**Compering Word2vec and Glove model**

Chen Sheiner and Shay Shemesh

Natural Language Processing, by Yair Horesh PhD

Summer semester 2022

**Abstract**

TBD

**Introduction**

Word embedding is one of the most popular ways of text analysis. The method stated that each word would get a N vector representation in which context is taken into account. That means that similar words get vectors that are close in value (usually this is measured in cosine distance). By representing words by numeric vectors, we are able to use regular known architectures for various problems. There are many word embedding technics but, in this paper, we will focus on only the main two methods, Word2Vec and GloVe.

Word2Vec

Word2Vec was first developed by Tomas Mikolov, et al. at Google in 2013. In his paper he described a 2-layer neural network in which there is one hidden layer. The input of this architecture is a positional vector for each word in a sentence, meaning zeroes vector of length as the length of the sentence and a “1” where the target ward is positioned. There are 2 sub-architectures used to encode each word to a vector that represents its context and meaning.

CBoW

Continuous Bag of Words (CBoW) uses the context in the sentence in order to predict the target word. This is done by taking each word in a sentence (or couple of words based on the window size parameter given to the model) and letting the network try to learn what is the target word that the input context is referring to. The weights of the hidden layer are bean trained and adjusted each run and the output layer are the probability for each word in the corpus to be the target word. When the network finished training the vector representing each word is the weights of the model when that word is the target.

Skip-Gram

Skip-Gram model is similar to CBoW but upside down. In this model we try by a known target word predict the context words that correspond to the target word. To do this the model’s input have two variances: negative or positive input. Positive inputs are giving the model the target word and a context word that is related to the target word. While negative input is giving the model the target word and a random word from the corpus, while also providing the model if this pair is a viable context pair. These methods allow the model to adjust the weights for embedding each of the input words and in the end the embedding of each word is harvested from the weights of the model.

GloVe

Global Vectors (GloVe, developed by Pennington, et al. at Stanford,2014) tries to correct a fundamental flaw in Word2Vec which is that the formal model only takes in consideration the local context words of the target word. This may hurt the performance of the model in which there are further words that have strong context with the target word. Glove corrects this flaw by taking into account the global statistics and local statistics in a corpus. This is done firstly by creating a co-occurrence of the vocabulary of words a in the corpus. From that matrix we can get the probability to see two words together. TBD

In this paper we will compare these model (specifically Skip-Gram and GloVe ) by training them on the same data with equal number of epochs and comparing them on a test dataset.

**Methods**

Training Data

Using the Harvard USPTO Patent Dataset (HUPD) we gather 16153 patents. In each patent we gather the abstract, background, summary and description as a full text to train on. By gathering all 16153 patents together we get a big training text to train our models on.

With this dataset we cleaned each patent text with tokenization and drop every non-word in the text. These data are used to train our two models.

Model Architecture

For each of the models we use similar parameters shown below.

Word2Vec

For the Word2Vec training we use these parameters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | Vector\_Size | Window\_Size | Workers | sg | Learning Rate |
| Value | 300 | 2 | 4 | 1 | 0.05 |

GloVe

For the GloVe training we use these parameters:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Vector\_Size | Window\_Size | Workers | Learning Rate |
| Value | 300 | 2 | 4 | 0.05 |

When trained each model was saved in epoch number [10,30,50,100] and the training prosses was timed.

Each of the models than was tested on the Synonyms (WordNet) dataset. In this dataset there are target words and a list of synonyms words for each of them. our goal is to compare in which model there were more synonyms words of the target word in the top 5 words decided by the model based on similar number of epochs trained and training duration.

**Future work**

TBD

**Bibliography**

Aucouturier, J.-J., Defreville, B., & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, *122*(2), 881–891.

Carter, D. (2003). The surgeon as a risk factor: Determinants of outcome include technical skill, volume of work, and case mix. In *BMJ* (Vol. 326, Issue 7394, pp. 832–833). British Medical Journal Publishing Group.

Chowdhury, L., Kamal, M., Hasan, N., & Mohammed, N. (2021). Curricular SincNet: Towards Robust Deep Speaker Recognition by Emphasizing Hard Samples in Latent Space. *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–4.

Cuschieri, A. (2003). Medical errors, incidents, accidents and violations. *Minimally Invasive Therapy & Allied Technologies*, *12*(3–4), 111–120.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 30–42.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hummel, D. W., Lanz, O. I., & Werre, S. R. (2010). Complications of cementless total hip replacement. *Veterinary and Comparative Orthopaedics and Traumatology*, *23*(06), 424–432.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4814–4818.

Ravanelli, M. (2017). Deep learning for distant speech recognition. *ArXiv Preprint ArXiv:1712.06086*.

Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, 1021–1028.

Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671–1675.

Sarker, S. K., & Vincent, C. (2005). Errors in surgery. *International Journal of Surgery*, *3*(1), 75–81.

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine*, *23*(2), 133–141.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, *17*(10), 1733–1746.

Wang, S., Zhao, H., Wang, W., Di, H., & Shu, X. (2017). Improving deep crowd density estimation via pre-classification of density. *International Conference on Neural Information Processing*, 260–269.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, *51*(6B), 2044–2056.