CROSS-MODAL TRANSFER WITH NEURAL WORD VECTORS FOR IMAGE FEATURE LEARNING

Go Irie, Taichi Asami, Shuhei Tarashima, Takayuki Kurozumi, Tetsuya Kinebuchi

NTT Corporation

ABSTRACT

Neural word vector (NWV) such as `word2vec` is a powerful text representation tool that can encode extensive semantic information into compact vectors. This ability poses an interesting question in relation to image processing research – Can we learn better semantic image features from NWVs? We empirically explore this question in the context of semantic content-based image retrieval (CBIR). In this paper, we consider cross-modal transfer learning (CMT) to improve initial convolutional neural network (CNN) image features by using NWVs. We first show that NWVs can improve semantic CBIR performance compared to classical word vectors, even if it is with simple CMT models, i.e., canonical correlation analysis (CCA). Next, inspired by a characteristic property of NWVs, we propose a new CMT model and demonstrate that it can improve CBIR performance even further.

Index Terms— image retrieval, cross-modal transfer, word vector

1. INTRODUCTION

Media processing is in a new era with the recent revival of neural representation learning. Convolutional neural networks (CNNs) have been breaking record after record in the ImageNet competition series [1, 2, 3], and deep neural networks are now essential tools for speech recognition [4]. In the field of natural language processing, neural word vectors (NWVs) such as `word2vec` [5] have gained much attention recently. Our focus in this paper is the application of NWVs to semantic content-based image retrieval (CBIR) problems.

NWVs have been proven that they can capture semantically richer information of text data, compared with classical ones like singular value decomposition (SVD). In particular, NWVs have a surprising property called additive compositionality; basic algebraic operations on vectors can recover semantic relationships of words, e.g., vec("Berlin") − vec("Germany") + vec("France") ≈ vec("Paris") [5]. Driven by this observation, researchers have started exploring applications of NWVs to image processing problems such as image captioning [6]. However, its application to CBIR remains almost untouched, even though it is a central topic in the image processing field. Several studies propose to apply CNN features to CBIR tasks [7, 8], but application of NWVs has not yet been investigated very well, as far as we know. Rich semantic information carried by text data may be useful to guide learning of image features, and it is expected to be much more effective when the texts are represented by NWVs.

This motivates us to investigate leveraging NWVs for image feature learning for CBIR tasks. We specifically consider cross-modal transfer learning (CMT) [9, 10] in this paper; assuming that a set of images and their associated texts represented by NWVs are available *only* in the offline training stage (but not in the online test stage), we consider learning a new image feature by leveraging the image-text correlations. A few recent papers study similar CMT problems. However, most of them assume that there exist additional supervised resources such as a large collection of weakly labeled images [11] or object detectors [12], which may not always be available. Our focus is on a pure unsupervised scenario where images and texts are not associated with any supervised labels.

Our contributions of this paper can be summarized as follows. (i) We demonstrate that NWVs can train better image features for CBIR compared with SVD, even when simple CMT models such as canonical correlation analysis (CCA) or kernel CCA (KCCA) are used. (ii) We propose a new CMT model to improve CBIR performance even further. The new model is inspired by the additive compositionality of NWVs and is designed to discover a new image feature space that preserves the subspace structure of the NWV space. Experiments show that our model can improve CBIR performance.

2. PROTOCOL

We first describe the evaluation protocol used throughout this paper. The task of CBIR is to retrieve semantically relevant images to a given image query based on the image feature similarity. Note that text data is available only in the training stage, but not in the test (retrieval) stage. We use the following two popular datasets, each consisting of image-text pairs and semantic labels as ground truth.

Wiki [13] consists of $2,866$ image-text pairs on 10 topics. We follow the public training/test split which leads to $2,173$ training and 693 test samples.

MSCOCO [14] contains 120K images of 80 types of objects

Table 1. CBIR performance in terms of mAP of the initial image representations. Numbers with rightarrows are dimensions after PCA.

