Clinical-Longformer: Whole Document Embedding and Classification for the Clinical Domain

Simon Levine-Gottreich
M.S. Computational Biology '21
Carnegie Mellon University
Pittsburgh, PA 15213
simonlevine@cmu.edu

Serena Abraham
M.S. Computational Biology '21
Carnegie Mellon University
Pittsburgh, PA 15213
smabraha@andrew.cmu.edu

Abstract

Using novel pre-training of Transformer encoders, this project tackles whole-document embedding for the clinical domain. Additionally, we propose a fine-tuning process on electronic healthcare records for transformer models and a novel medical coding classification benchmark task, on which we achieve state of the art performance. We release our best-performing encoder model and suggest future investigation with regard to the natural language tasks in this domain.

1 Introduction and Literature Review

Recently, Transformer (Vaswani et al. 2017) models have proven incredibly effective in language modeling tasks and downstream use, including unsupervised "pre-training" of a language model through a token masking process, and supervised "fine-tuning" of a model with additional classification architecture . Within the clinical domain, several publicly available Transformer models have been trained to generate useful document embeddings for such tasks (Gururangan et al. 2020), incorporating the unique biomedical corpora within data such as electronic health records, as well as used for downstream tasks.

However, the utility of existing models is often limited in clinical natural language processing. Electronic health records, such as reports, discharge summaries, and nursing notes, are often a sufficient and necessary feature for natural language tasks, such as International Classification of Disease (ICD) coding. These documents contain all relevant information for humans to assign medical codes, yet to date are not well suited as training instances for deep feed-forward networks. This is because such documents are rich and expressive but can be incredibly lengthy. The average tokenized length of an inpatient discharge summary, for instance, can number in the many thousands. Simultaneously, however, classic techniques of natural language instance feature regularization are ill-suited; reduction of document size via the omission of classic stop-words corpora, for instance, would also be an omission of valuable semantic context specific to this domain and associated tasks.

As such, recent work with deep models such as Transformers in the clinical domain has shown to be impressive, though largely involves ad-hoc, idiosyncratic pre-processing (Zhang, Liu, and Razavian 2020; Alsentzer et al. 2019). For instance, clinical document instances are often truncated and/or chunked when used as inputs to existing transformer language models, such as Bidirectional Encoder Representations from Transformers (BERT), sometimes down to tokenized lengths as small as 128 (Alsentzer et al. 2019).

Not only is the token ceiling (512 for BERT, usually) a poor match for clinical documentation datasets from a logistical perspective, but training in such a fashion likely results sub-optimal convergence in any downstream classification task. Specifically, the attention mechanism of these models would ignore any global semantic dependencies present in large instances. Such global information may be critical for tasks in the medical field, such as medical coding, where human workers often must carefully parse long document instances in their entirety to reconstruct a patient encounter and assign relevant code(s).

In this project[1], we present a state-of-the-art encoding model for practical use in the clinical domain. To generate our novel bioclinical-roberta-long model, we pre-train an existing RoBERTa (A Robustly Optimized BERT Pretraining Approach) model checkpoint (allenai/biomed_roberta_base) on the concatenated corpora of two massive two massive clinical natural language datasets (MIMIC-III and MIMIC-CXR). We then convert this pretrained RoBERTA encoder to a Long-Document Transformer , or Longformer (Beltagy, Peters, and Cohan 2020), adding a novel 4096-token $\mathcal{O}(n)$ global attention mechanism per “Longformer” specifications via sliding $\mathcal{O}(n^2)$ 512-token local attention. This allows for a linear scaling of the Transformer across the large document instances often useful for clinical machine learning.

Next, using our pre-trained Longformer as an encoder, we train classifiers end-to-end on the MedNLI natural language inference task against existing models and propose a new whole-document medical coding task to test whole-document embedding.

With regard to the latter task, we train classifiers end-to-end on whole-document instances of discharge summaries from MIMIC-III, with labels as the primary diagnostic ICD-9-CM code assigned for that particular patient visit. We aim to emulate the work of a human medical coder using natural language inference and linear classifiers. Using this pipeline, we benchmark our encoder bioclinical-roberta-long against three existing models end-to-end, including our pre-training checkpoint (bert-base-uncased, Bio_ClinicalBERT, allenai/biomed_roberta_base). We propose this task as a useful benchmark for whole-document tasks in the clinical field.

Finally, we make available a fork of the Taming Pretrained Transformers for eXtreme Multi-label Text Classification (X-Transformer) repository (Chang et al. 2020), modified for use with MIMIC-III datasets and our associated transformers for future study of automated ICD coding.

2 Dataset

2.1 MIMIC-III

’Medical Information Mart for Intensive Care III’ (MIMIC) is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital (A. Johnson, Pollard, and Mark 2020). It consists of de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Over the summer of 2020, a project team member completed the required data usage training course and worked extensively over the summer of 2020 on the OctoberChang/X-Transformer fork of this project.

For the purposes of this project, we use the entirety of the note events table (about 2 million clinical notes and physician studies in total) as one of the two concatenated corpora for masked language modeling. For the purposes of ICD-label classification, we subset the dataset on the 50 most frequent ICD codes[2] present with a corresponding hospital discharge summary as the feature. This filtering was done since we are mainly interested in evaluating the masked language modeling of our Longformer, though we highly encourage further exploration of the extreme multilabel problem of ICD code assignment from natural language features.

2.2 MIMIC-CXR

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 (A. E. W. Johnson et al. 2019 is a large publicly available dataset of chest radiographs, and crucially for our case with free-text radiology re-

[1] All code is view able at on our GitHub repository, github.com/simonlevine/clinical-longformer

