Sketch-based Image Retrieval using Generative Adversarial Networks

Longteng Guo[1,2], Jing Liu[1], Yuhang Wang[1,2], Zhonghua Luo[3], Wei Wen[3], Hanqing Lu[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Samsung R&D Institute, Beijing, China
{longteng.guo,jliu,yuhang.wang,luhq}@nlpr.ia.ac.cn,{zhonghua.luo,wei.wen}@samsung.com

ABSTRACT

For sketch-based image retrieval (SBIR), we propose a generative adversarial network trained on a large number of sketches and their corresponding real images. To imitate human search process, we attempt to match candidate images with the *imaginary* image in user's mind instead of the sketch query, i.e., not only the shape information of sketches but their possible content information are considered in SBIR. Specifically, a conditional generative adversarial network (cGAN) is employed to enrich the content information of sketches and recover the imaginary images, and two VGG-based encoders, which work on real and imaginary images respectively, are used to constrain their perceptual consistency from the view of feature representations. During SBIR, we first generate an imaginary image from a given sketch via cGAN, and then take the output of the learned encoder for imaginary images as the feature of the query sketch. Finally, we build an interactive SBIR system that shows encouraging performance.

1 INTRODUCTION

Sketch-based image retrieval (**SBIR**) is a natural way for image searching, since free-hand sketches are more expressive than textual queries, and easier to obtain than image queries. During SBIR, a user draw a sketch of the shape of the real object in his/her mind, which we call the *imaginary* image, and then submits it to search the wanted images. In fact, a user measures the relevance of the search results by comparing them against the imaginary image instead of the drawn sketch. However, as a common solution in previous works[2, 6], only the shape information is utilized, while the detailed content information is neglected. Therefore, the attempt to infer the imaginary image in user's mind, which can be seen as the inverse process of human sketching, is promising to enhance the performance of SBIR.

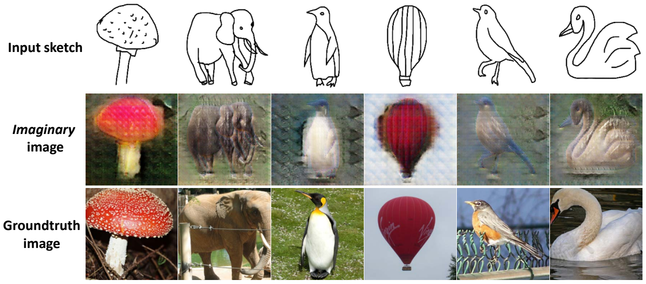

Figure 1: Examples of the generated *imaginary* images.

In this paper, we propose a learning framework based on generative adversarial networks in order to utilize a large number of sketches and their corresponding real images to enrich the content information of sketches and recover the imaginary images. The proposed framework consists of a conditional generative adversarial network (cGAN)[3] and two VGG-based encoders. CGAN, conditioned on the sketches, is leveraged to model the conditional distribution of real images and thus generate the imaginary images. The encoders take as input real and imaginary images respectively, and a feature reconstruction loss is defined on the produced features to achieve their perceptual consistency. Simultaneously, the encoder for imaginary images can also be employed to learn a suitable feature representation of sketches for SBIR. Based on the framework, we build an interactive SBIR system that shows encouraging performance.

2 METHODS

Given an input sketch I^S, our goal is to first infer the *imaginary* image I^M from I^S, and then encode a content-enriched feature from I^M such that the distance between the features of I^M and the target photo I^P is minimized. The framework of our proposed method is shown in Figure 2.

Our system consists of a **cGAN** and two VGG-based encoder networks ($E1$, $E2$). The cGAN contains two *adversarial* parts: the generator G and the discriminator D. The generator takes I^S and a noise z as input and outputs I^M. The discriminator conditioned on I^M and I^P to distinguish between real images and the generated imaginary images. The encoder networks work on I^M and I^P respectively to produce their semantic features and define a feature reconstruction loss L_{feat}, which is to guarantee the high-level perceptual consistency of I^M and I^P. Since I^P is a natural image, we

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10.
https://doi.org/10.1145/3123266.3127939

