Research Proposal

Nirbhay P. Tandon
968675
Email: ULPNTAND@ljmu.ac.uk
Project Supervisor: Mr. Ankit Jha

Abstract

Attention based Transformer architectures have become the norm of modern day Natural Language Processing. Google began this trend back in 2017 with their paper *Attention Is All You Need*[1], by introducing the Transformer architecture that works solely on attention mechanisms. The purpose of our work will be to explore a new kind of Transformer architecture. Compare & contrast its performance against the SQuAD 2.0 Dataset[2] based other architectures such as BERT[3], RoBERTa[4], etc. Through our research we aim to produce a new Transformer Architecture that allows **Add additional details around how neural nets & lstms compare to transformers. Fill in the final objective of the research proposal*

Contents

1 Introduction

2 Background & Related Research

2.1 Background

..is this just literature review now?

2.2 Related Research

Below is just papers that have been read and their summary.

- Transformers were produced as a result of a simplification process carried out in [1].

2.3 Related Research

*** Some papers have messed up formatting of references. Correct before final.**

1. A series of experiments have shown that these iterative methods provide enhanced performance in the word boundary detector which results in a significant improvement in speech recognition accuracy.[16]

2. Long short-term memory[6] and gated recurrent[9] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation. Aligning the positions to steps in computation time, they generate a sequence of hidden states ht, as a function of the previous hidden state ht - 1 and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. The goal of reducing sequential computation forms the foundation of the Extended Neural GPU, ByteNet and ConvS2S, all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations. On the WMT 2014 English-to-German translation task, the big transformer model (Transformer in Table 2) outperforms the best previously reported models by more than 2.0 BLEU, establishing a new state-of-the-art BLEU score of 28.4. On the WMT 2014 English-to-French translation task, the big model achieves a BLEU score of 41.0, outperforming all of the previously published single models, at less than 1/4 the training cost of the previous state-of-the-art model. The authors set the maximum output length during inference to input length + 50, but terminate early when possible[**?**].The Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, the authors achieve a new state of the art. In the former task the best model outperforms even all previously reported ensembles. Making generation less sequential is another research goals of ours[1]

3. Self-training methods such as ELMo (Peters et al, 2018), GPT (Rad-

ford et al, 2018), BERT (Devlin et al, 2019), XLM (Lample and Conneau, 2019), and XLNet (Yang et al, 2019) have brought significant performance gains, but it can be challenging to determine which aspects of the methods contribute the most. The authors present a replication study of BERT pretraining (Devlin et al, 2019), which includes a careful evaluation of the effects of hyperparmeter tuning and training set size. The two segments are presented as a single input sequence to BERT with special tokens delimiting them: [CLS], x1, . BERT uses the ubiquitous transformer architecture (Vaswani et al, 2017), which the authors will not review in detail. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. The authors use a transformer architecture with L layers. Each block uses A self-attention heads and hidden dimension H[4]

4. Increasing model size when pretraining natural language representations often results in improved performance on downstream tasks. At some point further model increases become harder due to GPU. Proposed methods lead to models that scale much better compared to the original BERT, a self-supervised loss that focuses on modeling inter-sentence coherence. A Lite BERT (ALBERT) architecture that has significantly fewer parameters than a traditional BERT architecture. These results show that weight-sharing has an effect on stabilizing network parameters. Inter-sentence modeling is an important aspect of language understanding, but we propose a loss based primarily on coherence. We have convincing evidence that sentence order prediction is a more consistently-useful learning task that leads to better language representations, we hypothesize that there could be more dimensions not yet captured by the current self-supervised training losses that could create additional representation power for the resulting representations. 4.2.2 DOWNSTREAM EVALUATION Following Yang et al (2019) and Liu et al (2019), we evaluate our models on three popular benchmarks: The General Language Understanding Evaluation (GLUE) benchmark (Wang et al, 2018), two versions of the Stanford Question Answering Dataset (SQuAD; Rajpurkar et al, 2016; 2018), and the ReAding Comprehension from Examinations (RACE) dataset (Lai et al, 2017). [19]

