QA System with Transformer-XL

Zhuzhu Wang, Yizhi Zhang

Stanford University

Problem Statement

Goal: Build a question-answering system on SQuAD 2.0 with

- BiDAF (character-level embedding)
- QANet
- Transformer-XL

Current Challenges:

- Switch LSTMs to the transformers to enable faster training
- Process the long context paragraphs with Transformer-XL

Problem Setup

Dataset: SQuAD 2.0, including unanswerable questions

Size:

- Training size: ~130k
- Eval size: 6078
- Test size: 5915

Input: Paragraph + Question

Output: Start and end positions that define the answer range

Related Work

BiDAF [1]:

- Word-level embedding layer, bidirectional LSTM encoding layer, C2Q+Q2C attention layer, modelling layer, output layer

QANet [2]:

- Replace the recurrent structure with convolutions and attentions; Faster in training

Transformer-XL [3]:

- Segment-level recurrence mechanism + relative positional encoding; Address the gradient vanishing and explosion problem; Model long-term dependencies

Related Work

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016

[2] 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. arXiv preprint arXiv:1804.09541, 2018

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

Methods

BiDAF

- Word embedding layer (the baseline model does not include character embedding)
- Embedding encoder layer (bidirectional LSTM)
- Attention layer (context-query bidirectional attention flow)

$$a_i = \sum_{j=1}^{M} softmax(S_{i,:})_j \cdot q_j$$

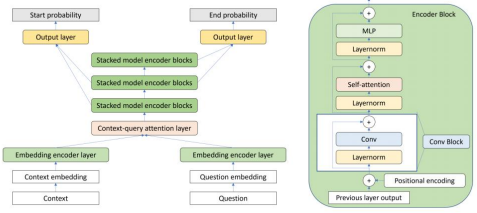
- Modeling layer (bidirectional LSTM)
- Output layer (bidirectional LSTM)

QANet

- Substituted RNNs with encoder blocks, both in embedding encoder and modeling layers

- QANet encoder block:

- Depthwise separable convolution layers (2 conv layers in embedding encoder layer, 4 conv layers in modeling layer)
- Multiheaded self-attention layer + MLP
- Layer dropout – applied after each layer, scaling dropout probability based on layer depth



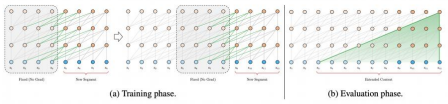
Transformer-XL

Segment-level recurrence:

$$\tilde{\mathbf{h}}_{\tau+1}^{n-1} = \left[\mathrm{SG}(\mathbf{h}_{\tau}^{n-1}) \circ \mathbf{h}_{\tau+1}^{n-1}\right],$$
$$\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n} = \mathbf{h}_{\tau+1}^{n-1}\mathbf{W}_{q}^{\top}, \tilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_{k}^{\top}, \tilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_{v}^{\top},$$
$$\mathbf{h}_{\tau+1}^{n} = \text{Transformer-Layer}\left(\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n}\right).$$

Relative positional encoding:

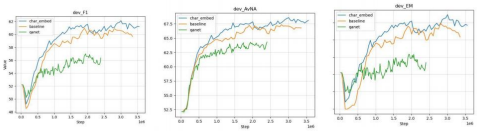
$$\mathbf{A}_{\tau,i,j}^{n} = \mathbf{q}_{\tau,i}^{n\top}\mathbf{k}_{\tau,j}^{n} + \mathbf{q}_{\tau,i}^{n\top}\mathbf{W}_{k,R}^{n}\mathbf{R}_{i-j} + u^{\top}\mathbf{k}_{\tau,j} + v^{\top}\mathbf{W}_{k,R}^{n}\mathbf{R}_{i-j}$$



(a) Training phase. (b) Evaluation phase.

Experiments

Results from BiDAF, QANet, Transformer-XL



Model	Test Set		Dev Set		
	EM	F1	EM	F1	AvNA
Baseline	N/A	N/A	57.60	60.94	67.23
Char embed	59.20	62.74	58.93	62.11	68.58
QANet	49.77	52.14	53.32	55.74	63.07
Transformer-XL	N/A	N/A	50.92	51.01	54.56

Analysis

Question: Who ordered Loudoun to defend Louisbourg?
Context: Loudoun, a capable administrator but a cautious field commander, …He was then ordered by William Pitt, the Secretary of State responsible for the colonies, to attack Louisbourg first. …
Prediction: William Pitt **Answer:** N/A

Question: Thousands of madrasahs spawned what organization?
Context: The Taliban were spawned by the thousands of madrasahs the Deobandi movement established for impoverished Afghan refugees and supported by governmental and religious groups in neighboring Pakistan. …
Prediction: Deobandi **Answer:** The Taliban

Conclusions

- Quantitative results did not demonstrate absolute advantage of QANet, which could be due to an increasing number of learnable parameters. But we observe the transformers are better than the RNNs at capturing long-term correlations in the sequence.
- For future work, we expect to: 1) Try different hyperparameters as well as more regularization techniques when training deep transformer networks like the QANet and Transformer-XL. 2) Pretrain on the NMT task then finetuning on the question answering dataset, which might allow the model to better capture the grammatical information in language.