Journal of Electrical Engineering & Technology (2024) 19:4587–4593
https://doi.org/10.1007/s42835-024-01889-6

ORIGINAL ARTICLE

Sketch Classification and Sketch Based Image Retrieval Using ViT with Self-Distillation for Few Samples

Sungjae Kang[1] · **Kisung Seo**[1]

Received: 23 October 2023 / Revised: 4 March 2024 / Accepted: 14 March 2024 / Published online: 25 March 2024

Abstract
Sketch-based image retrieval (SBIR) with Zero-Shot are challenging tasks in computer vision, enabling to retrieve photo images relevant to sketch queries that have not been seen in the training phase. For sketch images without a sequence of information, we propose a modified Vision Transformer (ViT)-based approach that enhances or maintains the performance while reducing the number of sketch training data. First, we add a token for retrieval and integrate auxiliary classifiers of multiple branches ViT network. Second, self-distillation is applied to enable fast transfer learning of sketch domains for our ViT network incorporating addition of classifiers and embedding vectors to each intermediate layers in the network. Third, to address the challenge of overfitting due to reduced input data pairs in training with large datasets, we integrate KL-Divergence, capturing distribution differences between sketches and photos, into the triplet loss, thereby mitigating the impact of limited sketch-photo samples. Experiments on the TU-Berlin and Sketchy dataset demonstrate show that our method performs a significant improvement over other similar methods on sketch classification and sketch-based image retrieval.

Keywords Sketch-based Image-Retrieval · Knowledge Distillation

1 Introduction

One definition of a sketch is a rough drawing that represents the main features of an object or idea, often produced by preliminary study. Sketch images can communicate information that is difficult to describe using text. There are various levels, ranging from the rough sketches like thumbnail sketch to precise ones like style sketches. Most sketch recognition studies have focused on rough levels. Sketch representation and interpretation remain open issues, especially in the context of sketch-based image retrieval (SBIR) [1–3], which has garnered significant attention. Research directions also include Zero-Shot SBIR [4, 5], where a model aims to generalize across disjoint training and test classes, reducing annotation costs. However, existing studies still require large amounts of sketch-photo pairs for training, and the availability of such paired data is relatively small. Approaches incorporating transformers for performance enhancement, such as SketchBert [2], Sketchformer [3], and TVT [6], are being attempted.

In this paper, we introduce a modified ViT [7] network to effectively learn a small number of sketch-photo pairs and enhance the performance of the Zero-Shot SBIR scheme. Inspired by DeiT [8], which added a distillation token to receive knowledge from a CNN-based teacher network, we introduce retrieval tokens to the existing ViT structure. These tokens provide additional cues for sketch-photo pairs to the model, effectively enhancing the model's learning and representation capabilities.

Recently, Transformer-based Self Knowledge Distillation techniques have been introduced [9, 10]. However, these methods also utilize the output of the last layer, hindering fast transfer learning through training in early layers. To solve the problem, we apply efficient self-distillation [11, 12] to enable fast transfer learning of sketch domains for a ViT network pre-trained with a large amount of RGB images. Unlike existing self-distillation methods, we incorporate classifiers and embedding vectors to each intermediate layer, not just the last layer, to improve performance.

To address overfitting problems associated with using few sketch-photo pairs with existing triplet loss, we apply

✉ Kisung Seo
ksseo@skuniv.ac.kr

Sungjae Kang
sungjae1132@skuniv.ac.kr

[1] Department of Electronics Engineering, Seokyeong University, Seoul, South Korea

KL-Divergence between sketch and photo to the triplet loss. The proposed KL-Divergence-based triplet loss provides Evidence of Lower Bound (ELBO), mitigating such problems. Additionally, we incorporate the Spherical loss [13], which integrates Euclidean distance and angular distance, and augment it with center loss [14] to mitigate inter-class variation problems. We perform joint training based on these loss functions.

