IMAGE-PRODUCT TRANSFORMATION BASED ON CROSS-DOMAIN GENERATIVE ADVERSARIAL NETWORKS

Yunfei Ge (yg2523), Xingzhi Li (xl2680)

Columbia University
School of Engineering and Applied Science
116th St and Broadway, New York, NY 10027

ABSTRACT

Retrieving product information from real-world image can have many promising applications. We propose a cross-domain generative adversarial model to address the challenge of generating clothes image from a dressed person. We train the neural network with paired images of fashion models (source domain) and products (target domain). To improve generated image quality and similarity, we introduce an additional association discriminator into original Generative adversarial network (GAN) model during training. We further enhance network performance by replacing convolutional layers with residual blocks. Experimental results show that we succeed in extracting clothes image from photos with great quality.

Index Terms— GAN, Computer Vision, Domain Adaption, Residual Network

1. INTRODUCTION

Generative adversarial network (GANs) [1] are a class of neural networks which can learn a deep generative model from high-dimensional distributions of data. This powerful technique has shown great success on various applications, such as super-resolution image generation (SRGAN) [2], image-to-image translation (CycleGAN) [3], realistic face image generation [4], etc.

Inspired by the versatility of GAN framework, we hope to apply this technique into cross-domain image transformation. The motivation of our work derives from daily life situations where we see an amazing outfit on fashion models and want to obtain the detail of the clothes. Given real-world pictures, our goal is to establish a generative neural network model which can transfer real-world photo in source domain into actual target domain product image in pixel level.

Cross-domain image transformation, also called as domain adaption, has been proposed and utilized in several previous works [5, 6, 7]. However, the adaption of these works concentrates on the feature space, without directly generating target image output. Our goal is to transfer the knowledge in source domain into realistic and associated pixel-level target image. The challenge lies on the problem that the output target is not deterministic [8]. Generated results may not preserve the semantic meaning with the original input photo.

In order to tackle the challenge, we improve the basic GAN framework with an additional association discriminator network to supervise the generator, producing target image that is corresponding to the input. The two discriminators jointly optimize the generator to produce realistic and associated output image.

Our main contribution to the work are as follows:

1. Propose an improved GAN architecture with two discriminators to realize cross-domain image-product transformation.
2. Adopt residual blocks into the generator and improve model performance.
3. Conduct training on actual photo-product pairs and successfully generated product image from real-world photos.

2. RELATED WORK

2.1. Generative Adversarial Network

Generative adversarial network was first proposed by Goodfellow et al. in [1]. In their work, they construct a new framework to estimate generative model by introducing another discriminative model. These two model compete with each other during the process of training. The discriminative model was designed to maximize the probability of assigning correct labels to both training examples and generated examples, while the generative model simultaneously minimize the value of this probability. One of the most popular GAN design is DCGAN[9] which bridge the gap of deep convolution neural network and GAN without using max pooling or fully connected layers. DCGAN used a network with downsampling and upsampling and achieved convincing results on various image datasets.

2.2. Residual network

Network depth is of great importance in architecture design, but the degradation problem still limits the extension of net-

