Zero-Shot Sketch-Based Image Retrieval via Graph Convolution Network

Zhaolong Zhang,[1] Yuejie Zhang,[1] Rui Feng,[1] Tao Zhang,[2] Weiguo Fan[3]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, P.R. China 200433
{18210240044, yjzhang, fengrui}@fudan.edu.cn
[2]School of Information Management and Engineering, Shanghai Key Laboratory of Financial Information Technology,
Shanghai University of Finance and Economics, Shanghai, P.R. China 200433
taozhang@mail.shufe.edu.cn
[3]Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa, USA, 52242
weiguo-fan@uiowa.edu

Abstract

Zero-Shot Sketch-based Image Retrieval (ZS-SBIR) has been proposed recently, putting the traditional Sketch-based Image Retrieval (SBIR) under the setting of zero-shot learning. Dealing with both the challenges in SBIR and zero-shot learning makes it become a more difficult task. Previous works mainly focus on utilizing one kind of information, i.e., the visual information or the semantic information. In this paper, we propose a SketchGCN model utilizing the graph convolution network, which simultaneously considers both the visual information and the semantic information. Thus, our model can effectively narrow the domain gap and transfer the knowledge. Furthermore, we generate the semantic information from the visual information using a Conditional Variational Autoencoder rather than only map them back from the visual space to the semantic space, which enhances the generalization ability of our model. Besides, feature loss, classification loss, and semantic loss are introduced to optimize our proposed SketchGCN model. Our model gets a good performance on the challenging *Sketchy* and *TU-Berlin* datasets.

Introduction

Sketch-based Image Retrieval (SBIR), which aims at retrieving the relevant images by using the free-hand sketches, has been studied for several years (Yu et al. 2016; Sangkloy et al. 2016; Song et al. 2017; Huang et al. 2018). Compared with the traditional text-image cross-modal retrieval, it may become the first choice for users when it is hard to provide a textual description but easy to sketch what they want. The difficulty for SBIR lies in the huge differences between sketches and images, which is widely known as the "domain gap", for sketches are often abstract and lose the texture information and they come from different modalities. In recent years, many researchers introduce deep neural networks into this field (Yu et al. 2016; Song et al. 2017). With the help of their strong representation ability, researchers can narrow the domain gap significantly and achieve good performance.

However, the explosive growth of the amount of multimedia content on the internet makes it very possible that the category of a wanted image does not appear in the training set in real life. The conventional SBIR methods can hardly handle this situation. Thus, researchers put the SIBR task under the condition of zero-shot learning (Yelamarthi et al. 2018) where the training and testing categories are disjoint, and propose the Zero-Shot Sketch-based Image Retrieval (ZS-SBIR) task. Compared with the traditional SBIR, ZS-SBIR is more challenging for the model not only needs to deal with the visual difference between sketches and images but also needs to establish the relations between the seen categories and the unseen categories. Therefore, the main challenges in ZS-SBIR can be summarized as narrowing the domain gap between the different modalities and transferring the knowledge from the seen categories to the unseen categories.

To handle these challenges, prior works mainly used generative models, Conditional Variational Autoencoders or Generative Adversarial Networks, to generate the image features corresponding to the given sketch features (Kumar Verma et al. 2019; Dutta and Akata 2019). Although the generative methods outperform the conventional SBIR methods, they still suffer from several problems. The crux of the matter for zero-shot learning is to transform the knowledge from the seen categories to the unseen ones through the side information which indicates their relations. However, generating the possible image features to the given sketch features does not make effective use of the side information. Furthermore, due to its instability during the training phase, it is hard to get the best performance for a generative model.

In this paper, we propose a SketchGCN model to alleviate the above shortcomings. Our SketchGCN model contains three sub-networks, i.e., an encoding network, a semantic preserving network, and a semantic reconstruction network. The encoding network tries to embed the sketches and images into a common semantic space, while the semantic preserving network takes the features as input and utilizes the side information to force them to maintain their category-level relations. The use of side information is essential for zero-shot tasks. With the help of graph convolution network, we can effectively use the side information to build and handle the relations among the features. The semantic reconstruction network further forces the extracted features to preserve their semantic relations. Different from the previous methods which adopt a Multilayer Perceptron to reconstruct

the semantic information, we introduce a Conditional Variational Autoencoder to generate the semantic information from the extracted features, which can enhance the generalization ability of our model. With the help of these components, our model produces good retrieval performance on two ZS-SBIR datasets, *Sketchy-Extended* (Sangkloy et al. 2016) and *TU-Berlin-Extended* (Eitz, Hays, and Alexa 2012).

