **C**are **U**nit
LSTM	**L**ong **S**hort-**T**erm **M**emory
ML	**M**achine **L**earning
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**rocessing
QA	**Q**uestion **A**nswering
ReLU	**R**ectified **L**inear **U**nit
RNN	**R**ecurrent **N**eural **N**etwork

Dedicated to Mother and Father for their unconditional suppport in this endeavour.

Chapter 1

Introduction

Natural Language Processing (NLP) involves processing and analysing of large quantities of natural language data with computers in order to extract information and insights contained in collections of this kind of data. Vast amounts of natural language data, in the form of text, are available in the clinical domain due to the growing usage of Electronic Health Records (EHR), these are designated "clinical notes".

Current state-of-the-art methods for NLP tasks are based on a Machine Learning technique called Deep Learning, which enables computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction (Salakhutdinov, 2014).

The Transformer (Vaswani et al., 2017) is increasingly the Deep Learning model and architecture of choice for NLP tasks like text classification (Yang et al., 2019), machine translation (Conneau and Lample, 2019) and summarization (Lewis et al., 2020). The model is shown to be superior in quality while being more parallelizable than previous state-of-the-art models. However, there is a limitation on the amount of data that can be processed due to its computational complexity.

Longformer (Beltagy, Peters, and Cohan, 2020) is a transformer-based model with a drop-in replacement for the standard attention that makes it easy to work with longer sequences. It outperforms other transformer models on long document tasks based in general domain text, yet there has not been reports of its usage in the clinical domain.

This research aims to evaluate a model with a more computationally efficient implementation of the attention mechanism, which enables processing of longer sequences than was possible with the standard Transformer. This is achieved by using evaluation metrics to assess the accuracy of such model on a text classification task using clinical notes as its input.

This chapter consists of introductory material to the thesis, first it describes electronic health records, followed by NLP research with clinical notes, the Transformer, Longformer, research questions, tools and approach, contributions and at the end an overview on the structure of the dissertation.

1.1 Electronic Health Records

Electronic health records (EHR) consist of different types of patient data: demographics, diagnoses, laboratory test results, prescriptions, clinical notes and images (Xiao, Choi, and Sun, 2018). This data is captured across many healthcare institutions, resulting in large amounts of diverse information that can be analysed for diagnosis, prognosis, treatment and prevention of disease. It mainly consists of structured and unstructured data, the latter being more abundant as it is easy for humans to create.

Clinical notes are unstructured data written in natural language that contain information about patients. Examples of types of clinical notes are: nursing, radiology and discharge summaries. Compared to structured data, they provide a more detailed overview of the patient.

The following text is an extract from a discharge summary where personal identifiable information has been removed in order to protect the patient's privacy.

```
Admission Date:  [**2163-9-10**]              Discharge Date:   [**2163-9-12**]

Date of Birth:  [**2079-1-17**]             Sex:   M

Service: MEDICINE

Allergies:
Penicillins / Erythromycin Base / Streptomycin / Citric Acid /
Atenolol / Torsemide / Heparin Agents

Attending:[**First Name3 (LF) 99**]
Chief Complaint:
Bleeding from colostomy and foley


Major Surgical or Invasive Procedure:
none

History of Present Illness:
84 year old male with multiple co-morbidities including rectal
cancer s/p resection and radiation in [**2157**] now with colostomy,
coronary artery disease s/p stents, systolic CHF, dilated
cardiomyopathy, atrial fibrillation not on [**Year (4 digits) **], cardiac
arrest and complete heart block s/p AICD/pacemaker, recent
trach/peg after prolonged hospitalization for rib
fractures/flail chest s/p fall who presents with large amount of
bleeding from colostomy and Foley. The patient also endorsed
increased shortness of breath, weakness and fatigue.
```

1.2 NLP Research with Clinical Notes

Clinical notes have been used in NLP research for various tasks such as text summarization and information extraction. McInerney et al. (McInerney et al., 2020) created a model to extract relevant text snippets to provide a summary to aid physicians considering diagnoses, while Wiegreffe et al. (Wiegreffe et al., 2019) explored ways to improve linking spans of text to concepts in a detailed domain ontology using deep learning.

Additionally, they have been used for predicting chronic diseases (Liu, Zhang, and Razavian, 2018), readmission (Huang, Altosaar, and Ranganath, 2020), mortality (Ghassemi et al., 2014), medical codes (Ford et al., 2016) and length of stay (Rajkomar et al., 2018).

While notes are commonly combined with structured data to use as features for ML models, sometimes information related to diseases is mostly found in text (Escudié et al., 2017). Models using text can outperform models using just structured data (Liu, Zhang, and Razavian, 2018; Hsu et al., 2020).

Clinical text can contain a lot of noise introduced during the annotation process and often contains heavy amounts of abbreviations and acronyms, a challenge for NLP models. It also suffers from imbalanced data distribution, where one class can contain many more instances than the other classes, which can cause an ML model to be biased to the dominant class (Nguyen and Patrick, 2016).

1.3 Predicting Hospital Readmission

Hospital readmission occurs when a patient is discharged from a hospital and later is admitted again within a specific time frame. Unplanned hospital readmissions are a significant contributor to unfavourable patient clinical outcome and high healthcare costs. It can be used to identify quality-of-care problems and can influence healthcare funder's decisions (Dreyer and Viljoen, 2019; Benbassat and Taragin, 2000).

Readmission prediction can aid healthcare-providers in deciding if a patient is ready for discharge. If not deemed ready, he may go through additional interventions before being considered ready for discharge, resulting in a lower readmission rate and reduced healthcare costs.

In terms of Machine Learning, it can be framed as a binary classification problem where the positive class symbolizes a readmission while the negative no readmission. This work will make use of this task in order to evaluate the performance of the machine learning models developed.

1.4 Transformer

Recently, transformer-based models revolutionized the NLP field. The Transformer is a simple neural network architecture, based on an attention mechanism. It was shown to be superior in quality while requiring less time to train (Vaswani et al., 2017).

At the time of its introduction, this architecture achieved state-of-the-art results on machine translation tasks, and was later applied successfully in other modalities such as image processing.

Given an input sequence, the attention mechanism provides a means of context around every position, allowing the Transformer to read the sequence elements without any specific order. This allows more parallelization than previous state-of-the art models, in turn reducing training times and allowing training on larger datasets.

The main disadvantage of the model is the attention mechanism's computational complexity that is quadratic with respect to the input sequence length. This characteristic makes it so that implementations have to settle on a reasonable input sequence length, thus limiting the context size around each element.

1.5 Longformer

The Longformer is a transformer-based model with an attention mechanism proposed as a drop-in replacement for the standard Transformer's. It has a linear complexity, making it possible to work with longer sequences (Beltagy, Peters, and Cohan, 2020).

Longformer is part of a family of efficient Transformer variants that define a form of sparse attention pattern to avoid computing the full quadratic attention, it builds on the work presented in other models like Sparse Transformers (Child et al., 2019) and BlockBERT (Qiu et al., 2020).

The standard transformer uses an attention mechanism where each position in the sequence attends to all other positions. In the Longformer this is replaced with a fixed sized window in the element neighbourhood. The receptive field can be increased without decreasing performance by dilating the window with the introduction of fixed size gaps.

This model achieved state-of-the art results on some long document question answering tasks.

1.6 Research Questions

Unplanned hospital readmission can have significant consequences for both patient and healthcare institutions. For patients, it often contributes to unfavourable outcomes and for healthcare institutions it can be seen as an indicative of quality-of-care problems.

One way of lowering readmission is to correctly assess readiness for discharge and act on patients that are most at risk. Readmission rates have been reported to decline after the implementation of reviews preceding discharge and improved follow-up after discharge (Benbassat and Taragin, 2000). Interventions aimed at preventing readmission can be cost-saving at reasonable success rates (Safran and Phillips, 1989).

Clinical notes have been used to predict hospital readmission, however they have disproportional contribution to the task. Discharge summaries were found to be especially predictive for readmission, as they aggregate information from the entire admission (Hsu et al., 2020).

Assessing the chance of readmission early in treatment is an important factor for the applicability of the model in a clinical setting. Discharge summaries are only available after the patient has been discharged. Models designed to be used early in the hospital stay may serve as a real-time decision support tool for clinicians to target high risk patients for discharge planning and other preventive interventions (Tabak et al., 2017). Previous work has shown that it is possible to get accurate predictions using subsets of all clinical notes produced during a hospital stay (Hsu et al., 2020; Huang, Altosaar, and Ranganath, 2020).

In order to produce an improved Transformer-based model that is able to process longer sequences of clinical text, this work focuses on the following questions:

1. Will using longer text sequences (1K, 2K and 4K tokens) with the Longformer model result in improved 30-day readmission prediction than the baseline models or Clinical-BERT?
2. How many days after the first admission can we predict readmission using all the notes available, without a drop in performance of more than 10%, compared to using only discharge summaries?

1.7 Contributions

This work focuses on the evaluation of a transformer model with the Longformer's attention mechanism on clinical domain text. Along with this model additional models are evaluated, a standard attention Transformer model (Huang, Altosaar, and Ranganath, 2020), a Deep Averaging Network (Iyyer et al., 2015) and a Long short-term memory network (Hochreiter and Schmidhuber, 1997). A clinical notes dataset is processed and divided into three distinct sets for training, evaluation and testing.

The outcome of this work can be applied in developing more accurate machine learning models for clinical domain tasks. These models can aid healthcare professionals in decision-making and potentially save lives.

All source code for data processing, model training and evaluation is available under an open-source licence.

1.8 Structure of the Dissertation

Chapter 2 consists of the literature review where aspects related to Deep Learning, baseline models and Transformers are described in more detail. The methodology is then explained in chapter 3. Chapter 4 presents the results of the experiments done in order to evaluate the model's performance and provides a discussion on results obtained. Finally, chapter 5 provides concluding remarks and proposes future work.

Chapter 2

Literature Review

2.1 Deep Learning

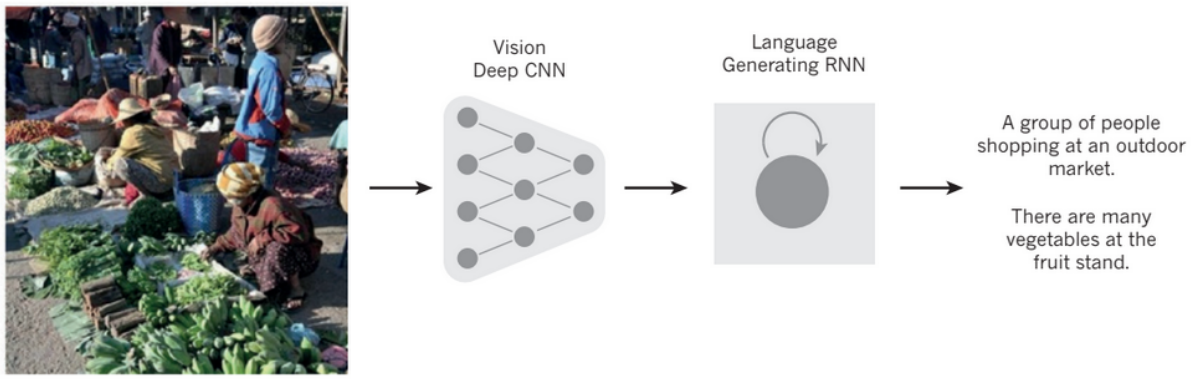


FIGURE 2.1: CRNN trained to 'translate' high-level representations of images into captions (Salakhutdinov, 2014).

Deep learning shifted the machine learning paradigm from manual feature engineering to automatic feature learning. It improved state-of-the-art models in speech recognition, visual object recognition, object detection, drug discovery and genomics (Salakhutdinov, 2014).

These kinds of models are mostly based on multi-layer neural networks and used on supervised learning tasks. Optimization is done using the gradient descent algorithm with backpropagation of gradients from the output layer, through hidden layers all the way to the input layer.

With representation learning, a neural network can be fed with raw data and automatically learn the features needed for different tasks, dispensing the need for domain experts who can label the data. This network can then be joined with another task specific network, be it object detection or classification. The whole deep network is then fine-tuned using backpropagation. The feature learner can be trained separately from the task specific network and the resulting weights used to initialize the deep network, this process is called pre-training.

A form of representation learning prevalent in deep learning for NLP are word embeddings. These are built by a network that is fed sequences of text and learns how to encode the vocabulary into a vector space. This learned space has real-valued components and semantically related words, or those that are used in similar ways, end up close to each other by having similar representations (Sutskever, Vinyals, and Le, 2014; Mikolov et al., 2013). This vector representation can be obtained by means of a model specifically designed to learn word embeddings, such as word2vec (Mikolov et al., 2013) or GloVe (Pennington, Socher, and Manning, 2014).

One drawback of word embedding is that the vectors generated are static and dissociated with the context in which the word is used, e.g. the word "bat" would have the same embedding either referring to the animal or to a baseball bat. To overcome this limitation, new models that learn contextual word embeddings were proposed (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). These are called language models, which are models that predict how likely a sequence of words is to appear in a given language (Gomez-Perez, Denaux, and Garcia-Silva, 2020).

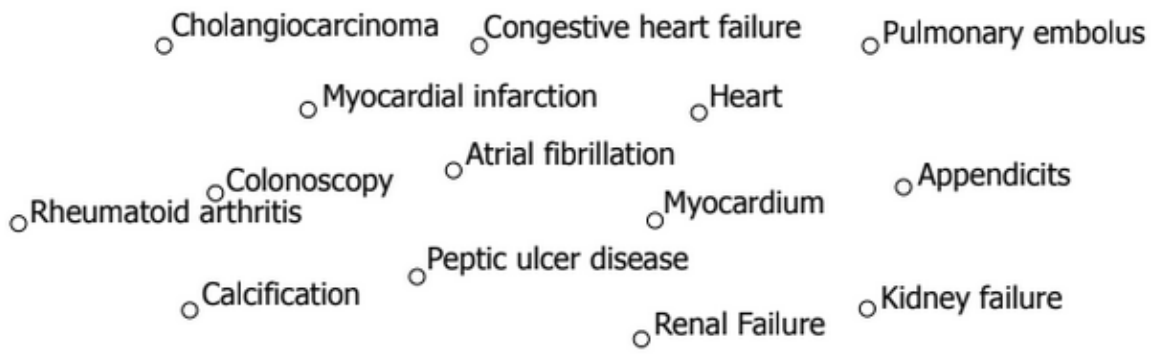


FIGURE 2.2: Word embeddings of clinical terms relating to heart conditions (Huang, Altosaar, and Ranganath, 2020).

Common architectures for deep learning are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The following are examples of state-of-the-art models at the time Salakhutdinov, 2014 was published: ImageNet (Krizhevsky, Sutskever, and Hinton, 2012) used CNNs to almost halve the error rate for object recognition; Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), which have become a crucial ingredient in recent advances with RNNs because they are good at learning long-range dependencies.

2.2 Deep Averaging Networks

The Deep averaging network (DAN) is a Bag-of-Words like model, meaning that it does not take word order and sentence structure into account. It shows significant improvement over previous models by deepening the network and applying a novel variant of dropout (Iyyer et al., 2015).

While many existing deep learning models for NLP tasks focus on learning the compositionality of their inputs, which requires many expensive computations, for some tasks nonlinearly transforming the input is more important.

Considering the text classification task: map an input sequence of tokens X to one of k labels. Iyyer et al., 2015 defined the neural Bag-of-Words model as follows: first, a composition function g is applied to the sequence of word embeddings v_w for $w \in X$. The output of this composition function is a vector z that serves as input to a logistic regression function. In this particular instantitation the compositon function g averages word embeddings

$$z = g(w \in X) = \frac{1}{|X|} \sum_{w \in X} v_w. \tag{2.1}$$

Feeding z to a softmax layer induces estimated probabilities for each output label

$$\hat{y} = softmax(W_s \cdot z + b), \tag{2.2}$$

where the softmax function is

$$softmax(q) = \frac{exp\, q}{\sum_{j=1}^{k} exp\, q_j}. \tag{2.3}$$

W_s is a $k \times d$ matrix for a dataset with k output labels, and b is a bias term.

The model is trained to minimize the cross-entropy error, which for a single training instance with ground-truth label y is

$$l(\hat{y}) = \sum_{p=1}^{k} y_p log(\hat{y}_p). \tag{2.4}$$

The DAN model is obtained by adding depth to the neural Bag-of-Words model. In Equation 2.1, z is computed by averaging the word vectors $v_{w \in X}$, it is the vector representation for the input text X. Instead of directly passing this representation to an output layer, we can further transform z by adding more layers before applying the softmax layer. Suppose we have n layers, $z_{1 \cdots n}$. Each layer is computed as:

$$z_i = g(z_{i-1}) = f(W_i \cdot z_{i-1} + b_i), \tag{2.5}$$

then the final layer's representation, z_n, is fed to a softmax layer for prediction (Figure 2.3).

The intuition behind deep feed-forward neural networks is that each layer learns a more abstract representation of the input than the previous one (Bengio, Courville, and Vincent, 2013). The same concept is applied in order to obtain the DAN with the expectation that each layer will increasingly magnify small but meaningful differences in the word embedding average. Although the model does not incorporate word order and syntax like in other deep learning models, in practice it can outperform such models on some tasks while taking only a fraction of the training time.

