

# UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS

## 0. Abstraction

최근에, CNN으로 이뤄진 지도학습은 컴퓨터 비전에서 큰 발전을 이룩하였다. 반면에 비지도학습은 상대적으로 주목을 받지 못하였다. 이 논문이 지도학습과 비지도학습을 위한 CNN의 성공 간격을 연결해주길 바란다. 우리는 특정 구조적 제약조건으로 비지도학습의 강력한 후보인 DCGAN을 소개하고자 한다. 다양한 이미지 데이터들로부터, Discriminator와 Generator가 물체부터 장면까지의 계층적 표현을 학습하는 것을 확인할 수 있었다. 게다가 이를 일반화할 수 있는 가능성도 증명하였다.

## 1. Introduction

라벨링되지 않은 데이터셋으로부터 재사용할 수 있는 feature representation을 학습하는 것은 활발한 연구분야이다. 컴퓨터 비전 분야에서 라벨링되지 않은 이미지나 비디오로부터 좋은 representation을 학습할 수 있다면, 이를 활용하여 이미지 분류와 같은 다양한 지도학습 문제에 적용할 수 있다. 이를 위한 하나의 방법으로 GAN을 사용하지만, GAN은 안정적인 학습이 어렵고, 결과가 엉성하다. 따라서 본 논문에서는 ▲ 안정적인 학습을 위한 Convolutional GAN의 구조적 위상에 대한 제한사항들을 제안하고 평가한다. ▲ 학습된 discriminator를 이용하여 이미지 분류 작업을 시행한다. ▲ GAN의 필터를 시각화하여 실험적으로 특정 필터들이 물체를 그리는 데 사용됨을 보여준다. ▲ generator는 생성된 샘플들의 문맥을 쉽게 다룰 수 있는 벡터의 연산 성질을 지니고 있다.

## 2. Related Work

라벨링되지 않은 데이터에 대한 비지도 학습은 주로 데이터를 클러스터링하고, 각 클러스터에 분류 점수를 부과하는 접근법이 대중적이다. 이미지에서는 계층적 클러스터링을 통해 강력한 이미지 표현법을 학습할 수 있다. 또 다른 유명한 방법은 이미지를 압축시킨 후, 이를 다시 가능한 정확하게 이미지로 복원하는 오토인코더를 학습시키는 것이다.

생성 이미지 모델은 크게 parametric, non-parametric 두 가지 방식이 있다. non-parametric model은 텍스쳐 합성, 고해상도, in-painting 등에 사용되는 방식으로 기존의 이미지 데이터베이스에서 patch를 매칭시키는 것이다. parametric model은 아직까지 실제 이미지를 만드는데 큰 성공을 거두지 못하였다. 기존의 방식들은 해상도가 좋지 못하고, 이해할 수 없는 결과를 보였다. RNN 접근법과 Deconvolution network 접근법은 실제 이미지와 유사하게 생성을 하였지만 지도학습으로 응용할 수는 없었다.

NN을 사용하는 것의 비판점 중 하나는 그들이 블랙박스모델이라는 점이다. 따라서 Deconvolution을 이용하여 각 필터에서 최대화하고자 하는 부분을 Gradient Descent를 이용하여 시각화하였다.

## 3. Approach and Model Architecture

CNN을 이용하여 GAN의 해상도를 높이기 위한 시도는 많았지만 성공적이지 못하였다. 본 논문에서는 수많은 모델 연구 끝에 데이터셋의 종류에 관계없이 고해상도와 deep한 생성모델의 안정적인 학습을 돋는 구조에 대해 알아냈다. 이 접근법의 핵심은 최근에 검증된 세가지 CNN 구조를 수정하고 적용한 것이다.

첫째, Pooling을 CNN with stride로 대체하면, 네트워크가 고유의 spatial downsampling을 학습할 수 있다. 둘째, Fully Connected Layer를 제거하는 것이다. Global Average Pooling은 모델의 안정성을 높이지만 수렴속도에 악영향을 끼친다. convolutional feature를 각각 O generator와 discriminator에 연결하는 것이 효과가 좋았다. GAN의 첫번째 레이언은 유니폼한 노이즈를 input으로 가지기 때문에 Matrix Multiplication이지만 결과적으로는 4 차원의 텐서로 변환하여 convolutional stack을 시작하게 된다. Discriminator의 마지막 레이언은 flatten이고 single sigmoid output을 가진다. 셋째, Batch Normalization은 input을 정규하여 안정적인 학습을 돋는다. 이것은 초기화와 관련된 문제를 해소하고 깊은 모델에서 gradient flow를 돋는다. 모든 레이어에 batchnorm을 적용하면

sample oscillation과 모델 불안정성을 오히려 높인다. 따라서, generator의 output layer와 discriminator의 input layer에는 batchnorm을 적용하는 것을 피해야 한다.