The main contributions of this work are as follows: 1) We propose the SketchGCN model for the zero-shot SBIR task, using a graph convolution network and a Conditional Variational Autoencoder. With the help of these parts, our model can transfer the information from the seen categories to the unseen categories effectively. 2) Instead of only taking semantic information into account, our graph convolution model with a learnable adjacency matrix considers both visual and semantic information to solve the challenges in ZS-SBIR. 3) A Conditional Variational Autoencoder is implemented to enhance the generalization ability of our proposed model by generating the semantic embedding from the visual features. The SketchGCN model successfully produces good retrieval performance under two widely used datasets (i.e., *Sketchy-Extended* and *TU-Berlin-Extended*), which shows the effectiveness of our model.

Related Work

Sketch-based Image Retrieval. Sketch-based Image Retrieval (SBIR) is aimed at answering how similar the given sketches and the candidate images are, and the core solution is embedding the sketches and the images into a common feature space to narrow the domain gap between them. In the early years of this task, the features are extracted from the edge maps of natural images and the sketches by well-designed descriptors. Some hand-crafted descriptors like SHOG (Eitz et al. 2010), Gradient Field HOG (Hu and Collomosse 2013), etc. are proposed successfully at that time. However, because the dramatic differences between sketches and images and extracting edge maps from images may bring noises, this task still needs further study.

In recent years, the deep neural network has developed rapidly and attracts the attention of researchers for its strong representation ability. Some end-to-end neural network models have been proposed to narrow the domain gap between sketches and images. Qi et al. (2016) used a Siamese Network to gather the sketch and image features of the same category and separate the features of different categories. Sangkloy et al. (2016) employed a Triplet Network to force the negative images to be farther than the positive images, which relaxed the condition in the Siamese Network. No matter it is a Siamese Network or a Triplet Network, the relations they can handle at once is finite. For Siamese Network, it is the relations between the sketch-image pairs, and for Triplet Network, it is the relations among the triplets. In contrast, our proposed model can handle a larger number of relations effectively by building up a graph in a mini-batch.

Zero-Shot Sketch-based Image Retrieval. The Zero-Shot Sketch-based Image Retrieval (ZS-SBIR) is proposed to solve the problem where the training data cannot include all the possible queries (Yelamarthi et al. 2018; Shen et al. 2018). As ZS-SBIR is the combination of zero-shot learning and sketch-based image retrieval, we first briefly introduce the related work of zero-shot learning. In order to let the deep learning models obtain the ability to recognize objects with very little direct supervision like humans (Lampert, Nickisch, and Harmeling 2009), zero-shot learning is proposed. The early works of zero-shot learning used the attributes to infer the labels of unseen classes, while other zero-shot learning approaches mapped the image and semantic feature into a common feature space, a semantic space or another common intermediate space (Xian et al. 2019). Recently, a few methods using generative models are proposed to solve this problem. CVAE-ZSL generated samples from the given attributes and used them for the classification of unseen classes (Mishra et al. 2018). f-CLSWGAN used a Wasserstein GAN to synthesize CNN features conditioned on class-level semantic information (Xian et al. 2018). Furthermore, GDAN built up a bidirectional generative model, which could generate visual features from class embedding features and reconstruct their corresponding class embedding back (Huang et al. 2019).

Inspired by these generative zero-shot learning methods, Yelamarthi et al. (2018) used two generative models, Conditional Variational Autoencoder and Conditional Adversarial Autoencoder, which could narrow the domain gap by generating additional details from the latent prior vector and the given sketches. Dutta and Akata (2019) introduced a generative model with a cycle consistency constraint, where the aligned sketch-image pairs were not required. Besides these generative methods, Shen et al. (2018) proposed a Zero-Shot Sketch-Image Hashing (ZSIH) model, a three-network model for deep generative hashing, mapping the sketches and images into a common semantic space with attention model, Kronecker fusion, and graph convolution. Dey et al. (2019) proposed a discriminate model with triplet loss, and overcame the domain gap through a Gradient Reversal Layer (GRL) which could help the model to capture the modality agnostic features. Different from the generative model, our proposed SketchGCN uses a graph convolution network to map the sketches and images into a common semantic space, which does not suffer the mode collapse problem. The ZSIH model also applies a graph convolution, but its adjacency matrix is pre-computed and fixed. In contrast, our proposed model can learn the adjacency matrix during training, which can narrow the domain gap and transfer the knowledge by taking both the visual information and the semantic information into account.

