Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis

Ryan Cotterell  Adam Poliak  Benjamin Van Durme  Jason Eisner
Center for Language and Speech Processing
Johns Hopkins University
{ryan.cotterell,azpoliak,vandurme,jason}@cs.jhu.edu

Abstract
The popular skip-gram model induces word embeddings by exploiting the signal from word-context cooccurrence. We offer a new interpretation of skip-gram based on exponential family PCA—a form of matrix factorization. This makes it clear that we can extend the skip-gram method to tensor factorization, in order to train embeddings through richer higher-order cooccurrences, e.g., triples that include positional information (to incorporate syntax) or morphological information (to share parameters across related words). We experiment on 40 languages and show that our model improves upon skip-gram.

1 Introduction
Over the past years NLP has witnessed a veritable frenzy on the topic of word embeddings: low-dimensional representations of distributional information. The embeddings, trained on extremely large text corpora such as Wikipedia and Common Crawl, are claimed to encode semantic knowledge extracted from large text corpora.

Numerous methods have been proposed—the most popular being skip-gram (Mikolov et al., 2013) and GloVe (Pennington et al., 2014)—for learning these low-dimensional embeddings from a bag of contexts associated with each word type. Natural language text, however, contains richer structure than simple context-word pairs. In this work, we embed \( n \)-tuples rather than pairs, allowing us to escape the bag-of-words assumption and encode richer linguistic structures.

As a first step, we offer a novel interpretation of the skip-gram model (Mikolov et al., 2013). We show how skip-gram can be viewed as an application of exponential-family principal components analysis (EPCA) (Collins et al., 2001) to an integer matrix of cooccurrence counts. Previous work has related the negative sampling estimator for skip-gram model parameters to the factorization of a matrix of (shifted) positive pointwise mutual information (Levy and Goldberg, 2014b). We show the skip-gram objective is just EPCA factorization.

By extending EPCA factorization from matrices to tensors, we can consider higher-order cooccurrence statistics. Here we explore incorporating positional and morphological content in the model by factorizing a positional tensor and morphology tensor. The positional tensor directly incorporates word order into the model, while the morphology tensor adds word-internal information. We validate our models experimentally on 40 languages and show large gains under standard metrics.\(^1\)

2 Matrix Factorization
In this section, we briefly explain how skip-gram is an example of EPCA. We are given data in the form of a matrix \( X \in \mathbb{R}^{n_1 \times n_2} \), where \( X_{ij} \) is the number of times that word \( j \) appears in context \( i \) under some user-specified definition of “context.”

Principal components analysis (Pearson, 1901) approximates \( X \) as the product \( C^T W \) of two matrices \( C \in \mathbb{R}^{d \times n_1} \) and \( W \in \mathbb{R}^{d \times n_2} \), whose columns are \( d \)-dimensional vectors that embed the contexts and the words, respectively, for some user-specified \( d < \min(n_1, n_2) \). Specifically, PCA minimizes\(^2\)

\[
\|X - C^T W\|_F^2 = \sum_{ij} (X_{ij} - c_i \cdot w_j)^2 \tag{1}
\]

\[
= \sum_j \left\| x_j - C^T w_j \right\|^2 \tag{2}
\]

where \( c_i, w_j, x_j \) denote the \( i \)th column of \( C \) and the \( j \)th columns of \( W \) and \( X \), and \( c_i \cdot w_j \) denotes an inner product of vectors (sometimes called “cosine

\(^1\)The code developed is available at https://github.com/azpoliak/skip-gram-tensor.
\(^2\)Singh and Gordon (2008) offer a comprehensive discussion of PCA and other matrix factorization techniques in ML.
Figure 1: Comparison of the graphical model for matrix factorization (either PCA or EPCA) and 3-dimensional tensor factorization. Priors are omitted from the drawing.

EPCA allows us to suppose that each $x_j$ was drawn from a multinomial—a more appropriate family for drawing a count vector. Our observation is that skip-gram is precisely multinomial EPCA with the canonical link function (Mohamed, 2011), where the count vector $x_j$ comes from a multinomial with canonical parameters $C^\top w_j$. It chooses embeddings $C, W$ to maximize

$$
\sum_j \sum_i x_{ij} \log p(\text{context } i \mid \text{word } j)
$$

This is the log-likelihood (plus a constant) if we assume that for each word $j$, the context vector $x_j$ was drawn from a multinomial with canonical parameter vector $C^\top w_j$ and count parameter $N_j = \sum_i x_{ij}$. This is the same model as in Figure 1a, but with a different conditional distribution for $x_j$, and with $x_j$ taking an additional observed parent $N_j$ (which is the token count of word $j$).

