CPSC 8430: Deep Learning HW2 report

Name: Siqi Zheng

Date: 03/24/2022

Director: Feng Luo

Introduction

This project is required to do the video caption generation using the sequence 2 sequence model. Normally, the input of the model should be a short video and output of the model be the corresponding caption that depicts the video. However, in this project, the feature of the video has already been extracted by the CNN so that the input of the model becomes the feature of the video which is an 80×4096 NumPy array. To get the accurate results, attention and schedule sample methods have been used in this model and BLEU evaluation method is used as the evaluation criterion. After several times attempts, the model can generally give a relative accurate caption.

Data processing

In this project, the video data has already been extracted, the caption data is the only data required to pre-process. Each video has a group of the captions, in this project the shortest caption is selected. After a series of operations such as removing punctuation marks and changing uppercase to lowercase, the desired title is obtained. To encode and decode the label, the index to word and word to index vocabulary are obtained and stored in the /.pkl files, in the meanwhile the processed data is stored in the /.cvs files. All these are in the processed data folder for later use.

Sequence to sequence model

The S2VT model used in this project is based on Subhashini Venugopalan paper, it is a two layers LSTM, the detail structure of the model is as the figure 1. The top LSTM layer (colored red) models visual feature inputs. The second LSTM layer (colored green) models language given the text input and the hidden representation of the video sequence. < BOS> is used to indicate begin-of-sentence and <EOS> for the end-of-sentence tag. < pad> is used when there is no input at the time step.

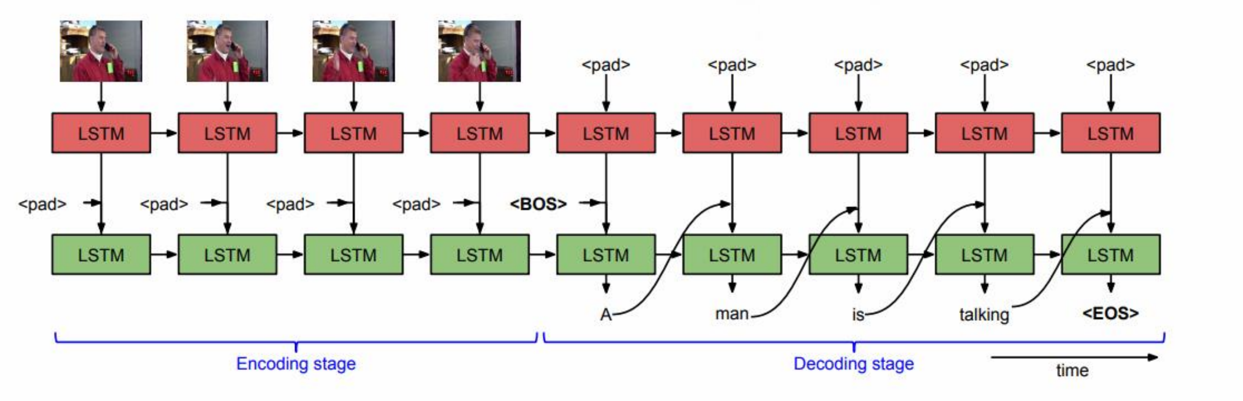


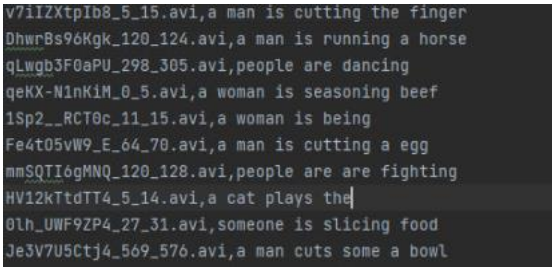
Fig. 1. structure of the S2VT

As for programming the model, “torch.nn.LSTM” function is used here. The following table describes the hyper parameter used in this model and results after training.

