Effective Approaches to Attention-based Neural Machine Translation

Link

arXiv

Summary

- The paper studies two novel attention based models with simplicity and effectiveness in mind

  1. Global attention where all source words are attended. A variable-length alignment vector $\mathbf{a}_t$ whose size equals the number of time steps on source side is calculated by comparing current target hidden state $\mathbf{h}_t$ and each source hidden state $\overline{\mathbf{h}}_s$. Given the alignment vector the context vector c_t is computed as the weighted average over all the source hidden states.

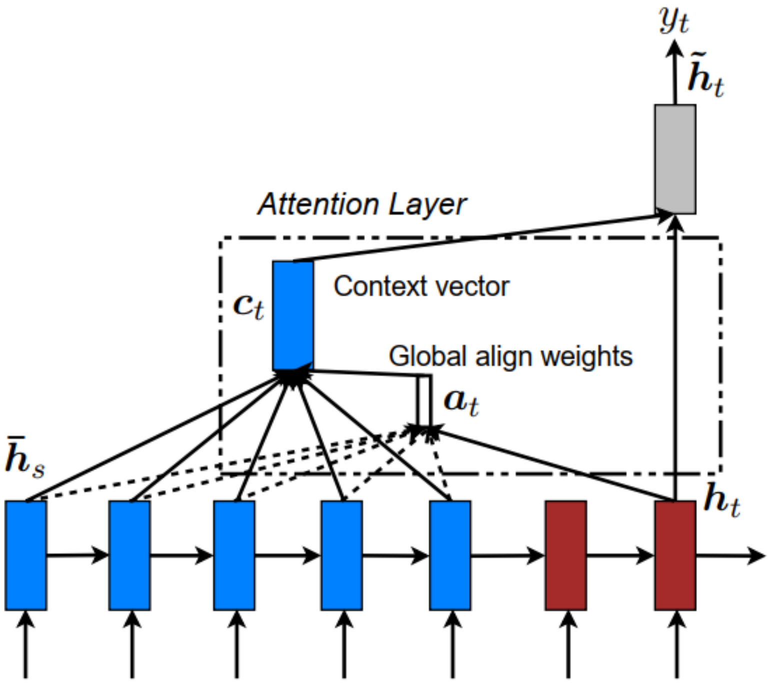

Figure 1: Global attention model

  2. Global attention model which chooses to focus only on a small subset of source input positions per target words. This avoids the expensive computation incurred by global attention and is easier to train than hard attention which selects only one position per target and is non-differentiable. For local attention first an aligned position p_t is generated for each target word at time t. The context vector is then derived as weighted average of the source hidden states withing window $[p_t - D, p_t + D]$.

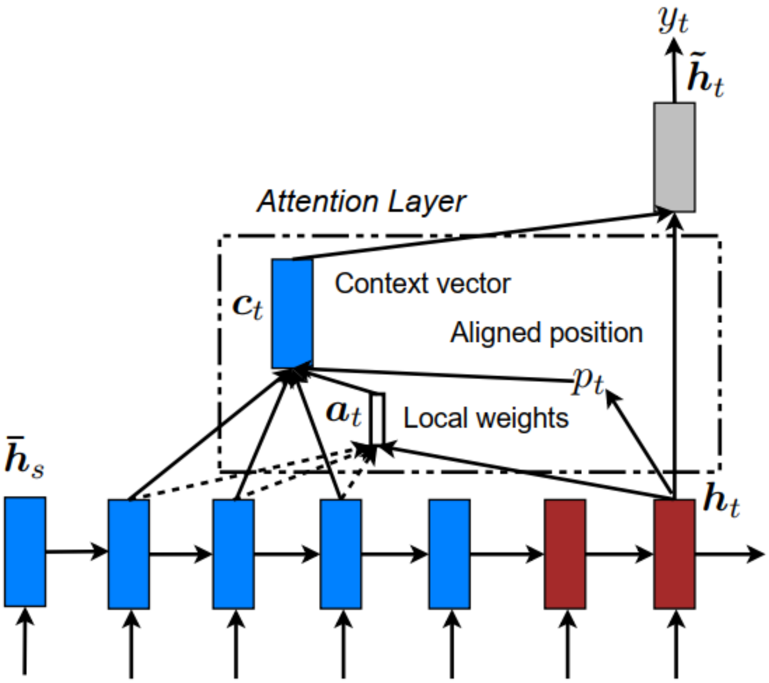

Figure 2: Local attention model

- To make alignment decisions dependent on past decisions, they use an input-feeding approach in which attention vectors are concatenated with inputs at the next time steps.

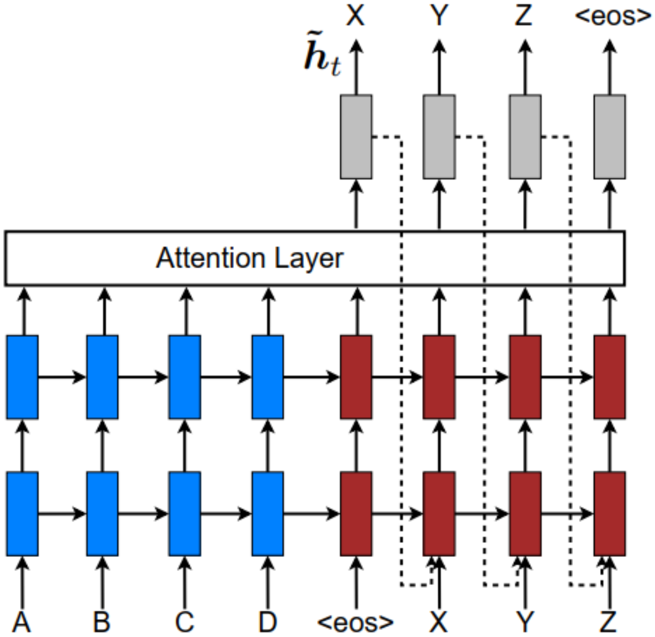

Figure 3: Input-feeding approach

- Local attention approach perform better than global attention approach which works significantly better than no attention approach. Source sentence reversing and input feeding approach also improves performance. Unknown replacement approach (replacing words not present in vocabulary with $< unk >$ further improves result. This shows that the attention models can learn useful alignment for unknown words.