CSE 628 – Assignment 1

Name - Gaurav Ahuja
SBU Id – 111500318

1. Hyperparameters explored

1.1 Batch Size

Batch size affects the robustness of estimates of gradient.
In the denominator of Cross entropy loss, ideally, we should have used summation over all vocabulary, but in our implementation, we only used summation over the batch. Hence varying batch size should affect the loss function and its gradient. So I explored three settings for batch size, one value which is more than the default value and one less than the default setting.

batch_sizes = [64, 128, 256]

1.2 Skip window and number of samples in a window

Size of skip window and number of samples in a window affect the data which gets used for training the model. Hence I tried varying these window parameters. One setting smaller than the default value and one greater was used.

Window parameters (skip_window, num_skips) = [(2, 4), (4, 8), (8, 16)]

Total = 3x3 = 9 configurations

1.3 Number of negative samples for NCE loss

Negative samples are used to estimate the the following term in NCE loss function:

$$\mathbf{E_{Pn}[log(1 - sigmoid(\Delta s_{\theta}\ (w, context))]}$$

As number of negative samples increase, the estimate become more robust, at the cost of increase in training time.
The following settings for negative samples were tried:
num_negative_samples = [32, 64, 128]

Total Configurations:

For Cross entropy – 3x3 = 9 configurations

For NCE – 3x3x3 = 27 configurations

2. Results

Cross Entropy			
Batch size	Num skips	Skip window	Accuracy
64	2	4	33.6
64	4	8	33.9
64	8	16	33.9
128	2	4	33.7
128	4	8	34
128	8	16	33.9
256	2	4	33.7
256	4	8	33.9
256	8	16	33.8

NCE				
Batch size	Num skips	Skip window	Negative Samples	Accuracy
64	2	4	32	34.2
64	2	4	64	33.9
64	2	4	128	34.1
64	4	8	32	34
64	4	8	64	34.2
64	4	8	128	34.2
64	8	16	32	34.4
64	8	16	64	34.1
64	8	16	128	33.6
128	2	4	32	34
128	2	4	64	34
128	2	4	128	33.5
128	4	8	32	34.4
128	4	8	64	33.9
128	4	8	128	34.1
128	8	16	32	34.1
128	8	16	64	34.3
128	8	16	128	34.2
256	2	4	32	33.9
256	2	4	64	33.5
256	2	4	128	33.1
256	4	8	32	34.3
256	4	8	64	34
256	4	8	128	33.3
256	8	16	32	34.3
256	8	16	64	33.9
256	8	16	128	34.1

Observations:

- For cross entropy, batch size 128 gives better accuracy than 64 and 256.
- For cross entropy, num skips 4 gives better accuracy than 2 and 8.
- For NCE, when batch size is 64 and 128 and num skips in 2 and 4, increasing number of negative samples leads to increase in accuracy. For batch size 64 and 128 and num skips 8, increasing number of negative samples leads to decrease in accuracy.
- For NCE, when batch size is 256 and num skips is 2 and 4, increasing number of negative samples leads to decrease in accuracy.
- It is difficult to find a consistent trend when one parameter is increased and all the others are kept fixed. The accuracy does not always exhibits a consistently increasing or decreasing trend. For example when all other things are kept fixed, increasing number of negative samples leads to increase in accuracy for (64,2,4) setting and follows a decreasing trend for (64,8,16) setting. So we can not in general say that increasing number of negative samples leads to increase in accuracy or decrease in accuracy.

3. Similar Words

Top 20 similar words to {first, american, would}:

- For “american”, the similar words discovered by model are other nationalities, and other words used with american like “american player”, “american comedian” etc.
- For “would” also, the model discovers similar words like “could”, “might”, and other words used with would like “would been”, “would believed”, “would seems” etc.
- Same behavior is observed with “first”. (Similar words – last, most, original etc), and words used with first like “first city”, “first book”, “ first kingdom” etc.
- Similar words found using cross entropy make more sense as compared to nce.

Cross Entropy		
FIRST	**AMERICAN**	**WOULD**
last	german	will
most	british	could
name	french	might
following	english	must
original	italian	said
same	its	we
during	russian	been
end	european	did
best	borges	does
until	canadian	not
second	eu	they
main	international	do
city	irish	seems
before	war	you
rest	united	who
next	trade	should
largest	comedian	believed
book	advisory	if
after	player	may
kingdom	barzani	only

NCE		
FIRST	**AMERICAN**	**WOULD**
most	french	that
at	english	not
was	british	will
he	german	could
after	its	they
during	eight	been
which	three	might
until	s	so
before	UNK	this
use	four	which
from	seven	it
th	from	but
where	at	had
s	by	he
this	war	only
century	six	does
that	nine	who
is	international	where
on	five	what
end	one	said

4. Justification behind using Noise Contrastive Estimation loss

- In cross entropy loss, the probability of a word given context is normalized (i.e. the probabilities of word given context for all words in the vocabulary sum to 1).

- During training, it is expensive to compute this normalizing constant by doing a summation over the whole vocabulary.

- Noise Contrastive Estimation optimizes a different loss, but the aim is still to maximize the similarity between words which occur close to each other in text.

- NCE achieves this by formulating a binary classification problem, in which the words occur together are positive examples while those which don’t form the negative examples. The negative samples are assumed to come from a noise distribution. It is also assumed that noise samples are k times more frequent that positive samples.

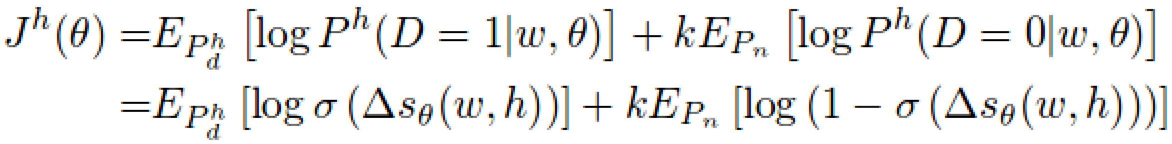
- NCE uses Logistic Regression as the classifier.

- Let h be the context and w be the outer word.

  P(this sample belongs to data distribution | w, context) =

  P(sample comes from data dist | w, h) / (P(sample comes from data dist | w, h) + k*P(w comes from noise dist))

This leads to the following objective function:



$$\begin{aligned} J^h(\theta) =& E_{P_d^h}\left[\log P^h(D=1|w,\theta)\right] + kE_{P_n}\left[\log P^h(D=0|w,\theta)\right] \\ =& E_{P_d^h}\left[\log \sigma\left(\Delta s_\theta(w,h)\right)\right] + kE_{P_n}\left[\log\left(1-\sigma\left(\Delta s_\theta(w,h)\right)\right)\right] \end{aligned}$$

Illustration 1: Equation taken from research paper

- To estimate the second term, k negative samples are chosen from noise distribution and the term is computed. Hence instead of going over the whole vocabulary, the NCE loss can be computed using small number of noise samples.

- Since NCE loss is fast to compute and tries to maximize the similarity score of words which occur together, it is a good alternate to use over cross entropy loss.