Zero-Shot Sketch Based Image Retrieval

Thesis to be submitted in partial fulfillment of the requirements for the degree

of

Bachelor of Technology

by

Ruchika Chavhan
170260011
Department of Physics

Under the guidance of

Prof. Biplab Banerjee
Centre of Studies in Resources Engineering

Prof. Anshuman Kumar
Department of Physics

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

ACKNOWLEDGEMENTS

I would like to thank Prof. Biplab Banerjee for his guidance throughout this project. I am extremely grateful to him for the encouragement, motivation and the opportunities that he has provided in the span of two years that I have worked with him. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am also thankful to Prof. Anjan Dutta and Prof. Zeynep Akata for the several helpful discussion sessions. I would also like to thank Prof. Anshuman Kumar for his interest in my project. Finally, I am also grateful to Ushashi Chaudhuri, who is the co-author of the manuscript currently under review.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. I am thankful to my sister Vishakha and my friend Abhishek for their constant support.

Ruchika Chavhan
IIT Bombay

ABSTRACT

Sketch Based Image Retrieval (SBIR) allows a user to search images with a free-hand sketch. A common bottleneck in SBIR is the scarcity of data occurring due to the laborious task of drawing sketches. This motivates the concept of Zero-shot learning, where a learner observes samples from classes that were not observed during training. In Zero-shot sketch-based image retrieval (ZS-SBIR), human sketches are used as queries to conduct retrieval of photos from unseen categories. The task of ZS-SBIR is challenging due to the fine-grained nature of the task, large domain gap between sketches and images and high intra-class variance of many categories. In this thesis, we propose a novel ZS-SBIR framework performing a bi-level domain adaptation of the sketch and image features using adversarial learning. This framework alleviates the above problems by providing modality-independent features and a class-discriminative latent space. Experimental results on the extended versions of the Sketchy, TU-Berlin, and QuickDraw datasets exhibit sharp improvements over the state-of-the-art. This manuscript is currently under review.

Keywords: Sketch Based Image Retrieval, Zero-shot learning, Domain Adaptation

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sketch Based Image Retrieval

Machine Learning has slowly become a ubiquitous component in modern software. Moreover, image content on the internet is increasing exponentially with the advent of social media and e-commerce. Most users search for an image using a textual description or by providing another image similar to the desired image. But, it is often difficult to describe images using a textual description but finding a visual equivalent is easier. Finding images belonging to the same domain as the desired image is often tedious. This motivates the concept of sketches as a visual query as they can be easily drawn on a touch-based device.

Sketch Based Image Retrieval (SBIR) is the task of retrieving natural images corresponding to a hand drawn sketch. Due to the high variance in man-made sketches and domain-sensitivity of machine learning models, classification based image retrieval is prone to highly erroneous results. Thus, SBIR framework should learn to associate salient components in the sketch with the corresponding components in the image having similar characteristics.

1.2 Zero-Shot Learning

Obtaining enough hand-made sketches that capture all variations for all possible objects in the world is arduous task. This calls for Zero-shot learning, a setup in which samples from unseen classes are presented to the learner. Naturally, some sort of side information is required for the zero-shot classes is required. This information can be available as:

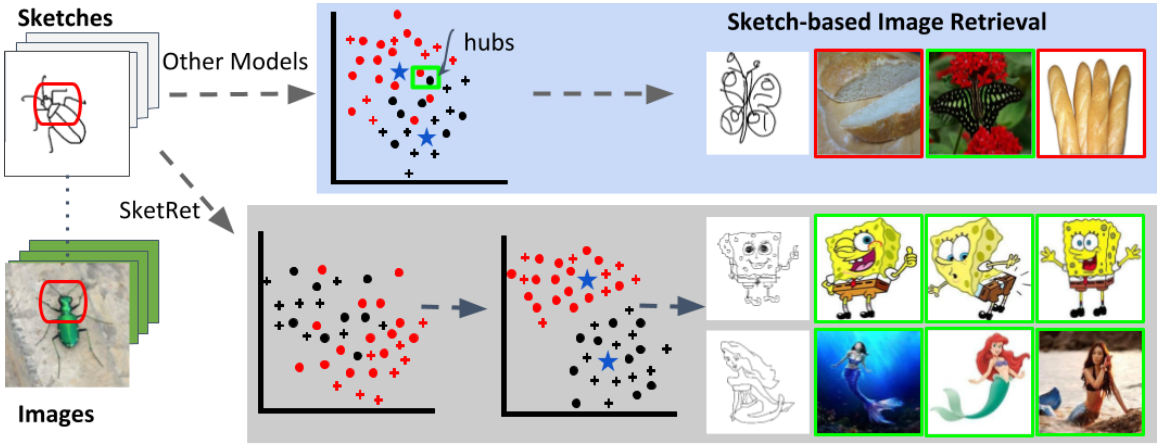

Figure 1.1: Given a cluttered feature space for images and sketches, a sub-optimal domain alignment fails to produce a discriminative latent space which is affected by problems like hubness and negative transfer (blue region). Feature adaptation at multiple feature scales together with a discriminative feature space learning ensures that zero-shot testing can be performed well in SketRet (gray region).

- **Attributes:** Certain attributes or class names are available. Eg: "black and white stripes", "long neck" etc.
- **Textual description:** Categories or images are associated with a sentence which aptly describes the content.
- **Class-class similarity:** The learner learns a continuous embedding for the categories. It maps the unseen category to the nearest class to provide a prediction.

For ZS-SBIR, semantic side information is available as class attributes/labels. One cannot trivially map the images/sketches to the semantic space as this will neglect the complex distributions of different local regions within the images. In ZS-SBIR, only natural images belonging to the unseen classes are available. However, in real-time image retrieval, one cannot differentiate between seen and unseen examples. The availability of both seen and unseen classes will make the retrieval task more confusing. This setting is called Generalized Zero-shot SBIR (GZS-SBIR). The tasks of ZS-SBIR and GZS-SBIR are extremely challenging due to the following:

- **Variance:** As sketches are drawn by various artists, the ZS-SBIR models need to overcome a substantial within-category variance.

- **Domain-gap:** Besides, the domain gap between sketches and images is considerable, given the disparity in spectral, spatial, and texture properties between the two modalities. Therefore, feature space should not be forcefully aligned; otherwise, the model becomes vulnerable to the negative transfer of irrelevant knowledge.
- **Hubness:** Another critical challenge in ZS-SBIR is the hubness problem, which occurs when a model has a training bias and retrieves images only from a subset of the available categories. This occurs as some embedding vectors of images (also called as "hubs") appear in the nearest neighborhood of many test query sketches.

