QANet with Universal Transformer

Kongphop Wongpattananukul

Department of Energy Resources Engineering | Email: kongw@stanford.edu

Stanford CS224n

Problem

Transformer encoder and decoder architecture that compose of only self-attention and feed-forward recently gain a lot of traction recently due to its outstanding result especially in machine translation. QANet (Yu et al., 2018), one of many application inspired by Transformer encoder, is built to tackle a reading comprehension task by replacing all recurrent neural network (RNN) which is a standard design for a reading comprehension task to use only convolution and self-attention for local and global interaction respectively. This architecture show an impressive performance by achieving state-of-the-art EM/F1 score in SQuAD 1.1 and speed-up training and testing time significantly over model with RNN. However, Dropping of RNN in favor of Transformer-based encoder in QANet also discard its advantage which is an inductive bias from iterative learning that is crucial for language understanding task. Universal Transformer (Dehghani et al., 2018) is introduced to tackle this issue with recursive transformation through depth.

Data/Task

The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset curated from Wikipedia article. Each example composes of context, question (query) and answer triplet that test machine reading comprehension skill to locate answer within the context. The latest version of SQuAD is 2.0 that also include question with no answer. SQuAD 2.0 contains 141.9K context-query pairs with 129.9K for training, 6.1K for development and 5.9K for testing.

Context: In England, the period of Norman architecture immediately succeeds that of the Anglo-Saxon and precedes the Early Gothic. In southern Italy, the Normans incorporated elements of Islamic, Lombard, and Byzantine building techniques into their own, initiating a unique style known as Norman-Arab architecture within the Kingdom of Sicily.

Question: What architecture type came after Norman in England?

Answer: Early Gothic

Approach

QANet with Universal Transformer composes of 5 major components as shown in Figure 1 (a) similar to QANet.

1. Embedding layer – Map word in context and question to vector representation with word / char embedding
2. Embedding encoder layer – Refine embedded vector with local contextual in word sequence
3. Context-query attention layer – Build a set of question-aware feature vector for words in context
4. Model encoder layer – Scan context for answer
5. Output layer – Provide answer to question in context

The only difference is that we modify our encoder block to encompass recurrent transformation from timestep encoding and weight sharing between convolution and feed-forward layer introduced in Universal Transformer (Dehghani et al., 2018) as shown in Figure 1 (b). The two-dimension coordinate embedding (over position and time) are

$$P^t_{pos,2j} = \sin\left(pos/10000^{2j/d}\right) \oplus \sin\left(t/10000^{2j/d}\right)$$
$$P^t_{pos,2j+1} = \cos\left(pos/10000^{2j/d}\right) \oplus \cos\left(t/10000^{2j/d}\right)$$

Where pos is position in sequence, t is recursion in encoder stack and d is hidden size of model

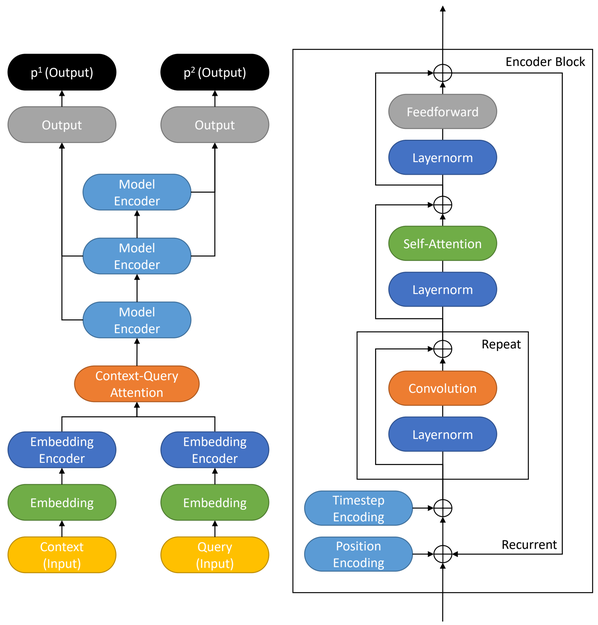


Figure 1: QANet with Universal Transformer (a), Embedding and model encoder block (b)