Graph Convolution Network. Exploring the methods of representing, handling, and operating graph-based data has received increasing attention. Among these methods, there is a kind of method called Graph Convolution Network (GCN), which tries to apply the convolution on graphs directly. Kipf and Welling (2016) explored an approach for semi-supervised learning on graph-structured data. They provided a fast convolution on graphs, and it could be stacked up to build a deep graph neural network. Because of its excellent performance, GCN has been applied to several computer vision tasks, such as Image Captioning (Yao et al. 2018) and Visual Question Answering (Norcliffe-Brown, Vafeias, and

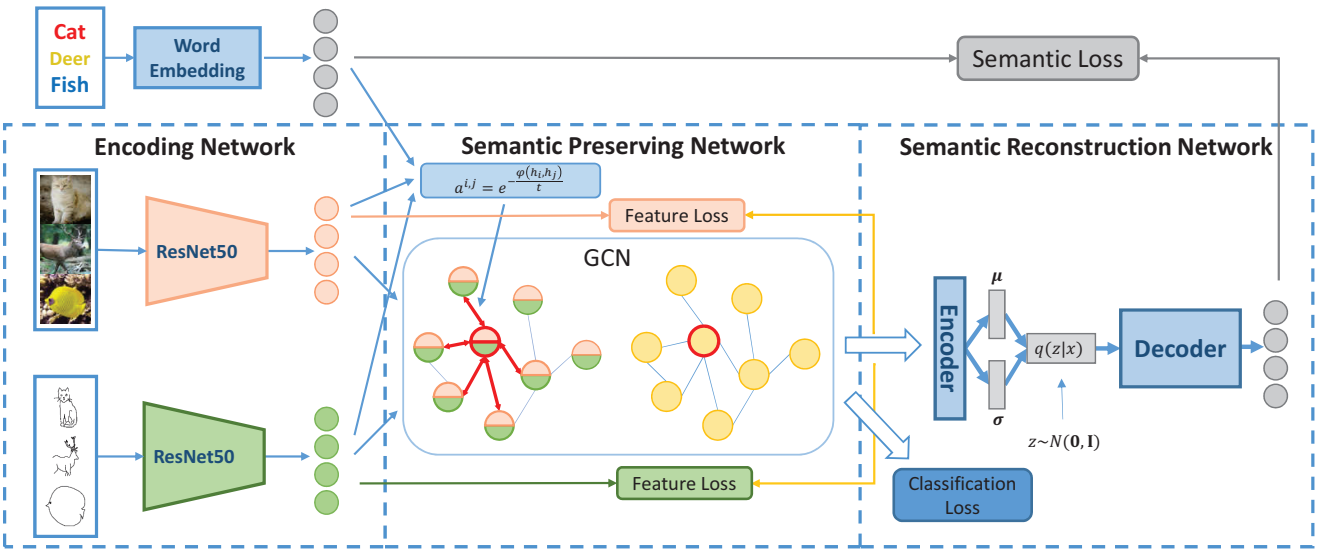


Figure 1: An overview of the architecture of our proposed SketchGCN for ZS-SBIR, which contains encoding network, semantic preserving network, and semantic reconstruction network. The encoding network maps the sketches and images into a common space, while the semantic preserving network takes both the visual information and the semantic information as input and uses a graph convolution network to build and handle their relations. Besides, the model is optimized by the feature loss, classification loss, and semantic loss.

Parisot 2018). Their main idea is to build up a relation graph of the items (i.e., the objects in an image or the images from several categories). In our proposed model, the relations can be built as the similarities between the sketches and the images naturally.

Methodology

In this paper, we propose the SketchGCN model to solve the problem of ZS-SBIR, where sketch and image data is divided into *seen* categories and *unseen* categories. The model is trained only on the sketches and images from *seen* categories and needs to retrieve the images given a sketch belongs to the *unseen* categories. The architecture of our SketchGCN model is illustrated in Figure 1, which contains encoding network, semantic preserving network, and semantic reconstruction network.

We first give a brief problem definition of ZS-SBIR. Let $D = \{(x_i, y_i, l_i)|l_i \in \mathcal{L}\}$ be the dataset consisting of sketches x_i, images y_i, and category labels l_i. $S = \{s_i\}$ represents the set of semantic embedding, i.e., the side information. The dataset is divided into a training set D_{train} and a testing set D_{test} according to whether the label is *seen* ($l_i \in \mathcal{L}^s$) or *unseen* ($l_i \in \mathcal{L}^u$), where $\mathcal{L}^s \cap \mathcal{L}^u = \varnothing$. During the training phase, the model needs to obtain the ability to narrow the domain gap between the sketches and the images and transfer the knowledge from the seen classes to the unseen ones with the help of the side information S. At the test stage, the model is supposed to retrieve the related images given a sketch x from the testing set D_{test}.