2.1 Related work

Levy and Goldberg (2014b) also interpreted skip-gram as matrix factorization. They argued that skip-gram estimation by negative sampling implicitly factorizes a shifted matrix of positive empirical pointwise mutual information values. We instead regard the skip-gram objective itself as demanding EPCA-style factorization of the count matrix $X$: i.e., $X$ arose stochastically from some unknown matrix of multinomial parameters (column $j$ of $X$ generated from parameter column $j$), and we seek a rank-$d$ estimate $C^\top W$ of that matrix.

pLSI (Hofmann, 1999) similarly factors an unknown matrix of multinomial probabilities, which is multinomial EPCA with the identity link function. In contrast, our unknown matrix holds the “canonical parameters” for a multinomial—arbitrarily shifted log-probabilities.

Our EPCA interpretation applies equally well to the component distributions that are used in hierarchical softmax (Morin and Bengio, 2005), which is an alternative to negative sampling. Additionally, it yields avenues of future research using Bayesian (Mohamed et al., 2008) and maximum-margin (Srebro et al., 2004) extensions to EPCA.

The canonical parameters of the distribution. We use the so-called canonical link, meaning that these two steps are inverses of each other and thus the canonical parameters are themselves a linear function of $w_j$.

$^3$The graphical model further suggests that the $c_i$ and $w_j$ vectors are themselves drawn from some prior. Specifying this prior defines a MAP estimate of $C$ and $W$. If we take the prior to be a spherical Gaussian with mean $0 \in \mathbb{R}^d$, the MAP estimate corresponds to minimizing (2) plus an $L_2$ regularizer, that is, a multiple of $\|C\|^2 + \|W\|^2$. We do indeed regularize in this way throughout all our experiments, tuning the multiplier on a held-out development set. However, regularization has only minor effects with large training corpora, and is not in the original word2vec implementation of skip-gram.

$^4$EPCA extends PCA in the same way that generalized linear models (GLMs) extend linear regression. The maximum-likelihood interpretation of linear regression supposes that the dependent variable $x_j$ is a linear function $C$ of the independent variable $w_j$ plus Gaussian noise. The GLM, like EPCA, is an extension that allows other exponential-family distributions for the dependent variable $x_j$. The difference is that in EPCA, the representations $w_j$ are learned jointly with $C$.

$^5$In the general form of EPCA, that column is passed through some “inverse link” function to obtain the expected feature values under the distribution, which in turn determines similarity”). Note that rank$(C^\top W) \leq d$, whereas rank$(X) \leq \min(n_1, n_2)$. Globally optimizing equation (1) means finding the best approximation to $X$ with rank $\leq d$ (Eckart and Young, 1936), and can be done by SVD (Golub and Van Loan, 2012).

By rewriting equation (1) as (2), both Roweis (1997) and Tipping and Bishop (1999) observed that the optimal values of $C$ and $W$ can be regarded as the maximum-likelihood parameter estimates for the Gaussian graphical model drawn in Figure 1a. This model supposes that the observed column vector $x_j$ equals $C^\top w_j$ plus Gaussian noise, specifically $x_j \sim N(C^\top w_j, I)$. Equation (2) is this model’s negated log-likelihood (plus a constant).

However, recall that in our application, $x_j$ is a vector of observed counts of the various contexts in which word $j$ appeared. Its elements are always non-negative integers—so as Hofmann (1999) saw, it is peculiar to model $x_j$ as having been drawn from a Gaussian. EPCA is a generalization of PCA, in which the observation $x_j$ can be drawn according to any exponential-family distribution (log-linear distribution) over vectors. The canonical parameter vector for this distribution is provided by the $j$th column of $C^\top W$, that is, $C^\top w_j$.
3 Tensor Factorization

Having seen that skip-gram is a form of matrix factorization, we can generalize it to tensors. In contrast to the matrix case, there are several distinct definitions of tensor factorization (Kolda and Bader, 2009). We focus on the polyadic decomposition (Hitchcock, 1927), which yields a satisfying generalization. The tensor analogue to PCA is rank-$d$ tensor approximation, which minimizes

$$
||X - C \otimes W \otimes R||_F^2 
= \sum_{ijk} (X_{ijk} - 1 \cdot (c_i \circ w_j \circ r_k))^2 
$$

(5)

$$
= \sum_{jk} \left| \left| X_{-j,k} - C^\top (w_j \circ r_k) \right| \right|^2 
$$

(6)

Given a tensor $X \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, this objective tries to predict each entry as the three-way dot product of the columns $c_i, w_j, r_k \in \mathbb{R}^d$, thus finding an approximation to $X$ that factorizes into $C, W, R$. This polyadic decomposition of the approximating tensor can be viewed as a Tucker decomposition (Tucker, 1966) that enforces a diagonal core.

