LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID ****

Abstract

pass

1. Introduction

The importance of our research area

Some progress in sketch based image retrieval

Difficulty in Zero-shot setup and some possible solutions

Our proposed methods and their advantages

itemize our contributions in this paper

2. Related Work

2.1. Sketch-based image retrieval

2.2. Zero-Shot Learning

2.3. Cross-modal domain translation

3. Methodology

There will be five parts in this section. Sec. 3.1 defines the our targeted problem and briefly introduce our framework. Sec.3.2 introduce the feature extractor. Sec. 3.3 introduce the image cVAE-GAN. Sec. 3.4 introduce the sketch cVAE-GAN. Sec. 3.5 introduce the semantic preservation module. Sec. 3.6 introduce the design of loss functions during training procedure.

3.1. Problem Definition

In this paper, we focus on solving the problem of hand-free sketch-based image retrieval under zero-shot setup, where only the sketches and images from seen class are used during training stage. Our proposed framework is expected to use the sketches to retrieve the images, the categories of which have never appeared during training.

We first provide a definition of the SBIR in zero-shot setting. Given a dataset $S = \{(x_i^{img}, x_i^{ske}, x_i^{sem}, y_i) | y_i \in \mathcal{Y}\}$, where x_i^{img}, x_i^{ske}, x_i^{sem} and y_i are corresponding to the image, sketch, semantic representation and class label. Following the zero-shot setting in [2], we split all classes $\mathcal{Y}$ into $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$ according to whether the label exists in ImageNet[1], where no overlap exists between two label set, i.e. $\mathcal{Y}_{train} \cap \mathcal{Y}_{test} = \emptyset$. Based on the partition of label set $\mathcal{Y}$, we split dataset into S_{train} and S_{test}. Our model need to disentangle structure representations of image using data in S_{train}. During test, given x^{ske} from S_{test}, our model need to retrieve several images from test images candidate.

Our goal is to learn a two-way map between image feature domain to sketch feature domain. To this end, we propose a new deep network (shown in Figure **??**), which contains two structure encoders $\{E_s^{img}, E^{ske}\}$, one appearance encoder E_a^{img}, two feature decoder $\{G^{img}, G^{ske}\}$, a semantic decoder and two domain discriminators $\{D^{img}, D^{ske}\}$. Note that, the overall model can be regarded as two cVAE-GANs working parallel, which target to reconstruct sketch features from image and reconstruct image features from both sketches and images. To better capture the semantics information inside the sketches and images, we also add a semantic decoder to preserve semantics information while reconstructing the image features.

3.2. Feature Extractor

Considering the abstractness and visual sparsity of sketch, it is challenging extract feature from sketch. To alleviate this issue, multi-channel and multi-scale model was proposed to extract more saint features [3]. Motivated by the visualization in [4], where different layers capture visual features at different levels, we follow [] and build our feature extractor using a multi-layer feature fusion network enrich feature representation capacity without adding any additional parameters.

3.3. Image cVAE-GAN

3.4. Sketch cVAE-GAN

3.5. Semantics Preservation

3.6. Loss Function

4. Experiment

4.1. Experiment Setup

4.1.1 Dataset

4.1.2 Implementation Details

4.2. Comparison

4.3. Ablation Study

4.4. Case study

5. Conclusion

6. To Discuss

- Whether to generator the whole image/sketch.
- If the poses between the image and the sketch are different, can the model learn the sketch information between image and sketch.
- Where to add the semantics information to further supervise the model's training.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 1

[2] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *European Conference on Computer Vision*, pages 316–333. Springer, 2018. 1

[3] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017. 1

[4] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1