Wiki						
Dim.	Org.	$\rightarrow 256$	$\rightarrow 128$	$\rightarrow 64$	$\rightarrow 32$	$\rightarrow 16$
FV	0.156	0.156	0.156	0.157	0.157	0.158
AlexNet	0.175	0.194	0.196	0.198	0.203	0.201
VGG	0.197	0.200	0.204	0.208	0.217	0.223
MSCOCO						
Dim.	Org.	$\rightarrow 256$	$\rightarrow 128$	$\rightarrow 64$	$\rightarrow 32$	$\rightarrow 16$
FV	0.396	0.397	0.396	0.394	0.392	0.390
AlexNet	0.455	0.476	0.482	0.485	0.484	0.479
VGG	0.525	0.544	0.549	0.554	0.556	0.545

and associated short texts. We randomly extract 10K image-text pairs from the 80K "training" images for training and use 40K "validation" images for test.

As in [13], we use normalized correlation (cosine similarity) to measure the similarities between image features. Retrieval performance is evaluated by using mean average precision (mAP). In MSCOCO, since one or more of the labels are possibly assigned to one image, we judge the retrieval to be successful iff the retrieved image has at least one label in common with the query image.

Initial image features. We consistently use $4,096$-dimensional CNN features extracted from fc6 layer of the VGG network with 19-layers [15] pre-trained on the ILSVRC dataset. As shown in Table 1, we found that this VGG feature is higher in CBIR performance than other competitors that are commonly used, including AlexNet fc6 features [1] and 32K-dimensional Fisher Vectors (FV) [16]. We also found that performance is improved by applying a PCA dimensionality reduction to the VGG features. This may be because the distributions of the pre-trained features are adapted to the target dataset through PCA. We hereafter use the VGG feature reduced to 64-dimensions as our initial image representation.

Word vectors. Among various word vectors, we choose one popular classical method, SVD, and two major NWVs, i.e., skip-gram with negative sampling (SGNS) [5] and GloVe [17]. We train each type of word vector using 1M short texts in SBU1M dataset [18]. As for SVD, we construct a positive pointwise mutual information matrix and then obtain 300-dimensional word vectors by SVD. For SGNS and GloVe, we use the codes published by their authors and obtain 300-dimensional NWVs. To obtain a text feature, i.e., a feature vector for a text that consists of multiple words, we compute the mean vector of the words appearing in the text [5].

3. CROSS-MODAL TRANSFER WITH NEURAL WORD VECTORS

We first analyze if it is possible to improve the initial image feature (VGG) by using CMT with NWVs. In this section, we assume CCA and KCCA as the CMT models. These models are simple but still the state-of-the-art for unsupervised CMT [10, 11]. We evaluate CBIR performance when texts are represented by the three different types of word vectors.

3.1. CMT with CCA/KCCA

CCA. CCA learns a set of linear projections from each of image and text feature spaces to a common subspace such that the image-text data correlation is maximized. From the view point of CMT, CCA can be interpreted as a way to discover a new image feature (sub)space most correlated to a text feature (sub)space. Suppose we have n data pairs of images and texts. We denote the initial feature matrices of the images and texts by $X := [\mathbf{x}_1 \ \dots \ \mathbf{x}_n] \in \mathbb{R}^{D_x \times n}$ and $Y := [\mathbf{y}_1 \ \dots \ \mathbf{y}_n] \in \mathbb{R}^{D_y \times n}$, respectively, where D_x and D_y are the dimensions of image and text features. $\mathbf{x}_i$ and $\mathbf{y}_i$, $\forall i$, are assumed to be semantically relevant (e.g., an image and its caption). The goal is to find two sets of linear projections $A \in \mathbb{R}^{D_x \times d}$ and $B \in \mathbb{R}^{D_y \times d}$ to a common d-dimensional subspace such that $d \leq \min\{D_x, D_y\}$, for X and Y, respectively. This is achieved by solving the following problem.

