Information-Theory Interpretation of the Skip-Gram Negative-Sampling Objective Function

Oren Melamud
IBM Research
Yorktown Heights, NY, USA
oren.melamud@ibm.com

Jacob Goldberger
Faculty of Engineering
Bar-Ilan University, Israel
jacob.goldberger@biu.ac.il

Abstract

In this paper, we define a measure of dependency between two random variables, based on the Jensen-Shannon (JS) divergence between their joint distribution and the product of their marginal distributions. Then, we show that *word2vec*'s skip-gram with negative sampling embedding algorithm finds the optimal low-dimensional approximation of this JS dependency measure between the words and their contexts. The gap between the optimal score and the low-dimensional approximation is demonstrated on a standard text corpus.

1 Introduction

Continuous word representations, derived from unlabeled text, have proven useful in many NLP tasks. Such word representations (or embeddings) associate a low-dimensional, real-valued vector with each word, typically induced via neural language models or matrix factorization.

Substantial benefit arises when embeddings can be efficiently trained on large volumes of data. Hence the recent considerable interest in the continuous bag-of-words (CBOW) and skip-gram with negative sampling (SGNS) models, described in (Mikolov et al., 2013), as implemented in the open-source toolkit *word2vec*. These models are based on a relatively simple log-linear method and avoid hidden layers typical to neural networks. Consequently, they can be trained to produce high-quality word embeddings on large corpora like the entirety of English Wikipedia in several hours, compared to days or even weeks in the case of other continuous models. Recent studies obtained state-of-the-art results by using skip-gram embeddings on a variety of natural language processing tasks, such as named entity extraction (Passos et al., 2014) and dependency parsing (Bansal et al., 2014). In recent years, there were several attempts to mathematically interpret word embedding models (Arora et al., 2016; Pennington et al., 2014; Stratos et al., 2015). Our study pursues this established line of work, attempting to explain the objective function of the SGNS word embedding algorithm.

In the SGNS model, the energy function takes the form of a dot product between the vectors of an observed word and an observed context. The objective function is a binary logistic regression classifier that treats a word and its observed context as a positive example, and a word and a randomly sampled context as a negative example. Levy and Goldberg (2014) offered a motivation for this function by showing that it obtains its global maximum value at the word-context pointwise mutual information (PMI) matrix. In this study, we take their analysis one step further and provide an information-theoretical interpretation of the SGNS objective function. In Section 2, we define a new measure of mutual information between random variables based the Jensen-Shennon divergence (Lin, 1991) instead of the KL divergence. In Section 3, we show that the value of the SGNS objective computed at the PMI matrix is this information measure. We then derive an explicit expression for the information loss caused by the low-dimensional embedding learned by the SGNS algorithm. Finally, in Section 4, we illustrate this by computing the information loss caused by actual SGNS embeddings learned on a standard text corpus.

2 A Dependency Measure based on Jensen-Shannon

In this section, we define a dependency measure between two random variables, which is based on the Jensen-Shannon divergence. Later, in Section 3, we show how it relates to the SGNS objective function.

There are several standard methods of measuring the distance between two discrete probability distributions, defined on a given finite set $\mathcal{A}$. The Kullback-Leibler (KL) divergence of a distribution p from a distribution q is defined as follows: $KL(p||q) = \sum_{i \in \mathcal{A}} p_i \log \frac{p_i}{q_i}$. The mutual information between two jointly distributed random variables X and Y is defined as the KL divergence of the joint distribution $p(x,y)$ from the product $p(x)p(y)$ of the marginal distributions of X and Y, i.e. $I(X;Y) = \text{KL}(p(x,y)||p(x)p(y))$.

