Joint Image-Text Representation by Gaussian Visual-Semantic Embedding

Zhou Ren[1]* Hailin Jin[2] Zhe Lin[2] Chen Fang[2] Alan Yuille[1,3]
[1]University of California, Los Angeles [2]Adobe Research, San Jose, CA
[3]Departments of Cognitive Science and Computer Science, Johns Hopkins University, Baltimore, MD
zhou.ren@cs.ucla.edu {hljin,zlin,cfang}@adobe.com alan.yuille@jhu.edu

ABSTRACT

How to jointly represent images and texts is important for tasks involving both modalities. Visual-semantic embedding models have been recently proposed and shown to be effective. The key idea is that by learning a mapping from images into a semantic text space, the algorithm is able to learn a compact and effective joint representation. However, existing approaches simply map each text concept to a single point in the semantic space. Mapping instead to a *density distribution* provides many interesting advantages, including better capturing uncertainty about each text concept, and enabling better geometric interpretation of concepts such as inclusion, intersection, *etc.* In this work, we present a novel Gaussian Visual-Semantic Embedding (GVSE) model, which leverages the visual information to model text concepts as Gaussian distributions in semantic space. Experiments in two tasks, image classification and text-based image retrieval on the large scale MIT Places205 dataset, have demonstrated the superiority of our method over existing approaches, with higher accuracy and better robustness.

Keywords

Visual-semantic embedding; gaussian embedding; image classification; text-based image retrieval

1. INTRODUCTION

Joint image-text representation is essential for tasks involving both images and texts, such as image captioning [8, 17], text-based image retrieval [7, 14], image classification [1, 3, 13], *etc.* In recent years, Visual-Semantic Embedding (VSE) models [2, 6, 10, 12] have shown impressive performance in those tasks. By leveraging the semantic information contained in unannotated text data, VSE models explicitly map images into a rich semantic space, with the goal that images of the same category are mapped to nearby locations around the corresponding text label, and text labels are embedded in such a smooth space with respect to some measure of similarity, that is, similar text concepts should be mapped into nearby locations.

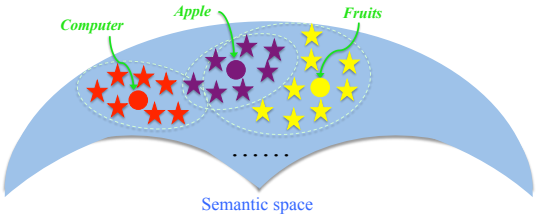

Figure 1: Existing Visual-Semantic Embedding models simply represent the text concepts as single points in the semantic space, as indicated by the circular symbols (we use different colors to indicate different categories). We propose the Gaussian Visual-Semantic Embedding (GVSE) model, which however leverages the visual information indicated as the pentagram symbols (pentagram symbols of the same color indicate various images of the same category) to model text concepts as Gaussian distributions. Modeling text concepts as densities innately represents uncertainties of text concepts, and has the benefit of geometric interpretation such as inclusion and intersection.

Although VSE models have shown remarkable results, representing text concepts as single points in semantic space carries some important limitations. Firstly, using an embedded vector to represent a text does not naturally express uncertainty about its concept with which the corresponding images may be associated. Even though various images may belong to the same text category, uncertainty is associated with them, which may reflect different aspects of a certain text concept. For instance, as shown in Fig.1, the *Computer* images may be from different brands, viewpoints, *etc.* Moreover, it is intrinsically problematic to map all images of a certain text label to a single point in semantic space, which would confuse the embedding function and thus harm the joint representation.

This paper advocates moving beyond modeling text concepts as single points to that as *densities* in semantic space. In particular, we explore Gaussian distributions (currently with diagonal covariance), in which the means are learned from unannotated text data online and variances are learned from the visual image data. Gaussians innately embody un-

*Part of this work was done when the author was an intern at Adobe Research.

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2967212

certainty, and have a geometric interpretation, such as an inclusion or intersection relationships between text concepts, as shown in Fig.1. We name the proposed method Gaussian Visual-Semantic Embedding (GVSE) model.

To evaluate the proposed method, we have conducted experiments in two tasks on the large scale MIT Places205 dataset. In the image classification task, our model outperforms the VSE baseline [2] by 1-5%. In the task of text-based image retrieval, we illustrate the robustness of our method and the capability in generalizing to untrained texts.

