Abstract

Image to image translation은 어떤 입력 이미지에 대해 출력 이미지와 mapping 시키는 것을 목표로 하는 컴퓨터 비전, graphics task이다. 어떤 특성을 갖고있는 이미지를 다른 특성을 갖는 이미지로 변경하는 task에 대한 연구([1], [4])는 새로운 architecture와 적절한 loss function을 manual하게 연구되었다. 이와 달리 GAN[2]을 기반으로한(정확히는 cGAN[5], Pix2Pix[3]는 다양한 tasks에 대해서 공통적으로 적용할 수 있는 일반적인 방법론을 제시하였다.Pix2Pix model과 Pix2Pix의 한계점을 극복하여 아래 Figure1과 같이 멋진 결과를 낸 CycleGAN[6]에 대해서 알아본다.

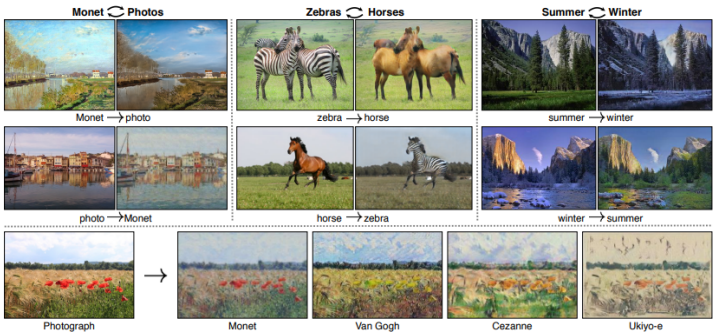

Figure 1: Result - CycleGAN

1 Background

1.1 Benefit of Generative Adversarial Network

Convolution neural networks는 image prediction tasks에서 손실함수를 최소화하는 방식으로 학습을 진행한다. 만약 naive하게 prediction과 ground truth의 Euclidean distance(L2 norm)을 손실함수로 사용하는 경우, 모델이 생성하는 결과 이미지는 semantic하지 않고 blurry한 결과를 얻게 된다. 하지만 우리가 원하는 이미지가 sharp하고 realistic한 이미지라면, 이를 위해 더 좋은 손실함수를 찾아야 하고 이는 manual하게 찾아야하므로 많은 노력이 들어가게 된다. 하지만 GAN은 판별자의 개념을 도입하여 prediction과 ground truth를 손실함수로 정량화하지 않고 생성자가 만들어낸 결과를 판별하고 피드백을 주어 생성모델을 학습하게 된다. 이렇게 생성된 결과는 sharp하며 현실에 있을 법하게 보인다.

1.2 conditional GAN : Base of Pix2Pix model

$$L_{cGAN}(G,D) = \mathbf{E}_{x,y}[logD(x,y)] + \mathbf{E}_{x,z}[log(1 - D(x,G(x,z))] \quad (1)$$

위는 cGAN의 목적함수로, cGAN은 GAN과 달리 입력으로 Latent vector 'z'만을 받는 것이 아니라 어떤 조건 x를 포함하여 (x, z)를 입력으로 한다. 즉 특정 클래스 혹은 특정 조건을 x로하여 x에 대한 조건에 맞도록 이미지를 만들어 내고자 할 때, cGAN을 사용할 수 있다.

2 Pix2Pix

Pix2Pix는 cGAN이 condition으로 특정 class와 같이 저차원의 간단한 정보를 입력으로 하지 않고, 이미지 자체를 조건으로 넣는 것이 큰 특징이다. 다시 말해, 특정한 이미지가 조건으로 주어졌을 때 만들어지는 결과의 다양성을 목표로 하는 것이 아니라, 조건에 맞는 적절한 이미지를 만들어 내는 것이 목적으로 한다. 이렇게 모델을 생성하였을 때, 만들어지는 이미지의 deterministic한 문제는 네트워크의 dropout을 사용함으로써 만들어지는 이미지에 random한 특징을 추가하여 해결할 수 있다. 이제 Pix2Pix의 구조와 목적함수에 대해 자세히 알아본다.