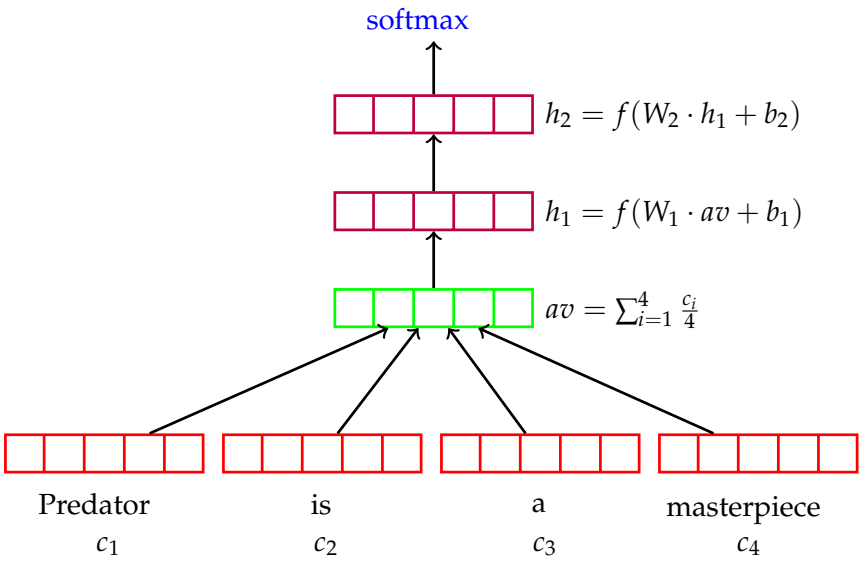


FIGURE 2.3: A two-layer DAN (Iyyer et al., 2015).

2.3 Recurrent Neural Networks

RNNs are networks for modelling sequences. They process the input one element at a time, while maintaining a hidden state that contains information about all the elements that it has processed.

Typical uses for RNNs are predicting the next character in a word or the next work in a sentence. They have also been used in sequence to sequence prediction or sequence transduction, where the goal is no generate an output sequence based on a different input sequence. One instance of this problem is language translation (Sutskever, Vinyals, and Le, 2014).

Training traditional RNNs has been problematic because over many iterations the gradients often explode or vanish, there is an inherent trade-off between learning and capturing longer dependencies between elements in the input sequence (Bengio, Simard, and Frasconi, 1994).

In order to moderate this trade-off, LSTMs were introduced. The main idea is to add an explicit memory to the network, LSTMs are composed of cells which act as the memory and each cell contains an input, output and forget gate that learn to decide if the memory should be accumulated or cleared (Hochreiter and Schmidhuber, 1997).

LSTM based models have been successfully used in an encoder-decoder architecture to perform language translation (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014). This type of model consists of two RNNs, where one encodes a sequence of text into a vector representation and the other decodes the representation into a different sequence of text. In Cho et al., 2014 it is shown that the model learns a semantically and syntactically meaningful representation of linguistic phrases.

2.3.1 Long short-term memory (LSTM)

A LSTM (Hochreiter and Schmidhuber, 1997) is an RNN architecture. Each LSTM block is composed of memory cells and gate units. The gate units, input, output and forget gates, regulate the passage of information in and out of the memory cells.

The LSTM was built by analysing the error flow in other RNN architectures, that led to the discovery that long sequences were beyond the capacity of existing architectures, due to the backpropagation of error that can either get too large or too small (Graves and Schmidhuber, 2005).

For each element in the input sequence, the following functions are computed

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
\tag{2.6}
$$

where h_t is the hidden state at time t, c_t is the cell state at time t, x_t is the input at time t, h_{t-1} is the hidden state of the layer at time $t-1$ or the initial hidden state at time 0, and i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively. σ is the sigmoid function, and $\odot$ is the Hadamard product.

The LSTM blocks are analogous to memory chips in computers in the sense that the input, output and forget gates of the cell are similar to write, read and reset operations of a memory chip. During its processing the cell's input is multiplied by the input gate's activation function, the previous cell's output by the forget gate, finally the network's output is multiplied by the activation function of the output gate (Graves and Schmidhuber, 2005). An illustration of this process is shown in Figure 2.4.

LSTMs have been used in handwriting recognition and generation, language modelling and translation, acoustic modelling of speech, speech synthesis, protein secondary structure prediction, analysis of audio, and video data among others. In many of these applications, they were able to advance the state-of-the-art (Greff et al., 2015). An example of a successful application of LSTMs on NLP is ELMo (Peters et al., 2018), which is a model for learning contextual word embeddings that was used to improve the state-of-the-art on some NLP tasks.

2.4 Transformer

Before the introduction of the Transformer, machine learning models for sequence transduction were predominantly based on recurrent or convolutional neural networks, with the best

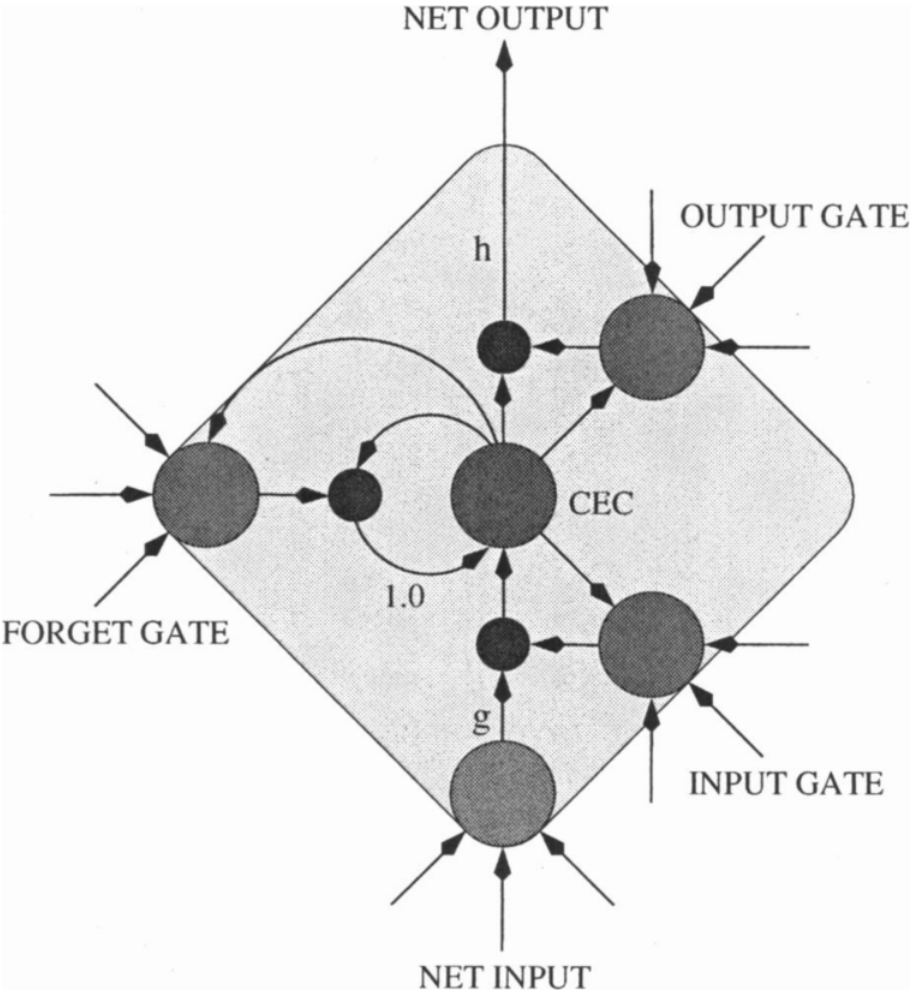


FIGURE 2.4: LSTM memory block with one cell. The internal state of the cell is maintained with a recurrent connection of weight 1.0. The three gates collect activations from inside and outside the block, and control the cell via multiplicative units (small circles). The input and output gates scale the input and output of the cell, while the forget gate scales the internalstate (Graves and Schmidhuber, 2005).

performing of these using an encoder-decoder structure connected through an attention mechanism (Vaswani et al., 2017).

Recurrent models typically perform computation in a way that is aligned with the processing of input: one step at a time. This trait of the model prevents parallelization within training examples, which becomes necessary when using longer sequence lengths. Attention, first introduced by Bahdanau, Cho, and Bengio, 2015, is a mechanism that allows modelling of dependencies without regard to their distance in the input or output sequences.

The transformer rids of recurrence and convolutions, and it relies exclusively on the attention mechanism, it is shown to be superior in quality while being more parallelizable, the only downside being the per-layer complexity of $O(n^2 \times d)$, where n is the sequence length and d the dimension of the learned representation.

2.4.1 Model Architecture

The model consists of an encoder-decoder structure. The encoder maps an input sequence of symbol representations $(x_1, \cdots, x_n)$ to a sequence of continuous representations $z = (z_1, \cdots, z_n)$. Given z, the decoder then generates an output sequence $(y_1, \cdots, y_m)$ of symbols one element at a time. At each step the model consumes the previously generated symbols as additional input when generating the next.

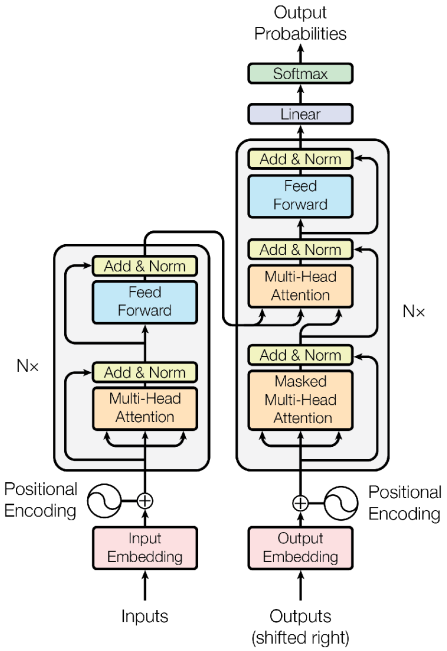


FIGURE 2.5: The Transformer model architecture (Vaswani et al., 2017).

2.4.2 Encoder and Decoder

Encoder: The encoder is composed of a stack of identical layers, in the original implementation of the Transformer $N = 6$.

Each layer has two sub-layers, a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The first encoder layer receives a sequence of input embeddings which are then processed by the self-attention sub-layer followed by the feed-forward network, the output of this network is then passed to the next encoder layer.

The self-attention mechanism is a means of relating different elements of a sequence in order to compute a representation of the whole sequence. It is described with more detail in section 2.4.3.

The fully connected feed-forward network consists of two linear transformations with a Rectified Linear Unit (ReLU) activation function in between, it is applied to each element from the self-attention sub-layer separately and identically.

$$ffn(z) = max(0, zW_1 + b_1)W_2 + b_2 \tag{2.7}$$

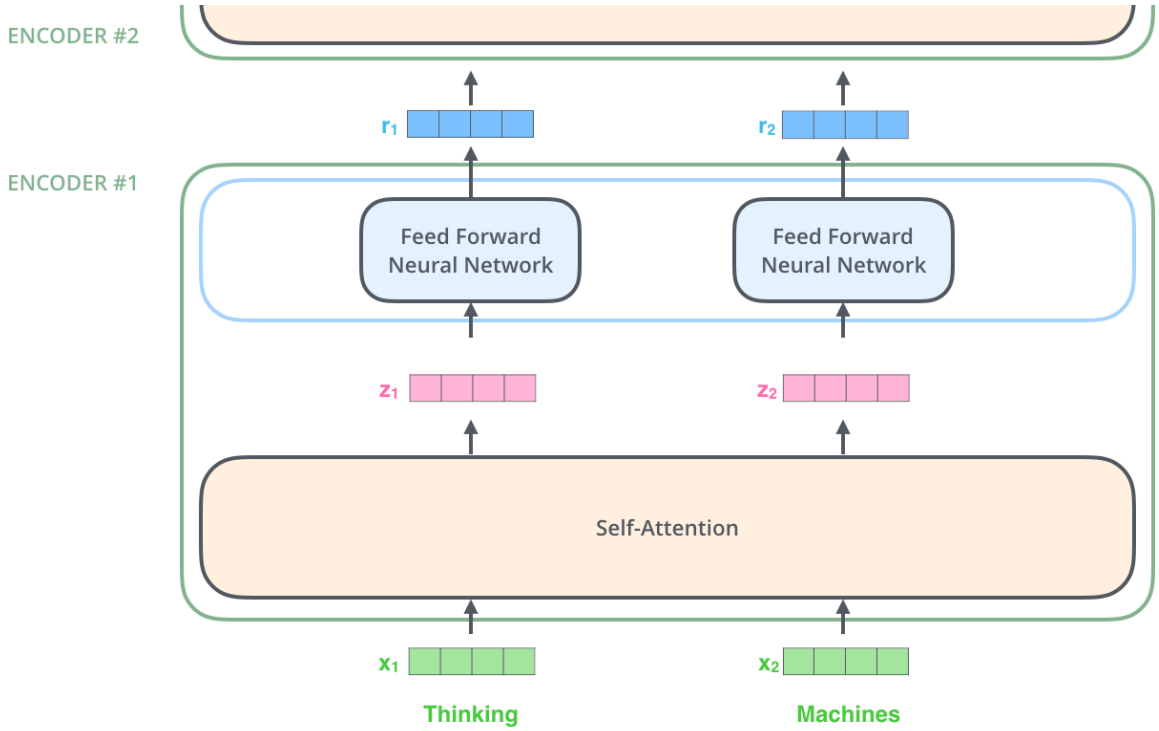


FIGURE 2.6: Encoder sub-layers (Alammar, 2018).

A residual connection is utilized around each of the two sub-layers, followed by layer normalization. This means that input from each sub-layer is combined with its output and the resulting vector of this operation is normalized before passing to the next layer. This process is one way to reduce training time and improve generalization performance (Ba, Kiros, and Hinton, 2016).

Decoder: The decoder is similar to the encoder. In addition to the two sub-layers in the encoder, it inserts another layer, which is a modified multi-head attention layer for calculating the attention scores over the output of the encoder. This structure is generally using in sequence-to-sequence settings like machine translation or question answering. Because it is not needed when implementing a classifier, minimal details are provided.

2.4.3 Attention

The attention mechanism's purpose is to encode an understanding of each element in the input sequence in relation to all the others.

It works by first computing a query, key and value vectors by multiplying the input embeddings with three weight matrices W^Q, W^K and W^V that are learned during training. Prior to

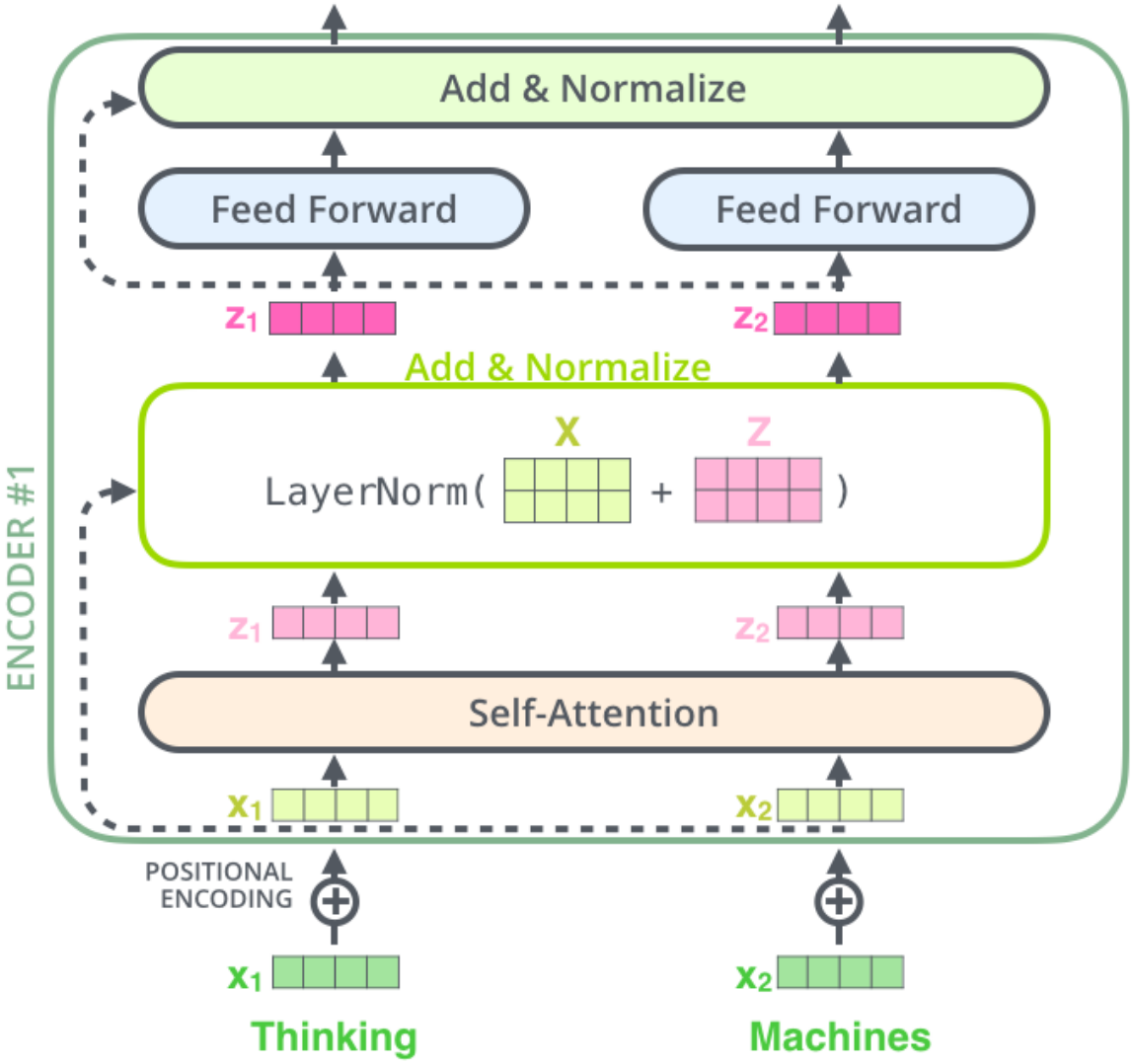


FIGURE 2.7: Encoder sub-layers (Alammar, 2018).

the multiplication, these embeddings are packed into a matrix X

$$\begin{aligned} Q &= X \times W^Q \\ K &= X \times W^K \\ V &= X \times W^V. \end{aligned} \tag{2.8}$$

The resulting query (Q) and key (K) vectors have a dimension of d_k, and the value (V) vectors have a dimension of d_v.