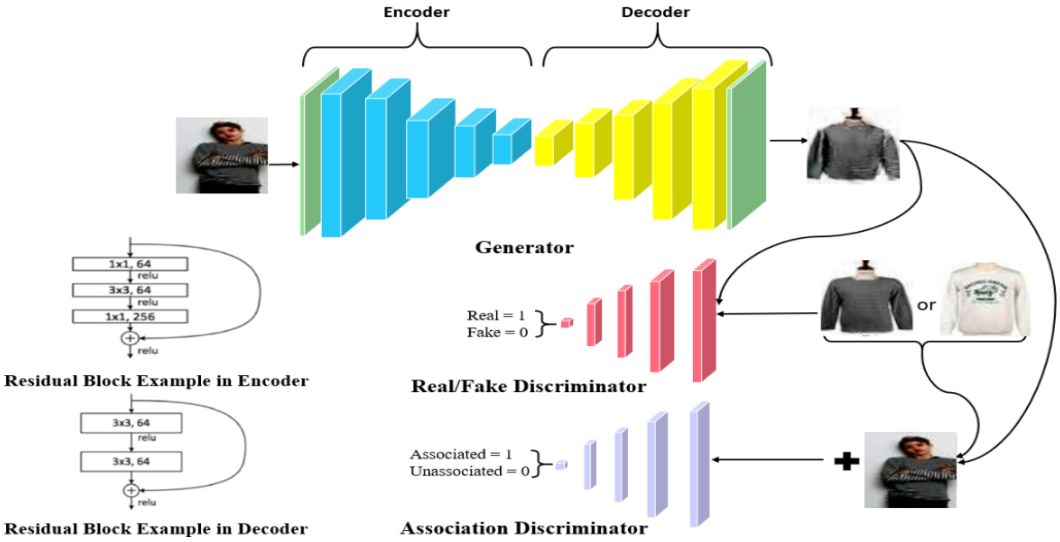

Fig. 1. Full architecture for image-product transformation.

work depth: as the depth of neural network increases, the accuracy gets saturated and degrades quickly. [10] introduced a deep residual leaning network to address this problem. They wisely hypothesized a mapping: $\mathcal{F}(x) = \mathcal{H}(x) - x$ where $\mathcal{H}(x)$ is the underlying mapping and thus is transformed into $\mathcal{F}(x) + x$. This can be recognized as a "shortcut connections" which skips one or more layers. The implementation of the structure substantially increases the depth of network and ease optimization comparing to its corresponding plain network. ResNet has won first prize in 2015 classification competition.

2.3. Domain adaptation

Domain adaptation aims at learning a target domain from a source domain which has a different data distribution such as transferring painting image to real image. Although there are methods like estimating a common embedding between two domains, GAN is one of the most successful way to tackling domain adaptation problem. CycleGAN[3] was able to transfer two domains without providing a pair of them. They proposed a novice way with two GANs transferring from source to target and transferring from target to source.

3. APPROACH

Our main goal is to extract product image from fashion image which has a different data distribution from target domain by modeling a mapping function $G : X \rightarrow Y$. The full architecture is composed of one generator and two discriminators. Two styles of residual blocks are implemented separately in encoder and decoder as shown in figure 1. In this section, We will describe in details the full architecture of our framework and explicitly design the training process.

3.1. Architecture

3.1.1. Generator

The Generator takes a 64×64 3-channel RGB fashion image as input, and outputs a 64×64 3-channel RGB product image. The generator is a combination of two parts: encoder and decoder. The encoder is used to analyze input image and extract features, we adopt encoder from ResNet-50 which composes of several repeated residual blocks and has an increasing kernel depth, but we get rid of the max-pooling layer of ResNet-50 based on the thinking from DCGAN[9] and the average-pooling layer. The encoder is continually downsampling the image size from 64×64 to 4×4, but at the same time increasing the depth from 3 to 2048. In order to implement the upsampling structure in decoder, we adopt ResNet-34 in a reversed way. Still, we get rid of the max-pooling layer and average pooling layer, and add a fully connected layer with $7 \times 7 \times 3$ filter and a tanh layer at the end of the generator network to reconstruct a image.

3.1.2. Real/fake discriminator

As in [1], the real/fake discriminator is assigned to compete against generator in the minmax game to generate a real product. Each of the input fashion image is matched up with its corresponding product image. The real/fake discriminator is composed of 4 convolutional layers with a binary output. The real/fake discriminator downsampling the image size from 64×64 to 4×4 with increasing kernel depth from 128 to 1024. The generated image is supposed to be distinguished from real product image taken a $False$ label, but at the same time the generator try to minimize the accuracy of real/fake discriminator by confusing it to give generated image a $True$ label.