In this thesis, we introduce SketRet, a discriminative deep framework for performing ZS-SBIR. We combat the negative transfer and hubness problems of ZS-SBIR by performing an improved feature adaptation while designing an insightful semantic projection network combining a neural network and a Graph Convolutional Network (CNN) which brings in a structural semantic consistency of the latent features. Extensive experiments are conducted on the Sketchy (Extended), TU-Berlin (Extended), and the QuickDraw (Extended) datasets by exploring a number of semantic spaces and feature backbones.

The report is divided into 6 sections. Section 2 contains a brief description of previous works in the field of SBIR. Section 3 describes the proposed methodology of our SketRet framework. Section 4 contains the Experimental Setup and Implementation details. In Section 5, we present the qualitative and quantitative results of our model. Finally, In section 6 we summarize the SketRet framework.

Chapter 2

Related Work

In this section, we review prior works on zero-shot SBIR.

As already stated, the major obstacle in solving the SBIR task stems from the fact that the distributions difference between sketch and image data is exceedingly large. Early works in this area include conventional pattern recognition methods for retrieval by engineering hand-crafted visual features [14, 22]. The proposition behind such approaches is to solve the SBIR problem by obtaining the edge-map of the natural images and to further match them with sketches arising from the same categories. As expected, the low-level SIFT [18], SURF [1], or HoG [4] based descriptors are unable to properly encode the regional variations of the sketch data, resulting in an inferior cross-modal matching. The performance measures of deep CNN based SBIR models have witnessed a massive enhancement lately, thanks to the data-driven feature learning capabilities of CNN. Since the retrieval performance benefits from a discriminative feature space, several endeavors rely on distance metric learning strategies like contrastive-loss [3], triplet-loss [23], and HOLEF-based loss [26], to name a few. As opposed to the real-valued feature embedding, hash-code based representations are also considered in this regard which offers a trade-off between performance and storage [16]. The generative models for cross-modal style transfer are also explored [9] in this regard.

Zero-shot Sketch-based Image Retrieval (ZS-SBIR): The ZS-SBIR literature consists of both the discriminative and generative deep learning based techniques. Under the generative umbrella, [24] proposes a hashing network for the semantic knowledge reconstruction (ZSIH). Similarly, [27] introduces a conditional generative

model for ZS-SBIR based on variational learning. The stacked auto-encoder (SAN) method proposed in [20] deploys a generative framework based on stacked-adversarial networks within a Siamese architecture. The paired cyclic consistency loss proposed in SEM-PCYC [7, 8] helps in aligning the sketches and images in an encoded semantic space using adversarial training. On a different note, [9] borrows ideas from the style transfer literature and develops a style-guided image to image translation model for ZS-SBIR. On the other hand, the discriminative model of [6] uses a triplet-based network to solve the task at hand. [10] highlights the implications of data and class imbalance in ZS-SBIR and introduces an adaptive margin diversity regularizer (AMD-reg) to combat the same. While all the techniques showcase their performance on ZS-SBIR, a few works [7, 8, 20] also demonstrate their experiments for the GZS-SBIR setting.

Chapter 3

Proposed Methodology

3.1 Preliminaries

Let $\mathcal{Z}^s = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{C}^s, \mathcal{W}^s\}$ be a multi-modal training dataset consisting of images $\mathcal{A}^s$ and sketches $\mathcal{B}^s$ obtained from the $|\mathcal{C}^s|$ *seen* visual categories. Additionally, we have access to semantic side information $\mathcal{W}^s$ which typically corresponds to the distributed word-vector embeddings of the individual category names. During inference, image and sketch data $\mathcal{Z}^u = \{\mathcal{A}^u, \mathcal{B}^u\}$ from a non-overlapping set of previously *unseen* classes $\mathcal{C}^u$ are considered ($\mathcal{C}^u \cap \mathcal{C}^s = \varnothing$) in the zero shot SBIR setup.
We deal with the unpaired dataset setting in $\mathcal{Z}^s$ where the number of sketch and image instances in $\mathcal{A}^s$ and $\mathcal{B}^s$ are different: $\{a_i^s\}_{i=1}^N \in \mathcal{A}^s$ and $\{b_i^s\}_{i=1}^M \in \mathcal{B}^s$. The model is trained to reduce the distribution mismatch between $\mathcal{A}^s$ and $\mathcal{B}^s$ and subsequently to transfer the knowledge from $\mathcal{Z}^s$ to $\mathcal{Z}^u$ with the help of the semantic information $\mathcal{W}^s$. The testing phase concerns the retrieval of images with similar semantic categories from $\mathcal{A}^u$ given the sketch queries from $\mathcal{B}^u$. In contrast to ZS-SBIR, GZS-SBIR assumes the presence of images from $\mathcal{A}^s \cup \mathcal{A}^u$ during testing for unseen-class sketch queries coming from $\mathcal{B}^u$, however, only $\mathcal{Z}^s$ is used during training in both the cases.

3.2 Overview

The goal of SketRet is to align the images and sketches from the same class in a semantically meaningful shared latent space. It is composed of cross-modal triplets where the sketch data from $\mathcal{B}^s$ serves as the anchor (α) while the positive (p) and negative (n) counterparts are selected from $\mathcal{A}^s$ (Fig. 3.1). We denote the label and semantic prototype for (α, p) by y^+ and w^+ while y^- is the label for n, respectively.