Then the attention scores are calculated by a method called "Scaled Dot-Product Attention", it consists of computing the dot products of the query with all keys, dividing each by $\sqrt{d_k}$, and

finally applying a softmax function to the result

$$attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{2.9}$$

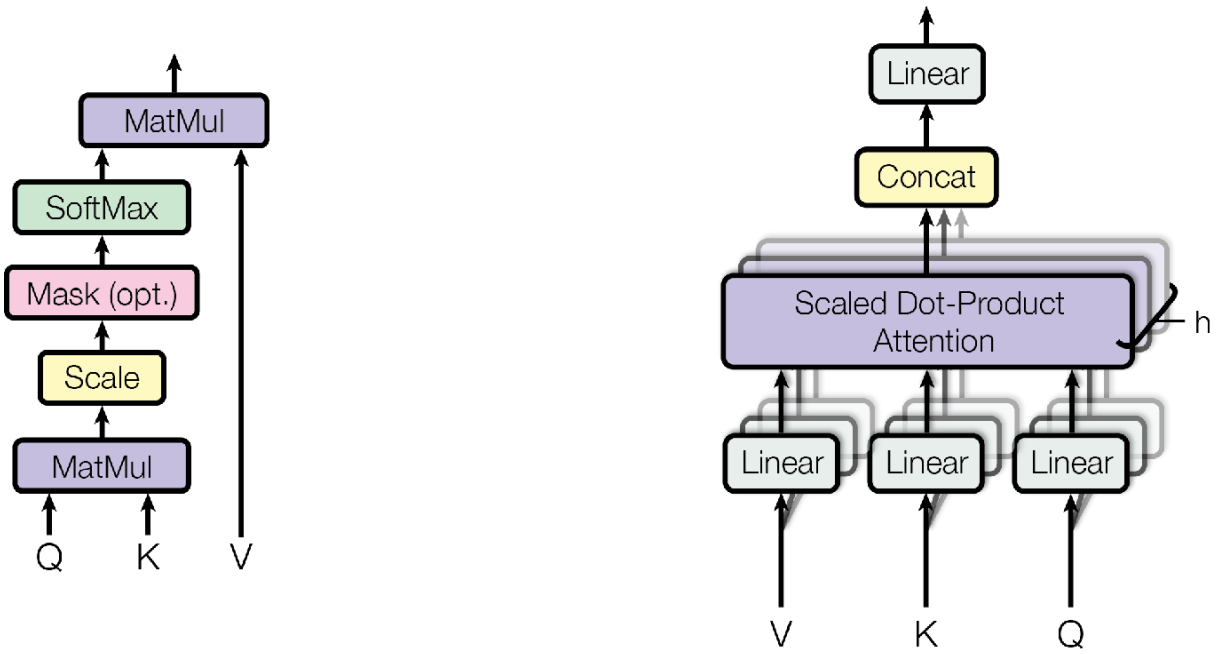


FIGURE 2.8: Scaled Dot-Product Attention (left) and Multi-Head Attention (right) (Vaswani et al., 2017).

The attention computation is done in parallel over h attention "heads", this means that the queries, keys and values matrix are computed using h different learned weight matrices, before being concatenated and multiplied with a final weight matrix

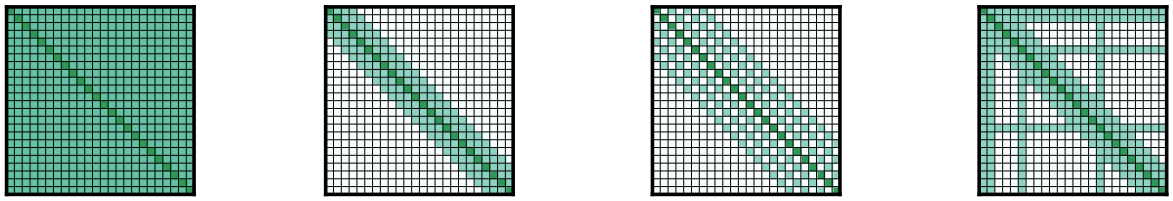
$$multiHead(Q, K, V) = concat(head_1, \cdots, head_h)W^O \tag{2.10}$$

where $head_i$ is calculated using Equation 2.9.

The computational complexity of Equation 2.9 is $O(n^2 \cdot d)$ (Vaswani et al., 2017).

2.5 Longformer

The Longformer is proposed as a drop-in replacement for the standard transformer self-attention mechanism, it has linear complexity making it possible to work with longer sequences (Beltagy, Peters, and Cohan, 2020).



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

FIGURE 2.9: Comparison of attention mechanisms (Beltagy, Peters, and Cohan, 2020).

In previous work, Sparse Transformers (Child et al., 2019) used sparse factorizations of the attention matrix devised by visualizing the attention patterns learned by a standard Transformer on an image dataset, while BlockBERT (Qiu et al., 2020) sparcified the attention matrix by means of a sparse block masking matrix.

The standard transformer uses an attention mechanism where each element in the sequence attends to all other elements. In Longformer this is replaced with a fixed sized window in the element neighbourhood (Figure 2.9). For each token, the attentions scores need only to be computed for itself and a few preceding and succeeding tokens in the sequence.

Given an input of size n and window size w, each element attends to $\frac{1}{2}w$ neighbour elements, resulting in a complexity of $O(n \times w)$.

The receptive field can be increased without increasing complexity by dilating the sliding window by introducing gaps of fixed size.

One drawback of the sliding window approach is that it is not flexible enough to learn task-specific representations, in order to combat this symmetric global attention is added on pre-selected input locations. For classification, global attention is used for the class token while for question-answering in all question tokens.

2.6 Bidirectional Encoder Representations from Transformers (BERT)

A common practice in deep learning is pre-training where a model is optimised on large amounts of data and then the resulting weights are used for initializing another model. The process of initializing from a pre-trained model is called fine-tuning. Normally, pre-training is done using amounts of data that could not be processed using a single conventional computer. The advantage of this approach is that during fine-tuning fewer weights need to be adjusted, resulting in more efficient training and more performant models.

BERT (Devlin et al., 2019) is a Transformer-based model designed to pre-train bidirectional representations from unlabelled text and fine-tuned, without significant architecture alterations, to create models for more specific tasks, such as classification, Question Answering (QA) and Natural Language Inference (NLI).

To make the model able to handle various fine-tuning tasks, the input representation is able to represent both a single sentence and a pair of sentences in one token sequence without ambiguity. The first token of every sequence is always a special classification token (`[CLS]`), it is used as a sequence representation in the final Transformer layer, and it passed to a classification layer. In the case of sentence pairs, after the `[CLS]` token comes the first sentence followed by the separation token (`[SEP]`) and then the second sentence as illustrated in Figure 2.10.

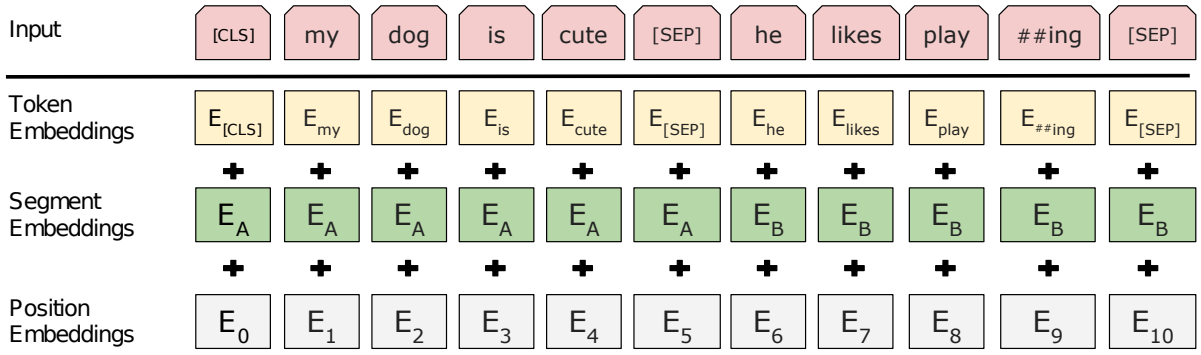


FIGURE 2.10: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings (Devlin et al., 2019).

Pre-training is done using two unsupervised tasks, masked language modelling and next sequence prediction.

In the first task, a percentage of the input tokens are masked, and the model is trained to predict those masked tokens. This enables the model to learn a bidirectional representation of the input, in contrast with standard language models that can only be trained left-to-right or right-to-left.

The second task each example consists of sentence pairs, the model is then trained to predict if the second sentence follows the first. This task results in a model that understands sentence relationships, the basis of QA and NLI. The model was trained on BooksCorpus (Zhu et al., 2015) (800M words) and English Wikipedia (2,500M words) which have 13 GB plain text combined.

For fine-tuning, BERT is able to model many tasks involving single sentence or sentence pairs by plugging in the appropriate inputs and outputs. Sentence `A` and sentence `B` in Figure 2.10 would be two distinct sentences in the paraphrasing task, a hypothesis and premise in the entailment task, a question and a passage in question answering. Finally, for classification or sequence tagging, `A` would be a sequence of text and `B` would be empty.

2.7 NLP Research with Clinical Notes

As mentioned in Section 1.2, there are many examples of usage of clinical notes in NLP research.

In the study "Query-Focused EHR Summarization to Aid Imaging Diagnosis" (McInerney et al., 2020), the authors aimed to create a system to conditionally extract summaries of clinical notes to serve as evidence to support different diagnosis given by radiologists (Figure 2.11). With this system, the radiologist is able to provide a potential diagnosis and receive snippets of text that are relevant to the query. This enables an efficient use of the unstructured patient history, as due to the volume of information it can be nearly impossible to thoroughly consult in order to make a timely diagnosis.

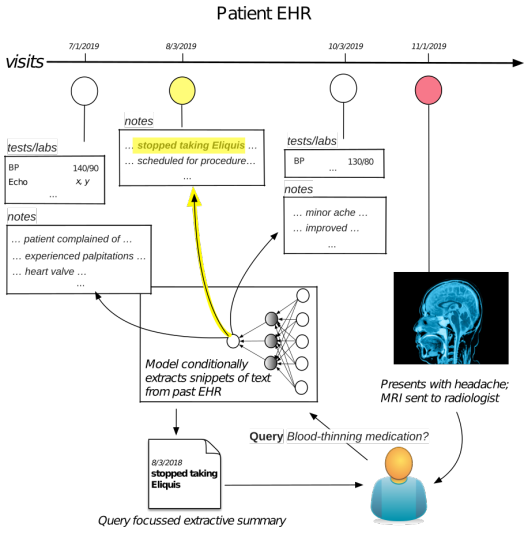


FIGURE 2.11: Overview of the model for generating extractive summaries from patient clinical notes in order to aid radiologist diagnoses (McInerney et al., 2020).

The study used EHR data from patients from Brigham and Women's Hospital (BWH) who had undergone magnetic resonance imaging of the brain, and also performed experiments using data from MIMIC-III (Johnson et al., 2016). Due to the lack of annotated data, the authors used International Classification of Diseases (ICD) codes present in the patient's clinical notes as proxies for diagnostic labels. The codes were organised in a hierarchy in order to relate codes that are similar and train models that are capable of giving both general and specific diagnosis. After data processing, the examples consist of $\{x, q, y\}$ triples where x is a list of sentences, q a list of ICD codes, and y is a list of labels.

The model takes as input clinical notes in the form of sentences x and a query q, then it should produce a summary of x that are relevant to q. This requires producing a distribution of relevance scores over x. Three models were compared, two unsupervised baseline models and a supervised model. The first baseline model encodes all sentences from a patient's clinical notes along with the query into Term Frequency-Inverse Document Frequency (TF-IDF) Bag-of-Words vectors, and uses the cosine similarity as a relevance score. The second model, similar to the first, also uses the cosine similarity. However, instead of TF-IDF Bag-of-Words vectors,

it uses contextual embeddings produced by a pretrained BERT model (Huang, Altosaar, and Ranganath, 2020). The third model is the only one that is actually trained, and the process consists of fine-tuning a BERT model for classification, as described in Section 2.6.

To evaluate the summaries produced by the models, the authors performed an assessment by domain experts. An interface was developed to allow radiologists to annotate snippets within notes with respect to their relevance to clinical queries. The authors found a 10-fold improvement in precision of the summaries over all the patient notes. Models trained on MIMIC-III fared well when evaluated on the BWH dataset, suggesting that models trained on one EHR may be deployed in a different setting. The study did not evaluate the model when used in practice, nevertheless the authors propose this as future work.

2.8 Summary

The shift from manual feature engineering to automatic feature learning using Deep learning has contributed to the improvement of state-of-the art ML models in a variety of fields, including NLP. Deep learning NLP models are often pre-trained on large amounts of unlabelled data and then fine-tuned on task-specific labelled data.

Pre-training results in models that are able to understand the syntactic and semantic properties of the training text. However, simpler models that do not incorporate such properties, like DANs can outperform more expensive models on some tasks while requiring only a fraction of the resources.

State-of-the-art NLP models were commonly based on RNNs using an encoder-decoder structure, where one encodes a sequence of text into a vector representation and the other decodes the representation into a different sequence of text. The prevalent problem with RNNs is the vanishing gradients during training, which limits their learning capabilities, LSTMs managed to solve it to some extent.

The Transformer improves on the shortcomings of previous RNN encode-decoder NLP models by removing recurrence and relying solely on an attention mechanism. BERT, based on the Transformer architecture, was designed to be pre-trained on unlabelled text and fine-tuned on specific tasks without significant alterations.

The per-layer attention complexity of the attention mechanism prevents the Transformer from working with longer sequences of text. In order to overcome this, the Longformer is proposed as a drop-in replacement for the standard attention. By employing a fixed-size window in the sequence element neighbourhood, the computational complexity is reduced from quadratic to linear, thus becoming feasible to work with longer sequences.

Chapter 3

Methodology

This chapter describes the methods employed to evaluate the performance of a Transformer-based model with a more efficient attention mechanism that enables processing of long text sequences. The dataset and it's processing is described. Implementation details of models used, including their configuration and training procedure, are also given. For Transformer-based models, there is a section on pre-training. Finally, details on the hyperparameter tuning procedure and the evaluation metrics are presented.

3.1 Dataset

3.1.1 Medical Information Mart for Intensive Care (MIMIC-III)

MIMIC-III integrates de-identified, comprehensive clinical data of patients admitted to critical care units.

The data is composed of vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. It contains data associated with 53,423 distinct hospital admissions for adult patients (aged 16 years or above) admitted to critical care units between 2001 and 2012.

The dataset is accessible to researchers internationally under a data use agreement, its open nature allows clinical studies to be reproduced and improved in ways that would not otherwise be possible (Johnson et al., 2016).

3.1.2 Data processing

MIMIC-III (Johnson et al., 2016) is a relational database, tables are linked by identifiers which usually have the suffix `ID`. Admissions are stored in the `ADMISSIONS` table, which contain the `ADMITTIME` and `DISCHTIME` columns for the date and time the patient was admitted and discharged from the hospital, respectively. For example, `HADM_ID` refers to a unique hospital admission and `SUBJECT_ID` refers to a unique patient. Notes are stored in the `NOTEEVENTS` table, which is linked to the admissions via the `HADM_ID` identifier.

Pre-processing of MIMIC-III clinical notes for fine-tuning is based on ClinicalBERT (Huang, Altosaar, and Ranganath, 2020). First all admissions are associated with the next unplanned admission belonging to the same patient, the ones that fall within the 30-day window and the remaining are labelled with the positive or negative class — readmitted or not readmitted — respectively. Then newborn admissions and those resulting in death are removed from the cohort. The next step is to associate admissions with notes. In the notes, words are converted to lowercase and line breaks and carriage returns removed. De-identified brackets and special characters like ==, - are also be removed.

Table 3.1 shows some clinical note's statistics after pre-processing. For each type there is the number of notes, the percentage of admissions with this type of note, length in number of tokens and average length.

TABLE 3.1: Dataset statistics after pre-processing. PERCENT is the percentage of admissions with this type of note. LEN is total length in tokens and AVG the average length.

TYPE	COUNT	PERCENT	LEN	AVG
Nursing/other	348,592	52.8%	53,697,159	154
Radiology	307,945	82.9%	57,903,879	188
Nursing	185,878	18.0%	47,497,919	256
ECG	119,904	85.4%	3,541,882	30
Physician	115,640	17.9%	96,071,741	831
Discharge summary	50,072	96.8%	72,973,487	1,457
Echo	28,458	44.2%	9,413,880	331
Respiratory	24,432	7.4%	3,696,010	151
Nutrition	7,419	5.9%	2,415,596	326
General	5,938	5.6%	1,340,358	226
Rehab Services	4,926	4.5%	2,092,918	425
Social Work	1,969	2.4%	587,892	299
Case Management	825	1.2%	111,723	135
Pharmacy	84	0.1%	25,215	300
Consult	68	0.1%	50,122	737

The previous pre-processing step results in examples that consist of triples of admission identifier, note text and readmission label $\{\texttt{HADM\_ID}, (x_0, ..., x_n), label\}$.