We conduct an ablation study of the loss configuration and also perform an extensive search for the connection cases of self-distillation suitable for ViT structures. Our proposed method is evaluated on the TU-Berlin and Sketchy datasets and compared to similar approaches on sketch classification, sketch-based image retrieval, and zero-shot sketch-based image retrieval.

Our contributions are summarized as follows. First, integration of a retrieval token and auxiliary classifiers across multiple branches within the network to improve its architectural design. Second, implementation of self-distillation for the ViT network adding classifiers and embedding vectors to intermediate layers for enhanced performance. Third, incorporation of KL-Divergence into the triplet loss function to mitigate overfitting challenges arising from reduced input data.

2 Methodology

2.1 Background

In this section, we briefly describe ViT Transformer [7] which is the basis network of the proposed method. A transformer's encoder consists of several blocks. The operation process in a block is as follows

$$SSA_i\left(qW_i^q, kW_i^k, vW_i^v\right) = softmax\left(\frac{qW_i^q\left(kW_i^k\right)^T}{\sqrt{d_k}}\right)vW_i^v \tag{1}$$

$$\begin{aligned} MSA(q,k,v) = Concat(&SSA_1\left(qW_0^q, kW_0^k, vW_0^v\right), \ldots, \\ &SSA_h\left(qW_h^q, kW_h^k, vW_h^v\right))W^o \end{aligned} \tag{2}$$

where, q, k, v are the query, key, and value respectively, $qW_i^q, kW_i^k, vW_i^v \in \mathbb{R}^{d_{hidden} \times d_k}$, $d_k = \frac{d_{hidden}}{h}$, h is the number of head, these are used as a input for Single Head Self Attention (SSA). The SSA module calculates the similarity between image patches using a query and key. These SSA modules are configured by h and processed in parallel. This is called Multi-head Self Attention (MSA). The MSA concatenates the output of SSA modules, and $W^o \in \mathbb{R}^{d_{hidden} \times d_k}$.

$$MLP(z) = GELU\left(0, \bar{z}W_1^f + b_1^f\right)W_2^f + b_2^f \tag{3}$$

$$F(x) = MSA\left(\bar{x}, \bar{x}, \bar{x}\right) + x \tag{4}$$

$$B(x) = MLP\left(\bar{F}(x)\right) + F(x) \tag{5}$$

where, $\bar{x}$ is the normalized x MLP consists of two fully connected layers and one GELU activation function. ViT transformer has several blocks. After passing through all blocks, the output value is as follows.

$$E(x) = B_L(\ldots\left(B_L(x)\right)) \tag{6}$$

Here, L is the number of blocks. Finally, classification is performed by adding one fully connected layer to the result of Eq. 6.

2.2 Proposed Method

2.2.1 Modified Vision Transformer

The structure of the proposed model for sketch classification and retrieval is shown in Fig. 1. As mentioned earlier, ViT [7] has one token for performing classification in addition to tokens for N input data. We add another token to perform for retrieval as shown in the right of Fig. 1. Therefore, the total number of input tokens is $N + 2$. ViT has much less image-specific inductive bias than Convolutional Neural Network (CNN). However, the Transformer network requires a process of pre-training with a large dataset. Fortunately, an off-the-shelf model for performing our task can be found in DeiT [8]. DeiT used the Knowledge distillation (KD) [15] technique by adding a distillation token to the ViT structure. In [8], a CNN network that can perform pre-training more easily is used as a teacher, and the distilled knowledge is transferred to the Transformer network, which is a student. In consideration of our limited computing environment, we adopt the DeiT-small model.

