**William Godel, NLP HW1**
All code can be found at: https://github.com/wpg205/NLP_HW
**Please note all tables contain validation accuracy percentage**

As per instructions, I have divided the training data into a 20,000 observation training data set and a 5,000 observation validation set (see HW1). Following this I set about training an initial model. The hyperparameters of this model are as follows:

- Ngram Length = 1
- Embedded Dimensions = 100
- Max sentence length = 200
- Max Vocab = 10000
- Optimizer = Adam
- Learning rate = .01
- Tokenization:
  - Spacy Tokenizer: Text
  - Lowercasing
  - Remove Punctuational = Yes
- Epochs = 2
- Linear Annealing = No

Validation accuracy for this model varied but was around 86%.

# 1 Lemma

First, I lemmaized the tokens using the same spaCy tokenizer (see HW1_lemma). I continued to lowercase and remove punctuation as it does not seem likely that leaving in either of those features would help the model. I experimented with a variety of learning rates and embedding dimensions, but found no improvement over the basic model.

Figure 1: Performance of the Basic Model

(a) (b)

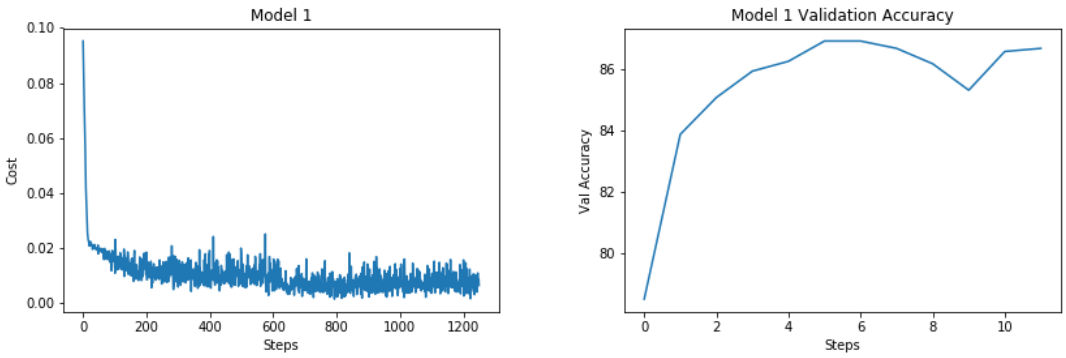
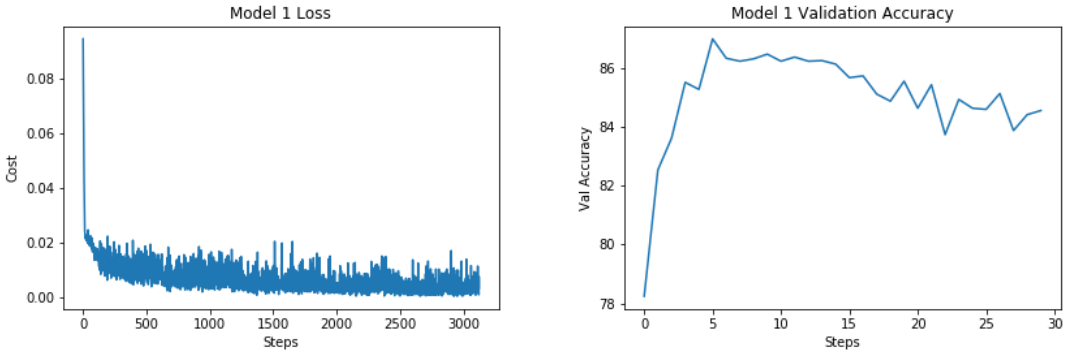


Figure 2: Performance of the Basic Model but with Lemmas Rather Than Text

(a) (b)




## 2 Embedding Dimensions and Learning Rates

I next experimented with different embedding dimensions and learning rates. Below is the performance table. From this analysis, it seems that a larger embedding dimension and lower learning rate are increasing performance, which makes intuitive sense in that a larger embedding size allows for a more complex model. See the chart in "Learning rates and Embedding Dimension" in the HW1 notebook for loss during training.