Table 1. sequence to sequence model parameters

Batch size	64	**Learning rate**	4e-4
Vocabulary size	1066	**Learning rate decay**	5e-3
Hidden layer dimension	512	**epoch**	200
Word dimension	512	**Loss**	0.033
Loss function	nn.NLLLoss	**Blue**	0.598

Partial results of the last test dataset are shown in the following figure 2.



```
v7iIZXtpIb8_5_15.avi,a man is cutting the finger
DhwrBs96Kgk_120_124.avi,a man is running a horse
qLwgb3F0aPU_298_305.avi,people are dancing
qeKX-N1nKiM_0_5.avi,a woman is seasoning beef
1Sp2__RCT0c_11_15.avi,a woman is being
Fe4tO5vW9_E_64_70.avi,a man is cutting a egg
mmSQTI6gMNQ_120_128.avi,people are are fighting
HV12kTtdTT4_5_14.avi,a cat plays the
0lh_UWF9ZP4_27_31.avi,someone is slicing food
Je3V7U5Ctj4_569_576.avi,a man cuts some a bowl
```

Fig. 2. S2VT result

Sequence to sequence model with attention model

In the traditional S2VT model, the encoder reads the feature only once and encodes it. At each timestep, the decoder uses only the corresponding embedding and produce a new word. However, when people understand the content, they will extract information from the entire object to understand not just a word or a feature. To simulate this situation, attention mechanisms is used in the S2VT model. The figure 3 shows the structure of the S2VT model with the attention.

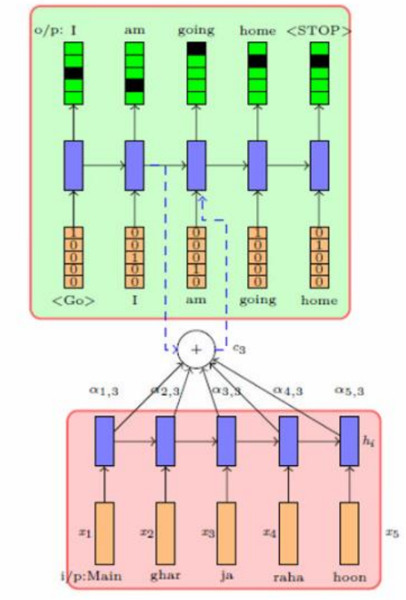


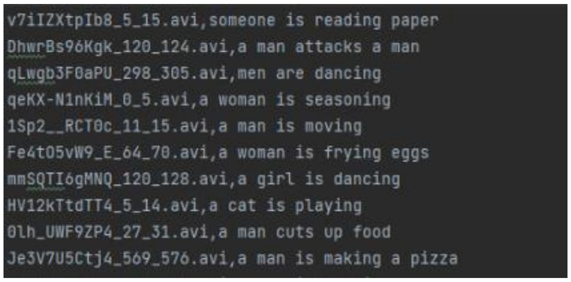
Fig. 3. S2VT with attention

The following table shows the parameter and results of the model

Table 2. parameter of the S2VT model with attention

Batch size	64	**Learning rate**	4e-4
Vocabulary size	1066	**Learning rate decay**	5e-3
Hidden layer dimension	512	**epoch**	200
Word dimension	512	**Loss**	0.09
Loss function	nn.NLLLoss	**Blue**	0.606

Partial results of the last test dataset are shown in the following figure 4.



```
v7iIZXtpIb8_5_15.avi,someone is reading paper
DhwrBs96Kgk_120_124.avi,a man attacks a man
qLwgb3F0aPU_298_305.avi,men are dancing
qeKX-N1nKiM_0_5.avi,a woman is seasoning
1Sp2__RCT0c_11_15.avi,a man is moving
Fe4tO5vW9_E_64_70.avi,a woman is frying eggs
mmSQTI6gMNQ_120_128.avi,a girl is dancing
HV12kTtdTT4_5_14.avi,a cat is playing
0lh_UWF9ZP4_27_31.avi,a man cuts up food
Je3V7U5Ctj4_569_576.avi,a man is making a pizza
```

Fig. 4. ATT-S2VT model result

Sequence to sequence model with attention and schedule sample model

Refer to the content in the paper “Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks”, there are two choices of inference input to the second layer, ground truth or the last time step output and it proposed a sampling mechanism to decide which input is selected during training. At the beginning of training, the model is not mature, sampling from the model would yield a random token and converge will be slow. So, selecting more often the ground truth will be helpful for fast converge. At the end of the training, selecting the last time step output will be helpful because it is the true inference situation. In this project, the exponential decay schedule is used, and the parameter of the k is chosen to be 0.97.