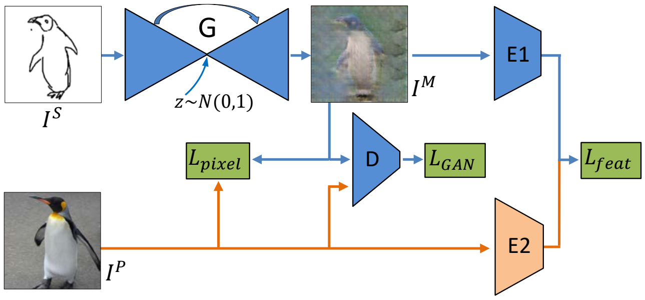

Figure 2: The framework of our cGAN based SBIR system. G and D denote the generator and discriminator respectively. $E1$ and $E2$ are two VGG-based encoders. While $E2$'s parameters are fixed during training.

initialize $E2$ with ImageNet pretrained model and keep its weights fixed. We specifically design the feature reconstruction loss L_{feat} as the MSE between the features extracted from $E1$ and $E2$. To capture the low-level visual similarity, we also adopt an average pixel-wise $L1$ loss L_{pixel}. Thus, combining the above losses with the adversarial loss L_{GAN} produced by cGAN, our final objective function becomes:

$$L = \lambda_1 L_{GAN} + \lambda_2 L_{pixel} + \lambda_3 L_{feat}.$$

The above cGAN and encoder network could be jointly trained on sketch/photo pairs. As for the network architecture, we base the generator on the *U-Net* [4] architecture, which skip connects layers in the encoder with the corresponding layers in the decoder. The discriminator uses the same structure as that in [1]. And the 16-layer VGG network is used for the two encoders.

3 IMPLEMENTATION AND PERFORMANCE

We perform experiments on the **Sketchy** dataset [5], which contains approximately 600 sketch/photo pairs for each of the 125 object categories. We randomly select 30 sketch/photo pairs for each category as the test set, and use the rest as the train set. The resolution of all the input and output images in our system are set to 256 × 256. The feature reconstruction loss is computed at the high-level *relu7_7* layer together with the mid-level *relu3_3* layer of the VGG networks. We adopt an alternate training mechanism for training G, D and $E1$. During the test phase, a sketch is fed successively through G and $E1$ to obtain a feature representation, and features of images in the database are extracted directly from $E2$. And the features from *relu7_7* layer of $E1$ and $E2$ are used.

Figure 1 shows the qualitative results of the generated *imaginary* images. As we can see, the generated images are not only semantically consistent with the input sketch but also share similar content information with the target photo. The retrieval results are shown in Figure 3. We can observe that the retrieved images correspond closely to the query sketches. To quantitatively evaluate the performance, the precision@20 metric of category-level SBIR is calculated. Our system achieves a score of 82.8%.

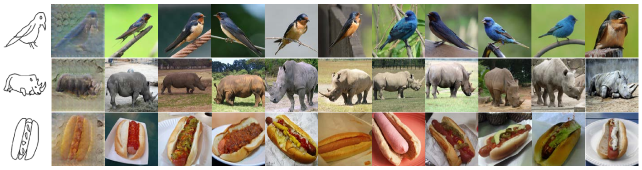

Figure 3: Examples of retrieval results. The first and second columns represent query sketches and the generated *imaginary* images respectively. The rest columns show the top 10 retrieved images.

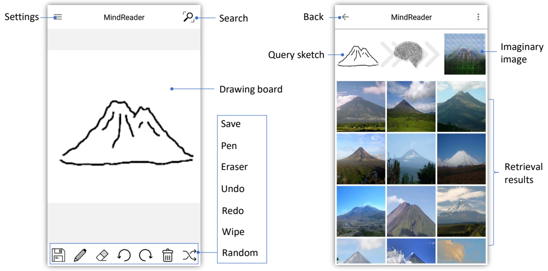

Figure 4: The GUI of our smart phone application.

4 APPLICATION

We develop a smart phone application for SBIR based on the above method. The screenshots in Figure 4 present the GUI of our smart phone application: **MindReader**, which contains two screens. The left screenshot is the input screen, which provides a canvas for users to draw sketches with fingers. Below the canvas provides a simple toolbox. After a user finishes the sketch and touch the "search" button, the system would return results and jump to the result screen, i.e. the right screenshot, which shows the query sketch and the generated *imaginary* image on the top, and displays the retrieved images on the middle.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No.61332016 and No.61472422).

REFERENCES

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).

[2] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. 2016. Sketch-based image retrieval via Siamese convolutional neural network. In *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2460–2464.

[3] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 234–241.

[5] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 119.

[6] Changcheng Xiao, Changhu Wang, Liqing Zhang, and Lei Zhang. 2015. Sketch-based image retrieval via shape words. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 571–574.