A TIME-RESTRICTED SELF-ATTENTION LAYER FOR ASR

Daniel Povey[1,2], Hossein Hadian[1], Pegah Ghahremani[1], Ke Li[1], Sanjeev Khudanpur[1,2]
[1]Center for Language and Speech Processing,
[2]Human Language Technology Center of Excellence, Johns Hopkins University.

Layer Overview

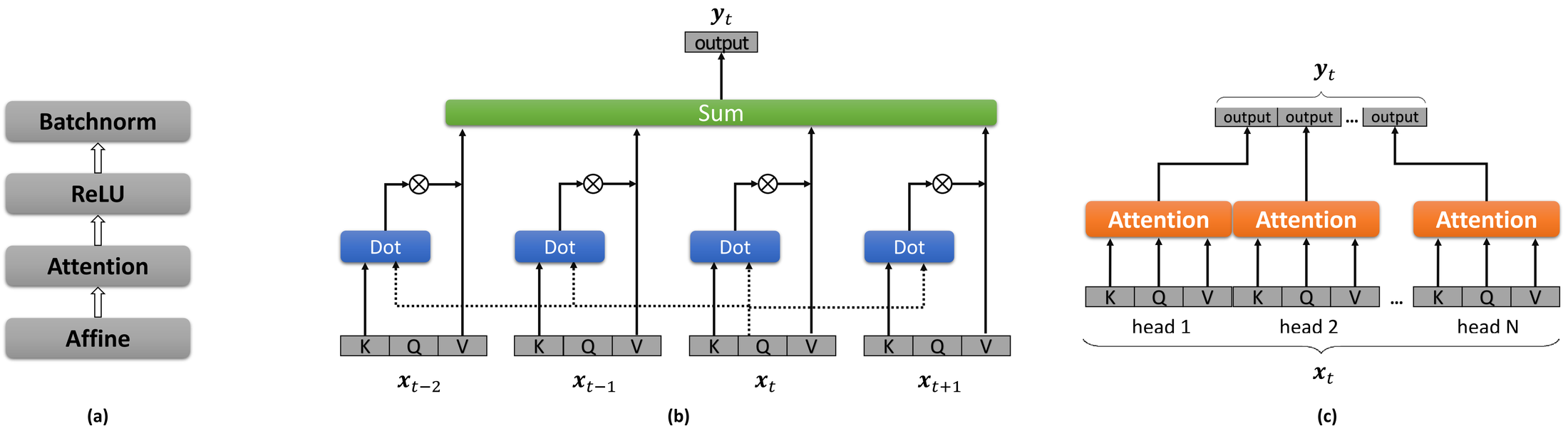


Figure 1: (a) The proposed self-attention layer and the comprising components. (b) A single-head attention component. Left and right context sizes are 2 and 1 respectfully. For clarity, positional-encoding and the softmax (which is applied to the dot-products) are not shown. (c) A multi-head attention component (which is used in the attention layer) using single-head attention blocks. K, Q, and V respectively mean key, query, and value.

Background

- Attention: A mechanism which allows the network to focus on the relevant part of the input at each time step.
- Self-attention: An attention mechanism where the input and output sequence lengths are the same.
- Attention-based models have recently been successfully applied to a variety of tasks such as machine translation, caption generation, and phoneme recognition.
- Multi-head attention: An attention layer with multiple heads which can jointly attend to different subspaces of the input representation. Recently proposed and used for NMT [1].

Time-restricted Self-Attention

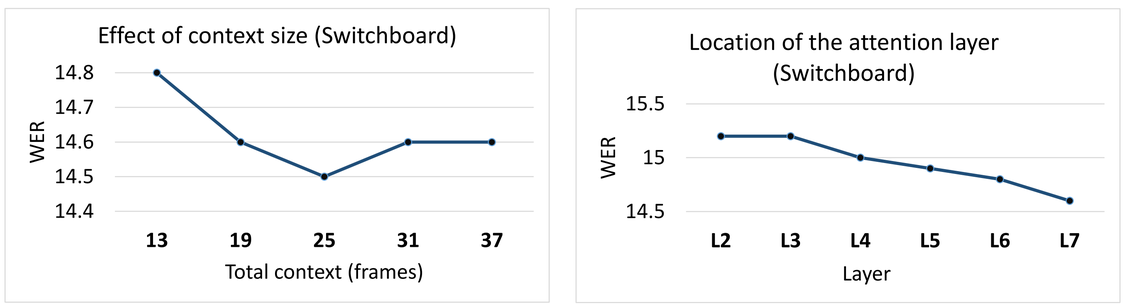
- We propose a self-attention layer which is time restricted, i.e. it is suitable for ASR where the input sequence can be very long (also see Figure 2 on the right).
- The attention component interprets its input $\boldsymbol{x_t}$ as being three things appended together: $\boldsymbol{q_t}$ and $\boldsymbol{k_t}$ and $\boldsymbol{v_t}$ which are the query, key and value respectively. The order in which we divide the input to key/query/value is not important as long as we do it consistently.
- The output $\boldsymbol{y_t}$ is a weighted sum (over time) of the values $\boldsymbol{v_t}$, where the weights are determined by dot products of the queries and the keys (normalized via softmax):

$$\boldsymbol{y_t} = \sum_{\tau=t-L}^{t+R} c_t(\tau)\boldsymbol{v_\tau} \qquad (1)$$

- Since our attention mechanism is time-restricted, we use a one-hot encoding of the relative position of τ versus t to enable positional encoding.
- This is not "attention-based speech recognition" – we are not using attention to replace the left-to-right alignment of the HMM. It is simply an alternative to TDNN and LSTM layers in our model topology.

