Grayson Byrd
HW2 Report
Github: https://github.com/gbbyrd

Project Description:
The purpose of this project is to develop an AI model that can assist in video captioning. The model will be shown a video and create a caption describing what is happening in the video. This is a very interesting problem in deep learning, and there are many different architectures and combinations of micro architectures that can solve this problem to varying degrees of success. At its core, this is a sequence to sequence problem. A sequence to sequence problem is characterized by both the input and output being sequences. For this particular case, the sequences will be of variable lengths and we will discuss techniques to solve this problem later on in the report.

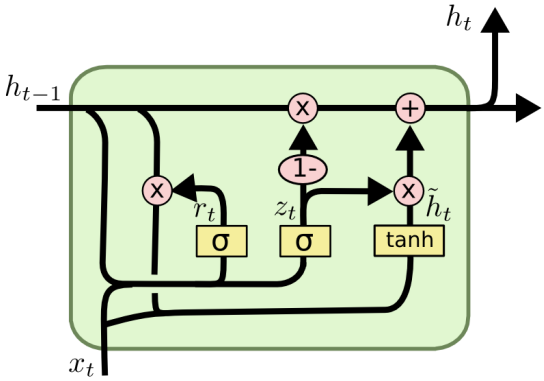
Technology Review

There are many techniques and technologies that have been researched in the past to improve the performance of models on sequence to sequence problems. I will discuss the specific techniques used in this model in this section.

Architecture

RNN Cell Type
The basis for my model is the Gated Recurrent Unit (GRU). This GRU is appropriate for encoder-decoder models, which make up the majority of models used to solve sequence to sequence problems. The diagram and equations for the GRU can be seen in Figure 1 below:

Figure 1: Diagram and Equations of Gated Recurrent Unit (GRU)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

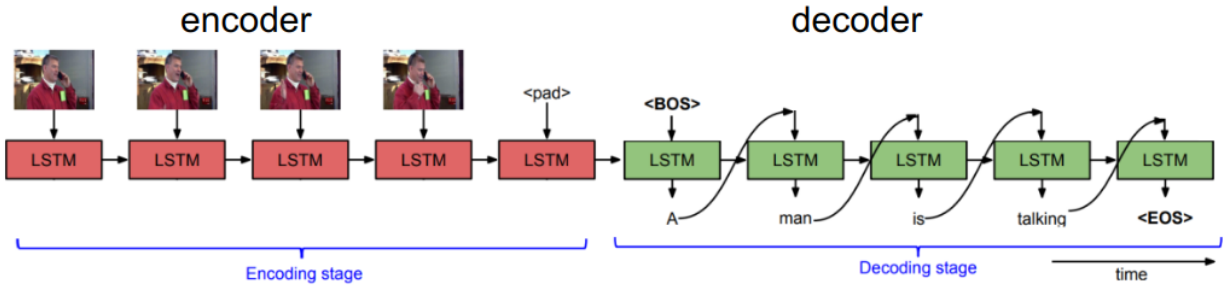
The GRU has a few features that make it favorable for encoder-decoder models, namely its ability to pass on Long Short Term Memory. The fundamental property of the GRU is its hidden state. This hidden state is passed along temporally as the model encodes and decodes sequences. The value of this hidden state, which is often passed as a vector, is decided by

multiple “gates”. The parameters of these gates filter out unimportant information and, after ample training, allow the GRU to contain a hidden state that only holds valuable information that can be used to predict the next word based on the previous encoded words and words predicted by previous time steps in the decoder.

Encoder Decoder Model

Adding multiple GRUs in a sequence together allows us to create more complicated architectures. Figure 2 below depicts an overview of the encoder decoder network that was used in this project.

Figure 2: Basic Overview of the Encoder-Decoder network used in this project



*** Note: This architecture replaces the LSTMs with GRUs ***

In the encoding stage, image frames encoded into a 4096 dimensional vector space are passed as input into the GRU. Once all of the frames have been temporally decoded, we are left with a hidden output from the last GRU in the encoder chain. This hidden state is then passed into the first GRU of the decoder chain. In this way, we effectively convey important information from every frame in the encoding sequence by leveraging the Long Short Term Memory property of the GRU. In the decoding phase, a <BOS> tag is passed as input into the first GRU, and then in each subsequent GRU the output of the previous GRU (which is the predicted word) is then passed as input. This continues until a GRU produces the <EOS> tag as the predicted word. The predicted sentence is the temporal combination of every predicted word in the decoding sequence.