2.2.2 Training Transformer via Self-Distillation

We performed transfer learning to use the pre-trained model for sketch task. Because pre-training is performed using photo images, whereas our task targets sketch images, thus, domain gap exists. Self-distillation techniques are used to enable the model to adapt it efficiently. In the self-distillation, the teacher model does not exist separately, unlike the standard KD method. Instead of it, the model distills knowledge within network itself. Among the existing self-distillation techniques, we are inspired by [11]. The network is divided into several sections. Then the knowledge in the deeper portion of the networks is squeezed into the shallow ones. During the training, each

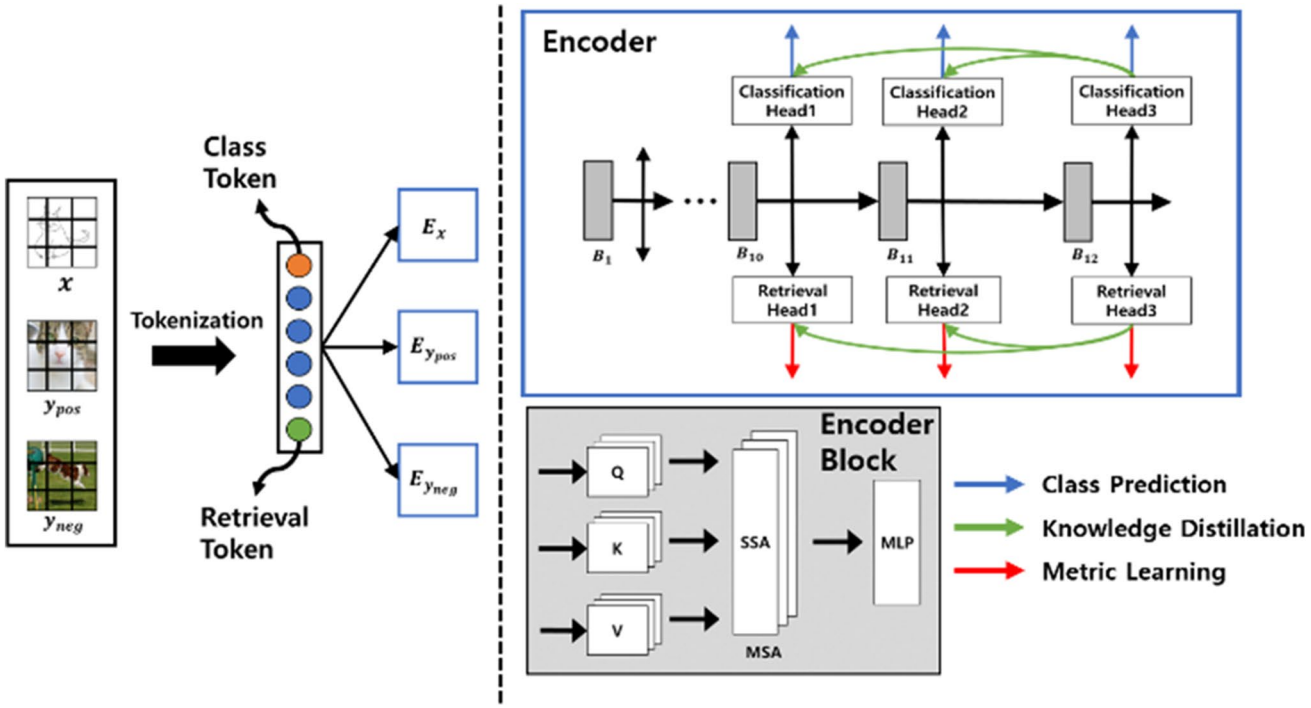

Fig. 1 The structure of ViT-based SBIR (Sketch-Based Image Retrieval) model

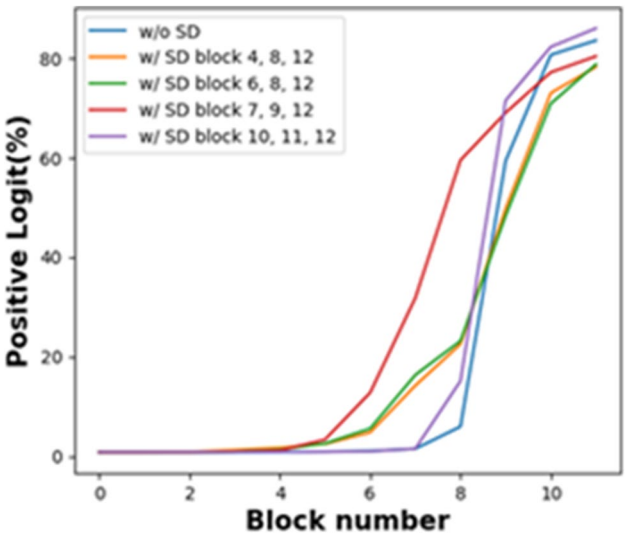

Fig. 2 Comparison of positive logits for each block in ViT for self-distillation

section of the shallow layer operates as a student, and the deepest layer acting as a teacher. We adopt the same mechanism for our ViT based network for self-distillation. The knowledge of the last encoder block is transferred to the former blocks. The KD loss for this is defined as follows (Fig. 2).

$$L_{soft} = p(\hat{z},\theta)\log\left(\frac{p(\hat{z},\theta)}{p(z^1,\theta)}\right) + p(\hat{z},\theta)\log\left(\frac{p(\hat{z},\theta)}{p(z^2,\theta)}\right) \quad (7)$$

$$L_{sim} = \frac{1}{d}(\|f^1(\epsilon) - \hat{f}(\epsilon)\|^2 + \|f^2(\epsilon) - \hat{f}(\epsilon)\|^2) \quad (8)$$

$$L_{kd} = L_{soft} + L_{sim} \quad (9)$$

In Eq. 7, $\hat{z}$ is the logit of the last block, z^1, z^2 are the logits of the two selected blocks, $p(z_i,\theta) = softmax\left(\frac{z_i}{\tau}\right)$, τ is the temperature value. L_{soft} makes the classifier of the selected blocks more accurately by minimizing the KL divergence between the softmax of the last