Adaptive Margin Diversity Regularizer for handling Data Imbalance in Zero-Shot SBIR

Titir Dutta, Anurag Singh, and Soma Biswas

Indian Institute of Science, Bangalore, India
{titird, anuragsingh2, somabiswas}@iisc.ac.in

Abstract. Data from new categories are continuously being discovered, which has sparked significant amount of research in developing approaches which generalizes to previously unseen categories, i.e. zero-shot setting. Zero-shot sketch-based image retrieval (ZS-SBIR) is one such problem in the context of cross-domain retrieval, which has received lot of attention due to its various real-life applications. Since most real-world training data have a fair amount of imbalance; in this work, for the first time in literature, we extensively study the effect of training data imbalance on the generalization to unseen categories, with ZS-SBIR as the application area. We evaluate several state-of-the-art data imbalance mitigating techniques and analyze their results. Furthermore, we propose a novel framework AMDReg (Adaptive Margin Diversity Regularizer), which ensures that the embeddings of the sketch and images in the latent space are not only semantically meaningful, but they are also separated according to their class-representations in the training set. The proposed approach is model-independent, and it can be incorporated seamlessly with several state-of-the-art ZS-SBIR methods to improve their performance under imbalanced condition. Extensive experiments and analysis justifies the effectiveness of the proposed AMDReg for mitigating the effect of data imbalance for generalization to unseen classes in ZS-SBIR.

1 Introduction

Sketch-based image retrieval (SBIR) [15][35], which deals with retrieving natural images, given a hand-drawn sketch query, has gained significant traction because of its potential applications in e-commerce, forensics, etc. Since new categories of data are continuously being added to the system, it is important for algorithms to generalize well to unseen classes, which is termed as Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) [6][5][16][7]. Majority of ZS-SBIR approaches learn a shared latent-space representation for both sketch and image, where sketches and images from same category come closer to each other and also incorporate additional techniques to facilitate generalization to unseen classes.

One important factor that has been largely overlooked in this task of generalization to unseen classes is the distribution of the training data. Real-world data, used to train the model, is not always class-wise or domain-wise well-balanced. When training and test categories are same, as expected, class imbalance in the training data results in severe degradation in testing performance, specially for the minority classes. Many seminal approaches have been proposed to mitigate this effect for the task of image classification [14][11][2][4], but the effect of data imbalance on generalization to unseen classes is relatively unexplored, both for single and cross-domain applications. In fact, both of the two large-scale datasets, widely used for SBIR/ ZS-SBIR, namely Sketchy Extended [25] and TU-Berlin Extended [8] have data imbalance. In cross-domain data, there can be two types of imbalance: 1) domain imbalance - where the number of data samples in one domain is significantly different compared to the other domain; 2) class imbalance - where there is a significant difference in the number of data samples per class. TU-Berlin Ext. exhibits imbalance of both types. Although a recent paper [5] has attributed poor retrieval performance for TU-Berlin Ext. to data imbalance, no measures have been proposed to handle this.

Here, we aim to study the effect of class imbalance in the training data on the retrieval performance of unseen classes in the context of ZS-SBIR, but interestingly we observe that the proposed framework works well even when both types of imbalances are present. We analyze several state-of-the-art approaches for mitigating the effect of training data imbalance on the final retrieval performance. To this end, we propose a novel regularizer termed **AMDReg - Adaptive Margin Diversity Regularizer**, which ensures that the embeddings of the data samples in the latent space account for the distribution of classes in the training set. To facilitate generalization to unseen classes for ZS-SBIR, majority of the ZS-SBIR approaches impose a direct or indirect semantic constraint on the latent-space which ensures that the sketch and image samples from *unseen* classes during testing are embedded in the neighborhood of its related *seen* classes. But merely imposing a semantic constraint does not account for the training class imbalance. The proposed AMDReg, which is computed from the class-wise training data distribution present in sketch and image domains helps to appropriately position the semantic embeddings. It tries to enforce a broader margin / spread for the classes for which less number of training samples are available as compared to the classes which have larger number of samples. Extensive analysis and evaluation on two benchmark datasets validate the effectiveness of the proposed approach. The contributions of this paper have been summarized below.

1. We analyze the effect of class-imbalance on generalization to unseen classes for the ZS-SBIR task. To the best of our knowledge, this is the first work in literature which addresses the data-imbalance problem in the context of cross-domain retrieval.
2. We analyze the performance of several state-of-the-art techniques for handling data imbalance problem for this task.
3. We propose a novel regularizer termed **AMDReg**, which can seamlessly be used with several ZS-SBIR methods to improve their performance. We have observed significant improvement in the performance of three state-of-the-art ZS-SBIR methods.
4. We obtain state-of-the-art performance for ZS-SBIR and generalized ZS-SBIR for two large-scale benchmark datasets.

2 Related Work

Here, we discuss relevant work in the literature for this study. We include recent papers for sketch-based image retrieval (SBIR), zero-shot sketch-based image retrieval (ZS-SBIR), as well as the class-imbalanced problems in classification.

