Spectral Graph-Based Method of Multimodal Word Embedding *

Kazuki Fukui[♠♡] and **Takamasa Oshikiri**[◇♡] and **Hidetoshi Shimodaira**[♠♡]
[♠] Department of Systems Science, Graduate School of Informatics, Kyoto University
[♡] Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project
[◇] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University
k.fukui@sys.i.kyoto-u.ac.jp
oshikiri@sigmath.es.osaka-u.ac.jp
shimo@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a novel method for multimodal word embedding, which exploit a generalized framework of multi-view spectral graph embedding to take into account visual appearances or scenes denoted by words in a corpus. We evaluated our method through word similarity tasks and a concept-to-image search task, having found that it provides word representations that reflect visual information, while somewhat trading-off the performance on the word similarity tasks. Moreover, we demonstrate that our method captures multimodal linguistic regularities, which enable recovering relational similarities between words and images by vector arithmetic.

1 Introduction

Word embedding plays important roles in the field of Natural Language Processing (NLP). Many existing studies use word vectors for various downstream NLP tasks, such as text classification, Part-of-Speech tagging, and machine translation. One of the most famous approaches is skip-gram model (Mikolov et al., 2013), which is based on a neural network, and its extensions have also been widely studied as well.

There are alternative approaches depending on a spectral graph embedding framework (Yan et al., 2007; Huang et al., 2012) for word embedding. For examples, Dhillon et al. (2015) proposed a method based on Canonical Correlation Analysis (CCA) (Hotelling, 1936), while a PCA based word embedding method was proposed in Lebret and Collobert (2014).

* This work was partially supported by grants from Japan Society for the Promotion of Science KAKENHI (16H02789) to HS.

In recent years, many researchers have been actively studying the use of multiple modalities in the fields of both NLP and computer vision. Those studies combine textual and visual information to propose methods for image-caption matching (Yan and Mikolajczyk, 2015), caption generation (Kiros et al., 2014), visual question answering (Antol et al., 2015), quantifying abstractness (Kiela et al., 2014) of words, and so on.

As for word embedding, multimodal versions of word2vec (Mikolov et al., 2013) have been proposed in Lazaridou et al. (2015) and Kottur et al. (2016). The first one jointly optimize the objective of both skip-gram model and a cross-modal objective across texts and images, and the latter uses abstract scenes as surrogate labels for capturing visually grounded semantic relatedness. More recently, Mao et al. (2016) proposed a multimodal word embedding methods based on a recurrent neural network to learn word vectors from their newly proposed large scale image caption dataset.

In this paper, we introduce a new spectral graph-based method of multimodal word embedding. Specifically, we extend Eigenwords (Dhillon et al., 2015), a CCA-based method for word embedding, by applying a generalized framework of spectral graph embedding (Nori et al., 2012; Shimodaira, 2016). Figure 1 shows a schematic diagram of our method.

In the rest of this paper, we call our method **Multimodal Eigenwords** (MM-Eigenwords). The most similar existing method is Multimodal Skip-gram model (MMskip-gram) (Lazaridou et al., 2015), which slightly differ in that our model can easily deal with many-to-many relationships between words in a corpus and their relevant images, while MMskip-gram only considers one-to-one relationships between concrete words and images.

Using a corpus and datasets of image-word rela-

Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, ACL 2017, pages 39–44,
Vancouver, Canada, August 3, 2017.

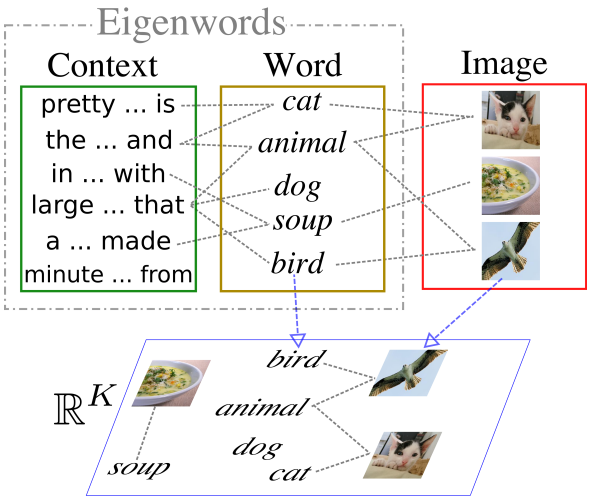


Figure 1: Our proposed method extends a CCA-based method of word embedding by means of multi-view spectral graph embedding frameworks of dimensionality reduction to deal with visual information associated with words in a corpus.

tionships, which are available in common benchmark datasets or on online photo sharing services, MM-Eigenwords jointly learns word vectors on a common multimodal space and a linear mapping from a visual feature space to the multimodal space. Those word vectors also reflect similarities between words and images.

