사전 학습된 이미지 생성 모델을 이용한 Semantic Segmentation 복원

안천수, 설상훈
고려대학교 전기전자공학부

csahn@mpeg.korea.ac.kr, sull@korea.ac.kr

Semantic Segmentation Restoration Using Pretrained Image Synthesis Model

CheonSu Ahn, Sanghoon Sull
School of Electrical Engineering, Korea University

요 약

최근 pixel 단위로 annotation 된 semantic segmentation mask 를 이용해 이미지를 생성하는 다양한 모델이 제시되고 있다. 하지만 사전 학습된 모델을 사용하면 주어진 mask 가 불완전할 경우 제대로 된 이미지를 생성하지 못하는 문제점이 있다. 본 논문은 mask 로부터 이미지를 생성하는 사전 학습된 모델을 활용하여 품질이 떨어지는 불완전한 segmentation mask 를 완전한 mask 로 복원하는 모델을 제시한다. 이는 사전 학습된 이미지 생성 모델에 대한 추가 학습 없이 mask 생성 모델에 대한 학습을 통해서만 이루어진다. 학습 과정에서는 생성한 mask 와 완전한 segmentation mask 를 비교하고 복원한 mask 를 기반으로 생성한 이미지와 실제 이미지를 비교하여 mask 복원 성능을 높일 수 있도록 했다. 도로 주행 영상 데이터 셋인 Cityscapes 에 대해 제시한 모델을 이용하여 sparse 한 mask 와 sampling 을 통해 8 배만큼 저화질로 만든 mask 를 실제 완전한 mask 로 복원하는 실험을 진행하였다.

Ⅰ. 서 론

최근 semantic segmentation mask 가 입력으로 주어졌을 때 실제와 같은 이미지를 생성하는 다양한 모델이 제시되고 있다[1,2]. 그중에서 대표적으로 SPADE[3] 구조를 활용한 모델이 있다.

이런 semantic mask 를 입력으로 받아 이미지를 생성하는 모델에서는 학습할 때 입력으로 받는 segmentation mask 를 고화질의 pixel 단위의 annotation 으로 사용하는 경우가 많다. 이를 실제 모델을 활용할 때 불완전한 mask 에 대해서는 이미지를 제대로 생성하지 못하는 단점이 있다. 이것을 해결하기 위해서는 불완전한 mask 에 대해서 네트워크를 다시 처음부터 학습시켜야 한다. 하지만 이미지 생성 모델은 대체로 복잡하기에 처음부터 학습시키기 위해서는 컴퓨터 연산이 많이 필요하고 학습 시간이 길다는 문제점이 발생한다.

이런 문제점을 해결하기 위해 본 논문에서는 사전 학습된 이미지 생성 모델을 학습 과정에 활용하여 불완전한 mask 를 완전한 mask 로 복원하는 mask 생성 모델을 제안한다. 학습 시 이미지 생성 모델은 loss 에 대한 gradient 만 전파할 뿐 업데이트를 하지 않아 컴퓨터 연산량과 소요 시간을 절감할 수 있다. 학습 과정을 통해서 전체 모델은 실제와 유사한 복원된 mask 를 만들 수 있게 되고 이를 통해 사전 학습된 이미지 생성 모델이 실제와 같은 이미지를 생성할 수 있게 해준다.

Ⅱ. 본론

2.1 모델 구조

전체 모델 구조는 그림 1 과 같은 구조를 사용하였다. Baseline 모델로 SPADE[3]를 이용한 이미지 생성 모델과 같은 형태로 이미지와 mask 생성 모델을 각각 구성하였으며 학습 과정에서 이미지 생성 모델은 사전에 학습된 모델을 사용하여 업데이트 하지 않고 loss 에 대한 gradient 만 전파하며 mask 생성 모델만 업데이트 하게 된다.

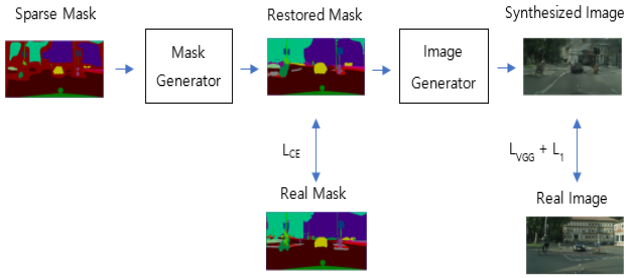


그림 1. 전체 모델 구조

Mask 에 관한 loss 는 생성한 mask 와 실제 mask 의 cross entropy loss 를 사용하였으며 복원한 mask 를 r, 실제 mask 를 x 라고 할 때 $L_{mask} = L_{CE}(r,x)$로 표현된다. 이미지 생성에 관한 loss 는 기존 [3]과 유사하게 VGG loss 와 추가적으로 L1 loss 를 사용하였으며 생성한 이미지를 s, 실제 이미지를 y 라고 할 때 $L_{image} = L_{VGG}(s,y) + L_1(s,y)$로 표현된다. 따라서 학습 과정에서 사용되는 전체 loss 는 $L_{total} = L_{mask} + L_{image}$ 가 된다.

앞에서 제시한 loss 를 통해 학습이 이루어지면 mask 생성 모델은 실제 mask 와 유사하면서도 이미지 생성 모델로 하여금 실제 이미지와 유사한 이미지를 만들도록 하는 mask 를 얻을 수 있다.

2.2 실험 결과

모델 학습과 평가를 위해 도로 주행 데이터 셋인 Cityscapes[4]를 사용하였다. Semantic segmentation mask 는 35 개의 class 로 구분되어 있고 3000 개의 학습 데이터와 500 개의 검증 데이터로 구성되어 있다.

