Ontology Mediated Attention for Biomedical Documents

Aditya Dave
Univ. of S. California
daveadit@usc.edu

Alexander Chang
Univ. of S. California
changam@usc.edu

Robert Steele
Univ. of S. California
rjsteele@usc.edu

Slava Zinevich
Univ. of S. California
zinevich@usc.edu

1 Introduction

The transformer (Vaswani et al., 2017) has improved state-of-the-art across numerous sequence modeling tasks. However, its effectiveness comes at the expense of a quadratic computational and memory complexity with respect to the sequence length, hindering its adoption. Recently, researchers have directly addressed the Transformer's limitation by designing lower-complexity alternatives such as the Longformer (Beltagy et al., 2020), Reformer (Kitaev et al., 2020), Linformer (Wang et al., 2020), and Performer (Choromanski et al., 2020). Domain specific data has also been shown to be essential to achieve state-of-the-art results in language related to a specific language domains. In the biomedical domain language models, Clinical-Longformer and Clinical-BigBird (Li et al., 2022) which are based on Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) respectively have proven to outperform Clinical-BioBERT (Alsentzer et al., 2019) and other short-sequence-transformers in all downstream tasks. We hope to further build on this by using biomedical knowledge bases like UMLS (Unified Medical Language System) (Bodenreider and Olivier, 2004). We will be using QuickUMLS (Soldaini and Goharian, 2016) for it, building an informed attention mechanism.

2 Methods

2.1 Datasets

For this study we are relying on two different sources of data. The first is our source of training data, MIMIC-III, a de-identified dataset of doctors notes (Johnson et al., 2016). The second will be used for analyzing performance, and will be provided by the i2b2 challenge (Sun et al., 2013).

MIMIC-III is a large clinical dataset comprised of documents covering patients admitted to the intensive care unit at the Beth Israel Deaconess Medical Center. It covers over 50,000 admissions with over 2 million total documents. It is a common pretraining dataset, having previously been used in in Clinical-BERT and Clinical-Longformer, both of which are baselines for this study (Alsentzer et al., 2019; Li et al., 2022).

i2b2 is a benchmark dataset for clinical entity recognition. It has also been used to derive a question and answering dataset known as emrQA which we will be using for evaluation. This is done in a semi-automatic manner by using the location of entities in a document and building sets of questions around those entities. For example, if the medication Tylenol is identified as an entity, emrQA will generate the question "Is this patient taking Tylenol" and use the line with the entity as the answer (Pampari et al., 2018). This makes emrQA an extractive question answering dataset similar to SQUAD. We selected Medication, Relations, and Heart Disease as the topics for question answering due to their use in previous papers.

We followed preprocessing and sampling techniques identified in (Yue et al., 2020). They showed under-sampling leads to no change in performance for Medication and Relation tasks, as there is a large amount of redundancy in each of the questions from the automatic generation mechanism. When documents were longer than the allowed input sequences, we broke them up into smaller documents for each question, and allowed for "no answer" to be possible in sections without the answer. We also built in overlap to these sections to ensure that the full answer would appear in at least one document and not be cut off.

For evaluation we select two common metrics used for question answering, Exact Match (EM) and f1-score. Exact Match operates just as it sounds: an exact match between the answer string and the predicted string. F1-score, on the other

hand, is derived from token level performance, and measures the overlap between the predicted sequence and the actual sequence. Training, validation, and test sets were built in a ratio of 7:1:2 for each dataset.

2.2 Ontology Mediated Attention

The attention mechanism used in BigBird (Zaheer et al., 2020) can be described in terms of the generalized attention mechanism. That is, for an input sequence $\boldsymbol{X} \in \mathcal{R}^{n \times d}$ and a directed graph D comprised of n nodes, the ith output vector for the generalized attention can be formulated as

$$A_D(\boldsymbol{X})_i = \boldsymbol{x}_i + \sum_{h=1}^{H} \sigma\left(Q_h(\boldsymbol{x}_i) K_h(\boldsymbol{X}_{N(i)})^T\right) \cdot V_h(\boldsymbol{X}_{N(i)}) \quad (1)$$

where Q_h, K_h, and V_h are the query, key, and value matrices respectively, σ is the softmax function, H is the number of heads, and $N(i)$ is the set of out-neighbors for node i in D. The graph of D is not fully-connected for sparse attention, in contrast to that of full attention, which is fully-connected. For our implementation of attention, we would be replacing the edges that were added via random attention with edges that are between nodes selected using ontological information.

