Proposal for DSA5204 Project

Yanrong Chen (A0251099Y) Shuyuan Shen (A0251155M) Yuhua Wang (A0251546E)
Geer Zhang (A0251176H) Junhao Huang (A0251318L)

1 Citation

In this project, we would reference these publications:

- *Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks*[1].
- *Deep Generative Adversarial Networks for Image-to-Image Translation: A Review*[2].
- *Analyzing and improving the image quality of stylegan*[3].
- *GANs trained by a two time-scale update rule converge to a local nash equilibrium*[4].
- *Image-to-image translation with conditional adversarial networks*[5].
- *Improved techniques for training GANs*[6].
- *Very deep convolutional networks for large-scale image recognition*[7].
- *The Spatially-Correlative Loss for Various Image Translation Tasks*[8].

More specifically, we plan to reproduce and do some further researches mainly on the first one, and the type of project is **application**.

2 Introduction to research problem

This paper is aimed at solving the unpaired image-to-image translation problem and applying the technique in many tasks like collection style transfer, season transfer, photo enhancement, etc., where the paired training data does not exist.

Image-to-image translation is one of the main research areas in computer vision, which aims to produce an output image reflecting the style of target domain while keeping unrelated contents of the input source image unchanged. Common image-to-image translation tasks are multi-modal, single-modal and single-image image-to-image translation. Since Generative Adversarial model(GAN) has been proposed, it has been often used in image-to-image translation tasks, and has shown good performance.

When solving these image-to-image tasks, supervised setting with paired input-output training data is the most commonly used. However, in practice, obtaining paired data can be hard, for example it may be time-consuming and the desired output may not be well-defined sometimes, like the translation between photographs and pictures in Monet style. Therefore, in this paper, they have learned the relationship between domains, not pairs, i.e., the goal is to learn the mapping G: $X \longrightarrow Y$ given training samples $\{x_i\}_{i=1}^{N} \in X$ and$\{y_j\}_{j=1}^{M} \in Y$ in order to make distribution of images from $G(X)$ indistinguishable from the distribution of Y.

Although GAN is widely used in solving image-to-image problems, in the case of unpaired training data, it fails and needs modifications. The mapping G and adversarial loss can just ensure that the distributions of $\hat{Y}$ and Y are the same, but the individual x and y may not be paired up, and the optimization is difficult to make progress due to cases of mode collapse, for example, different inputs may get the same output. Therefore, a "cycle consistent" loss is added to the adversarial loss to prevent the learned mappings from contradicting each other. In the system, there is a translator $G : X \longrightarrow Y$and its inverse $F : Y \longrightarrow X$, they are trained simultaneously with two discriminators D_X and D_Y. The objective is the combination of Adversarial Loss and Cycle-Consistency Loss.

3 Background

The traditional Image-to-image translation can be trace back to texture transfer because image style can be seen as a texture, and if we keep some semantic information while compositing image. Then, we can obtain the result of image-to-image translation. However, at this stage, the transformation was based on pixels, the basic image feature, and didn't include any semantic information. Thus, the performance wasn't ideal.

As it was mentioned above, image-to-image translation can be divided into two parts, image texture extraction and image reconstruction, but, for traditional Image-to-image translation, it can hardly solve the reconstruction part.

With the rapid development of deep neural network, researchers found that it can be utilized on image recognition, and CNN (e.g., VGG19 [7]) can extract the best feature instead of dividing the image into pixel.

After GAN has been proposed, it obtained great performance on image-to-image translation. What make GAN different from the traditional neural network is the design of the loss function. GAN is composed of generator and discriminator, which used to reconstruct the image and distinguish the image respectively (e.g., CycleGAN, which proposed a new loss, Cycle Consistency Loss[1]).

Here are the main knowledges we need to learn:

1. The structure of VGG19 and how it was applied in feature extraction.
2. The different between the loss function in traditional neural network and the one in GAN.
3. The structure of GAN, such as the structure of different discriminator (e.g., PatchGAN discriminator[2], StyleGAN discriminator[3]).
4. The loss proposed by CycleGAN, why adding this loss can improve the performance.
5. The metric for automatically evaluating the quality of image generative models[4][6].

4 Plan

We intend to reproduce the Cycle-Consistent GAN model from the selected application article[1], testing its performance on another dataset-selfie2anime. In the dataset, 69,926 animated character images are first obtained from Anime-Planet1.27,023 face images were extracted using Anime Face Detector2. After selecting only female character images and manually removing monochrome images, we collected two datasets of female anime face images of size 3400 and 100 for training and testing data, respectively. All the images are resized to 256 x 256 by applying CNN-based image super-resolution algorithm in the end.

More specifically, as part of the extension, we plan to explore redefining the loss function part of the model. In the original paper, it is proposed to combine adversarial loss and cycle consistency loss as the full objective for training and learning. Since cycle consistency losses do not explicitly decouple structure and appearance, there will be false learning of missing features. Furthermore, the model is trained using a fixed ImageNet, in which is full of pre-trained network. And it cannot be adaptively applied to other fields. We plan to introduce the self similarity of two unpaired samples of images from the perspective of image segmentation to describe the characteristics of deep features as a loss function. [8] Finally we rebuild an extended model to test performance, and compare the model results in the dataset used in the original paper.

5 Evaluation

We use the dataset ImageNet for reproduction and selfie2anime for additional research which studies the performance of image-to-image models in anime face image generation.

- **Datasets**: Selfie2anime comprises 3500 unpaired samples of female selfies and female anime face images, 3400 unpaired samples of which are training data and the rest are test data. All images are of same size (256×256).
- **Baseline models**: We choose the same baseline models as [1], i.e., CoGAN, Pixel loss + GAN, Feature loss + GAN, BiGAN/ALI and pix2pix.
- **Evaluation metrics**: For selfie2anime, the Frechet inception distance (FID) score is the most commonly used metrics. A lower FID suggests a shorter distance between the distributions of generated image and real image, thus generated images of better quality[2]. We also select the kernel inception distance (KID) as a metric, which gives unbiased estimates.
- **Ablation study**: To study the effect of each component of loss function, we remove GAN loss and cycle loss respectively and compare their generated images with that of CycleGAN.

6 Division of work

- Yanrong Chen: The analysis of ablation study.
- Shuyuan Shen: The reproduction of the original model, evaluate the performance of the model whose loss function is based on adversarial loss+cycle-consistency loss on the selfie2anime dataset.
- Yuhua Wang: The comparison of the extended model and the original model on the Zebra dataset.
- Geer Zhang: The extension model of the loss function covering the self similarity feature, evaluate the performance on the selfie2anime dataset.
- Junhao Huang: The performance of other models i.e., CoGAN, Pixel loss+ GAN in the baseline on the selfie2anime dataset.

References

[1] Jun-Yan Zhu et al. "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks". In: Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 2223–2232.

[2] Aziz Alotaibi. "Deep Generative Adversarial Networks for Image-to-Image Translation: A Review". In: *Symmetry* 12.10 (2020).

[3] Tero Karras et al. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119.

[4] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[5] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[6] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems* 29 (2016).

[7] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[8] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. "The Spatially-Correlative Loss for Various Image Translation Tasks". In: *CoRR* abs/2104.00854 (2021).