Different versions of the dataset are used to train the models, these contain essentially the same examples as they refer to the same set of admissions, however the difference is in the length of the textual data. For example, a particular entry in the dataset $\{\texttt{HADM\_ID}_1, (x_0, ..., x_n), 0\}$ may have its note text divided thus being substituted with other entries each with a part of the full text while maintaining the admission identifier and readmission label. Depending on the maximum length (l) of the text permitted for the dataset, $\{\texttt{HADM\_ID}_1, (x_0, ..., x_n), 0\}$ may be substituted by $\{\texttt{HADM\_ID}_1, (x_0, ..., x_l), 0\}$ and $\{\texttt{HADM\_ID}_1, (x_{l+1}, ..., x_n), 0\}$ if $l < n$. The resulting model predictions are later recombined as described in Section 3.2. The five versions of the datasets by text length are full length, 512, 1,104, 2,048 and 4,096. The resulting examples maintain the original admission identifier and label. At this step we have five different datasets,

the original full-length in addition to the four resulting from division, each one of these datasets are further divided into training, validation and tests sets. For training, 80% of examples are chosen at random, 10% for validation and the remaining 10% are used for the test set.

Discharge summaries were found to be especially predictive for readmission as they aggregate information from the entire admission (Hsu et al., 2020), these types of note will be exclusively of interest for research question 1. Figure 3.1 illustrates the distribution of discharge summary lengths, most have between 1000 and 2000 tokens. For research question 2, with the goal being to produce a model that can be more applicable in clinical settings by making accurate predictions early in patient treatment and with discharge summaries often available only after a patient has been discharged, all notes available before discharge should be processed. Clinical notes created up to a week of a patient's patient's admission are used.

The pre-training dataset is built using physician and nursing notes, as per Huang, Altosaar, and Ranganath (Huang, Altosaar, and Ranganath, 2020). Text pre-processing step similar to the fine-tuning datasets pre-processing, first words are lower cased, then de-identified brackets and special characters are removed. The resulting file is around 161 million tokens long and occupies 941 MB of storage.

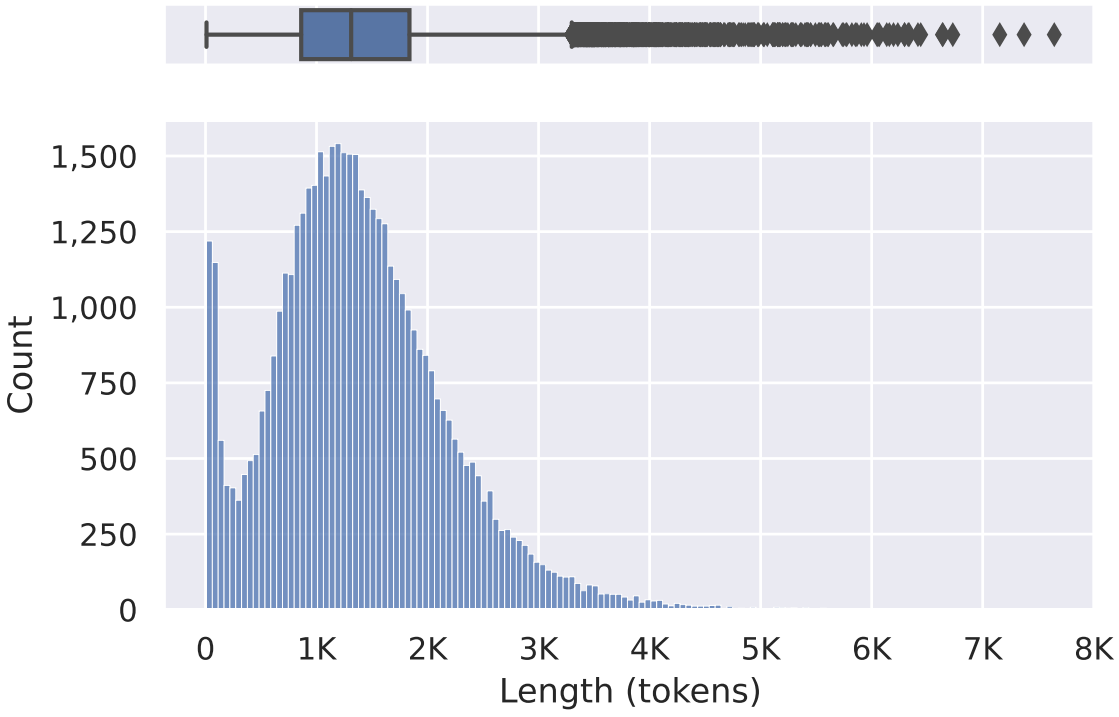


FIGURE 3.1: Distribution of Discharge summary lengths. Average: 1,457 tokens. Median: 1,377 tokens.

3.2 Experimental setup

The experimental setup consists of two baseline models – a DAN and a LSTM – a standard attention Transformer-based model and three Longformer attention models. The models go through hyperparameter search before training, and their performance is then evaluated using appropriate metrics. The hyperparameter search method is called Bayesian optimization, and it is further expanded in Chapter 3.8. Transformer-based models are first pre-trained on a clinical domain dataset and then fine-tuned for the 30-day hospital readmission task that can be interpreted as binary classification. In all models, the final classification layer uses a sigmoid activation function, which outputs a value between zero and one that we interpret as being the probability of a patient being readmitted.

The models are trained on the dataset with the maximum length they are able to process, DAN and LSTM are trained on the full length, the standard attention model on the 512 dataset, and the Longformer models on 1,024, 2,048 and 4,096 respectively. For examples that are divided into many (n), this method results in as many predictions for the same patient, in this case the predictions are combined to form the final prediction using the same formula as Huang, Altosaar, and Ranganath, 2020

$$P(readmit = 1|p_{id}) = \frac{p^n_{max} + p^n_{mean}{}^{n/c}}{1 + {}^{n}/_{c}}, \tag{3.1}$$

where p^n_{max} and p^n_{mean} are the maximum and mean probabilities of readmission over n subsequences and c a scaling factor. The threshold value of 0.5 is used to decide if the final classification is positive or negative. Source code for data processing and model training is made available via online repository [1], and datasets are available by request through the MIMIC-III website [2].

3.3 DAN

The Deep averaging network (DAN), described in Section 2.2, is a Bag-of-Words like model that applies a composition function to a sequence of input embeddings before feeding it to a logistic regression function. There is no limitation on the maximum sequence length it can process. In this work the implementation is based on Iyyer et al., 2015 and it uses GloVe (Pennington, Socher, and Manning, 2014) pre-trained word embeddings which have been trained on Wikipedia and Newswire data. GloVe embeddings are available in four number of dimensions, 50, 100, 200 and 300.

In order to improve model robustness, a regularization technique called word dropout is employed. It consists of randomly removing embeddings from the average computation with a given probability, similar to dropout (Hinton et al., 2012), which removes input or hidden units instead.

[1] https://github.com/ynurmahomed/clinical-longformer
[2] https://physionet.org/content/mimiciii/1.4/

The model uses a sigmoid activation function in the final layer and is optimized using the Adagrad (Duchi, Hazan, and Singer, 2011) algorithm with binary cross entropy loss as the objective function.

Hyperparameters used for tuning are learning rate, batch size, embedding dimension, number of hidden layer, word dropout rate and weight decay. Table 3.2 summarizes the hyperparameter value ranges used.

TABLE 3.2: Hyperparameter values used for tuning DAN model.

Hyperparameter	Searched values
Learning rate	$1 \cdot 10^{-4}$ to $5 \cdot 10^{-2}$
Batch size	32, 64, 128
Embedding dimension	50, 100, 200, 300
Hidden layers	1 to 3
Word dropout	$1 \cdot 10^{-1}$ to $5 \cdot 10^{-1}$
Weight decay	$1 \cdot 10^{-2}$ to $5 \cdot 10^{-2}$

3.4 LSTM

TABLE 3.3: Hyperparameter values used for tuning the LSTM model.

Hyperparameter	Searched values
Learning rate	$1 \cdot 10^{-5}$ to $1 \cdot 10^{-3}$
Batch size	32, 64, 128
Embedding dimension	50, 100, 200, 300
Hidden layer dimension	100, 200, 300
Dropout	$1 \cdot 10^{-1}$ to $5 \cdot 10^{-1}$

The LSTM (Section 2.3.1) is an RNN architecture where each block is composed of memory cells and gate units.

The model implemented in this work is a recurrent neural network based on Lai et al., 2015. Similar to DAN, it processes GloVe embeddings as inputs and there is no limit on the maximum sequence length that can be given. The model consists of a bidirectional RNN with LSTM cells followed by a max pooling layer before a final two layer feed-forward network with an ReLU activation function in between.

In the final layer it uses a sigmoid activation function and is optimized using the Adam (Kingma and Ba, 2015) algorithm with binary cross entropy loss as the objective function.

Hyperparameters used for tuning are learning rate, batch size, embedding dimension, hidden layer dimension and dropout. These are show in Table 3.3.

3.5 ClinicalBERT

ClinicalBERT (Huang, Altosaar, and Ranganath, 2020) is a BERT (Section 2.6) model that is initialized from the $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) checkpoint and further pre-trained on MIMIC-III (Johnson et al., 2016) clinical text.

Like other standard attention models, ClinicalBERT has fixed maximum input sequence length. To deal with this limitation the input is split to the maximum length, then the predictions are computed by aggregating the predictions from each sub-sequence like described in Section 3.2.

The model can be downloaded and fine-tuned for any task-specific application. However, as we have no control on the data used for pre-training and in order to prevent that the model is evaluated on text that it has already processed, we did not use the existing ClinicalBERT. In this work we replicated the methodology of the original by further pre-training $\text{BERT}_{\text{BASE}}$ on our version of MIMIC-III and then fine-tuning and evaluating, this guarantees that data leakage will not occur.

The classification layer is a three layer neural network with ReLU activation functions in the first two layers and a sigmoid activation in the final layer. This layer receives as input the special `[CLS]` token that is used as a sequence representation in the final Transformer layer.

Optimization is done using the Adam with decoupled weight decay (AdamW) (Loshchilov and Hutter, 2019) algorithm, it uses a linear learning rate scheduler with 10% warmup.

Hyperparameters used for tuning are learning rate, batch size and dropout as shown in Table 3.4.

TABLE 3.4: Hyperparameter values used for tuning the BERT model.

Hyperparameter	Searched values
Learning rate	$2 \cdot 10^{-5}$, $3 \cdot 10^{-5}$, $5 \cdot 10^{-5}$
Batch size	32, 48, 56, 64
Dropout	$1 \cdot 10^{-1}$ to $3 \cdot 10^{-1}$

3.6 Longformer

The Longformer (Beltagy, Peters, and Cohan, 2020), described in Section 2.5, has been pre-trained starting from the RoBERTa (Liu et al., 2019) checkpoint, which is a replication study of BERT (Devlin et al., 2019) pretraining that measures the impact of many key hyperparameters and training data size. With RoBERTa, Liu et al. found that BERT was significantly under trained, and were able to match or exceed the performance of every model published after it. It essentially removes the next-sentence pretraining objective and trains the model with larger mini-batches and learning rates.

In order to make a fair comparison to ClinicalBERT, we replicated the source-code for the implementation of BERT, the only difference being that for the attention module it uses the Longformer version. This removes any improvements to the model's performance that might be

due to RoBERTa's improved pre-training. The Huggingface's Transformers (Wolf et al., 2019) library, used in this work, provides control over specific models by allowing that custom models be built from a few base classes.

This custom Longformer model is pre-trained from the $\text{BERT}_{\text{BASE}}$ checkpoint. Similar to Beltagy, Peters, and Cohan, 2020, the initial checkpoint is converted to the long model before pre-training. The conversion process consists of allocating larger weight matrices in accordance to the desired maximum sequence length and copying the values from the staring checkpoint to these newly created matrices. Three models with differing maximum sequence length – 1,024, 2,048 and 4,096 – are converted from $\text{BERT}_{\text{BASE}}$. The attention window hyperparameter defines how many tokens in its neighbourhood each token should attend to, and it should be less than the maximum sequence length.

Similar to ClinicalBERT (Chapter 3.5), the classification layer is a three layer neural network with ReLU activations and a sigmoid in the final layer and optimization is done using AdamW with a linear learning rate scheduler with 10% warmup.

Hyperparameters used for tuning are learning rate, batch size, attention window and dropout as shown in Table 3.5.

TABLE 3.5: Hyperparameter values used for tuning the Longformer model.

Hyperparameter	Searched values
Learning rate	$2 \cdot 10^{-5}$, $3 \cdot 10^{-5}$, $5 \cdot 10^{-5}$
Batch size	32, 48, 56, 64
Attention window	512, 1,024, 2,048, 4,096
Dropout	$1 \cdot 10^{-1}$ to $3 \cdot 10^{-1}$

3.7 Model Pre-training

Pre-training for the ClinicalBERT and Longformer models is similar to BERT. The models are optimized on the masked language modelling and next sentence prediction objectives using the pre-training dataset obtained from processing MIMIC-III clinical notes (Chapter 3.1.2).

The model is trained for a total of 100,000 steps. First the models are pre-trained for 90,000 steps on a maximum sequence length of 128 tokens and then for an additional 10,000 steps with the largest sequence length supported [3].

In terms of hyperparameters, the batch-size is set to 64 and the learning rate is set to $2 \cdot 10^{-}5$. The AdamW (Loshchilov and Hutter, 2019) algorithm is used for optimization with $\beta_1 = 0.9$, $\beta_2 = 0.999$.

[3]This is the procedure recommended by BERT's authors in https://github.com/google-research/bert#pre-training-tips-and-caveats.

3.8 Hyperparameter Tuning

Hyperparameter tuning is the process by which a machine learning model's hyperparameters are chosen. A model's performance depends not only on its architecture, but also on the hyperparameter values. Standard models, when properly regularised, outperform more recent models (Melis, Dyer, and Blunsom, 2018).

One method of doing hyperparameter tuning is called grid search. It works by trying every possible combination of discrete hyperparameter values. This is a simple method, however as the number of hyperparameters increases it becomes more costly to try every possible combination.

Random search is another method, it replaces exhaustive search over every possible combination by trying the values randomly. This method generalizes to continuous values and can be less costly than grid search because not all hyperparameters contribute in the same proportion to the model's performance, it is not optimal to repeat combinations of values that contribute less.

Both of the previous methods ignore past results, therefore they fail to notice trends in hyperparameter values that guide to better models. A much more effective method would be to do an informed search by looking at the results obtained so far and try to predict what combinations are more likely to lead to better model performance.

This is possible to do by means of Bayesian optimization with a Gaussian process prior, an effective method for determining the maxima and minima of expensive to compute objective functions (Brochu, Cora, and Freitas, 2010), such as training a machine learning model. It uses the Bayes theorem, which in simplified terms states that the posterior probability of a model M given evidence E is proportional to the likelihood of E given M multiplied by the prior probability of M:

$$P(M|E) \propto P(E|M)P(M). \tag{3.2}$$

We can define x_i as the ith sample, in this case a combination of hyperparameters, and $f(x_i)$ as the observation of the objective function at x_i or the result of training the model with that combination of hyperparameters. In practice, we collect samples and observations $D_{1...t} = \{x_{1...t}, f(x_{1...t})\}$ and combine with the prior with the likelihood function $P(D_{1...t}|f)$ to obtain the posterior:

$$P(f|D_{1...t}) \propto P(D_{1...t}|f)P(f). \tag{3.3}$$

The goal is to maximize the model's output using evidence and prior knowledge at each stage. There are many forms of modelling priors that can be used, a Gaussian Process one of them (Brochu, Cora, and Freitas, 2010). A Gaussian process is a generalization of the Gaussian probability distribution, in essence it assumes that similar inputs give similar outputs, and are able to

predict the expected outcome of a particular hyperparameter combination (Snoek, Larochelle, and Adams, 2012). In addition, for each hyperparameter they learn the appropriate scale for measuring the similarity between two values, making it possible to evaluate if one particular value tends to give results that are very similar or dissimilar to another value.

By using a Gaussian process model for hyperparameter tuning, we are able to make an informed decision on which values to try next, which can lead to more accurate models.

In this work, an open source tool for hyperparameter optimisation based on Gaussian process models is used (Biewald, 2020) [4].

Resulting hyperparameters are shown in Tables 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8.

3.9 Evaluation Metrics

A convenient way of summarizing the performance of classifiers is to create a confusion matrix by doing a cross-tabulation between actual and predicted results for each class. An example of a confusion matrix for a binary classification problem is shown in Table 3.6.

TABLE 3.6: Confusion matrix of a binary classification problem.

		Predicted class: Positive	Predicted class: Negative	
Actual class	Positive	TP	FN	POS
	Negative	FP	TN	NEG
		PPOS	PNEG	

From the confusion matrix, different evaluation metrics can be derived. Some of this metrics are the True Positive Rate (TPR) or Recall

$$TPR = \frac{TP}{POS} \tag{3.4}$$

False Positive Rate (FPR)

$$FPR = \frac{FP}{NEG} \tag{3.5}$$

and Precison

$$Precison = \frac{TP}{PPOS}. \tag{3.6}$$

When the output of the classifier is a numeric score instead of positive of negative predictions, we may choose metrics that would allow us to evaluate the performance without committing to a specific threshold for deciding between positive or negative classes. Each of the models is evaluated using the following metrics:

- Area under the receiver operating characteristic curve (AUC-ROC):

  obtained by calculating the area under the plot of the True Positive Rate versus the False Positive Rate at different prediction thresholds. The closer the curve is to the top left of

[4] https://docs.wandb.ai/guides/sweeps

the plot, the better the classifier, as shown in Figure 3.2. The metric can be interpreted as the probability that the scores given by a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Fernández et al., 2018).

