Literature Survey

Student Name: Molly Hayward
Supervisor Name: Dr Noura Al-Moubayed
19/10/2020

Project title: An Analysis of Question Answering in the Biomedical Domain

I INTRODUCTION

A Problem Background

The volume of data emerging from the biomedical field is increasing rapidly. This data is very valuable, but due to the large volumes of emerging data, we run the risk of overlooking poignant medical discoveries and information vital for developing new treatments. This highlights the scope for biomedical-domain question answering, allowing users to query large biomedical corpora to efficiently retrieve an answer pertaining to the question encoded in the query.

One significant challenge of automating biomedical data extraction is the importance of delivering reliable health information in response to queries. This is vital for protecting consumers from receiving misleading, or even harmful, information. However, some questions may contain entities that are medically significant, but do not pertain to the question being asked by the user. For example, people may disclose additional medical information to provide context, but this may be insignificant to the focus of the question. Furthermore, medical questions tend to be lengthy, and can be very noisy with erroneous spelling. This makes automatic analysis more difficult.

The current trend within question answering, as well as the wider field of NLP, is to train increasingly large neural language models on vast datasets, requiring excessive computational resources. Large datasets, such as SQuAD have enabled advancements in open-domain question answering. However, in a closed-domain setting such as biomedicine, gold-standard datasets are typically far smaller. We investigate modifications that can be made to such language models to advance the state-of-the-art in Biomedical Question Answering, without reliance on copious amounts of data which is unattainable in closed-domain settings.

Keywords — Biomedical Question Answering, Transfer Learning, Transformers, Attention, Natural Language Processing, Machine Learning, Deep Learning

II THEMES

A *Background on Transfer Learning*

With transfer learning, a neural network is first pre-trained on a data-rich problem to produce a language model with an intricate knowledge of a given language (most often English). It is then fine-tuned on a downstream task, such as question answering, to tweak its parameters to fit the problem. Previously, this knowledge would be incorporated via pre-computed word embeddings (continuous word representations), such as those of Word2Vec [5] and GloVe [6].

Transformers [10] are a fundamental component of recent models harnessed for transfer learning, initially designed for sequence-to-sequence conversion problems, such as language translation. The introduction of the transformer demonstrated that an architecture utilising only an attention mechanism could contend with the previous state-of-the-art recurrent models, namely recurrent networks (RNNs). While RNNs rely on recurrent sequential processing of the input data, transformers can calculate attention values and process input tokens simultaneously, allowing for parallelisation in training.

Transformers have an encoder-decoder architecture, encoding an input sequence to produce an intermediate representation, which is then decoded to produce an output sequence. The attention mechanism utilised by the transformer allows the model to attend to states at earlier points in the input, drawing knowledge from the most relevant words in the document. The encoder and decoder blocks consist of identical stacked layers of encoders and decoders, respectively. Encoder layers consist of two sub-components - a self-attention mechanism and small feed-forward network. They generate encodings from the input, incorporating information supplied by the attention mechanism to highlight relevant tokens, before propagating this data to the next encoding layer. The final encoder passes this data to the decoder layers, which produce an output sequence. Decoder layers contain an attention mechanism which draws relevant knowledge from the outputs of previous decoders.

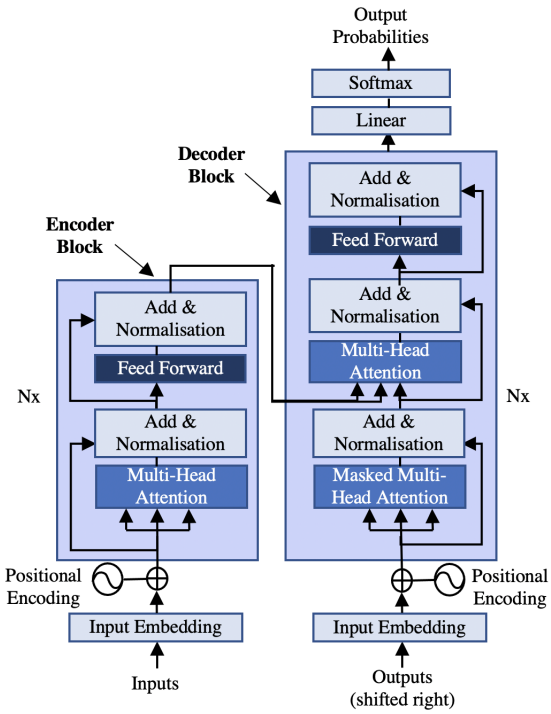

Figure 1: Transformer Architecture [10]

Introduced in Devlin et al. (2019), BERT (Bidirectional Encoder Representations from Transformers) is a hugely popular model based on the transformer architecture. BERT uses a predictive pre-training strategy, known as Masked Language Modelling (MLM), whereby approximately 15% of the input tokens are masked. Several stacked encoder-only blocks are used to read the input sequence and produce an continuous intermediate representation (sequence vector) which is used to predict the original values of the masked tokens. BERT uses Masked LM, so before sequences of inputs are given to BERT, 15% of the tokens are masked (i.e. replaced with a

[MASK] token). The model aims to predict the masked tokens given the remaining sequence of unmasked tokens. BERT is designed to encapsulate the left and right contexts in all layers, avoiding the shallow concatenation of two unidirectional representations as used in ElMo [7]. With the addition a single output-layer, BERT can be fine-tuned on downstream tasks (e.g. question answering). BERT achieved an F1 score of 93.2% and 83.1% on SQuAD version 1 and 2, respectively. However, as BERT is an extractive question-answering model, answers are restricted to snippets of the context paragraphs.