In our setting, the new matrix $R \in \mathbb{R}^{d \times n_3}$ embeds types of context-word relations. The tensor $X$ can be regarded as a collection of $n_2 n_3$ count vectors $X_{i,j,k} \in \mathbb{N}^{n_1}$: the fibers of the tensor, each of which provides the context counts for some (word $j$, relation $k$) pair. Typically, $X_{i,j,k}$ counts which context words $i$ are related to word $j$ by relation $k$.

We now move from third-order PCA to third-order EPCA. Minimizing equation (6) corresponds to maximum-likelihood estimation of the graphical model in Figure 1b, in which each fiber of $X$ is viewed as being generated from a Gaussian all at once. Our higher-order skip-gram (HOSG) replaces this Gaussian with a multinomial. Thus, HOSG attempts to maximize the log-likelihood

$$
\sum_{ijk} X_{ijk} \log p(\text{context } i \mid \text{word } j, \text{relation } k) 
$$

(7)

$$
= \sum_{ijk} X_{ijk} \log \frac{\exp (1 \cdot (c_i \circ w_j \circ r_k))}{\sum_{i'} \exp (1 \cdot (c_{i'} \circ w_j \circ r_k))} 
$$

(8)

Note that as before, we are taking the total count $N_{jk} = \sum_i X_{ijk}$ to be observed. So while our embedding matrices must predict which words are related to word $j$ by relation $k$, we are not probabilistically modeling how often word $j$ participates in relation $k$ in the first place (nor how often word $j$ occurs overall). A simple and natural move in future would be to extend the generative model to predict these facts also from $w_j$ and $r_k$, although this weakens the pedagogical connection to EPCA.

We locally optimize the parameters of our probability model—the word, context and relation embeddings—through stochastic gradient ascent on (7). Each stochastic gradient step computes the gradient of a single summand $X_{ijk} \log p(i \mid j, k)$. Unfortunately, this requires summing over $n_1$ contexts in the denominator of (8), which is problematic as $n_1$ is often very large, e.g., $10^7$. Mikolov et al. (2013) offer two speedup schemes: negative sampling and hierarchical softmax. Here we apply the negative sampling approximation to HOSG; hierarchical softmax is also applicable. See Goldberg and Levy (2014) for an in-depth discussion.

HOSG is a bit slower to train than skip-gram, since $X$ yields up to $n_3$ times as many summands as $X$ (but $\ll n_3$ in practice, as $X$ is often sparse).

4 Two Tensors for Word Embedding

As examples of useful tensors to factorize, we offer two third-order generalizations of Mikolov et al. (2013)’s context-word matrix. We are still predicting the distribution of contexts of a given word type. Our first version increases the number of parameters (giving more expressivity) by conditioning on additional information. Our second version decreases the number of parameters (giving better smoothing) by factoring the word type.

4.1 Positional Tensor

When predicting the context words in a window around a given word token, Mikolov et al. (2013) uses the same distribution to predict each of them. We propose to use different distributions at different positions in the window, via a “positional tensor”: $X_{(\text{dog, ran, } -2)}$ is the number of times the context word dog was seen two positions to the left of ran. We will predict this count using $p(\text{dog} \mid \text{ran, } -2)$, defined from the embeddings of the word ran, the position $-2$, and the context word dog and its competitors. For a 10-word window, we have $X \in \mathbb{R}^{V \times V \times 10}$. Considering word position should improve syntactic awareness.

4.2 Compositional Morphology Tensor

For Mikolov et al. (2013), related words such as ran and running are monolithic objects that do not share parameters. We decompose each word into a lemma $j$ and a morphological tag $k$. The
contexts $i$ are still full words. Thus, we predict the count $X_{(d_{og,run,t})}$ using $p(d_{og} \mid run, t)$, where $t$ is a morphological tag such as $[\text{pos}=v, \text{tense}=\text{PAST}]$.