block and the one of the student model. In Eq. 8, $\hat{f}$ is the projected vector of feature that extracted from the last block, f^1, f^2 are the ones from the two selected blocks. ϵ is the series of input, d is the dimension of vectors. Features in the shallower blocks is guided by deepest block through L_{sim}.

2.2.3 Joint Training

Our proposed model aims to enable the network to efficiently recognize sketch and to clearly distinguish the latent space of sketch and photo using few sketch-photo pairs. In this section, we explain the loss functions used for joint

training. Firstly, we define three input pairs as follows. That is, a sketch image x, a photo image y_+ belongs to the same class of the sketch image x, and a photo image y_- belongs to another class. We use the triplet loss to perform the retrieval task. It makes the distance between sketch and image belonging to the same class closer, and the other class farther. The triplet loss is as shown in Eq. (10).

$$L_t(x, y_+, y_-) = max(0, g\left(f^i(x), f^i\left(y_+\right)\right) - g\left(f^i(x), f^i\left(y_-\right)\right) + m) \tag{10}$$

Here, g is used to measure the Euclidean distance between two features, and m is a margin. it enables to differentiate between a pair of sketch-photo of the same class and that of different classes.. Our modified triplet loss is represented in Eq 11

$$\begin{aligned} L_T(x, y_+, y_-) = \max[&(\lambda KL(\sigma(x), \sigma(y_+)) + g(f^i(x), f^i(y_+))) \\ &- (\lambda KL(\sigma(x), \sigma(y_-)) + g(f^i(x), f^i(y_-))) + \mathrm{m}] \end{aligned} \tag{11}$$

Here, KL is a KL-Divergence function, and σ is a softmax function. We consider spherical loss [13] which consists of the angular distance to improve the classification performance.

$$L_s\left(x_i\right) = \frac{1}{N} \sum_i log \frac{e^{\|x_i\| \delta(\theta_{y_i}, i)}}{e^{\|x_i\| \delta(\theta_{y_i}, i)} + \sum_{k \neq y_i} e^{\|x_i\| \delta(\theta_k, i)}} \tag{12}$$