Sketch-based Image Retrieval (SBIR): The primary goal of these approaches is to bridge the domain-gap between natural images and hand-drawn sketches. Early methods for SBIR, such as HOG [12], LKS [24] aim to extract hand-crafted features from the sketches as well as from the edge-maps obtained from natural images, which are then directly used for retrieval. The advent of deep networks have advanced the state-of-the-art significantly. Siamese network [22] with triplet-loss or contrastive-loss, GoogleNet [25] with triplet loss, etc. are some of the initial architectures. Recently a number of hashing-based methods, such as [15][35] have achieved significant success. [15] uses a heterogeneous network, which employs the edge maps from images, along with the sketch-image training data to learn a shared representation space. In contrast, GDH [35] exploits a generative model to learn the equivalent image representation from a given sketch and performs the final retrieval in the image space.
Zero-shot Sketch-based Image Retrieval (ZS-SBIR): The knowledge-gap encountered by the retrieval model, when a sketch query or database image is from a previously unseen class makes ZS-SBIR extremely challenging. ZSIH [26], generative-model based ZS-SBIR [32] are some of the pioneering works in this direction. However, as identified by [6], ZSIH [26] requires a fusion-layer for learning the model, which shoots up the learning cost and [32] requires strictly paired sketch-image data for training. Some of the recent works, [5][6][7][16] have reported improved performance for ZS-SBIR over the early techniques. [6] introduces a further generalization in the evaluation protocol for ZS-SBIR, termed as generalized ZS-SBIR; where the search set contains images from both the sets of seen and unseen classes. This poses even greater challenge to the algorithm, and the performances degrade significantly for this evaluation protocol [6][7]. Few of the ZS-SBIR approaches are discussed in more details later.
Handling data Imbalance for Classification: Since real-world training data are often imbalanced, recently, a number of works [14][11][2][4] have been proposed to address this problem. [14] mitigates the problem of foreground background class imbalance problem in the context of object recognition and proposes a modification to the traditional cross-entropy based classification loss. [4] introduces an additional cost-sensitive term to be included with any classification loss, designed on the basis of effective number of samples in a particular class. [2] and [11] both propose a modification in the margin of the class-boundary learned via minimizing intra-class variations and maximizing inter-class margin. [17] discusses a dynamic meta-embedding technique to address classification problem under long-tailed training data scenario.

Equipped with the knowledge of recent algorithms for both ZS-SBIR and single domain class imbalance mitigating techniques, we now move forward to discuss the problem of imbalanced training data for cross-domain retrieval.

3 Does Imbalanced Training Data Effect ZS-SBIR?

First, we analyze what is the effect of training data imbalance on generalization to unseen classes in the context of ZS-SBIR. Here, for ease of analysis, we consider only class imbalance, but our approach is effective for the mixed imbalance too, as justified by the experimental results later. Since both the standard datasets for this task, namely Sketchy Ext. [25] and TU-Berlin Ext. [8] are already imbalanced, to systematically study the effect of imbalance, we create a smaller balanced dataset, which is a subset of Sketchy Ext. dataset. This is termed as ***mini*-Sketchy Dataset** and contains sketches and images from 60 classes, with 500 images and sketches per class. Among them, randomly selected 10 classes are used as *unseen* classes and the rest 50 classes are used for training.

To study the effect of imbalance, motivated by the class-imbalance literature in image classification [14][11], we introduce two different types of class imbalance: 1) Step imbalance - where few of the classes in the training set contains less amount of samples compared to other classes; 2) Long-tailed imbalance - where the number of samples across the classes decrease gradually following the rule, $n_k^{lt} = n_k \mu^{\frac{k}{C_{seen}-1}}$; where n_k^{lt} is the available samples for k^{th} class under long-tailed distribution and n_k is the number of original samples of that class (=500 here). Here, $k \in \{1, 2, ..., C_{seen}\}$, i.e. C_{seen} is the number of training classes and $\mu = \frac{1}{p}$. We define imbalance factor p for a particular data-distribution to be the ratio of the highest number of samples in any class to the lowest number of samples in any class in that data and higher value of p implies more severe training class imbalance. Since the analysis is with class-imbalance, we assume that the data samples in image and sketch domain is the same.

As mentioned earlier, the proposed regularizer is generic and can be used with several baseline approaches to improve their performance in presence of data imbalance. For this analysis, we choose one recent auto-encoder based approach [7]. We term this as *Baseline Model* for this discussion, since the analysis is equally applicable for other approaches as well. We systematically introduce both the step and long-tailed imbalances for two different values of p and observe the performance for each of them. The results are reported in Table 1.

As compared to the balanced setting, we observe significant degradation in performance of the baseline whenever any kind of imbalance is present in the training data. This implies that training data imbalance not only effects the test performance when the classes remain the same, it also adversely effects the generalization performance significantly. This is due to the fact that unseen classes are recognized by embedding them close to their semantically relevant seen classes. Data imbalance results in (1) latent embedding space which is not sufficiently discriminative and (2) improperly learnt embedding functions, both of which negatively affects the embeddings of the unseen classes. The goal of the proposed AMDReg is to mitigate these limitations, which in turn will help in better generalization to unseen classes (Table 1 bottom row). Thus we see, that if the imbalance is handled properly, it may reduce the need for collecting large-scale balanced training samples.