Encoding Network

The encoding network adopts a Siamese network architecture to learn two embedding functions $f(\cdot)$ and $g(\cdot)$, which maps the sketches and images into a common embedding space. In the training stage, these two embedding functions do not share weights because the sketches and images are from different modalities. By the semantic preserving network, the embedding functions are guided to learn modality-free representation to narrow the domain gap. We use ResNet50 (He et al. 2016) to model these two embedding functions respectively, and it can be replaced by any other kind of neural networks.

Semantic Preserving Network

The goal of the semantic preserving network is to generate fusion representations for sketches and images, which can narrow the domain gap between them. In the traditional SBIR task, a Contrastive Loss or Triplet Loss is applied to handle the differences between sketches and images by pulling them together if they belong to the same category and pushing them away if not.

However, these methods only consider the visual information which can not deal with all the challenges under the zero-shot setting. The critical problem for zero-shot learning is how to generalize the knowledge obtained from the seen categories to infer the unseen categories. There are many zero-shot learning approaches using the semantic embeddings which imply the relations between the categories to transfer the knowledge. Inspired by the ZSIH model (Shen et al. 2018) and the natural graph structure in the SBIR task (i.e., the nodes are the features of the sketches and the im-

ages, while the edges indicate their similarities), we introduce graph convolution network into our model.

Graph Convolution Network. Following the description in ZSIH (Shen et al. 2018), we build up a graph for a batch of data $\{x_i, y_i, s_i\}_{i=1}^{m}$, where the nodes represent the sketch-image pairs and the edges indicate the relations between the pairs. Let $H^{(l)} = (h_1^{(l)}, h_2^{(l)}, ..., h_m^{(l)})^T$ denotes the feature matrix of the nodes in the l-th GCN layer and $A \in \mathbb{R}^{m \times m}$ denotes the graph adjacency matrix. The node features are computed by concatenating the sketch features and the image features. The convolutional operation in GCN follows the below layer-wise propagation rule:

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}) \tag{1}$$

where $\hat{A}$ is a normalized version of the graph adjacency matrix A; $W^{(l)}$ is a parameter matrix; σ is a non-linear operation like ReLU. Since the semantic embeddings imply the relations among all the categories, each element $a^{i,j}$ in the adjacency matrix A is decided by the semantic embeddings S, which is computed as:

$$a^{i,j} = e^{-\frac{\|s_i - s_j\|_2^2}{t}} \tag{2}$$

where t is a regulation factor and $a^{i,j}$ indicates the similarity between the node h_i and the node h_j. Therefore, the nodes will be affected by the nodes which are similar to them on the semantic space, and their semantic relations are then obtained by the model.

Graph Convolution Network with Metric Learning. Determining the graph adjacency matrix only by the semantic embedding may be arbitrary. Besides the challenge of transferring the knowledge, how to narrow the domain gap between sketches and images is another challenge inheriting from the traditional SBIR task. Therefore, we take the visual information into account when calculating the adjacency matrix. Inspired by the successful combination of metric learning and graph convolution network in few-shot learning (Garcia and Bruna 2017), we introduce a trainable adjacency matrix into our model. The details are shown in Figure 2. We first compute their semantic distance on the semantic space. The semantic distance $d_{i,j}$ between the nodes h_i and h_j is computed as:

$$d_{i,j} = \|s_i - s_j\|_1 \tag{3}$$

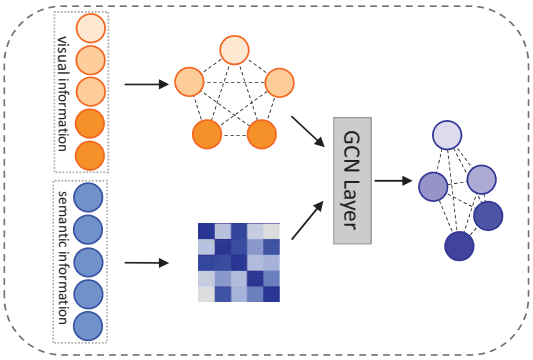
We then use a metric function φ which considers both the visual information and the semantic information to compute the distance between the nodes. The metric function φ is modeled by a Multilayer Perceptron (MLP), which is computed as Eq. (4).

$$\varphi(h_i, h_j) = \text{MLP}([abs(h_i - h_j), d_{i,j}]) \tag{4}$$

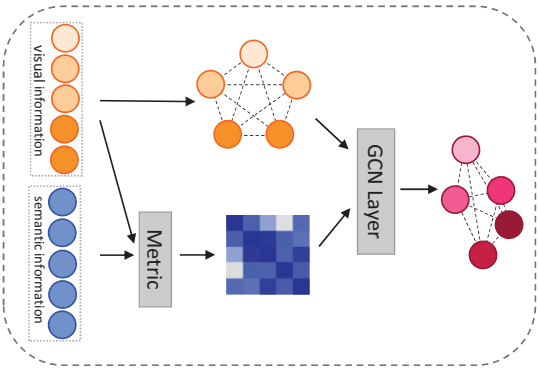
where $[\cdot, \cdot]$ is the concatenation operation. Each element of the adjacency matrix A can be computed as:

$$a^{i,j} = e^{-\frac{\varphi(h_i, h_j)}{t}} \tag{5}$$

where t is a regulation factor. In this way, the visual information and the semantic information are both considered in the adjacency matrix. Therefore, the model can deal with both the visual and semantic relations between sketches and images. Finally, the features of the nodes through the GCN represent the fusion embedding of the sketch-image pairs.