Here, $\delta\left(\theta_{y_i}, i\right) = (-1)^t \cos\left(m\theta_{y_i}, i\right) - 2t, \theta_{y_i}, i \in \left[\frac{t\phi}{\alpha}, \frac{(t+1)\phi}{\alpha}\right], t \in [0, \alpha - 1]$. Despite of the separability of the features, a performance degradation can occur if the inter-class variation is not sufficiently discriminative. It is important to learn a center of deep features of each class and penalize the distance between the deep features and their corresponding classes. It makes the feature more distinguishable [14]. The center loss is defined as follows.

$$L_{ct} = \frac{1}{2} \sum_{k=1}^{m} \|x_k - c_{y_k}\|^2 \tag{13}$$

Here, m is mini-batch size, c_{y_k} is the center of each class y_i. After setting the center of each class, the samples belonging to the corresponding class are positioned close to the center. As mentioned above, the proposed model has an auxiliary layer for performing task in a total of three encoder blocks. For our final loss, we calculate all losses and basic cross-entropy loss for classification in each block, multiply each coefficient value, and add them all together.

$$L_{total} = \sum_{b=1}^{3} (\gamma^1 L_{ce}^b + \gamma^2 L_T^b + \gamma^3 L_s^b + \gamma^4 L_{ct}^b) + \gamma^5 L_{kd} \tag{14}$$

3 Experiments

3.1 Experimental environments

To evaluate our model, we use two kinds of datasets: TU-Berlin Extension [16] and Sketchy (extended) [17]. TU-Berlin Extension is a data set consisting of 20,000 images per category in 250 categories and 204,480 photo images in the same number of categories as the TU-Berlin provided in [17]. Sketchy (extended) is a dataset with 125 categories, 75,000 sketch images and 73,000 photo images. In experiments for on both datasets, sketch images in each category are split 80 : 20 for training and retrieval evaluation respectively. We resize input images to 224 × 224, and random matching sketch and photo images for SBIR. We conduct experiments on a few sketch samples. For ZS-SBIR, we follow the setting in [5] on sketchy and split the dataset into 104 categories for training and 21 categories for testing, and make making sure that the testing categories do not appear in the 1,000 categories of ImageNet. We randomly choose 30 categories that contain at least 400 images for testing and the rest 220 categories for training.

We utilize a pre-trained model with ImageNet. During training, we extracted logit and features for performing task from the classification and retrieval head of each selected encoder block. And all of our networks are trained for 5 epochs. We adopt the SGD (Stochastic Gradient Descent) optimizer and cosine learning rate decay, and the learning rate was is set to 0.05 and the weight decay to 0.000005. The hyper-parameters required for loss calculation (Eq. 14.) were set to 1.0, 0.15, 1.5, 0.15, and 0.1, respectively, and these values could be found through the grid search method.

To evaluate the sketch classification task, top-1 accuracy is applied, For the SBIR task, we use mean average precision (mAP) and precision at top-rank 200 (P@200) we use mean average precision (mAP) and at top-rank 200 (P@200). Also mAP and top-rank 100 (P@100) are used for the ZS-SBIR. We test each experiment using 10%, 25%, 50%, and 100% of entire training images to validate the efficiency of the proposed method even though smaller usage of data.

3.2 Results and Analysis

3.2.1 SBIR

For evaluating the proposed method on SBIR task, we compare our method to StyleMeup [18] for both datasets. Table 1 shows results for SBIR task. mAP of our model

Table 1 Comparative results of our model against other methods on SBIR (SD → Self-distillation methods and 10%, 25%, 50%, 100% → amount of sketch data for training.)

