**Exploring similarity measurements for embeddings generated using Cross Entropy Loss function:**

Train using default parameters (except increased number of steps to 400001)

Components in different measures:

1. pairwise\_cossim ->
   1. cosine similarity in a given pair, (student, essay) -> cosine(student, essay)
   2. A naïve estimate of by how much quantity a pair is similar to each other
   3. The more it is, higher the probability of a given pair to be analogous to each other **(directly proportional)**
2. absolute\_length\_diff ->
   1. For a given average of difference vectors and a particular vector, calculates their absolute length difference
   2. Estimates how close our the lengths of two related pairs
   3. The less it is, more analogous are the pair to each other **(inversely proportional)**
3. diff\_length\_difference ->
   1. For a given average length for a set of difference vectors, calculate their difference with a length of query difference vector
   2. The less it is, more analogous are the pair to each other **(inversely proportional)**
4. avg\_cossim
   1. average of cosine similarity on a list of true related pairs
5. diff\_cossim
   1. cosine similarity between average difference vector and a particular difference vector
   2. The more it is, higher the probability of a given pair to be analogous ot each other **(directly proportional)**

Note: 2 and 3 may seem similar, but are different.

(2) tries to match the magnitude of the average of true difference vectors to the magnitude of query difference vector,

(3) tries to match the average of magnitudes of true difference vectors to the magnitude of query difference vector

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment #** | **Similarity measure experiments** | **Least** | **Most** | **Overall** |
| 1 | 1.0/dist\_diff | 35.9 | 34.9 | 35.4 |
| 2 | 1.0/ absolute\_euclid\_dist\_diff | **36.2** | **33.8** | **35.0** |
| 3 | pairwise\_cossim | **36.2** | **33.8** | **35.0** |
| 4 | diff\_cossim | 28.4 | 27.7 | 28.1 |
| 5 | 1.0/abs(avg\_cossim - pairwise\_cossim) | 35.9 | 34.7 | 35.3 |
| 6 | 1.0/(absolute\_euclid\_dist\_diff \* abs(avg\_cossim - pairwise\_cossim) ) | 35.2 | 34.2 | 34.7 |
| 7 | pairwise\_cossim/(absolute\_euclid\_dist\_diff \* abs(avg\_cossim - pairwise\_cossim) ) | **37.3** | **36.0** | **36.7** |
| 8 | pairwise\_cossim / (absolute\_euclid\_dist\_diff) | **36.2** | **33.8** | **35.0** |

**Conclusion:**

Initially

**Hyper-Parameter Tuning (Loss function: Cross-Entropy)**

Default parameters:

batch\_size=128 skip\_window=4 num\_skips=8 num\_sampled=64 max\_num\_steps=400001 learning\_rate=1.0

**Accuracy after training cross-entropy from scratch on default parameters:**

**Least illustrative accuracy: 30.3 %, Most illustrative accuracy: 32.6 %, Overall Accuracy: 31.5%**

Tuning models after training from scratch

Similarity measure use:

pairwise\_cossim/(absolute\_euclid\_dist\_diff \* abs(avg\_cossim - pairwise\_cossim))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment #** | **Hyper parameters** | **Least** | **Most** | **Overall** |
| 1 | batch\_size=64  sample nearest words are not close to each other,  e.g: digit words like zero doesn’t get one, two, three etc | 30.9 | 31.9 | 31.4 |
| 2 | batch\_size=256  sample nearest words are not close to each other,  e.g: digit words like zero doesn’t get one, two, three etc | **31.6** | **31.6** | **31.6** |
| 3 | skip\_window=2 num\_skips=4  sample nearest words are close to each other than the model set with default parameters | 28.2 | 33.7 | 31.0 |
| 4 | skip\_window=8 num\_skips=16 | **31.5** | **31.6** | **31.6** |
| 5 | learning\_rate=2.0  (Loss keeps overshooting between 4.8x and 4.9x)  Nearest words logged are not closer:  For e.g: six -> ‘mcdonagh’, ‘randomized’ as neighbours | **34.2** | **32.3** | **33.3** |
| 6 | embedding\_size=256, num\_steps=900001  Nearest words logged are somewhat closer:  For e.g: zero, six, eight, nine | **30.3** | **32.2** | **31.2** |
|  | Best model:  embedding\_size=256, num\_steps=900001 | **30.3** | **32.2** | **31.2** |

