Supplementary Material: Deep Attentive Ranking Networks for Learning to Order Sentences

Detailed Hyperparameters

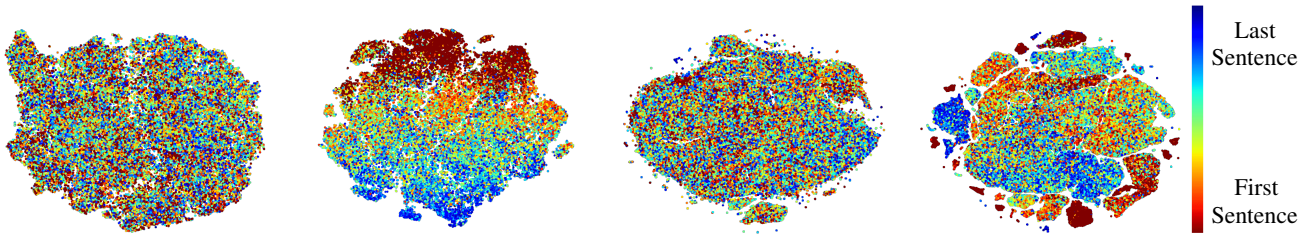
In this section, we provide more details about the hyperparameters. For sentence encoder, the pre-trained $\text{BERT}_{\text{BASE}}$ model with 12 Transformer blocks, the hidden size as 768, and 12 self-attention heads has been used. The feed-forward intermediate layer size is 4×768, i.e., 3072. All the layers in the sentence encoder and the paragraph encoder use dropouts with probability 0.1 and gelu activation (Hendrycks and Gimpel 2016). We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. However, we take initial learning rates of $5e-5$ for both the BERT based sentence encoder and the Transformer based paragraph encoder, and $5e-3$ for the feed-forward neural network decoder. The paragraph encoder is a Transformer Network having 2 Transformer blocks, with hidden size 768 and a feed-forward intermediate layer size of 4×768, i.e., 3072. We experimented with 2, 4 and 8 Transformer blocks on ROCStory, and we found that our model was performing well on ROCStory dataset with 4 Transformer blocks. For arXiv dataset we experiment with 2 and 8 Transformer blocks. The Transformer gives 768-dimensional sentence representation. The decoder is a five layer feed-forward network with ReLU non-linearity in each layer with hidden size of 200, and a 1-dimensional output layer for the score. We experimented with various batch sizes and found 400 to be a reasonable number.

t-SNE Embeddings Visualization

Fig. 1 shows the t-SNE embeddings visualtions of sentence representations obtained from pre-trained BERT (Devlin et al. 2018)/sentence encoder (before training) and trained sentence encoder for arXiv abstracts and NSF abstracts datasets. Clearly, we can see that the representations of sentences in first and last positions in the ordered paragraph are clustered together, respectively. All the visualizations are shown for sentences from the unseen test set, showing better generalizability of our model. These visualizations correspond to the model (among our approaches) showing the best results in terms of Kendall's tau score.

Visualization of Word Attention

To visualize the word level interaction, we show the self attention among words for a sentence in a paragraph belonging to the test set from ROCStory dataset (Mostafazadeh et al. 2016) in Fig. 2 and Fig. 3. Our model's sentence encoder is able to focus on the first and last tokens in a sentence as can be seen in Fig. 2. Since, ROCStory corpus paragraphs have some clear sentiment attached to them, we can see in Fig. 3 that one of the attention heads tries to focus on a word with a strong sentiment.



(a) arXiv: Untrained (b) arXiv: Trained (c) NSF: Untrained (d) NSF: Trained

Figure 1: t-SNE embeddings of sentence representations arXiv abstracts and NIPS abstracts datasetss on sentence ordering task. Colors correspond to position of the sentences in the original paragraph. **Untrained:** Sentence embeddings before fine-tuning (same as pre-trained BERT). **Trained:** Sentence embeddings after fine-tuning.

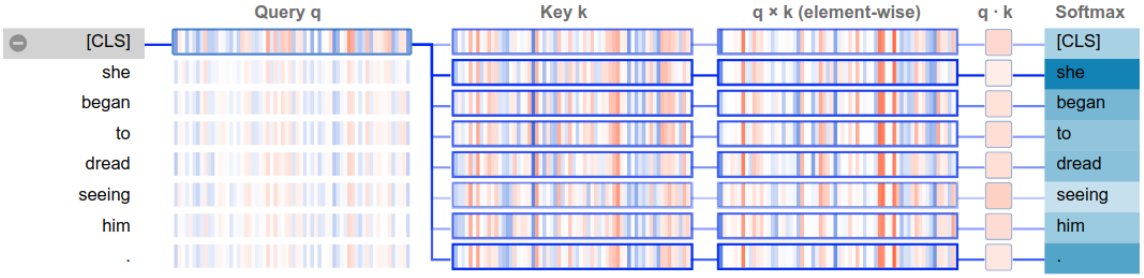


Figure 2: Visualization of the all neuron values for query and key for computing self-attention among words in the sentence encoder layer. We can see that this particular attention head is focusing on the first and last tokens in the sentence.

Figure 3: Visualization of the all neuron values for query and key for computing self-attention among words in the sentence encoder layer. This attention head is focusing on the word *dread* indicating that it is detecting the word corresponding to a strong sentiment.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hendrycks, D., and Gimpel, K. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.