Method	TU-Berlin (ext)		Sketchy (ext)	
	mAP	P@200	mAP	P@200
StyleMeup[18]	0.778	0.795	0.905	0.927
ours (w/o SD)	0.754	0.811	0.890	0.905
ours (w/ SD) 10%	0.620	0.686	0.821	0.859
ours (w/ SD) 25%	0.672	0.731	0.867	0.881
ours (w/ SD) 50%	0.752	0.779	0.883	0.895
ours (w/ SD) 100%	0.781	0.835	0.911	0.922

(w/ SD, 100%) is 0.781 which is slightly higher than that of StyleMeup for TU-Berlin. P@200 of our model (w/ SD, 100%) is 0.835, which is higher than 0.795 of StyleMeup for Sketchy data. For SBIR task on Sketchy data, the mAP of ours (w/ SD, 100%) is 0.911, which is slightly higher than that of StyleMeup. Our P@200 on Sketchy is 0.922, only 0.005 lower than that of StyleMeup.

3.2.2 SBIR

The comparison results of Zero-Shot SBIR, which is the most challenging task, are shown in Table 2. GZS-SBIR [19] proposes a generative model based on an inverse autoregressive flow based variational auto-encoder to solve the retrieval task on unseen classes. SkechGCN [20] proposes a sketch representation using GCN, which simultaneously considers both the visual information and the semantic information. BDA-SketRet [21] introduced the Jensen-Shannon divergence, a symmetric KL divergence, to perform effective alignment among multi-domain features.

Our proposed model is superior to other similar methods on both TU-Berlin and Sketchy dataset consistently. Specifically, mAP of our model (w/ SD, 100%) is much higher than mAPs of GSZ-SBIR and SketchGCN, by 0.146 and 0.060, respectively on TU-Berlin dataset. On the other hand, on Sketchy dataset, our model performs much better than the same models by 0.210 and 0.117, respectively. Especially, Ours (w/ SD, 10%), using only 10% of TU-Berlin data for training, show 0.020 higher for mAP than SketchGCN. Therefore, it is clear that our proposed method works well for with very small data. We supplement the retrieval performance of our proposed model qualitatively on TU-Berlin and Sketchy datasets.

Table 4 Results of ablation study of block selection for self-distillation

Method	Classification	SBIR	ZS-SBIR
$[B_4, B_8, B_{12}]$	87.37	0.856	0.463-
$[B_6, B_8, B_{12}]$	86.35	0.882	0.481
$[B_7, B_9, B_{12}]$	88.31	0.892	0.477
$[B_{10}, B_{11}, B_{12}]$	90.45	0.911	0.499

Table 2 Comparative results of our model against other methods on ZS-SBIR (SD → Self-distillation methods and 10%, 25%, 50%, 100% → amount of sketch data for training.)

Method	TU-Berlin (ext)			Sketchy (ext)		
	mAP	P@200	P@100	mAP	P@200	P@100
GZS-SBIR[19]	0.238	-	0.334	0.289	-	0.358
SketchGCN [20]	0.324	0.478	0.505	0.382	0.487	0.538
BDA-SketRet[21]	0.375	-	0.504	0.437	-	0.514
ours (w/o SD)	0.362	0.481	0.519	0.453	0.507	0.550
ours (w/ SD) 10%	0.297	0.360	0.391	0.402	0.468	0.489
ours (w/ SD) 25%	0.352	0.446	0.472	0.443	0.497	0.525
ours (w/ SD) 50%	0.368	0.450	0.484	0.478	0.511	0.546
ours (w/ SD) 100%	0.384	0.507	0.531	0.499	0.552	0.581

Table 3 Results of ablation study of the impact of losses

	L_{ce}	L_t	L_T	L_s	L_{ct}	L_{kd}	Top-1	mAP
1)	O	O					85.73	0.642
2)	O	O		O	O		89.21	0.860
3)	O		O				89.04	0.658
4)	O		O	O	O		89.42	0.890
5)	O		O	O	O	O	90.45	0.911