$$\varepsilon_i = k^i,$$

$$k = 0.97, i \ is \ the \ number \ of \ epoch,$$

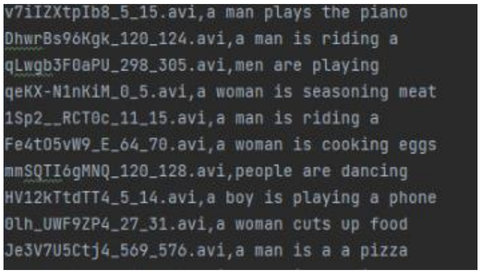
$$\varepsilon_i is \ the \ possibility \ to \ choose \ the \ ground \ truth$$

The following table shows the parameter and results of the model

Table 3. parameter of the S2VT model with attention and schedule sample

Batch size	64	**Learning rate**	4e-4
Vocabulary size	1066	**Learning rate decay**	5e-3
Hidden layer dimension	512	**epoch**	200
Word dimension	512	**Loss**	0.072
Loss function	nn.NLLLoss	**Blue**	0.632

Partial results of the last test dataset are shown in the following Fig. 5.



```
v7iIZXtpIb8_5_15.avi,a man plays the piano
DhwrBs96Kgk_120_124.avi,a man is riding a
qLwgb3F0aPU_298_305.avi,men are playing
qeKX-N1nKiM_0_5.avi,a woman is seasoning meat
1Sp2__RCT0c_11_15.avi,a man is riding a
Fe4tO5vW9_E_64_70.avi,a woman is cooking eggs
mmSQTI6gMNQ_120_128.avi,people are dancing
HV12kTtdTT4_5_14.avi,a boy is playing a phone
0lh_UWF9ZP4_27_31.avi,a woman cuts up food
Je3V7U5Ctj4_569_576.avi,a man is a a pizza
```

Fig. 5. ATT-SS-S2VT result

The loss versus the training epochs have been provided in Fig. 6

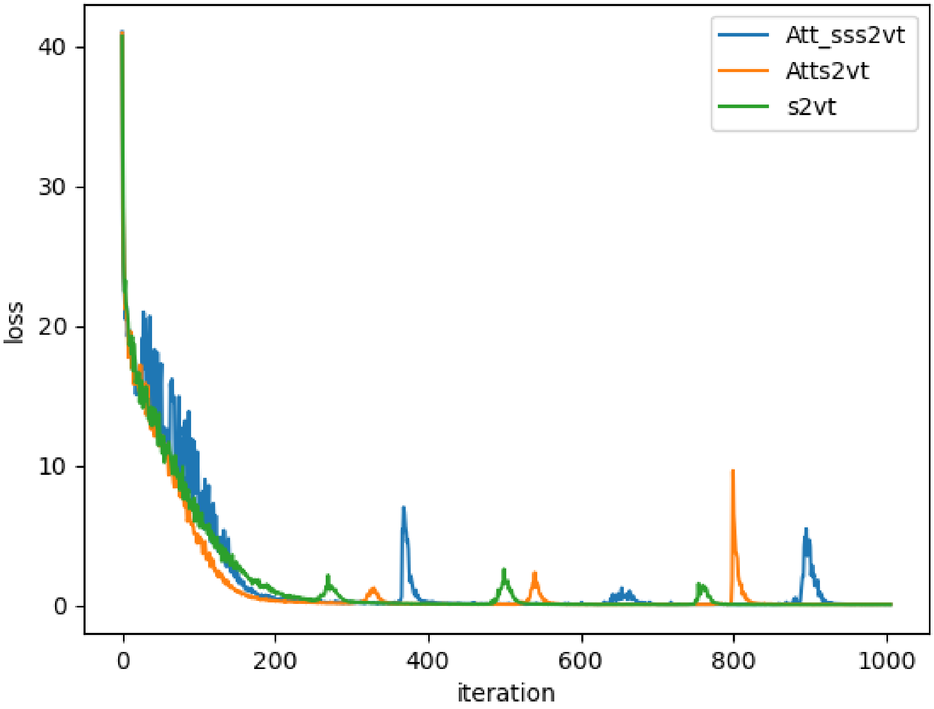


Fig. 6. Loss versus iteration