SQuAD 2.0 Question Answering with Reformer

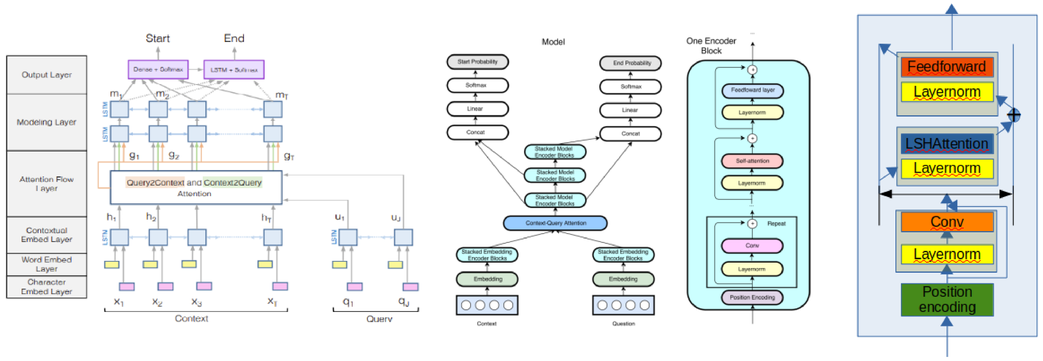
Haiyuan Mei
Stanford University, Department of Computer Science

Overview

- **Motivation**: Training Transformer based models is prohibitively expensive, LSH based attention mechanism is able to reduce the complexity from $O(L^2)$ to $O(L \log L)$.
- **Approach:** Replicate and experiment Reformer [1] based model and two baseline models: BiDAF [2], QANet [3] and evaluate the effectiveness of Reformer model on SQuAD 2.0 dataset.
- **Training and evaluation:** Use SQuAD V2.0 (Stanford Question Answering Dataset) [4] to all three models, then evaluate three metrics: Exact Match (EM), F1 and AvNA with Dev dataset, and the training speed.

Model

- **BiDAF:** The BiDAF(Bidirectional attention flow for machine comprehension [2]) model BiDAF is a closed-domain, extractive QA model that can only answer factoid questions. The hierarchical multi-stage architecture models the representations of the context paragraph at different levels of granularity, and finally feed the output layer with Query-aware Context representation.



- **QANet:** The QANet [3] applies the same 6 layer hierarchy as BiDAF model. The key improvement which makes QANet outperform BiDAF is that it uses full convolution and transformer based self attention mechanism in the embedding encoder and model encoder layers, where convolution models local interaction, and self-attention models global interaction. It also uses DCN xiong2016dynamic for query-context attention.
- **Reformer:** Reformer(the efficient transformer [1]) applies Memory-efficient attention and locality-sensitive hashing attention to reduce the memory and computation complexity from $O(l^2)$ to $O(l \log l)$, and reversible residual layers (RevNets [5]) to avoid storing n copies of actions for n attention layers.

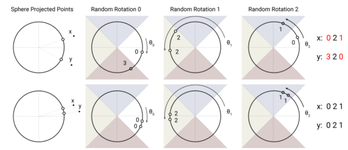
Training

- **Loss:** The training loss sums up the two cross-entropy loss with respect to the true start and end index of theanswer, averaged over all examples:

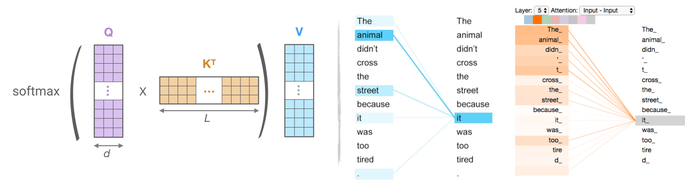
$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(p^1_{y^1_i}) + \log(p^2_{y^2_i}) \quad (1)$$

LSH Attention

- **Locality Sensitive Hashing:** Map high dimensional vector to a hash bucket.



- **LSH Attention:** Achieve $O(L \log L)$ complexity for attention calculation.

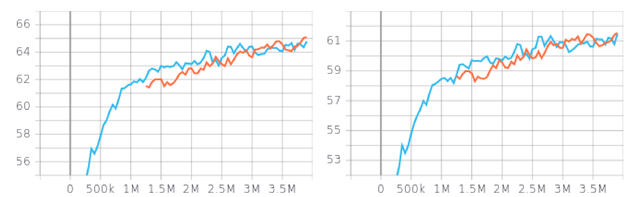


Experiment

- BiDAF, character embedding based Reformer and QANet. Reformer based QANet model have similar F1/EM score as the QANet, with much slower the training speed:

Experiments	BiDAF	QANet	QANet with Reformer
F1	< 60.5	65.1	64.78
EM	< 57.5	61.54	61.47
AvNA	-	71.89	70.71
Training time	-	<8 hrs	60hrs

- Evaluation F1/EM of Complex Reformer and QANet).



Discussion

- Reformer can be an alternative choice for attention based models, especially when long sequence cannot be handled by limited GPU resource.
- Reformer's ability to handle long sequence is with the cost of computation parallelism.

References

[1] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020.
[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension, 2016.
[3] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension, 2018.
[4] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018.
[5] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations, 2017.