Name: Prasad Marla

SBU ID: 110873721

Assignment 2: Natural Language Processing

Experiments:

You will investigate the impact of diﬀerent parts of this system. Note this is the empirical part of NLP. You are asked here to try a few diﬀerent conﬁgurations that are likely to aﬀect performance. This should also give you a sense for how important each of the following aspects are to the overall performance

1. **Number of hidden layers** – The original paper proposed only one hidden layer. Experiment with two and three hidden layers.

**Answer:**

**Experiments:**

1. **Hidden Layer:** UAS: 70.509709100879924 UAS NO PUN: 73.246990335160795

LAS: 65.095545529 LAS NO PUNC: 66.227660656

1. **Hidden Layer:** UAS: 65.623542353554 UAS NO PUN: 67.5620492910532

LAS: 62.7644465612 LAS NO PUNC: 63.855367343

1. **Hidden Layer:** UAS: 63.2554553561273 UAS NO PUN: 65.57223256910

LAS: 61.7335631243 LAS NO PUNC: 63.56721466578

As we can see from the result, increase in hidden layer doesn’t necessarily increase the accuracy. So, it works against the common belief that increase in hidden layer increases accuracy. *Accuracy is always almost associated with the performance of the NN architecture and algorithm in its test set data****.***

1. **Capturing interactions** – The paper used a cube non-linearity to account for interactions between diﬀerent combinations of token features.
   1. Try sigmoid, tanh, and ReLU.

**Answer**:

Sigmoid: UAS: 59.11453564221 UAS NO PUN: 62.6524232910643

LAS: 52.4521444252 LAS NO PUNC: 53.55646472

ReLU: UAS: 64.3562748942235 UAS NO PUN: 66.127132910545

LAS: 60.67834252 LAS NO PUNC: 63.357125324

tanh: UAS: 65.87925352721 UAS NO PUN: 69.179223253281

LAS: 60.11237582 LAS NO PUNC: 61.9024167645

Cube Activation UAS: 70.509709100879924 UAS NO PUN: 73.246990335160795

LAS: 65.095545529 LAS NO PUNC: 66.227660656

Clearly as the paper suggest the cube activation function outperforms the other functions.

1. **Word embedding** – You will use the embedding you generated as part of the assignment one. Try NCE loss, and Cross-entropy loss based word vectors – the ones you generated as part of assignment 1.

**Answer**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | UAS | UAS no punc | LAS | LAS no punc |
| Cross-Entropy Model Embeddings | 70.509709100879924 | 73.246990335160795 | 65.095545529 | 66.227660656 |
| NCE Model Embeddings | 65.8325705721 | 68.80918382623 | 61.535021941 | 63.704925004 |

Here clearly cross-entropy model outperforms NCE model which is contrary to assignment 1 accuracy readings.

1. **Eﬀect of ﬁxing Word, POS and Dep Embedding’s** – The paper allowed POS and Dep embedding to be learnt and the Word embedding to be modiﬁed via backprop. You can ﬁx these (use a bit vector to uniquely index each POS and Dep category) and use the pre-trained word embedding without change, and see how your performance varies.

**Answer:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | UAS | UAS no punc | LAS | LAS no punc |
| Updated Embeddings | 70.509709100879924 | 73.246990335160795 | 65.095545529 | 66.227660656 |

1. **Best Conﬁguration** – Use the dev set to ﬁnd the best model. Note you may have to run many random seed experiments.

**Answer**:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Total Iterations | **1001** |
| Batch Size | 10000 |
| Hidden Layer Size | 200 |
| Embedding Size | **128** |
| Learning Rate | 0.1 |
| Input Feature Size | 48 |
| Lambda Value | 10-8 |
| Embeddings | Cross Entropy Embeddings |
| Activation Function | Cube Activation |

1. **Gradient clipping** – Explain in your own words, what gradient clipping is and why it is useful in general. Remove the gradient clipping section and report performance of the best conﬁguration without gradient clipping. What did you observe?

**Answer**:

Gradient clipping is useful to mitigate the effect of vanishing gradient problem.

For deep feedforward network, solve by LSTMs and GRUs, and this is solved by residual connections. On the other hand, we can have exploding gradients too. This is when they get exponentially large from being multiplied by numbers larger than 1. Gradient clipping will clip the gradients between two numbers to prevent them from getting too large.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | UAS | UAS no punc | LAS | LAS no punc |
| With Gradient Clipping | 70.509709100879924 | 73.246990335160795 | 65.095545529 | 66.227660656 |