3.1.3. Association discriminator

Even though we are able to generate a real product through a framework with one generator and one real/fake discriminator, the source domain can be mapped to a random product in target domain. But the main goal of our work is to resemble the product dressed on human in input image. Different from classical GAN framework, a new discriminator is added to network: association discriminator[11]. This discriminator aims to compete with generator through evaluating the association relationship between source domain and target domain. We adopt the same network in association discriminator as the real/fake discriminator. We pile up real product image with its corresponding fashion image and generated image with its input image. The reformed input pair size is $64 \times 64 \times 6$. The discriminator is supposed to distinguish the generated association and real association, while the generator attempts to confuse its decision.

3.2. Adversarial training

In classical GAN[1] framework, z represents source domain which has a distribution $p_z(z)$, and x represents the target domain with distribution $p_x(x)$. $D(x)$ shows the probability of x comes from target data distribution. $G(z)$ shows the mapping from z to estimated x. We train D to maximize the probability of assigning correct label to images from both target distribution and estimated distribution, and at the same time train G to minimize the maximum probability. The loss function can be defined as:

$$\max_G \min_D \mathcal{L}(D,G) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[1 - \log D(G(z))]$$

In our work, however, we have two discriminators. The real/Fake discriminator has a loss function:

$$\min_{D_D} \mathcal{L}(D_D) = \mathbb{E}_x[\log D_D(x)] + \mathbb{E}_z[1 - \log D_D(G(z))]$$

where x is the real product image, and z is the input fashion image which basically adopt the classical loss function. We label real product image as $True$ and generated image as $False$.

On the other hand, the association discriminator is formulated as:

$$\min_{D_A} \mathcal{L}(D_A) = \mathbb{E}_x[\log D_A(x,z)] + \mathbb{E}_z[1 - \log D_A(G(z),z)]$$

where we pile up real product image x and its corresponding fashion image z, and generated image $G(z)$ and its input fashion image z to the input of discriminator. The association between real product and its match-up fashion image is labeled as $True$, while the association between generated image and input fashion image is labeled as $False$.

After we trained the real/fake discriminator and the association discriminator separately, we utilize a compound loss function of G to minimize the maximum probability of both discriminators:

$$\max_G \mathcal{L}(G) = \min_{D_D, D_A} \frac{1}{2}\mathcal{L}(D_D) + \frac{1}{2}\mathcal{L}(D_A)$$

where we assign equal weight to both discriminators.

Algorithm 1 Adversarial training

for $i \leq IterationNumber$ **do**
  for $j \leq BatchNumber$ **do**
    Sample minibatch of input image $z^{(1)}, \ldots, z^{(m)}$
    Load corresponding real product image x of z
    Update real/fake discriminator:
    $\theta_D \leftarrow \theta_D - \Delta_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(D_D)$
    Update association discriminator:
    $\theta_A \leftarrow \theta_A - \Delta_{\theta_A} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(D_A)$
    Update generator:
    $\theta_G \leftarrow \theta_G + \Delta_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}\mathcal{L}(D_A) + \frac{1}{2}\mathcal{L}(D_D)$

4. EVALUATION

In this section, we test the performance of our framework with real-world model and paired images of their clothes. We conduct a set of experiments with different environment settings to evaluate the quality of our results.

4.1. Evaluation Methods

To realize cross-domain image transformation, we conduct experiments on LookBook dataset [11], which consists of over 50,000 fashion-related images collected from online shopping sites. In average, there are 8 images of the same model in different poses corresponding to one product image in tops category.

We use the following two criteria to evaluate the results. Root Mean Square Error (RMSE) calculate the pixel-level dissimilarity between two images. The value presents the dissimilarity between generated output and original image. RMSE is simple to implement, but can run into problems, because large distances between pixel intensities do not necessarily mean the contents of the images are dramatically different.

Structural Similarity Index (SSIM) measures the perceptual difference between two similar images [12]. SSIM is able to perceive the change in structural information of the image by comparing local regions of the image instead of globally. This methods is known to be more consistent with human perception.

