Infinite Dimensional Word Embeddings

Eric Nalisnick *
Department of Computer Science
University of California, Irvine
Irvine, CA 92697, USA
enalisni@uci.edu

Sachin Ravi *
Department of Computer Science
Princeton University
Princeton, NJ 08540, USA
sachinr@cs.princeton.edu

1 Introduction

(Neural) Word Embeddings (WEs) (Bengio et al., 2003; Mnih & Hinton, 2009; Turian et al., 2010; Mikolov et al., 2013) have received wide-spread attention for their ability to capture surprisingly detailed semantic information without supervision. However, despite their success, WEs still have deficiencies. One potential flaw is that the vectors, since their dimensionality is fixed across the vocabulary, do not accurately reflect each word's semantic complexity. For instance, the meaning of the word *race* varies with context (ex: competition vs anthropological classification), but the meaning of *regatta* is rather specific and invariant. It seems unlikely that *race* and *regatta*'s representations could contain the same number of parameters without one overfitting or underfitting.

To better capture the semantic variability of words, we propose a novel extension to the popular *Skip-Gram* and *Continuous Bag-of-Words* models (Mikolov et al., 2013) that allows vectors to have stochastic, data-dependent dimensionality. By employing the same mathematical tools that allow the definition of an *Infinite Restricted Boltzmann Machine* (Côté & Larochelle, 2016), we define two log-bilinear energy-based models named *Infinite Skip-Gram* (iSG) and *Infinite Continuous Bag-of-Words* (iCBOW) after their fixed dimensional counterparts. During training, iSG and iCBOW allow word representations to grow naturally based on how well they can predict their context. This behavior, among other things, enables the vectors of specific words to use few dimensions (since their context is reliably predictable) and the vectors of vague or polysemous words to elongate to capture as wide a semantic range as needed. As far as we are aware, this is the first word embedding method that allows each vector's dimensionality to be learned. We demonstrate the usefulness of iSG and iCBOW qualitatively and qualitatively in the experiments.

2 Infinite Skip-Gram

We begin by describing the Infinite Skip-Gram (iSG) model, which, when given a word, predicts one of its neighboring words within a fixed context window. Let word vectors $\mathbf{w}_i \in \mathbb{R}^\infty$ and context vectors $\mathbf{c}_k \in \mathbb{R}^\infty$ be infinite dimensional, and define the iSG model to be the following joint Gibbs distribution over $\mathbf{w}_i$, $\mathbf{c}_k$, and a random positive integer $z \in \mathbb{Z}^+$ denoting the maximum index over which to compute the vector inner product: $p(w_i, c_k, z) = \frac{1}{Z} e^{-E(\mathbf{w}_i, \mathbf{c}_k, z)}$ where $Z = \sum_{\mathbf{w}} \sum_{\mathbf{c}} \sum_z e^{-E(\mathbf{w}, \mathbf{c}, z)}$, also known as the partition function. Define the energy function as $E(\mathbf{w}_i, \mathbf{c}_k, z) = z \log a - \sum_{j=1}^{z} w_{i,j} c_{k,j} - \lambda w_{i,j}^2 - \lambda c_{k,j}^2$ where $1 < a < \infty$, $a \in \mathbb{R}$ and λ is a weight on the L2 penalty. The $\log a$ term, the same as used in (Côté & Larochelle, 2016)'s iRBM, is essential because it defines the infinite sum over dimensions to be a convergent geometric series. See the Appendix for the assumptions underlying and derivation of the partition function.

Learning. For iSG's learning objective, ideally, we would like to integrate out z:

$$\log p(c_k|w_i) = \log \sum_{z=1}^{\infty} p(c_k, z|w_i) = \log \left[\sum_{z=1}^{l} p(c_k, z|w_i) + \frac{a}{a-1} p(c_k, l|w_i) \right]. \tag{1}$$

The problem is that this quantity is computable only if l is known. It must exist under the sparsity/optimization assumption that makes the partition function finite (see Appendix), but that

*Authors contributed equally.

assumption gives no information about l's value. One work-around is to set l to some static value, but that would subvert the motivation for using an infinite dimensional model.

