Attention을 이용한 비식별화 이미지 dataset 복원을 위한 GAN의 시각화 비교

Visualization Comparison of GAN for Reconstructing De-identified Image Dataset using Attention

요 약

최근 CCTV를 통해 수집된 이미지 데이터들을 통해 사회문제를 해결하려는 움직임들이 보이고 있다. 이러한 경향에 따라 개인 정보의 침해 없이 CCTV로부터 수집된 데이터를 수집할 수 있는 방안에 대한 문제가 발생한다. 이를 해결하기 위해서 촬영된 이미지들의 얼굴 부분을 모자이크 처리를 해주고 있다. 하지만 이러한 이미지 데이터들은 비식별화 되어있고 이는 image자체로도 사용하기 힘들고 학습데이터로도 사용하기 힘들다는 단점이 있다. 이를 해결해주기 위해 비식별화 되어있는 얼굴부분을 새로 reconstruct 시켜줄 수 있다. 기존의 방법을 사용하여 비식별화된 이미지를 reconstruct하면 얼굴부분이 제대로 reconstruct되지 않고 이에 다른 method를 추가해주어 이 문제를 해결해주어야 한다. 본 연구에서는 reconstruct 과정에서 GAN모델에 Attention module을 추가해주었을 때 모델이 reconstruct하기 위해 attention하고 있는 부분을 Grad-CAM을 통해 시각화 하였다.

1 Introduction

최근 사회문제들을 해결하기 위해서 여러 대의 CCTV로 부터 수집한 Image data들을 사용하려는 노력이 있다. 지속적으로 영상이 생성되기 때문에 Deep Neural Networks (DNN) 학습용 영상 및 Image dataset을 많이 만들어낼 수 있다는 장점이 있지만 개인정보 보호의 이슈가 해결되지 않은 영상 및 Image data는 사용할 수 없다는 문제가 있다. 이러한 문제를 해결하기 위하여 이미지를 비식별화 할 수 있는데 이 과정에서 영상 및 Image data의 변화가 생겨나고 결론적으로 비식별화 되어 있는 부분에 의해 효과적인 학습 dataset으로 사용하기 힘들다. 이러한 문제를 해결하기 위해서 GAN [1]을 사용하여 비식별화 되어있는 모자이크 부분을 reconstruct해주는 방법을 사용할 수 있다.

기존에 Image를 reconstruct 해주는 방법으로는 DCGAN [2], Pix2Pix [3], CycleGAN [4]등이 있다. DCGAN [2]은 deep convolutional network를 GAN [1]에 적용시킨 연구로 image를 reconstruct 할 수 있다는 가능성을 보여주었다. 하지만 한계점으로는 image-to-image reconstruction이 되지 않는다는 단점이 있고 이는 복원을 위한 network를 위하여 사용하기 힘들다는 한계점을 지니게 된다. 이러한 문제를 해결해주기 위하여 등장한 image-to-image transfer GAN [1]으로는 Pix2Pix [3], CycleGAN [4]등이 있다.

우선 Pix2Pix [3]는 label y에 대응되는 image x가 pair로 존재하는 dataset이 있을때 이를 가지고 학습할 수 있다. Generator가 생성하는 reconstruct된 이미지를 기존의 image y와 비교하는 L1 loss를 Adversarial GAN loss와 합쳐서 loss function을 새로 만들어 학습의 결과를 높여주었다. Image-to-image reconstruction은 진행이 되었고 기존의 방법보다 좋은 성능을 보이지만 label과 대응되는 image가 pair로 존재해야 된다는 단점이 있고 dataset을 만드는데에 어려움이 존재한다.

CycleGAN [4]은 pair로 된 이미지가 존재해야 된다는 단점을 가진 위의 Pix2Pix [3]의 문제점을 해결한 network이다. Generator와 Discriminator를 2개씩 만들어주었고 각각의 Generator가 역함수 관계를 가지게하여 2개의 Generator를 거치게 되면 생성되는 이미지와 원래 이미지에 관한 loss를 뜻하는 identity loss를 기존의 Adversarial GAN loss에 추가해 주었다. Pair로 존재해야 된다는 한계를 해결하였지만 attention이 되어있지않아 원하는 부분에 관한 reconstruction이 되지는 않는다.

본 연구에선 기존의 GAN에 attention module을 추가하여 주었을때 attention이 되어 적용이 되는지를 시각적으로 확인 시켜주었다. Attention을 구현하기 위하여 CBAM [5] 모듈을 사용해주었다. 이는 Channel Attention module과 Spatial Attention module을 직렬로 연결시켜주어 좋은 attention rate을 보여준다.