마지막 레이어는 Tanh를 사용하고 나머지 레이어들은 ReLU activation을 사용한다. 이를 통해 빠르게 학습하고, 트레이닝 분포의 color space를 커버한다. discriminator는 leaky rectified activation이 고해상도 모델링에 효과가 좋다.

## 4. Details of Adversarial Training

Large-scale Scene Understanding(LSUN), Imagenet-1k, Faces dataset 세가지 종류에 대해 학습을 진행하였다. tanh activation의 범위인  $[-1, 1]$ 로 스케일링하는 것을 제외하고는 별도의 전처리는 적용하지 않았다. 모든 모델은 mini-batch SGD를 이용하였으며, 미니배치의 사이즈는 128이다. 모든 가중치는  $N(0, 0.02^2)$ 으로 초기화하였다. LeakyReLU에서는 0.2값을 사용하였다. Adam optimizer( $lr=0.001$ )을 사용하고 모멘텀항을 추가하였다.

LSUN. 이미지 생성모델의 퀄리티가 증가함에 따라, over-fitting과 memorization 문제가 걱정되었다. 따라서 해당 문제가 모델에서 발생하지 않음을 증명하기 위해 3M장 이상의 LSUN bedroom 데이터셋에 대해 학습을 시켰다. 1에폭 뒤의 결과를 통해 모델이 단순히 학습 데이터를 오버피팅하거나 메모라이징을 통해 좋은 퀄리티의 샘플을 결과로 내는 것이 아님을 알 수 있다.



1 epoch 후에 생성된 사진들이다. 작은 learning rate의 SGD로 학습된 결과물을 통해 memorization에 의한 결과가 아님을 실험적으로 알 수 있다.

메모라이징 문제를 개선하기 위해서 간단한 de-duplication process를 실시하였다. 3072-128-3072 denoising dropout regularized RELU autoencoder를 학습시킨 후, semantic-hashing을 통해 약 27500개의 duplicates들을 탐지하고 제거하였다.

랜덤한 웹이미지 쿼리를 통해 10K명의 사람 이름과 3M개의 얼굴 이미지에 대한 데이터셋을 구축하였다. OpenCV face detector를 이용해 고화질로 350,000개의 face box를 얻을 수 있었다. 이것을 학습에 사용하였다.

Imagenet-1k데이터셋에서 32\*32 center crop하여 학습을 진행하였다.

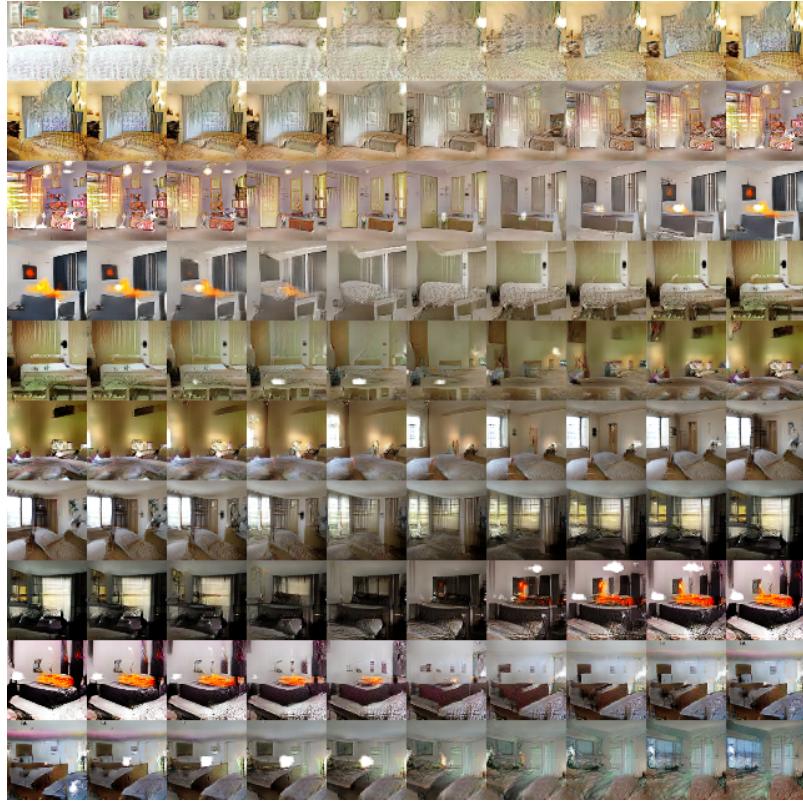
## 5. Empirical Validation of DCGANs Capabilities

비지도학습의 퀄리티를 평가하기 위해서 지도학습 데이터셋에 feature extractor의 역할로 사용하였다. 먼저 DCGAN을 Imagenet-1k 데이터셋으로 학습시키고, discriminator의 convolutional feature들을 maxpooling 하여 4\*4 spatial grid를 생성하였다. 이 피쳐를 flatten하여 L2-SVM classifier에 적용한 결과 82.8%의 정확도를 얻을 수 있었다. 이는 CNN(84.4%)보다는 살짝 떨어지지만, K-means의 성능(80.6%)을 상회한다.