The Jensen-Shannon (JS) divergence (Lin, 1991) between distributions p and q is:

$$\begin{aligned} \text{JS}_\alpha(p,q) &= \alpha \text{KL}(p||r) + (1-\alpha)\text{KL}(q||r) \quad (1) \\ &= H(r) - \alpha H(p) - (1-\alpha)H(q) \end{aligned}$$

such that $0 < \alpha < 1$, $r = \alpha p + (1-\alpha)q$ and H is the entropy function (i.e. $H(p) = -\sum_i p_i \log p_i$). Unlike KL divergence, JS divergence is bounded from above and $0 \leq \text{JS}_\alpha(p,q) \leq 1$.

We next propose a new measure for mutual-information using the JS-divergence between $p(x,y)$ and $p(x)p(y)$ instead of the KL-divergence. We define the Jensen-Shannon Mutual information (JSMI) as follows:

$$\text{JSMI}_\alpha(X,Y) = \text{JS}_\alpha(p(x,y), p(x)p(y)). \quad (2)$$

It can be easily verified that X and Y are independent if and only if $\text{JSMI}_\alpha(X,Y) = 0$.

We next derive an alternative definition of the JSMI dependency measure. Assume we choose between the two distributions, $p(x,y)$ and the product of marginal distributions $p(x)p(y)$, according to a binary random variable Z, such that $p(Z=1) = \alpha$. We first sample a binary value for Z and next, we sample a r.v. W as follows:

$$p(W=(x,y)|Z) = \begin{cases} p(x)p(y) & \text{if} \quad Z=0 \\ p(x,y) & \text{if} \quad Z=1. \end{cases} \quad (3)$$

The divergence measure $\text{JSMI}_\alpha(X,Y)$ can be alternatively defined in terms of mutual information between W and Z. The mutual-information between W and Z is:

$$\begin{aligned} I(W;Z) &= H(W) - \sum_{i=0,1} p(Z=i)H(W|Z=i) \\ &= H(\alpha p(x,y) + (1-\alpha)p(x)p(y)) \\ &\quad -\alpha H(p(x,y)) - (1-\alpha)H(p(x)p(y)). \end{aligned}$$

Eq. (1) thus implies that:

$$\text{JSMI}_\alpha(X,Y) = I(W;Z). \quad (4)$$

Applying Bayes rule we obtain:

$$\begin{aligned} &p(Z=1|W=(x,y)) \quad (5) \\ &= \frac{\alpha p(x,y)}{\alpha p(x,y) + (1-\alpha)p(x)p(y)} \\ &= \frac{1}{1+\exp(-\log(\frac{\alpha p(x,y)}{(1-\alpha)p(x)p(y)}))} = \sigma(\text{pmi}_{x,y}) \end{aligned}$$

such that $\sigma(u) = \frac{1}{1+\exp(-u)}$ is the sigmoid function and

$$\text{pmi}_{x,y} = \log \frac{p(x,y)}{p(x)p(y)} + \log \frac{\alpha}{1-\alpha} \quad (6)$$

is a shifted version of the PMI function. Equations (4) and (5) imply that:

$$\begin{aligned} &\text{JSMI}_\alpha(X,Y) = H(Z) - H(Z|W) \quad (7) \\ &= h(\alpha) + \alpha \sum_{x,y} p(x,y) \log \sigma(\text{pmi}_{x,y}) \\ &+(1-\alpha) \sum_{x,y} p(x)p(y) \log \sigma(-\text{pmi}_{x,y}) \end{aligned}$$

such that $h(\alpha) = -\alpha \log(\alpha) - (1-\alpha)\log(1-\alpha)$ is the binary entropy function.

3 The Skip-Gram Embedding Algorithm

The SGNS embedding algorithm (Mikolov et al., 2013) represents each word x and each context y as d-dimensional vectors $\vec{x}$ and $\vec{y}$, with the purpose that words that are "similar" to each other will have similar vector representations. We can represent a given d-dimensional embedding by a matrix m, such that $m(x,y) = \vec{x} \cdot \vec{y}$. The rank of the embedding matrix m is (at most) d.

