CS F425: Deep Learning

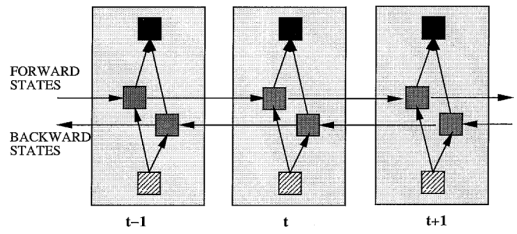
25 Bi-dir RNN Transformer

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Mar 24, 2023 ON-CAMPUS Campus @ BITS-Pilani [Jan-May 2023]

http://ktiwari.in/dl

Bidirectional LSTM/RNN [1]



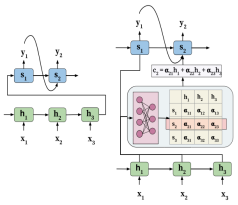
- No limitation on using input just up to a preset future frame
- Simultaneous training in positive and negative time direction
- Modified one can estimate conditional posterior probability of complete symbol sequences without assumption on distribution
- Useful to real data, classification experiments

[1] **cite. 8304** Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks.", IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.

Attention Model [2]

Our visual system tends to focus selectively on some parts of the image, while ignoring other irrelevant information

- Attention was first introduced in Machine Translation
- Attention weights are learned by additional feed forward NN



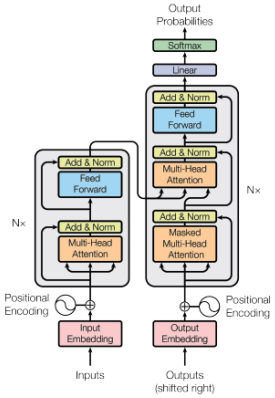
[2] **cite. 27787** Bahdanau, Kyunghyun, and Yoshua. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
Chaudhari, Sneha, et al. "An attentive survey of attention models." ACM Tr Intelligent Systems and Technology (TIST) 2021

Transformer [3]

Sequence transduction models include an encoder and a decoder that are connected through attention

- Transformer is simple, solely based on attention mechanism
- It takes significantly less time to train
- Simultaneous input of sentence handles issus related to exploding/vanishing gradients

Training is sequential and end-to-end



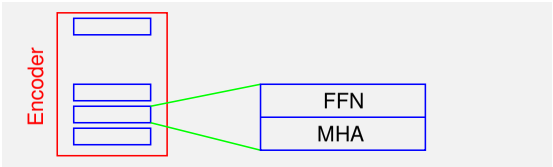
[3] **cite. 69087** Vaswani, et al. "Attention is all you need.", Advances in neural information processing systems (NIPS) 2017.

Overview

- Has Encoder (self attention) and Decoder (self+cross attention)
- Issue with sequential and recurrent data
- Things are squential and dependent

  `Dr Kalam was president. He also worked at DRDO`

  Word2Vec[4] or GLOVE[5] can used to get representations (512D)
- 6 blocks of feed forward network, and multi head attention

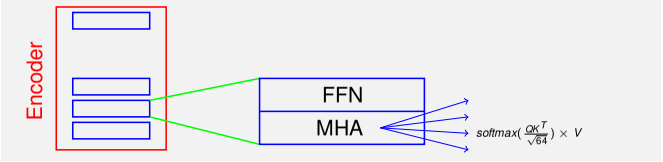


[4] **cite. 33782** Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
[5] **cite. 33108** Pennington, Jeffrey, et al. "Glove: Global vectors for word representation." Proceedings of the conference on empirical methods in natural language processing (EMNLP). 2014.

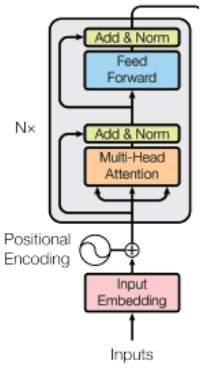
Multi Head Attention (MHA)

- Multiple set of three dedicated NN circuit Query, Key, and Value
- Converts 512D vector to 64D size
- Attention? connection strength between different words

  `The dog eat food becaue it was hungry.`
- Self attention and cross attention
- Consider vectors X_{it}^{Q}, X_{it}^{K}, and X_{it}^{V} for `I am good.`
- Word embedding $\rightarrow$ $softmax(\frac{QK^T}{\sqrt{64}}) \times V$
- Concatinate multiple ones and apply FC to get 64D vector



Transformer Multi Head Attention (MHA)



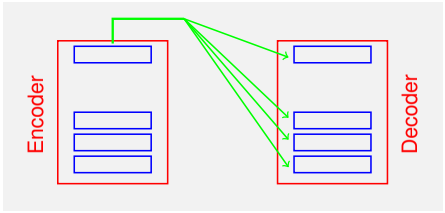
Output of MHA would be same for different permutation of the same sentence

- Positional embedding is added

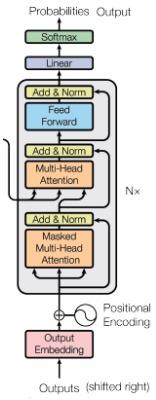
$$X = embedding + P$$

- P being sinusoidal encoding
- Residual connections are added and layer normalization is applied

Decoder



- Masking is needed while training for self attention
- Not for cross attention
- Residual connections and layer normalization are added



Thank You!

Thank you very much for your attention!