2.1 Objective

Pix2Pix의 목적함수는 cGAN의 목적함수에 traditional loss로 L1 norm을 더하여 사용하며, 아래와 같다.

$$G^* = \underset{G}{argmin}\underset{D}{max} L_{cGAN}(G,D) + \lambda L_{L1}(G) \quad (2)$$

L1 loss는 생성자가 만들어내는 이미지가 ground truth와 유사할 수 있도록 encourage하는 역할을 수행한다. 즉, 위에서 말했듯이 L2 norm은 blurring한 결과를 유발하기 때문에 더 sharp한 결과를 출력할 수 있도록 하는 L1 norm을 사용한다. 이때, L1 norm은 아래와 같다.

$$L_{L1}(G) = \mathbf{E}_{x,y,z}[||y - G(x,z)||_1] \quad (3)$$

2.2 Architecture - Generator

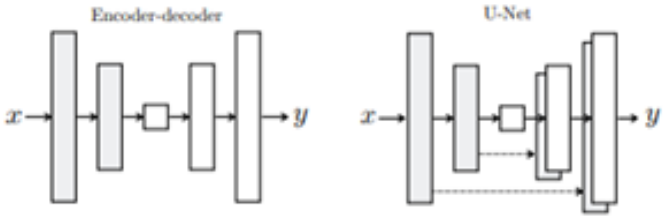

Figure 2: Architecture of Pix2Pix Generator

Pix2Pix의 generator는 인코더-디코더 구조에 스킵커넥션이 추가되어 있는 U-Net이 사용된다. 즉, 입력으로 들어오는 이미지와 출력으로 나가는 이미지가 같은 구조를 렌더링하는 형태의 구조이다. 기존 인코더-디코더 구조에 bottleneck 이후 layer에 bottleneck 이전 layer의 정보를 넘겨주는 skip-connection을 사용함으로써 low-level의 정보를 공유할 수 있게 되어 정보 손실을 줄일 수 있다. 새롭게 구성한 objective function과 U-Net구조를 활용하였을 때의 성능을 Figure3에서 확인할 수 있다.

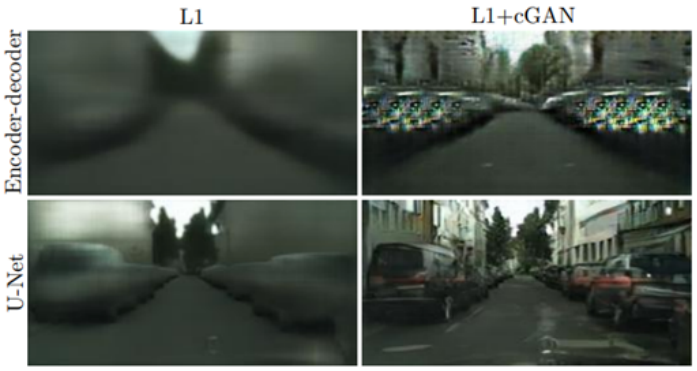

Figure 3: Performance of Objective and U-Net

추가적으로, Pix2Pix2의 판별자는 convolutional PatchGAN을 사용하여, 생성된 이미지 전체에 대해 판별하지 않고 이미지 내 패치 단위로 진위여부를 판별한다. 이에 따라 high-frequency의 이미지를 얻을 뿐만 아니라, 큰 이미지에 대해서 상대적으로 더 적은 파라미터를 가지고 빠른 속도로 학습할 수 있다.