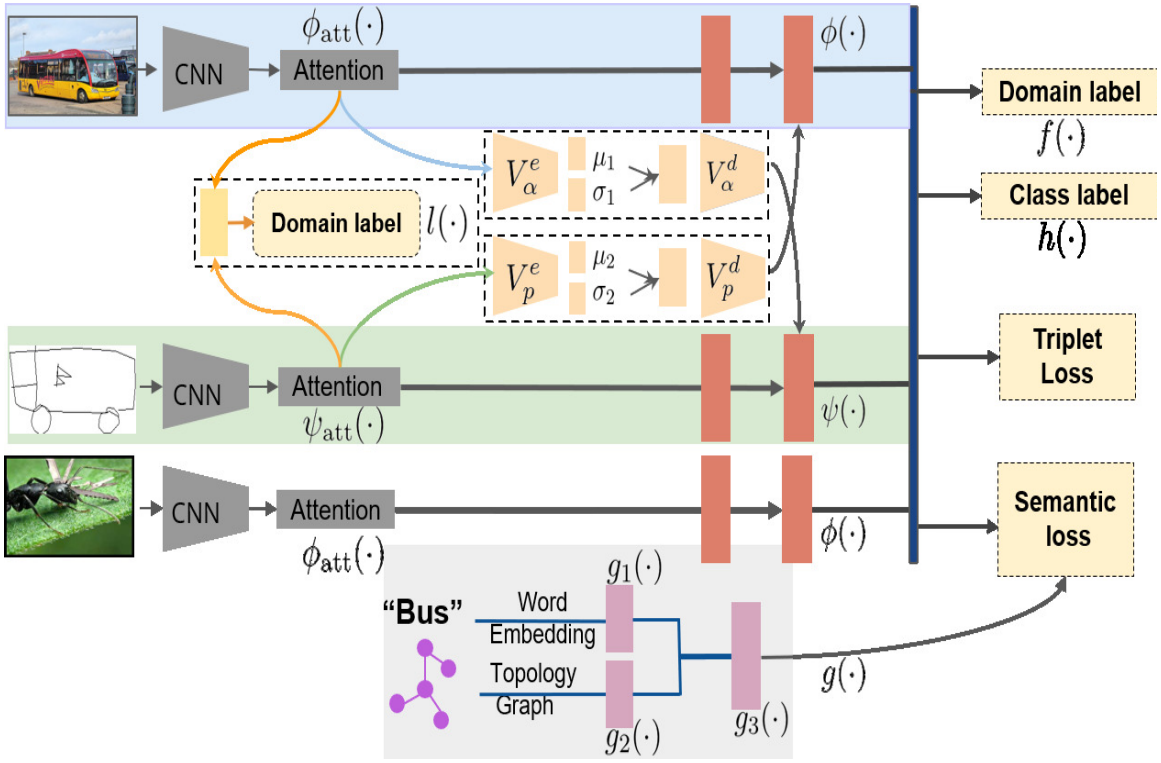

Figure 3.1: A depicton of our SketRet architecture. The images and sketches undergo two rounds of domain adaptation at the outputs of (ϕ_{att}, ψ_{att}) and (ϕ, ψ), respectively. The cross-modal encoder-decoder modules $\mathcal{V}_\alpha = (\mathcal{V}_\alpha^e, \mathcal{V}_\alpha^d)$ and $\mathcal{V}_p = (\mathcal{V}_p^e, \mathcal{V}_p^d)$ (light blue and light green arrows) aid in learning improved cross-modal features. The semantic projection network $g(\cdot)$ (light orange arrows) embeds the prototypes and the semantic topology graph into the latent space. At test time, images and sketches from the unseen classes are projected into the latent space through $\phi(\cdot)$ and $\psi(\cdot)$ and the nearest neighbor based retrieval subsequently takes place.

The feature networks for $\mathcal{A}^s$ and $\mathcal{B}^s$ are defined by $\phi(\cdot)$ and $\psi(\cdot)$ which are convolutional neural networks with integrated attention sub-networks $\phi_{att}(\cdot)$ and $\psi_{att}(\cdot)$ ($\phi_{att} \subset \phi, \psi_{att} \subset \psi$). The attention block outputs are simultaneously projected to the local adversarial domain classifier $l(\cdot)$ to highlight spatially indistinct features of the same-class samples from $\mathcal{A}^s$ and $\mathcal{B}^s$ and to the shared latent space.

We further introduce two cross-modal feature reconstruction modules ($\mathcal{V}_\alpha(\cdot), \mathcal{V}_p(\cdot)$) which aim to reconstruct $\phi(p)$ from $\psi_{att}(\alpha)$ and $\psi(\alpha)$ from $\phi_{att}(p)$ through variational bottlenecks. On the other hand, the outputs of $\phi(\cdot)$ and $\psi(\cdot)$ need to be synchronized for defining the shared embedding space. In this regard, the global domain adaptation on $\phi(p/n)$ and $\psi(\alpha)$ is carried out considering a combination of the domain classifier $f(\cdot)$ and a multi-class category classifier $h(\cdot)$. A semantic sub-network $g(\cdot, \cdot)$ comprising of an MLP $g_1(\cdot)$ and a graph CNN $g_2(\cdot, \cdot)$ is used to non-linearly project the semantic vectors into the shared space.

Loss Functions: There are mainly three learning objectives that together govern the training of SketRet, namely, (i) domain loss ($\mathcal{L}_{domain}$), (ii) cross-model triplet loss ($\mathcal{L}_{triplet}$), and (iii) semantic loss ($\mathcal{L}_{semantic}$), respectively. The loss functions are detailed in the following.

3.3 Domain Losses

To adapt the features of $\mathcal{A}^s$ and $\mathcal{B}^s$, we aspire the latent visual embeddings to be driven by shared region-level features and to suppress the effects of irrelevant domain-specific concepts. In our SketRet framework, a two-level feature adaptation is carried out to accomplish the fine-grained domain alignment between $\mathcal{A}^s$ and $\mathcal{B}^s$. Especially, we gradually ensure the local domain invariance by first learning shared spatial features at the outputs of $\phi_{att}(\cdot)$ and $\psi_{att}(\cdot)$ followed by globally aligning the latent visual embeddings obtained from $\psi(\cdot)$ and $\phi(\cdot)$, respectively. Local feature adaptation may be affected by the large data variation within each of the modalities. Hence, we introduce two variational encoder-decoder modules connecting $\phi_{att}(p)$ to $\psi(\alpha)$ and $\psi_{att}(\alpha)$ to $\phi(p)$, respectively, which contribute to maximizing the correlation between the bi-modal features through cross-modal feature reconstruction. It is assured that the domain losses leverage the class labels effectively.

3.3.1 Global Adaptation

The global domain adaptation loss in SketRet follows an adversarial training strategy where the discriminator tries to minimize the domain confusion and the feature extractors aim to maximize the domain disagreements in the shared latent space produced by (ϕ, ψ). This, in turn, makes the latent space agnostic to both the modalities. We note that the coarse-level distribution matching between the modalities using only the domain discriminator f may induce a trivial solution where all the samples may be confined in a few modes. As a remedy, we need to ensure that the samples maintain the class labels in the adapted space. Accordingly, the discriminator in SketRet is defined in terms of the domain classifier f and the label predictor h, while ϕ and ψ serve as the feature extractors.