Table 1. Evaluation (MAP@200) of Baseline Model [7] for ZS SBIR on *mini-Sketchy* dataset. Results for long-tailed and step imbalance with different imbalance factors are reported. The final performance using the proposed AMDReg is also compared.

Experimental Protocol	Balanced data	Imbalanced Data			
		Long-tailed		Step	
		$p = 10$	$p = 100$	$p = 10$	$p = 100$
Baseline [7]	0.395	0.234	0.185	0.241	0.156
Baseline [7] + AMDReg		**0.332**	**0.240**	**0.315**	**0.218**

4 Proposed Approach

Here, we describe the proposed Adaptive Margin Diversity Regularizer (AMDReg), which when used with existing ZS-SBIR approaches can help to mitigate the adverse effect of training data imbalance. We observe that majority of the state-of-the-art ZS-SBIR [6][16][7] approaches have two objectives: (1) projecting the sketches and images to a common discriminative latent space, where retrieval can be performed; (2) to ensure that the latent space is semantically meaningful so that the approach generalizes to unseen classes. For the first objective, a classification loss is used while learning the shared latent-space, which constraints the latent-space embeddings of both sketches and images from same classes to be clustered together, and samples from different classes to be well-separated. For the second objective, different direct or indirect techniques are utilized to make the embeddings semantically meaningful to ensure better generalization.

Semantically Meaningful Class Prototypes: Without loss of generality, we again chose the same baseline [7] to explain how to incorporate the proposed AMDReg into an existing ZS-SBIR approach. Let us consider that there are C_{seen} number of classes present in the dataset, and d is the latent space dimension. The baseline model has two parallel branches $F_{im}(\theta_{im})$ and $F_{sk}(\theta_{sk})$ for extracting features from images and sketches, $\{f^{(m)}\}$, where $m \in \{im, sk\}$, respectively. These features are then passed through corresponding content encoder networks to learn the shared latent-space embeddings for the same, i.e. $z^{(m)} = E_m(f^{(m)})$. In [7], a distance-based cross-entropy loss is used to learn these latent embeddings such that the embeddings is close to the semantic information. As is widely used, the class-name embeddings $h(y)$ of the *seen*-class labels $y \in \{1, 2, ..., C_{seen}\}$ are used as the semantic information. These embeddings are extracted from a pre-trained language model, such as, word2vec [18] or GloVe [20]. Please refer to Fig. 1 for illustration of the proposed AMDReg with respect to this baseline model.

The last fully connected (fc) layer of the encoders is essentially the classification layer and the weights of this layer, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{C_{seen}}], \mathbf{p}_i \in \mathbb{R}^d$ can be considered as the shared class-prototypes or the representatives of the corresponding class [21]. To ensure a semantically meaningful latent representation, one can learn the prototypes ($\mathbf{p}_i$'s) such that they are close to the class-name embeddings, or the prototypes can themselves be set equal to the semantic embeddings, i.e. $\mathbf{p}_i = h(y)$ and kept fixed. If the training data is imbalanced, just ensuring that the prototypes are semantically meaningful is not sufficient, we should also ensure that they take into account the label distribution of the training data. In our modification, to be able to adjust the prototypes properly, instead of fixing them as the class-embeddings, we initialize them using these attributes. Since the output of this fc layer is given by $\mathbf{z}^{(m)} = [z_1^{(m)}, z_2^{(m)}, ..., z_{C_{seen}}^{(m)}]$; the encoder with the prototypes is learnt using standard cross-entropy loss as,

$$\mathcal{L}_{CE}(\mathbf{z^{(m)}}, y) = -\log \frac{exp(z_y^{(m)})}{\sum_{j=1}^{C_{seen}} exp(z_j^{(m)})} \tag{1}$$

Now, with this as the background, we will describe the proposed regularizer, AMDReg, which ensures that the prototypes are modified in such a way that they are spread out according to their class representation in the training set.

Adaptive Margin Diversity Regularizer: Our proposed AMDReg is inspired from the recently proposed Diversity Regularizer [11], which addresses data imbalance in image classification by adjusting the classifier weights (here prototypes) so that they are uniformly spread out in the feature space. In our context, it can be enforced by the following regularizer

$$\mathcal{R}(\mathbf{P}) = \frac{1}{C_{seen}} \sum_{i<j} [||\mathbf{p}_i - \mathbf{p}_j||_2^2 - d_{mean}]^2, \ \forall j \in \{1, 2, ..., C_{seen}\} \tag{2}$$

Here d_{mean} is the mean distance between all the class prototypes and is computed as

$$d_{mean} = \frac{2}{C_{seen}^2 - C_{seen}} \sum_{i<j} ||\mathbf{p}_i - \mathbf{p}_j||_2^2, \ \forall j \in \{1, 2, ..., C_{seen}\} \tag{3}$$

The above regularizer tries to spread out all the class prototypes, without considering the amount of imbalance present in the training data. As has been observed in many recent works [2], due to insufficient number of samples of the minority classes, it is more likely that their test samples will have a wider spread instead of being clustered around the class prototype during testing. For our problem, this implies greater uncertainty for samples of unseen classes, which are semantically similar to the minority classes in the training set.