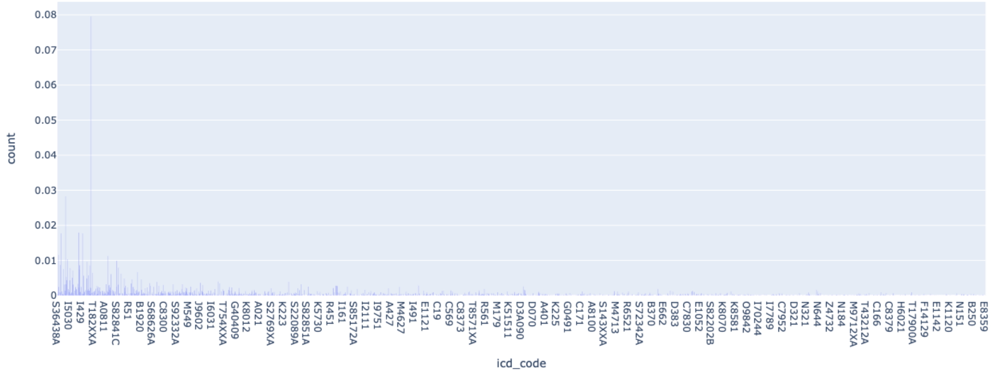
[2] Selected codes for classifier training and inference included: ’41401’, ’0389’, ’41071’, ’V3001’, ’4241’, ’51881’, ’V3000’, ’V3101’, ’431’, ’4240’, ’5070’, ’4280’, ’41041’, ’41011’, ’5789’, ’430’, ’486’, ’1983’, ’99859’, ’5849’, ’5770’, ’43491’, ’5712’, ’99662’, ’4271’, ’03842’, ’99811’, ’V3401’, ’42731’, ’56212’, ’4373’, ’43411’, ’4321’, ’41519’, ’51884’, ’85221’, ’570’, ’03811’, ’53140’, ’03849’, ’4412’, ’42823’, ’44101’, ’42833’, ’0380’, ’85220’, ’4210’, ’4414’, ’51919’, ’5715’

ports. The dataset contains 227,835 radiographic free-text reports corresponding to studies performed at the Beth Israel Deaconess Medical Center in Boston, MA.

We concatenate all notes to those in MIMIC-III for the purposes of language modeling only. Though a radiology practice-specific ICD classifier based on these notes alone is a compelling prospect, we feel this data alone is a poor fit, likely being of insufficient size for masked language modeling and as well representing patient visits with highly variant and sparse ICD coding.

For the purposes of this project, we leverage the free-text physician study notes only, ignoring radiographic images, though we would encourage future study to ensemble these data. To this end, we include a Jupyter notebook linking MIMIC-CXR patient visits and all associated radiographic and text features to their associated MIMIC-IV ICD-10-CM ("clinical modification", or diagnostic codes) in our repository.

Figure 1: Diagnostic ICD codes associated with MIMIC-CXR are highly variant and sparse



2.3 MIMIC-IV

MIMIC-IV (A. Johnson, Bulgarelli, et al. 2020), an update to MIMIC-III, incorporates contemporary data and improves on numerous aspects of MIMIC-III, such as using the now-standard ICD-10 code set. This dataset is currently on a rolling release and free-text data was unavailable at the time, but this dataset would easily be incorporated into our pipeline at a future date. We thus use MIMIC-IV solely for the provision of the MIMIC-CXR to ICD-10 code assignment notebook previously mentioned. We suggest masked language modeling using this corpus once released.

2.4 Preprocessing

We preprocessed our corpus minimally. Specifically, we remove MIMIC-specific de-identificatio tokens, administrative language [3], and make slight spelling and grammar adjustments to ensure maximal overlap with the corpus of the checkpointed tokenizer used. For instance, we convert all instances of "w/" to "with", remove trailing spaces, remove dashes and pound symbols, and so forth.

2.5 Benchmarks

We present benchmarks of our model against baselines using the MedNLI natural language inference task, the MIMIC-III-based diagnostic ICD prediction.

[3] removed administrative language included `"Admission Date"`, `"Discharge Date"`, `"Date of Birth"`, `"Phone"`, `"Date/Time"`, `"ID"`, `"Completed by"`, `"Dictated By"`, `"Attending"`, `"Provider:  "`, `"Provider"`, `"Primary"`, `"Secondary"`, `" MD Phone"`, `" M.D. Phone"`, `" MD"`, `" PHD"`, `" X"`, `" IV"`, `" VI"`, `" III"`, `" II"`, `" VIII"`, `"JOB"`, `"JOB: cc"`, `" Code"`, `" x"`, `" am"`, `" pm"`, etcetera.

2.5.1 MedNLI

MedNLI (Shivade 2017) a dataset derived from MIMIC-III annotated by physicians, performing a natural language inference task (NLI), grounded in the medical history of patients. In this case, a sentence from a past medical history of the patient is provided, as well as a sentence from an annotating expert asked to:

“

- *Write one alternate sentence that is definitely a true description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient suffers from a chronic condition".*
- *Write one alternate sentence that might be a true description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient has hypertension".*
- *Write one sentence that is definitely a false description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "The patient’s insulin levels are normal without any medications".*

”

As such, the labels for this task are one and only one of “entailment”, “contradiction”, or “neutral”. We tokenize each concatenated sentence pair and attempt to predict the associated label using our Longformer and a linear classification head.

It should be noted here that the training, testing, and validation instances here are derivative of MIMIC-III, and thus there are likely a small number of patients that are represented in these data in the masked language modeling step. However, we follow the example of ClinicalBERT in noting that this is likely insufficient to cause significant overfitting (Alsentzer et al. 2019).

2.5.2 ICD Labeling

It is difficult to benchmark directly against current state of the art models for previously-devised ICD tasks as:

1. the data and/or data pre-processing steps are altered or proprietary (such as in Zhang, Liu, and Razavian 2020)
2. whole - instance ICD classification in the clinical domain has, to the best of our knowledge, not been attempted.

So, for this task, we construct a simple but extensible ICD classifier using categorical labels on the subset of MIMIC-III discharge summary notes previously described.

An example[4] of the ICD classification task is to infer the following class label from the document instance below:

X:

```
{PATIENT; .........D.O.B:........; CHI: ..........  Admission:  Specialty
-....; Ward - xxConsultant:  Date of Admission - xx/xx/xxxxDate of
Discharge - xx/xx/xxxx; Discharged to:  [ .  .  ]Follow Up:  [ ] Clinical
Comments:  Diagnosis:  Musculoskeletal chest pain Ischaemic heart
diseaseType II diabetes mellitus Hypertension Previous CVAObesityThis
[..]  year old woman was admitted with a complaint of recurrent [chest
pain].  There is a background of ischaemic heart disease with previous
[.  .  ]  myocardial infarction and [.  .  .  ]Other history is of
hypertension, cerebral vascular disease, type II diabetes mellitus and
obesity.  Cardiac examination [ .  .  ]  ECG showed sinus rhythm with old
[.  .  ]  infarction.  There were no sequential changes and troponin was
not raised.  I felt that her symptoms were consistent with musculoskeletal
origin.  [ .  .  ].  Yours sincerely, Dr [...]},
```

[4] www.isdscotland.org

y:

```
{786.59 - Chest pain (unspecified) (central) (includes chest discomfort,
pressure, and tightness}
```

3 Methods

All code is housed on GitHub and contains all preprocessing steps and classification steps. We use Pytorch and Huggingface third-party packages for organizing deep learning code and as an API for loading and hosting Transformer models, respectively.

3.1 Architecture

We leverage Transformers to embed clinical text into usable features for downstream classification. We then build classification pipelines for ICD code prediction and the MedNLI task.