We evaluated the multimodal word representations obtained by our model through word similarity task and concept-to-image search, having found that our model has ability to capture both semantic and word-to-image similarities. We also found that our model captures multimodal linguistic regularities (Kiros et al., 2014), whose examples are shown in Figure 2b.

2 Multi-view Spectral Graph Embedding

A spectral graph perspective of dimensionality reduction was first proposed in Yan et al. (2007), which showed that several major statistical methods for dimensionality reduction, such as PCA and Eigenmap (Belkin and Niyogi, 2003), can be written in a form of graph embedding frameworks, where data points are nodes and those points have weighted links between other points. Huang et al. (2012) extended this work for two-view data with many-to-many relationships (or links) and showed that their two-view graph embedding framework includes CCA, one of the most popular method for multi-view data analysis, as its special cases. However, available datasets may have more than two views with complex graph structures, which are unmanageable for CCA or Multiset CCA (Kettenring, 1971) whose inputs must be fed in the form of n-tuples.

Shimodaira (2016) further generalized the graph embedding frameworks to deal with many-to-many relationships between any number of views, and Nori et al. (2012) also proposed an equivalent method for multimodal relation prediction in social data. This generalized framework is used to extend Eigenwords for cross-lingual word embedding (Oshikiri et al., 2016), where vocabularies and contexts of multiple languages are linked through sentence-level alignment. Our proposed method also makes use of the framework of Shimodaira (2016) to extend Eigenwords for multimodal word embedding.

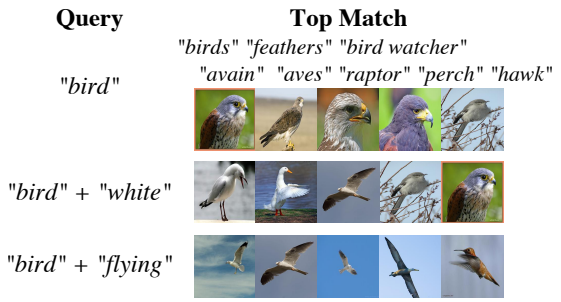
3 Eigenwords (One Step CCA)

Canonical Correlation Analysis (Hotelling, 1936) is a multivariate analysis method for finding optimal projections of two sets of data vectors by maximizing the correlations. Applying CCA to pairs of raw word vectors and raw context vectors, Eigenwords algorithms attempt to find low-dimensional vector representations of words (Dhillon et al., 2015). Here we explain the simplest version of Eigenwords called One Step CCA (OSCCA).

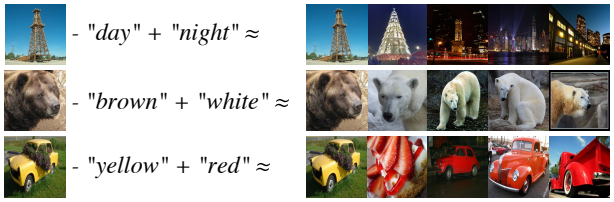
We have a corpus consisting of T tokens; $(t_i)_{i=1,\ldots,T}$, and the vocabulary consisting of V word types; $\{v_i\}_{i=1,\ldots,V}$. Each token t_i is drawn from this vocabulary. We define a word matrix $\mathbf{V} \in \{0,1\}^{T\times V}$ whose i-th row encodes the token t_i by 1-of-V representation; the j-th element is 1 if the word type of t_i is v_j, 0 otherwise.

Let h be the size of context window. We define context matrix $\mathbf{C} \in \{0,1\}^{T\times 2hV}$ whose i-th row represents the surrounding context of the token t_i with concatenated 1-of-V encoded vectors of $(t_{i-h},\ldots,t_{i-1},t_{i+1},\ldots,t_{i+h})$.