Towards this end, we propose to adjust the class prototypes adaptively, which takes into account the data imbalance. Since there can be both class and domain imbalance in the cross-domain retrieval problem, we propose to use the total number of sketch and image samples per class in the training set, and we refer to this combined number for k^{th}-class as the *effective* number of samples, n_k^{eff}, in this work. We then define the imbalance-based margin for the k^{th} class as,

$$\Delta_k = \frac{K}{n_k^{eff}} \tag{4}$$

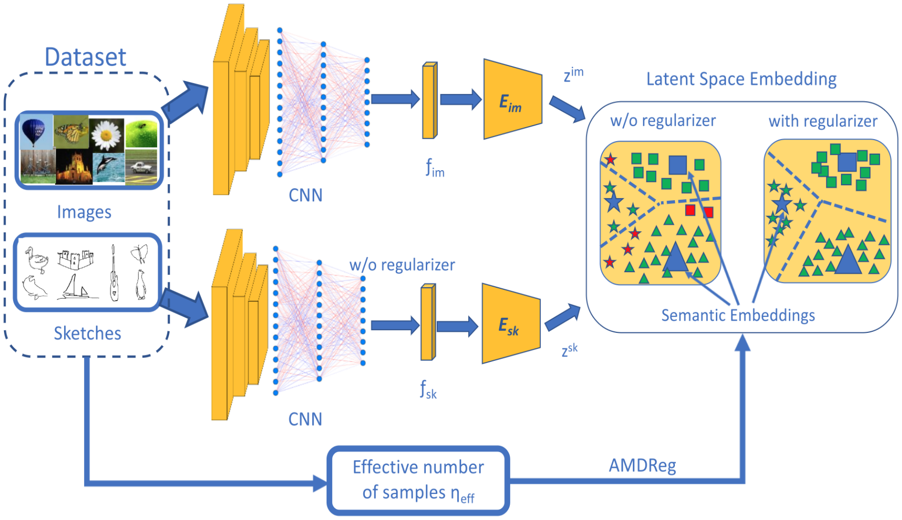

Fig. 1. Illustration of the proposed Adaptive Margin Diversity Regularizer (AMDReg). The AMDReg ensures that the embeddings of the shared prototypes of the images and sketches are not only placed away from each other, but also account for the increased uncertainty when the training class distribution is imbalanced. This results in better generalization to unseen classes.

This is similar to the inverse frequency of occurrence, except for the experimental hyper-parameter K. Thus the final AMDReg is given by

$$\mathcal{R}_{\Delta}(\mathbf{P}) = \frac{1}{C_{seen}} \sum_{i<j} [||\mathbf{p}_i - \mathbf{p}_j||_2^2 - (d_{mean} + \Delta_j)]^2, \ \forall j \in \{1, 2, ..., C_{seen}\} \quad (5)$$

Thus, we adjust the relative distance between $\mathbf{p}_i$'s such that they are atleast separated by a distance which is more than the mean-distance by the class imbalance margin. This ensures that the prototypes for the minority classes have more margin around them, which will reduce the chances of confusion for the semantically similar unseen classes during testing. Finally, the encoder with the prototypes are learnt using the CE loss along with the AMDReg as

$$\mathcal{L}_{CE}^{AMDReg} = \mathcal{L}_{CE} + \lambda \mathcal{R}_{\Delta} \quad (6)$$

where λ is an experimental hyper-parameter, which controls the contribution of the regularizer towards the learning.

Difference with Related Works: Even though the proposed AMDReg is inspired from [11], there are significant differences, namely (1) [11] addresses the imbalanced classification task for a single domain, while our work address generalization to unseen classes in the context of cross-domain retrieval (ZS-SBIR); (2) While [11] ensures that the weight vectors are equally spread out, AMDReg accounts for the training data distribution while designing the relative distances between the semantic embeddings; (3) Finally, [11] works with the max-margin loss, but AMDReg is used with the standard CE loss while learning the semantic class prototypes.

The proposed approach also differs from another closely related work LDAM [2]. LDAM loss is a modification on the standard cross-entropy or Hinge-loss to incorporate class-wise margin. In contrast, proposed AMDReg is a margin-based regularizer with adaptive margins between class-prototypes, based on the corresponding representation of classes in the training set. Thus, while [2] is inspired from margin-based generalization bound, the proposed AMDReg is inspired from the widely used inverse frequency of occurance.

4.1 Analysis with standard & SOTA imbalance-aware approaches

Here, we analyze how the proposed AMDReg compares with several existing state-of-the-art techniques used for addressing the problem of imbalance in the training data mainly for the task of image classification. These techniques can be broadly classified into two categories, (1) re-sampling techniques to balance the existing imbalanced dataset and (2) cost-sensitive learning or modification of the classifier. For this analysis also, we use the same retrieval backbone [7]. In this context, we first compute the average number of samples in the dataset. Any class which has lesser number of samples than the average are considered *minority* classes, and the remaining are considered *majority* classes.