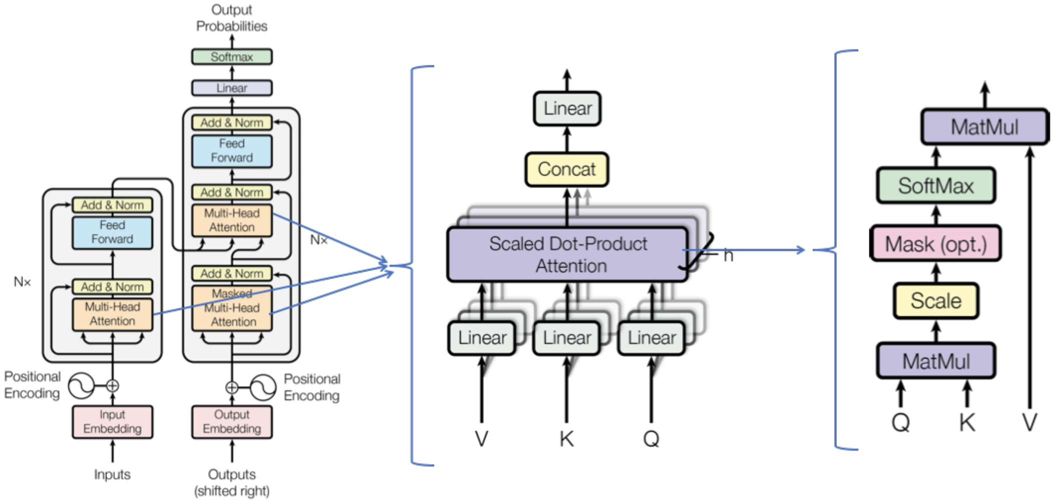


Figure 2: The original Transformer self-attention mechanism as illustrated in Vaswani et al. 2017.

A pre-trained checkpoint of the RoBERTa transformer model is the basis for our Longformer used in both masked language modeling and the two classification tasks. The RoBERTa model embeds text via byte-pair encoding (BPE), a sub-word hybrid of word-level and character-level token representation.

Apart from this, RoBERTa as a derivative of BERT is distinguishable as it is trained with dynamic masking, full-length sentences (or clinical text instances, in our case), no next-sentence prediction (NSP), and larger mini-batches.

The model checkpoint we use, BioMed-RoBERTa-base, achieves state of the art performance on various tasks in the biomedical domain[5] Prior to our use, the model was trained on a 7.55 billion token, 47GB corpus, derived from 2.68 million instances of Semantic Scholar literature.

Additionally, RoBERTa generally outperforms BERT. Thus, rather than elongating a pre-trained clinical or biomedical BERT model with global attention, we use BioMed-RoBERTa-base.

Finally, our elongation process re-instantiating a RoBERTa class with the global attention mechanism as described by Beltagy, Peters, and Cohan 2020. While the original Transformer model (Vaswani et al., 2015) uses Attention$(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$, a $\mathcal{O}(n^2)$ self-attention mechanism for a given instance of length n, the Longformer self-attention matrix over the entirety of a given instance is regularized for sparsity. That is, two sets of key, value, and query vector projections are computed: $(Q_s, K_s, V_s) \propto$ sliding local attention, and $(Q_g, K_g, V_g) \propto$ attention scores for the global instance.

[5] biomedical benchmark tasks for the pre-trai ned biomed-roberta-base checkpoint included RCT-180K, ChemProt, JNLPBA, BC5CDR, and NCBI-Disease, as of writing this.

This results in a linear scaling to sequences longer than the original self-attention window (512 in this case), such that masked language modeling and subsequent tasks on large instances are far more tractable.

3.2 Masked Language Modeling

3.2.1 Preparation and Assumptions

We use the popular HuggingFace API of Transformer models to load the pre-trained `allenai/biomed_roberta_base` model checkpoint, as well as the baseline `bert-base-uncased` and `Bio_ClinicalBERT` models.

For `allenai/biomed_roberta_base` *only*, we implement a naive masked language modeling pipeline. Specifically, all 2.3 million free-text notes in MIMIC-III and MIMIC-CXR are concatenated into a single string and the RoBERTa model encoder is trained on tokenized samples. Note that the checkpointed encoder distribution includes an associated "biomed" tokenizer, and in consideration of our document pre-processing steps, we did not bother to address any vocabulary exclusive to MIMIC-III or MIMIC-CXR versus the tokenizer.

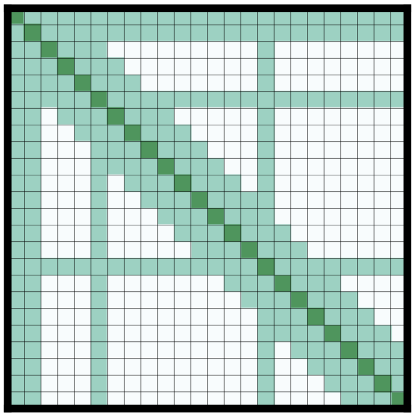


Figure 3: The global and sliding local self-attention mechanism of the Long-Document Transformer architecture, as illustrated in Beltagy, Peters, and Cohan 2020. Local $\mathcal{O}(n^2)$ attention was set to 512 tokens, and $\mathcal{O}(n)$ global attention was set to 4096 given a tokenized instance of length n.

Additionally, in performing masked language modeling on essentially one immense string object, we did not separate masked language modeling instances in a document-wise fashion. However, the original RoBERTa paper authors make a compelling case for just such a naïve approach [6] by doing away with Next Sentence Prediction (NSP) loss [7].

3.2.2 Training

The RoBERTa model[8] was pre-trained on the concatenated corpora for 1000 epochs[9]. Training involved the following hyperparameters using the Huggingface Trainer API: 500 warmup steps, a batch size of 8, gradient checkpointing enabled, a learning rate of 3e$-$5, floating-point 16 precision, the Adam optimizer, a maximum gradient normalization of 5, and 32 gradient accummulation steps. This largely follows from Huggingface masked language modeling tutorials.

After pre-training, the biomed-RoBERTa object was re-instantiated with the 4096-token global attention mechanism from the Longformer repository in order to accommodate the large number of tokens for our classification task. [10]

[6] RoBERTa: A Robustly Optimized BERT Pretraining Approach

[7] *" NSP is a binary classification loss for predicting whether two segments follow each other in the original text. Positive examples are created by taking consecutive sentences from the text corpus. Negative examples are created by pairing segments from different documents. Positive and negative examples are sampled with equal probability. "*

[8] Note that this stage still involved a maximum attention window of 512 tokens. This was a matter of circumstance, as attempting pretraining on an elongated model with global attention at 4096 tokens resulted in out-of-memory errors on our 187 gigabyte machine (even with a batch size of 1, gradient checkpointing, FP16, etc.), and incredibly long training estimates.

