Assignment 1: Word Vectors

CSE 538 FALL 2019
Viraj Kamat
SBU ID : 112818603

Task 1: Hyper-parameters description

max_num_steps : This parameter controls the number of steps to train the model, ideally higher the number of steps better is the model trained. One possible drawback of large number of steps is overfitting by the model.

batch_size: Indicates the number of input vectors to consider when building the loss function to train the model per iteration. Rather than pass a single input vector (the numerator of the loss function) at a time, the batch_size parameter enables us to have multiple input vectors processed at the same time. This make the computation efficient.

skip_window: Determines the number of context vectors to the left and right of the observed word when choosing our desired context word. If the skip-window is too big then it will have a negative effect on training as it will consider words farther off from the center word.

num_skips : The number of context words to consider within the window_size (skip_window*2 +1) to train our model with respect to our input vector. It desirable the num_skips is not too large and the context vectors are closer to the center word while adding them to the context vector list.

num_sampled : Exclusively used by Noise contrastive estimation loss function that determines the amount of noise to introduce when training the model. Increasing the number of negative samples achieved better results.

Task 2: Analogy task

Several trial and error approaches were taken in fine tuning the hyper parameters to train the model.
The table below outlines the configurations from the best of training approaches taken:

Batch-size	Embedding-size	Negative Samples	Learning rate	skip-window	num_skips	max_num_steps	Accuracy	Model
128	128	N/A	0.2	2	4	200001	33.3	CE
128	128	64	0.2	2	4	200001	33.3	NCE
256	128	N/A	1	2	4	300001	33.5	CE
256	128	128	1	2	4	300001	34.2	NCE
256	128	N/A	0.5	1	2	500001	32.7	CE
256	128	256	0.5	1	2	500001	33.2	NCE
256	128	N/A	0.8	1	2	300001	32.8	CE
256	128	128	0.8	1	2	300001	33.4	NCE
256	128	N/A	1	2	4	500001	33.5	CE
256	128	128	1	2	4	500001	33.8	NCE
256	128	N/A	1	4	8	500001	33.4	CE
256	128	128	1	4	8	500001	34	NCE
256	**128**	N/A	1	8	16	**500001**	33.7	CE
256	**128**	128	1	8	16	**500001**	34.2	NCE

- Increasing the batch_size by some amount did improve the performance, increasing it by a large margin adversely affected the accuracy of the model. In most cases it was restricted to a size of 256.
- Increasing the number of epochs had a positive result in training the model, the larger the number of steps means the more iterations the optimizer goes through when training the model and achieves better results.
- In general, it is better to have the context words (targets) chosen close to the center word, if sample context are chosen farther off from the input word the training is adversely affected.
- For noise contrastive estimation increasing the number of negative samples helped to a certain extent, increasing it by a large amount adversely affected the accuracy of the model and it was restricted to 128 only.

Task 3: 20 similar words

Results for the Noise Contrastive Estimation Model

american	would	first
than	conformably	where
t	no	zero
insult	using	armand
zero	bloodbath	english
early	everyone	war
seven	try	early
at	join	seven
three	my	at
prisons	groups	de
armand	complex	bakunin
he	her	three
other	amount	he
UNK	only	american
english	right	UNK
war	together	expound
voyages	links	more
de	see	abuse
where	point	s
more	over	i
rique	number	b

With respect to finding the top 20 words here are some of the predictions that can be made (since the NCE model has about 34% accuracy, we can expect about 7 words to be relevant):

1. American: Here the words “english” emerges, implying the use of the language English. America is also known for overseas peace-keeping efforts hence the term “war” and the term “voyages” maybe a reference to the fact that America was found through voyages.
2. Would: “try” is a word commonly following would, the word “no” also follows would when used in a question, several other words can be seen that would commonly be used with the word would mostly in question-based sentences.
3. First: “war” maybe a word possibly implying the first world war. Most people also wake up at “7:00” hence the presence of the word seven – first time of the day.

Results for the Cross-Entropy Model

american	would	first
scholl	argued	penetrate
superchargers	believed	departs
canadian	nonfuture	best
keneally	wo	last
freedoms	legality	doubted
french	autocad	folksong
topicslist	appears	crunch
electro	may	shoulders
calligraphers	can	withdrawing
da	we	crs
russian	wanted	suppressing
blige	seems	nycfoto
snowy	said	riskier
vasily	does	krak
minefield	should	pools
barzani	did	latter
eu	must	souffle
italian	might	same
violinists	could	most
german	will	studebaker

With respect to finding the top 20 words here are some of the predictions that can be made (since the model has about 34% accuracy, we can expect about 7 words to be relevant):

American: “superchargers” are found prominently in American cars, the word “canadian” is also relevant since Canada is either in America or a neighbor of the USA. “freedoms” could be a reference to the freedom of speech in America.

Would: Would – “argued” and Would “believed” are very common in daily sentence usage. So are the words “said” as in “Would have said”, and in case of the word “might”- would and might are used interchangeably, as in “She would have” or “She might have”.

First: “letter” could refer to the use in sentence “First letter” to a person, also departs could be used as in first departs. Not many contexts were observed with the word first.

Task 4: Summary of NCE loss

The Noise contrastive estimation loss method or NCE works by including random negative samples in training our model. We consider the unigram probabilities of these samples to achieve better training by creating a stronger relation between center and relevant context words.

In cross entropy, for a given word its association with a context word is normalized i.e the probability of all possible context words sums up to 1. This computation is expensive. Noise contrastive estimation eliminates this by considering only the unigram probabilities of some noise samples.

The advantage here is that rather than using the entire set of contexts unlike cross entropy, simply choosing a set of noise samples in our training and evaluating against true word pairs is cost efficient.

Noise contrastive estimation is a Logistic regression classifier that aims to compute binary classification i.e find the probability of a context word given a center word. It introduces k negative samples from the distribution called noise which indicates that it is k times more frequent than our true center-context word pair.

The following are the steps in training:
1. Draw k negative samples called the noise from the corpus and a true target word (context word)

2. With these samples and true target word create an augmented training instance with respect to input word

3. We define our loss function as follows

$$J(\theta, Batch) = \sum_{w_o, w_c \in Batch} -\left[log\ P_r(D = 1, w_o|w_c) + \sum_{x \in V^k} log\ (1 - P_r(D = 1, w_x|w_c)) \right]$$

Where:

$log\ P_r(D = 1, w_o|w_c)$ is the likelihood that a true target w_o word appears alongside a context w_c word

$log\ (1 - P_r(D = 1, w_x|w_c))$ is the probability that a target w_o word appears along noisy samples given by drawn from the negative w_x samples, the summation of these probabilities is then taken into account for all the negative samples.

4. We train the model to optimize this loss function likelihood that true target word appears alongside the context word.