We apply CCA to T pairs of row vectors of $\mathbf{V}$ and $\mathbf{C}$. The objective function of CCA is constructed using $\mathbf{V}^\top\mathbf{V}$, $\mathbf{V}^\top\mathbf{C}$, $\mathbf{C}^\top\mathbf{C}$ which represent occurrence and co-occurrence counts of words and contexts. In Eigenwords, however, we use $\boldsymbol{\mathcal{C}}_{VV} \in \mathbb{R}_+^{V\times V}$, $\boldsymbol{\mathcal{C}}_{VC} \in \mathbb{R}_+^{V\times 2hV}$, $\boldsymbol{\mathcal{C}}_{CC} \in \mathbb{R}_+^{2hV\times 2hV}$ with the following preprocessing of these matrices before constructing the objective function. First, centering-process of $\mathbf{V}$ and $\mathbf{C}$ is



(a) Word-to-Image Search.



(b) Examples of Multimodal Linguistic Regularities.

Figure 2: Examples of word-to-image search (a) and demonstrations of vector arithmetics between words and images (b). We chose $\eta = 10^6$ in these examples.

omitted, and off-diagonal elements of $\mathbf{C}^\top\mathbf{C}$ are ignored for simplifying the computation of inverse matrices. Second, we take the square root of the elements of these matrices for "squashing" the heavy-tailed word count distributions. Finally, we obtain vector representations of words as $\boldsymbol{\mathcal{C}}_{VV}^{-1/2}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$, where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K \in \mathbb{R}^V$ are left singular vectors of $\boldsymbol{\mathcal{C}}_{VV}^{-1/2}\,\boldsymbol{\mathcal{C}}_{VC}\,\boldsymbol{\mathcal{C}}_{CC}^{-1/2}$ corresponding to the K largest singular values.

For the fast and scalable computation, Dhillon et al. (2015) employed the method of Halko et al. (2011) which use random projections to compute singular value decomposition of large matrices.

4 Multimodal Eigenwords

In this section, we introduce Multimodal Eigenwords (MM-Eigenwords) by extending the CCA based model of Eigenwords to obtain multimodal representations across words and images.

Suppose we have N_{vis} images, and each image is associated with multiple tags (or words). These associations are denoted by $\tilde{w}_{ij} \geq 0$ $(1 \leq i \leq V, 1 \leq j \leq N_{vis})$, whose value represents the strength of a relationship between the i-th word and the j-th image. In this study, for example, $\tilde{w}_{ij} = 1$ if the j-ith image has the i-th word as its tag, whereas $\tilde{w}_{ij} = 0$ otherwise, and we define a matrix $\widetilde{\mathbf{W}}_{VX} = (\tilde{w}_{ij})$. In addition, we denote a image feature matrix by $\mathbf{X}_{vis} \in \mathbb{R}^{N_{vis} \times p_{vis}}$ and its i-th row vector $\boldsymbol{x}_i$, as well as row vectors of $\mathbf{V}$, $\mathbf{C}$ by $\boldsymbol{v}_i$, $\boldsymbol{c}_i$ respectively. Here, the goal of MM-Eigenwords is to obtain multimodal representations by extending the CCA in Eigenwords with generalized frameworks of multi-view spectral graph embedding (Nori et al., 2012; Shimodaira, 2016), which include CCA as their special cases. In these frameworks, our goal can be attained by finding an optimal linear mappings to the K-dimensional multimodal space $\mathbf{A}_V$, $\mathbf{A}_C$, $\mathbf{A}_{vis}$ that minimize the following objective with a scale constraint.

$$\sum_{i=1}^{T} \|\boldsymbol{v}_i\mathbf{A}_V - \boldsymbol{c}_i\mathbf{A}_C\|_2^2 + \sum_{i=1}^{T}\sum_{j=1}^{N_{vis}} \eta w_{ij} \|\boldsymbol{v}_i\mathbf{A}_V - \boldsymbol{x}_j\mathbf{A}_{vis}\|_2^2, \quad (1)$$

where $w_{ij} = (\mathbf{V}\widetilde{\mathbf{W}}_{VX})_{ij}$, and the multimodal term coefficient $\eta \geq 0$ determines to which extent the model reflects the visual information. Considering a scale constraint, Eq. (1) can be reformulated as follows:

We first define some matrices

$$\mathbf{X} = \begin{pmatrix} \mathbf{V} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{C} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{X}_{vis} \end{pmatrix}, \mathbf{W} = \begin{pmatrix} \mathbf{O} & \mathbf{I}_T & \mathbf{W}_{VX} \\ \mathbf{I}_T & \mathbf{O} & \mathbf{O} \\ \mathbf{W}_{VX}^\top & \mathbf{O} & \mathbf{O} \end{pmatrix},$$
$$\mathbf{M} = \mathrm{diag}(\mathbf{W}\mathbf{1}), \mathbf{A}^\top = (\mathbf{A}_V^\top, \mathbf{A}_C^\top, \mathbf{A}_{vis}^\top), \mathbf{W}_{VX} = (\eta w_{ij}),$$

then the optimization problem of Eq. (1) can be written as

$$\max_{\mathbf{A}} \mathrm{Tr}\left(\mathbf{A}^\top\mathbf{X}^\top\mathbf{W}\mathbf{X}\mathbf{A}\right) \quad \text{subject to } \mathbf{A}^\top\mathbf{X}^\top\mathbf{M}\mathbf{X}\mathbf{A} = \mathbf{I}_K. \quad (2)$$

Similar to Eigenwords, we squash $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ and $\mathbf{X}^\top\mathbf{M}\mathbf{X}$ in Eq. (2) by replacing them with $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{G}}$ respectively, which are defined as follows.

$$\boldsymbol{\mathcal{H}} = \begin{pmatrix} \mathbf{O} & \boldsymbol{\mathcal{C}}_{VC} & \eta\boldsymbol{\mathcal{C}}_{VV}\widetilde{\mathbf{W}}_{VX}\mathbf{X}_{vis} \\ \boldsymbol{\mathcal{C}}_{VC}^\top & \mathbf{O} & \mathbf{O} \\ \eta\mathbf{X}_{vis}^\top\widetilde{\mathbf{W}}_{VX}^\top\boldsymbol{\mathcal{C}}_{VV} & \mathbf{O} & \mathbf{O} \end{pmatrix},$$
$$\boldsymbol{\mathcal{G}} = \begin{pmatrix} \boldsymbol{\mathcal{G}}_{VV} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\mathcal{C}}_{CC} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \boldsymbol{\mathcal{G}}_{vis} \end{pmatrix},$$

where $\mathrm{diag}(\boldsymbol{v})$ is a diagonal matrix aligning $\boldsymbol{v}$ as its diagonal elements, $\mathrm{sqrt}(\cdot)$ represents element-wise square root, the vectors $\boldsymbol{m}, \boldsymbol{n}$ are defined as $\boldsymbol{m} = \mathrm{sqrt}(\mathbf{V}^\top \mathbf{1}), \boldsymbol{n} = \eta \widetilde{\mathbf{W}}_{VX} \mathbf{1}$, $\circ$ represents element-wise product, and

$$\boldsymbol{\mathcal{G}}_{VV} = \boldsymbol{\mathcal{C}}_{VV} + \mathrm{diag}(\boldsymbol{m} \circ \boldsymbol{n}),$$
$$\boldsymbol{\mathcal{G}}_{vis} = \eta \mathbf{X}_{vis}^\top \mathrm{diag}(\widetilde{\mathbf{W}}_{VX}^\top \boldsymbol{m}) \mathbf{X}_{vis}.$$

Consequently, our final goal here is to find an optimal linear mapping which maximizes $\mathrm{Tr}(\mathbf{A}^\top \boldsymbol{\mathcal{H}} \mathbf{A})$ subject to $\mathbf{A}^\top \boldsymbol{\mathcal{G}} \mathbf{A} = \mathbf{I}_K$, and this problem reduces to a generalized eigenvalue problem $\boldsymbol{\mathcal{H}} \boldsymbol{a} = \lambda \boldsymbol{\mathcal{G}} \boldsymbol{a}$. Hence, we can obtain the optimal solution as $\hat{\mathbf{A}}^\top = (\hat{\mathbf{A}}_V^\top, \hat{\mathbf{A}}_C^\top, \hat{\mathbf{A}}_{vis}^\top) = \boldsymbol{\mathcal{G}}^{-1/2}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$, where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ are eigenvectors of $(\boldsymbol{\mathcal{G}}^{-1/2})^\top \boldsymbol{\mathcal{H}} \boldsymbol{\mathcal{G}}^{-1/2}$ for the K largest eigenvalues. Note that we obtain the word representations as the rows of $\hat{\mathbf{A}}_V$, as well as a linear mapping from the visual space to the common multimodal space $\hat{\mathbf{A}}_{vis}$, and that when visual data $\mathbf{X}_{vis}$ is omitted from the model, Eq. (2) is equivalent to CCA, namely, the ordinary Eigenwords. There are several ways to solve a generalized eigenvalue problem. In this study, we employed a randomized method for a generalized Hermitian eigenvalue problem proposed in Saibaba et al. (2016).