2 Methods

비식별화 이미지를 복원하기 위한 Generator와 생성 품질을 측정하기 위한 Discriminator로 CycleGAN [4]에서 사용한 네트워크 구조와 동일한 모델을 사용하였다. 그 대신 Attention을 각 객체에 적용하기 위해 CBAM [5] 모듈을 각 ResNet [6] Block 사이에 적용하여 변화를 주었다. 추가된 Attention module이 객체에 집중

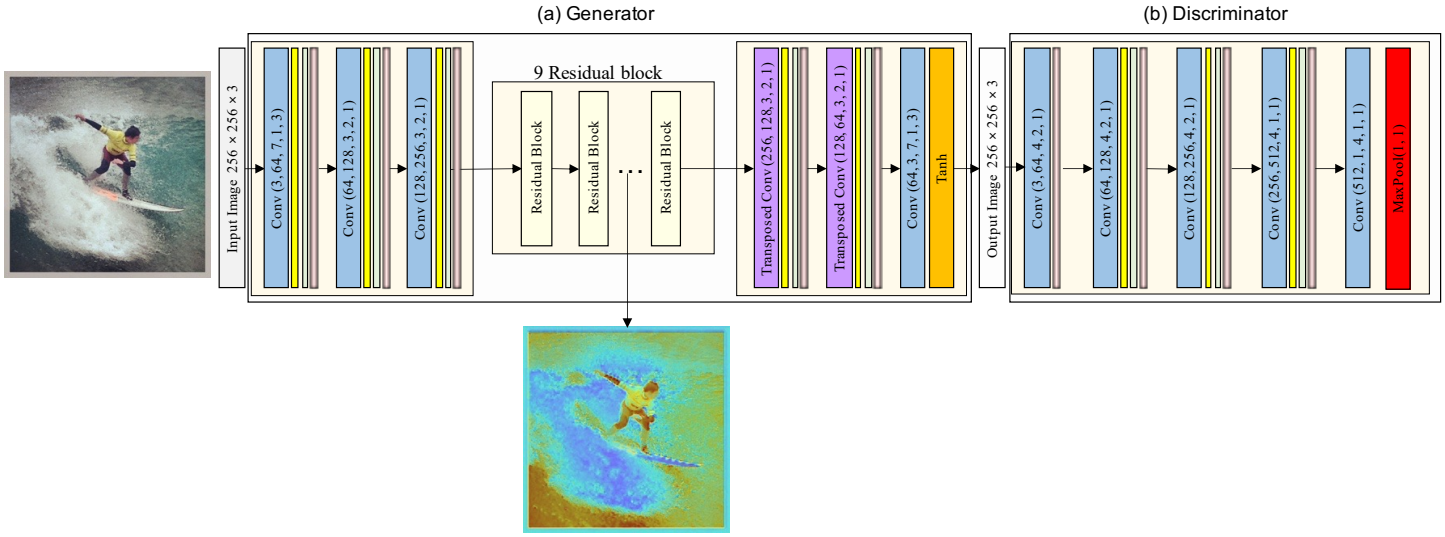


그림 1: Network architecture of Generator and Discriminator

하여 작동되고 있는지 확인하기 위해 Grad-CAM [7]을 이용하여 feature map을 시각화하였다.

2.1 Network Training

전체적인 네트워크의 그림은 Fig. 1와 같다. 비식별화 된 이미지는 이를 복원할 수 있는 네트워크 G를 먼저 통과하여 비식별화된 부분을 복원하게 된다. 그 이후 복원된 네트워크를 Discriminator D에 통과 시켜 해당 이미지의 진위 여부를 판단하도록 하였다. 이에 네트워크 학습을 위한 Adversarial loss는 다음과 같이 주어진다.

$$\mathcal{L}_{adv} = \mathbb{E}_{y\sim\mathcal{Y}}[\log(D(y))] + \mathbb{E}_{y\sim\mathcal{Y}}[\log(1 - D(G(x)))], \quad (1)$$

이때 비식별화된 데이터는 x, 비식별화 되지 않은 실제 이미지는 y로 표현하였다. 그리고 데이터셋이 pair를 이루기 때문에 L_1 loss 도 적용하여 학습에 활용하였다. 이때 해당 loss는 다음과 같다.

$$\mathcal{L}_{dist} = \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[|G(x) - y|] \quad (2)$$

따라서 네트워크 학습을 위한 최종 loss는 다음과 같고

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{dist} \quad (3)$$

학습의 최종 목표는 최적화된 G^*와 D^*를 찾는 것으로 다음과 같이 표현될 수 있다.

$$G^*, D^* = \arg\min_G \max_D \mathcal{L} \quad (4)$$

이번 실험에서 최적화된 G^*와 D^*를 학습 완료하고 이들 weight를 고정 후 feature map을 시각화하였다.

2.2 Visualization

Grad-CAM [7]은 입력 이미지에 label이 주어질 경우를 가정하고 구현되었기 때문에 G와 D를 모두 통과 시켜 출력으로 진위 여부를 판별하도록 하였다. 결과적으로 이미지 생성을 classification으로 변화시켜 feature map을 시각화 할 수 있도록 하였다. 이때 시각화한 feature map은 CBAM [5]가 적용된 레이어에 적용하였다. 이를 통해 Attention module이 비식별화 된 객체에 집중하여 이미지를 복원할 수 있도록 설계되었음을 검증할 수 있다.