$$\max_{A,B} \ \mathrm{trace}(A^\top X Y^\top B) \tag{1}$$

$$\text{s.t. } A^\top X X^\top A = I_d, \ \ B^\top Y Y^\top B = I_d, \tag{2}$$

where I_d is an identity matrix of size d. This can be transformed to a generalized eigenproblem that can easily be solved. Since our task is CBIR, we assume that texts are available only in the training stage, but not in the test stage; we keep only image-side projections A and discard B. The new image representation $\mathbf{z} \in \mathbb{R}^d$ can be obtained as $\mathbf{z} = A^\top \mathbf{x}$.

KCCA. CCA is kernelizable by computing the kernel Gramian on a subset of image and text features, which leads to KCCA.

$$\max_{A,B} \ \mathrm{trace}(A^\top K_x K_y^\top B) \tag{3}$$

$$\text{s.t. } A^\top K_x K_x^\top A = I_d, \ \ B^\top K_y K_y^\top B = I_d, \tag{4}$$

where K_x/K_y is an $s \times n$ kernel Gramian matrix (we assume Gaussian kernel in this paper due to its simplicity) for X/Y, and A/B is an $s \times d$ coefficient matrix (s is the number of samples used to compute K_x/K_y). This can also be transformed into a generalized eigenproblem thus can be solved.

3.2. Results

We evaluate the CBIR performance of the image features learned by CCA/KCCA-based CMT with the three types of word vector. The kernel parameters for KCCA (standard deviation and s) are tuned in a grid search manner. The results for various feature dimensions d are reported in Table 2. We found that CCA/KCCA improves the performance of the initial VGG feature, and the two NWVs (SGNS and GloVe) yield better performance compared with SVD, especially when KCCA is used.

Table 2. CBIR performance in terms of mAP of image features learned by CCA/KCCA-based CMT.

Wiki					
↓CMT Dim.→	16	24	32	48	64
None (init VGG)	0.223	0.218	0.217	0.211	0.208
CCA-SVD	0.250	0.247	0.244	0.242	0.241
CCA-SGNS	0.257	0.256	0.253	0.251	0.249
CCA-GloVe	0.255	0.254	0.252	0.250	0.248
KCCA-SVD	0.266	0.262	0.259	0.256	0.254
KCCA-SGNS	**0.271**	**0.276**	**0.271**	0.264	0.260
KCCA-GloVe	0.267	0.274	0.270	**0.265**	**0.261**
MSCOCO					
↓CMT Dim.→	16	24	32	48	64
None (init VGG)	0.545	0.552	0.556	0.553	0.553
CCA-SVD	0.571	0.579	0.581	0.583	0.582
CCA-SGNS	0.568	0.580	0.582	0.583	0.583
CCA-GloVe	0.571	0.579	0.582	0.583	0.583
KCCA-SVD	0.580	0.578	0.577	0.586	0.587
KCCA-SGNS	**0.588**	0.596	0.600	0.605	0.606
KCCA-GloVe	0.587	**0.597**	**0.601**	**0.606**	**0.607**

4. SUBSPACE PRESERVING TRANSFER FOR ADDITIVE COMPOSITIONALITY

So far we have confirmed that CMT with NWVs can improve semantic CBIR performance even with simple CCA/KCCA models. We next aim at improving CBIR performance even further by modifying CMT models. Our model is inspired by additive compositionality of NWVs, i.e., semantic relationships between words can be well approximated by linear algebraic operations on NWVs [5, 17]. This implies that words in the same semantic category are more likely to be distributed in low-dimensional linear subspace(s) in the NWV space [5], which leads us to assume that an NWV space consists of a union of (hidden) subspaces. On the basis of this, we propose a new CMT model named Subspace Preserving Transfer (SPT) that aims at retaining the inherent subspace structure of the NWV space in a common subspace.