| | | Embedding Dimension | | | |
|---|---|---|---|---|---|
| | | 100 | 200 | 500 | 1000 |
| Learning Rate | .001 | 86.36 | 86.96 | 87.18 | **87.52** |
| | .005 | 85.98 | 85.68 | 84.94 | 84.64 |
| | .01 | 84.90 | 84.82 | 84.64 | 84.96 |
| | .05 | 84.76 | 84.90 | 83.70 | 83.42 |
| | .1 | 79.90 | 82.96 | 83.28 | 82.48 |

## 3 Optimization and Learning Annealing

I then test SGD versus the Adam Optimizer (with the learning rate and embed dimension of 1000). I also simultaneously tested different rates of learning annealing (specifically torch.optim.lr_scheduler.StepLR). This analysis is in the HW1 file. Adam outperformed SGD at every level of annealing, and showed little variation in performance associated with annealing. I will use a gamma of .1 in the final model due it its superior performance here. For charts, see the end of the HW1 file.

| | Annealing Gamma | | | |
|---|---|---|---|---|
| Optimizer | 1 | .5 | .1 | .01 |
| SGD | 65.34 | 64.20 | 65.18 | 64.72 |
| Adam | 87.00 | 87.14 | **87.24** | 86.94 |

# 4 NGrams Variation

For NGrams, I experimented with NGram of length 2,3,4 (the base model had only length 1). I loaded my Ngrams in the HW1 file, but ran my analysis in HW1_NGRAMS.

| | Embedding Dimension | | | |
|---|---|---|---|---|
| | 100 | 200 | 500 | 1000 |
| Bigrams | 85.30 | 86.06 | 86.22 | 86.12 |
| Trigrams | 85.94 | 86.22 | 86.56 | 86.02 |
| Quadragrams | 85.32 | 85.40 | 85.96 | 85.98 |

None of the additional Ngrams seemed to add significantly to performance.

# 5 Max Vocab Size and Max Sentence Length

Finally, I proceeded to vary the vocabulary size and sentence length. I did this only to the base model (see HW1_vocab_sentence) using only unigrams. I would have liked to have tried with bigrams and trigrams, but without a GPU the compute times were simply too slow. Larger vocabs and length were both superior.

| | Max Sentence Length | | |
|---|---|---|---|
| Vocabulary Size | 100 | 300 | 500 |
| 10,000 | 82.88 | 88.02 | 88.42 |
| 50,000 | 84.02 | 88.90 | **89.54** |

# 6 Final Model Performance

After varying hyper parameters along all the suggested dimension, and also trying sentence length (unlisted in original assignment), I tried a final model (see HW1_final) with the following characteristics: Adam Optimizer, learning annealing (step LR gamma of .1), unigrams, embedding dimension of 1000, max sentence length of 500, max vocab of 50000, with a base tokenization scheme (no lemmas) with lowercase and punctuation removed. This model achieves:

| Final Model Accuracy | | |
|---|---|---|
| Train | Validation | Test |
| 98.02 | 87.62 | **85.22** |

**Correct Validation Classifications:**

Obviously written for the stage. Lightweight but worthwhile. How can you go wrong with Ralph Richardson, Olivier and Merle Oberon.(true class = pos)

One of my favorite scenes is at the beginning when guests on a private yacht decide to take an impromptu swim - in their underwear! Rather risqué for 1931! (true class = pos)

Go, Igor, go, you are the proof that Slovenian films may, should and must be different. There's soul in it, and this is rare. Don't let anybody put you down! (true class = pos)

**Incorrect Validation classifications:**

wow...this has got to be the DUMBEST movie I've ever seen. We watched it in english class...and this movie made ABSOLUTELY no sense. I would never, EVER watch this movie again...and my sympathy to those who have ever PAID to see it. (true class = neg)

Summer Phoenix did a great performance where you really feel what she's not able to feel and you just cannot understand what she has on her mind. Besides, she portrays a jewish girl who behaves really confronting the status quo of that century. (true class = pos)

Laughs, adventure, a good time, a killer soundtrack, oscar-worthy acting, and special effects/ animitronics like none other, what else could you want in a movie? If you see this will be on the telly, WATCH IT, otherwise, run out now to RENT IT!!! (true class = pos)