Silberer and Lapata (2012) also uses CCA to obtain multimodal representations, which associates term-document matrix representing word occurrences in documents and perceptual matrix containing scores on feature norms (or attributes) like "*is_brown*", "*has_fangs*", etc. This model is not considering any recent developments in word embedding. In addition, the feature norms are expensive to obtain, and hence we cannot expect them for a large number vocabularies. Besides, images relevant to a given word are more easy to collect.

5 Experiments

5.1 Dataset

In our experiment, we used English Wikipedia corpus (2016 dump)[1], which consists of approximately 3.9 billion tokens. We first used the script provided by Mahoney[2] to clean up the original dump. Afterward, we applied word2phrase (Mikolov et al., 2013) to the original corpus twice with a threshold value 500 to obtain multi-term phrases.

As for visual data, we downloaded images from the URLs in the NUS-WIDE image dataset (Chua et al., 2009), which also provides Flickr tags of each image. Although Flickr tags associated with each image could be very noisy and have varying abstractness, they provides a rich source of many-to-many relationships between images and words. Since we were interested in investigating if the large, but noisy web data would play a role as a helpful source for multimodal word representations, we omitted preprocessing like manually removing noisy tags or highly abstract tags.

The images were converted to 4096-dim feature vectors using the Caffe toolkit (Jia et al., 2014), together with a pre-trained[3] AlexNet model (Krizhevsky et al., 2012). These feature vectors are the output of the fc7 layer on the AlexNet. We randomly selected 100k images for a training set.

5.2 Word Similarity Task

We compared MM-Eigenwords against Eigenwords and skip-gram model through word similarity tasks, a common evaluation method of vector word representations. In our experiments, we used MEN (Bruni et al., 2014), SimLex (Hill et al., 2015), and another semantic similarity (Silberer and Lapata, 2014) denoted as SemSim, which provide 3000, 999, and 7576 word pairs respectively. These datasets provide manually scored word similarities, and the last one also provides visual similarity scores of word pairs denoted as VisSim. As for model-generated word vectors, the semantic similarity between two word vectors was measured by cosine similarity, and we quantitatively evaluated each embedding method by calculating Spearman correlation between model-based and human annotated scores.

5.3 Concept-to-Image Search

We also evaluated the accuracy of concept-to-image search to investigate the extent to which our multimodal word representations reflect visual information. In this experiment, we used 81 manually annotated concepts provided in NUS-WIDE dataset as queries. In addition, we randomly selected 10k images which are absent during the training phase as test-images and used $\hat{\mathbf{A}}_{vis}$ to

[1] https://dumps.wikimedia.org/enwiki/
[2] http://mattmahoney.net/dc/textdata.html
[3] https://github.com/BVLC/caffe/tree/master/models/

Method	η	Word Simliarity Task MEN	SimLex	SemSim	VisSim	Concept-to-Image Search P@1	P@5	P@10
Skip-gram		0.77	0.40	0.67	0.54			
Eigenwords		0.75	**0.45**	0.68	**0.58**			
MM-Eigewords	0.01	0.77	0.41	0.71	0.57	0.21	0.23	0.22
MM-Eigewords	0.1	**0.78**	0.38	**0.72**	0.57	0.14	0.14	0.14
MM-Eigewords	1	0.74	0.34	**0.72**	0.57	0.12	0.14	0.14
MM-Eigewords	10^4	0.66	0.21	0.37	0.34	0.44	0.39	0.37
MM-Eigewords	10^6	0.61	0.20	0.29	0.29	**0.53**	**0.47**	**0.49**

Table 1: Spearman correlations between word similarities based on the word vectors and that of the human annotations, and the right part shows the accuracies of concept-to-image search evaluated by precision@k.

project them to the textual space, on which top-match images were found by cosine similarities with the query vectors. We evaluated the accuracies of image search by precision at 1, 5, and 10, averaged over all query concepts, while varying the value of the multimodal term coefficient η in Eq. (1).