Initial Experiments

- The number of heads did not significantly affect the performance. We used 30 in the experiments.
- We found that a key/value dimension ratio of 0.5 works best.



Analysis

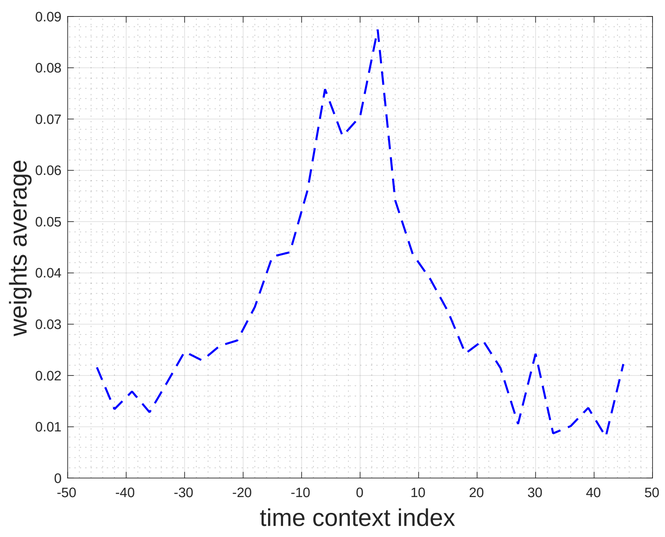


Figure 2: Weight vector $\boldsymbol{c_t}$ averaged over all heads for an attention layer with 150 heads and context $[-45, 45]$.

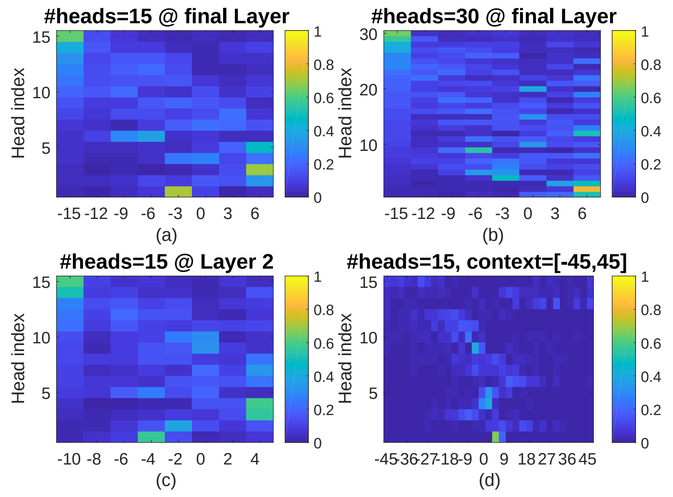
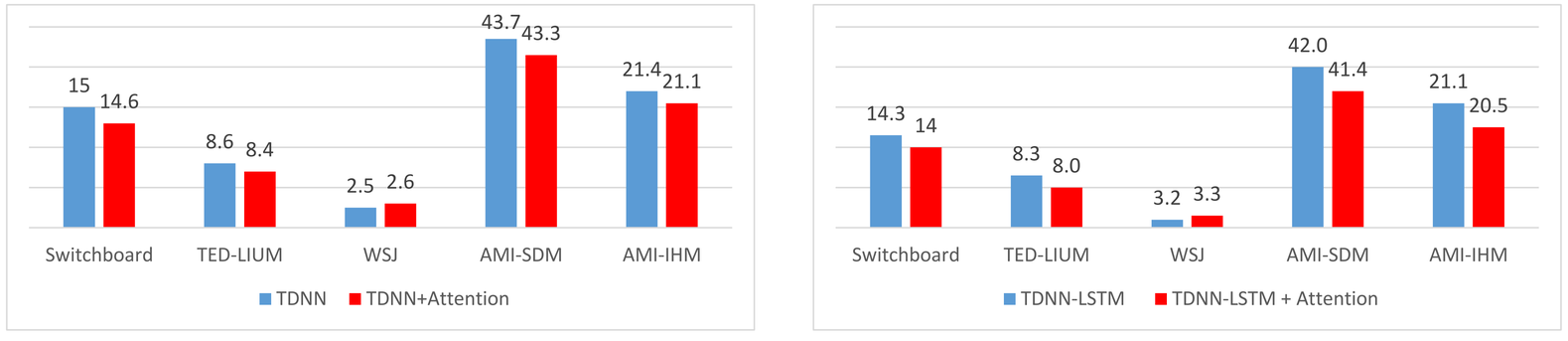


Figure 3: Attention weight vector $\boldsymbol{c_t}$ for different attention configurations. The horizontal axis shows time.

Results



Conclusion

We introduced a time-restricted self-attention layer and used it in our state-of-the-art LF-MMI neural networks, replacing a TDNN or LSTM layer.

- We found that using a single self-attention layer towards the end of the network can improve the WER by 0.2-0.6 in our TDNN and TDNN-LSTM setups (except on WSJ).
- In TDNN-LSTMs, it can also speed up decoding by 20%.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017