1.1 Related Work

Multi-modal embedding models utilize information from multiple sources, such as images and texts. By leveraging the abundant textual data available on the Internet, several lexically distributed representations of texts have been proposed to capture the semantic meaning among texts, *e.g.*, the word2vec model [9] and GloVe model [11]. Image-sentence embedding models [5, 7] were proposed for image captioning. Recently, visual-semantic embedding models for image classification were proposed by leveraging the distributed representation of labels, *e.g.*, Frome *et al.* [2] and Norouzi *et al.* [10]. Ren *et al.* [12] proposed multiple instance visual-semantic embedding for multi-label image annotation. Vilnis *et al.* [16] presented Gaussian embedding model for word representation. The main difference between [16] and our model is that [16] is only for word representation learned from word data alone, while our model is for joint image-text representation.

2. OVERVIEW OF VISUAL-SEMANTIC EMBEDDING

We first review the background on visual-semantic embedding. Given an image dataset $\mathcal{D} \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, each image is represented by a d-dimensional feature vector, $\mathbf{x}_i \in \mathcal{X} \overset{def}{=} \mathbb{R}^d$, and each image is associated with a text label, y_i. There are totally n distinct text labels in dataset $\mathcal{D}$, *i.e.*, $y_i \in \mathcal{Y} \equiv \{1, 2, ..., n\}$.

Previous methods [1, 3] associate images and texts as a classification problem, which predefine a fixed set of text labels $\mathcal{Y}$, and learn to predict the labels given image input, *i.e.*, $\mathcal{X} \rightarrow \mathcal{Y}$. However, these classification-based approaches do not provide a joint representation, thus they have limited applications. Moreover, they lack the ability of generalizing to unseen labels, and need to be retrained when a new label emerges. For example, given a training dataset $\mathcal{D}$ as above, and a test dataset $\mathcal{D}' \equiv \{(\mathbf{x}'_j, y'_j)\}_{j=1}^{N'}$ where $\mathbf{x}'_j \in \mathcal{X}$ and all test labels are distinct from the training labels in dataset $\mathcal{D}$, *i.e.*, $y'_j \in \mathcal{Y}' \equiv \{n+1, ..., n+n'\}$. The test labels are untrained as $\mathcal{Y} \cap \mathcal{Y}' = \emptyset$. Clearly, without side information about the relationships between labels in $\mathcal{Y}$ and $\mathcal{Y}'$, it is infeasible to generalize a classification-based method to unseen text labels without retraining it.

Fortunately, visual-semantic embedding models [2, 10, 12] have been proposed to address this issue. Instead of learning a mapping from images to labels ($\mathcal{X} \rightarrow \mathcal{Y}$), it aims to construct a continuous semantic space $\mathcal{S} \equiv \mathbb{R}^e$ which captures the semantic relationship among all text labels in $\mathcal{Y} \cup \mathcal{Y}'$, and explicitly learn the embedding function from images to such space, $f : \mathcal{X} \rightarrow \mathcal{S}$. The semantic space $\mathcal{S}$ is constructed such that two labels y and y' are semantically similar if and only if their semantic embeddings $s(y)$ and $s(y')$ are close in $\mathcal{S}$, where $s(y)$ is the semantic embedding vector of label y in $\mathcal{S}$. Thus the trained and unseen test labels become related via the semantic space $\mathcal{S}$. Once $f(\cdot)$ is learned, it can be applied to a test image $\mathbf{x}'$ to obtain $f(\mathbf{x}')$, and this image embedding vector of $\mathbf{x}'$ is then compared with the unseen label embedding vectors, $\{s(y'); y' \in \mathcal{Y}'\}$, to search for the most relevant test labels. This allows us to generalize the visual-semantic embedding models to unseen labels.

There are two key components of visual-semantic embedding models: One is how to learn the embedding function $f(\cdot)$. Various approaches [2, 10, 12] have validated the effectiveness of using ranking distance to learn $f(\cdot)$. The other key component is how to construct the continuous semantic space $\mathcal{S}$ of text labels.

3. GAUSSIAN VISUAL-SEMANTIC EMBEDDING

3.1 Constructing the semantic text space

We firstly introduce how we construct $\mathcal{S}$. Distributed representations [9, 11] has shown the capacity to provide semantically meaningful embedding features for text terms (including words and phrases), by learning from unannotated text data from the Internet. This method is able to learn similar embedding vectors for semantically related words because of the fact that those words are more likely to appear in similar semantic contexts.