Despite the state-of-the-art performance achieved by BERT in many NLP tasks; pre-training on general domain corpora limits the performance of BERT on tasks in the biomedical domain. There is a shift in the distribution of words used in biomedical corpora, due to the confined nature of the topics discussed in these documents, compared with general corpora. BioBERT [4] aims to overcome this challenge, as they hypothesise that state-of-the-art word representation models, trained on biomedical texts, would be more effective in biomedical text mining tasks. BioBERT significantly outperforms BERT on three popular NLP tasks in the biomedical domain, including question answering, exceeding the MRR score achieved by BERT by a margin of 12.24%.

B Generative Language Models

Given that many approaches to Question Answering focus on extracting answers from documents containing the relevant knowledge, unanswerable questions can arise a context paragraph does not contain the answer. Recent advancements in generative language models, which can produce original samples of text given an initial prompt or question, allow us to generate original answers without relying on an accurate context document.

Introduced in Radford et al. (2019), GPT-2 [8] is a flexible and generalisable generative language model, used for question-answering and many other problems. The aim of this model is to perform down-stream tasks without any parameter or architectural modification i.e. in a zero-shot setting; hence, unsupervised training is conducted on 40GB of web-text, with the aim of learning from task-specific training data unintentionally embedded in such a large and diverse corpus. For instance, a corpus containing documents with questions followed by answers could incorporate task-specific knowledge of question answering, without explicit training on this task. On account of its deep knowledge of language, GPT-2 is able to produce a realistic continuation of a snippet of text; hence, GPT-2 can generate an answer, given a prompting question. On the Natural Questions dataset [3], a corpus of factoid questions, GPT-2 answers 4.1% of questions correctly when evaluated using the *exact match* metric and has an accuracy of 63.1% on the 1% of questions it had the highest confidence in. This highlights the scope for potential improvements of GPT-2 by fine-tuning on the downstream question answering task, rather than evaluating GPT-2 as a generalisable model.

Raffel et al. (2019) [9] conducted an analysis of transfer learning techniques used in research settings in recent years, combining their insights to create the Text-To-Text Transformer (T5). T5 formulates every text processing task as a text-to-text problem; allowing it to tackle a diverse range of tasks without requiring architectural modifications. Furthermore, as T5 is trained in a generative setting (i.e. producing strings of text), it can produce unique answers to questions,

informed by its knowledge of language. This is in contrast to BERT-style models which select a pre-defined span of the input. Unsupervised training was conducted on a very large corpus of de-noised text, extracted from the web, followed by supervised fine-tuning on a range of tasks (including Question Answering). When fine-tuned on SQuAD T5 achieved an F1 score of 88.81, and an EM score of 90.88.

C Efficiency in Transfer Learning

Large scale models with millions, or even billions, of hyper-parameters require a large allocation of resources in order to achieve state-of-the-art results. Numerous studies have sought to improve the efficiency in training Transformer-based models; either by allowing the models to consume less compute and memory resources, or by improving the efficiency of the earning process to necessitate fewer training epochs.

Allocated equivalent compute resources, ELECTRA [1] outperforms models that rely upon Masked Language Modelling (MLM) during pre-training, such as BERT. With MLM, approximately 15% of input tokens are masked, then the model learns from predicting the true identity of the masked tokens. ELECTRA utilises Replaced Token Detection (RTD) - a more sample-efficient training approach. A discriminative model predicts whether each token has been replaced with an alternative token; allowing the model to learn from all input tokens, rather than just 15%. When fine-tuned on the SQuAD dataset, a benchmark Question Answering dataset, ELECTRA outperforms BERT in the Question Answering task (given equivalent resources), achieving an EM score of 88.0 and an F1 score of 90.6.

Due to the memory and computational expense of computing dot product attention ($O(L^2)$ on sequences of length L), the Transformer model can only focus on a limited window of context. However, a question answering model may be required to attend to text encountered many lines prior in order to extract relevant knowledge. The Reformer [2] model replaces dot product attention with Locality-Sensitive Hashing, reducing computational and memory complexity to O(L), enabling the model to expand its context window. Memory requirements are further reduced with the use of reversible residual layers, to avoid storing the inputs to each layer in memory during back-propagation. Application of the Reformer model to Question Answering has not yet been explored and this is likely due to a lack of availability of pre-trained model weights.

D Performance Evaluation

Equation 1

$$Precision = \frac{Relevant\ Retrieved\ Documents}{Retrieved\ Documents}$$

Equation 2

$$Recall = \frac{Relevant\ Retrieved\ Documents}{Relevant\ Documents}$$

Equation 3

$$F_1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Equation 4

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

In the domain of question answering, precision and recall are used to evaluate the quality of context documents retrieved in response to a query. Precision refers to the proportion of documents that are considered *relevant*, and recall describes the proportion of relevant documents that are actually *retrieved*. Mean Reciprocal Rank (MRR) is an evaluation metric for questions producing a list of candidate answers; the average of the reciprocal ranks of results for queries Q. MRR is zero in the case where no correct answers are returned.

References

[1] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

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

[3] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[7] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.