5. Machine reading comprehension has become a central task in natu-

ral language understanding, fueled by the creation of many large-scale datasets. Question 2: “What was the name of the 1937 treaty?” Plausible Answer: Bald Eagle Protection Act even produced systems that surpass human-level exact match accuracy on the Stanford Question Answering Dataset (SQuAD), one of the most widely-used reading comprehension benchmarks. These systems are still far from true language understanding. Models only need to select the span that seems most related to the question, instead of checking that the answer is entailed by the text. The authors evaluated three existing model architectures: the BiDAF-No-Answer (BNA) model proposed by Levy et al (2017), and two versions of the DocumentQA No-Answer (DocQA) model from Clark and Gardner (2017), namely versions with and without ELMo (Peters et al, 2018) These models all learn to predict the probability that a question is unanswerable, in addition to a distribution over answer choices. The authors find this strategy does slightly better than taking the argmax prediction, possibly due to the different proportions of negative examples at training and test time. A state-of-the-art model achieves only 66.3% F1 score when trained and tested on SQuAD 2.0, whereas human accuracy is 89.5% F1, a full 23.2 points higher. When evaluating on the test set, the authors use the threshold that maximizes F1 score on the development set. Following Rajpurkar et al (2016), the authors report average exact match and F1 scores. SQuAD 2.0 forces models to understand whether a paragraph entails that a certain span is the answer to a question. Relation extraction systems must understand when a possible relationship between two entities is not entailed by the text (Zhang et al, 2017). Jia and Liang (2017) created adversarial test examples that fool models trained on SQuAD 1.1. Models that are trained on similar examples are not fooled by their method. The adversarial examples in SQuAD 2.0 are difficult even for models trained on examples from the same distribution. [2]

6. Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al, 2018a; Radford et al, 2018; Howard and Ruder, 2018). The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations. The authors present BERT fine-tuning results on 11 NLP tasks. 4.1 GLUE. The General Language Understanding Evaluation (GLUE) benchmark (Wang et al, 2018a) is a collection of

diverse natural language understanding tasks. The authors present BERT fine-tuning results on 11 NLP tasks. The General Language Understanding Evaluation (GLUE) benchmark (Wang et al, 2018a) is a collection of diverse natural language understanding tasks. To finetune on GLUE, the authors represent the input sequence as described, and use the final hidden vector C equals RH corresponding to the first input token ([CLS]) as the aggregate representation. The only new parameters introduced during fine-tuning are classification layer weights W equals RK x H , where K is the number of labels. The best performing method concatenates the token representations from the top four hidden layers of the pre-trained Transformer, which is only 0.3 F1 behind fine-tuning the entire model. This demonstrates that BERT is effective for both finetuning and feature-based approaches. Using only the RND strategy performs much worse than the strategy as well. Recent empirical improvements due to transfer learning with language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems. These results enable even low-resource tasks to benefit from deep unidirectional architectures. The authors' major contribution is further generalizing these findings to deep bidirectional architectures, allowing the same pre-trained model to successfully tackle a broad set of NLP tasks.[3]

7. Recurrent networks can in principle use their feedback connections to store representations of recent input events in form of activations (short-term memory”, as opposed to long-term memory” embodied by slowly changing weights). This is potentially significant for many applications, including speech processing, non-Markovian control, and music composition (e.g., Mozer 1992). With conventional Back-Propagation Through Time” (BPTT, e.g., Williams and Zipser 1992, Werbos 1988) or Real-Time Recurrent Learning”. The authors use 3 different, randomly generated pairs of training and test sets. With each such pair the authors run 10 trials with different initial weights. The difficult task is of a type that has never been solved by other recurrent net algorithms. It shows that LSTM can solve long time lag problems involving distributed, continuous valued representations. Long Short-Term Memory Once block 1 is called and closed, this fact will become visible to block 2. The efficient truncated backprop version of the LSTM algorithm will not solve problems similar to strongly delayed XOR problems”, where the goal is to compute the XOR of two widely separated inputs that previously occurred somewhere in a noisy se-

quence. By generating an appropriate negative connection between memory cell output and input, LSTM can give more weight to recent inputs and learn decays where necessary.Each memory cell's internal architecture guarantees constant error ow within its constant error carrousel CEC, provided that truncated backprop cuts o error ow trying to leak out of memory cells This represents the basis for bridging very long time lags. It will be interesting to augment sequence chunkers. [15]