(a) Only semantic information is used to compute the adjacency matrix.



(b) Both semantic information and visual information are used to compute the adjacency matrix.

Figure 2: An illustration of the graph convolution network in our model.

Semantic Reconstruction Network

In order to further force the model to retain the category-level relations in the semantic space, a few methods utilize a decoder network to reconstruct the semantic information from the generated embedding (Dey et al. 2019; Shen et al. 2018). These methods commonly model the decoder network as an MLP, but such MLP can only remember the semantic information that it has met during the training phase. This makes the model have a poor ability to generalize the knowledge to infer the unseen classes during the testing phase. The MLP can be easily formalized as:

$$s = \psi(x) \tag{6}$$

where ψ is the reconstruction net modeled by MLP; s and x are the feature vector of semantic embedding and visual embedding respectively.

In contrast, we apply a Conditional Variational Autoencoder (CVAE) (Sohn, Lee, and Yan 2015) to generate the se-

mantic information from the embedded features. The CVAE can enhance our model, for it can generate the semantic information given the sketches from the unseen classes. It contains an encoder network and a decoder network. Specifically, given the fusion embedding of sketch and image x^{fus}, and the semantic embedding s, an encoder network parameterized by ϕ approximates the variational distribution $q_\phi(z|x^{fus})$, and a decoder network parameterized by θ models the conditional distribution $p_\theta(s|z, x^{fus})$. The variational lower bound can then be written as:

$$\begin{aligned}\mathcal{L}_{KL}(\phi,\theta;s,x^{fus}) = \\ &- D_{KL}(q_\phi(z|x^{fus},s)||p_\theta(z|x^{fus})) \\ &+ \mathbb{E}[\log p_\theta(s|z,x^{fus})]\end{aligned} \tag{7}$$

The encoder network takes the fusion features output from the previous network and tries to embed them into a latent space. The decoder network reconstructs the semantic embedding from the latent embedding under the condition of the fusion features. It can be formalized as:

$$\hat{s} = D([z, x^{fus}]) \tag{8}$$

where $[\cdot,\cdot]$ is the concatenation operation. The random variable z is reparameterized using a differentiable transformation $z = g_\phi(x^{fus}, \epsilon), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (Kingma and Welling 2013). With the help of CVAE, the generalization ability of our model can be further enhanced.

Learning objectives

The learning objective of the proposed model consists of the feature loss $\mathcal{L}_f$, the classification loss $\mathcal{L}_{cls}$, the semantic loss $\mathcal{L}_{sem}$, and the above mentioned KL divergence loss $\mathcal{L}_{KL}$.

Feature Loss. The feature loss aims to narrow the distance between the encoding network outputs and the semantic preserving network outputs. The feature loss can be computed as:

$$\mathcal{L}_f = \frac{1}{2N}\sum_{i=1}^{N}(\left\|x_i^{sketch} - x_i^{gcn}\right\|_2^2 + \left\|x_i^{image} - x_i^{gcn}\right\|_2^2) \tag{9}$$

where x_i^{sketch} and x_i^{image} represent the sketch feature and image feature output from the encoding network respectively; and x_i^{gcn} denotes the fusion feature output from the semantic preserving network.

Classification Loss. We connect a linear classifier with the parameter θ_c to the output of the semantic preserving network, which can ensure the output fusion feature preserves the discriminate characters within each training category. We implement the classification loss as a Cross-Entropy Loss, which can be computed as:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} - \log P(l_i|x_i^{gcn};\theta_c) \tag{10}$$

where l_i represents the ground truth label and x_i^{gcn} is the output fusion feature.

Semantic Loss. The semantic loss is used to restrain the generated semantic embedding, which is computed as:

$$\mathcal{L}_{sem} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{s}_i - s_i\|_2^2 \tag{11}$$

where $\hat{s}_i$ and s_i represent the generated semantic embedding and the ground truth respectively.

The overall objective function is combined with the above four parts, which is defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_f + \lambda_2\mathcal{L}_{cls} + \lambda_3\mathcal{L}_{sem} + \mathcal{L}_{KL} \tag{12}$$

where λ_i $(i = 1, 2, 3)$ denotes the parameter to balance the different parts of the objective function.