A better option is to sample z values and rely on the randomness to make learning dynamic yet tractable. This way l can grow arbitrarily (i.e. its the observed maximum sample) while the vectors have a finite number of dimensions. Thus we write the loss in terms of an expectation

$$\mathcal{L}_{iSG} = \log p(c_k|w_i) = \mathbb{E}_{z|c_k,w_i}\left[\log p(c_k, z|w_i) - \log p(z|c_k, w_i)\right]. \tag{2}$$

Notice that this is the evidence bound widely used for variational inference except here there is equality, not a bound, because we have set the variational distribution $q(z)$ to the posterior $p(z|w, c)$, which is tractable. The sampling we desire then comes about via a score function estimate of the gradient:

$$\frac{\partial}{\partial w_i}\mathcal{L}_{iSG} \approx \frac{1}{S}\sum_{s=1}^{S}\frac{\partial}{\partial w_i}\log p(c_k, \hat{z}_s|w_i) + \left[\log p(c_k|w_i) - 1\right]\frac{\partial}{\partial w_i}\log p(\hat{z}_s|c_k, w_i) \tag{3}$$

where S samples are drawn from $\hat{z}_s \sim p(z|c_k, w_i)$. Note the presence of the $p(c_k|w_i)$ term—the very term that we said was problematic in Equation 1 since l was not known. We can compute this term in the Monte Carlo objective by setting l to be the largest $\hat{z}$ value sampled up to that point in training. The presence of $p(c_k|w_i)$ is a boon because, since it does not depend $\hat{z}$, there is no need for control variates to stabilize the typically high variance term $\frac{\partial}{\partial w_i}\log p(\hat{z}_s|c_k, w_i)$.

Yet there's still a problem in that $\hat{z} \in [1, \infty)$ and therefore a very large dimensionality (say, a few thousand or more) could be sampled, resulting in the gradient incurring painful computational costs. To remedy this situation, if a $\hat{z}$ value greater than the current value of l is sampled, we set $\hat{z} = l + 1$, restricting the model to grow only one dimension at a time (just as done for the iRBM). Constraining growth in this way is computationally efficient since $\hat{z}$ can be drawn from a $(l + 1)$-dimensional multinoulli distribution with parameters $\Theta = \{\theta_1 = p(z = 1|w, c), \ldots, \theta_{l+1} = \frac{a}{a-1}p(z = l|w, c)\}$. The intuition is the model can sample a dimension less than or equal to l if l is already sufficiently large or draw the $(l + 1)$th option if not, choosing to increase the model's capacity. The hyperparameter a controls this growing behavior: as a approaches one (from the right), $P(z > l|w)$ approaches infinity.

3 Infinite Continuous Bag-of-Words

We next describe the Infinite Continuous Bag-of-Words (iCBOW) model, which predicts a word given multiple surrounding context words. Let word vectors $\mathbf{w}$ and context vectors $\mathbf{c}$ be defined just as for iSG. The iCBOW is a conditional Gibbs distribution over a center word $\mathbf{w}_i$ and a random positive integer $z \in \mathbb{Z}^+$ denoting the maximum index as before, given multiple context words $\mathbf{c}_k$: $p(w_i, z|c_{i-K}, \ldots, c_{i+K}) = \frac{1}{Z_{w,z}}e^{-\frac{1}{2K-1}\sum_j E(\mathbf{w}_i, \mathbf{c}_j, z)}$ where $Z_{w,c} = \sum_{\mathbf{w}}\sum_z e^{-\frac{1}{2K-1}\sum_j E(\mathbf{w}, \mathbf{c}_j, z)}$. The energy function is defined just as for the iSG and admits a finite partition function using the same arguments. The primary difference between the iSG and iCBOW is that the latter assumes all words appearing together in a window have the same vector dimensionality. The iSG, on the other hand, assumes just word-context *pairs* share dimensionality.

