NLP Assignment 4 [Name: HSU, Chia hong; STID: 20562937]

1 Introduction

Files used for training: “train.txt” under the data directory (the smaller training dataset).
Files for validation, testing: “valid.txt”, “test.txt”

2.1 Preprocessing

Methods used for cleaning data:

1. Clean “<unk>”
2. Clean mentions
3. Clean URL
4. Clean hashtags
5. Clean punctuations
6. Clean digits
7. Lemmatize words
8. Clean stopwords

3 Statistics

sentence number: 120115
vocab number: 23755
word number: 947836
average sentence length: 7.89

4.1 Hyper-parameters

Regularization technique: Layer Normalization + Dropout (=0.2) + grad_clipping (=0.1)
Other parameters: early_stop = 3, batch_size = 32, embed_dim = 150, max_grad_norm = 0.1

Learning Rate(layer_num=1, hidden_dim=128) (I decided not to test lr=0.0001 is because it would literally take DECADES for the loss to converge)

	Validation Perplexity
lr=0.1	4568.1191
lr=0.01	4515.3291
lr=0.001	11259.6479 (haven’t converge after 40 epochs)

Hidden Dimension(lr=0.01, layer_num=1)

	Validation Perplexity
hidden_dim=128	4515.3291
hidden_dim=256	4637.4830
hidden_dim=512	4389.7425

Layer Number(lr=0.01, hidden_dim=128)

	Validation Perplexity
layer_num=1	4511.3818
layer_num=2	4323.9126

Best Test Perplexity Score: 4347.6137 (lr=0.1, hidden_dim=128, layer_num=2, other parameters remain the same)