Experiments

Dataset and Implementation Details

We evaluate our proposed model on two popular SBIR datasets, i.e., *Sketchy* (Sangkloy et al. 2016) and *TU-Berlin* (Eitz, Hays, and Alexa 2012) by conducting extensive experiments. These two datasets are first proposed for SBIR and extended by Liu et al. (2017) for large-scale sketch-image retrieval.

Sketchy-Extended is a large-scale dataset which contains 125 categories. There are 100 images and at least 600 sketches in each category. Liu et al. (2017) collected 60,502 natural images from *ImageNet* (Deng et al. 2009) to extend this dataset. Therefore, *Sketchy-Extended* totally contains 73,002 images and 75,479 sketches. We follow the setting in Yelamarthi et al.'s work (2018) and split the dataset into 104 categories for training and 21 categories for testing, making sure that the testing categories do not appear in the 1,000 categories of *ImageNet*.

TU-Berlin-Extended totally contains 20,000 unique sketches evenly distributed over 250 categories. In order to make this dataset suitable for SBIR, Liu et al. (2017) extended this dataset by collecting 204,489 images. We adopt this extended dataset into our experiment. Following Shen et al. (2018), we randomly choose 30 categories that contain at least 400 images for testing and the rest 220 categories for training.

Implementation Details Our proposed SketchGCN model is implemented with the popular deep learning toolbox Pytorch and trained on 4 TITAN Xp graphics cards. We use ResNet50 (He et al. 2016) pre-trained on *ImageNet* to model the encoding networks $f(\cdot)$ and $g(\cdot)$, and they are fine-tuned during training. Each network outputs the 2048-D features for sketches and images. We apply a single-layer graph convolution network for the semantic preserving network, which takes the 4096-D concatenation features as input and outputs the 2048-D fusion features. The MLP to model the metric function φ is implemented by stacking 4 fully connection layers, and each layer except the last layer is followed by a batch normalization and an activation function Leaky ReLU with 0.01 negative slope. The last layer is followed by ReLU, for the distance needs to be no less than zero. For the semantic reconstruction network, a Conditional Variational Autoencoder is implemented. An encoder first

Table 1: The comparison results against the recently published ZS-SBIR methods.

Method	*Sketchy-Extended*				*TU-Berlin-Extended*			
	mAP	P@100	P@200	mAP@200	mAP	P@100	P@200	mAP@200
ZSIH (Shen et al. 2018)	0.254	0.340	-	-	0.220	0.291	-	-
CVAE (Yelamarthi et al. 2018)	0.196	-	0.333	0.225	0.005	-	0.003	0.009
GZS-SBIR[1](Kumar Verma et al. 2019)	0.253	0.305	-	-	0.187	0.281	-	-
GZS-SBIR[2](Kumar Verma et al. 2019)	0.289	0.358	-	-	0.238	0.334	-	-
SEM-PCYC (Dutta and Akata 2019)	0.349	0.463	-	-	0.297	0.426	-	-
Doodle2Search (Dey et al. 2019)	0.369	-	0.370	0.461	0.109	-	0.121	0.157
SkechGCN (Ours)	**0.382**	**0.538**	**0.487**	**0.568**	**0.324**	**0.505**	**0.478**	**0.528**

[1] Feedback-Auto
[2] Feedback-VAE

maps the fusion feature into a 1024-D latent feature space. A decoder then utilizes three fully connection layers to reconstruct the 300-D word embedding from that latent space. The first two fully connection layers are followed by a batch normalization and a ReLU activation function. As for the side information, we use GloVe (Pennington, Socher, and Manning 2014) to embed the category label into a 300-D word embedding. When the category label does not appear in the word dictionary, we split the category label into words and use their average instead. The model is optimized by the Adam algorithm with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ across all the datasets, and the learning rates for the three part networks are set as $lr_1 = 0.00001, lr_2 = lr_3 = 0.0001$ respectively. The weights of the loss are set as $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 0.1$. The whole model is trained end-to-end during the training phase, and during the testing phase only the encoding networks will function.

Comparison with Existing Methods

To verify the superiority of our proposed model, we make the comparison with the other five recently published methods on ZS-SBIR, i.e., ZSIH (Shen et al. 2018), CVAE (Yelamarthi et al. 2018), GZS-SBIR (Kumar Verma et al. 2019), SEM-PCYC (Dutta and Akata 2019), and Doodle2Search (Dey et al. 2019), and two main paradigms of ZS-SBIR methods are both taken into concerned. CVAE, GZS-SBIR, and SEM-PCYC are the methods from the first paradigm, which generate the corresponding image features according to the sketch features and category labels. ZSIH and Doodle2Search are from the second paradigm, which maps the sketches and images into a common feature space. The details are shown in Table 1. As different methods adopt different metrics, we provide all the results based on these metrics to compare. Specifically, Mean Average Precision (mAP, mAP@200) and Precision considering top 100 and 200 (P@100, P@200) are calculated to make evaluations.