4.2. Experimental results

Fig.1 demonstrates some examples of generation results from our cross-domain transformation network. The three columns

Fig. 2. Examples of Generation Result

presents the inputs, correlated targets, and generated images respectively. Generation results indicate that the model can learn the shape, color, pattern and print of the clothes and produce similar product image.

To verify the effectiveness and accuracy of our framework, we observe the training process by monitoring the average RMSE / SSIM value of the system over time.

4.2.1. Different network architecture

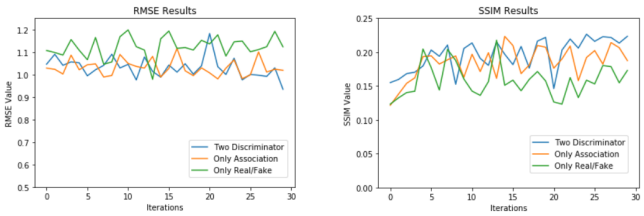

Fig. 3. Comparison of Different Network Architecture

Three different network models are discussed in this experiment: (1) Generator with only real/fake discriminator; (2) Only association discriminator; (3) Two discriminators (our model). Fig.2 illustrates the average RMSE / SSIM in different network during training process.

Along with the training process, the RMSE value gradually decreases, while the SSIM value increases. Our model with two discriminators can outperform the other two structures with the minimum RMSE value and maximum SSIM value. It is noticeable that the network model with only association discriminator can achieve better results than the model with only real/fake discriminator. The result indicates that without supervision of relevant input pairs, the generated image may look realistic, but possibly fails to represent the features of the original item.

4.2.2. Generator with/without residual blocks

The addition of residual blocks enable the system to train a deeper neural network without accuracy degradation. Meanwhile, the residual structure requires more iterations to fully train the network to perform our tasks.During the experiment, note that we applied stochastic gradient descent (SGD) and Adam optimizer in normal generator, while used only SGD optimizer in residual network to reduce computational cost.

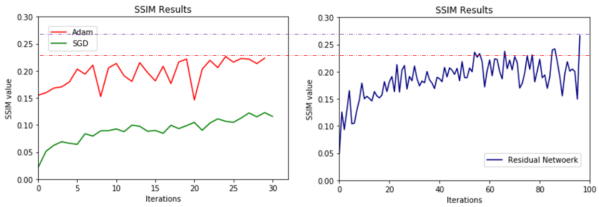

Fig. 4. Comparison of Normal Generator and Residual Generator

Fig.3 demonstrates that with longer training process, the generator with residual blocks can achieve the best performance among all other structures and conditions with the largest SSIM score. However, since the improvement of residual network is relatively trivial comparing to the training time, the trade-off between output accuracy and training efficiency still remains to be discussed.

4.2.3. Quantitative evaluation

Table 1. Evaluation Scores

Condition	RMSE	SSIM
Only A	0.98	0.22
Only D	1.0	0.20
A + D (SGD)	1.03	0.12
A + D (Adam)	0.94	0.23
A + D (ResNet)	**0.92**	**0.28**

Table 1 demonstrates the final evaluation score of different conditions, where "A" and "D" indicate association and real/fake discriminator. We can observe that our model can outperform all other models in the task of cross-domain image transformation.

5. CONCLUSION

In this paper, we present a cross-domain image transformation GAN model to tackle the task of generating product image from real-world human photos. We modify the traditional GAN framework with an additional discriminator to ensure the association between generated output and the target. To improve output accuracy, we introduce residual blocks into the generator, increasing the depth of the network to achieve better performance. Experimental results show that we succeed in producing clothes (product) image from natural human photos with great quality. In future work, we consider fine tune the parameter settings in neural network, trying to improve the generation image quality. We expect to train the framework with different datasets, extending the idea to other applications related to cross-domain image transformation.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network.," in *CVPR*, 2017, vol. 2, p. 4.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[5] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[6] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070.

[7] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*. Springer, 2010, pp. 213–226.

[8] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon, "Pixel-level domain transfer," in *European Conference on Computer Vision*. Springer, 2016, pp. 517–532.

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.