3 CycleGAN

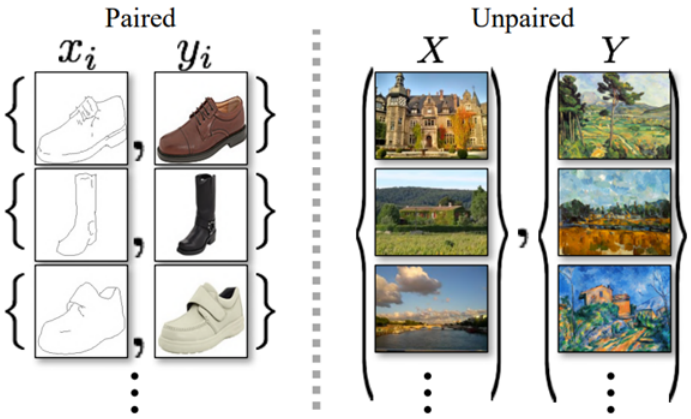

Figure 4: Architecture of Pix2Pix Generator

Pix2Pix의 한계점은 Figure4의 Paired와 같이 (x, y) = (조건 이미지, 정답 이미지)로 묶인 데이터셋이 존재할 때만 사용할 수 있다. 따라서 정답 이미지를 target으로 하기위한 적절한 조건 이미지를 직접 생산해내야 한다. 이에 따라 [3]의 결과 이미지는 결과 이미지를 간단히 스케치하여 그것을 조건 이미지로 한 colorization task를 한 이미지가 주를 이룬다(같은 배경을 다른 시간에 찍은 사진이나 이미지의 일정 부분을 삭제한 조건 이미지도 존재한다). 따라서, 데이터셋이 Unpaired인 상황에서도 Image to image translation task를 수행할 수 있는 CycleGAN을 제시한다.

3.1 Two Generator

cGAN[5]이 아닌 기존 GAN[2]의 목적함수를 사용한다면 입력 x와 y = G(x)가 의미있는 방식으로 짝지어지는 것을 보장할 수 없다. 즉, GAN의 목적함수는 input x를 통해 만들어낸 y가 ground truth의 분포(Y domain)를 따르게 하는 것이므로 mode collapse(input와 관계 없이 모두 같은 이미지로 mapping되어 optimization에 실패하는 것)와 같은 문제가 발생할 수 있다. 이러한 문제를 해결하기 위해 본 논문의 저자는 cycle consistent의 개념을 도입한다. 즉, 입력이미지 x가 2개의 생성자를 통해 2번 translation 되어 다시 원본이미지처럼 돌아올 수 있도록 모델을 학습시키는 것이다.

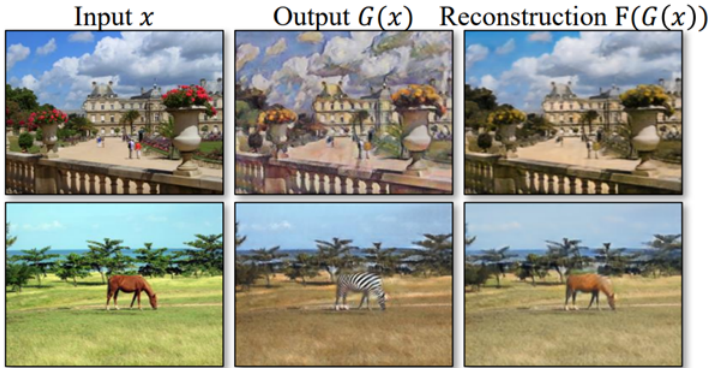

Figure 5: Objective : Reconstruction

Figure5의 위 3개의 이미지를 보면, 실제 카메라로 촬영한 풍경 이미지를 생성자 G를 통해 그림과 같은 이미지로 바꾸고, 다시 생성자 F를 통시 원본 이미지처럼 돌려 놓는 것이다. 아래 말을 얼룩말로 translation하고 다시 말로 바꾸는 것을 통해 더 쉽게 이해할 수 있다.
결과적으로 원본이미지의 content는 보존된 상태로 domain과 관련된 특성만을 바꿀 수 있도록 학습되는 것이다. 따라서 G(x)를 F에 넣었을 때, 다시 원본 이미지로 복구될 수 있는 형태로 objective funtion을 구성해야 함을 추측할 수 있고 이에 대해 자세히 알아본다.