Our model is essentially a version of the skipgram method (Mikolov et al., 2013) that parameterizes the embedding of the word $\text{ran}$ as a Hadamard product $w_j \odot r_k$, where $w_j$ embeds $\text{RUN}$ and $r_k$ embeds $\text{TAG}$. This is similar to the work of Cotterell et al. (2016), who parameterized word embeddings as a sum $w_j + r_k$ of embeddings of the component morphemes. Our Hadamard product embedding is in fact more general, since the additive embedding $w_j + r_k$ can be recovered as a special case—it is equal to $(w_j, 1) \odot (1, r_k)$, which uses twice as many dimensions to embed each object.

## 5 Experiments

We build HOSG on top of the HYPERWORDS package. All models (both skip-gram and higher-order skip-gram) are trained for 10 epochs and use 5 negative samples. All models for §5.1 are trained on the Sept. 2016 dump of the full Wikipedia. All models for §5.2 were trained on the lemmatized and POS-tagged WaCky corpora (Baroni et al., 2009) for French, Italian, German and English (Joubarne and Inkpen, 2011; Leviant and Reichart, 2015). To ensure controlled and fair experiments, we follow Levy et al. (2015) for all preprocessing.

### 5.1 Experiment 1: Positional Tensor

We postulate that the positional tensor should encode richer notions of syntax than standard bag-of-words vectors. Why? Positional information allows us to differentiate between the geometry of the cocurrence, e.g., the $i$ is found to the left of the noun it modifies and is—more often than—close to it. Our tensor factorization model explicitly encodes this information during training.

To evaluate the vectors, we use QVEC (Tsvetkov et al., 2016), which measures Pearson’s correlation between human-annotated judgements and the vectors using CCA. The QVEC metric will be higher if the vectors better correlate with the human-annotated resource. To measure the syntactic content of the vectors, we compute the correlation between our learned vector $w_j$ for each word and its empirical distribution $g_i$ over universal POS tags (Petrov et al., 2012) in the UD treebank (Nivre et al., 2016). $g_i$ can be regarded as a vector on the $|\mathcal{T}| - 1$-dimensional simplex, where $\mathcal{T}$ is the tag set. We report results on 40 languages from the UD treebanks in Table 1, using 4-word or 10-word symmetric context windows (i.e., $c \in \{2, 5\}$). We find that 77.5% of the languages, our positional tensor embeddings outperform the standard skip-gram approach on the QVEC metric.

We highlight again that the positional tensor exploits no additional annotation, but better exploits the signal found in the raw text. Of course, our HOSG method could also be used to exploit annotations if available: e.g., one would get different embeddings by defining the relations of word $j$ to be the labeled syntactic dependency relations in which it participates (Lin and Pantel, 2001; Levy and Goldberg, 2014a).

### 5.2 Experiment 2: Morphology Tensor

Since the compositional morphology tensor allows us to share parameters among related word forms, we get a single embedding for each $\text{lemma}$, i.e., all the words $\text{ran}, \text{run}$ and $\text{running}$ now con-

---

Table 1: The scores for QVEC-CCA for 40 languages. All embeddings were trained on the complete Wikipedia dump of September 2016. We measure correlation with universal POS tags from the UD treebanks.

<table>
<thead>
<tr>
<th>Language</th>
<th>ar</th>
<th>bg</th>
<th>ca</th>
<th>cs</th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>et</th>
<th>eu</th>
<th>fa</th>
<th>fi</th>
<th>fr</th>
<th>ga</th>
<th>gl</th>
<th>he</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOSG</td>
<td>0.15 ± 0.04</td>
<td>0.19 ± 0.03</td>
<td>0.20 ± 0.03</td>
<td>0.21 ± 0.01</td>
<td>0.22 ± 0.02</td>
<td>0.22 ± 0.01</td>
<td>0.23 ± 0.01</td>
<td>0.24 ± 0.01</td>
<td>0.25 ± 0.02</td>
<td>0.26 ± 0.02</td>
<td>0.27 ± 0.02</td>
<td>0.28 ± 0.02</td>
<td>0.29 ± 0.02</td>
<td>0.30 ± 0.02</td>
<td>0.31 ± 0.02</td>
<td>0.32 ± 0.02</td>
<td>0.33 ± 0.02</td>
<td>0.34 ± 0.02</td>
</tr>
<tr>
<td>$c = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOSG</td>
<td>0.15 ± 0.03</td>
<td>0.16 ± 0.03</td>
<td>0.17 ± 0.03</td>
<td>0.18 ± 0.03</td>
<td>0.19 ± 0.03</td>
<td>0.20 ± 0.03</td>
<td>0.21 ± 0.03</td>
<td>0.22 ± 0.03</td>
<td>0.23 ± 0.03</td>
<td>0.24 ± 0.03</td>
<td>0.25 ± 0.03</td>
<td>0.26 ± 0.03</td>
<td>0.27 ± 0.03</td>
<td>0.28 ± 0.03</td>
<td>0.29 ± 0.03</td>
<td>0.30 ± 0.03</td>
<td>0.31 ± 0.03</td>
<td>0.32 ± 0.03</td>
</tr>
</tbody>
</table>