For our transformer implementation, we have forked the BigBird code from Hugging Face transformers (Wolf et al., 2019). We chose to use this over the code written by the original authors of the BigBird paper since this code was written in PyTorch (Paszke et al., 2019). As mentioned in our previous report, the defining feature of the BigBird model is its sparse attention unit, which utilizes three different types of attention: random, windowed, and global. In our implementation, we are replacing random and global attention with our ontology-mediated attention. In this regard, we will use QuickUMLS to determine which of the input tokens are included in the ontology, and then we would attend to a random subset of these.

To do this, we will first create a modified pair of input key and value matrices for the attention layer which will only include the vectors from the previous layer that are included in the ontology (as determined by QuickUMLs). This, along with the unmodified key and value matrices, will be used as input for the attention unit. Within the attention unit, we will pass the modified key and value matrices into the function for producing the random attention adjacency matrix. This will generate a random subset of the vectors included in the ontology to attend to. The unmodified key and value matrices will be used to compute windowed and global attention as reported above.

The attention unit for the BigBird model does not operate on individual input vectors, but rather on blocks of input vectors. This is because computations over random elements at sporadic locations of a matrix are not conducive to parallelization. Thus, to effectively leverage GPU computing, the attention mechanism performs operations on blocks of size 64. During the generation of the random attention adjacency matrix, a fixed number of random blocks are selected from each row of the input, and this parameter is set to 3 by default. Although we will use these values for the initial configuration of our ontology-mediated attention mechanism, these are potential hyperparameters that we can adjust during the tuning of our final model.

In addition to random attention, we have also considered modifying global attention. BigBird's implementation of global attention attends to two blocks from the input sequence. The first block of the input sequence is always one of the two globally attended blocks. The second globally attended block is selected based on the size of the input sequence, but is usually either the fifth or the tenth block. For our potential implementation for global attention, we will also attend to the first block in the input sequence. However, the second globally attended block will probably be replaced with the input block containing the most ontologically relevant terms.

2.3 QuickUMLS

While the model learns structure, it has no informed prior information. We aim to use an ontology to generate a structure heuristic to improve upon the learning speed and accuracy. The biomedical ontology that will be used is called the UMLS, or the Unified Medical Language System. It provides a graph-like structure compromised of medical vocabularies, along with connections between them. In itself, it can act as a knowledge base, and can be used to provide structure that helps relate any number of unstructured inputs or a subset thereof. We aim to modify the attention mechanism input by adding a preprocessing step that will aim to create a fast, efficient heuristic based on the ontology. To do that, we leverage QuickUMLS (Soldaini and

Goharian, 2016).

QuickUMLS is a fast, unsupervised algorithm that uses simstring matching to process documents and extract medical concepts based on the UMLS database. For an arbitrary text input, an efficient matching algorithm extracts potential UMLS-related substrings, and looks for matching concept labels from the knowledge base. It then returns candidate concepts (represented as unique concept ids) as well as string similarity scores for each one. These similarity scores provide a fast heuristic for the model to relate input subsets that could inform one another.

With the generated heuristic, we can constrain the attention as mentioned above to a subset of the input document based on the heuristic. This allows for less vectorized attention operations, while aiming to minimize the loss of information that could propagate with the reduction from global attention. QuickUMLS is uniquely applicable for the task due its relatively minimal complexity, speed of processing, and reliable results. Additionally, we can leverage the similarity heuristic to control the broadness of concepts that the algorithm returns.