Let $p(x,y)$ be the normalized number of co-occurrences of word x and context-word y in a given corpus and let $p(x)$ and $p(y)$ be the corresponding unigram distributions. Consider a binary classifier that treats a word and its observed context as a positive example, and a word and a randomly sampled context as a negative example. The classification is made based on the embedding in such a way that the probability that (x,y) is a positive example is $\sigma(\vec{x} \cdot \vec{y})$. The objective function ideally maximized by the SGNS word embedding

algorithm is the expectation of the log-likelihood function of the embedding:

$$S(m) = h(\frac{1}{k+1}) + \frac{1}{k+1}\sum_{x,y} p(x,y)\log\sigma(\vec{x}\cdot\vec{y}) + \frac{k}{k+1}\sum_{x,y} p(x)p(y)\log\sigma(-\vec{x}\cdot\vec{y}). \quad (8)$$

Note that the term $h(\frac{1}{k+1})$, which does not appear in the original SGNS objective function (Mikolov et al., 2013), is a constant number that was added here to simplify the following presentation.

The sparsity of $p(x,y)$ (which is obtained as normalized counts from a given learning corpus) makes it feasible to compute the second term of (8). The number of summed-over elements in the third term of (8), however, is quadratic in the size of the vocabulary, making it hard to compute. Therefore, in practice, we can approximate the expectation by sampling of 'negative' examples. The actual SGNS score, then, is:

$$S(m) \approx h(\frac{1}{k+1}) + \frac{1}{k+1}\cdot\frac{1}{n}\sum_{t=1}^{n}(\log\sigma(\vec{x}_t\cdot\vec{y}_t) + \sum_{i=1}^{k}\log\sigma(-\vec{x}_t\cdot\vec{y}_{ti})). \quad (9)$$

such that t goes over all the word-context pairs in a given corpus. The negative examples y_{ti} are created for each pair (x_t, y_t) by drawing k random contexts from the context-word distribution $p(y)$.

As pointed out in (Levy et al., 2015), k has two distinct functions in the SGNS objective function. First, it is used to better estimate the distribution of negative examples. Second, it is used as a weight on the probability of observing a positive example versus a negative example; a higher k means that negative examples are more probable.

We can compute the SGNS score function $S(m)$ for every real-valued matrix $m = (m_{x,y})$. Levy and Goldberg (2014) showed that the function achieves its global maximal value when for each word-pair (x,y) the inner product of the embedding vectors $\vec{x}\cdot\vec{y}$ is equal to $\text{pmi}(x,y)$. In other words they showed that $S(m) \leq S(\text{pmi})$ for every matrix m. We next show that the value of the function $S(m)$ at its maximum point, the PMI matrix, has a concrete interpretation, namely it is exactly the Jensen-Shannon Mutual Information (JSMI) between words and their contexts.

Theorem 1: The value of the SGNS score with k negative samples (8) at the PMI matrix satisfies:

$$S(\text{pmi}) = \text{JSMI}_\alpha(X,Y)$$

such that $\alpha = \frac{1}{k+1}$.

Proof: It can be easily verified that by substituting $\alpha = \frac{1}{k+1}$ in the definition of JSMI (Eq. (7)), we exactly obtain the SGNS score (8) at the PMI matrix. □

Levy and Goldberg (2014) showed that SGNS's objective achieves its maximal value at the PMI matrix. However, this result reveals nothing about the more interesting lower dimensional case, where the PMI matrix factorization is forced to compress the joint distribution and thereby learn a meaningful embedding. We next derive an explicit description of the approximation criterion that quantifies the gap between $S(m)$ and $S(\text{pmi})$.

Given the word co-occurrences joint distribution $p(x,y)$, we obtained in Eq. (5) a conditional distribution on the alphabet of (Z,W) as follows:

$$p(Z=1|W=(x,y)) = \sigma(\text{pmi}_{x,y}).$$

In a similar way, given any matrix m, we can define a conditional distribution p_m on the alphabet of (Z,W) as follows:

$$p_m(Z=1|W=(x,y)) = \sigma(m_{x,y}).$$

Note that in the special case where m is the PMI matrix, $p_{\text{pmi}}(z|w)$ coincides with the original $p(z|w)$ that was defined in Eq. (5).