Training Techniques

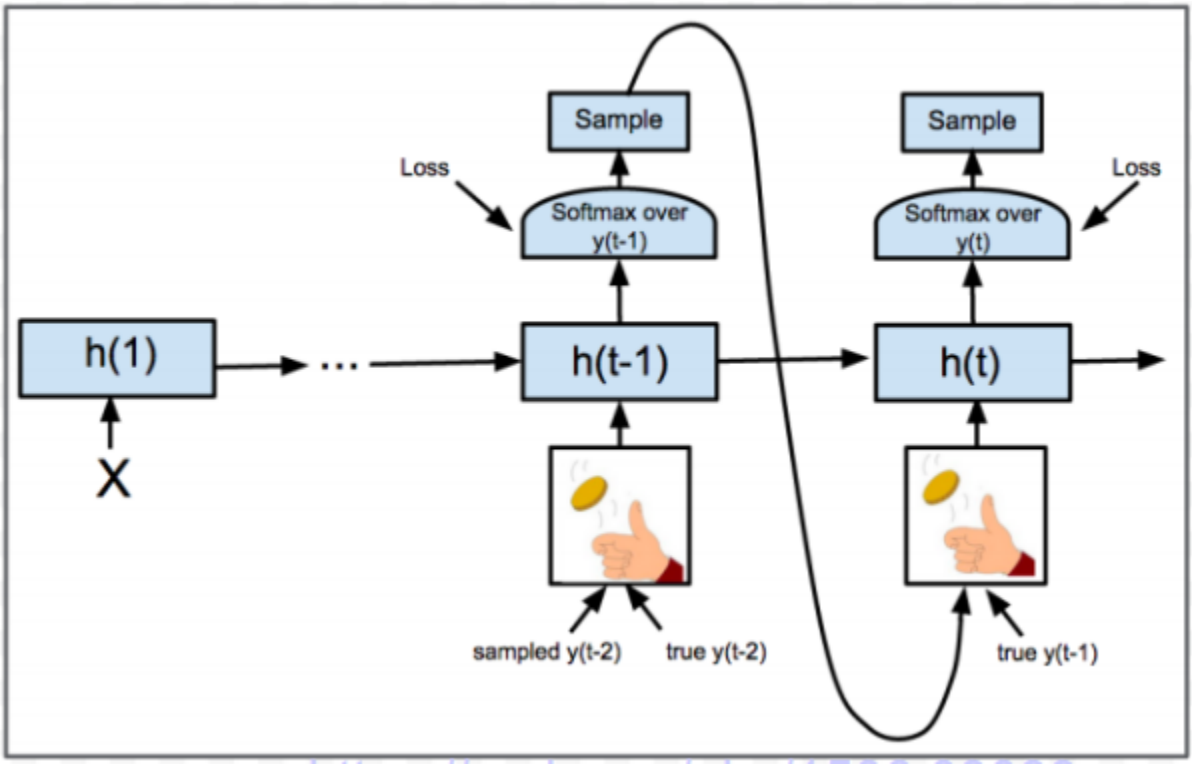
No loss is calculated during the encoding portion of the model for this project. In this model, the cross entropy loss was calculated comparing the predicted sentence with the target sentence. The model parameters are then adjusted using the Adam optimizer. A few more advanced techniques were used to improve the performance. These are explained below.

Scheduled Sampling

Scheduled sampling is an additional training measure used to improve the performance of this model. Scheduled sampling is a way to introduce gaussian noise to the system. The general idea behind scheduled sampling is to, at each time step in the decoding sequence, randomly determine if you will pass the predicted output from the previous GRU or the ground truth word that should have been output by the previous GRU. If we passed the ground truth at every

single time step in the decoding sequence, the model would not be able to learn how to produce correct sequences as it would not be learning to form a sentence as a whole, and instead would be learning to predict a word based on a previous word. However, randomly adding target words into the decoding sequence has been found to improve the training performance of the model and reduce training time. This concept is explained in Figure 3 shown below.