Thus, we utilize the GloVe model [11] to construct a 300-dim text semantic space $\mathcal{S}$ which embodies the semantic relationship among labels.

3.2 Modeling text concepts as Gaussian distributions

Motivated by the success of ranking loss in state-of-the-art visual-semantic embedding [2, 10, 12], we employ ranking loss to learn the embedding function $f : \mathcal{X} \rightarrow \mathcal{S}$. The intuition is to encourage the embedding of an image to be closer to its ground truth text label than other negative labels:

$$L_{\text{GVSE}}(\mathbf{x}_i, y_i) = \sum_{y_n \in \mathbf{y}_i^-} \max(0, m + D\left(f(\mathbf{x}_i), y_i\right) - D\left(f(\mathbf{x}_i), y_n\right)), \tag{1}$$

where m is the ranking loss margin that we cross-validate, $f(\mathbf{x})$ is the embedding vector of image $\mathbf{x}$ in $\mathcal{S}$, $\mathbf{y}_i^-$ denotes the negative labels excluding the ground truth label y_i, *i.e.*, $\mathbf{y}_i^- = \mathcal{Y}/y_i$, and $D(f(\mathbf{x}), y)$ is the distance measure between an image embedding point and a text concept. We will introduce how we compute $D(f(\mathbf{x}), y)$ later.

Existing visual-semantic embedding methods model texts as single points in the semantic space $\mathcal{S}$, thus $D(f(\mathbf{x}), y)$ in those methods is just the Euclidean distance between $f(\mathbf{x})$ and the text embedding vector $s(y)$. However, as claimed in the introduction, they are limited in representation capacity. Intrinsically, it is beneficial to model text concepts as densities, which can embody the uncertainty of concepts and also have better geometric interpretation.

In this paper, we model text concepts as Gaussian distributions with diagonal covariances, *i.e.*, $y_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$. Thus, the distance between an image embedding vector $f(\mathbf{x}_i)$ and a text label y_i can be measured by Mahalanobis distance

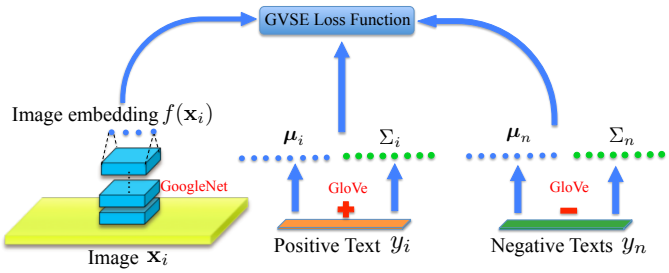

Figure 2: Training framework of the proposed Gaussian Visual-Semantic Embedding model.

as follows:

$$D(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (f(\mathbf{x}_i) - \boldsymbol{\mu}_i). \quad (2)$$

Thus, our GVSE model could be effectively learned by the loss function in Eq.1.

3.3 Training and inference with GVSE

The learning framework is shown in Fig.2. We jointly learn the embedding function $f(\cdot)$ and the text density parameters $\{\boldsymbol{\mu}_i\}_{i=1}^n$, $\{\Sigma_i\}_{i=1}^n$ in two steps.

Firstly, the mean text embedding vectors $\{\boldsymbol{\mu}_i\}$ are learned using the unannotated text data online alone, as introduced in Section 3.1. Thanks to the similarity between text concepts captured in semantic space $\mathcal{S}$, our model is able to generalize to unseen text concepts, which will be illustrated in Section 4.3. Secondly, we learn $f(\cdot)$ and $\{\Sigma_i\}$ by end-to-end training. Note that we use GoogleNet [15] to extract image features $\mathbf{x}_i$. On top of the convolutional layers of GoogleNet, we add a fully connected layer to model the embedding function $f(\cdot)$. On the other hand, since Σ_i is a 300×300 diagonal matrix, we use a $300\times n$ fully connected layer to encode all parameters in $\{\Sigma_i\}_{i=1}^n$, where each column of the weight corresponds to a covariance matrix Σ_i.

Given a trained GVSE model, it is straightforward to do inference on either a query image or a query text, depending on the tasks. From either direction, we map the query and testing entries into the semantic space $\mathcal{S}$. And Mahalanobis distance between the query and each testing entry can be computed by Eq.2. Finally the result is computed based on such distances.