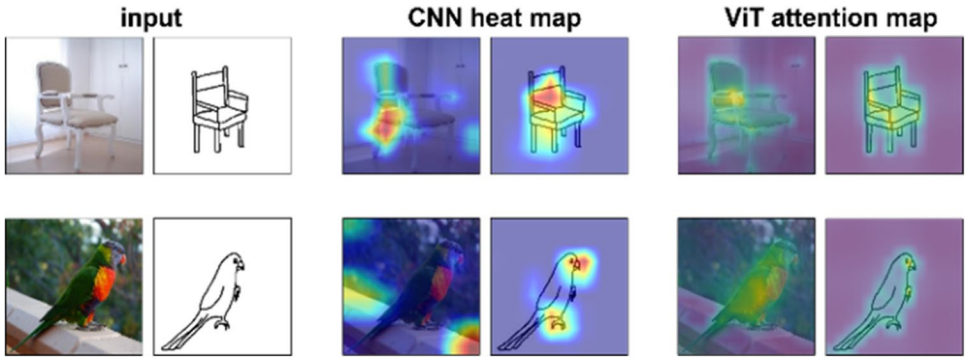

Fig. 3 Results of CNN heat map and ViT attention map

3.3 Ablation Study and Visualization

3.3.1 Combination of Multiple Loss Functions

To analyze the impact of losses, an ablation study is executed for SBIR task on Sketchy dataset as shown Table 3. No. 1 (Cross-Entropy and Triplet) shows the lowest performance. No. 3 and No. 4, which include Sphere and Center losses, have much better than that of No. 1. On the other hand, experiment No. 5 which exploited combination of most of losses including our proposed losses of Our Triplet and KD represents best scores.

3.3.2 Selecting the Proper Encoder Blocks for Self-Distillation

Since our ViT-based network has twelve blocks, there are numerous combinatorial cases to consider. Therefore, it is necessary to conduct an ablation study to select the blocks for self-distillation. For the experiment, we utilized the Sketchy dataset and maintained the same hyper-parameter settings as in the main experiments. The results of the ablation study on block selection for self-distillation are presented in Table 4. Among the combinations of blocks, [B10, B11, B12] exhibited the best performance across all three tasks. Furthermore, it was observed that blocks closer to the final block in the student network tended to yield better performance compared to those located further back. The earliest blocks, such as B1 and B2, are not included in the table as their performance is inferior.

In addition, comparison results of positive logits, predicted percentage of accurate class, for each block are shown in Fig. 2. It's noteworthy that combinations like [B4, B8, B12], [B6, B8, B12], and [B7, B9, B12] outperform the performance without self-distillation for intermediate blocks (from B4 to B8), but the last block's performance is lower. Conversely, [B10, B11, B12] shows lower performances for intermediate blocks (from B4 to B8) compared to [B4, B8, B12], [B6, B8, B12], and [B7, B9, B12], yet blocks nearer to the last block, such as B10 and B11, exhibit higher performance. This implies that distilling blocks closer to the last block in self-distillation has a more direct impact on the final performance.

3.3.3 Qualitative Impact of ViT in Sketch

Sketch images consists of a contour lines without background. Figure 3 shows heat maps extracted from the CNN model and attention maps extracted from our ViT model for photo and sketch images. CNN model is used to generate heat maps, region of interest is extract from the heatmap using the Grad-Cam. Extracting the region of interest with a CNN model shows that the map is formed based on the center of the object, as shown in the middle of the Fig. 3. On the other hand, extracting the region of interest for the ViT-trained network shows that the map is formed around the contour of the object as shown in the right of the Fig. 3. This shows that the ViT model is a better network for sketch recognition compared to the CNN model, by focusing on contour rather the center of the object, when performing the sketch image classification or retrial.

4 Conclusion

We proposed a modified Vision Transformer (ViT) based model with self-distillation for sketch based image retrieval. First, we add a token for retrieval and integrate auxiliary classifiers of multiple branches to improve the structure of ViT network. Second, self-distillation is integrated to enable fast transfer learning of sketch domains for ViT network with the proposed scheme of adding classifiers and embedding vectors to each intermediate layers for improving performance. Third, incorporation of KL-Divergence into the triplet loss function to mitigate overfitting problems arising from few input data. Experiments on the TU-Berlin and Sketchy dataset show our method to outperform similar methods significantly on sketch classification, SBIR, and zero-shot SBIR. Even the proposed method shows competitive incase of using only a small percentage of training data for zero-shot SBIR especially.