모델을 평가하기 위해 데이터셋에 존재하는 sparse mask 를 pixel 단위로 완전히 annotation 된 완전한 mask 로 복원하는 실험과 8 배만큼 저화질로 만든 mask 를 다시 완전한 mask 로 복원하는 실험을 진행하였다. 각 실험에 대한 주요 결과는 그림 2 와 같다.

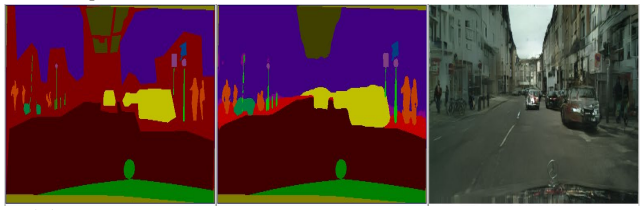
표 1 에 평가 결과를 정리하였다. 생성 이미지와 실제 이미지 분포의 차이를 측정하는 Frechet Inception Distance(FID)는 낮을수록 좋고 실제 mask 와의 유사도를 측정하는 pixel accuracy (accu), mean Intersection-over-Union (mIoU)는 높을수록 좋다.

먼저 sparse mask 를 입력으로 받아 실제 mask 와 비슷하게 복원하는 실험을 진행하였다. 제안한 모델을 사용할 경우 복원 과정을 거치지 않을 경우에 비해 FID 가 낮아지며 mask 에 대한 accu 와 mIoU 도 크게 증가함을 확인할 수 있었다. Loss 를 변경하며 결과를 비교해 보면 L_{CE} 에 L_{VGG} 와 L1 loss 를 적용했을 때 전체적으로 가장 좋은 성능을 보였다.

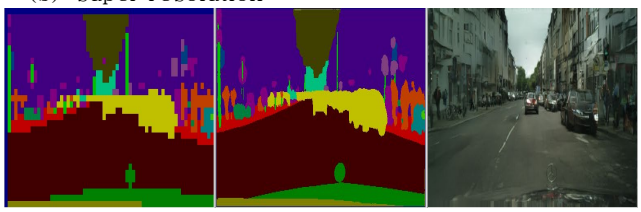
그리고 nearest 방식으로 8 배만큼 저화질로 만든 mask 를 다시 원래의 고화질 mask 로 복원하는 실험을 진행하였다. 복원 과정을 거치면서 FID 는 낮아지고 accu 와 mIoU 는 높아지는 것을 확인할 수 있었다. Loss 를 변경하며 결과를 비교해 보면 L_{CE} 에 L1 을 추가했을 때 mIoU 가 가장 높았으며 전체 loss 를 적용했을 때 FID 가 가장 낮았다.

전체 실험 결과를 볼 때 단순히 mask 에 대한 loss 인 L_{CE} 만 적용하는 것 보다 이미지에 대한 loss 인 L_{VGG} 와 L1 loss 를 적절하게 사용하면 전체 loss 를 최소화하는 과정에서 성능에 도움이 됨을 확인할 수 있었다.

(a) sparse→fine densification



(b) super resolution



(c) Real mask & image

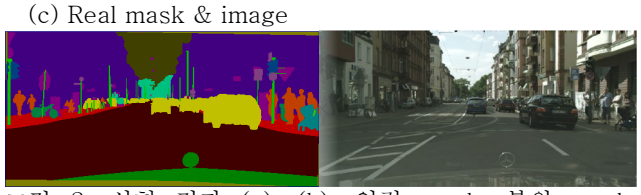


그림 2. 실험 결과 (a), (b): 입력 mask, 복원 mask, 생성 이미지 (c): 실제 mask 및 이미지

Task	Loss	FID	accu	mIoU
synthesis	-	62.9	-	-
sparse→fine densification	-	157.7	68.5	39.4
	CE+ VGG+ L1	68.4	**86.5**	**52.2**
	CE	70.7	86.1	42.1
	CE+ VGG	71.1	85.8	37.9
	CE+ L1	**68.2**	86.2	44.3
super-resolution	-	78.1	84.2	57.6
	CE+ VGG+ L1	**65.4**	**93.8**	65.2
	CE	69.2	93.3	69.5
	CE+ VGG	65.5	93.7	65.1
	CE+ L1	66.4	93.7	**70.6**

표 1. Cityscapes 데이터 셋 평가 결과. Synthesis 에서는 실제 온전한 mask 로 이미지를 생성했다. Loss 가 존재하지 않는 경우는 mask 복원 과정 없이 이미지를 생성하고 실제 이미지 및 mask 와 비교한 경우이다.

Ⅲ. 결론

본 논문에서는 사전 학습된 이미지 생성 모델을 활용하여 불완전한 semantic mask 가 주어졌을 때 완전한 mask 로 복원하는 모델을 제시하였다. 실험을 통해 mask 생성 모델을 학습할 때 실제 mask 와 비교할 뿐만 아니라 복원한 mask 로 생성한 이미지와 실제 이미지를 비교하는 방식을 이용하면 mask 를 더 잘 복원할 수 있을 뿐만 아니라 사전 학습된 이미지 생성 모델이 더 사실과 같은 이미지를 생성할 수 있음을 확인할 수 있었다. semantic segmentation task 와 같이 데이터가 부족한 상황에서 data augmentation 이 필요할 때 sparse 하거나 저화질의 semantic mask 만 있더라도 본 논문에서 제시한 방법을 통해 실제와 유사한 데이터 셋을 만들어 낼 수 있을 것이다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2021-2016-0-00464)

참 고 문 헌

[1] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

[2] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al, “learning to predict layout-to-image conditional convolutions for semantic image synthesis,” Advances in Neural Information Processing Systems (NeurIPS), 2019.

[3] Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y, “Semantic image synthesis with spatially-adaptive normalization,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016