As per the principle, f considers the initial domain labels for $\psi(\alpha)$ and $(\phi(p), \phi(n))$ to be 1 and 0, respectively. Likewise, the global domain loss $\mathcal{L}_{dom}^{global}$ and the multi-class classification loss $\mathcal{L}_{class}$ are defined as,

$$\mathcal{L}_{dom}^{global} = \mathbb{E}_{\alpha \in \mathcal{B}^s, p,n \in \mathcal{A}^s}[\log(1 - f(\psi(\alpha))) + \log f(\phi(p)) + \log f(\phi(n))] \tag{3.1}$$

$$\mathcal{L}_{class} = \mathbb{E}_{\alpha \in \mathcal{B}^s, p,n \in \mathcal{A}^s}[-y^+ \log h(\psi(\alpha)) - y^+ \log h(\phi(p)) - y^- \log h(\phi(n))] \tag{3.2}$$

The corresponding adversarial objective function is then:

$$\mathcal{L}_1 = \min_{\phi,\psi,h} \max_{f} \mathcal{L}_{dom}^{global} + \mathcal{L}_{class} \tag{3.3}$$

3.3.2 Local Adaptation

The cost function for the local adaptation ($\mathcal{L}_2$) is defined in Eq. 3.4 and it encourages the learning of abstract local spatial concepts common to $\mathcal{A}^s$ and $\mathcal{B}^s$. This is done by adversarially adapting the spatially average-pooled feature-map outputs of $\phi_{att}(p)$ and $\psi_{att}(\alpha)$. For example, if the output dimensionality of ψ_{att} and ϕ_{att} is $512 \times 7 \times 7$, we note that the pooled feature-map has a spatial resolution of 7×7 and each of the 49 cells summarizes the properties of different local regions of the input data.

$$\mathcal{L}_2 = \min_{\phi_{att},\psi_{att}} \max_{l} \mathbb{E}_{\alpha \in \mathcal{B}^s, p \in \mathcal{A}^s}[\log(1 - l(\psi_{att}(\alpha))) + \log l(\phi_{att}(p))] \tag{3.4}$$

In the adversarial setup, $l(\cdot)$ denotes the binary discriminator while ϕ_{att} and ψ_{att} represent the feature encoders, respectively. Since α and p share the category label, the domain labels are considered to be 1 for sketches and 0 for images.

3.3.3 Cross-modal Reconstruction

Although $\mathcal{L}_2$ adapts the intermediate feature maps of $\mathcal{A}^s$ and $\mathcal{B}^s$, however, it does not guarantee that the feature maps of a given modality are aware of the final latent feature distributions for the other modality. In order to better equip the feature maps with the cross-modal information, we introduce the classwise cross-modal reconstruction loss using generative modelling. Specifically, the cross-modal encoder-decoder modules $\mathcal{V}_\alpha = (\mathcal{V}_\alpha^e, \mathcal{V}_\alpha^d)$ and $\mathcal{V}_p = (\mathcal{V}_p^e, \mathcal{V}_p^d)$ reconstruct the latent feature embedding of sketch anchor $\psi(\alpha)$ given the outcome of $\phi_{att}(p)$ and vice-versa. Both $\mathcal{V}_\alpha^e$ and $\mathcal{V}_p^e$ are designed to be stochastic encoders and their outputs follow the standard normal distributions as per the principles of variational learning. We define the respective loss functions as follows, where D_{KL} is the Kullback-Leibler divergence.

$$\mathcal{L}_{KL}(\mathcal{V}, F, x) = D_{KL}(q(\mathcal{V}(F(x)))\|\mathcal{N}(0,1)) \tag{3.5}$$

$$\mathcal{L}_{rec}^1 = \mathcal{V}_p(\phi_{att}(p)) - \psi(\alpha)^2 + \mathcal{L}_{KL}(\mathcal{V}_p^e, \phi_{att}, p) \tag{3.6}$$

$$\mathcal{L}_{rec}^2 = \mathcal{V}_\alpha(\psi_{att}(\alpha)) - \phi(p)^2 + \mathcal{L}_{KL}(\mathcal{V}_\alpha^e, \psi_{att}, \alpha) \tag{3.7}$$

$$\mathcal{L}_3 = \min_{\mathcal{V}_p, \mathcal{V}_\alpha, \phi, \psi} \mathbb{E}_{\alpha \in \mathcal{B}^s, p \in \mathcal{A}^s}[\mathcal{L}_{rec}^1 + \mathcal{L}_{rec}^2] \tag{3.8}$$

The overall domain loss function is written as,

$$\mathcal{L}_{domain} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \tag{3.9}$$

3.4 Cross modal triplet loss

While the domain losses contribute towards making the two domains highly indistinguishable in the latent space, we simultaneously introduce a cross-modality triplet loss function for the triads $\{(\psi(\alpha), \phi(p), \phi(n))\}$ to ensure that the images and sketches form class-wise dense clusters in the latent space. $\mathcal{L}_{class}$ alone may not be able to guarantee this, considering the high intra-class variance for both the modalities and a high degree of data imbalance inherent to the task itself. Maintaining a margin among the class boundaries helps in combating both the problems.

By definition, the cross-modal triplet loss $\mathcal{L}_{triplet}$ aims to bring the same class sample $\phi(p)$ from the image modality closer to a given sketch anchor $\psi(\alpha)$ while pushing the

negative image sample $\phi(n)$ far from $\psi(\alpha)$ at least by a fixed margin of μ as per Eq. 3.10 using the Euclidean distance metric D.

$$\mathcal{L}_{triplet} = \min_{\phi,\psi} \mathbb{E}_{\alpha \in \mathcal{B}^s, p, n \in \mathcal{A}^s} [\max\{0, \mu + D(\psi(\alpha), \phi(p)) - D(\psi(\alpha), \phi(n))\}] \tag{3.10}$$

We note that the selection of triads $\{(\alpha, p, n)\}$ is carried out carefully in order to avoid any learning bias. Since we aspire to construct a dense cluster in the latent space for each of the classes in $\mathcal{C}^s$, sketches and images of the same class which are far from each other should be brought closer. Hence, we ensure to continuously select a p which is away from α among the nearest neighbors in the image feature space. Similarly while selecting n, we ensure to select the pairs $\{(\alpha, n)\}$ which are close in the feature space but share different annotations. We integrate this strategy with offline random sampling to construct the pool of triads.

3.5 Semantic Loss

The semantic side information is obtained by taking various combinations of text-based and hierarchical word embeddings for the category names. For the distributed word-vector models, we consider the pre-trained Word2Vec [19], GloVe [21] , and fasttext [2] while the Jiang-Conrath [15] and path similarity are utilized as the hierarchical encoding.

