On Scalar Embedding of Relative Positions in Attention Models

Technical Appendix

Artificial Task

In this section, we give a detailed illustration of the data generation on Reber and Process-50 tasks. What is more, we give the heat maps of the prior score, context score, and final attentive score for a concrete example on the Process-50 task.

Reber We generate embedded reber grammar sequence as shown in Figure 1. The dotted sub-figure is one Reber grammar module. In this work, the last character of the sequence is identical to the second character (T or P) of the sequence.

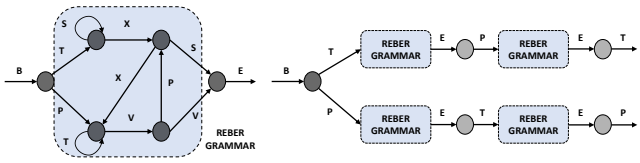


Figure 1: Embedded Reber grammar sequence.

Process Classification We generate binary sequence using the 2-state Markov chain with transition pattern shown in Figure 2. The stationary state distribution is: $[\frac{\beta}{\beta+\alpha}, \frac{\alpha}{\beta+\alpha}]$. To make sure the model focus on learning the state transition information, we generate two kinds of sequences with the same one frequency and zero frequency. The data generation setting is as follows: The positive parameter is $\alpha = 0.2, \beta = 0.4$; The negative parameter : $\alpha = 0.3, \beta = 0.6$; (0, 1) stationary distribution: $(2/3, 1/3)$.

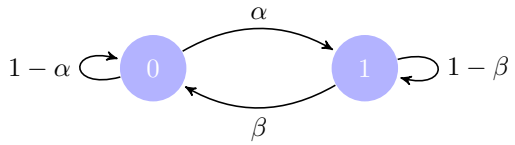
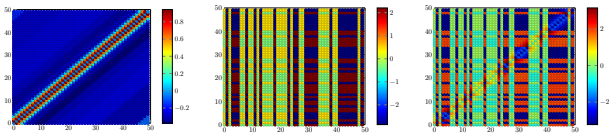
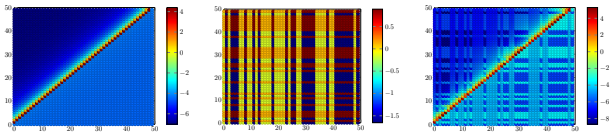


Figure 2: State transition



(a) Prior score (b) Context score (c) Final score

Figure 3: Heat map of prior score, context score and final score learned by T5.



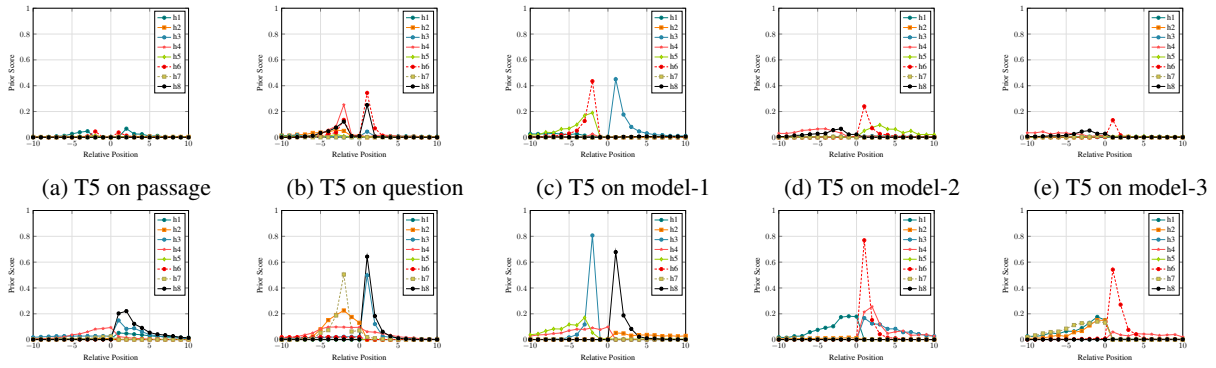
(a) Prior score (b) Context score (c) Final score

Figure 4: Heat map of prior score, context score and final score learned by AT5.

For Process-50, the sequence length is 50. Figure 3 and Figure 4 show the heat maps for the prior score, context score, and final score of one head in T5 and AT5 on one concrete sequence *00010011010000000110100000000000110111100000000100* with a positive label in the test dataset. From Figure 3a and Figure 4a, we can find that both T5 and AT5 have learned the useful prior score pattern, which encourages the model to focus on neighbor tokens. However, we can conclude that AT5 learned a more informative and powerful prior score than T5 by comparing Figure 3c and Figure 4c. Finally, in our experiment, T5 makes a wrong classification for this sequence, while AT5 makes the correct classification.

Quantitative Analysis on Question Answering

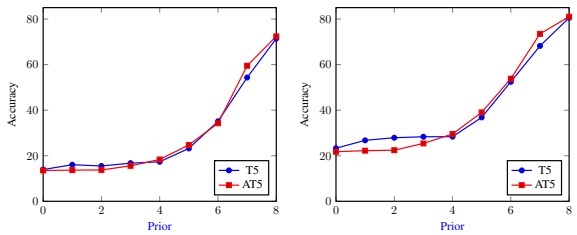
Due to the page limitation, we do not give the quantity analysis for question answering (QA). In this section, we give a detailed quantitative analysis of QA task. Specifically, QANet includes five encoder modules: passage encoder layer, question encoder layer, and three model encoder layers, where the input for three model encoder layers is the output of passage-question attention. Specifically, the concatenation of the first



(a) T5 on passage (b) T5 on question (c) T5 on model-1 (d) T5 on model-2 (e) T5 on model-3

(f) AT5 on passage (g) AT5 on question (h) AT5 on model-1 (i) AT5 on model-2 (j) AT5 on model-3

Figure 5: The prior probabilities of each head at passage self-attention layer, question self-attention layer, and three model encoder self-attention layer learned by T5 and AT5.



(a) QA EM (b) QA F1

Figure 6: Prior probability ablation study for T5 and AT5 on question answering. EM: exact match.

and second model encoder outputs is used to predict the start position of the answer. And the concatenation of output of first and third model layers is used to predict the end position of the answer.

Figure 5 shows the prior score of five modules in QANet for T5 and AT5 on SQuAD-v1. On the QA task, both T5 and AT5 successfully learned the difference in relative positions. However, AT5 learned better distribution for the prior probabilities and achieved higher EM and F1 than T5.

The ablation study in Figure 6 illustrates the changes of the test performance for T5 and AT5 when the heads are added into the QANet one by one according to their maximum probabilities among all relative positions (from heads with lower maximum probabilities to heads with higher maximum probabilities). From the Figure 6, we find that the models without adding prior score perform worse. This study indicates that positional information is crucial for QANet. This phenomenon is also found at Process-50 and MT (machine translation) tasks. From the figure, we observe that when the lower-probability heads are added into the QANet, the T5 performs better than AT5. When the higher-probability heads are added into the QANet, the performance of AT5 improves faster than T5 and eventually outperforms T5. These phenomenons are identical to the ablation study on other tasks.