4. EXPERIMENTS

In this section, we report our experiments on image classification and text-based image retrieval, comparing the proposed GVSE model with visual-semantic embedding models.

4.1 Dataset and implementation

We test on a large-scale image dataset, MIT Places205 dataset [18], which has 2,448,873 images from 205 scene categories. We follow the train-test split provided in the dataset.

We use Caffe [4] to implement our model. The optimization of our network is achieved by Stochastic Gradient Descent with a momentum term of weight 0.9 and with mini-batch size of 100. The initial learning rate is set to 0.1, and we update it with the "steps" policy. A weight decay of 0.0005 is applied.

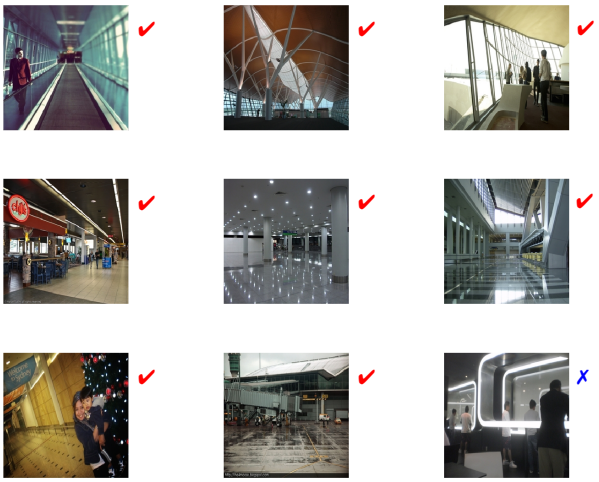

Figure 3: The top-9 image retrieval results of searching trained text "*airport terminal*". The incorrect image shares very similar visual appearance.

Figure 4: The top-9 image retrieval results of searching trained text "*kitchen*". The incorrect images belong to "*kitchenette*" which is a very close concept.

4.2 Image classification

To quantify the image classification performance, we use mAP@k as the evaluation metric, which measures the mean average precision of returning the ground truth label within top-k of the prediction list.

The results are shown in Table 1. We compare with two baseline models: one is visual-semantic embedding (VSE) model trained with L2 loss, the other is the DeViSE [2] model trained with ranking loss. These two models are state-of-the-art visual-semantic embedding models. As we see, with the basic nearest neighbor search in testing, the proposed GVSE model outperforms the best baseline for 0.93% on average, which validates the benefit of modeling text concepts as densities.

Moreover, we train a SVM classifier on top of the embedding representation that GVSE learned, and such classifying technique boosts the performance for 4.48% in mAP@1

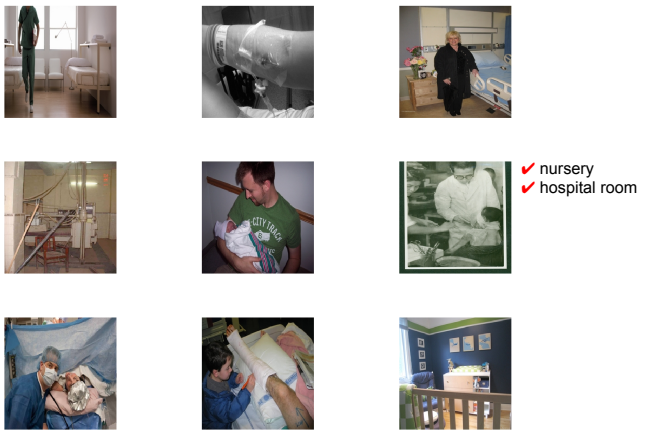

Figure 5: The top-9 retrieval results of searching untrained text "*infant*". The returned images belong to "*nursery*" and "*hospital room*", which are conceptually related.