We project the semantic prototypes $\mathcal{W}^s$ together with the topology information of the original semantic space to the shared latent space. The topology information, which is found to bring in a regularization effect into the latent space, is encapsulated in the weighted semantic adjacency matrix $\Gamma_{|\mathcal{C}^s| \times |\mathcal{C}^s|}$ defined in terms of the pairwise cosine dissimilarity among the semantic prototypes of the seen classes. The joint information is found to perform better zero-shot inference than the semantic prototypes alone. Ideally, the outputs of $g_1(\mathcal{W}^s)$ and $g_2(\Gamma, \mathcal{W}^s)$ are concatenated and subsequently projected onto the latent space by another MLP $g_3(\cdot)$: $g(\mathcal{W}^s, \Gamma) = g_3([g_1(\mathcal{W}^s), g_2(\Gamma, \mathcal{W}^s)])$ where $[\cdot, \cdot]$ defines the vector concatenation operation. The semantic reconstruction loss brings $\phi(p)$ and $\psi(\alpha)$ closer to the projected class embedding $g(w^+, \Gamma)$ while maximizing the divergence between $\phi(n)$ and $g(w^+, \Gamma)$. This is accomplished through the semantic loss $\mathcal{L}_{semantic}$ as follows:

$$\mathcal{L}_{semantic} = \min_{g,\phi,\psi} \mathbb{E}[S(\psi(\alpha), g(w^+, \Gamma), 1) + S(\phi(p), g(w^+, \Gamma), 1) + S(\phi(n), g(w^+, \Gamma), 0)] \tag{3.11}$$

where the distance S between the vectors $(\mathbf{x}, \mathbf{y})$ is defined in terms of the cosine distance for a given threshold t as, $S(\mathbf{x}, \mathbf{y}, t) = \frac{1}{2}(t - \frac{\mathbf{x}\mathbf{y}^T}{\mathbf{x}\,\mathbf{y}})$.

The overall objective function for the SketRet framework can now be put forward as:

$$\mathcal{L}_{total} = \mathcal{L}_{domain} + \mathcal{L}_{triplet} + \mathcal{L}_{semantic} \tag{3.12}$$

Chapter 4

Experiments

4.1 Datasets

We validate the efficacy of the SketRet framework by performing experiments on the benchmark Sketchy-extended [23], TU Berlin-extended (TUB) [11], and the newly introduced QuickDraw-extended [6] datasets, respectively. The Sketchy dataset consists of 125 different categories of unpaired sketch and photo images. We use the two conventional train-test splits on the Sketchy dataset. In the first split (S1) we randomly select 25 classes as the unseen test data, while in the second split (S2) we use $|\mathcal{C}^s| : |\mathcal{C}^u| = 104 : 21$ as mentioned in [27] where the 21 unseen classes are carefully chosen not to be part of ImageNet [5]. The TU-Berlin dataset, on the other hand, contains 250 different classes of images and sketches and $|\mathcal{C}^s| : |\mathcal{C}^u| = 220 : 30$ is considered randomly such that the classes in $\mathcal{C}^u$ contain at least 400 images each. Finally, the large-scale QuickDraw dataset has samples from 110 classes and $|\mathcal{C}^s| : |\mathcal{C}^u| = 80 : 30$ is considered randomly.

4.2 Evaluation Protocol

We select around 5000 triplets in each training iteration based on the aforementioned triplet-mining protocol. $\mu = 1$ is selected for the triplet loss in Eq. 3.10. We use batch-normalization and leaky-ReLU non-linearity after each of the newly introduced layers to ensure a stable training. $\mathcal{L}_{total}$ is optimized using the stochastic gradient descent (SGD) with momentum as the optimizer with a mini-batch size of 32. An initial learning rate of 0.0001 and a momentum of 0.9 are set. We find that $\mathcal{L}_{total}$ converges for all the datasets within 50 iterations. We report the performance of

SketRet in terms of mAP@all (mean average precision), mAP@200, P@100, and P@200, respectively, where P stands for precision.

4.3 Implementation Details

The feature backbone networks $\phi(\cdot)$ and $\psi(\cdot)$ are the ImageNet pre-trained VGG-16 model [25]. Two modality specific spatial attention learning modules consisting of convolution kernels with sigmoid non-linearity are applied on the outputs of the final convolution layer (conv-5) of ϕ and ψ, and the networks upto the attention blocks are denoted as $\phi_{att}(\cdot)$ and $\psi_{att}(\cdot)$ each producing 512 feature maps of size 7×7. The attention blocks are eventually followed by three new dense layers which project the attended feature maps onto the final latent space with dimensions $\mathbb{R}^{256}$. Besides, a spatial average pooling across the channels is applied on the outputs of (ϕ_{att}, ψ_{att}) to obtain a single channel feature map of resolution 7×7. The encoder and decoder modules of $\mathcal{V}_\alpha$ and $\mathcal{V}_p$ are designed in terms of a single dense layer each with an encoder space dimensionality of 128.

The local domain classifier $l(\cdot)$, the global domain classifier $f(\cdot)$, and the multi-class category classifier $h(\cdot)$ are designed in terms of one dense layer each. As far as the semantic projection network $g(\cdot)$ is concerned, g_1 and g_3 are implemented in terms of three dense layers each. g_2 is defined considering a graph convolution layer, followed by the pooling and flattening layers. During training, VGG-16 parameters prior to conv-5 are frozen while the newly introduced layers are only updated.

Chapter 5

Results

5.1 Comparative Study

We choose the following state of the art methods, ZSIH [24], CVAE [27], SEM-PCYC [7], Doodle2search [6], Style-guide [9], AMD regularizer with SEM-PCYC and Style-guide [10], respectively, for analyzing our performance. Similar to SketRet, these techniques report their performances using the VGG-16 [25] based feature backbone networks. While SEM-PCYC, and Style-guide are based on adversarial training, CVAE and ZS-SBIR utilizes variational encoder-decoder networks. AMD regularizer can be used together with any of the ZS-SBIR methods and it helps in tackling the data/class imbalance between the training and test sets. The performance of our full SketRet model on the Sketchy dataset with its two splits, the Tu-Berlin dataset and Quickdraw-extended datasets in comparison to the state of the art is reported in Table 5.1.

For Sketchy, split 2 is considered to be more difficult than split 1 as it consists of test classes which are unseen to the ImageNet pre-trained networks. Amongst the other competing techniques, we find the inclusion of AMDreg boosts the performance of the baseline ZS-SBIR systems. For example, SEM-PCYC + AMDreg (39.7) and Style-guide + AMDreg (33.0) are found to be superior than the standalone SEM-PCYC (34.9) and Style-guide (37.6). In spite of this, SketRet beats SEM-PCYC + AMD by a margin of $4\% - 6\%$ on the mAP value. We achieve a mAP value of 43.7 and 43.5 for split 1 and 2 in this regard.

