Homework 2 – Updated!

Joshua Lumpkin

Vocabulary:

Saved as a JSON, vocab size of 3528


```
▼ root:
  ▼ itos:
      0: "<PAD>"
      1: "<BOS>"
      2: "<EOS>"
      3: "<UNK>"
      4: "a"
      5: "horse"
      6: "on"
      7: "woman"
      8: "the"
      9: "her"
      10: "head"
      11: "head."
      12: "under"
      13: "and"
      14: "she"
      15: "between"
      16: "legs"
      17: "gets"
      18: "goes"
      19: "is"
      20: "horse."
      21: "pooped"
```


Example of preprocessing, tokenizing. I do append BOS and EOS to the caption as well

```
Original: A woman goes under a horse.
Tokenized: ['a', 'woman', 'goes', 'under', 'a', 'horse.']
Numericalized: [4, 7, 18, 12, 4, 20]
Max caption length: 42
Batch video features shape: torch.Size([32, 80, 4096])
Batch captions shape: torch.Size([32, 42])
```

Training setup:

- Epochs = 30
- Learning rate = .001
- Batch size = 32

```
S2VTModel(
  (encoder_lstm): LSTM(4096, 500, num_layers=2, batch_first=True, dropout=0.5)
  (decoder_lstm): LSTM(500, 500, num_layers=2, batch_first=True, dropout=0.5)
  (fc): Linear(in_features=500, out_features=3529, bias=True)
)
```

Results:

I am continually working on improving the output. I am displeased with the results. I could not get it to not be fixated on singular terms. It seemed to always converge on a similar sentence structure, like what is shown in the picture. I attempted 5-6 model variations but could not get past this. Some variations I tried: model with attention, singular feature to caption, extensive training with all captions per video, simple models, stepLR. I did not have time to implement beam search.

```
Predicted: ['a', 'man', 'is', 'a', 'a', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<
EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>',
'<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>']
Ground Truth: ['a', 'woman', 'is', 'carrying', 'a', 'bag', 'of', 'cans.', '<EOS>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PA
D>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>',
'<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>']
Example video features shape: (80, 4096)
Epoch [6/30], Loss: 4.591127748074739
Example video features shape: (80, 4096)
Epoch [7/30], Loss: 4.545964199563731
Example video features shape: (80, 4096)
Epoch [8/30], Loss: 4.494312213814777
Example video features shape: (80, 4096)
Epoch [9/30], Loss: 4.510790938916414
Example video features shape: (80, 4096)
Epoch [10/30], Loss: 4.483419620472452
Epoch: 11
Predicted: ['a', 'man', 'is', 'a', 'a', 'a', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS
>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<
EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>']
Ground Truth: ['someone', 'is', 'brushing', 'a', 'cats', 'hair.', '<EOS>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<P
AD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>',
'<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>']
Example video features shape: (80, 4096)
```

```
Example video features shape: (80, 4096)
Epoch: 1
Predicted: ['wave.', 'wave.', 'wave.', 'wave.', 'falling.', 'falling.', 'falling.', 'form', 'falling.', 'form', 'falling.', 'angrily', 'falling.', 'ou
tfit', 'falling.', 'falling.', 'form', 'falling.', 'falling.', 'falling.', 'form', 'falling.', 'falling.', 'angrily', 'form', 'falling.', 'falling.',
'falling.', 'falling.', 'falling.', 'falling.', 'falling.', 'outfit', 'falling.', 'form', 'outfit', 'outfit', 'form', 'falling.', 'falling.', 'fallin
g.']
Ground Truth: ['women', 'are', 'leaving', 'a', 'cab.', '<EOS>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD
>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<
PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>']
Epoch [1/30], Loss: 7.303551539130833
Example video features shape: (80, 4096)
Epoch [2/30], Loss: 5.13406401095183
Example video features shape: (80, 4096)
Epoch [3/30], Loss: 4.891730940860251
Example video features shape: (80, 4096)
Epoch [4/30], Loss: 4.754633126051529
Example video features shape: (80, 4096)
Epoch [5/30], Loss: 4.623692668002585
Epoch: 6
Predicted: ['a', 'man', 'is', 'is', 'a', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>',
'<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS
>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>', '<EOS>']
Ground Truth: ['a', 'woman', 'seated', 'on', 'a', 'chair', '<UNK>', 'a', 'knife', '<UNK>', '<EOS>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD
>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<
PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>', '<PAD>']
Example video features shape: (80, 4096)
Epoch [6/30], Loss: 4.595051039820132
```