3 Experimental Results

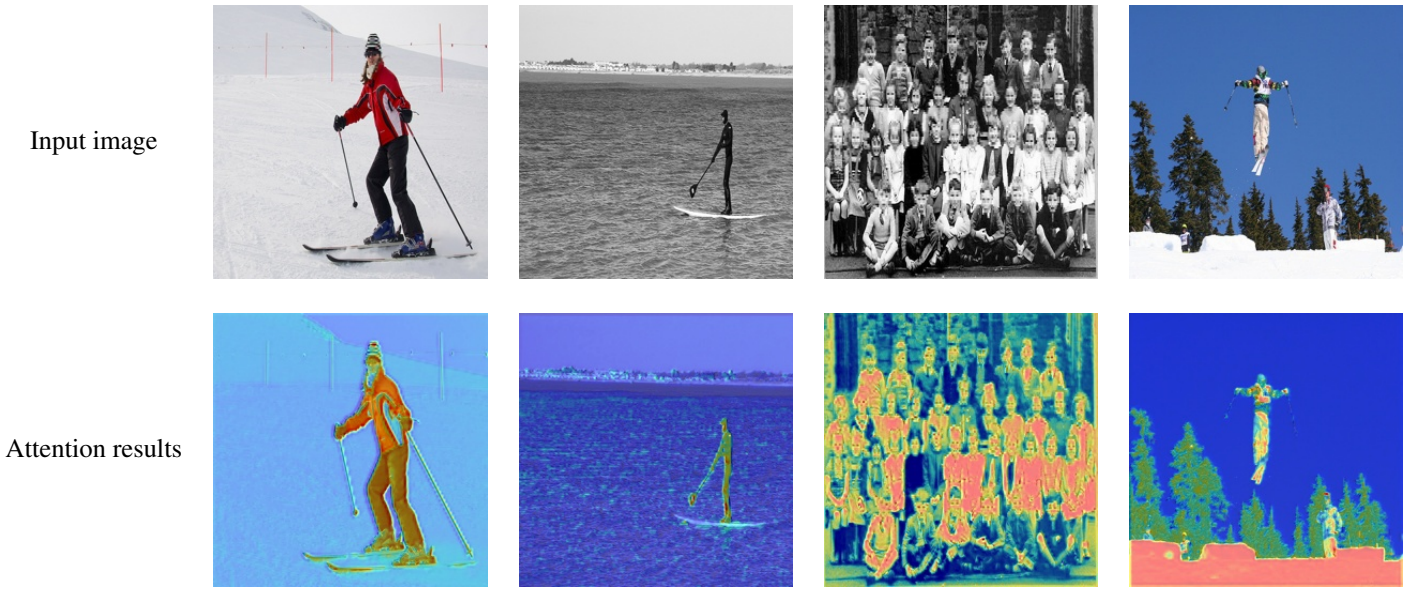
Fig. 1에서 볼 수 있듯이 Grad-CAM [7]을 통과시킨 이미지를 살펴보면 다른 배경과 사람 사이의 명확한 색차이를 확인할 수 있다. 이는 Attention 모듈이 image에 있는 사람 부분에 attention을 걸은 것임을 확인할 수 있다. 추가적으로 다양한 사람들이 다양한 색깔 및 배경에서 찍힌 image에서의 결과를 얻어보았다. 이는 이미지 복원 네트워크 G에서 feature map을 Grad-CAM [7]을 통해 시각화한 결과이다. 결과는 아래 Fig. 2와 같다.

결과적으로 network에서의 attention이 사람 부분에 걸렸고 이는 image에서의 사람부분을 reconstruct하는데에 있어 의미있는 영향을 주고 있다는 사실을 알 수 있다.

4 Conclusion

CCTV 데이터를 이용하여 여러 네트워크를 학습하고 이를 활용해 시스템을 개발할 수 있지만 학습 데이터를 구축하는 것은 개인정보 침해의 우려가 있어 매우 조심스러운 일이다. 따라서 CCTV 데이터셋을 공개하기 위해 얼굴과 같이 개인 정보가 포함된 영역을 모자이크 처리하여 공개 또는 배포한다. 본 논문에서는 기존

그림 2: Attention 모듈에서의 feature 시각화 실험 결과



의 image-to-image translation 방식을 이용하여 비식별화된 이미지를 복원하여 네트워크 학습에 이용할 수 있도록 하는 가능성을 검증하였다. 실험에서 Generator를 학습 시키고 Discriminator로그 품질을 평가했을때 복원 CBAM [5] 모듈에서의 feature를 Grad-CAM [7]을 이용하여 측정해보았다. 실험 결과 학습된 네트워크는 개인 정보가 포함된 사람 영역에 집중하여 복원하는 모습을 확인할 수 있었다.

참고 문헌

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv*, 2016.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv*, 2015.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.