Table 1 shows that our model outperforms all the other existing methods on almost all the metrics, which shows the effectiveness of our model. The generative model mainly uses the sketch features to obtain the corresponding image features with the category label. This may use the visual information effectively, but does not fully utilize the semantic information which indicates their relations between the seen categories and the unseen categories. In contrast, our model can make good use of the semantic information by using the graph convolution network to handle the implied relations. It is worth noting that ZSIH also adopts the GCN in their model, but its adjacency matrix is precomputed and only determined by the semantic embedding of the category labels. This may ignore the visual information which also plays an important role in the ZS-SBIR task. In contrast, our model also utilizes the visual information when modeling the graph. This can significantly narrow the gap between different modalities and alleviate the problems in the ZS-SBIR task. Furthermore, the CVAE applied in our model also helps a lot. The *TU-Berlin-Extended* dataset is a more challenging and more realistic dataset compared with the *Sketchy-Extended* dataset. *TU-Berlin-Extended* contains a larger number of categories and the sketches are more abstract, while the sketches in *Sketchy-Extended* are well-drawn. Therefore, all the other existing methods deliver worse results on *TU-Berlin-Extended*. Despite this, our proposed model can still get a good performance on the challenging *TU-Berlin-Extended* dataset. In conclusion, our model mainly benefits from utilizing all the information effectively with the GCN and CVAE.

Ablation Study

To further evaluate the effectiveness of each component in our proposed model, we conduct several experiments on the *Sketchy-Extended* and *TU-Berlin-Extended* datasets. Different components are compared by modifying some parts of the SketchGCN.

We implement two basic models that do not contain the semantic reconstruction network. One implements the GCN with a fixed adjacency matrix A only, where the adjacency matrix A is computed following Eq. (2), while the other one implements the GCN with a learnable adjacency matrix A only, where the adjacency matrix A is learned following Eqs. (3-5). Both of their regulation factors t are set as $t = 0.1$. Besides, a model adopting an MLP as the semantic reconstruction network is also implemented to illustrate the effectiveness of the CVAE applied in our model, and the MLP is modeled by stacking two fully connection layers. The experimental results are shown in Table 2.

Discussion on *Sketchy-Extended*. Our basic model with

Table 2: The ablation study results on our proposed model.

Description	*Sketchy-Extended*			*TUBerlin-Extended*		
	mAP@200	P@200	mAP	mAP@200	P@200	mAP
Only GCN with a fixed A	0.549	0.472	0.366	0.517	0.465	0.307
Only GCN with a learnable A	0.556	0.472	0.367	0.508	0.456	0.296
MLP as reconstruction network	0.555	0.479	0.375	0.506	0.457	0.300
SkecthGCN (full model)	**0.568**	**0.487**	**0.382**	**0.528**	**0.478**	**0.324**

GCN only obtains 0.549 mAP@200 and 0.472 P@200 on the *Sketchy-Extended* dataset, which outperforms the triplet-network-based method, Doodel2Search (Dey et al. 2019). This can be attributed to the strong ability to handle the relations of GCN. Compared with the triplet network, GCN can deal with more relations at one time. Comparing the results obtained by the GCN with a learnable A and fixed A, we can draw a conclusion that taking the visual information into account when determining the adjacency matrix can figure out the differences between the sketches and the images and lead to better performance. The results in Rows 4 and 5 of Table 2 show the effectiveness of using a semantic reconstruction network. Furthermore, our full model which uses a CVAE to generate the semantic information from the visual embedding is superior to the model using MLP. This shows that our model can gain a certain generalization ability which benefits from the CVAE.

Discussion on *TU-Berlin-Extended*. Our basic model still outperforms the triplet-network-based method, which obtains 0.436 mAP@200 and 0.396 P@200. However, the models with a learnable A and MLP as the reconstruction network perform worse than the basic model. This may because the *TU-Berlin-Extended* dataset contains many categories that have substantial visual similarities (Dutta and Akata 2019), and such similarities can confuse the model which lead to poor performance. The results in Row 4 of Table 2 show that using an MLP leads to no improvement. As we argued before, simply mapping the visual embedding back to the semantic embedding does no good to the generalization ability of the model. However, the results in Row 5 of Table 2 show that our full model outperforms the basic model and can alleviate the consequence caused by the above mentioned visual similarity. This again proves the effectiveness of the CVAE which helps our model further generalize the knowledge to infer the unseen categories.