4.1. Subspace Preserving Transfer

Again, let us assume that we have n pairs of image and text features, X and Y. Similar to the case of CCA/KCCA, the goal is to find two sets of projections A and B. SPT achieves this goal in two steps: i) discover an inherent NWV subspace structure, and ii) learn A and B which simultaneously preserve the discovered structure and image-text correlations.

i. NWV Subspace Discovery. What we want to do here is to discover the inherent subspace structure of NWVs, without any supervision. To this end, our idea is following. Generally, if we have a set of vectors lying in the same subspace, each vector can linearly and sparsely be reconstructed by a few other vectors [19]. This suggests that the subspace structure can approximately be captured by computing sparse linear reconstructions of the NWVs. However, since NWVs are usually trained on a separate corpus (SBU1M in this paper) independently from the text data (from Wiki or MSCOCO in this paper), the subspaces extracted from only NWVs may not well reflect the distributions of the text features, resulting in undesired results. We therefore compute the sparse linear reconstructions over both of the NWVs and the text features[1] so that can find the subspaces adapted to the text data.

Let us denote by $V := [\mathbf{v}_1 \ \dots \ \mathbf{v}_m] \in \mathbb{R}^{D_y \times m}$ a set of m NWVs for m words. For brevity, we define $\tilde{Y} := [Y \ \ V] \in \mathbb{R}^{D_y \times (n+m)}$, i.e., $\tilde{Y} := [\tilde{\mathbf{y}}_1 \ \dots \ \tilde{\mathbf{y}}_l] := [\mathbf{y}_1 \ \dots \ \mathbf{y}_n \ \mathbf{v}_1 \ \dots \ \mathbf{v}_m]$ $(l = n + m)$. In order to discover the subspace structure in $\tilde{Y}$, we solve the following sparse linear reconstruction problem [19].

$$\min_W \|W\|_1 \quad \text{s.t. } \tilde{Y} = \tilde{Y}W, \quad \text{diag}(W) = \mathbf{0}. \tag{5}$$

where the i-th column $\mathbf{w}_i$ of the solution $W = [\mathbf{w}_1 \ \dots \ \mathbf{w}_l] \in \mathbb{R}^{l \times l}$ consists of the sparse coefficients to recover $\tilde{\mathbf{y}}_i$. On the basis of the above observation, $\mathbf{w}_i$, $\forall i$, is expected to have sparse non-zero elements that correspond to other features in the same subspace as $\tilde{\mathbf{y}}_i$. Hence, the desired subspace structure can be captured in W.

ii. Projection Learning. Next, we compute the projections A and B that simultaneously maximize the correlation between the image and text features and preserve the NWV subspace structure captured in W as much as possible. This problem can be formulated as follows.

$$\max_{A,B} \ \text{trace}(A^\top XY^\top B) - \lambda \sum_{i=1}^{l} \|B^\top \tilde{\mathbf{y}}_i - \sum_{j \neq i} w_{ij} B^\top \tilde{\mathbf{y}}_j\|_2^2 \tag{6}$$

$$\text{s.t. } A^\top XX^\top A = I_d, \ \ B^\top \tilde{Y}\tilde{Y}^\top B = I_d \tag{7}$$

The first term is the same as the objective of CCA (Eq. (1)) which is to maximize the correlation between the image and text features in the common subspace. The second term is to ensure that W is retained in the common subspace, i.e., it ensures a set of features in the same NWV subspace are in a low-dimensional subspace in the common subspace. The second term can be rewritten in matrix form as

$$\textstyle\sum_{i=1}^{l} \|B^\top \tilde{\mathbf{y}}_i - \sum_{j \neq i} w_{ij} B^\top \tilde{\mathbf{y}}_j\|_2^2 \tag{8}$$

$$= \text{trace}(B^\top \tilde{Y} L \tilde{Y}^\top B). \tag{9}$$

where $L = (I_l - W)^\top (I_l - W)$. Then Eq. (6) becomes

$$\max_{A,B} \ \text{trace}(A^\top XY^\top B) - \lambda \text{trace}(B^\top \tilde{Y} L \tilde{Y}^\top B) \tag{10}$$

$$\text{s.t. } A^\top XX^\top A = I_d, \ \ B^\top \tilde{Y}\tilde{Y}^\top B = I_d. \tag{11}$$

[1] A text feature is originally computed as a mean vector of the words in the text (see Sec. 2), but note that the sparse reconstruction by using the whole NWVs and the text features generally does not recover it and is preferably determined to be optimized to the data.

