ON SEMI-SUPERVISED LEARNING OF WORD EMBEDDINGS

Wenchao Du
David R. Cheriton School of Compute Science
University of Waterloo
Waterloo, ON, CANADA
w8du@uwaterloo.ca

Stephen Vavasis
Department of Combinatorics and Optimization
University of Waterloo
Waterloo, ON, CANADA
vavasis@uwaterloo.ca

ABSTRACT

We investigated learning word embeddings with supervised information of word analogies based on GloVe, one of the state-of-the-art models. We showed that training with a small subset of analogies improved the performance significantly.

1 INTRODUCTION

Word representation learning is the task of mapping words to Euclidean space so that the word vectors preserves linguistic regularities reflected from the training corpus. Such representations are useful for information retrieval, document classification, and text generation.

Perhaps the most interesting property of word vectors is that semantic information of words can be measured with metrics of Euclidean space, typically cosine distance. It is known that semantic similarity can be measured by cosine distance of word vectors. Recently, Mikolov et al.(2013) discovered that word analogies are also measurable with cosine distances.

Matrix factorization methods have been proven effective in word representation learning. In fact, the current state-of-the-art model, GloVe(Pennington, Socher, and Manning) falls into this category. Next we will discuss some aspects of matrix factorization.

2 MATRIX FACTORIZATION: A PROBABILISTIC VIEW

This section provides some background on matrix factorization approach to word embedding learning. Suppose we are given a corpus of n words and we are to learn the vector space representations of size k for these words, matrix factorization approach generally solves for the optimization problem

$$\underset{U,V}{\text{minimize}} \quad \|W \circ (A - UV)\|_F \tag{1}$$

where A is $n \times n$ matrix of statistics from the corpus (for example, Point-wise Mutual Information (PMI) or harmonic sum of distance of words), V is $k \times n$ matrix whose columns are word vectors, U is $n \times k$ whose rows are the context vectors associated with words, W is $n \times n$ of weights, and $\circ$ is element-wise multiplication or Hadamard product. For W to be all-one matrix, this problem has exact solution provided by singular value decomposition (SVD) for any k. For general W, the problem is NP-hard, and gradient methods are commonly adopted.

There is in fact a probabilistic interpretation of (1). Consider regression model

$$A_{ij} = \sum_{l=1}^{k} U_{il} V_{lj} + \epsilon_{ij} \tag{2}$$

where $\epsilon_{ij} \sim N(0, \frac{1}{W_{ij}^2})$ are Gaussian noise of non-constant variance. Then with some simple calculations, the maximum likelihood estimate of noises

$$\underset{U,V}{\text{maximize}} \quad \sum_{i,j} \log P(\epsilon_{ij} | A, U, V) \tag{3}$$

Table 1: Experimental results

Model	Dimension	Semantic	Syntactic	Total
GloVe	50	0.284	0.204	0.237
GloVe + SSL	50	0.511	0.453	0.465
GloVe	100	0.434	0.261	0.333
GloVe + SSL	100	0.690	0.563	0.589

is equivalent to (1). Why is non-uniform noise important? Because in the problem of word representation learning, those entries of A associated with rare words have less certainty than the others. They provide relatively inaccurate information on distributions of word co-occurrences, and this is countered by assigning smaller weights to these entries in (1), which is equivalent to assuming larger variances in (2).

3 METHOD

Word analogy task is the problem of finding the best word to fill in the blank of "X to Y is analogous to Z to ___". This is done by solving

$$\underset{v}{\text{minimize}} \quad \|v_a - v_b - v_c + v\| \tag{4}$$

The idea of semi-supervised learning of word embedding is to inject the distance of analogous words into any objective. Specifically, given a learning objective $f(A, W, U, V)$, we pick a subset S of analogies and optimize

$$\underset{U,V}{\text{minimize}} \quad f(A, W, U, V) + \lambda \sum_{(a,b,c,d)\in S} \|V_a - V_b - V_c + V_d\| \tag{5}$$

where V_i is the i^{th} column of V, i.e. the vector representation of i^{th} word.

How do we pick analogies S so we are confident that the algorithm is not over-fitting? From standard optimization theory we know there exist $\mu_{(a,b,c,d)}$ for any $(a, b, c, d) \in S$ such that (4) is equivalent to

$$\begin{aligned} &\underset{U,V}{\text{minimize}} \quad f(A, W, U, V) \\ &\text{subject to} \quad \|V_a - V_b - V_c + V_d\| \leq \mu_{(a,b,c,d)} \; \forall (a, b, c, d) \in S \end{aligned} \tag{6}$$

Having this equivalence, consider the case where we have (a, b, c, d) and (a, b, e, f) in S and (c, d, e, f) for testing. Then by triangular inequality we have

$$\|V_c - V_d - V_e + V_f\| \leq \|V_a - V_b - V_c + V_d\| + \|V_a - V_b - V_e + V_f\| = \mu_{(a,b,c,d)} + \mu_{(a,b,e,f)} \tag{7}$$

If $\mu_{(a,b,c,d)}$ and $\mu_{(a,b,e,f)}$ are small enough, then V_d is the solution to (4). This means analogy (c, d, e, f) may be implied from analogies (a, b, c, d) and (a, b, e, f). To eliminate the possibility of such situations, we enforce the constraint that no two analogies in S have more than one common word.

4 EXPERIMENTS

We experimented with f to be the weighted least squares objective of GloVe. We randomly pick analogies from each of the analogy sets provided by GloVe. This yields a training set of 256 analogies and a testing set of 19,288 analogies. We trained on a corpus of 17 million tokens and used window size of 15. See Table 1 for results. We could not reproduce the results at the same level as in GloVe. This is maybe because the corpus we used for training is much smaller.

5 CONCLUSION AND FUTURE WORK

We explore the semi-supervised method of word representation learning with word analogies. The method is effective. We plan to experiment the same idea on other model such as CBOW and skip-gram. We will also evaluate the word vectors learned on other semantic tasks draw further conclusions.

REFERENCES

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–43, 2014.