Theorem 2: The difference between the SGNS score at the PMI matrix and the SGNS score at a given matrix m can be written as:

$$S(\text{pmi}) - S(m) = \text{KL}(p_{\text{pmi}}(Z|W)||p_m(Z|W)) \quad (10)$$

Proof:

$$S(\text{pmi}) - S(m) = \sum_{x,y}(\alpha p(x,y)\log\frac{\sigma(\text{pmi}_{x,y})}{\sigma(m_{x,y})} + (1-\alpha)p(x)p(y)\log\frac{\sigma(-\text{pmi}_{x,y})}{\sigma(-m_{x,y})})$$

$$= \sum_{x,y}(\alpha p(x,y)\log\frac{p_{\text{pmi}}(Z=1|x,y)}{p_m(Z=1|x,y)} + (1-\alpha)p(x)p(y)\log\frac{p_{\text{pmi}}(Z=0|x,y)}{p_m(Z=0|x,y)})$$

$$= \sum_{w,z} p(W=w, Z=z) \log \frac{p_{\text{pmi}}(Z=z|W=w)}{p_m(Z=z|W=w)}$$

$$= \text{KL}(p_{\text{pmi}}(Z|W)||p_m(Z|W)).\square$$

The KL divergence between two distributions is always non-negative and is zero only if the two distributions are the same. Therefore, we re-derive the results of (Levy and Goldberg, 2014) that $S(\text{pmi}) = \max_m S(m)$. Theorem 2 can be viewed as an instance of the well-known connection between maximizing log-likelihood and minimizing KL divergence between the estimated and the true data-generating distribution. In this case, the true distribution is the pmi-based classifier $p_{\text{pmi}}(Z|W)$.

Combining theorems 1 and 2 we obtain that $S(m) \leq \text{JSMI}_\alpha(X,Y)$ for every low-dimensional embedding matrix. The difference $\text{JSMI}_\alpha(X,Y) - S(m)$ is the information loss caused by the low-dimensional embedding. We can view it as a Jensen-Shannon variant of the information bottleneck principle (Tishby et al., 1999; Globerson et al., 2007) that is defined in terms of the KL divergence. The optimal d-dimensional embedding, is the best d-dimensional approximation of the JSMI dependency measure in the sense that it minimizes the information loss. The JSMI is the upper bound that any embedding can obtain. To illustrate that, in the next section we compute the JSMI between words and their contexts based on a standard text corpus and show the information gap between the JSMI and the actual SGNS score as a function of the embedding dimension d.

From Theorem 2 we can also derive an explicit information-theoretic interpretation of the score function $S(m)$ (7) as the difference between two KL-divergence terms:

$$S(m) = S(\text{pmi}) - (S(\text{pmi}) - S(m)) =$$

$$I(Z;W) - (S(\text{pmi}) - S(m)) =$$

$$\text{KL}(p(Z|W)||p(Z)) - \text{KL}(p(Z|W)||p_m(Z|W))$$

The word embedding problem can be also viewed as a factorization of the PMI matrix. Previous works suggested other criteria for matrix factorization such as least-squares (Eckart and Young, 1936) and KL-divergence between the original matrix and the low-rank matrix approximation (Lee and Seung, 2000). We have shown that the SGNS algorithm factorizes the PMI matrix based on the JSMI-based criterion stated in Eq. (10).

