Transformer-XL in SQuAd QA System

Using an Attentive Language Model in Question Answering

Designed by Chenkai Mao and Qinghong Zheng

Our team combined Transformer-XL and QANet for the SQuAD QA system.

1. Background

Stanford **Qu**estion **A**nswering **D**ataset (**SQuAD**):

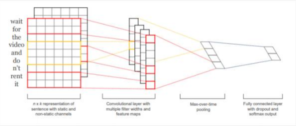
- Reading comprehension dataset (Question Answering)
- 100,000 Questions, over 50,000 unanswerable

2. Methods

(a) Baseline: Bidirectional Attention Flow (BiDAF)

— Embedding Layer: Projection & Highway
— Encoder Layer: bidirectional LSTM
— Attention Layer: bidirectional attention flow
— Modeling Layer: bidirectional LSTM
— Output Layer: bidirectional LSTM

(b) Convnet character-level embedding[1]


wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

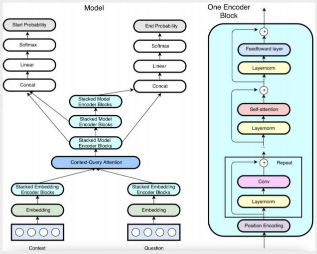
Convnet uses a sliding window to produce a feature map with each feature being

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b)$$

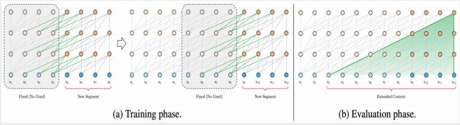
and is max-pooled to take the maximum value. Convnets are also used for word-level embeddings QANet and Transformer-XL.

(c) QANet[2]

—Transformer
—Self-attention
—Weight-sharing



(d) Transformer-XL[3]



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

Key contributions:

- Segment-level recurrence mechanism
- Novel relative positional encoding scheme

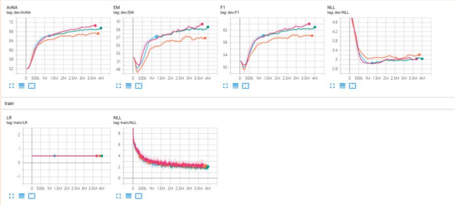
$$
\begin{aligned}
\widetilde{\mathbf{h}}_\tau^{n-1} &= \left[\mathrm{SG}(\mathbf{m}_\tau^{n-1}) \circ \mathbf{h}_\tau^{n-1}\right] \\
\mathbf{q}_\tau^n, \mathbf{k}_\tau^n, \mathbf{v}_\tau^n &= \mathbf{h}_\tau^{n-1}{\mathbf{W}_q^n}^\top, \widetilde{\mathbf{h}}_\tau^{n-1}{\mathbf{W}_{k,E}^n}^\top, \widetilde{\mathbf{h}}_\tau^{n-1}{\mathbf{W}_v^n}^\top \\
\mathbf{A}_{\tau,i,j}^n &= {\mathbf{q}_{\tau,i}^n}^\top \mathbf{k}_{\tau,j}^n + {\mathbf{q}_{\tau,i}^n}^\top \mathbf{W}_{k,R}^n \mathbf{R}_{i-j} \\
&\quad + u^\top \mathbf{k}_{\tau,j} + v^\top \mathbf{W}_{k,R}^n \mathbf{R}_{i-j} \\
\mathbf{a}_\tau^n &= \text{Masked-Softmax}(\mathbf{A}_\tau^n)\mathbf{v}_\tau^n \\
\mathbf{o}_\tau^n &= \text{LayerNorm}(\text{Linear}(\mathbf{a}_\tau^n) + \mathbf{h}_\tau^{n-1}) \\
\mathbf{h}_\tau^n &= \text{Positionwise-Feed-Forward}(\mathbf{o}_\tau^n)
\end{aligned}
$$

The segment-level recurrence shown above is added to QANet to realize the Transformer-XL model. The context is divided into segments by memory sequence lengths.

Limitations for QA Tasks:

- Segment context in QA tasks could break context continuity
- Previous segments have no information of later segments

3. Quantitative Results



Metric	Baseline	Char-Embed	QANet-Small	QANet-Large	Transformer-XL(12 epochs)
F1	60.86	63.43	65.31	67.79	65.20
EM	57.69	60.14	62.76	63.97	62.55
AvNA	67.13	70.01	72.19	74.16	70.25
NLL	3.05	3.00	2.97	2.96	2.50

- Character-level embedding improves the performance of the BiDAF model
- Transformer model (QANet) improves on the BiDAF model, but the training time is longer.
- Transformer-XL model did not perform as well as QANet, based on the limitations of the model on reading comprehension tasks.

4. Prediction Error Analysis

- Insufficient understanding in sentence context
- Ambiguities in context-question matching
- Imprecise answer boundaries

Context	Question	Answer	Prediction
. . . Typically each committee corresponds with one (or more) of the departments (or ministries) of the Scottish Government. The current Subject Committees in the fourth Session are: Economy, Energy and Tourism; Education and Culture; Health and Sport; Justice; Local Government and Regeneration; Rural Affairs, Climate Change and Environment; Welfare Reform; and Infrastructure and Capital Investment.	Economy, Energy and Tourism is one of the what?	current Subject Committees	N/A
. . . The UK subsequently adopted the main legislation previously agreed under the Agreement on Social Policy, the 1994 Works Council Directive, which required workforce consultation in businesses, and the 1996 Parental Leave Directive. . . .	Which directive mentioned was created in 1994?	Works Council Directive	Parental Leave Directive
. . . Their combined work informed the study of imperialism and it's impact on Europe, as well as contributed to reflections on the rise of the military-political complex in the United States from the 1950s. . . .	When was the military-political complex reflected upon within the scope of understanding imperialism?	the 1950s	1950s

References

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hananneh Hajishirzi. Bi-directional attention flow for machine comprehension. ICLR conference 2017, 2017.

[2] Adams W. 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.