CSC 321: Assignment 1

Zeeshan Qureshi <g0zee@cdf.toronto.edu>

1. Train Model

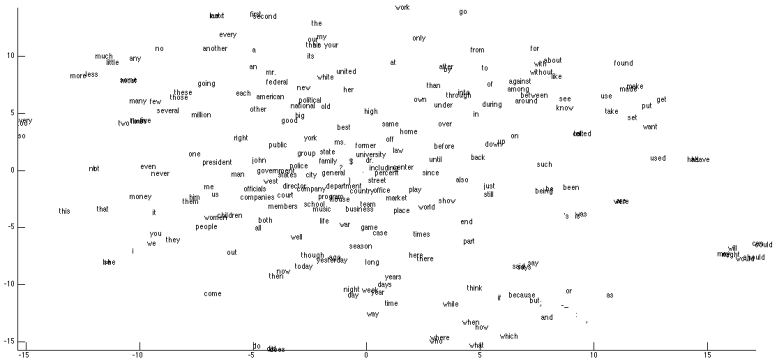
Table 1. Results

Model(d, hid)	Training	Validation	Test	Epochs
model(8, 32)	2.869	2.894	2.891	5
model(8, 256)	2.783	2.821	2.823	5
model(32, 64)	2.701	2.751	2.753	8
model(32, 256)	2.547	2.638	2.646	8

As we increase the number of dimensions in the distributed representation and the number of hidden units, the cross entropy error goes down. We run the risk of over-fitting our network to the data if we increase its capacity continuously but since our models do early stopping we are disregarding it and chosing **model(32, 256)** as the best since it has the lowest validaton and test cross entropy error.

2. Experiment with Model

2.1. t-SNE Plot



2.2. Closest Words

Candidate	Nearest	Second	Third
company	department	general	program
will	might	may	would

Candidate	Nearest	Second	Third
government	states	west	john
police	west	general	officials
women	children	companies	officials

I've only listen some of the more interesting ones here because of the space constraint. We see that for some words, like *will*, *police* and *company* the closest words make a lot of sense and are very likely to appear in the same sentence. For *women* and *government* the closest word makes a lot of sense but the rest don't seem to be at anything.

2.3. Sentence Prediction

Sentence	First	Second	Third
women also had	to	a	been
government of united	.	states	life

The model seems to have extracted some features of grammar and semantics of the english language but not much, for *women also had* the first 3 predictions are grammatically correct, but for *government of united* only the second one makes sense. If we had mor data to train the network on it would probably have better prediction of sentences since the 4-gram probabilities would be better for the not so frequently occurring terms; over here punctuations seem to have higher probabilities than words.

2.4. Word Distances

Pair	Distance	Pair	Distance
(university, dr)	.55	(university, $)	.53
(court, company)	1.08	(court, companies)	1.35
(money, million)	2.81	(money, music)	2.00

Wee see that *doctor* is closer to *university* but the symbol *$* is even closer, which doesn't make sense. The word *court* is close to both *company* and *compaies* which means that the net thinks that they are similar. *money* is closer to *music* than to *million* which is somewhat expected (pop culture) but then *million* is not even in the top 10 closest words for *money* which is a bit strange. It is probably because our training sentences don't contain any data related to financial news or politics.

3. Summary

Based on the analysis of the t-SNE plot and sentence prediction we can see that the distributed representation of the words helps the model extract features from the sentences which would not be possible in a 1-of-n representation. We also see that with the increade in the embedding size and hidden units the model generalizes pretty well. It remains to be seen how the model would extract meaning from a much larger training set and if we would see any strange neighbouring words.