Table 3. CBIR performance in terms of mAP of image features learned by SPT/KSPT-based CMT.

Wiki					
↓CMT Dim.→	16	24	32	48	64
None (init VGG)	0.223	0.218	0.217	0.211	0.208
SPT-SVD	0.251	0.249	0.246	0.244	0.243
SPT-SGNS	0.270	0.271	0.271	0.271	0.271
SPT-GloVe	0.271	0.272	0.272	0.272	0.272
KSPT-SVD	0.256	0.249	0.247	0.244	0.243
KSPT-SGNS	0.301	**0.302**	**0.302**	**0.302**	**0.302**
KSPT-GloVe	**0.302**	**0.302**	**0.302**	**0.302**	**0.302**
MSCOCO					
↓CMT Dim.→	16	24	32	48	64
None (init VGG)	0.545	0.552	0.556	0.553	0.553
SPT-SVD	0.570	0.579	0.582	0.583	0.583
SPT-SGNS	0.563	0.574	0.581	0.582	0.583
SPT-GloVe	0.566	0.580	0.582	0.584	0.584
KSPT-SVD	0.592	0.597	0.596	0.600	0.599
KSPT-SGNS	0.590	0.602	0.607	0.610	0.611
KSPT-GloVe	**0.600**	**0.611**	**0.614**	**0.617**	**0.618**

which gives the problem of SPT. The objective (Eq. (10)) can further be aggregated into a single term. Let us introduce the following matrices.

$$P = \begin{bmatrix} A \\ B \end{bmatrix}, \quad Q = \begin{bmatrix} X & 0 \\ 0 & \tilde{Y} \end{bmatrix}, \quad E = \begin{bmatrix} 0 & \frac{1}{2} I_{n/l} \\ \frac{1}{2} I_{n/l}^\top & -\lambda L \end{bmatrix}, \quad (12)$$

where $I_{n/l}$ is an $n \times l$ matrix defined as $I_{n/l} = [I_n \; 0]$. Substituting these matrices into Eq (10), we get

$$\text{trace}(A^\top X Y^\top B) - \lambda \text{trace}(B^\top \tilde{Y} L \tilde{Y}^\top B) \quad (13)$$

$$= \text{trace}(P^\top Q E Q^\top P) \quad (14)$$

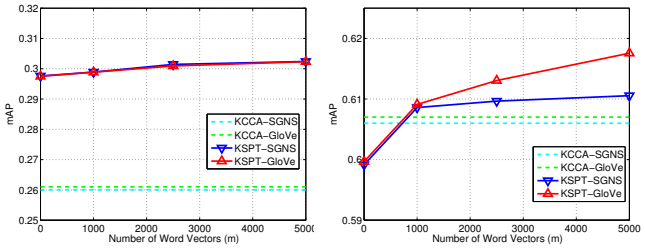
Hence, similar to CCA, this problem can be transformed into a generalized eigenproblem, and the solution P consists of the eigenvectors corresponding to the d largest eigenvalues. The optimal A can be obtained by extracting the first D_x rows of P. A new feature vector $\mathbf{z}$ for the initial image feature vector $\mathbf{x}$ can be computed as $\mathbf{z} = A^\top \mathbf{x}$.