Results

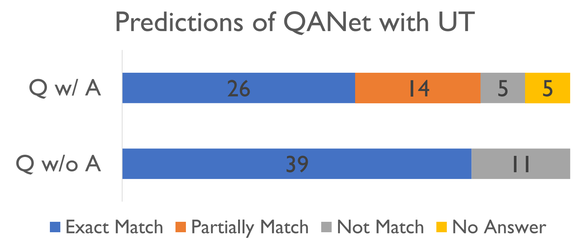
From baseline model, we implement a simplified QANet system (with hidden size of 80) and benchmark with published result on question with answer. It show significant improvement over baseline on SQuAD 2.0 development set due to CharCNN and Transformer-based encoder but only little in question with answer compare the original QANet (still it is already very close). Then, we introduce QANet with Universal Transformer using recursive transformation and gain EM/F1 score of +0.8/+1.1 and +2.8/+2.9 for the same and bigger network respectively.

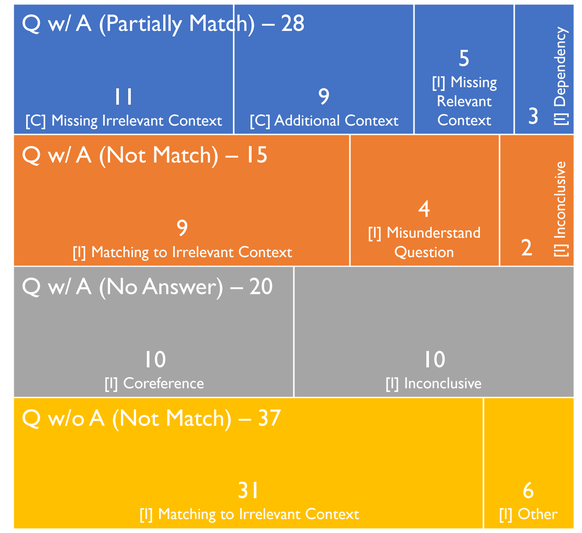
Model	Dataset	Q w/ A		Overall	
		EM	F1	EM	F1
Baseline	Dev set	68.9	77.9	58.4	61.7
[S] QANet	Dev set	69.5	78.5	63.2	66.7
[S] QANet with UT	Dev set	67.9	78.3	64.0	67.8
[L] QANet with UT	Dev set	67.5	77.0	**66.0**	**69.6**
[L] QANet with UT	Test set	-	-	62.4	66.3
QANet (Yu et al., 2018)	Dev set*	**73.6**	**82.7**	-	-

*SQuAD 1.1 Development Set

Analysis

100 predictions and 100 errors is randomly selected from result in SQuAD 2.0 development set for further investigation. The model seem to do fairly well at identifying type of question but only 26 (Q w/ A) and 39 (Q w/o A) of them get the perfect answer. The error analysis reveal that the cause of the problem is boiled down to dependency structure. The model incorrectly identifies relationship between each word/phrase and generate incorrect prediction.





Conclusions

- QANet with Universal Transformer prove that recursive transformation through depth like Universal Transformer is also useful for reading comprehension task like SQuAD.
- The model incorporate RNN's inductive bias with weight sharing of convolution and feed-forward layer and time encoding in encoder block.
- Error analysis reveal that model still struggle with incorrect dependency of question and context and additional feature from dependency or long-term dependency might be able to solve this issue.

References

1. 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. CoRR, abs/1804.09541, 2018. URL http://arxiv.org/abs/1804.09541.
2. Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. CoRR, abs/1807.03819, 2018. URL http://arxiv.org/abs/1807.03819.