1) **Re-balancing the dataset:** Re-sampling is a standard and effective technique used to balance out the dataset distribution bias. The most common methods are under-sampling of the majority classes [1] or over-sampling of minority classes [3]. We systematically use such imbalance data-sampling techniques on the training data to address the class imbalance for ZS-SBIR as discussed below. Here, the re-sampled / balanced dataset created by individual re-sampling operations described below is used for training the baseline network and reporting the retrieval performance.

1. ***Naive under-sampling:*** Here, we randomly select $1/p$-th of total samples per class for the majority classes and discard their remaining samples. Naturally, we loose a significant amount of important samples with such random sampling technique.
2. ***Selective Decontamination*** [1]: This technique is used to intelligently under-sample the majority classes instead of randomly throwing away excess samples. As per [1], we also modify the Euclidean distance function $d_E(\mathbf{x}_i, \mathbf{x}_j)$ between two samples of c^{th} class, $\mathbf{x}_i$ and $\mathbf{x}_j$ as,

$$d_{modified}(\mathbf{x}_i, \mathbf{x}_j) = (\frac{n_c}{N})^{(1/m)} d_E(\mathbf{x}_i, \mathbf{x}_j) \qquad (7)$$

   where n_c and N are the number of samples in c^{th} class and in all classes, respectively. m represents the dimension of the feature space. We retain only those samples in the majority classes for which the classes of majority of samples in top-K nearest neighbors agree completely.

3. ***Naive over-sampling:*** Here, the minority classes are augmented by repeating the instances (as in [35]) and using the standard image augmentation techniques (such as, rotation, translation, flipping etc.).
4. ***SMOTE*** [3]: In this intelligent over-sampling technique, instead of replacing the samples, the minority classes are augmented by generating *synthetic* features along the line-segment joining each minority class sample with its K-nearest neighbors.
5. ***GAN Based Augmentation*** [29]: Finally, we propose to augment the minority classes by generating features with the help of generative models, which have been very successful for zero-shot [29] / few-shot [19] / any-shot [30] image classification. Towards that goal, we use f-GAN [29] model to generate synthetic features for the minority classes using their attributes and augment those features with the available training dataset to reduce the imbalance.

2) **Cost-sensitive Learning of Classifier:** The goal of cost-sensitive learning based methods is to learn a better classifier using the original imbalanced training data, but with a more suitable loss function which can account for the data imbalance. To observe the effect of the different kinds of losses, we modify the distance-based CE-loss in the baseline model to the following ones, keeping the rest of the network fixed.

1. ***Focal loss:*** This loss [14] was proposed to address foreground-background class imbalance issue in the context of object detection. It is based on a simple yet effective modification of standard cross-entropy loss, such that while computation, the *easy* or well-classified samples are given less weights compared to the difficult samples.
2. ***Class-balanced Focal Loss:*** It is a variant of focal loss, recently proposed in [4], which incorporates the effective number of samples for a class in the imbalanced dataset.
3. ***Diversity Regularizer:*** This recently proposed regularizer [11] ensures that both the majority and minority classes are at equal distance from each other in the latent-space and reported significant performance improvement for imbalanced training data for image classification.
4. ***LDAM:*** [2] proposes a margin-based modification of standard cross-entropy loss or hinge loss, to ensure that the classes are well-separated from each other.

The retrieval performance obtained with these imbalance-handling methods are reported in Table 2. We observe that all the techniques result in varying degree of improvement over the base model. Among the data augmentation techniques, GAN-based augmentation outperforms the other approaches. In general, all the cost-sensitive learning techniques performs quite well, specially the recently proposed diversity regularizer and the LDAM cross-entropy loss. However, the proposed AMDReg outperforms both the data balancing and cost-sensitive learning approaches, giving the best performance across all types and degrees of imbalance.

Table 2. ZS-SBIR performance (MAP@200) of different kinds of imbalance handling techniques applied on the Baseline Model [7] for the *mini-Sketchy* dataset. Results of the original Baseline Model is also reported for reference.

Imbalance Handler	Methods	Long-tailed		Step	
		$p = 10$	$p = 100$	$p = 10$	$p = 100$
	Baseline Model [7]	0.234	0.185	0.241	0.156
Data balancing methods	Naive under-sampling	0.235	0.191	0.256	0.159
	Naive over-sampling	0.269	0.219	0.258	0.155
	Selective decontamination [1]	0.268	0.221	0.251	0.164
	SMOTE [3]	0.269	0.217	0.269	0.183
	GAN-based Augmentation [29]	0.305	0.229	0.274	0.188
Loss-Modification Techniques	Focal loss [14]	0.273	0.228	0.289	0.195
	Class-balanced Focal Loss [4]	0.299	0.236	0.296	0.210
	Diversity-Regularizer [11]	0.296	0.222	0.285	0.207
	LDAM-CE loss [2]	0.329	0.234	0.310	0.213
	Proposed AMDReg	**0.332**	**0.240**	**0.315**	**0.218**

5 Experimental Evaluation on ZS-SBIR

Here, we provide details of the extensive experiments performed to evaluate the effectiveness of the proposed AMDReg for handing data imbalance in ZS-SBIR.

Datasets Used and Experimental Protocol: We have used two large-scale standard benchmarks for evaluating ZS-SBIR approaches, namely, Sketchy Ext. [25] and TU-Berlin Ext. [8].