|  |  |
| --- | --- |
| **Query word** | **Cross Entropy Top 20 nearest words (excluding self)** |
| first | **most**, same, an, spit, a, **second**, metallic, microphone, neuropathologist, attendant, fingernail, **last**, veered, federalist, vacate, paixhans, zora, next, hopewell, regulatory |
| american | **french**, **english**, maharaj, **hollander**, focal, montanari, **british**, concatenation, bocce, cheaper, ecusa, berlinguer, forking, atomists, **australian**, monogamous, basket, infractions, musaf, mathrm |
| would | **will**, **may**, **can**, **could**, **had**, **was**, **did**, were, is, does, , **should** must, might, has, are, agitator, pele, garr, communicator, unbuilt |

Default training parameters (using pretrained), exploring distance measures with cross entropy model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Hyper parameters** | **Least** | **Most** | **Overall** |
| NCE | 1.0/dist\_diff | **32.2** | **35.0** | **33.6** |
| NCE | 1.0/ absolute\_euclid\_dist\_diff | 29.9 | 31.3 | 30.6 |
| NCE | pairwise\_cossim | 29.9 | 31.3 | 30.6 |
| NCE | diff\_cossim | 28.2 | 28.4 | 28.3 |
| NCE | 1.0/abs(avg\_cossim - pairwise\_cossim) | **32.4** | **35.7** | **34.0** |
| NCE | 1.0/dist\_diff \* abs(avg\_cossim - pairwise\_cossim) ) | **32.3** | **35.4** | **33.9** |

Hyper-Parameter Tuning

Default parameters:

batch\_size=128 skip\_window=4 num\_skips=8 num\_sampled=64 max\_num\_steps=200001 learning\_rate=1.0

**Accuracy after training cross-entropy from scratch on default parameters:**

**Least illustrative accuracy: 32.4 %, Most illustrative accuracy: 35.7 %, Overall Accuracy: 34.0%**

Tuning models using pretrained (Using measure: 1.0/(abs(avg\_cossim - pairwise\_cossim) \* dist\_diff)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Hyper parameters** | **Least** | **Most** | **Overall** |
| NCE | batch\_size=64 | 31.5 | 35.2 | 33.4 |
| NCE | batch\_size = 32 (accuracy is pathetic because loss did not converge around 1.x) | 29.8 | 31.4 | 30.6 |
| NCE | batch\_size=256 | 32.1 | 34.7 | 33.4 |
| NCE | skip\_window=2 num\_skips=4 | 31.8 | 35.3 | 33.6 |
| NCE | skip\_window=8 num\_skips=16 | **32.2** | **34.4** | **33.3** |
| NCE | num\_sampled=32 | 33.3 | 35.1 | 34.2 |
| NCE | num\_sampled=16 (loss convergence at 0.22x) | 32.4 | 33.8 | 33.1 |
| NCE | num\_sampled=8 (loss convergence at 0.21x)  (mikolov recommends smaller num\_sampled for large data-sets) | **32.5** | **35.8** | **34.1** |
| NCE | num\_steps=400001 (loss convergence at 1.41)  loss did not converge any better by increasing steps,  so no point in increasing number of steps any more | 32.2 | 33.3 | 34.4 |
| NCE | learning\_rate=0.5  increase in learning rate might overshoot the loss, hence exploring lower learning\_rate than the standard 1.0.  loss did not converge any better (1.45), hence exploring more lower learning rate would not be intuitive | 32.1 | 34.7 | 33.4 |
| NCE best model | batch\_size = 128 skip\_window = 8  num\_skips = 16 num\_sampled = 8  max\_num\_steps = 200001 learning\_rate = 1.0 | **32.6** | **36.5** | **34.6** |

|  |  |
| --- | --- |
| **Query word** | **NCE Top 20 nearest words (excluding self)** |
| first | **name**, **last**, following, during, same, original, of, **second**, end, **most**, book, after, until, united, th, city, was, beginning, **best**, |
| american | german, british, italian, d, russian, understood, its, european, irish, actor, might, december, heterodox, war, player, english, canadian, international, autres, writer |
| would | will, could, said, india, must, we, do, does, did, you, they, not, who, families, if, may, so, i, believed, even |