[9] terminated after approximately 4 days on an AWS EC2 `g4dn.12xlarge` instance, a quad Nvidia T4 machine

[10] This model, as with all other used, is available for use at the Huggingface model zoo.

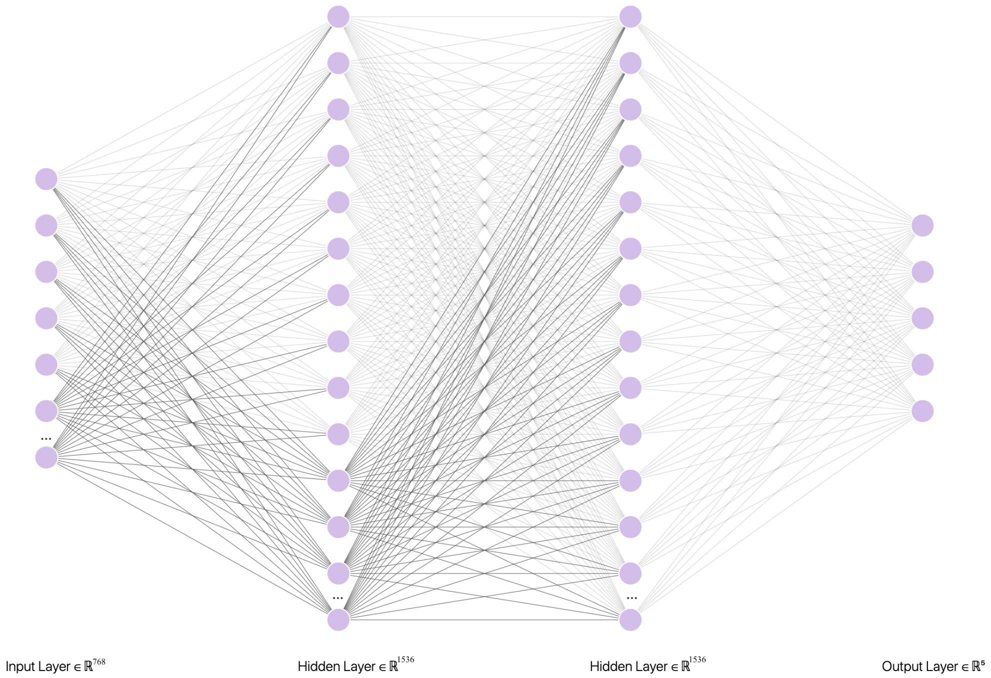
3.3 Classification

We test two approaches to classification using our novel Longformer and benchmark Transformers.

3.3.1 End-to-End Automated Medical Coding

In our first approach, we train each of the four encoders with a sequence classification head on top. A 3-layer fully connected multilayer perceptron takes in the pooled output (768 dimensions) of the Transformer models. Each internal layer is 1536 neurons, and each layer uses hyperbolic tangent activation. Cross-entropy loss is used as a loss criterion in backpropagation, where we attempt to model the minimal divergence between predicted ICD code and the ground-truth in MIMIC-III.

Figure 4: The multilayer perceptron classifier trained using Transformer embeddings as features and primary diagnosis ICD codes as labels. Biases not shown.


...
...
...
Input Layer $\in \mathbb{R}^{768}$
Hidden Layer $\in \mathbb{R}^{1536}$
Hidden Layer $\in \mathbb{R}^{1536}$
Output Layer $\in \mathbb{R}^{5}$

Again, in all cases, we train on the discharge summaries and associated primary ICD-9-CM labels corresponding to the fifty most frequent codes present in MIMIC-III. The final data processing step in this case was to set aside 80% of about 14,000 total discharge summary examples for training, 15% for validation during development, and 5% for testing.[11]

For all models, we trained for 10 epochs end-to-end (unfrozen encoder) with an encoder learning rate or $1e{-}5$, a linear classifier learning rate of $3e{-}5$, 2 batches of gradient accumulation prior to backward passes, and an Adam optimizer for both encoder and multilayer perceptron learning.

In the case of our Longformer, we employ a maximum token length of 4096, where a 512-token local attention window slides across each instance. We truncate each instance longer than this maximum length and pad those that are shorter. Gradient checkpointing and a batch size of 2 was used for the Longformer and 12 otherwise. Otherwise, training hyperparameters are equivalent, such that all encoders but our Longformer only use the first 512 tokens of each instance.

[11] Note that we provide preprocessing logic suited to but untested on the procedural ICD code set, since diagnostic coding is hypothetically entirely contingent upon discharge summaries and not separate operative reports.

Figure 5: Label Legend

🟥 allenai/bio_med_roberta_base	🟧 bert-base-uncased
🟦 emilyalsentzer/Bio_ClinicalBERT	🟦 **simonlevine/bioclinical-roberta-long**

Figure 6: Training loss as a function of time. Note the local instability but global convergence in all cases, with Longformer taking significantly longer due to lower batch sizes and general complexity.

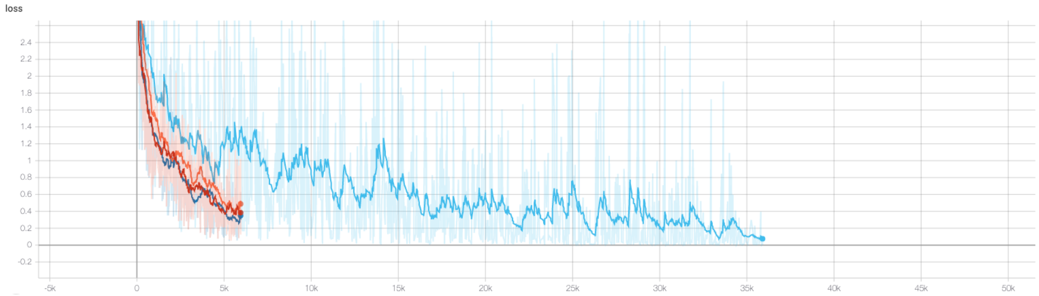
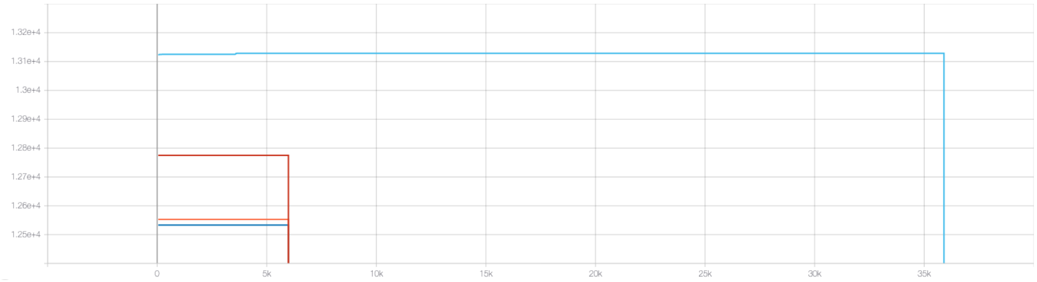