Sketchy Ext. [25] dataset originally contained approximately 75,000 sketches and 12,500 images from 125 object categories. Later, [15] collected and added additional 60,502 images to this dataset. Following the standard protocol [6][16], we randomly choose 25 classes as *unseen*-classes (sketches as query and images in the search set) and the rest 100 classes for training.

TU-Berlin Ext. [8] originally contained 80 hand-drawn sketches per class from total 250 classes. To make it a better fit for large-scale experiments, [34] included additional 2,04,489 images. As followed in literature [6] [7], we randomly select 30-classes as *unseen*, while the rest 220-classes are used for training.

The dataset statistics are shown in Fig. 2, which depicts data imbalance in both the datasets. This is specially evident in TU-Berlin Ext., which has huge domain-wise imbalance as well as class-wise imbalance. These real-world datasets reinforce the importance of handling data imbalance for the ZS-SBIR task.

5.1 State-of-the-art ZS-SBIR approaches integrated with AMDReg

As already mentioned, the proposed AMDReg is generic and can be seamlessly integrated with most state-of-the-art ZS-SBIR approaches for handling the training data imbalance. Here, we have integrated AMDReg with three state-of-the-art approaches, namely (1) Semantically-tied paired cycle-consistency based

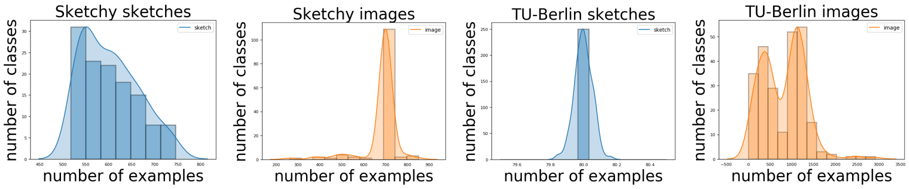

Fig. 2. Dataset statistics of sketches and images of Sketchy-extended and TU-Berlin-extended are shown in the first two and last two plots respectively in that order.

network (SEM-PCYC) [6]; (2) Semantic-aware knowledge preservation for ZS-SBIR (SAKE) [16]. (3) Style-guided network for ZS-SBIR [7]. Now, we briefly describe the three approaches along with the integration of AMDReg.

SEM-PCYC [6] with AMDReg: SEM-PCYC is a generative model with two separate branches for image and sketch; for visual-to-semantic mapping along with cyclic consistency loss. Further, to ensure that the semantic output of the generators is also class-discriminative, a classification loss is used. This classifier is pre-trained on *seen*-class training data and kept frozen while the whole retrieval model is trained. We modify the training methodology by enabling the classifier to train along with the rest of the model, by including the AMDReg with the CE-loss. Here, the semantic information is enforced through an auto-encoder, which uses a hierarchical and a text-based model as input, and thus the weights are randomly initialized. Please refer to [6] for more details.

SAKE [16] with AMDReg: This ZS-SBIR method extends the concept of domain-adaptation for fine-tuning a pre-trained model on ImageNet [23] for the specific ZS-SBIR datasets. The network contains a shared branch to extract features from both sketches and images, which are later used for the categorical classification task using the soft-max CE-loss. Simultaneously, the semantic structure with respect to the ImageNet [23] classes are maintained. Here also, we modify the CE-loss using the proposed AMDReg to mitigate the adverse effect of training data imbalance. The rest of the branches and the proposed SAKE-loss remain unchanged. Please refer to [16] for more details of the base algorithm.

Style-guide [7] with AMDReg: This is a two-step process, where the shared latent-space is learnt first. Then, the latent-space content extracted from the sketch query is combined with the styles of the relevant images to obtain the final retrieval in the image-space. While learning the latent-space, a distance-based cross-entropy loss is used, which is modified as explained in details earlier. Please refer to [7] for more details of the base algorithm.

Implementation Details The proposed regularizer is implemented using Pytorch. We use a single Nvidia GeForce GTX TITAN X for all our experiments.

Table 3. Performance of several state-of-the-art approaches for ZS-SBIR and generalized ZS-SBIR.

	Algorithms	TU-Berlin extended		Sketchy-extended	
		MAP@all	Prec@100	MAP@all	Prec@100
SBIR	Softmax Baseline	0.089	0.143	0.114	0.172
	Siamese CNN [22]	0.109	0.141	0.132	0.175
	SaN [33]	0.089	0.108	0.115	0.125
	GN Triplet [25]	0.175	0.253	0.204	0.296
	3D shape [28]	0.054	0.067	0.067	0.078
	DSH (binary) [15]	0.129	0.189	0.171	0.231
	GDH (binary) [35]	0.135	0.212	0.187	0.259
ZSL	CMT [27]	0.062	0.078	0.087	0.102
	DeViSE [10]	0.059	0.071	0.067	0.077
	SSE [36]	0.089	0.121	0.116	0.161
	JLSE [37]	0.109	0.155	0.131	0.185
	SAE [13]	0.167	0.221	0.216	0.293
	FRWGAN [9]	0.110	0.157	0.127	0.169
	ZSH [31]	0.141	0.177	0.159	0.214
Zero-Shot SBIR	ZSIH (binary) [26]	0.223	0.294	0.258	0.342
	ZS-SBIR [32]	0.005	0.001	0.196	0.284
	SEM-PCYC [6]	0.297	0.426	0.349	0.463
	SEM-PCYC + AMDReg	**0.330**	**0.473**	**0.397**	**0.494**
	Style-guide [7]	0.254	0.355	0.375	0.484
	Style-guide + AMDReg	**0.291**	**0.376**	**0.410**	**0.512**
	SAKE [16]	0.428*	0.534*	0.547	0.692
	SAKE + AMDReg	**0.447**	**0.574**	**0.551**	**0.715**
Generalized Zero-shot SBIR	Style-guide [7]	0.149	0.226	**0.330**	0.381
	SEM-PCYC [6]	0.192	0.298	0.307	0.364
	SEM-PCYC + AMDReg	**0.245**	**0.303**	0.320	**0.398**