Visualization and Qualitative Analysis

In this section, we visualize some retrieval results delivered by our proposed model. The top-10 retrieval results of the given sketches are shown in Figure 3.

It can be observed that the retrieved images have visual similarities with the query sketches. The query sketches from *cow*, *rhinoceros*, and *giraffe* can retrieve the items from each other sometimes. This is probably because the visual and semantic similarities among them. For *cow* and *rhinoceros*, they have a similar outline, and all of those three animals live on the grassland. This kind of phenomenon can also be observed from *songbird* and *seagull*, *window* and *door*, etc. In general, the category that has a unique shape tends to get a better result like *helicopter* in *Sketchy-Extended* and *microscope* in *TU-Berlin-Extended*. However, having a common outline leads to a poor result sometimes. For example, *donut* may retrieve all the images that have a circle outline.

(a) On *Sketchy-Extended*.

(b) On *TU-Berlin-Extended*.

Figure 3: Some examples of the top-10 retrieval results on (a) *Sketchy-Extended* and (b) *TU-Berlin-Extended*. The correct retrieved images are marked by green and the incorrect are marked by red.

Conclusion

In this paper, we propose the SketchGCN model for the ZS-SBIR task, where we combine a GCN model and a CVAE model successively. The ability to handle relations of GCN helps us to deal with the similarities between sketches and images. Our model can leverage both the visual information and the semantic information effectively which benefits from the method of computing the adjacency matrix. The CVAE helps our model to gain the generalization ability to infer the unseen categories. In future work, we will consider exploring a more effective way to model the graph structure in the ZS-SBIR task.

Acknowledgment

This work was supported in part by the Naional Natural Science Foundation of China (No. 61976057 and

No. 61572140), in part by the Shanghai Municipal R&D Foundation (No. 17DZ1100504 and No. 16JC1420401), in part by the Shanghai Natural Science Foundation (No. 19ZR1417200), and in part by the Humanities and Social Sciences Planning Fund of Ministry of Education of China (No. 19YJA630116). The work of Weiguo Fan was supported by the Henry Tippie Endowed Chair Fund from the University of Iowa. Yuejie Zhang and Tao Zhang are corresponding authors.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR 2009*, 248–255. IEEE.

Dey, S.; Riba, P.; Dutta, A.; Llados, J.; and Song, Y.-Z. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of CVPR 2019*, 2179–2188. IEEE.

Dutta, A., and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of CVPR 2019*, 5089–5098. IEEE.

Eitz, M.; Hildebrand, K.; Boubekeur, T.; and Alexa, M. 2010. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics* 17(11):1624–1636.

Eitz, M.; Hays, J.; and Alexa, M. 2012. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)* 31(4):44:1–44:10.

Garcia, V., and Bruna, J. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, 770–778. IEEE.

Hu, R., and Collomosse, J. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding* 117(7):790–806.

Huang, F.; Jin, C.; Zhang, Y.; Weng, K.; Zhang, T.; and Fan, W. 2018. Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition* 76:537–548.

Huang, H.; Wang, C.; Yu, P. S.; and Wang, C.-D. 2019. Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of CVPR 2019*, 801–810. IEEE.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kumar Verma, V.; Mishra, A.; Mishra, A.; and Rai, P. 2019. Generative model for zero-shot sketch-based image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of CVPR 2019*, 951–958. IEEE.

Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of CVPR 2017*, 2862–2871. IEEE.

Mishra, A.; Krishna Reddy, S.; Mittal, A.; and Murthy, H. A. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2188–2196.

Norcliffe-Brown, W.; Vafeias, S.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. In *Proceedings of NeurIPS 2019*, 8334–8343.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, 1532–1543.

Qi, Y.; Song, Y.-Z.; Zhang, H.; and Liu, J. 2016. Sketch-based image retrieval via siamese convolutional neural network. In *Proceedings of ICIP 2016*, 2460–2464. IEEE.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35(4):119.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-shot sketch-image hashing. In *Proceedings of CVPR 2018*, 3598–3607. IEEE.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *Proceddings of NeurIPS 2015*, 3483–3491.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of ICCV 2017*, 5551–5560. IEEE.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *Proceedings of CVPR 2018*, 5542–5551. IEEE.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(9):2251–2265.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of ECCV 2018*, 684–699. Springer.

Yelamarthi, S. K.; Reddy, S. K.; Mishra, A.; and Mittal, A. 2018. A zero-shot framework for sketch based image retrieval. In *Proceedings of ECCV*, 316–333. Springer.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C.-C. 2016. Sketch me that shoe. In *Proceedings of CVPR 2016*, 799–807. IEEE.