8. Relation extraction systems populate knowledge bases with facts from an unstructured text corpus. When the type of facts are predefined, one can use crowdsourcing (Liu et al, 2016) or distant supervision (Hoffmann et al, 2011) to collect examples and train an extraction model for each relation type. The relation educated at(x, y) can be mapped to "Where did x study?" and "Which university did x graduate from?". The authors' goal is to find a set of text spans A in s for which R(e, a) holds for each a equals A. To understand how well the method can generalize to unseen data, the authors design experiments for unseen entities, unseen question templates, and unseen relations. The authors test the method's ability to generalize to new descriptions of the same relation, by holding out a question template for each relation during training. Some recent QA datasets were collected by expressing knowledge-base assertions in natural language. The Simple QA dataset (Bordes et al, 2015) was created by annotating questions about individual Freebase facts (e.g. educated at(T uring, P rinceton)), collecting roughly 100,000 natural-language questions to support QA against a knowledge graph. To the best of the knowledge, this is the first robust method for collecting a question-answering dataset by crowd-annotating at the schema level.The authors showed that relation extraction can be reduced to a reading comprehension problem, allowing them to generalize to unseen relations that are defined on-the-fly in natural language. To support future work in this avenue, the authors make the code and data publicly available.[13]

9. Question answering is an important end-user task at the intersection of natural language processing (NLP) and information retrieval (IR). Results presented in Tables 2 and 3 clearly demonstrate the strength of the FastQA + char-emb. (FastQA) system. It is very competitive to previously established stateof-the-art results on the two datasets and even improves those for NewsQAWe introduced a simple, context/type

matching heuristic for extractive question answering which serves as guideline for the development of two neural baseline system. FastQA, our recurrent neural network (RNN)-based system turns out to be an efficient neural baseline architecture for extractive question answering. It combines two simple ingredients necessary for building a currently competitive QA system: a) the awareness of question words while processing the context and b) a composition function that goes beyond simple bag-of-words modeling. We argue that this important finding puts results of previous, more complex architectures as well as the complexity of recent QA datasets into perspective[21]

10. As machine reading comprehension tasks with unanswerable questions stress the importance of answer verification in Machine reading comprehension modeling, this paper devotes itself to better verifieroriented MRC task-specific design and implementation for the first time.[18] Machine reading comprehension (MRC) aims to teach machines to answer questions after comprehending given passages (Hermann et al 2015; Joshi et al 2017; Rajpurkar, Jia, and Liang 2018), which is a fundamental and longstanding goal of natural language understanding (NLU). In terms of powerful enough PrLMs like ALBERT and ELECTRA, the Retro-Reader significantly outperforms the baselines with p-value ¡ 0.01,8 and achieves new state-of-the-art on the SQuAD2.0 challenge.. The authors' method shows consistent improvements over the baselines and achieves new state-of-the-art results Retro-Reader on ALBERT and Retro-Reader on ELECTRA denote the final models, which are respectively the ALBERT and ELECTRA based retrospective reader composed of both sketchy and intensive reading modules without question-aware matching for simplicity. The authors make the following observations: 1) The authors' implemented ALBERT and ELECTRA baselines show the similar EM and F1 scores with the original num- All HasAns NoAns. EM F1 EM F1 EM F1 BERT + E-FV + RV ALBERT[12]

11. Pre-trained word representations (Mikolov et al, 2013; Pennington et al, 2014) are a key component in many neural language understanding models. Learning high quality representations can be challenging They should ideally model both (1) complex characteristics of word use, and (2) how these uses vary across linguistic contexts. The authors use vectors derived from a bidirectional LSTM that is trained with a coupled language model (LM) objective on a large text corpus. For

> this reason, the authors call them ELMo (Embeddings from Language Models) representations. The authors learn a linear combination of the vectors stacked above each input word for each end task, which markedly improves performance over just using the top LSTM layer. Adding ELMo establishes a new state-of-the-art result, with relative error reductions ranging from 6 - 20% over strong base models. This is a very general result across a diverse set model architectures and language understanding tasks. The authors' baseline model (Clark and Gardner, 2017) is an improved version of the Bidirectional Attention Flow model in Seo et al (BiDAF; 2017). A 11 member ensemble pushes F1 to 87.4, the overall state-of-the-art at time of submission to the leaderboard. The increase of 4.7% with ELMo is significantly larger the 1.8% improvement from adding CoVe to a baseline model (McCann et al, 2017). The authors have introduced a general approach for learning high-quality deep context-dependent representations from biLMs, and shown large improvements when applying ELMo to a broad range of NLP tasks. Through ablations and other controlled experiments, the authors have confirmed that the biLM layers efficiently encode different types of syntactic and semantic information about wordsin-context, and that using all layers improves overall task performance [17]

