Image Captioning

2017cs10359
Prashit Raj

Part1- Non-Compatative part

CNN Encoder:

1. Three convolutional layers with max_pooling
2. One fully connected linear layer with relu activation
3. Takes in image tensor of size (3,224,224)
4. Give output as a feature vector of size 512

RNN Decoder:

1. Used nn.Embedding to embed the word vector of captions.
2. Used nn.LSTM to implement the LSTM model
3. One fully connected linear layer with relu activation

Used Cross-Entropy loss to create the loss function

Vocabulary:
Taken all the words with occurrence more than 10 into the vocabulary. Used the index of the word in the sorted vocabulary for vectorizing the captions.

Results:
Getting the same caption for each test image however, the loss function seemed to converge.

Part2- Compatative part

CNN Encoder:

1. Used the first two layers of xgg16 model
2. One fully connected linear layer with relu activation
3. Takes in image tensor of size (3,224,224)
4. Give output as a feature vector of size 512

RNN Decoder:

Same as in part-1

Used Cross-Entropy loss to create the loss function

Vocabulary:

Taken all the words with occurrence more than 10 into the vocabulary. Used the index of the word in the sorted vocabulary for vectorizing the captions.

Results:

Getting different for each test image however, the loss function seemed to converge.

Didn't have enough time to do proper analysis or make a proper report as the training took a lot of time.Apologies for that.

References:

Teacher Forcing