Figure 7: Memory usage (Mb) over time



3.3.2 Medical Natural Language Inference (MedNLI)

MedNLI is a sequence level task. In Kanakarajan et al. 2019, model pre-training is conducted on the binarized next sentence prediction. We employ the same mechanism with all our models and compare its performance with each other. The objective, as stated before, is to predict how the first sentence is related to the second sentence in terms of entailment, contradiction or neutral.
There are three major occurrences to take note of after fine tuning our models on the MedNLI task:

- The attention is distributed all over the second sentence as opposed to majority of the attention focused on the delimiting token of the second sentence.
- Attention is greater on words similar to the source word in the preceding sentence
- Before fine tuning, OOV (Out Of Vocabulary) words are split into multiple tokens that do not receive attention. After training, such tokens may now be

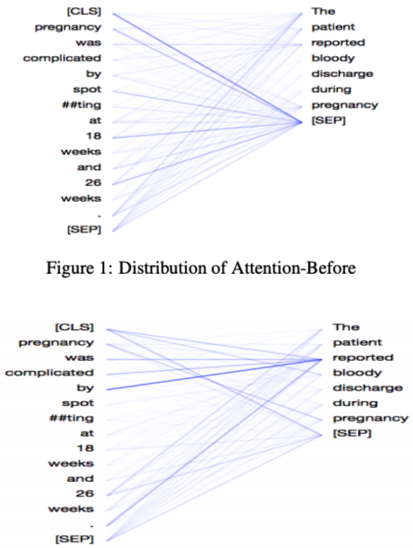


Figure 1: Distribution of Attention-Before

Figure 2: Distribution of Attention-After

Figure 8: Attention

the focus of stronger attention as there is a strong attention flow between tokenized words across two sentences.

4 Results

4.1 Classification

We report class-averaged accuracy, F1, recall, precision, and final loss scores for each of the classification tasks. We use these benchmarks as they are customary and follow previous literature.

4.1.1 End-to-End Automated Medical Coding

The results of the ICD coding task on validation and test data subsets are as follows. Note that final loss is given, and all other metrics are computed as weighted class averages.

Validation	bert-base-uncased	Bio_ClinicalBERT	biomed_roberta_base	**bioclinical-roberta-long**
Loss	1.3842	1.3450	1.3297	**1.2189**
F1	0.6560	0.6710	0.6824	**0.7184**
Precision	0.6817	0.6941	0.7114	**0.7216**
Recall	0.6563	0.6733	0.6791	**0.7183**
Accuracy	0.6563	0.6733	0.6791	**0.7183**

Test	bert-base-uncased	Bio_ClinicalBERT	biomed_roberta_base	**bioclinical-roberta-long**
Loss	1.3429	1.3401	1.2950	**1.1927**
F1	0.6556	0.6636	0.6784	**0.7219**
Precision	0.6789	0.6900	0.7050	**0.7260**
Recall	0.6586	0.6655	0.6773	**0.7214**
Accuracy	0.6586	0.6655	0.6773	**0.7214**

4.1.2 MedNLI

Test	F1 Scores	Precision	Weighted Accuracy	Recall	Loss
Allen_ai/biomed_RoBERTa_base	0.8253	0.8484	0.827	0.827	0.7997
BERT-base-uncased	0.7596	0.7933	0.7651	0.7651	1.169
Bio_ClinicalBERT	0.8011	0.8305	0.8059	0.8059	0.9654
Clinical_longformer	0.8291	0.8537	0.8333	0.8333	0.8594

Validation	F1 Scores	Precision	Weighted Accuracy	Recall	Loss
Allen_ai/biomed_RoBERTa_base	0.8367	0.8484	0.8618	0.8394	0.6721
BERT-base-uncased	0.7731	0.7933	0.8086	0.7842	0.9263
Bio_ClinicalBERT	0.8156	0.8305	0.8492	0.8265	0.8694
Clinical_longformer	0.8361	0.8537	0.8563	0.8373	0.7446

5 Discussion

5.1 Masked Language Modeling

It is difficult to discern the effect of masked language modeling directly on the model performance for practical use. As such, we lack quantitative benchmarks to compare the pre-trained (but not fine-tuned) model and instead point to our classification benchmarks as a proxy for pre-training performance. Additionally, we feel this step is less important given that instantiating global attention was only done post pre-training for practical reasons as was previously mentioned.

5.2 Classification

5.2.1 Automated Medical Coding

We achieve state of the art performance relative to the most promising alternative models available. This is attributable mainly to our novel use of global attention in this context as well as the superior base RoBERTa architecture. Unlike past explorations, we are able to deliberately preprocess clinical notes as minimally as possible, assuming only that hospital encounter discharge summaries ought to contain all relevant information for coding at least, but not limited to, the primary diagnostic ICD-9 code in this dataset.

As for the other models, there are some valuable insights. Firstly, it is clear that the corpus in `base-bert-uncased` is clearly a performance bottleneck, as the pre-trained `Bio-Clinical-BERT` has shown both by its authors and in this benchmark. Secondly, RoBERTa appears to outperform BERT in a significant way. `Bio-Clinical-BERT` was pre-trained on MIMIC-III by its creators from the `bio-BERT` checkpoint (Alsentzer et al. 2019), whereas `biomed-roberta-base` was pre-trained on an extensive but far less specific corpus of biomedical literature.

In other words the architecture of RoBERTa is sufficient to outperform a highly domain-specific pre-trained BERT model even without pre-training. This remained a guiding assumption in choosing a RoBERTa for elongation with global attention, as our "Longformer" (`bioclinical-roberta-long`) benefits not only from a larger global window but a more robust local window, at least during end-to-end fine-tuning.

Next, as was shown by Alsentzer et al. 2019 and after comparing BERT-base to the other models in our tests, we note the small but significant impact a domain-specific corpus has on model performance (BERT-base has a much less specific tokenizer), such that biomedical and clinical tokens are present in the clinical model vocabulary.