For all the experiments, we set $\lambda = 10^3$ and $K = 1$. Adam optimizer has been used with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a learning rate of $lr = 10^{-3}$. The implementation of different baselines and the choice of hyper-parameters for their implementation has been done as described in the corresponding papers.

5.2 Evaluation for ZS-SBIR

Here, we report the results of the modifications to the state-of-the-art approaches for ZS-SBIR. We first train all the three original models (as described before) to replicate the results reported in the respective papers. We use the codes given by the authors and are able to replicate all the results for SEM-PCYC and Style-guide as reported. However, for SAKE, in two cases, the results we obtained are slightly different from that reported in the paper. So we report the results as we obtained, for fair evaluation of proposed improvement (marked with a star to indicate that they are different from the reported numbers in the paper).

We incorporate the proposed modifications for AMDReg in all three approaches and retrained the models. The results are reported in Table 3. All the

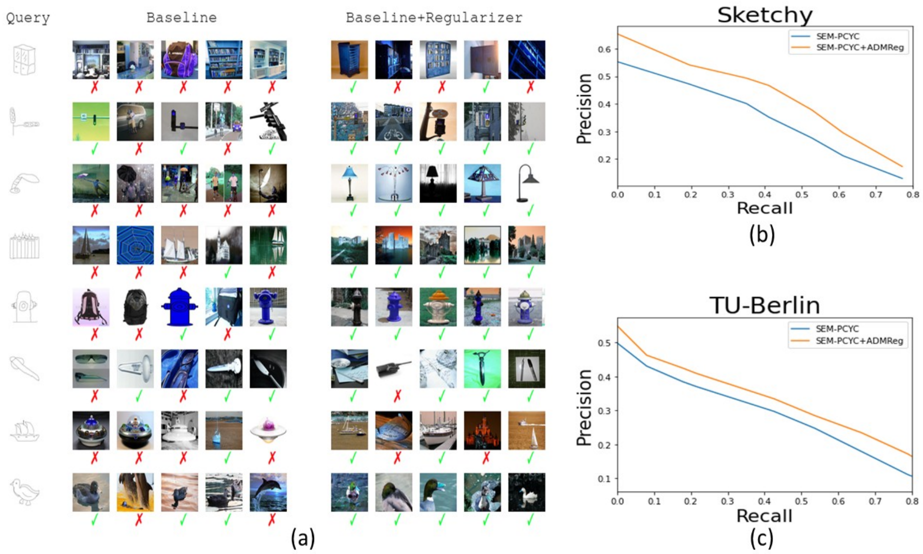

Fig. 3. Performance comparison of the base model (SEM-PCYC) and the modified base-model using proposed AMDReg: (a) Few examples of top-5 retrieved images against the given unseen sketch query from TU-Berlin dataset; (b) P-R curve on Sketchy dataset; (c) P-R curve on TU-Berlin dataset.

results of the other approaches are taken directly from [6]. We observe significant improvement in the performance of all the state-of-the-art approaches, when trained using the proposed regularizer. This experiment throws insight that by handling the data-imabalance, which is inherently present in the collected data, it is possible to gain siginificant improvement in the final performance. Since AMDReg is generic, it can potentially be incorporated with other approaches, developed for the ZS-SBIR task, to handle the training data imbalance problem.

Fig. 3 shows top-5 retrieved results for a few unseen queries (first column), using SEM-PCYC as the baseline model, without and with AMDReg, respectively. We observe significant improvement when AMDReg is used, justifying its effectiveness. We make similar observations from the P-R curves in Fig. 3.

5.3 Evaluation for Generalized ZS-SBIR

In real scenarios, the search set may consist of both the seen and unseen image samples, which makes the problem much more challenging. This is termed as the generalized ZS-SBIR. To evaluate the effectiveness of proposed AMDReg for this scenario, we follow the experimental protocol in [6] and SEM-PCYC [6] as the base model. From the results in Table 3, we observe that AMDReg is able to significantly improve the performance of the base model and yields state-of-the-art results for three out of the four cases. Only for Sketchy Ext., it performs slightly less than Style-Guide, but still improves upon its baseline performance.