Kernelization. Similar to KCCA, the kernelized version of SPT, i.e., KSPT, can be derived by computing a kernel Gramian over a subset of image and text features. One difference from KCCA is that only the image side is kernelized. This is to ensure that the extracted NWV subspace structure is not corrupted by any non-linearity induced by the kernelization. More specifically, on the basis of Eq. (10), the problem of KSPT can be formulated as

$$\max_{A,B} \; \text{trace}(A^\top K_x Y^\top B) - \lambda \text{trace}(B^\top \tilde{Y} L \tilde{Y}^\top B) \quad (15)$$

$$\text{s.t. } A^\top K_x K_x^\top A = I_d, \;\; B^\top \tilde{Y} \tilde{Y}^\top B = I_d. \quad (16)$$

where K_x is an $s \times n$ kernel Gramian matrix for X and A is an $s \times d$ projection matrix. Similar to the case of SPT, this problem can also be transformed into a generalized eigenproblem so thus can be solved efficiently.



(a) Wiki (b) MSCOCO

Fig. 1. CBIR performance of KSPT in terms of mAP when varying the number of word vectors for training m.

4.2. Results

We evaluate the semantic CBIR performance when the proposed SPT/KSPT is used as the CMT model. The number of NWVs used to discover the subspace structure, m, is set at $m = 5,000$ and the other parameters are tuned in a grid search manner. The results are reported in Table 3. We found that SPT and KSPT improve the initial VGG features. Comparing these results with those by CCA and KCCA (Table 2), we can see that SPT/KSPT is highly competitive with or better than CCA/KCCA. These results confirm the effectiveness of the proposed SPT/KSPT and show that they can improve CBIR performance even further. Especially, KSPT-SGNS and KSPT-GloVe are consistently better than KCCA counterparts and yield significant gains, e.g., KSPT-GloVe exceeds KCCA-GloVe by 4.1% for 64-dimensional features on Wiki. This may be because the nonlinearity induced by kernelization enables KSPT to better transfer the subspace structure to the common subspace while maintaining image-text correlations. Upon seeing the differences between the three types of word vector, SGNS and GloVe outperform SVD. This suggests that the structure of NWV subspaces may tend to be more correlated to the semantics, which can effectively be captured by using SPT or KSPT.

We also analyze the impact of varying the number of NWVs m on performance of KSPT. The results are reported in Fig. 1. The mAP values improve as m increases and exceed those of KCCA even for a small m ($m \geq 1,000$). This confirms the effectiveness of capturing the subspace structure of the NWVs for image feature learning.

5. CONCLUSIONS

We investigated CMT with NWVs for image feature learning in CBIR problems. We first showed that NWVs could improve image features when simple CMT models are used. We then proposed a new CMT model that explicitly takes into account the additive compositionality of NWVs and demonstrated that it could improve CBIR performance even further. An interesting future work may be to couple the proposed method with other word/sentence embedding techniques such as Fisher vector encoding of word vectors [20] and recurrent neural networks [21, 22].

6. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *TASLP*, vol. 20, no. 1, pp. 30–41, 2012.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[7] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *ECCV*, 2014.

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014, pp. 392–407.

[9] J. Costa Pereira and N. Vasconcelos, "Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems," *CVIU*, vol. 124, pp. 123–135, 2014.

[10] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *IJCV*, vol. 106, no. 2, pp. 210–233, 2014.

[11] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *ECCV*, 2014, pp. 529–545.

[12] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *NIPS*, 2014, pp. 1889–1897.

[13] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010.

[14] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv preprint arXiv:1409.1556*, 2014.

[16] F. Perronnin, J. Sánches, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.

[17] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[18] V. Ordonez, G. Kulkarni, and T. L. Berg., "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011.

[19] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *PAMI*, vol. 35, no. 11, pp. 2765–2781, 2013.

[20] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015, pp. 4437–4446.

[21] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, and S. Fidler:, "Skip-thought vectors," in *NIPS*, 2015, pp. 3294–3302.

[22] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *ACL*, 2015, pp. 1556–1566.