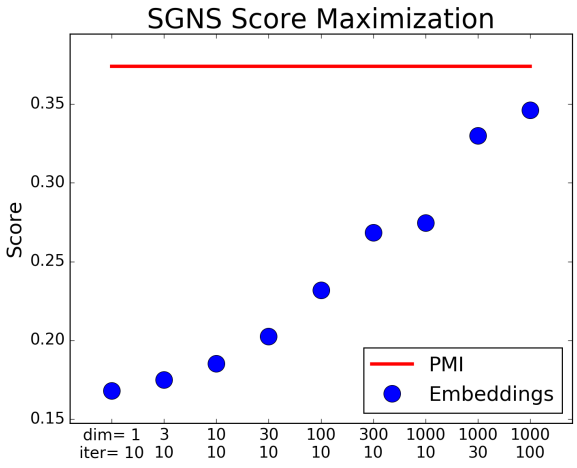

Figure 1: SGNS objective function score of trained embeddings models, compared to the optimal PMI-based score. *dim* and *iter* denote the dimensionality and training iterations used for each model.

4 Experiments

In this section we use *word2vec* to train real skip-gram with negative sampling (SGNS) embedding models. By measuring the value of their objective function and comparing it against the optimal one using exact PMI values, we demonstrate how a well-trained model minimizes the difference in Eq. (10). We note that this is an intrinsic measure that does not necessarily reflect the usefulness of the learned embeddings for other tasks.

We used the Penn Tree Bank (PTB), a popular small-scale corpus, for our experiments. A version of this dataset is available from Tomas Mikolov.[1] It consists of 929K training words with a 10K word vocabulary, which we used to train our models. To learn the SGNS word embeddings, we used word2vec's default parameter values: window-size = 5, min-count = 5, and number of negative samples $k = 5$. We varied the dimensionality of the embeddings and the number of training iterations performed. Once the models were trained, we measured their score (9) on the training corpus.

Based on the same learning corpus, we computed $S(\text{pmi}) = \text{JSMI}_\alpha(X,Y)$ for $\alpha = \frac{1}{k+1} = 1/6$. Note that $p(x,y) = 0$ implies that $\text{pmi}_{x,y} = -\infty$ and therefore $\log \sigma(-\text{pmi}_{x,y}) = 0$. Hence, as in the second term, to compute the third term of $S(m)$ (8) for the case of $m = \text{pmi}$, we can sum only

[1] http://www.fit.vutbr.cz/~imikolov/rnnlm/simple-examples.tgz

over the positive pairs (x, y) that actually appear in the corpus.[2] In other words, for the special case $m = \text{pmi}$, it is feasible to compute the exact score (8) and not just its approximation (9) that is based on negative sampling. Figure 1 illustrates the optimal PMI-based score, compared with the scores obtained by different models with varied embedding dimensionality and number of training iterations. As can be seen, the embeddings score gets close to the optimal value using higher dimensionality and more training iterations, but doesn't surpass it.

5 Conclusion

In this study, we developed a new correlation measure between random variables, denoted JSMI. This measure is based on the JS divergence and differs from the standard mutual information measure that is based on the KL divergence. We showed that the optimization of skip-gram embeddings with negative sampling finds the best low-dimensional approximation of the JSMI measure. Thus, we provided an information theory framework that hopefully contributes to a better understanding of this embedding algorithm. Furthermore, although we focused here on the case of word-context joint distributions, the connection we haven shown between the PMI matrix and the JSMI function is valid for every joint distribution of two random variables.

Acknowledgments

This work is supported by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics* 4:385–399.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Association for Computational Linguistics (ACL)*.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1:211–218.

[2] We used the exact same positive co-occurrence pairs sampled by word2vec during the training of the SGNS embeddings to compute $S(\text{pmi})$.

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftaly Tishby. 2007. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research* 8:2265–2295.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Trans. of the Association for Computational Linguistics* 3:211–225.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Conference on Natural Language Learning (CoNLL)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Karl Stratos, Michael Collins, and Daniel J Hsu. 2015. Model-based word embeddings from decompositions of count matrices. In *ACL (1)*. pages 1282–1291.

Naftaly Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Allerton Conf. on Communication, Control, and Computing*.