More concretely, by adjusting the simstring similarity score between the input and UMLS concepts, we can effectively control the labeling predictions. Higher similarity scores in QuickUMLS represent concept labels that match more closely with the string literal. Thus, a higher similarity threshold for acceptance will yield a shorter list of candidate concepts that will be used for attention, with each of the candidates being a closer match and, consequently, a more accurate possible-label prediction. The trade-off is that a mislabeling can cause the model to skip terms that are actually related, which is undesirable. Conversely, a lower similarity threshold decreases the chances of mislabeling completely, as the output candidate labels will be a superset of the output with a higher threshold, but will also increase the amount of possibly unrelated concept labels, which is also undesirable.

The process of generating concept labels for the input is thus a varying heuristic, which we postulate can both help with the model results as well as quicken and cheapen the learning process due to more selective attention.

3 Results

Results for each task is shown in Table 1. So far this is only a recreation of previous results to ensure that our process is matching that of previous works. ClincialBERT performed best on each of the tasks, with BERT base performing the worst, except when no predictions were made for BioBERT and BERT in emr-Relations and emr-Heart Disease.

4 Discussion

Our models under performed significantly from the baselines in the paper that first published the results (Li et al., 2022). However, they still follow the ranking in performance set by the original papers with ClinicalBERT and BioBERT performing better than BERT. We also found that the poor performance was not a product of limited compute. The changes in performance were negligible when training for 5 epochs vs 8 epochs, with the 8 epoch model having slightly better results. We have two theories as to why these performances were so different from the performances shown in the papers. The first hypothesis is rooted in the splitting that we did in our handling of documents that were longer than the allowable input sequence. This change creates a large amount of documents that have blank answers. In our experiments only 25% of test documents had answers after combining all splinter documents made for predicting on long documents. This implies that our model is learning to return "no answer" very often, leading to poor recall. The second is in the conversion of our data to the SQUAD-2 format. While this was done in (Yue et al., 2020), we are not sure if (Li et al., 2022) followed the same process. One limitation of the methods of (Yue et al., 2020) is that they only consider the first answer when there are multiple. It was not specified in (Li et al., 2022) how they handled this case so we defaulted to the methods in (Yue et al., 2020). We hope to explore these issues further in the coming weeks.

This leaves a few experiments left for our project. There are still two baselines left in Clinical-Longformer and Clinical-BigBird. We also have three different formulations in our changes to attention: Ontology Mediated Random Attention, Ontology Mediated Global Attention, and both combined.

References

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

Table 1: Performance of Baseline models tested so far

	emrQA-Medication		emrQA-Relation		emrQA-Heart Disease	
Model	EM	F1	EM	F1	EM	F1
BERT	0.1437	0.1704	0.1097	0.1249	0.	0.
BioBERT	0.1593	0.1974	0.	0.	0.	0.
ClinicalBERT	0.1660	0.2176	0.1239	0.1376	0.1526	0.2045

Beltagy, Iz, Peters, Matthew E, Cohan, and Arman. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bodenreider and Olivier. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Choromanski, Krzysztof, Likhosherstov, Valerii, Dohan, David, Song, Xingyou, Gane, Andreea, Sarlos, Tamas, Hawkins, Peter, Davis, Jared, Mohiuddin, Afroz, Kaiser, Lukasz, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kitaev, Nikita, Kaiser, Łukasz, Levskaya, and Anselm. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Li, Yikuan, Wehbe, Ramsey M, Ahmad, Faraz S, Wang, Hanyin, Luo, and Yuan. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. *academia: 29618874*.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, and Illia. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Sinong, Li, Belinda Z, Khabsa, Madian, Fang, Han, Ma, and Hao. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrqa dataset. *arXiv preprint arXiv:2005.00574*.

Zaheer, Manzil, Guruganesh, Guru, Dubey, Kumar Avinava, Ainslie, Joshua, Alberti, Chris, Ontanon, Santiago, Pham, Philip, Ravula, Anirudh, Wang, Qifan, Yang, Li, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.