The TU-Berlin dataset is challenging mainly due to the presence of classwise as well as domain-wise data imbalance. Similar to Sketchy, AMDReg is found to boost the performance of SEM-PCYC and Style-guide on TU-Berlin with mAP values of 33.0

Model	Sketchy-ext (S2) mAP	P@ 100	P@ 200	mAP @200	Sketchy-ext (S1) mAP	P@ 100	TU Berlin-ext mAP	P@ 100	Quickdraw-ext mAP	P@ 100
ZSIH [24]	25.4	34.0	-	-	-	-	22.0	29.1	13.1	18.8
CVAE [27]	19.6	-	33.3	22.5	-	-	00.5	-	00.3	-
ZS-SBIR [27]	-	-	-	-	19.6	28.4	00.5	00.1	00.6	00.1
SEM-PCYC [7]	-	-	37.0	45.9	34.9	46.3	29.7	42.6	17.7	25.5
Doodle2search [6]	36.9	-	37.0	46.1	-	-	10.9	-	07.5	-
Style-guide [9]	-	-	40.0	35.8	37.6	48.4	25.4	35.5	-	-
Style-guide+AMDReg [10]	-	-	-	-	41.0	51.2	29.1	37.6	-	-
SEM-PCYC+AMDReg [10]	-	-	-	-	39.7	49.4	33.0	47.3	-	-
SketRet (Ours)	**43.5**	**51.2**	**45.8**	**55.6**	**43.7**	**51.4**	**36.8**	**51.1**	**21.6**	**36.1**
ZS-SBIR [27]	-	-	-	-	14.6	19.0	00.3	00.1	00.2	00.1
SEM-PCYC [7]	-	-	-	-	30.7	36.4	19.2	29.8	14.0	22.1
SEM-PCYC+AMDReg [10]	-	-	-	-	32.0	39.8	**24.5**	30.3	-	-
Style-guide [9]	-	-	-	-	33.0	38.1	14.9	22.6	-	-
GZS-SketRet (Ours)	**22.7**	**25.1**	**22.6**	**33.7**	**33.8**	**41.3**	22.7	**38.1**	**15.4**	**28.6**

Table 5.1: Comparing our SketRet with the state of the art on ZS-SBIR (top) and GZS-SBIR (bottom) on both the splits of Sketchy-extended, TU Berlin-extended and Quickdraw-extended datasets. All models use VGG-16 feature backbone. The '-' represents the evaluation metrices which were not mentioned in the respective papers. The performances are reported in terms of mAP, mAP@200, P@100, and P@200, respectively, where P stands for precision. S1 and S2 are splits 1 and 2.

and 29.1, respectively. The performance of other competing techniques are extremely low, for example, Doodle2search produces a mAP value of 10.9. In contrast, we achieve a mAP value of 36.8 for the TU-Berlin dataset with a boost of 4% over the existing literature.

The QuickDraw dataset is excessively large-scale consisting of highly ambiguous sketches and is by far the most challenging dataset for the ZS-SBIR task. Here, the mAP scores achieved by ZSIH (13.1), Doodle2search (7.5), and SEM-PCYC (17.7) are extremely low. SketRet is able to further improve the state-of-the art performance by reporting a mAP score of 21.6, which is 4% higher than the aforesaid values.

Similar to ZS-SBIR, SketRet showcase overall improved performance measures for GZS-SBIR for all the datasets. While we observe a marginal degradation in mAP value for TU-Berlin (22.7) than SEM-PCYC + AMDReg (24.5), we are able to outperform the comparative techniques in all the other performance metrics. In particular, SketRet produces high P@100 values of 41.3 for Sketchy (split 1), 38.1 for TU-Berlin, and 28.6 for QuickDraw which are at least 2.5 % more than the nearest techniques from the literature. No prior approach report the GZS-SBIR score for split 1 for Sketchy yet. However, we find that our performance in this case is substantially high with a mAP value of 22.7.

5.2 Evaluating Effect of input modalities

We evaluate the effects of the semantic information and the visual feature backbones as the input modalities.

5.2.1 Effect of Semantic Information

In the ZS-SBIR setup, the significance of the semantic information is imperative in maneuvering the alignment of the multi-modal data in the latent space. Different models yield different topological alignment of the classes in the latent space, which effectively causes the similar classes to cluster in a short range, while pushing apart the faraway classes. For example, the embeddings of `armour` and `axe` are very close to each other in the Word2Vec space as it bags both the classes under the super class of metals, while they are far in fasttext.

We consider the individual textual (300-d) and hierarchical embeddings as well as their concatenations and report the mAP values in Table 5.2 for both Sketchy and TU-Berlin. We observe that there is a variation of up to 4%, ranging from 36.3 to 43.5 in Sketchy and 28.4 to 36.8 in TU-Berlin, in the performance of SketRet by using different semantic information. Among the individual semantic spaces, GloVe produces inferior results for both the datasets relatively. We obtain the best performance of Sketchy using the fasttext model with a mAP of 43.5, while for TU Berlin the Jiang-Conrath produces the best performance with a mAP of 36.8. It is found that the individual semantic spaces provide superior performance than their pairwise combinations. Since the neighborhood topology may not be consistent in different semantic spaces, the topology may vary abruptly when we concatenate them, leading to a slight degradation in the retrieval performance.

5.2.2 Effect of Visual Features

Similar to the semantic information, the chosen visual feature encoder affects the model performance considerably. Different backbone networks have been utilized by a few existing techniques in the literature for ZS-SBIR. We feel it is slightly unjust to directly compare them with the rest of the literary works which exploit the conventional VGG-16 framework to maintain a fair comparison. Hence in this subsection, we deploy different encoder networks to train SketRet and compare with

W2v	Glove	Fasttext	Path	Jin-Con	Sketchy (S2)	TUB	QuickDraw
✓					41.1	32.6	**21.6**
	✓				39.3	28.4	–
		✓			**43.5**	34.6	16.2
			✓		40.2	36.0	18.1
				✓	41.9	**36.8**	16.8
✓			✓		40.1	32.3	17.6
✓				✓	40.7	33.6	15.7
	✓		✓		36.3	33.0	–
	✓			✓	37.6	32.5	–
		✓	✓		39.1	34.2	**21.1**
		✓		✓	39.7	33.7	14.4

Table 5.2: Effects of different semantic information on the mAP value for TU-Berlin and Sketchy (split 2) datasets.

the respective approaches from the literature (table 5.3) to provide a base-lining for the future endeavors.