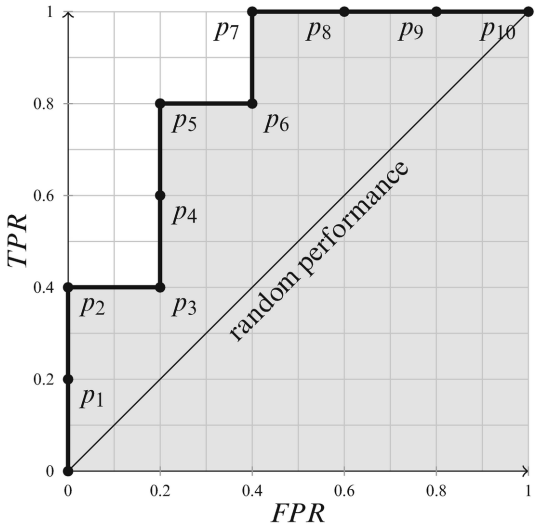


FIGURE 3.2: Illustration and comparison of the ROC curve for different classifiers, the upwards represents a classifier with random performance. (Fernández et al., 2018).

- Area under the precision-recall curve (AUC-PR):

  obtained by calculating the area under the plot of the precision versus the recall at different prediction thresholds. The closer the curve is to the top left of the plot, the better the classifier, as shown in Figure 3.3. A random classifier will be a horizontal line with a precision that is proportional to the number of positive examples in the dataset. For a balanced dataset, this will be 0.5.

- Recall at precision of 70% (RP70)

  The metric is obtained by selecting fixing the precision at 70% and using this threshold to calculate the recall. The goal is to produce models that minimize the false positive rate, in turn reducing the effect of alarm fatigue (Sendelbach and Funk, 2013).

The area under the curve (AUC) summarizes the ROC and PR curves into a single value that can be compared between different classifiers. One way of estimating this area is by using the trapezoidal rule, linear interpolation is used to connect points from the curve and then the AUC is computed by summing trapezoidal areas created between each point. While this method works for ROC curves, in the case of PR curves linear interpolation is a mistake that yields an overly-optimistic estimate of performance (Davis and Goadrich, 2006). A better estimate for AUC-PR, which avoids interpolation, is the average precision.

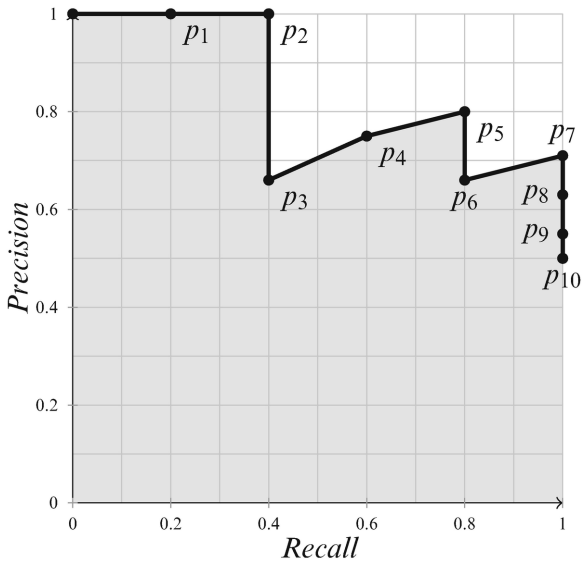


FIGURE 3.3: Illustration and comparison of the PR curve (Fernández et al., 2018).

3.10 Summary

In this work we implement, train and evaluate two baseline models, DAN and LSTM, a standard attention Transformer-based model and three Longformer-attention models.

The dataset consists of clinical notes from MIMIC-III, a relational database that integrates de-identified, comprehensive clinical data of patients admitted to critical care units. For observing the effect of using larger attention mechanisms, we focus on discharge summaries. For the early readmission task, all types of notes are processed. Different views of the dataset are created according to note length, full length, 1K, 2K and 4K. Where the note exceeds the maximum length, the example is split into multiple examples and the models combines the predictions on each example to form the final prediction. The datasets are each divided into 80% for training, 10% for validation and the final 10% for testing.

The models go through tuning on the validation dataset to find the optimal combination of hyperparameters before being evaluated on the test dataset using AUC-ROC, AUC-PR and RP70.

Chapter 4

Results and Discussion

This chapter presents the results obtained from hyperparameter tuning and from training and evaluating the models.

In this chapter the different Longformer models according to maximum length 1024, 2048 and 4096 are referred to as Longformer 1K, 2K and 4K respectively.

First, the results for each research question are presented comparably. Then the hyperparameter configuration for each model is given. Finally, a there is a brief discussion of the results.

4.1 Research Question 1

Will using longer text sequences (1K, 2K and 4K tokens) with the Longformer model result in improved 30-day readmission prediction than the baseline models or ClinicalBERT?

Table 4.1 summarizes the results obtained by training and evaluating the models on discharge summaries. The baseline LSTM model achieves the highest scores in terms of AUC-ROC, ROC-PR and RP70 with 0.65, 0.63 and 0.3 respectively. All the Longformer models have significantly better performance in terms of RP70 than ClinicalBERT, with Longformer 2K obtaining the best score between them of 0.30. ClinicalBERT does slightly worse than all the Longformer models in terms of AUC-PR, however it is better than Longformer 1K and Longformer 2K in terms of AUC-ROC. The baseline DAN model had the lowest scores of all the models.

TABLE 4.1: Results for the 30-day readmission prediction task with discharge summaries.

Model	AUC-ROC	AUC-PR	RP70
DAN	0.58	0.54	0
LSTM	**0.65**	**0.63**	**0.30**
ClinicalBERT	0.63	0.61	0.06
Longformer 1K	0.62	0.62	0.25
Longformer 2K	0.62	0.62	0.30
Longformer 4K	0.64	0.62	0.15

4.2 Research Question 2

How many days after the first admission can we predict readmission using all the notes available, without a drop in performance of more than 10%, compared to using only discharge summaries?

Results for the evaluation of Longformer 4K and ClinicalBERT on early clinical notes are presented in Table 4.2. For both Longformer 4K and ClinicalBERT the AUC-PR scores increase from the first day in ICU up to the fourth day, for the subsequent days they exhibit a slight decrease in performance. Overall ClinicalBERT has the better performance of the two over the period of seven days with the highest difference in AUC-PR score occurring on the fifth day in ICU.

TABLE 4.2: Comparison of results in AUC-PR from Longformer 4K using early notes and using discharge summaries for the 30-day readmission prediction task.

Days in ICU	Longformer 4K	Longformer 4K Discharge Summary	Difference	LSTM
1	0.58	0.62	**0.04**	0.58
2	0.59	0.62	0.03	0.58
3	0.59	0.62	0.03	0.59
4	**0.61**	0.62	0.01	0.59
5	0.59	0.62	0.03	**0.63**
6	0.59	0.62	0.03	0.60
7	0.60	0.62	0.02	0.63

4.3 Hyperparameter tuning

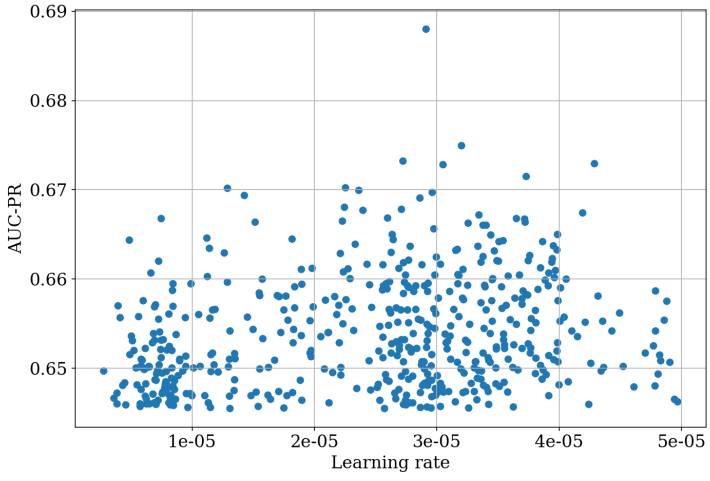


FIGURE 4.1: Relationship between the learning rate hyperparameter and the AUC-PR metric for the Longformer 1K model. The learning rate that resulted in the best performing model is close to $3 \cdot 10^{-5}$.

The hyperparameter tuning process is based on a Gaussian process model. The scatter plot illustrated in Figure 4.1 represent evaluations of different configurations of Longformer 1K on the evaluation dataset. The two top performing runs, in terms of AUC-PR, seem to centre

around the $3 \cdot 10^{-5}$ mark. This value was chosen as the learning rate for the 2K and 4K versions of the model.

4.3.1 DAN

Table 4.3 shows the hyperparameter values which resulted in the best performance on the validation dataset.

TABLE 4.3: Hyperparameter values used for training DAN model.

Hyperparameter	Value
Learning rate	$8 \cdot 10^{-4}$
Batch size	64
Embedding dimension	300
Hidden layers	2
Word dropout	$7 \cdot 10^{-1}$
Weight decay	$2 \cdot 10^{-2}$

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the DAN model. The model obtained 0.58 on AUC-ROC, 0.54 on AUC-PR and a score of 0 on RP70.

4.3.2 LSTM

Table 4.4 shows the hyperparameter values which resulted in the best performance on the validation dataset.

TABLE 4.4: Hyperparameter values used for training LSTM model.

Hyperparameter	Value
Learning rate	$8 \cdot 10^{-4}$
Batch size	64
Embedding dimension	300
Hidden dimension	200
Dropout	$5 \cdot 10^{-1}$

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the LSTM model. The model obtained 0.65 on AUC-ROC, 0.63 on AUC-PR and a score of 0.3 on RP70.

4.3.3 ClinicalBERT

Table 4.5 shows the hyperparameter values which resulted in the best performance on the validation dataset.

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the ClinicalBERT model. The model obtained 0.63 on AUC-ROC, 0.61 on AUC-PR and a score of 0.06 on RP70.

TABLE 4.5: Hyperparameter values used for training ClinicalBERT model.

Hyperparameter	Value
Learning rate	$2 \cdot 10^{-5}$
Batch size	32
Dropout	$2 \cdot 10^{-1}$

4.3.4 Longformer 1K

Table 4.6 shows the hyperparameter values which resulted in the best performance on the validation dataset.

TABLE 4.6: Hyperparameter values used for training Longformer 1K model.

Hyperparameter	Value
Learning rate	$3 \cdot 10^{-5}$
Batch size	32
Attention window	1,024
Dropout	$2 \cdot 10^{-1}$

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the Longformer 1K model. The Longformer 1K model obtained 0.62 on AUC-ROC, 0.62 on AUC-PR and a score of 0.25 on RP70.

4.3.5 Longformer 2K

Table 4.7 shows the hyperparameter values which resulted in the best performance on the validation dataset.

TABLE 4.7: Hyperparameter values used for training Longformer 2K model.

Hyperparameter	Value
Learning rate	$3 \cdot 10^{-5}$
Batch size	32
Attention window	1,024
Dropout	$2 \cdot 10^{-1}$

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the Longformer 2K model. The model obtained 0.62 on AUC-ROC, 0.62 on AUC-PR and a score of 0.3 on RP70.

4.3.6 Longformer 4K

Table 4.8 shows the hyperparameter values which resulted in the best performance on the validation dataset.

Figure 4.2 illustrates the precision-recall curve and Figure 4.3 the ROC curve for the Longformer 4K model. The model obtained 0.64 on AUC-ROC, 0.62 on AUC-PR and a score of 0.15 on RP70.

TABLE 4.8: Hyperparameter values used for training Longformer 4K model.

Hyperparameter	Value
Learning rate	$3 \cdot 10^{-5}$
Batch size	32
Attention window	1,024
Dropout	$2 \cdot 10^{-1}$

4.4 ROC and PR curves

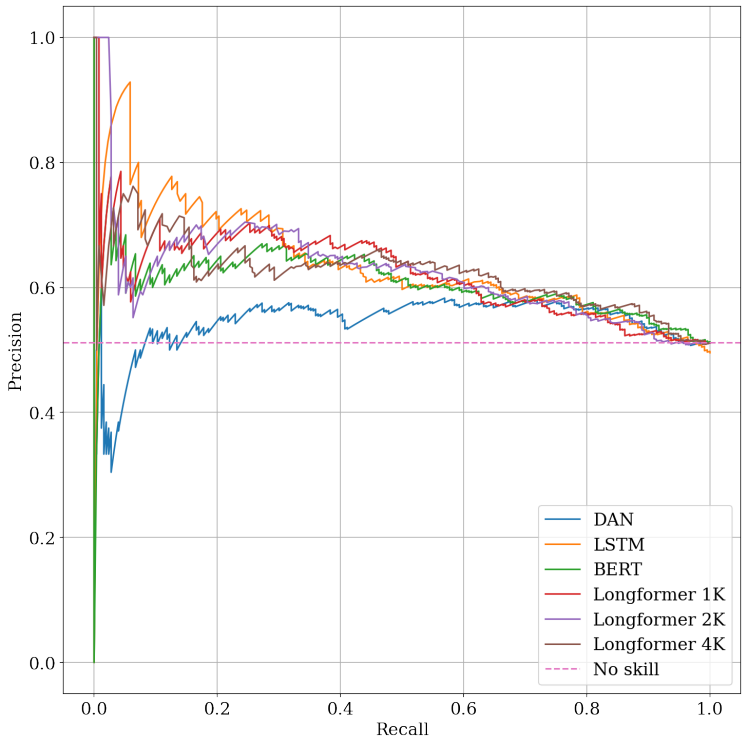


FIGURE 4.2: PR curves the 30-day readmission prediction task with discharge summaries. LSTM has the highest AUC-PR with 0.63 while DAN has the lowest with 0.54. The green vertical line is an inconsistency with BERT evaluation that is present as well in DAN and LSTM.

Figures 4.2 and 4.3 show the precision-recall curves and ROC curves for the different models.

The pink dashed line represents a classifier with no discriminative power between the classes. At recall values below 0.1 the DAN model has a precision below the no skill line, while the other models are above it.

As the recall gets closer to 0 the precision seems to not be well-defined for the LSTM, BERT and DAN models. While for the Longformer models this is not observed. The inconsistency manifests as a vertical line at recall 0, where it should be a point at precision 1.

In terms of ROC curves, similar to the precison-recall curves, the DAN model exhibits a true positive rate below the no skill line at false positive rates closer to 0.

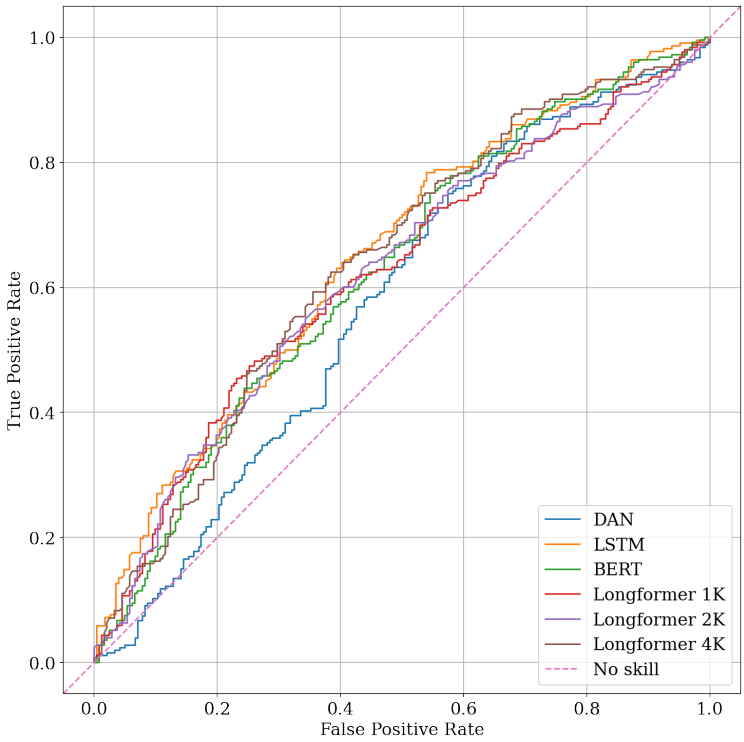


FIGURE 4.3: ROC Curves the 30-day readmission prediction task with discharge summaries. LSTM has the highest AUC-ROC with 0.65 while DAN has the lowest with 0.58.

4.5 Discussion

4.5.1 Research Question 1

Using longer text sequences with the Longformer model does result in improved 30-day readmission prediction, as can be seen in Table 4.1 where the results are better in two of the three chosen metrics.

The Longformer attention model with 4K maximum length was able to produce better results than the standard attention Transformer-based model (ClinicalBERT) in terms of AUC-ROC and AUC-PR. In terms of RP70, all the Longformer attention models significantly outperform the standard attention model, at most by a margin of 24%. ClinicalBERT has the lowest RP70 values, as illustrated in Figure 4.4 at a precision of 70% the recall is much lower than 20%.

This means that compared to the others, these models can identify a larger fraction of patients from the overall pool of those that are readmitted at the expense of miss-classifying at most 30%.

Having better recall makes them more useful in a clinical setting as it is important that all possible readmissions are identified, even if some turn out to be wrongly classified. Failing to properly identify a readmission can result in negative outcomes to patients.

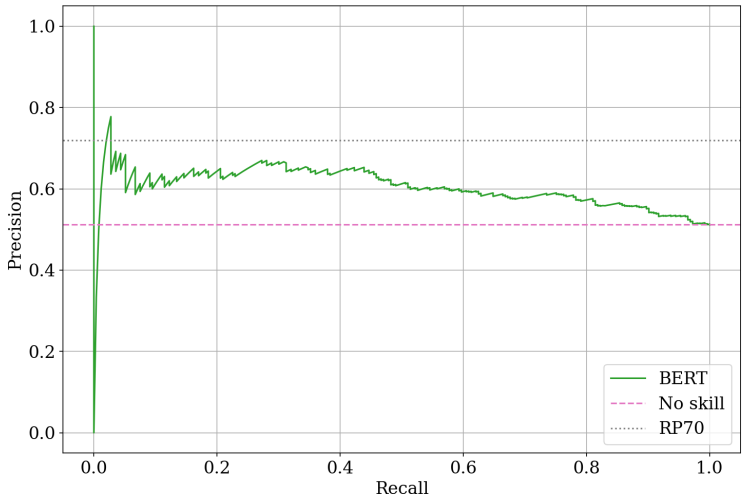


FIGURE 4.4: PR Curve for the ClinicalBERT model. The recall value at 70% precision, much smaller than 2% (0.06).