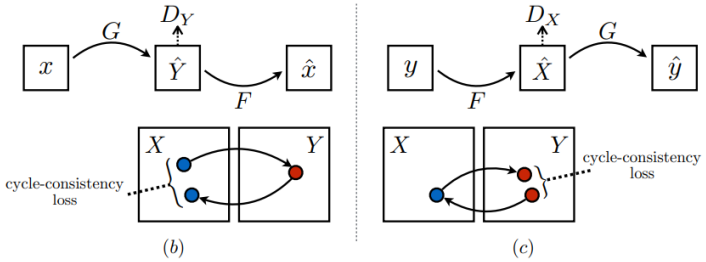

Figure 6: Translation of CycleGAN

3.2 Cycle-consistency loss

Cycle GAN의 translation은 Figure6과 같이 이루어진다. 이에 따른 Cycle-GAN의 전체 objective는 아래와 같다.

$$L_{GAN}(G,F,D_X,D_Y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,X,Y) + \lambda L_{cyc}(G,F) \tag{4}$$

$$\Longrightarrow$$

$$L_{GAN}(G,D_Y,X,Y) = \mathbf{E}_{y\sim p_{data}(y)}[logD_Y(y)] + \mathbf{E}_{x\sim p_{data}(x)}[log(1-D_Y(G(x)))] \tag{5}$$

$$L_{GAN}(F,D_X,Y,X) = \mathbf{E}_{x\sim p_{data}(x)}[logD_X(x)] + \mathbf{E}_{y\sim p_{data}(y)}[log(1-D_X(F(y)))] \tag{6}$$

$$L_{cyc}(G,F) = \mathbf{E}_{x\sim p_{data}(x)}[||F(G(x))-x||_1] + \mathbf{E}_{y\sim p_{data}(y)}[||G(F(y))-y||_1] \tag{7}$$

6와 5과 같이 기존의 GAN의 목적함수를 그대로 사용하여 target domain의 있을 법한 이미지를 생성하도록 의도하며, 7와 같이 Cycle-consistent loss를 통해 입력과 매칭되는 image-to-image translation 결과 이미지를 찾을 수 있도록 한다.

3.3 Architecture

CycleGAN의 architecture는 [4]의 아키텍쳐를 기반으로 한다. 네트워크 내에 3개의 convolutions가 포함되며 각각 residual blocks가 존재한다. 이 중 2개는 stride가 $\frac{1}{2}$인 fractionally-strided convolutions를 수행하고 나머지 하나의 convolution은 features를 RGB로 mapping하기 위해 사용된다. Residual blocks로는 128×128 이미지에 대해서는 6개의 blocks, 256×256 해상도 이상의 이미지에 대해서는 9개의 blocks이 필요하다. 판별 네트워크는 70×70의 patch를 기준으로 하는 patchGAN이 사용된다.

4 Conclusion

딥러닝에서 각 task의 기준에 맞는 모델을 평가하는 지표의 Score를 높이는 것도 중요하지만, GAN을 기반으로 한 모델이 생성하는 이미지는 human study가 꼭 필요한 지표라 생각된다.

Loss	Map → Photo % Turkers labeled *real*	Photo → Map % Turkers labeled *real*
CoGAN [32]	0.6% ± 0.5%	0.9% ± 0.5%
BiGAN/ALI [9, 7]	2.1% ± 1.0%	1.9% ± 0.9%
SimGAN [46]	0.7% ± 0.5%	2.6% ± 1.1%
Feature loss + GAN	1.2% ± 0.6%	0.3% ± 0.2%
CycleGAN (ours)	**26.8% ± 2.8%**	**23.2% ± 3.4%**

Figure 7: Scores of human study

위 표에서 CycleGAN은 이전의 GAN모델에 비해 human study에서 매우 큰 성취를 보여주었다. 하지만 context를 유지한채 style을 바꾸는 task에 가까워 데이터 샘플이 부족할 때, 원하지 않는 이미지가 출력될 수 있으며 모양을 바꿀 수 없는 한계가 존재한다.

[1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.