6 Results

For Eigenwords and MM-Eigenwords, we set the number of word types to $V \approx 140$k, including 30k most frequent vocabularies, words in the benchmarks, and Flickr tags associated with training-images, and we set the number of power iteration to 3. As for skip-gram model, we set the subsampling threshold to 10^{-5}, number of negative examples to 5, and training iterations to 5. In addition we fixed the dimensionality of word vectors to $K = 500$, and the context window size to $h = 4$ for every methods. As mentioned in Section 1, one of the most related methods is MMSkip-gram, against which we should compare MM-Eigenwords. However, since we could not find its code nor implement it by ourselves, a comparative study with MMSkip-gram is not included in this paper.

Table 1 shows the results of the word similarity tasks. As we can see in the table, with smaller η, the performance on word-similarity tasks of MM-Eigenwords is similar to that of Eigenwords or skip-gram model, whereas poor results on the concept-to-image search task. On the other hand, larger η helps improve the performance on the concept-to-image search while sacrificing the performances on the word similarity tasks. These results implies that too strongly associated visual information can distort the semantic structure obtained from textual data. Despite some similar existing studies showed positive results with auxiliary visual features (Lazaridou et al., 2015; Kiela and Bottou, 2014; Hill et al., 2014), our results achieved less improvements in the word-similarity tasks, indicating negative transfer of learning.

However, the visual informative word vectors obtained by our method enable not only word-to-word but also word-to-image search as shown in Figure 2a, and the many-to-many relationships between images and a wide variety of tags fed to our model contributed to the plausible retrieval results with the sum of two word vectors as their queries (e.g. "*bird*" + "*flying*" $\approx$ **images of flying birds**). Moreover, the word vectors learned with our model capture multimodal linguistic regularities (Kiros et al., 2014). We show some examples of our model in Figure 2b.

7 Conclusion

In this paper, we proposed a spectral graph-based method of multimodal word embedding. Our experimental results showed that MM-Eigenwords captures both semantic and text-to-image similarities, and we found that there is a trade-off between these two similarities.

Since the framework we used can be adopted to any number of views, we could further extend our method by considering image caption datasets through employing document IDs like Oshikiri et al. (2016) in our future works.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*. pages 2425–2433.

Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data

representation. *Neural computation* 15(6):1373–1396.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *JAIR* 49(2014):1–47.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*. ACM, page 48.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. 2015. Eigenwords: Spectral word embeddings. *JMLR* 16:3035–3078.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2):217–288.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *TACL* 2:285–296.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* pages 665–695.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Zhiwu Huang, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2012. Cross-view graph embedding. In *ACCV*. Springer, pages 770–781.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pages 675–678.

Jon R Kettenring. 1971. Canonical analysis of several sets of variables. *Biometrika* 58(3):433–451.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*. pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*. pages 835–841.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *CVPR*. pages 4985–4994.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. pages 1097–1105.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *HLT-NAACL*. pages 153–163.

Rémi Lebret and Ronan Collobert. 2014. Word embeddings through hellinger pca. In *EACL*. pages 482–490.

Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. 2016. Training and evaluating multimodal word embeddings with large-scale web annotated images. In *NIPS*. pages 442–450.

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

Nozomi Nori, Danushka Bollegala, and Hisashi Kashima. 2012. Multinomial relation prediction in social data: A dimension reduction approach. In *AAAI*. pages 115–121.

Takamasa Oshikiri, Kazuki Fukui, and Hidetoshi Shimodaira. 2016. Cross-lingual word representations via spectral graph embeddings. In *ACL*. pages 493–498.

Arvind K Saibaba, Jonghyun Lee, and Peter K Kitanidis. 2016. Randomized algorithms for generalized hermitian eigenvalue problems with application to computing karhunen–loève expansion. *Numerical Linear Algebra with Applications* 23(2):314–339.

Hidetoshi Shimodaira. 2016. Cross-validation of matching correlation analysis by resampling matching weights. *Neural Networks* 75:126–140.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *EMNLP-CoNLL*. pages 1423–1433.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*. pages 721–732.

Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *CVPR*. pages 3441–3450.

Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *TPAMI* 29(1):40–51.