It seems there is a sublinear relationship between increasing the supported input length and the performance on the 30-day readmission prediction task.

The results indicate that the sparse attention employed by the Longformer (Section 2.5) model improve classification, and this is consistent with results from other studies (Beltagy, Peters, and Cohan, 2020; Li et al., 2022).

4.5.2 Research Question 2

It was expected that at the beginning of the stay, the difference in performance between the models compared would be so much that it would require increasing the amount of data to process. This conjecture turned out not to be true, thus we cannot conclusively answer the research question.

The results in Table 4.2 show that during the 1st week of a patient's stay, the model trained using early notes does not have a difference in performance of more than 10% compared to using exclusively discharge summaries. Furthermore, it can make the best predictions at the 4th day. The highest difference in performance is recorder on the 1st day of stay, this is expected as there would not be much information available to make accurate predictions at that time, compared to the information contained on a discharge summary which is produced at the end of the stay. Results for the LSTM model trained on early notes are added for comparison.

This research question focuses on the applicability of the model developed in this work. Research question 1 relies on discharge summaries, which are only available after a discharge from the hospital.

In a real world setting, a model is more useful if it is able to determine the risk of a patient while undergoing treatment. With this in mind, the model was trained and evaluated on data that should be available while a patient is yet to have been discharged.

4.5.3 Training data

The Longformer's 4K loss curves, illustrated in Figure 4.5 suggests that the training data might not be appropriate for the task.

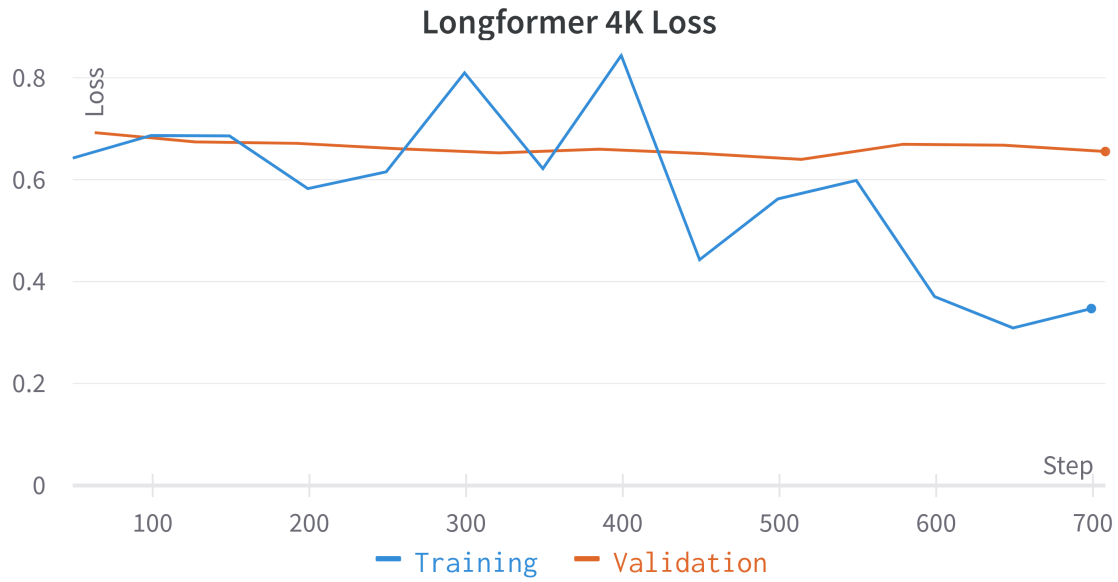


FIGURE 4.5: Comparison of loss curves for training and validation of Longformer 4K. The training curve shows improvement, nevertheless a large difference is found in comparison with the training curve.

BERT like models have achieved good results by leveraging pre-training on large general domain text datasets. Models like ClinicalBERT (Huang, Altosaar, and Ranganath, 2020), BioBERT (Lee et al., 2020) and BlueBERT (Peng, Yan, and Lu, 2019) are initialized with the original BERT model and then continue pre-training on domain-specific text. This process assumes that using more text will always be beneficial, even if it is out of the final application domain.

Gu et al., 2020 found that in the biomedical domain, this process, which they call mixed-domain pre-training, is inferior due to the original BERT's vocabulary not being representative of the target domain. To generate the vocabulary, BERT uses Byte-Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016), which encodes rare or unknown words as sequences of subword units. Having a vocabulary with missing biomedical terms causes the derived models to have to learn these terms using fragmented words. Gu et al., 2020 pre-trained a BERT like model from scratch using domain-specific text and obtained new state-of-the-art results for a wide range of biomedical NLP applications. Their work also led to the creation of a new benchmark for biomedical NLP.

The situation described in the previous paragraph is equally observed with the models produced in this work. The following example is a discharge summary that has been tokenized using the original BERT's tokenizer, in it many medical concepts have been fragmented into subwords:

```
[CLS] date of birth : sex : f service : medicine all ##er ##gies : pl ##avi
##x / he ##par ##in agents attending : chief complaint : hem ##op ##ty ##sis
, fever , inability to vent ##ila ##te major surgical or invasive procedure
: tr ##ache ##ost ##omy revision arterial line central ve ##nous line
peripheral ##ly inserted central cat ##het ##er history of present illness
: 73 ##f h / o d ##m , multiple cv ##as , di ##sse ##minated tb on tx since
, tr ##ache ##d for res ##p failure with prolonged int ##uba ##tion /
failure to we ##an living at mac ##u / rehab , developed res ##p distress
```

Hemoptysis is defined as the spitting of blood that originated in the lungs or bronchial tubes, and it can be a crucial symptom to determine a patient's need for special treatment before being discharged. In the previous example, the word hemoptysis is divided into subwords `hem`, `##op`, `##ty` and `##sis`. If the model is not able to effectively learn this concept, it can significantly hinder its ability to make a correct prediction.

To alleviate this problem, the model would have to be trained from scratch with in domain text.

4.5.4 Interpretability

Deep learning models often produce predictions that are difficult to understand due to their black-box nature. Understanding the effect of a certain input to a program's output is crucial to our ability to improve such program.

Integrated Gradients (Sundararajan, Taly, and Yan, 2017) is a method for attributing the prediction of a deep neural network to its inputs, which requires no modification to the original network. Using this method the attributions for a few examples have been computed, these include somewhere the model outputs the correct prediction and others where it does not.

On most of the examples where the model is able to correctly predict the label (Figures A.1, A.2, A.3, A.4), it attributes its output mostly to the tokens are subwords of tracheostomy, which is an opening created at the front of the neck, so a tube can be inserted into the windpipe (trachea) to help a patient breath. In another example (Figure A.5) the model attributes the prediction to the tokens `es` and `##rd`, which are derived from the abbreviation *esrd* that most probably stands for End-stage renal disease, a disease where the patient needs to do dialysis or a kidney transplant. Here, the model is able to understand that the patient will need to return for to the hospital to have a dialysis procedure.

Where the model is not able to make correct predictions, the two worst performing examples are of patients with issues of mental health (Figures A.6 and A.7). These include anxiety and depression associated with alcohol and drug use, there are also reports of past suicide attempts. In these examples, the model outputs a probability value much lower than the threshold for a positive classification (Equation 3.1).

4.5.5 Other clinical BERT models

Huang, Altosaar, and Ranganath, 2020 pre-trained and fine-tuned BERT using clinical notes from the MIMIC-III dataset. The model is able to uncover high-quality relationships between medical concepts, as judged by physicians. Additionally, it outperformed BERT, a Bag-of-words model and a bidirectional LSTM on the readmission prediction task. The model achieved an AUC-ROC of 0.71 and an AUC-PR of 0.70.

Another BERT model trained on clinical text was released by Alsentzer et al., 2019. Their work used approximately 2 million notes from the MIMIC-III to train two varieties, one that uses text from all note types and another that uses only discharge summaries. The models were applied in natural language inference and named entity recognition. The model showed improvements on three out of five tasks, where the remaining were de-identifications tasks. In one of the tasks, the resulting performance of 82.7% accuracy was an improvement over the previous state-of-the-art of 73.5%.

4.5.6 Time and memory requirements

The training times and memory requirements of the models are presented in Table 4.9. Memory requirements are shown in terms of maximum GPU memory required to successfully run the training procedure, as it was not possible to capture the actual memory usage. All experiments were conducted on NVIDIA A100 Tensor Core GPUs.

TABLE 4.9: Training time and memory usage

Model	Training time	Parameter count	Maximum GPU memory
DAN	1 m	28.5 M	5 GB
LSTM	3 m	29.2 M	10 GB
ClinicalBERT	29 m	112 M	40 GB
Longformer 1K	8 h 30 m	134 M	40 GB
Longformer 2K	9 h 41 m	135 M	40 GB
Longformer 4K	6 h 41 m	136 M	40 GB

The baseline models required the least amount of time and memory, both took less than five minutes to train. Memory wise, the DAN model required a maximum of five GB, while the LSTM required double the amount.

The BERT-based models are significantly more costly to train, all took more than six hours to fine-tune and required a maximum of 40 GB of memory. Here, the training time excludes pre-training, which took about two days with Longformer 4K model. Also, it is important to note that these models required using gradient accumulation and reduced precision.

In this particular instance, besides having much higher capacity, and in turn being costlier to run, the BERT-based models were not able to beat the LSTM baseline. The models have acceptable performance on the training set, however on the validation set it is much worse, suggesting that collecting more data is necessary to improve it (Goodfellow, Bengio, and Courville, 2016).

4.6 Summary

In this chapter, the hyperparameters used for training the models are presented along with the scores obtained by evaluating them on the test dataset.

On the 30-day hospital readmission using discharge summaries task, the LSTM model achieved the highest scores across all metrics. The 2K model achieved the highest score among the Longformer attention models, in terms of RP70. The Longformer models are better than ClinicalBERT in terms of AUC-PR, in terms of AUC-ROC except for Longformer 4K, they are worse.

On the 30 hospital readmission using early notes task, we observe an increase in prediction accuracy as the days go by for both the Longformer 4K and ClinicalBERT models. This trend is maintained until the fifth day, then there is a slight drop in performance. Overall, the ClinicalBERT model has better performance on this task, however the score difference is always below 1%.

Domain-specific models derived from BERT suffer from out of vocabulary words, which causes them to have to learn concepts using fragmented words. Pre-training from scratch using biomedical domain text was shown to lead to superior results by Gu et al., 2020.

Chapter 5

Conclusion and Future Work

This chapter concludes the study by summarizing the key findings in relation to the research questions, as well as the main contributions. It will also discuss the limitations and propose opportunities for future research.

This study aimed to evaluate a Transformer-based model with Longformer's attention mechanism on clinical domain text. Three Longformer attention models were pre-trained and fine-tuned on notes from the MIMIC-III dataset, along with those a standard attention model and two baseline models were also trained. The baseline models consist of a Deep Averaging Network (DAN) and a Long short-term memory (LSTM).

The results indicate that increasing the attention mechanism size on a Longformer-based model results in better classification performance. Further findings show that this model reaches its best performance on the 4^{th} day of a patient's admittance into an ICU, with a slight decrease thereafter.

Although the Longformer-based model had better performance than the standard attention model, it did not achieve better performance than the LSTM baseline. One of the reasons could be the effect of out-of-vocabulary words that can hinder the model's performance. Models initialized from BERT, which was trained on general domain text, tend to have many common biomedical concepts missing from their vocabulary. Due to the way Byte pair encoding (Sennrich, Haddow, and Birch, 2016) works, missing words are encoded as sequences of subword units. As a result, standard BERT models are forced to divert parametrization capacity and training bandwidth to model biomedical terms using fragmented subwords. Pre-training on domain specific text can significantly outperform mixed-domain pre-training (Gu et al., 2020). Further analysis of the test examples reveals that the model struggles to classify cases related to mental health, it could be that it has not seen enough of these types of cases and a better sampling strategy is necessary.

For future research it is recommended to do pre-training from scratch with in domain text as opposed to initializing the model from a BERT checkpoint. This would be useful to mitigate the problem with out-of-vocabulary words. Furthermore, using a benchmark dataset like the Biomedical Language Understanding and Reasoning Benchmark (Gu et al., 2020) would be a more robust evaluation strategy for the model as it encompasses diverse biomedical tasks and

has data that is manually annotated by experts. Additionally, other efficient attention models that try to reduce the quadratic complexity could be explored, models like Routing Transformer (Roy et al., 2021) and Sparse Sinkhorn Attention (Tay et al., 2020) have show promising results however they have yet been applied in the clinical domain.

Token attribution examples

Legend: Negative Neutral Positive Predicted 1 Actual 1

Word Importance

[CLS] date of birth : sex : m service : medicine all ##er ##gies : pen ##ici ##llins / er ##yt ##hr ##omy ##cin base / st ##re ##pt ##omy ##cin / ci ##tric acid / ate ##no ##lo ##l / tor ##se ##mide / he ##par ##in agents attending : chief complaint : bleeding from col ##ost ##omy and foley major surgical or invasive procedure : none history of present illness : 84 year old male with multiple co - mor ##bid ##ities including rec ##tal cancer s / p res ##ection and radiation in now with col ##ost ##omy , corona ##ry artery disease s / p ste ##nts , sy ##sto ##lic ch ##f , dil ##ated card ##iom ##yo ##pathy , at ##rial fi ##bri ##llation not on , cardiac arrest and complete heart block s / p ai ##cd / pace ##maker , recent tr ##ach / peg after prolonged hospital ##ization for rib fractures / fl ##ail chest s / p fall who presents with large amount of bleeding from col ##ost ##omy and foley . the patient also endorsed increased short ##ness of breath , weakness and fatigue . the patient had been recently admitted 8 / 21 - 25 / to the mic ##u for pneumonia and mrs ##a ba ##cter ##emia / sep ##sis and had been discharged on van ##com ##y ##cin and baby as ##pi ##rin (for his at ##rial fi ##bri ##llation and corona ##ry artery disease) . . in the ed , initial vital signs were : af ##eb ##ril ##e , hr ##70 , bp ##10 ##7 / 53 , rr ##24 , 91 % on tr ##ach vent . given his tender abdomen , there was initial concern for an intra ##b ##dom ##inal event . ac ##s / general surgery was consulted and ct abdomen / pe ##lvis was performed . as the imaging was negative , there were no acute surgical concerns ; ac ##s recommended tagged rb ##c if he continued to bleed . gi was also consulted and recommended two units pr ##bc for hc ##t 22 , iv pp ##i initiation . the patient also underwent peg lava ##ge which was clear but was passing small amounts of maroon colored stool ##s . there was also initial concern for di ##c but labs not suggest ##ive of this (the patient has history of he ##par ##in induced th ##rom ##bo ##cy ##top ##enia but plate ##lets normal) . the patient has had hem ##at ##uria in the past , with three - way foley in place . . upon arrival to the mic ##u , the patient was resting comfortably in bed with tr ##ach / peg , end ##ors ##ing fatigue and abdominal pain . per the patient ' s wife , the patient looked good on discharge two days ago but seemed under the weather on arrival to rehab . when she saw him at 6 ##pm yesterday evening , he was stable without signs of bleeding . . review of systems : (+) per hp ##i (-) fever , headache , cough , short ##ness of breath , or w ##hee ##zing . chest pain , chest pressure , pal ##pit ##ations , or weakness . nausea , vomiting . past medical history : - rec ##tal cancer s / p ex ##cision and x ##rt ( ) - cad s / p ste ##nts (?) - cv ##a in with residual right hand d ##yst ##hes ##ia - complete heart block s / p pace ##maker - h / o cardiac arrest (now with ai ##cd) - gi bleed secondary to angie ##cta ##sia ##s in the duo ##den ##um ( ) s / p ca ##uter ##ization via e ##g ##d - at ##rial fi ##bri ##llation - sy ##sto ##lic ch ##f (e ##f 40 - 45 %) - s / p fall with multiple rib fractures ( ) - mic ##u admission / for hem ##op ##ty ##sis , bleeding from tr ##ach - abd ##omi ##no ##per ##ine ##al res ##ection w / social history : resident of rehab w plans to return home ; previously had lived in with his wife , now w some depression about moving out of their 42 year home . has two children . retired computer science professor . - tobacco : 5 cigar ##s daily for 30 years , quit s / p cv ##a - alcohol : previously glasses / week , generally per wife " affects him quite a bit , " changing his mood and making him sick - illicit ##s : family history : father died in 80s from mi . mother died in 80s from pe . no family history of colon , breast , ut ##erine , or o ##var ##ian cancer . no family history of

FIGURE A.1: Example of a discharge summary where the model is able to make a correct prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. In this example, the model attributes its output mostly to subwords of tracheostomy, an opening created at the front of the neck, so a tube can be inserted into the windpipe (trachea) to help a patient breath.