We also note that our accuracy and recall results are identical for all models. This is likely a function of our subset of data, whereby we have good balance of examples of each class. These metrics indicate that sensitivity is equal to specificity on the ICD task, and thus accuracy is the same as recall.

In other words, selecting the 50 most frequent ICD codes has caused each model correctly classify positive samples at the same rate as it would correctly classify negative samples. This also explains how precision can be variant. We suggest future investigation to leverage more sparse and variegated code sets to challenge the model further.

5.2.2 Natural Language Inference (MedNLI)

We can infer from Figures 8,9,10 and 11, the performance of the four models including our own model, is nearly uniform on the MedNLI task. These results are not surprising considering the smaller token size (instances less than 512 tokens) prevalent in this task. With the smaller tokenized instance size, the global attention mechanism in our Longformer model is not leveraged and does not contribute a significant improvement in comparison to the other models considered. One somewhat surprising aspect of these uniform results is that the biomedical corpus – present in a broad sense in the BioClinical-BERT, Biomed-Roberta-base, and biomed-Roberta-long (Clinical-Longformer) – did not have a significant effect on classification performance in the MedNLI task.

Figure 9: Biomed RoBERTa Confusion matrix for the MedNLI task

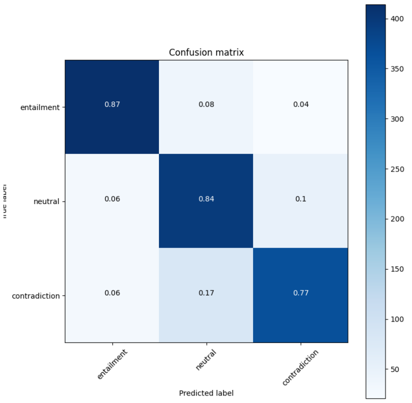


Figure 10: BERT base case uncased Confusion matrix for the MedNLI task

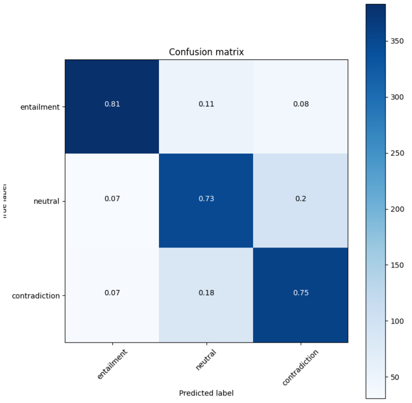


Figure 11: Bioclinical BERT for the MedNLI task

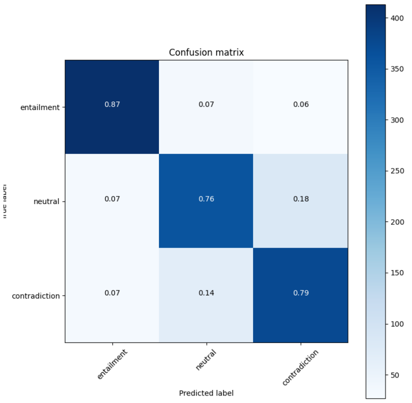
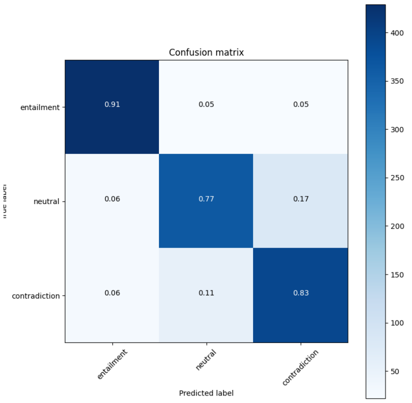


Figure 12: Clinical-Longformer Confusion matrix for the MedNLI task



6 Future Directions

6.1 Automated Medical Coding

Ultimately, we feel that our encoder is an excellent basis for a domain-specific classifier to solve one of the biggest inefficiencies in healthcare administration today: the problem of assigning any and all relevant ICD codes to novel patient visit documentation. We feel this falls under the category of extreme multilabel classification – as there are over 150,000 combined diagnostic and procedural codes in the International Classification of Disease handbook, 10th edition – and each patient encounter can have one or more labels, and they may be ordered by clinical and administrative importance. Again, this project focuses only on the *primary* diagnostic code, and only those (top 50) with high frequency in the dataset.

One suggested approach comes from the realm of information retrieval via deep neural networks. For this, we suggest modeling the extreme multilabel ICD classification problem through pure natural-language, as each ICD code is associated with a free-text *label* in addition to any features of clinical instances. As was previously mentioned, we provide a repository[12] for just this purpose.

We do not attempt training on MIMIC-III using this framework at this time. This is because, despite its relatively large size, the dataset is highly unlikely to contain all possible combinations of ICD codes across all patient visits in all possible orderings. There may be dozens of codes assigned for a particular visit.

If we assume naively for the sake of argument that each patient has exactly 12 associated ICD codes, ordered by importance, and there are $\sim 13,000$ potential ICD-9-CM codes in the dataset, that verges on a combinatorial permutations-with-replacement problem considering the size of the dataset. For a single patient in our base case, we have an extremely large number of permutations for a classifier to learn, let alone a dataset of the appropriate size:

$$N_{PotentialCodeAssignments} = P^R(n,r) = n^R = 13,000^{12} \approx 2.33 * 10^{49}$$

However, a smaller, more practice-specific dataset could make use of this framework.

We finally suggest some directions to overcome the inherent shortcomings of “flat” categorical classifiers. As was mentioned, ICD codes effectively have two representations: a free-text description similar to other extreme multilabel classification datasets (useful in X-Transformer), and also a 4-7 character hierarchical alphanumerical “code”. While previous work has attempted to use categorical label encodings (such as this project), or ICD label free-text with BERT and multilabel attention mechanisms (Zhang, Liu, and Razavian 2020, and X-Transformer-ICD), we feel the ICD *code*, rather than its description, is likely a richer learning point due to its explicit hierarchical nature. These codes in can be thus be processed into a hierarchical embedding (Mao et al. 2019), ideally to reduce the label space adequately.

[12] simonlevine/x-transformer-icd

We therefore suggest extending our Longformer to classify a hierarchical or graphical embedding of ICD labels (ICDCodex ; Fisher 2020)[13], Node2Vec, DeepWalk, etcetera). In the latter case, for instance, a regression head atop the Transformer logits could predict a continuous vector of a predicted graphical embedding of ICD code(s) given the embedding from Transformer encoder. It would then be feasible a nearest-neighbors search over this vector to find the most probable label. Then, the most probable label could be re-discretized (or pre-generated in a lookup dictionary of graphical embeddings) yielding a result to the extreme multilabel classification problem using the hierarchical information of ICD codes. We feel our categorical-label classifier could serve as a useful a baseline for this future work.