5.4 Evaluation for SBIR

Though the main purpose of this work is to analyze the effect of training data imbalance on generalization to unseen classes, this approach should also benefit standard SBIR in presence of imbalance. We observe from Table 4, that the performance of SBIR indeed decreases drastically with training data imbalance. Proposed AMDReg is able to mitigate this by a significant margin as compared to the other state-of-the-art imbalance handing techniques. We further analyze the performance of SEM-PCYC [6] on Sketchy Ext. dataset for standard SBIR protocol with and without AMDReg. We observe significant improvement when proposed AMDReg is used (MAP@all: 0.811; Prec@100: 0.897) as compared to the baseline SEM-PCYC (MAP@all: 0.771; Prec@100: 0.871).

Table 4. SBIR evaluation (MAP@200) of Baseline Model [7] on *mini-Sketchy*.

Balanced Data	Step Imb. ($p = 100$)	GAN-based Aug. [29]	CB Focal Loss [4]	Diversity Regularizer [11]	**Proposed** **AMDReg**
0.839	0.571	0.580	0.613	0.636	**0.647**

6 Conclusion

In this work, for the first time in literature, we analyzed the effect of training data imbalance for the task of generalization to unseen classes in context of ZS-SBIR. We observe that most real-world SBIR datasets are in-fact imbalanced, and that this imbalance does effect the generalization adversely. We systematically evaluate several state-of-the-art imbalanced mitigating approaches (for classification) for this problem. Additionally, we propose a novel adaptive margin diversity regularizer (AMDReg), which ensures that the shared latent space embeddings of the images and sketches account for the data imbalance in the training set. The proposed regularizer is generic, and we show how it can be seamlessly incorporated in three existing state-of-the-art ZS-SBIR approaches with slight modifications. Finally, we show that the proposed AMDReg results in significant improvement in both ZS-SBIR and generalized ZS-SBIR protocols, setting the new state-of-the-art result.

Acknowledgement

This work is partly supported through a research grant from SERB, Department of Science and Technology, Government of India.

References

1. Barandela, R., Rangel, E., Sanchez, J.S., Ferri, F.J.: Restricted decontamination for the imbalanced training sample problem. Iberoamerican Congress on Pattern Recognition, Springer (2003)
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. NeurIPS (2019)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y.: Class-balanced loss based on effective number of samples. CVPR (2019)
5. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: practical zero-shot sketch-based image retrieval. CVPR (2019)
6. Dutta, A., Akata, Z.: Sematically tied paired cycle consistency for zero-shot sketch-based image retrieval. CVPR (2019)
7. Dutta, T., Biswas, S.: Style-guided zero-shot sketch-based image retrieval. BMVC (2019)
8. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM TOG **31**(4), 44.1–44.10 (2012)
9. Felix, R., Kumar, V.B., Reid, I., Carneiro, G.: Multi-modal cycle-consistent generalized zero-shot learning. In: ECCV (2018)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013)
11. Hayat, M., Khan, S., Zamir, S.W., Shen, J., Shao, L.: Gaussian affinity for max-margin class imbalanced learning. ICCV (2019)
12. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. CVIU **117**(7), 790–806 (2013)
13. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR (2017)
14. Lin, T.Y., Goyal, P., Girshiick, R., He, K., Dollar, P.: Focal loss for dense object detection. arXiv:1708.02002 [cs.CV] (2018)
15. Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: fast free-hand sketch-based image retrieval. CVPR (2017)
16. Liu, Q., Xie, L., Wang, H., Yuille, A.: Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. ICCV (2019)
17. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR (2019)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. NeurIPS (2013)
19. Mishra, A., Reddy, S.K., Mittal, A., Murthy, H.A.: A generative model for zero-shot learning using conditional variational auto-encoders. CVPR-W (2018)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. EMNLP (2014)
21. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. CVPR (2018)
22. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. ICIP (2016)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Li, F.F.: Imagenet: large-scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)

24. Saavedra, J.M., Barrios, J.M.: Sketch-based image retrieval using learned keyshapes (lks). BMVC (2015)
25. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG (2016)
26. Shen, Y., Liu, L., Shen, F., Shao, L.: Zero-shot sketch-image hashing. CVPR (2018)
27. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: NeurIPS (2013)
28. Wang, M., Wang, C., Wu, J.X., Zhang, J.: Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. VLDB (2015)
29. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. CVPR (2018)
30. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. CVPR (2019)
31. Yang, Z., Cohen, W.W., Salakhutdinov, R.: Revisiting semi-supervised learning with graph embeddings. arXiv preprint arXiv:1603.08861 (2016)
32. Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A.: A zero-shot framework for sketch-based image retrieval. ECCV (2018)
33. Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net that beats humans. BMVC (2015)
34. Zhang, J., Liu, S., Zhang, C., Ren, W., Wang, R., Cao, X.: Sketchnet: sketch classification with web images. CVPR (2016)
35. Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., ad L. V. Gool, H.T.S.: Generative domain-migration hashing for sketch-to-image retrieval. ECCV (2018)
36. Zhang, R., Lin, L., Zhang, R., Zuo, W., Zhang, L.: Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Transactions on Image Processing **24**(12), 4766–4779 (2015)
37. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. In: CVPR (2016)