Legend: Negative Neutral Positive Predicted 1 Actual 1

Word Importance

[CLS] date of birth : sex : f service : medicine all ##er ##gies : pl ##avi ##x / he ##par ##in agents attending : chief complaint : hem ##op ##ty ##sis , fever , inability to vent ##ila ##te major surgical or invasive procedure : tr ##ache ##ost ##omy revision arterial line central ve ##nous line peripheral ##ly inserted central cat ##het ##er history of present illness : 73 ##f h / o d ##m , multiple cv ##as , di ##sse ##minated tb on tx since , tr ##ache ##d for res ##p failure with prolonged int ##uba ##tion / failure to we ##an living at mac ##u / rehab , developed res ##p distress at 05 ##30 on , rr - 36 - 40 ; hr 136 , t noted to be 8 , bp noted to be 125 / pt apparently was already on van ##c and im ##ipe ##nem starting for presumed va ##p p ##na . mrs ##a apparently growing out of the sp ##ut ##um . per ems report 500 ##cc ##s of blood su ##ction ##ed this am from air ##way was noted to be difficult to vent ##illa ##te . su ##ction ##ed , with tr ##ach changes without su ##cc ##es , still difficulty with diminished r sided lung sounds . difficulty with bag valve mask as well . staff at he ##bre ##b denies trauma . . on arrival to the , pt ' s bp 88 / 38 , rr 29 , 96 % sat ##ting . lev ##op ##hed started when sp ##b transient ##ly to the 50 ##s . iv ##f given , bp with press ##ors rebound ##ed to 110 ##s - 120 ##s . in ed , h ##yp ##ote ##ns ##ive transient ##ly , started on press ##ors . ip notified and tried to rep ##osition ##ed the tr ##ach . the patient was bro ##nched and 80 % o ##cc ##lusion of tr ##ach tube with gran ##ulation tissue was noted distal to the end of tr ##ach . ip unable to pass scope . pt was int ##uba ##ted from above . et ##t in ##ital ##ly placed into the r main ##ste ##m bro ##nch ##us with collapse of l lung noted . the tube was then pulled back with successful re - inflation of the l lung . pt given van ##c , ce ##ft ##ax , flag ##l and was placed on lev ##ph ##ed gt ##t to maps > fen ##t and verse ##d gt ##t were started as well . . pt previously admitted on for ata ##xia , was d ##x ' d with acute stroke , di ##sse ##minated tb , seizure ##rs , was tr ##ache ##d due to prolonged int ##uba ##tion / failure to we ##an , then x ##fer ##red on to for further care . from there , the pat was reportedly x ##fer ##red to mac ##u at he ##b re ##b where she has been residing then (apparently off precautions) . . id history : she was found to have pan - sensitive tb on 3 drug regime ##n (in ##h / et ##h / p ##za) . her course has been complicated by worse ##ning cv ##as (now bilateral) , h ##yp ##ox ##ic respiratory failure , continued fever ##s . she has had an extensive id work - up and all studies negative to date except for tb . bone marrow (as ##pi ##rate only) - with few blasts repeat b ##m - 10 - 20 % blasts , core bio ##psy md ##s . liver bio ##psy - nec ##rot ##ic with few gran ##ulo ##mas (no micro sent) . all micro stains on pathology negative only positive tests : ace level and hit antibody b - g ##lu ##can positive on , negative on repeat . had right va ##ts , right su ##pr ##ac ##lav ##ic l ##n , and liver bio ##psy done bio ##psy shows gran ##ulo ##ma suspicious for infectious process . tissue from lung growing afb . sp ##ut ##um c ##x also with afb when grew . tb is pan - sensitive . repeat ct chest / abd / pe ##lvis did not show marked improvement but did not show worse ##ning either . tb med ##s switched at beginning of - no response to holding ri ##fa ##mp ##in . then tried changing to st ##m / lev ##aq ##uin to see if was one of the other drugs ; no effect . trial of am ##bis ##ome ; no effect . has had lo ##cula ##ted e ##ff ##usions but now no fluid ; tapping them the first time showed no growth on c ##x ; the second time there was no fluids . . she had bc ##x with coa ##g ne ##g st ##ap ##h seen on with no subsequent growth to date . rec ' d em ##pi ##ric van ##co tx with f / up tee (tt ##e from was without

FIGURE A.2: Example of a discharge summary where the model is able to make a correct prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. The model's output is attributed to subwords of tracheostomy, tr and ##ache.

Legend: Negative Neutral Positive Predicted 1 Actual 1

Word Importance

[CLS] date of birth : sex : m service : medicine all ##er ##gies : fen ##tan ##yl attending : chief complaint : chest tight ##ness , short ##ness of breath major surgical or invasive procedure : none history of present illness : mr . is a 27 ##yo male with a history of ha ##j ##du - syndrome , chronic tr ##ach with restrictive and ob ##st ##ru ##ctive lung disease , multiple episodes of p ##na including pseudo ##mona ##l p ##na , who presents now with chest tight ##ness and progressive d ##ys ##p ##nea over the past week . . patient was hospitalized from with sep ##sis , and treated for p ##na empirical ##ly with van ##c / mer ##open ##em . was then read ##mit ##ted from , still on van ##c / mer ##open ##em , again for p ##na , with mini - bal culture at time positive for pseudo ##mona ##s resistant to mer ##open ##em and ce ##fe ##pi ##me . patient treated with to ##bra ##my ##cin and ce ##ft ##az ##adi ##me for planned 21 day course (starting) , and was followed by id in out ##patient setting following discharge . patient reports he had been doing well since that time , with no interval infections . has been off pre ##din ##son ##e for his ob ##st ##ru ##ctive lung disease since , but states he has still been on ba ##ct ##rim daily . . patient has h / o ha ##j ##du - syndrome , which is a rare auto ##som ##al dominant congenital connect ##ive tissue disorder characterized by severe and excessive bone res ##or ##ption leading to os ##te ##op ##oro ##sis and bony def ##or ##mit ##ies . also has h / o restrictive lung disease secondary to severe ky ##ph ##os ##col ##ios ##is , as well as component of ob ##st ##ru ##ctive lung disease . has ch ##rn ##ic tr ##ache ##oto ##my tube , and is on 4 ##l tr ##ach collar at home during day ; on sim ##v at night . however , over the past week , patient has noted chest tight ##ness and worse ##ning sob . no significant cough , though with cough patient producing thick , " sp ##ong ##y " white mu ##cous . no fever or chill ##s . patient ' s vent requirements increased , and he has been on vent both during day and at night over the past 1 - 2 days . reports symptoms are similar to those with previous pneumonia ##s . given progressive symptoms , came to ed for evaluation . . in the ed , initial vs were : t 8 p 102 bp 133 / 88 r 20 o ##2 100 % sat on vent settings ac 400 ##x ##20 pee ##p 5 , 35 % fi ##o c ##x ##r suggest ##ive of right - sided infiltrate . patient was given ce ##fe ##pi ##me and lev ##aq ##uin and a 500 ##cc bo ##lus of ns , and admitted to ic ##u for further evaluation and management . . on arrival to the ic ##u , patient reports continued sub ##ster ##nal chest tight ##ness , worse with inspiration , and short ##ness of breath , which have gradually worsened over past week . denies any significant cough at present . no recent sick contacts . recent choking with food or as ##piration events he can recall . . review of systems : (+) per hp ##i . rhino ##rr ##hea and nasal congestion . occasional con ##sti ##pati ##on for which he takes mira ##la ##x . no bloody stool ##s . occasional heart pal ##pit ##ations . chronic back pain , and increased left knee pain . also reports pr ##uri ##tis on back and left arm , denies any other rash ##es . (-) denies fever , chill ##s , headache , sin ##us tenderness . no changes in vision . denies nausea , vomiting , dia ##rr ##hea , abdominal pain , or changes in bow ##el habits . denies d ##ys ##uria , frequency , or urgency . denies my ##al ##gia ##s . past medical history : ha ##j ##du - syndrome (rare auto ##som ##al dominant congenital connect ##ive tissue disorder characterized by severe and excessive bone res ##or ##ption leading to os ##te ##op ##oro ##sis , also known as type vi id ##io ##pathic os ##te ##ol ##ysis) os ##te ##omy ##eli ##tis , right ole ##cr ##ano ##n (pressure - related) chronic ob ##st ##ru ##ctive / restrictive lung

FIGURE A.3: Example of a discharge summary where the model is able to make a correct prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. The model's output is attributed to subwords of tracheostomy, tr and ##ache.

Legend: Negative Neutral Positive Predicted 1 Actual 1

Word Importance

[CLS] date of birth : sex : m service : medicine all ##er ##gies : patient recorded as having no known all ##er ##gies to drugs attending : chief complaint : altered mental status / hyper ##car ##bic respiratory failure major surgical or invasive procedure : int ##uba ##tion for respiratory distress . right internal jug ##ular central ve ##nous cat ##het ##er placement and removal . history of present illness : 64 yo male with a pm ##h significant for mor ##bid obesity , d ##mi ##i , ge ##rd , h ##t ##n , hyper ##lip ##ide ##mia , and prior vent ##ila ##tory dependency (felt to be multi ##fa ##ctor ##ial in the setting of obesity , os ##a , and ? cop ##d) s / p tr ##ach removal for plug ##ged tr ##ach , who presented from nh with increased so ##m ##no ##lence . per ed report , he had been confused with hall ##uc ##inations over the past 3 - 4 days with let ##har ##gy and sob . he was treated with ba ##ct ##rim , lev ##of ##lo ##xa ##cin 750 mg , daily , and im ##ipe ##nem 500 mg iv q 6 ##hr for a ut ##i . he had no f / c / s , no cough , no focal ne ##uro change . he was ar ##ous ##able only with pain and movement . . in the ed , his vital ##s were : . t ##m 99 hr 81 - 95 bp 69 - 148 / 61 - 81 rr 16 - 20 sat 93 % 2 ##ln ##c . initial ab ##g was : 37 / 65 / lac ##tate was he did not respond to a dose of na ##rca ##n . later ab ##g was 25 / 74 / 42 and pt was noted to have increased so ##m ##no ##lence . serum and urine to ##x screens were negative . c ##x ##r was unchanged with a l hem ##idia ##ph ##rag ##matic elevation (not new) . t ##p ##n was 20 , which is baseline . ek ##g revealed old lb ##bb . ua appeared dirty , so pt was treated with van ##c and ct ##x . ic ##u course : after fiber ##op ##tic int ##uba ##tion for hyper ##car ##bic respiratory failure , sb ##p dropped to the 60s , so prop ##of ##ol gt ##t was stopped . he remained h ##yp ##ote ##ns ##ive so cv ##l was placed and 5 l ns were administered . he was also started on do ##pa ##mine gt ##t > lev ##op ##hed gt ##t , which was quickly tape ##red off , and he was able to be ex ##tub ##ated . past medical history : - respiratory failure on mechanical vent ##ila ##tor with tr ##ache ##ost ##omy since . chronic failure was felt to be multi ##fa ##ctor ##ial and related to cop ##d , os ##a , and obe ##sti ##y . complicated by sub ##gl ##otti ##c ste ##nosis requiring rigid bro ##nch ##os ##co ##py with ab ##lation of gran ##ulation tissue ; pt had cl ##og ##ged tr ##ache ##ost ##omy tube with patent airways above and below so tube was removed - d ##m - hyper ##cho ##les ##terol ##emia - h / o afi ##b in the past while in ic ##upt declined anti ##co ##ag ##ulation but did require ami ##oda ##rone - depression - ge ##rd - mor ##bid obesity - ne ##uro ##pathy - h / o ch ##f - h / o h ##yp ##ona ##tre ##mia - h / o cv ##a in with residual r sided weakness - d ##m social history : no et ##oh or tobacco use . lives at family history : non ##con ##tri ##bu ##tory physical exam : vital ##s : 0 p 87 bp 126 / 62 rr 18 sat 96 % on shovel mask gen : mor ##bid ##ly obe ##se male laying flat in bed hee ##nt : int ##uba ##ted , per ##rl , con ##jun ##ct ##iva ##e an ##ic ##ter ##ic / non ##in ##jected , mmm lungs : scattered ex ##pi ##rator ##y w ##hee ##zes cv : rr ##r , nl s ##1 , s ##2 , no m / r / g abd : soft , nt , n ##d , + bs ex ##tre ##m : no c / c / 1 + ed ##ema in le up to knees ne ##uro : awake , alert , and oriented x ##3 per ##tine ##nt results : 07 : 50 ##pm wb ##c - 0 rb ##c - 49 * h ##gb - 7 * hc ##t - 7 * mc ##v - 102 * mc ##h - 5 * mc ##hc - 8 rd ##w - 0 07 : 50 ##pm ne ##uts - 5 l ##ym ##phs - 8 mono ##s - 3 e ##os - 9 * bas ##os - 5 07 : 50 ##pm macro ##cy ##t - 2 + 07 : 50 ##pm pl ##t count - 195 07 : 50 ##pm pt - 7 * pt ##t - 8 in ##r (pt) - 3 * 07 : 50 ##pm urine color - yellow appear - clear sp - 01 ##9 07 : 50 ##pm urine blood - mod ni ##tri ##te -

FIGURE A.4: Example of a discharge summary where the model is able to make a correct prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. The model's output is attributed to subwords of tracheostomy, tr and ##ache.

Legend: Negative Neutral Positive Predicted 1 Actual 1

Word Importance

[CLS] date of birth : sex : f service : medicine all ##er ##gies : pen ##ici ##llins / per ##co ##ce ##t attending : chief complaint : hyper ##tens ##ive urgency major surgical or invasive procedure : hem ##od ##ial ##ysis history of present illness : from admission note : the pt is a 24 y . o . f with es ##rd on hd , sl ##e , mali ##gnant h ##t ##n , history of sv ##c syndrome admitted with h ##t ##n and sob in the setting of missed hd . the patient reported missing hd yesterday because she thinks she is being over ##di ##ures ##ed . she reports persistent pain at site of rec ##tus sheath hem ##ato ##ma . denies n / v / d . pt recently admitted from with consistent abdominal pain at the site of her known left abdominal wall hem ##ato ##ma , hyper ##tens ##ive to 230 ' s and hyper ##kal ##emi ##c to 2 after missing her last two dial ##ysis sessions . at this time the pt . was dial ##y ##zed , received a blood trans ##fusion , and was administered her daily anti ##hy ##per ##tens ##ive medications . pt . left ama after her trans ##fusion despite the primary team ' s concerns to look for an active area of bleeding . in the ed , patient complain of mild d ##ys ##p ##nea , satin ##g well on ra . c ##x ##r mild volume over ##load . ku ##b with no evidence of obstruction . she was started on a lab ##eta ##lo ##l gt ##t . ec ##g - ra ##d , l ##v ##h no change from prior . hc ##t stable at the renal team evaluated pt and recommended hd , however the patient refused . she was transferred to ic ##u for bp control . past medical history : systemic lu ##pus er ##ythe ##mat ##os ##us : - diagnosed (16 years old) when she had swollen fingers , arm rash and art ##hra ##l ##gia ##s - previous treatment with cy ##to ##xa ##n , cell ##ce ##pt ; currently on pre ##d ##nis ##one - complicated by uv ##eit ##is ( ) and es ##rd ( ) ck ##d / es ##rd : - dia ##gos ##ed - initiated dial ##ysis but refused it as of , has survived despite this - pd cat ##het ##er placement mali ##gnant hyper ##tension - baseline bp ##s 180 ' s - 120 ' s - history of hyper ##tens ##ive crisis with seizures - history of two intra ##par ##en ##chy ##mal hem ##or ##rh ##ages that were thought due to the posterior rev ##ers ##ible le ##uk ##oe ##nce ##pha ##lo ##pathy syndrome , associated with le par ##esis in that resolved th ##rom ##bo ##cy ##top ##enia : - tt ##p (got plasma ##pher ##esis ##is) versus mali ##gnant h ##t ##n th ##rom ##bot ##ic events : - sv ##c th ##rom ##bos ##is ( ) ; related to a cat ##het ##er - negative lu ##pus anti ##co ##ag ##ula ##nt (, ,) - negative anti ##card ##iol ##ip ##in antibodies i ##gg and i ##gm x ##4 (-) - negative beta - 2 g ##ly ##co ##pro ##tein antibody (,) hoc ##m : last noted on echo an ##emia history of left eye en ##uc ##lea ##tion for fungal infection history of va ##ginal bleeding lasting 2 months s / p de ##pop ##rove ##ra injection requiring trans ##fusion history of coa ##g negative st ##ap ##h ba ##cter ##emia and hd line infection - and th ##rom ##bot ##ic micro ##ang ##io ##pathy : may be et ##iology of episodes of worse hyper ##tension given appears quite lab ##ile ob ##st ##ru ##ctive sleep ap ##nea , auto ##cp ##ap / pressure setting , straight cp ##ap / pressure setting 7 ps ##h ##x : placement of multiple cat ##het ##ers including dial ##ysis . tons ##ille ##ct ##omy . left eye en ##uc ##lea ##tion in . pd cat ##het ##er placement in . s / p ex - lap for free air in abdomen , ex - lap normal social history : single and lives with her mother and a brother . she graduated from high school . the patient is on disability . the patient does not drink alcohol or smoke , and has never used recreational drugs . family history : negative for auto ##im ##mun ##e diseases including sl ##e , th ##rom ##bo ##phi ##lic disorders . maternal grandfather with h ##t ##n , mi , stroke in 70s . physical exam : gen : sleeping comfortably , easily

FIGURE A.5: Example of a discharge summary where the model is able to make a correct prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. The model's output is attributed to subwords of esrd, malignant, microangiopathy. The word complicated is also influential.