It may also be prudent for future investigators to concatenate, ensemble, or rank these label representations.

7 Division of Work

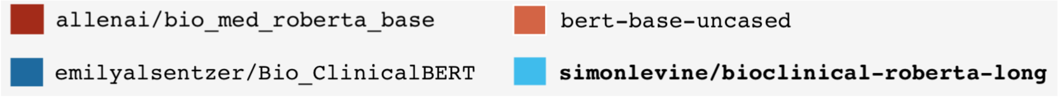
Team members both shared in the literature review/research process as well as the plan for implementation and benchmark tasks. Simon completed the PhysioNet training program in order to access the dataset and did exploratory analysis on MIMIC-III and MIMIC-CXR. Team members then pair-programmed MIMIC-III preprocessing steps. Simon implemented an original MLP classifier, and the team pair-programmed subsequent adaptation to both MedNLI and ICD tasks. Simon ran the MLM and ICD classifier training and collected benchmark statistics and logging. Similarly, Serena ran the MedNLI training and collected benchmark statistics and logging. Team members authored this document with mostly equal contribution.

[13] github.com/icd-codex

8 Appendix

8.1 Additional Training Plots

Figure 13: Label Legend



🟥 allenai/bio_med_roberta_base
🟧 bert-base-uncased
🟦 emilyalsentzer/Bio_ClinicalBERT
🟦 **simonlevine/bioclinical-roberta-long**

Figure 14: Validation accuracy, class-averaged

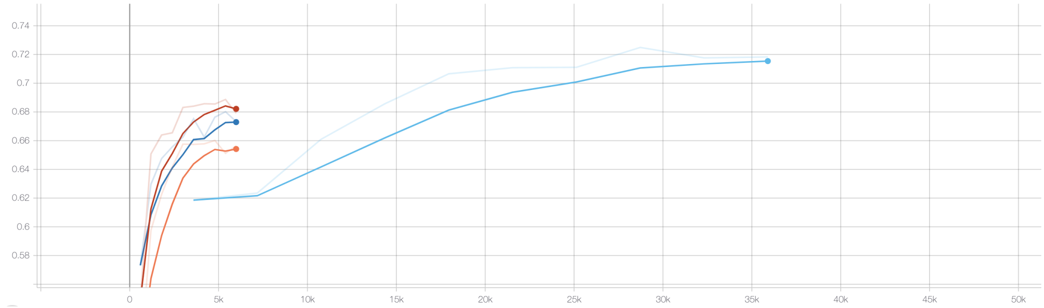


Figure 15: Validation precision as a function of time.

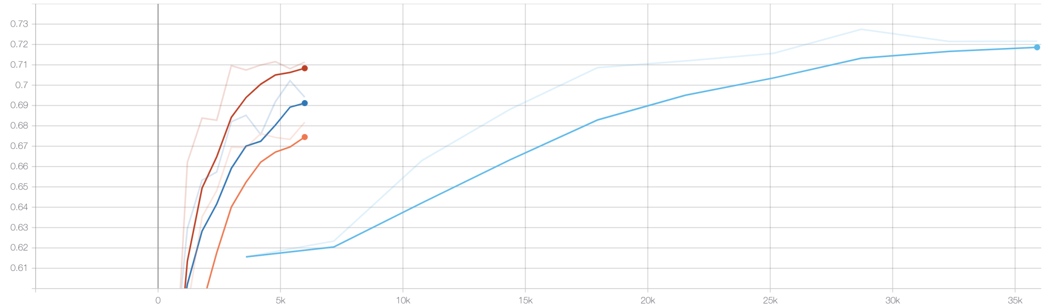


Figure 16: Validation loss, class-averaged, as a function of time.

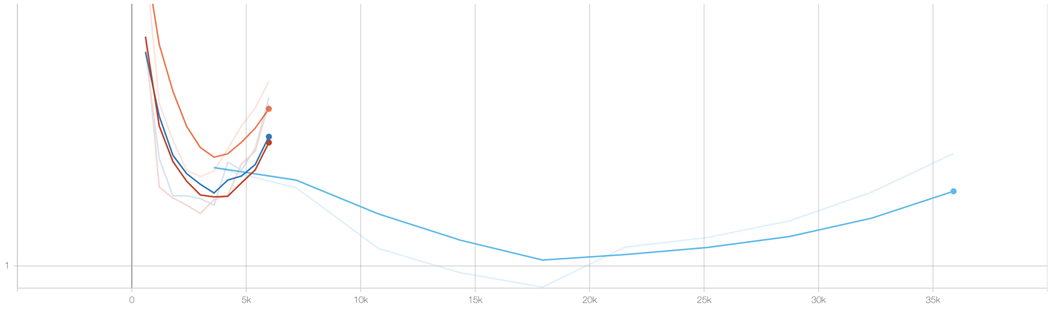


Figure 17: Validation recall, class-averaged, as a function of time.