Acknowledgements This Research was supported by Seokyeong University in 2023.

References

1. Liu L, Shen F, Shen Y, Liu X, Shao L (2017) Deep sketch hashing: Fast free-hand sketch-based image retrieval. CVPR
2. Lin H, Fu Y, Jiang YG, Xue X (2020) Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. CVPR
3. Ribeiro LSF, Bui T, Collomosse J, Ponti M (2020) Sketchformer: Transformer-based representation for sketched structure. CVPR
4. Dey S, Riba P, Dutta A, Llados J, Song YZ (2019) Doodle to search: Practical zero-shot sketch-based image retrieval. CVPR
5. Dutta A, Akata Z (2019) Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. CVPR, pp 1105–1113
6. Tian J, Xu X, Shen F, Yang Y, Shen HT (2022) Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. AAAI
7. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. ICLR
8. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H (2021) Training data-efficient image transformers ´ distillation through attention. ICML
9. Suh S, Rey VF, Lukowicz P (2023) TASKED: Transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation. Knowl-Based Syst 260:110143
10. Ma H, Wang J, Lin H, Zhang B, Zhang Y, Xu B (2023) A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. IEEE Transactions on Multimedia
11. Zhang L, Song J, Gao A, Chen J, Bao C, Ma K (2019) Be your own teacher: Improve the performance of convolutional neural networks via self distillation. ICCV, pp 3713–3722
12. Yun S, Park J, Lee K, Shin J (2020) Regularizing class-wise predictions via self-knowledge distillation. CVPR
13. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. CVPR
14. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. ECCV
15. Hinton G, Vinyals O, Dean J (2014) Distilling the knowledge in a neural network. NIPS
16. Eitz M, Hays J, Alexa M (2012) How do humans sketch objects? ACM TOG
17. Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG
18. Sain A, Bhunia AK, Yang Y, Xiang T, Song YZ (2021) Stylemeup: Towards style-agnostic sketch-based image retrieval. CVPR
19. Verma VK, Mishra A, Mishra A, Rai P (2019) Generative model for zero-shot sketch-based image retrieval. CVPR
20. Zhang Z, Zhang Y, Feng R, Zhang T, Fan W (2020) Zero-shot sketch-based image retrieval via graph convolution network. AAAI
21. Chaudhuri U, Chavan R, Banerjee B, Dutta A, Akata Z (2022) BDA-SketRet: Bi-level domain adaptation for zero-shot SBIR. Neurocomputing 514:245–255

Sungjae Kang received B.S. and M.S. degrees from Electronics Engineering from Seokyeong University, Seoul, Korea, in 2020 and 2022 repectively. He is currently Researcher in AI Matics. His research interests include deep learning, computer vision.

Kisung Seo received the BS, MS, and Ph.D degrees in Electrical Engineering from Yonsei University, Seoul, Korea, in 1986, 1988, and 1993 respectively. He became Full Time Lecturer and Assistant Professor of Industrial Engineering in 1993 and 1995 at Seokyeong University, Seoul, Korea. He joined Genetic Algorithms Research and Applications Group (GARAGe) and Case Center for Computer-Aided Engineering & Manufacturing, Michigan State University from 1999 to 2002 as a Research Associate. He was also appointed Visiting Assistant Professor in Electrical & Computer Engineering, Michigan State University from 2002 to 2003. He was a Visiting Scholar at BEACON (Bio/computational Evolution in Action CONsortium) Center, Michigan State University from 2011 to 2012. He is currently Professor of Electronics Engineering, Seokyeong University. His research interests include deep learning, evolutionary computation, fuzzy-neural networks, computer vision, intelligent robotics.