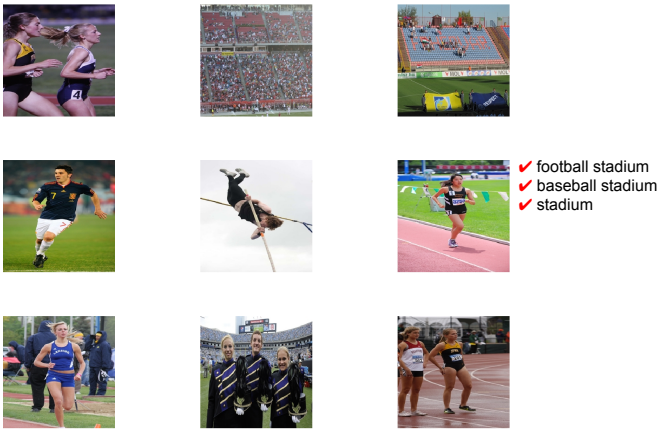

Figure 6: The top-9 retrieval results of searching untrained text "*sports*". The returned images belong to "*football stadium*", "*baseball stadium*" and "*stadium*", which are conceptually related.

Approach	mAP@1	mAP@5
1. nearest neighbor search		
VSE with L2 loss	41.33	68.02
DeViSE [2]	51.53	82.59
The proposed GVSE	52.38	83.60
2. SVM search		
The proposed GVSE	**56.86**	**84.46**

Table 1: Image classification results on the MIT Places205 dataset, shown in %.

and 0.86% in mAP@5. State-of-the-art classification-based method [18] (that uses deep learning and does not use explicitly modeling) reported mAP@1 performance 55.50%. Our performance is 1.36% higher. Note that classification-based methods [18] do not learn a joint representation and are not able to generalize to untrained classes, which is an advantage of our method, as shown in Fig.5, Fig.6 of Section 4.3.

4.3 Text-based image retrieval

Since GVSE model learns a joint embedding representation for both images and texts. We could either search labels given images (as in image classification of Section 4.2), or search images given text query. Thus we conduct a qualitative experiment on text-based image retrieval.

Thanks to the semantic space learned from unannotated text data, our model is able to search both trained or untrained texts, as discussed in Section 2 (searching by untrained texts is known as *zero-shot learning* [12], which is a challenging task.). Fig.3 and Fig.4 illustrate two examples of trained text searching, while Fig.5 and Fig.6 show two examples of untrained text searching.

As we see in Fig.3 and Fig.4, the retrieval results are very robust, except a few incorrect cases, which either share similar visual appearance with the query images or belong to very close text concept of the query. On the other hand, when we search untrained texts as shown in Fig.5 and Fig.6, the returned results are conceptually very related to the queries. For instance, "*sports*" is a text label that does not appear in the MIT Places205 training dataset. However, using the proposed GVSE model, the retrieved "*stadium*"s images are very related to "*sports*".

Both quantitative and qualitative experimental results validate the superiority of the proposed GVSE model in joint image-text representation.

4.4 Future work

From the experimental results above, we have validated the superiority of modeling text concepts as densities over single points. We explored our preliminary idea on Gaussian modeling. And for effective training, we constrained the densities to be diagonal Gaussians.

A straightforward improvement of our method is to model text concepts as more sophisticated density distributions, such as Gaussians with arbitrary covariances and Gaussian Mixture Models. Another future improvement over the proposed method is in the training process. We learned our model parameters by end-to-end training. However, with more complicated text modelings, such as GMM, iterative training can benefit the convergence.

5. CONCLUSION

In this paper, we have proposed a novel Gaussian Visual-Semantic Embedding model for joint image-text representation. Instead of modeling text concepts as single points in the semantic space, we move beyond to modeling that as densities. Effective end-to-end training framework and testing techniques have been introduced. Experiments on multimodal tasks in both directions of images and texts, including image classification and text-based image retrieval, have demonstrated that the proposed method outperforms existing visual-semantic embedding models with higher accuracy, better robustness, as well as the ability to generalize to untrained text concepts.

Acknowledgement

This work is partially supported by the Army Research Office ARO 62250-CS and a gift grant from Adobe Research. We also acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

6. REFERENCES

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierachical image database. In *CVPR*, 2009.

[2] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[3] Y. Gong, Y. Jia, T. K. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. In *ICLR*, 2014.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

[6] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *TACL*, 2015.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *TACL*, 2015.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015.

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

[10] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

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

[12] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Multi-instance visual-semantic embedding. In *arXiv preprint arXiv:1512.06963*, 2015.

[13] Z. Ren, C. Wang, and A. Yuille. Scene-domain active part models for object representation. In *ICCV*, 2015.

[14] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 1999.

[15] 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.

[16] L. Vilnis and A. McCallum. Word representations via gaussian embedding. In *ICLR*, 2015.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.