SkechGCN [28] considers the ResNet-50 [12] architecture while SAN [20] utilizes the ResNet 152 [12] model, both pre-trained on the Imagenet dataset. It is noticed that SketRet with ResNet-152 performs quite poorly with a mAP@200 value of 43.0 for Sketchy and 16.0 for TU-Berlin, in comparison to SketRet with VGG-16, which gives a high mAP@200 value of 55.6 and 54.4 for both the datasets, respectively. The consideration of ResNet-50 feature backbone, on the other hand, achieves a comparable mAP of 40.1 for Sketchy. We also test the performance of our framework using the trending SE-Resnet-50 feature extractor. In this regard, SAKE [17] uses a conditional squeeze and excitation CSE-ResNet 50 [13] architecture, while also exploiting an auxiliary ImageNet dataset [5] to aid the training. It is worth noting at this point that SE-ResNet is different from CSE-ResNet as it does not use any conditional variable. In this work, we focus on those comparative frameworks which do not utilize additional auxiliary information apart from the concerned datasets in order to ensure a fair comparison. Hence, SAKE is not included for comparison purposes. Overall. it can be observed that SketRet beats the concerned techniques consistently when adopting the respective visual feature extractors.

	SketRet		State-of-the-Art		
Pretrain	Sketchy (S2)	TUB	Reference	Sketchy (S2)	TUB
VGG-16	43.5	36.8	—- Table 5.1 —-		
ResNet-50	40.1	33.2	SkechGCN [28]	38.2	32.4
ResNet-152	43.0*	16.0*	SAN [20]	24.0*	14.0*
SE-ResNet50	44.0*	23.8		-	-

Table 5.3: Comparison of different visual backbones on SketRet and the corresponding state-of-the-art on the TU-Berlin and Sketchy (split 2) datasets. Values are reported in terms of mAP and * denotes mAP@200. Relevant literary work using SE-ResNet-50 is not found.

5.3 Ablation study

The full model consists of a group of sub-modules, each contributing in its own way to enhance the performance. In the ablation analysis, the baseline network comprises of the $\mathcal{L}_{triplet}+\mathcal{L}_{semantic}$, while the full network contains $\mathcal{L}_{total}$, and we analyze the effects of adding binary domain classifier $\mathcal{L}_{dom}^{global}$, global adaptation loss $\mathcal{L}_1$, local adaptation loss $\mathcal{L}_2$, and the cross-modal reconstruction loss $\mathcal{L}_3$ into the base model. We report the results in terms of mAP values on Sketchy and TU-Berlin datasets (table 5.4).
The global adaptation is performed on the latent features to reduce the domain-gap between the data from the two modalities by increasing the domain confusion. This is expected to yield a class-wise overlapping embedding space for sketches and images. Simply adding the the binary domain classifier without the label classifier leads to mode collapse and the performance is at par with the baseline. To avoid this and maintain class-wise discriminativeness, we add the full $\mathcal{L}_1$ loss and observe an increase in the overall performance (33.4 / 22.5) to that of the baseline framework (table 5.4). We then append the network with the local adaptation module applied on the intermediate feature maps, wherein we seek to highlight important local constructs common to both the modalities. We see a further boost in the overall performance (36.9 / 26.1) of the network and at this stage the results are comparable to the state-of-the-art. When we further go on to add the cross-modal reconstruction modules, we observe significant improvements in the results (43.5 / 32.6). As evident from table 5.4, the full model incurs a boost of $12-13\%$ on the mAP values for both the

	Experimental set up	Sketchy (S2)	TUB	QuickDraw
Losses	$\mathcal{L}_{triplet} + \mathcal{L}_{semantic}$	27.6	20.1	3.7
	$\mathcal{L}_{triplet} + \mathcal{L}_{semantic} + \mathcal{L}_{dom}^{global}$	28.5	21.4	7.1
	$\mathcal{L}_{triplet} + \mathcal{L}_{semantic} + \mathcal{L}_1$	33.4	22.5	11.2
	$\mathcal{L}_{triplet} + \mathcal{L}_{semantic} + \mathcal{L}_1 + \mathcal{L}_2$	36.9	26.1	13.4
	$\mathcal{L}_{triplet} + \mathcal{L}_{semantic} + \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$ ($\mathcal{L}_{total}$)	**43.5**	**32.6**	21.6
Model	F w/o GCN	41.2	31.1	18.8
	F w/o attention block in local DA	38.6	26.2	14.3
	F w/o local attention & GCN	34.2	23.7	13.8
	F w/o GCN (GZS-SBIR)	13.4	14.3	10.3

Table 5.4: Ablation of loss functions and model components in terms of mAP for both Sketchy (split 2) and TU-Berlin. Here F denotes the full model.

datasets than the baseline, ranging from 27.6 to 43.5 in the Sketchy and 20.1 to 32.6 in the TU-Berlin.

Further, we study the effects of the graph CNN module in $g(\cdot)$ and the spatial attention layers in ϕ_{att} and ψ_{att}, respectively. We observe a marginal performance drop of around $1-2\%$ from 32.6 mAP to 31.1 in TU-Berlin and 43.5 mAP to 41.2 in Sketchy, when the GCN layer is removed from the full SketRet. Similarly, the attention module is crucial in highlighting the domain-invariant mid-level features and SketRet without the attention layers is found to marginally degrade the performance. We notice a fall from 43.5 to 38.6 in the Sketchy (split 2) and 32.6 to 26.2 in the TU-Berlin, which is nearly $5-6\%$ drop, confirming their relevance. We also investigate the effect of removing the graph CNN module for the GZS-SBIR experiments and notice a drop in the mAP values to 13.4 and 14.3 in the Sketchy and TU-Berlin datasets, respectively.

5.4 Study of Hubness

In this section, we aim to show the role of the domain losses in retrieving more discriminative image samples given the sketch queries. We first train the model using $\mathcal{L}_{triplet} + \mathcal{L}_{semantic} + \mathcal{L}_1$, followed by training the entire model with all the domain loss functions. In the first case, we notice the presence of hubs. Precisely, the left column of Figure 5.1 shows a scenario where the instances of `rifle` class are retrieved for nearly many query samples as the embeddings of the `rifle` class are cluttered in the feature space with multiple classes, like `giraffe`, `tree`, and `windmill`. When we look at the quantitative results and study the query sample wise precision scores, it

