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

Abstract

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). 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. This survey shows the evolution of transformers, techniques used to improve their computational and memory complexities and their application in the biomedical domain.

1 Survey

The topic of solving seq2seq learning tasks using deep learning networks is not new, but in 2014 became hot when (Sutskever et al., 2014) published their breakthrough paper. The task discussed was machine translation, a problem yet unsolved at the time. They used LSTM networks, and those became the state-of-the-art. However, 3 years later a new architecture achieved better performance, using the Transformer model (Vaswani et al., 2017). Unlike LSTM, which was a combination of an RNN and a long-term context learner, the transformer used a mechanism of self-attention for the network to understand the context of words within a sentence. The concept of self-attention quickly popularized and the Transformer was soon superseded by the next generation model, a bi-directional architecture under the name BERT (Devlin et al., 2018).

There are two main issues that were present in the original proposed transformer based model: 1. the maximum length of the sequence was restricted to a sequence length of 512 tokens, and 2. the full self-attention mechanism of the transformer architecture leads to quadratic computational and memory complexity. There are many proposed methods and models to overcome these issues, some more successful than others. Such models are generally referred to as x-formers, built on top of transformers. The general idea was to limit the attention to a sub-scope of context, and various implementations arose such as the TransformerXL (Dai et al., 2019) which aimed at recurrence of local contexts, the sparse-transformer (Child et al., 2019) which was superlinear but still more efficient, and other techniques (Wu et al., 2019) (Ye et al., 2019).
The current state-of-the-art models Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) introduce sparsity into the attention matrix to decrease the time complexity and also increase the sequence length to 4096 tokens.

When looking at specific domains another issue arises. As these models rely heavily on word distributions and derived meaning, tasks based on lingual sub-domains might see lesser performance than general-purpose tasks. With the transition of health services to the Electronic Health Record (EHR), a vast amount of data is constantly being generated for tracking and evaluating patients. However, much of the relevant information for patients is interweaved in unstructured data sources such as text and image. For instance, it been shown that curated fields such as ICD codes are prone to error, decreasing their value to clinical research and intervention (Quan et al., 2008). This has necessitated the building of automatic methods to sift through and extract valuable information from these unstructured resources.

Natural Language Processing (NLP) models have been developed in attempt to resolve the above issues in unstructured text. Multiple papers have explored the utility of transformers for tasks in clinical NLP. BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019), two models based on the original BERT model, both showed that models further pretrained on Biomedical and clinical data performed better on a variety of tasks compared to their general-domain counter parts. This has been replicated across several different studies both within domain and for other domains (Gururangan et al., 2020; Si et al., 2019). Gu et al has since shown that given enough data, BERT models trained entirely on domain specific data outperform those pretrained on both general domain and domain specific data (Gu et al., 2021). Researchers have since focused on improving specific components of the model such as fine-tuned performance on a variety of tasks (Tinn et al., 2021). Most recently Li et al introduced ClinicalLongformer and ClinicalBigBird (Li et al., 2022) which pretrain the sparse transformer architectures with clinical data achieving state-of-the-art results.

Biomedical ontologies have been used in several transformer-based architectures to improve the performance of deep neural networks for medical applications. For example, one research group used ontological information to learn robust representations of medical codes in the face of data scarcity (Choi et al., 2017). Their model feeds in each medical concept along with its ancestors in the ontological graph into its own attention unit, which then attends to the concept at different levels of granularity depending on the availability of the token. The attention unit then outputs an ontology-aware representation of that token.

Another study uses clinical ontologies to improve transformer-based medical recommendation (Shang et al., 2019). Their model also produces an embedding of each medical concept with information on its ancestors in the ontology. However, in contrast to the model used by Choi et al. which used an attention unit to produce an ontology-aware embedding, the model designed by Shang et al. (2019) uses a graph neural network to produce the embedding, which is then used in a subsequent BERT-based architecture.

In their work, Sotudeh et al. also produce ontology-aware representations which are then fed into a transformer-based architecture (Sotudeh et al., 2020). Their approach involves creating a new vector representation containing information on which words from the source document are ontological terms. They then use this in a filtering gate function to filter the original embedded tokens. The filtered representations are then directly fed into an attention unit. Our intended usage of ontological information most closely resembles that of Sotudeh et al. (2020). However, in our case we aim to use ontological information to filter keys in the attention layer itself, inducing a sparsely connected attention unit.

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*.

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

Rewan Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. Gram: Graph-based attention model for healthcare representation learning. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795.

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*.

Zihang Dai, Zhilin Yang Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

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

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

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*.

Hude Quan, Bing Li, Duncan Saunders, Gerry Parsons, and Carolyn Nilsson. 2008. Assessing validity of icd-9-cm and icd-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, 43(4):1424–1441.

Shang, Junyuan, Ma, Tengfei, Xiao, Cao, Sun, and Jimeng. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Sotudeh, Sajad, Goharian, Nazli, Filice, and Ross W. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. *arXiv preprint arXiv:2005.00163*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *arXiv preprint arXiv:2112.07869*.

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*.

Angela Wu, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. 2019. Bp-transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv:1911.04070*.

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.