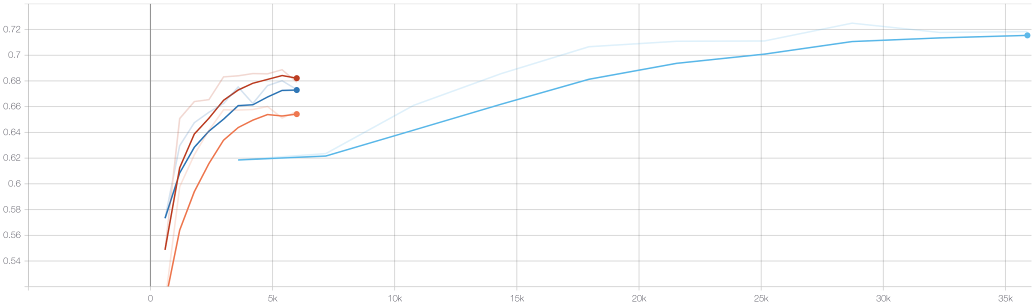
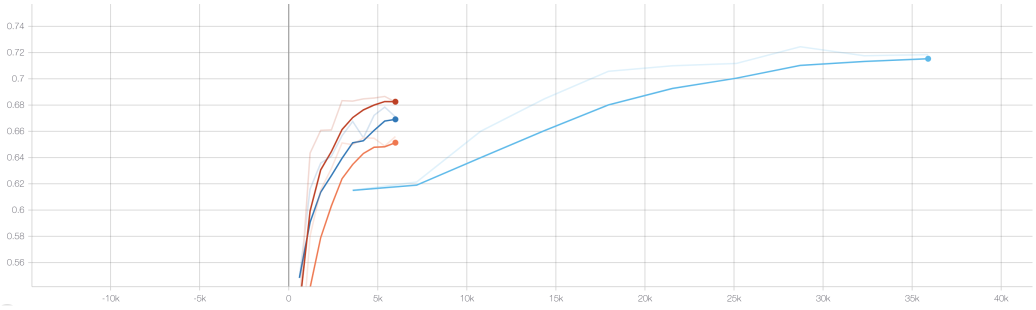
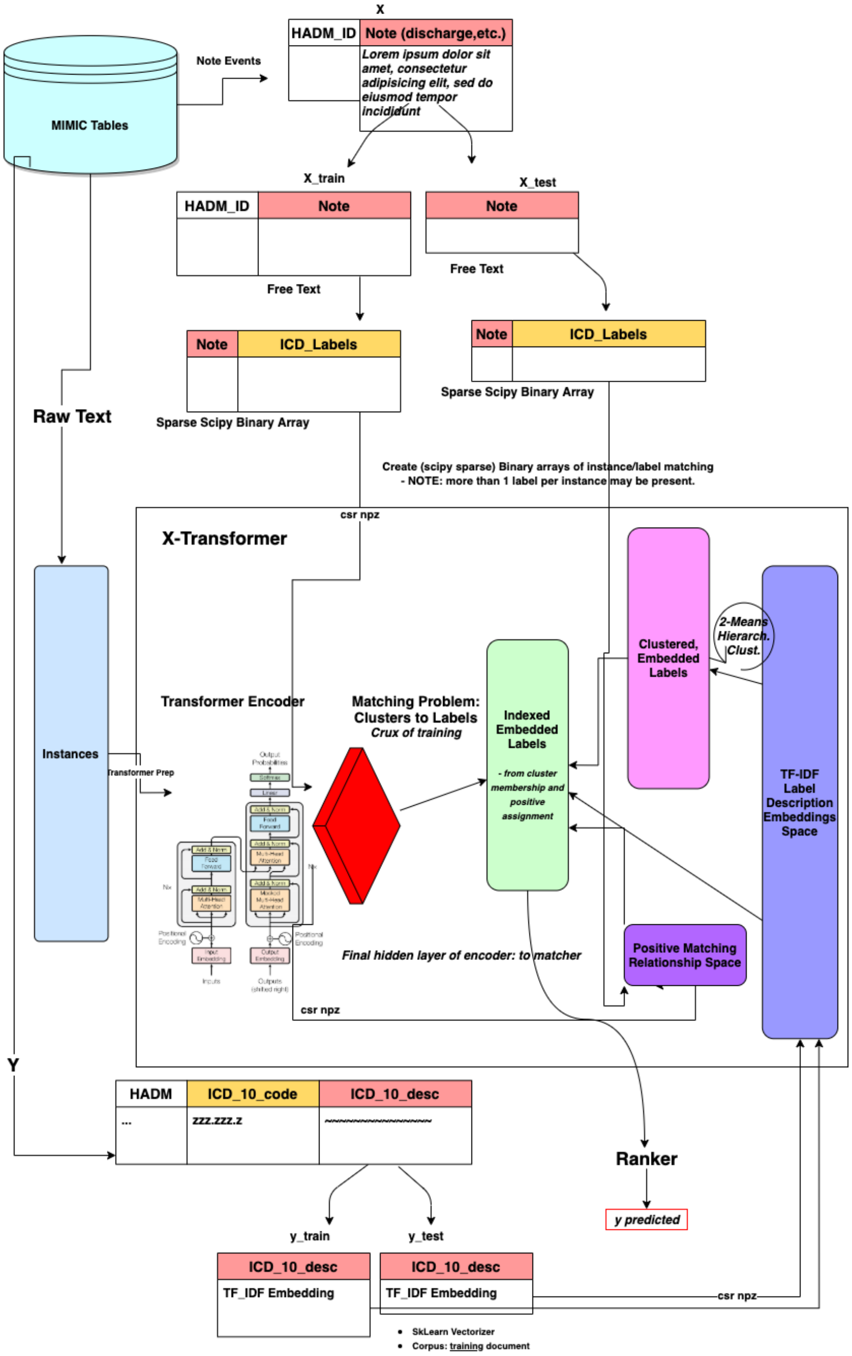


Figure 18: Validation F1, class-averaged, as a function of time.



8.2 X-Transformer-ICD

Figure 19: Any of the encoders used fit neatly into the X-Transformer-ICD pipeline for complete automated ICD coding at scale



References

[Als+19] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://www.aclweb.org/anthology/W19-1909.

[BPC20] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL].

[Cha+20] Wei-Cheng Chang et al. *Taming Pretrained Transformers for Extreme Multi-label Text Classification*. 2020. arXiv: 1905.02331 [cs.LG].

[Fis20] Jeremy Fisher. *icd-codex/icd-codex v0.4.5*. Version v0.4.5. Dec. 2020. DOI: 10.5281/zenodo.4300935. URL: https://doi.org/10.5281/zenodo.4300935.

[Gur+20] Suchin Gururangan et al. *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. 2020. arXiv: 2004.10964 [cs.CL].

[Joh+19] Alistair E. W. Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6.1 (Dec. 2019). DOI: 10.1038/s41597-019-0322-0. URL: https://doi.org/10.1038/s41597-019-0322-0.

[Joh+20] Alistair Johnson, Lucas Bulgarelli, et al. *MIMIC-IV*. 2020. DOI: 10.13026/A3WN-HQ05. URL: https://physionet.org/content/mimiciv/0.4/.

[JPM20] Alistair Johnson, Tom Pollard, and Roger Mark. *MIMIC-III Clinical Database*. 2020. DOI: 10.13026/C2XW26. URL: https://physionet.org/content/mimiciii/1.4/.

[Kan+19] Kamal Raj Kanakarajan et al. "Saama Research at MEDIQA 2019: Pre-trained BioBERT with Attention Visualisation for Medical Natural Language Inference". In: (2019). DOI: 10.18653/v1/W19-5055. URL: https://www.aclweb.org/anthology/W19-5055/.

[Mao+19] Yuning Mao et al. "Hierarchical Text Classification with Reinforced Label Assignment". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019). DOI: 10.18653/v1/d19-1042. URL: http://dx.doi.org/10.18653/v1/D19-1042.

[Shi17] Chaitanya Shivade. *MedNLI — A Natural Language Inference Dataset For The Clinical Domain*. 2017. DOI: 10.13026/C2RS98. URL: https://physionet.org/content/mednli/.

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[ZLR20] Zachariah Zhang, Jingshu Liu, and Narges Razavian. *BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining*. 2020. arXiv: 2006.03685 [cs.IR].