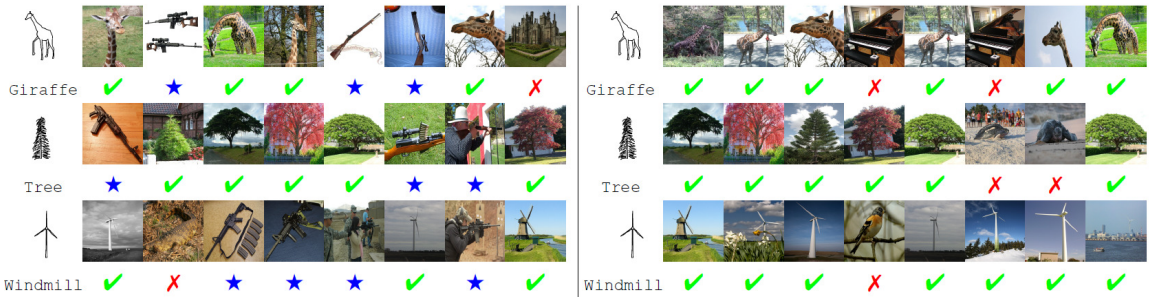

Figure 5.1: Left hand column shows the top-8 retrieval instances for a few sketch queries from the Sketchy dataset using $\mathcal{L}_{semantic} + \mathcal{L}_{triplet} + \mathcal{L}_1$ model. The green checks denote correctly retrieved classes, while the red crosses denote images from incorrect class. The blue stars denote the hub instances occurring repeatedly from a particular common class for most of the classes (`rifle` in this case). The right hand column shows the top-8 retrieval images for the same query sample using the full model. Notice that there are no hub instances generated here.

is noticed that while a few queries of those classes do yield high values (upto 40% P@100), others fail miserably and end up retrieving a large number of hubs. This results in a considerable amount of samples giving $< 10\%$ P@100 scores.

We further repeat the same analysis for the query samples when the full SketRet model is trained. In this case, we observe that the precision scores in terms of both P@100 and P@200 are uniformly distributed over all the samples for different classes. No particular class is visibly found to clutter the retrieval results in the latent space. The right column of figure 5.1 depicts that by jointly using the global and local adaptations and using cross-modal reconstruction modules, we are able to achieve hub-free retrieval results for the same set of query instances. This establishes our claim that the notion of fine-grained domain adaptation helps in obtaining a more discriminative latent space to combat hubness and negative knowledge transfer judiciously.

5.5 Qualitative Results

In this section, we provide qualitative results by showing the zero-shot retrieved photo instances for a few sample sketches.

Figure 5.2: Top-15 retrieval instances for a few sketch queries from the Sketchy dataset using the full model. The green checks denote correctly retrieved classes, while the red crosses denote images from incorrect class. Notice that there are no hub instances generated here.

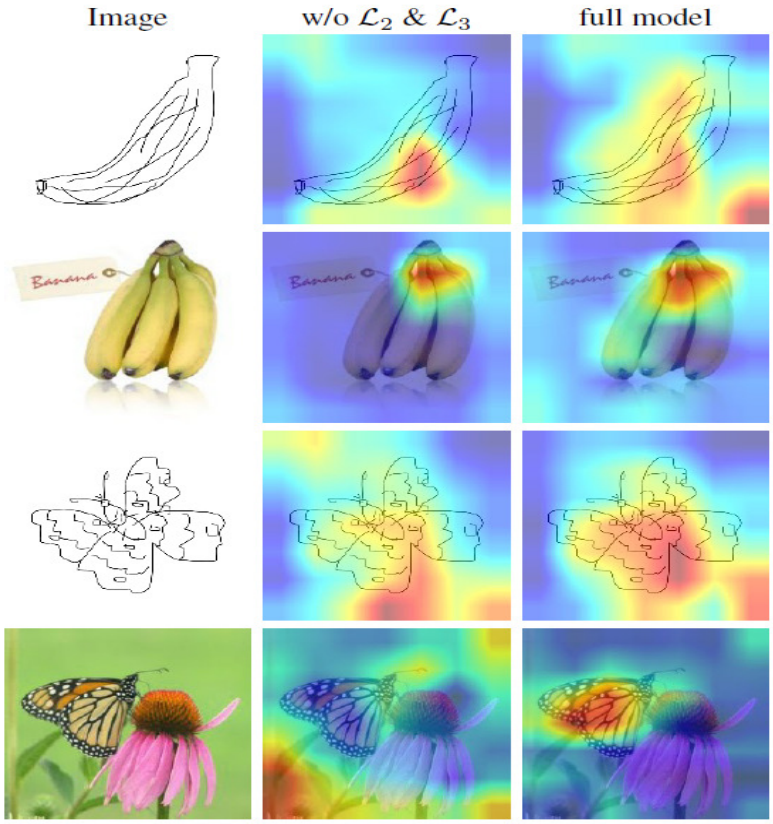

Figure 5.3: Grad-CAM plots highlighting the region of importance in a few sample sketch and photo images (left column) on the model trained without the L2 and L3 losses (middle column) and on the full model (right column).

5.6 Grad-CAM visualization

Gradient-weighted class activation mapping (Grad-CAM) primarily uses the gradients of the target class at the final convolution layer to synthesize an intermediate localization map which highlights the most important regions in the image. The Grad-CAM plots effectively helps in displaying the region which gets the most importance for any particular target-class. We train the model without the $\mathcal{L}_2$ and $\mathcal{L}_3$ losses and the full model and show the Grad-CAM plots of both the models for a few images in Fig. 5.3. We can see that the full model produces better highlight to the local constructs. Notice in the butterfly image, although the foreground consists of both the butterfly and the flower, the importance is layed properly on the concerned `butterfly` class. In Fig. 5.2, we show the zero-shot image retrieval results for a few sample sketches from the Sketchy-extended dataset.

Chapter 6

Conclusion

In this thesis, we discuss the methodology and results of a novel ZS-SBIR framework called SketRet. The main premise of this model is to perform improved alignment between the image and sketch features based on both the mid-level and high-level CNN based feature embeddings. Together, we introduce two generative cross-modal reconstruction modules to ensure the learning of robust modality-independent features. We further propose to project the semantic information into the shared latent space through a two-stream fusion network by jointly exploiting both the prototypes and the semantic class neighborhood. Overall, SketRet learns a discriminative and compact latent space and wisely tackles both the negative transfer and the hubness issues of domain adaptation and ZSL, respectively. Experimentally, we outperform the recent techniques in all the performance metrics on all the existing datasets.

Bibliography

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *ACL*, 2017.

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[7] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[8] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. *IJCV*, 2020.

[9] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. In *BMVC*, 2019.

[10] Titir Dutta, Anurag Singh, and Soma Biswas. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir. In *ECCV*, 2020.

[11] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[14] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 2013.

[15] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv*, 1997.

[16] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[17] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019.

[18] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, 1999.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[20] Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, and Hema Murthy. Stacked adversarial network for zero-shot sketch based image retrieval. In *WACV*, 2020.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[22] Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, 2014.

[23] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016.

[24] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

[26] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[27] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018.

[28] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Zero-shot sketch-based image retrieval via graph convolution network. In *AAAI*, 2020.