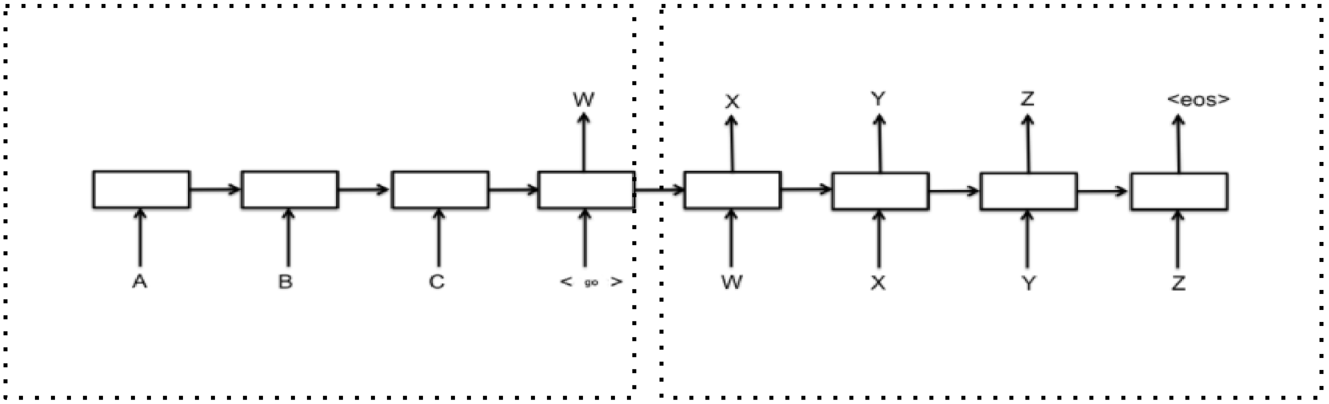
HW2 Report

Sisheng Liang

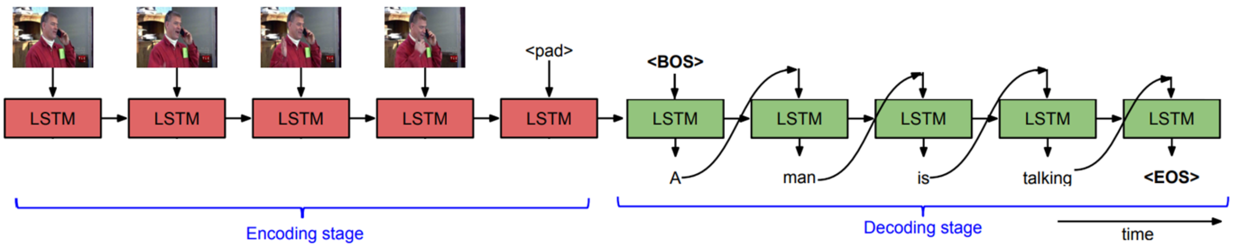
https://github.com/liang-cu/hw2_1.git

1. **Basic encoder decoder model:** The video caption problem is a sequence-to-sequence modeling problem. The basic model consists of an encoder that processes the input feature maps and a decoder that generates the caption. The encoder generates the latent space that is fed to the decoder. The word is encoded into vector and fed to the decoder with one-to N encoding. Some basic word encoding techniques are show as follows:
    a. Dictionary - most frequently word or min count
    b. Other tokens <PAD>, <BOS>, <EOS>, <UNK>
    -<PAD> : Pad the sentence to the same length
    -<BOS> : Begin of sentence, a sign to generate the output sentence.
    -<EOS> : End of sentence, a sign of the end of the output sentence.
    - <UNK> : Use this token when the word isn't in the dictionary or just ignore the unknown word.

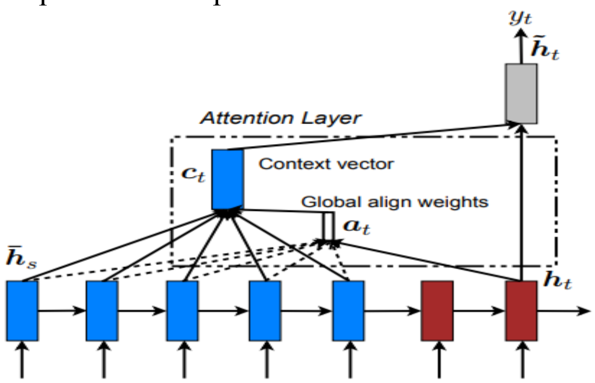
The generated caption vector is turned back into word caption. As for my implementation, one layer LSTM is used to implement the encoder and one-layer GRU is used to implement the decoder.



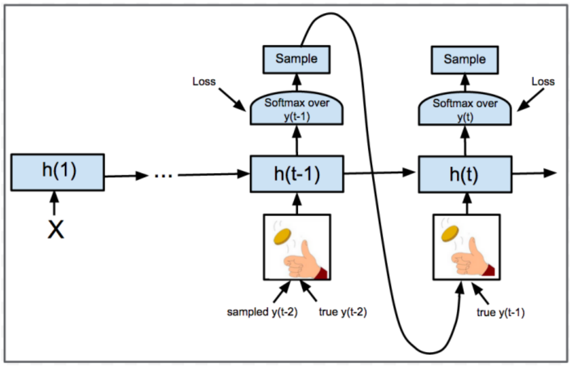
Encoder Decoder



2. **Attention**: Attention is added to the decoder to improve the performance of the basic decoder. As shown in the following picture from the slides. The attention layer can learn different weights for different part, therefore it put more weight from the more important part. This could improve the performance of the decoder. Intuitively, the attention in the decoder decides parts of the source sequence to pay attention to [1]. The attention layer is added to the decoder part of the implementation.



3. **Schedule Sampling** The current approach to training them consists of maximizing the likelihood of each token in the sequence given the recurrent state and the previous token. At inference, the unknown previous token is then replaced by a token generated by the model itself. This discrepancy between training and inference can yield errors that can accumulate quickly along the generated sequence. The schedule sampling is a curriculum learning strategy to gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead [2].



Evaluation:

Case 1: the BLEU score of this implementation is 0.653 with a learning rate of 0.001, data dropout ration 0.3.

Case 2: the BLEU score of this implementation is 0.691 with a learning rate of 0.001, data dropout ration 0.1.

[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
[2] Bengio, S., Vinyals, O., Jaitly, N. and Shazeer, N., 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, *28*.