---

7If one wanted to extend the model to decompose the context words $i$ as well, we see at least four approaches.

8Cotterell et al. (2016) made two further moves that could be applied to extend the present paper. First, they allowed a word to consist of any number of (unordered) morphemes—not necessarily two—whose embeddings were combined (by summation) to get the word embedding. Second, this sum also included word-specific random noise, allowing them to learn word embeddings that deviated from compositionality.
Table 2: Word similarity results comparing the compositional morphology tensor with the standard skip-gram model. Numbers indicate Spearman’s correlation coefficient $\rho$ between human similarity judgements and the cosine distances of vectors. For each language, we compare on several sets of human judgments as listed by Faruqui et al. (2016, Table 2).

6 Related Work

Tensor factorization has already found uses in a few corners of NLP research. Van de Cruys et al. (2013) applied tensor factorization to model the compositionality of subject-verb-object triples. Similarly, Hashimoto and Tsuruoka (2015) use an implicit tensor factorization method to learn embeddings for transitive verb phrases. Tensor factorization also appears in semantic-based NLP tasks. Lei et al. (2015) explicitly factorize a tensor based on feature vectors for predicting semantic roles. Chang et al. (2014) use tensor factorization to create knowledge base embeddings optimized for relation extraction. See Bouchard et al. (2015) for a large bibliography.

Other researchers have likewise attempted to escape the bag-of-words assumption in word embeddings, e.g., Yatbaz et al. (2012) incorporates morphological and orthographic features into continuous vectors; Cotterell and Schütze (2015) consider a multi-task set-up to force morphological information into embeddings; Cotterell and Schütze (2017) jointly morphologically segment and embed words; Levy and Goldberg (2014a) derive contexts based on dependency relations; PPDB (Ganitkevitch et al., 2013) employs a mixed bag of words, parts of speech, and syntax; Rastogi et al. (2015) represent word contexts, morphology, semantic frame relations, syntactic dependency relations, and multilingual bitext counts each as separate matrices, combined via GCCA; and, finally, Schwartz et al. (2016) derived embeddings based on Hearst patterns (Hearst, 1992). Ling et al. (2015) learn position-specific word embeddings (§4.1), but do not factor them as $w_j \odot r_k$ to share parameters (we did not compare empirically to this). As demonstrated in the experiments, our tensor factorization method enables us to include other syntactic properties besides word order, e.g., morphology. Poliak et al. (2017) also create positional word embeddings. Our research direction is orthogonal to these efforts in that we provide a general purpose procedure for all sorts of higher-order cooccurrence.

7 Conclusion

We have presented an interpretation of the skip-gram model as exponential family principal components analysis—a form of matrix factorization—and, thus, related it to an older strain of work. Building on this connection, we generalized the model to the tensor case. Such higher-order skip-gram methods can incorporate more linguistic structure without sacrificing scalability, as we illustrated by making our embeddings consider word order or morphology. These methods achieved better word embeddings as evaluated by standard metrics on 40 languages.

Acknowledgements

We thank colleagues, particularly Yanif Ahmad, and anonymous reviewers for helpful discussion and feedback. The first author was supported by a DAAD Long-Term Research Grant and an NDSEG fellowship. This work was supported by the JHU Human Language Technology Center of Excellence (HLTCOE), DARPA DEFT, and NSF Grant No. 1423276. The U.S. Government is authorized to reproduce and distribute reprints for its purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.
References


Dekang Lin and Patrick Pantel. 2001. Dirt @sbt@discovery of inference rules from text. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’01, pages 323–328, New York, NY, USA. ACM.