라벨링이 부족한 데이터셋인 StreetView House Numbers dataset(SVHN)에 대해서도 DCGAN의 discriminator의 feature를 사용하였다. 위의 실험과 비슷하게, 모델을 준비하였다. 결과는 테스트 에러 22.48%로 SOTA를 달성하였다. 추가적으로, 이 결과가 DCGAN 내부에 있는 CNN 모델에 의한 결과가 아님을 실험을 통해 밝혀냈다.(DCGAN 내부의 것과 구조가 같은 CNN 모델로 학습시킨 경우 밸리데이션 에러가 28.87%였다.)

## 6. Investigating and Visualizing the Internalss of the Networks

첫번째로, latent space의 landspace를 이해하기 위한 실험을 진행하였다. latent space에서의 변화가 이미지 생성에서의 semantic 변화로 이어진다면, 모델이 relevant, interesting representation을 배웠음을 시사한다.



1번째 행은 latent space에서 직선 상에 있는 랜덤한 9개의 점을 시각화한 것이다. 모두 침실을 의미하는 것을 볼 수 있다. 6번째 행은 방 안의 창문이 조금씩 큰 창문으로 변하는 것을 확인할 수 있다.

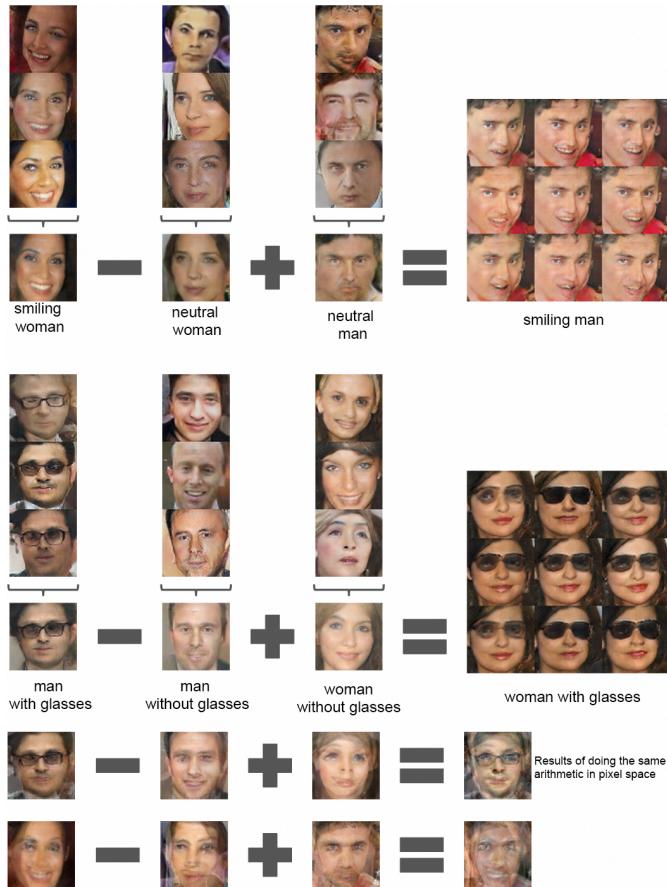
대량의 데이터셋으로부터 학습된 비지도학습 DCGAN은 계층적인 feature를 학습한다. guided backpropagation을 사용한 결과 아래 그림과 같이 discriminator가 학습한 feature들은 침대나 창문과 같이 침실의 특정한 부분들이 활성화되었다. 비교를 위해 random하게 초기화된 필터에서 활성화된 부분을 표시해보았지만 특별한 의미는 보이지 않는다.



이제, generator가 무엇을 표현하는지 알아보자. 150개의 샘플에 대해 52개의 창문에 bounding box를 수동으로 그린다. 두번쨰 CNN layer feature에 logistic regression을 적용하여 feature activation이 창문인지 아닌지를 예측한다. 그리고 window를 담당하는 필터를 제거한 결과, 창문은 사라지거나 문이나 거울과 같이 비슷한 모양을 가진 다른 물체로 대체되었다. 퀄리티는 조금 떨어지지만 나머지 부분은 거의 비슷하게 유지되었다.



워드 임베딩처럼 Z space에서 vector연산을 통해 얼굴의 포즈를 바꾸는 등 output을 조절할 수 있음을 확인하였다.



## 7. Conclusion and Future Work

본 연구에서는 GAN모델을 학습시키는 안정적인 구조를 제안하고, 모델이 지도학습을 위한 이미지의 좋은 representation을 학습할 수 있음을 보였다. 모델 불안정성 등의 문제가 남아있으나 후속연구가 이뤄질 것이라 기대한다. 이 프레임 워크를 비디오나 오디오 도메인에 확장시키는 것도 흥미로울 것이다.

참고

- <https://ysbsb.github.io/gan/2020/12/05/DCGAN.html>
- <https://kau-deeperent.tistory.com/79>