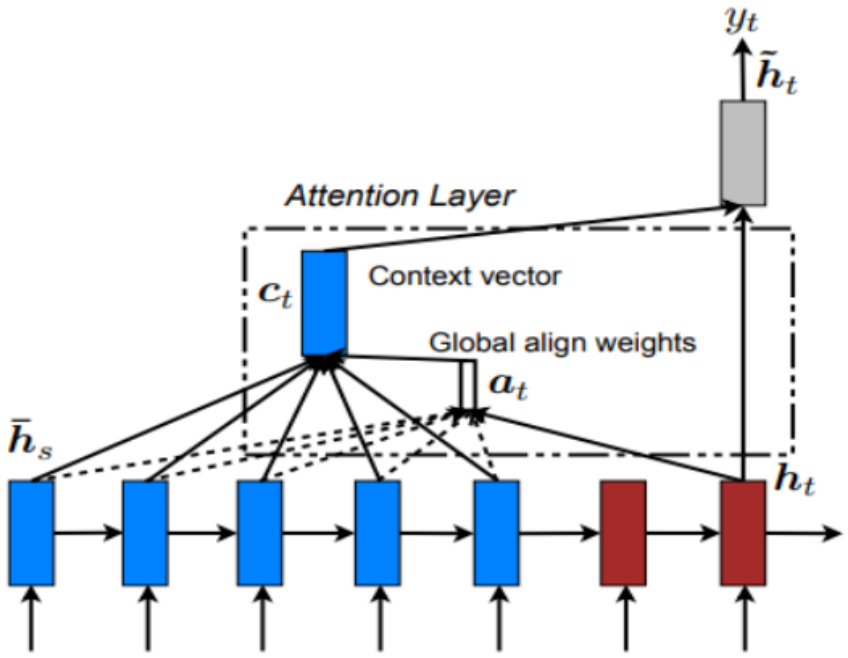
Figure 3: Diagram of scheduled sampling



Attention

Attention is a technique that is loosely based on how humans interpret sequences when reading and watching videos. When a human is reading a sentence, they only focus on a few key words to help them decode the sentence. Attention is at its core an vector that describes the importance of each of the outputs from the encoding sequence. At each time step in the decoding process, this attention is concatenated with the previous hidden state and passed as an input to the GRU to convey information about the importance of each encoded frame in the encoding portion of the model. A diagram of this technique is shown in Figure 4 below.

Figure 4: Diagram depicting the attention mechanism used in my model



Evaluating Performance

In order to measure the performance of the model on the testing set of the data and to gage its ability to generalize to videos outside of its training data, the BLEU evaluation was used. Figure 5 below is an overview of how the BLEU@1 evaluation is calculated.

Figure 5: BLEU@1 evaluation algorithm

Precision = correct words / candidate length

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

where c = candidate length, r = reference length

BLEU@1 = BP * Precision

e.g.:

Ground Truth : *a man is mowing a lawn*

Prediction : *a man is riding a man on a woman is riding a motorcycle*

BLEU: 1 * 4/13 = 0.308

The goal of this project is to create a model that can reach the defined baseline on the BLEU@1 evaluation criteria. The defined baseline for this project is: **0.6 average on the testing captions.**

Results

This model was trained for 100 epochs, evaluating each of the captions for each of the videos in the training set. The total training time was around 2 hours on the DGX2 (a subset of Palmetto).

The total performance of this model was 0.698110312762288 when using the BLEU@1 to evaluate.

References:

Bengio, Samy, et al. “Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks.” *ArXiv.org*, 23 Sept. 2015, https://arxiv.org/abs/1506.03099.

Bleu: A Method for Automatic Evaluation of Machine Translation. https://aclanthology.org/P02-1040.pdf.

Vaswani, Ashish, et al. “Attention Is All You Need.” *ArXiv.org*, 6 Dec. 2017, https://arxiv.org/abs/1706.03762.

Venugopalan, Subhashini, et al. “Sequence to Sequence -- Video to Text.” *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, https://doi.org/10.1109/iccv.2015.515.