Lightening the Load: DeLighT Blocks for Faster QA Training and Ensembling

Sam Lowe[1]

[1]*MSCS Student, Department of Engineering, Stanford University*

Stanford Engineering

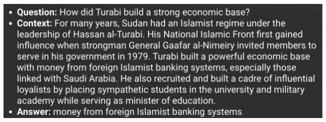
Abstract

Transformer models have proven to be one of the most dominant paradigms for NLP in recent years across a wide variety of problem domains. However, these networks tend to be extremely memory- and computationally-demanding models to train, leaving the door open to more efficient alternatives to standard transformer architectures. In this work, we explore the applicability of one such lightweight transformer alternative - the DeLighT block - to the task of SQuAD 2.0 question answering. We replace the Transformer-style RNN Encoder blocks in a standard QANet model with DeLighT blocks, consisting of a series of group-linear, expand-reduce layers, followed by self attention and modeling layers. This substitution provides two main advantages - greater flexibility over per-block parameter counts and parallelization during training time - which results in cheaper training costs in regards to both memory and compute time. We experiment with several different model configurations of varying architectures, attention styles, and training regimes to discover a lightweight model to use as the basis for ensembling methods, the next step for this work.

Stanford University

Background

Problem Setting. The problem addressed in this work is that of Stanford's SQuAD 2.0 Question Answering. The NLP domain of question answering is an fruitful area of study because it provides a strong indicator of a system's ability to understand what it reads, as the model must be able to first comprehend a context paragraph before ultimately attempting to answer the prompt. SQuAD examples consist of a context paragraph, associated question, and span answer from the context area (if it exists, which is an added difficulty in 2.0).



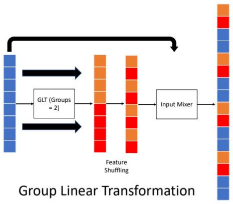
- **Question:** How did Turabi build a strong economic base?
- **Context:** For many years, Sudan had an Islamist regime under the leadership of Hassan al-Turabi. His National Islamic Front first gained influence when strongman General Gaafar al-Nimeiry invited members to serve in his government in 1979. Turabi built a powerful economic base with money from foreign Islamist banking systems, especially those linked with Saudi Arabia. He also recruited and built a cadre of influential loyalists by placing sympathetic students in the university and military academy while serving as minister of education.
- **Answer:** money from foreign Islamist banking systems

Sample SQuAD Data Point

The Parameter Boom. Since the seminal paper "Attention is All You Need", Transformer architectures have been behind some of the biggest gains in performance across NLP domains, but these gains have been accompanied by a ballooning in computational costs associated with training the models. One recent example of a state-of-the-art language model, GPT-3, contains a whopping 175 billion parameters. The performance gains enabled by these models are admirable, but they have come at the cost of raising the barriers to NLP research, meaning a lightweight alternative has the potential to not just improve performance, but democratize research efforts.

Methods

DeLighT Block. The architectural component at the center of this study is the DeLighT block, a lightweight Transformer alternative. The DeLighT block borrows its name from the expand-reduce DeLighT transformation at the start of the architecture. A DeLighT transformation consist of N layers of group linear transformations: the first N / 2 layers expand the input while increasing the number of groups, and the latter N / 2 reduce this back to a compressed representation. Additionally, feature shuffling and input mixing enable long-term dependency learning and residual connections. The rest of the DeLighT block architecture is very similar to Transformers: self attention followed by modeling layers.



Group Linear Transformation

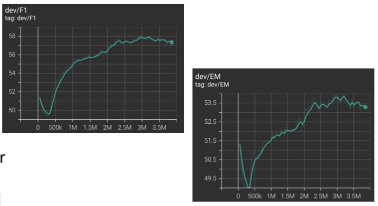
QANet + DeLighT. QANet is a Transformer-inspired architecture for question answering that has achieved state-of-the-art results on SQuAD 2.0. QANet uses a Transformer-style encoder block after its embedding and cross-attention layers, so we have substituted the DeLighT block directly for those two components.

Experiments

Models. We experimented with a wide variety of model configurations to find the one most amenable to the DeLighT block architecture. Initial experiments in utilizing the DeLighT block within the baseline BiDaF model were unsuccessful, so we attempted to train the same architecture but with a training regime more akin to QANet. Our final experiments utilize QANet as the base model and explore the applicability of differing attention variants (self attention and synthesizer attention).

Results.

Model	F1	EM
BiDaF (Baseline)	60.77	57.47
BiDaF + DeLighT	53.42	48.21
BiDaF + DeLighT (QANet Regime)	52.70	46.77
QANet + DeLighT (Synth Attention)	55.76	51.94
QANet + DeLighT (Self Attention)	58.10	53.97



Training Curves for QANet + DeLighT (Self Attention)

Analysis and Future Work

Analysis. All of the attempts at training a BiDaF-style model with DeLighT blocks struggled to manage any sort of learning, mainly settling for the "safe bet" of predicting no answer at all times. The QANet-style models, on the other hand, were capable of modest learning (though still not meeting the benchmark set by the BiDaF baseline), indicating the potential for our architecture to perform well on SQuAD given the right considerations around hyperparameters. In analyzing sample outputs from the model, it appears that it is best, and most confident, at predicting numerical answers, preferring numerical answers even on some samples where no answer is correct.

Conclusions. Our results demonstrate that while Transformer models are a popular choice for many applications today, they are certainly not the *only* choice when it comes to performant models, and the promise of lightweight and efficient alternates to Transformer models is a promising direction for future developments in question answering and other NLP domains.

Future Work. The main direction for the remaining work is to demonstrate the power of ensembling to offset some of the performance penalties observed for the DeLighT architectures. I will be training a series of models over different random seeds and experimenting with differing methods of combining their predictions (e.g. averaging, voting) to augment the shortcomings in each.