Legend: Negative Neutral Positive Predicted 0 Actual 1

Word Importance

[CLS] date of birth : sex : m service : medicine all ##er ##gies : patient recorded as having no known all ##er ##gies to drugs attending : chief complaint : suicide attempt , et ##oh withdrawal major surgical or invasive procedure : none history of present illness : mr . is a 60 year old male with a history of depression , anxiety , substance use , hc ##v , and prior suicide attempts who presents with alcohol withdrawal after a suicide attempt earlier today . he ran out of his psychiatric medications two weeks ago and has felt increasingly suicidal since . he attempted to hang himself with a belt from the ceiling and after kicking the chair out from under himself the ceiling fell and he landed on the ground . he then called a cab to bring him to the ed . he normally drinks a qu ##art to a half gallon of liquor . he drank about a qu ##art this morning . he end ##ors ##es prior history of dt ##s and seizures with alcohol withdrawal . . in the ed , initial vs were : pain 0 , t 8 , hr 97 , bp 180 / 111 , rr 22 , o ##2 sat 98 % ra . he was noted to have an ##iso ##cor ##ia and to be tre ##mu ##lous . he had no d ##ys ##phon ##ia or d ##ys ##pha ##gia . imaging of the head and neck showed no fracture , ich , or arterial abnormalities . patient was given val ##ium 10 mg iv x 2 and 5 mg iv x vital signs on transfer were : hr 81 , bp 154 / 93 , rr 13 , 100 % ra . he was admitted to the ic ##u due to concern for severe alcohol withdrawal . . on arrival to the ic ##u , the patient was tre ##mu ##lous and stated that he felt so bad that he wanted to die . later he denied any desire to kill himself and stated that he simply wanted help . . review of sy ##tem ##s : (+) per hp ##i . + 2 / 10 chest pain and abdominal pain . + chill ##s , + short ##ness of breath . + pain with ur ##ination . (-) denies fever , recent weight loss or gain . denies headache , sin ##us tenderness , rhino ##rr ##hea or congestion . denied cough . denied nausea , vomiting , dia ##rr ##hea , con ##sti ##pati ##on or abdominal pain . no recent change in bow ##el or bladder habits . denied art ##hra ##l ##gia ##s or my ##al ##gia ##s . no intercourse x 7 years . . past medical history : rei ##ter ' s syndrome hc ##v h ##x iv ##du , in met ##had ##one program , recent re ##la ##pse ##s h ##x suicidal ##ity depression and anxiety mit ##ral valve pro ##la ##pse os ##te ##oa ##rth ##rit ##is , chronic pain hyper ##tension bell ' s pal ##sy s / p left li ##s fran ##c or ##if social history : lives alone . former nurse , currently on disability . no tobacco . - gallon et ##oh daily . denies recent il ##ici ##t drug use but has past history of heroin use . states he bought k ##lon ##op ##in 2 mg # 15 tab ##s off the street and took them all this past weekend to " help me come down " . family history : mother had an alcohol problem until she was 48 and has since been sober . mother also had colon cancer . per om ##r , depression in maternal relatives . physical exam : vital ##s : t : 2 bp : 140 / 101 p : 82 r : 15 o ##2 : 100 % ra general : middle - aged caucasian male , tre ##mu ##lous , appears uncomfortable . hee ##nt : sc ##ler ##a an ##ic ##ter ##ic , mm dry , oro ##pha ##ryn ##x with thick , dry secret ##ions . neck : su ##pp ##le , j ##v ##p not elevated , no lad lungs : clear to aus ##cu ##lta ##tion bilateral ##ly , no w ##hee ##zes , ra ##les , ron ##chi cv : regular rate and rhythm , normal s ##1 + s ##2 , no murmurs , rubs , gallo ##ps abdomen : + bs , soft , mildly tender per ##ium ##bil ##ically and in the ep ##iga ##st ##rium , no rebound tenderness or guarding gu : no foley ex ##t : warm , well per ##fus ##ed , 2 + pulses , no club ##bing , cy ##ano ##sis or ed ##ema ne ##uro : a & o , cn ##ii - xii intact , per ##rl (4 - > 2) , moves all ex ##tre ##mit ##ies ps ##ych : responds minimal ##ly to questions . ? ? ? ? ? ? i feel like i ? ? ? ? ? ? m crawling out of my skin ? ? ? ? ? ? . appears depressed . end

FIGURE A.6: Example of a discharge summary where the model makes an incorrect prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. In this example, the probability of readmission (Equation 3.1) is of 0.1, a value too low for a positive classification. Most of the tokens that contribute positively punctuation marks, the model seems to not pay attention to any medical concept, perhaps due to subword tokenization (Section 4.5.3).

Legend: Negative Neutral Positive Predicted 0 Actual 1

Word Importance

[CLS] date of birth : sex : m service : medicine all ##er ##gies : hal ##do ##l attending : chief complaint : cc : major surgical or invasive procedure : int ##uba ##tion upper end ##os ##co ##py history of present illness : hp ##i : 38 yo male with h ##x of suicide attempts initially presented to after a suicide attempt with ? 12 cl ##ona ##ze ##pa ##m and end ##ors ##ing suicidal idea ##tion . while in the ed , he revealed that he had ing ##ested broken glass . of note , the patient was recently discharged from on after ing ##est ##ing broken glass and for ing ##est ##ing razor blades . he has a history of swallowing glass and razor blades . he had an e ##g ##d here in after he swallowed multiple razor blades . denied hem ##ate ##mes ##is or l ##gi ##b . denied any light ##head ##ed ##ness . he does note diffuse abdominal pain . denies nausea , vomiting . . in the ed : - urine to ##x - positive for cocaine - 2 ##mg iv mor ##phine for abdominal pain - portable c ##x ##r , abdomen negative for free air , or glass - pa and la ##t - no free air , no glass seen , no widened media ##sti ##num - gi saw patient and will scope tonight - admitted to mic ##u for upper end ##os ##co ##py past medical history : - suicide attempt - swallowing glass and razor blades - bipolar - depression - sp ##lene ##ct ##omy social history : the patient reports that he was born in the area and the youngest of 3 , he has 2 older sisters . his father was an abusive alcoholic and his mother was extremely abusive to the children . he said " she burned my hands and all kind of things . " his mother placed all her children in a catholic institution when he was 3 where he was sexually abused by a priest . sued the arch ##dio ##ces ##an and was awarded $ 350 , 000 which he says is in a trust . he was adopted at age 12 and was abused by his adoptive mother . at one point he went to to live with his biological mother and graduated from high school in after attending 3 high schools . he has worked in banks , retail , as a model , a strip ##per , selling cars but is now on disability . he is close with a family in who he refers to as his god ##par ##ents and also to one of his sisters . currently lives alone in , never married and no children . . - smoke ##s pp ##d , recreational cocaine and marijuana use ; occasional et ##oh use family history : father sister with depression and psychiatric hospital ##izations physical exam : vs : t : 7 bp : 102 / 66 hr : 66 rr : 12 o ##2 : 99 % on ra hee ##nt : norm ##oc ##ep ##hal ##ic , an ##ic ##ter ##ic , per ##rl ##a , e ##omi , no ph ##ary ##nge ##al injection , ex ##uda ##te , neck su ##pp ##le , dent ##ition fair , no lad , no thy ##rom ##ega ##ly cv : + s ##1 + s ##2 rr ##r pu ##lm : ct ##a b / l no rr ##w abd : ex ##t : no club ##bing , no ed ##ema per ##tine ##nt results : data : ek ##g : ns ##r @ normal axis and intervals . no st - t wave changes . . c ##x ##r : : no evidence of radio ##pa ##que foreign object . no sub ##dia ##ph ##rag ##matic free air . a lateral radio ##graph is recommended to further and more closely assess the es ##op ##ha ##gus . . abd x ##ray : : findings : on this single su ##pine view , the presence of free air cannot be directly assessed . the bow ##el gas pattern is un ##rem ##ark ##able . no air - fluid levels or di ##sten ##ded loops of small bow ##el are present . limited views of the hips and sac ##roi ##lia ##c joints are un ##rem ##ark ##able . 01 : 10 ##am glucose - 86 ur ##ea n - 16 cr ##ea ##t - 0 sodium - 143 potassium - 7 chloride - 102 total co ##2 - 23 an ##ion gap - 22 * 01 : 10 ##am alt (sg ##pt) - 63 * as ##t (sg ##ot) - 66 * ld (ld ##h) - 251 * al ##k ph ##os - 73 amy ##lase - 40 to ##t bi ##li - 3 01 : 10 ##am lip ##ase - 19 01 : 10 ##am album ##in - 6 01 : 10 ##am wb ##c - 4 * # rb ##c - 50 * h ##gb - 5 hc ##t - 6 mc ##v - 92 mc ##h - 2 * mc ##hc - 9 rd ##w - 3 01 : 10 ##am ne ##uts - 5 l ##ym

FIGURE A.7: Example of a discharge summary where the model makes an incorrect prediction. Highlighted in green are tokens that positively contribute towards the predicted class, while in red are tokens that contribute negatively. The probability of readmission is 0.1. Tokens that negatively impact are related to suicide, the single token with positive attribution is the word swallowing, in a passage where the patient is said having swallowed glass and razor blades.

Bibliography

Alammar, Jay (2018). *The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time*. URL: http://jalammar.github.io/illustrated-transformer/.

Alsentzer, Emily et al. (June 2019). "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton (2016). "Layer Normalization". In: *ArXiv preprint* abs/1607.06450. URL: https://arxiv.org/abs/1607.06450.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: *ArXiv preprint* abs/2004.05150. URL: https://arxiv.org/abs/2004.05150.

Benbassat, Jochanan and Mark Taragin (2000). "Hospital readmissions as a measure of quality of health care: advantages and limitations". In: *Archives of internal medicine* 160.8, pp. 1074–1081.

Bengio, Y, P Simard, and P Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". eng. In: *IEEE transactions on neural networks* 5.2, pp. 157–166. ISSN: 1045-9227.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828. ISSN: 01628828. DOI: 10.1109/TPAMI.2013.50. arXiv: 1206.5538.

Biewald, Lukas (2020). *Experiment Tracking with Weights and Biases*. Software available from wandb.com. URL: https://www.wandb.com/.

Brochu, Eric, Vlad M Cora, and Nando De Freitas (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*.

Child, Rewon et al. (2019). "Generating Long Sequences with Sparse Transformers". In: *ArXiv preprint* abs/1904.10509. URL: https://arxiv.org/abs/1904.10509.

Cho, Kyunghyun et al. (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al., pp. 7057–7067. URL: https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*. Ed. by William W. Cohen and Andrew W. Moore. Vol. 148. ACM International Conference Proceeding Series. ACM, pp. 233–240. DOI: 10.1145/1143844.1143874. URL: https://doi.org/10.1145/1143844.1143874.

Devlin, Jacob et al. (2019). "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, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dreyer, R and A J Viljoen (2019). "Evaluation of factors and patterns influencing the 30-day readmission rate at a tertiary-level hospital in a resource-constrained setting in Cape Town, South Africa". eng. In: *South African medical journal* 109.3, pp. 164–168. ISSN: 0256-9574.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61, pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html.

Escudié, Jean-Baptiste et al. (2017). "A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease". In: *BMC Medical Informatics and Decision Making* 17.1, p. 140. ISSN: 1472-6947. DOI: 10.1186/s12911-017-0537-y. URL: https://doi.org/10.1186/s12911-017-0537-y.

Fernández, Alberto et al. (2018). *Learning from Imbalanced Data Sets*, pp. 47–61. DOI: 10.1007/978-3-319-98074-4.

Ford, Elizabeth et al. (2016). "Extracting information from the text of electronic medical records to improve case detection: a systematic review". In: *Journal of the American Medical Informatics Association* 23.5, pp. 1007–1015.

Ghassemi, Marzyeh et al. (2014). "Unfolding physiological state: mortality modelling in intensive care units". In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. Ed. by Sofus A. Macskassy et al. ACM, pp. 75–84. DOI: 10.1145/2623330.2623742. URL: https://doi.org/10.1145/2623330.2623742.

Gomez-Perez, Jose Manuel, Ronald Denaux, and Andres Garcia-Silva (2020). "Understanding Word Embeddings and Language Models". In: *A Practical Guide to Hybrid Natural Language*

Processing: Combining Neural Models and Knowledge Graphs for NLP. Cham: Springer International Publishing, pp. 17–31. ISBN: 978-3-030-44830-1. DOI: 10.1007/978-3-030-44830-1_3. URL: https://doi.org/10.1007/978-3-030-44830-1_3.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM networks". In: *Proceedings of the International Joint Conference on Neural Networks* 4, pp. 2047–2052. DOI: 10.1109/IJCNN.2005.1556215.

Greff, Klaus et al. (2015). "LSTM: A Search Space Odyssey". In: *ArXiv preprint* abs/1503.04069. URL: https://arxiv.org/abs/1503.04069.

Gu, Yu et al. (2020). "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ArXiv preprint* abs/2007.15779. URL: https://arxiv.org/abs/2007.15779.

Hinton, Geoffrey E. et al. (2012). "Improving neural networks by preventing co-adaptation of feature detectors". In: *ArXiv preprint* abs/1207.0580. URL: https://arxiv.org/abs/1207.0580.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". eng. In: *Neural computation* 9.8, pp. 1735–1780. ISSN: 1530-888X.

Hsu, Chao-Chun et al. (2020). "Characterizing the Value of Information in Medical Notes". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2062–2072. DOI: 10.18653/v1/2020.findings-emnlp.187. URL: https://aclanthology.org/2020.findings-emnlp.187.

Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. arXiv: 1904.05342 [cs.CL].

Iyyer, Mohit et al. (2015). "Deep Unordered Composition Rivals Syntactic Methods for Text Classification". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 1681–1691. DOI: 10.3115/v1/P15-1162. URL: https://aclanthology.org/P15-1162.

Johnson, Alistair E.W. et al. (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1, p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://doi.org/10.1038/sdata.2016.35.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6980.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Ed. by Peter L. Bartlett et al., pp. 1106–1114. URL: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a6
Abstract.html.

Lai, Siwei et al. (2015). "Recurrent Convolutional Neural Networks for Text Classification". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. Ed. by Blai Bonet and Sven Koenig. AAAI Press, pp. 2267–2273. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745.

Lee, Jinhyuk et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTZ682. arXiv: 1901.08746. URL: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

Lewis, Mike et al. (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

Li, Yikuan et al. (2022). "CLINICAL-LONGFORMER AND CLINICAL-BIGBIRD: TRANSFORMERS FOR LONG CLINICAL SEQUENCES". In: URL: https://huggingface.co/yikuan8/Clinical-Longformer..

Liu, Jingshu, Zachariah Zhang, and Narges Razavian (2018). "Deep ehr: Chronic disease prediction using medical notes". In: *Machine Learning for Healthcare Conference*. PMLR, pp. 440–464.

Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv preprint* abs/1907.11692. URL: https://arxiv.org/abs/1907.11692.

Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

McInerney, Denis Jered et al. (2020). "Query-Focused EHR Summarization to Aid Imaging Diagnosis". In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 126. Proceedings of Machine Learning Research. PMLR, pp. 632–659. URL: https://proceedings.mlr.press/v126/mcinerney20a.html.

Melis, Gábor, Chris Dyer, and Phil Blunsom (2018). "On the State of the Art of Evaluation in Neural Language Models". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=ByJHuTgA-.

Mikolov, Tomás et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1301.3781.

Nguyen, Hoang and Jon Patrick (2016). "Text Mining in Clinical Domain: Dealing with Noise". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Balaji Krishnapuram et al. ACM, pp. 549–558. DOI: 10.1145/2939672.2939720. URL: https://doi.org/10.1145/2939672.2939720.

Peng, Yifan, Shankai Yan, and Zhiyong Lu (2019). "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, pp. 58–65. DOI: 10.18653/v1/W19-5006. URL: https://aclanthology.org/W19-5006.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Peters, Matthew E. et al. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Qiu, Jiezhong et al. (2020). "Blockwise Self-Attention for Long Document Understanding". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2555–2565. DOI: 10.18653/v1/2020.findings-emnlp.232. URL: https://aclanthology.org/2020.findings-emnlp.232.

Radford, Alec et al. (2018). "Improving Language Understanding by Generative Pre-Training". In: URL: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Rajkomar, Alvin et al. (2018). "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.1, pp. 1–10.

Roy, Aurko et al. (2021). "Efficient Content-Based Sparse Attention with Routing Transformers". In: *Transactions of the Association for Computational Linguistics* 9, pp. 53–68. DOI: 10.1162/tacl_a_00353. URL: https://aclanthology.org/2021.tacl-1.4.

Safran, Charles and Russell S Phillips (1989). "Interventions to prevent readmission: the constraints of cost and efficacy". In: *Medical care*, pp. 204–211.

Salakhutdinov, Ruslan (2014). "Deep learning". In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. Ed. by Sofus A. Macskassy et al. ACM, p. 1973. DOI: 10.1145/2623330.2630809. URL: https://doi.org/10.1145/2623330.2630809.

Sendelbach, S. and M. Funk (2013). "Alarm fatigue: a patient safety concern". In: *AACN Adv Crit Care* 24.4, pp. 378–386.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. DOI: 10.18653/v1/p16-1162. URL: https://doi.org/10.18653/v1/p16-1162.

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Ed. by Peter L. Bartlett et al., pp. 2960–

2968. URL: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani et al., pp. 3104–3112. URL: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Tabak, Ying P et al. (2017). "Predicting Readmission at Early Hospitalization Using Electronic Clinical Data: An Early Readmission Risk Score". eng. In: *Medical care* 55.3, pp. 267–275. ISSN: 0025-7079.

Tay, Yi et al. (2020). "Sparse Sinkhorn Attention". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9438–9447. URL: http://proceedings.mlr.press/v119/tay20a.html.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wiegreffe, Sarah et al. (2019). "Clinical Concept Extraction for Document-Level Coding". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, pp. 261–272. DOI: 10.18653/v1/W19-5028. URL: https://aclanthology.org/W19-5028.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv preprint* abs/1910.03771. URL: https://arxiv.org/abs/1910.03771.

Xiao, Cao, Edward Choi, and Jimeng Sun (2018). "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review". In: *Journal of the American Medical Informatics Association* 25.10, pp. 1419–1428. ISSN: 1527974X. DOI: 10.1093/jamia/ocy068.

Yang, Zhilin et al. (2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al., pp. 5754–5764. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Zhu, Yukun et al. (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *2015 IEEE International Conference on Computer*

Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, pp. 19–27. DOI: 10.1109/ICCV.2015.11. URL: https://doi.org/10.1109/ICCV.2015.11.