**Loads of stuff missing still. Need to list out paper by LSTM, YIH, etc.

3 Aims & Objectives

Through our research we aim to establish the efficiency of our new Transformer architecture. We shall implement the existing models that are available via libraries such as HuggingFace[5], PyTorch & Tensorflow, run the SQuAD 2.0[2], to obtain benchmark scores & then compare the results with our proposed architecture. We hope to establish our proposed transformer architecture as a competent enough contender to be used within both industry & academia.

4 Research Methodology

To implement this research we shall break the project down into 5 phases. These are outlined below.

4.1 Literature Review

In this phase we will review the research that has been published already around the different kinds of architectures, shortlist some of the most widely used ones, compare their results using the SQuAD2.0 [2] & outline the pros and cons of each of these architectures. The rationale is to review & understand as much of the research as possible so that we can avoid potential pitfalls, not duplicate our efforts by reinventing the wheel & organize a better approach to perform our research. *****to add some pointers about how existing research has been done*****

4.2 Research Benchmarks

Here we shall focus on obtaining benchmark scores for the shortlisted architectures above using the dataset[2]. The parameters used will be validation & test set scores of the models. The training shall be carried out on each of the models for a 100(***TBD, 100 epochs per model for Squad will take over 100 hours of training, not sure if its worth it***) epochs. We shall also look at the specificity/recall of these results to better understand if our work was done correctly or not.

4.3 Architecture Creation

4.4 Architecture Refinement

4.5 Research Findings

4.6 Conclusion

5 Expected Outcomes

6 Requirements & Resources

7 Research Plan

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan N, Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762.

[2] Rajpurkar, P., Jia, R. and Liang, P., 2018. Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.

[3] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Huggingface.co. 2021. Transformers — transformers 4.4.2 documentation. [online] Available at: ¡https://huggingface.co/transformers/¿ [Accessed 4 April 2021].

[6] Hochreiter, S., Bengio, Y., Frasconi, P. and Schmidhuber, J., 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[7] Akbik, A., Blythe, D. and Vollgraf, R., 2018, August. Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics (pp. 1638-1649).

[8] Al-Rfou, R., Choe, D., Constant, N., Guo, M. and Jones, L., 2019, July. Character-level language modeling with deeper self-attention. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 3159-3166).

[9] Zhou, J., Cao, Y., Wang, X., Li, P. and Xu, W., 2016. Deep recurrent models with fast-forward connections for neural machine translation. Transactions of the Association for Computational Linguistics, 4, pp.371-383.

[10] Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O. and Kaiser, L., 2015. Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114.

[11] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[12] Zhang, Y., Zhong, V., Chen, D., Angeli, G. and Manning, C.D., 2017, September. Position-aware attention and supervised data improve slot filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 35-45).

[13] Levy, O., Seo, M., Choi, E. and Zettlemoyer, L., 2017. Zero-shot relation extraction via reading comprehension. arXiv preprint arXiv:1706.04115.

[14] Yih, S.W.T., Chang, M.W., Meek, C. and Pastusiak, A., 2013. Question answering using enhanced lexical semantic models.

[15] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.

[16] Chung, M.I., Kushner, W. and Damoulakis, J., 1985, April. Word boundary detection and speech recognition of noisy speech by means of iterative noise cancellation techniques. In ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 10, pp. 1838-1838). IEEE.

[17] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[18] Zhang, Z., Yang, J. and Zhao, H., 2020. Retrospective reader for machine reading comprehension. arXiv preprint arXiv:2001.09694.

[19] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[20] Graves, A., 2013. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

[21] Weissenborn, D., Wiese, G. and Seiffe, L., 2017. Making neural qa as simple as possible but not simpler. arXiv preprint arXiv:1703.04816.

[22] Baevski, A. and Auli, M., 2018. Adaptive input representations for neural language modeling. arXiv preprint arXiv:1809.10853.

[23] Zhang, Z., Wu, Y., Zhou, J., Duan, S., Zhao, H. and Wang, R., 2020, April. Sg-net: Syntax-guided machine reading comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 9636-9643).