Learning. Like with iSG, learning iCBOW's parameters is done via a Monte Carlo objective. Define the iCBOW objective $\mathcal{L}_{iCBOW}$ as

$$\mathcal{L}_{iCBOW} = \log p(w_i|c_{i-K} \ldots c_{i+K}) = \mathbb{E}_z[\log p(w_i, z|c_{i-K} \ldots c_{i+K}) - \log p(z|w_i, c_{i-K} \ldots c_{i+K})]. \tag{4}$$

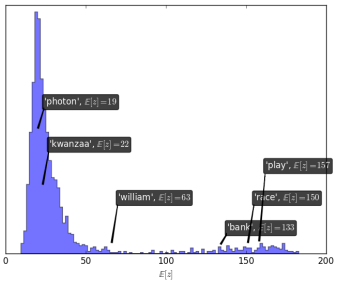
Again we use a score function estimate of the gradient to produce dynamic vector growth:

$$\begin{aligned}\frac{\partial}{\partial w_i}\mathcal{L}_{iCBOW} \approx \frac{1}{S}\sum_{s=1}^{S}&\frac{\partial}{\partial w_i}\log p(w_i, \hat{z}_s|c_{i-K}, \ldots, c_{i+K}) \\ &+ \left[\log p(w_i|c_{i-K}, \ldots) - 1\right]\frac{\partial}{\partial w_i}\log p(\hat{z}_s|w_i, c_{i-K}, \ldots)\end{aligned} \tag{5}$$

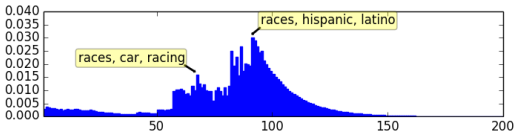
where S samples are drawn from $\hat{z}_s \sim p(z|w_i, c_{i-K}, \ldots, c_{i+K})$. Vectors are constrained to grow only one dimension at a time as done for the iSG by sampling from a $l + 1$th dimensional multinoulli with $\theta_{l+1} = \frac{a}{a-1}p(z = l|w_i, c_{i-K}, \ldots, c_{i+K})$.

	WordSim-353	MEN
SG-50	0.607	0.650
CBOW-50	0.609	0.659
iCBOW	0.614	0.660
iSG	0.620	0.674
CBOW-200	0.643	0.712
SG-200	**0.696**	**0.736**

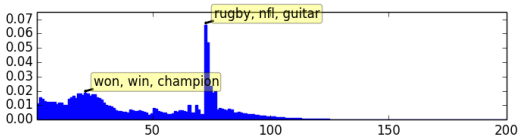
(a) Semantic similarity (Spearman's rank correlation)



(b) Histogram of the expected dimensionalities.



(c) $p(z|w =$ 'race')



(d) $p(z|w =$ 'player')

Figure 1: Subfigure (a) shows results for semantic similarity tasks. The Spearman's rank correlation between model and human scores are calculated for the WordSim-353 and MEN datasets. Subfigure (b) shows a histogram of the expected vector dimensionalities after training iCBOW. Subfigures (c) and (d) show the distribution over dimensionalities iSG learned for the words *race* and *player*.

4 Evaluation

We evaluate iSG and iCBOW quantitatively and qualitatively against original Skip-Gram (SG) and Continuous Bag-of-Words (CBOW). For all experiments, models were trained on a one billion word subset of Wikipedia (6/29/08 snapshot). The same learning rate ($\alpha = 0.05$ for CBOW, $\alpha = 0.025$ for SG), number of negative samples (5), context window size (6), and number of training epochs (1) were used for all models. iSG and iCBOW were initialized to ten dimensions.

Quantitative Evaluation. We test each model's ability to rank word pairs according to their semantic similarity, a task commonly used to gauge the quality of WEs. We evaluate our embeddings on two standard test sets: WordSim353 (Finkelstein et al., 2001) and MEN (Bruni et al., 2014). As is typical for evaluation, we measure the Spearman's rank correlation between the similarity scores produced by the model and those produced by the humans. The correlations for the proposed infinite dimensional models and for their finite dimensional analogs are reported in Subtable (a) of Figure 1. We see that the iSG and iCBOW perform better than their 50 dimensional counterparts but worse than their 200 dimensional counterparts. All scores are relatively competitive though, separated by no more than 0.1.

