Efficient Estimation of Word Representations in Vector Space

A. Sagirova V. Busovikov M. Pautov V. Goncharenko S. Konev
Mentor: Leyla Mirvakhabova

NLA course project
Skoltech, 2018

Outline

Introduction

In NLP words are represented as indices in a vocabulary.

- Advantages: simplicity and robustness.
- Disadvantage: in automatic speech recognition the performance is dominated by the size of the data.

Project goals

- Test techniques for measuring the quality of the resulting vector representations.
- We expect that not only similar words tend to be close to each other, but that words can have multiple degrees of similarity.
- The quality of words representation is measured in task of answering the query.

Answering the query

- Given a pair of words (a, b) and word c.
- Task is to find the word d, such that semantic similarity in pair (c, d) is the same as in pair (a, b).
- Examples of queries:

a	b	c	d
France	Paris	Germany	Berlin
Big	Bigger	Small	Smaller
Man	Brother	Woman	Sister

Neural network language model

- The sparse history h is projected into some continuous low-dimensional space, where similar histories get clustered.
- Thanks to parameter sharing among similar histories, the model is more robust: less parameters have to be estimated from the training data.

Dataset and quality estimation metric

- We have used a corpus of English Wikipedia articles for training the word vectors.
- This corpus contains about 13M tokens. We have restricted the vocabulary size to 100K most frequent words.
- As a measure of word closeness we have used cosine distance between word vectors.

Model architectures

- Feedforward Neural Net Language Model
- Recurrent Neural Net Language Model
- Continuous Bag-of-Words Model
- Continuous Skip-gram Model

Feedforward Neural Net Language Model

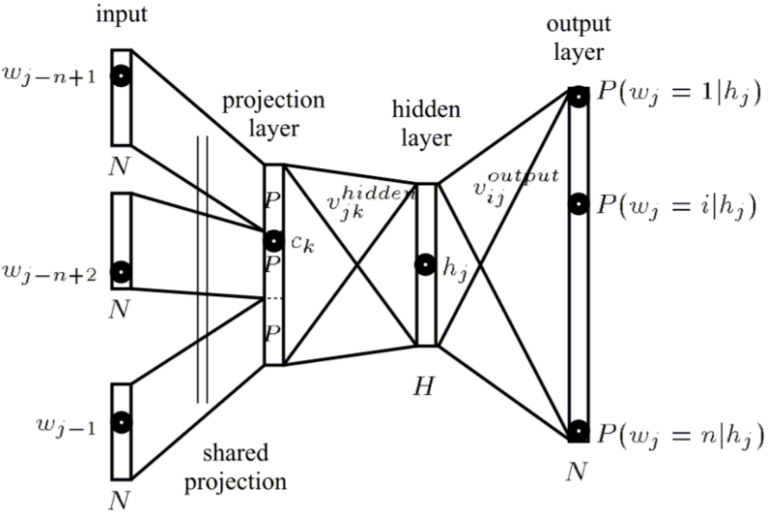
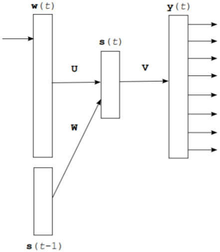


Figure 1: Feedforward neural network based LM used by Y. Bengio and H. Schwenk.[2]

Recurrent Neural Net Language Model



- Input layer w and output layer y have the same dimensionality as the vocabulary (10K - 200K)[1].
- Hidden layer s is orders of magnitude smaller (50 - 1000 neurons).
- U is the matrix of weights between input and hidden layer, V is the matrix of weights between hidden and output layer.
- Without the recurrent weights W, this model would be a bigram NNLM.
- Complexity per training example $Q = H \times H + H \times V$.

Continuous Bag-of-Words Model

- All words get projected into the same position.
- Task is to build a log-linear classifier with four future and four history words at the input.
- The training criterion is to correctly classify the current (middle) word.
- Training complexity is $Q = N \times D + D \times log_2(V)$.

Continuous Skip-gram Model

- Each current word is used as an input to a log-linear classifier with continuous projection layer.
- Words are predicted within a certain range before and after the current word.

Results

- Quality of models was estimated on set of 3k queries consisting of 4 words.
- Query is answered correctly only if model's guess is exactly the last word in query.
- Model score is a percentage of correctly answered queries.

Model	Score
SVD as word2vec	0%
Bag-of-Words	26%

Conclusion

- SVD is not suitable for our task because of the problem of extrapolating semantic similarity from one bigram to the other.
- Models as RNNLM or Feedforward overcome this as the embedding space for these models is linear.
- We used the embedding with dimensionality 3 times smaller than it was proposed in article in order to save time for training.
- Our Bag-of-Words outperformed Freeforward NNLM, RNNLM

References

[1] Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[2] Y. Bengio, R. Ducharme, P. Vincent A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.

[3] Y. Bengio, Y. LeCun. Scaling learning algorithms towards AI. In: Large-Scale Kernel Machines, MIT Press, 2007.

[4] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, 2007.