Qualitative Evaluation. Observing that the iSG and iCBOW models perform comparatively to finite versions somewhere between 50 and 200 dimensions, we qualitatively examine their distributions over vector dimensionalities. Subfigure (b) of Figure 1 shows a histogram of the expected dimensionality—i.e. $\mathbb{E}_{z|w,c}[z]$—of each vector after training the iCBOW model. As expected, the distribution is long-tailed, and vague words occupy the tail while specific words are found in the head. As shown by the annotations, the word *photon* has an expected dimensionality of 19 while the homograph *race* has 150. Note that expected dimensionality correlates with word frequency—due to the fact that multi-sense words, by definition, can be used more frequently—but does not follow it strictly. For instance, the word *william* is the 482nd most frequently occurring word in the corpus but has an expected length of 62, which is closer to the lengths of much rarer words (around 20-40 dimensions) than to similarly frequent words.

In subfigures (c) and (d) of Figure 1, we plot the quantity $p(z|w)$ for two homographs, *race* (c) and *player* (d), as learned by iSG, in order to examine if their multiple meanings are conspicuous in their distribution over dimensionalities. For *race*, we see that the distribution does indeed have at least two modes: the first at around 70 dimensions represents car racing, as determined by computing nearest neighbors with that dimension at a cutoff, while the second at around 100 dimensions encodes the anthropological meaning.

REFERENCES

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, 2014.

Marc-Alexandre Côté and Hugo Larochelle. An infinite restricted boltzmann machine. *Neural computation*, 2016.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 406–414, 2001.

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

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pp. 1081–1088, 2009.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394. Association for Computational Linguistics, 2010.

Appendix

A Finite Partition Function

iSG's partition function, containing a sum over all countably infinite values of z, would seem to be divergent and thus incomputable. However, it is not, due to two key properties first proposed by Côté & Larochelle (2016) to define a *Restricted Boltzmann Machine* with an infinite number of hidden units (iRBM). They are:

1. **Sparsity penalty**: The $L2$ penalty in $E(\mathbf{w}_i, \mathbf{c}_k, z)$ (i.e. the $w_{i,j}^2$ and $c_{k,j}^2$ terms) ensures the word and context vectors must have a finite two-norm under iterative gradient-based optimization with a finite initial condition. In other words, no proper optimization method could converge to the infinite solution if all $\mathbf{w}$ and $\mathbf{c}$ vectors are initialized to have a finite number of non-zero elements (Côté & Larochelle, 2016).
2. **Per-dimension constant penalty**: The energy function's $z \log a$ term results in dimensions greater than l becoming a convergent geometric series. This is discussed further below.

With those two properties in mind, consider the conditional distribution of z given an input and context word:

$$p(z|w, c) = \frac{e^{-E(\mathbf{w},\mathbf{c},z)}}{\sum_{z'=1}^{\infty} e^{-E(\mathbf{w},\mathbf{c},z')}}. \tag{6}$$

Again, the denominator looks problematic due to the infinite sum, but notice the following:

$$\begin{aligned} Z_z &= \sum_{z'=1}^{l} e^{-E(\mathbf{w},\mathbf{c},z')} + \sum_{z'=l+1}^{\infty} e^{-E(\mathbf{w},\mathbf{c},z')} \\ &= \sum_{z'=1}^{l} e^{-E(\mathbf{w},\mathbf{c},z')} + e^{-E(\mathbf{w},\mathbf{c},l)} \sum_{z'=0}^{\infty} \frac{1}{a^{z'}} \\ &= \sum_{z'=1}^{l} e^{-E(\mathbf{w},\mathbf{c},z')} + \frac{a}{a-1} e^{-E(\mathbf{w},\mathbf{c},l)}. \end{aligned} \tag{7}$$

The sparsity penalty allows the sum to be split as it is in step #2 into a finite term ($\sum_{z'=1}^{l} e^{-E(\mathbf{w},\mathbf{c},z')}$) and an infinite sum ($\sum_{z'=l+1}^{\infty} e^{-E(\mathbf{w},\mathbf{c},z')}$) at an index l such that $w_{i,j} = c_{k,j} = 0 \;\; \forall j > l$. After $e^{-E(\mathbf{w},\mathbf{c},l)}$ is factored out of the second term, all remaining $w_{i,j}$ and $c_{k,j}$ terms are zero. A few steps of algebra then reveal the presence of a convergent geometric series. Intuitively, we can think of the second term, $\frac{a}{a-1} e^{-E(\mathbf{w},\mathbf{c},l)}$, as quantifying the data's need to expand the model's capacity given w and c.