**특정인물 이미지 진위여부 판별을 위한 딥러닝 기법 적용**

**이영섭, 김현중, 송주원, 오현지**

*동국대학교 통계학과 데이터마이닝연구실*

*04620, 서울특별시 중구 필동로 1길 30*

*E-mail: yung@dongguk.edu, kimhj\_0220@daum.net*

#### 요약

*차세대 머신러닝 알고리즘으로 각광받고 있는 GAN*

*(Generative Adversarial Networks, 적대적 생성 네트워크)기술이 발전하면서 이미지 생성에 고도화된 인공지능(AI)이 사용되기 시작했다. 최근 연구 및 활용사례들은 GAN을 사용한 가짜 이미지 생성에 초점이 맞추어져 있고, 그에 따라 가짜 이미지들에 대해 우려를 표하는 시선들이 존재한다. 그러나 GAN을 다른 방면으로 활용하면 진짜 이미지와 유사한 가짜 이미지를 구분하는데 도움이 될 수 있을 것이다. 본 논문에서는 최신의 GAN 학습기를 활용하여 특정인물 판별에 대한 기존 CNN 분류기의 성능을 향상시키고자 하였다. 성능이 뛰어난 BEGAN 모델을 통해 진짜 같은 가짜 사진을 만들어 내었고, 이를 학습데이터로 사용하여 분류 모델을 만든 결과 기존의 타인 이미지만을 사용하여 만든 분류기의 경우보다 더 좋은 분류성능을 보임을 확인하였다.*

**키워드**

딥러닝; 전이학습; CNN; VGGNet; GAN; DCGAN; BEGAN;

##### 서론

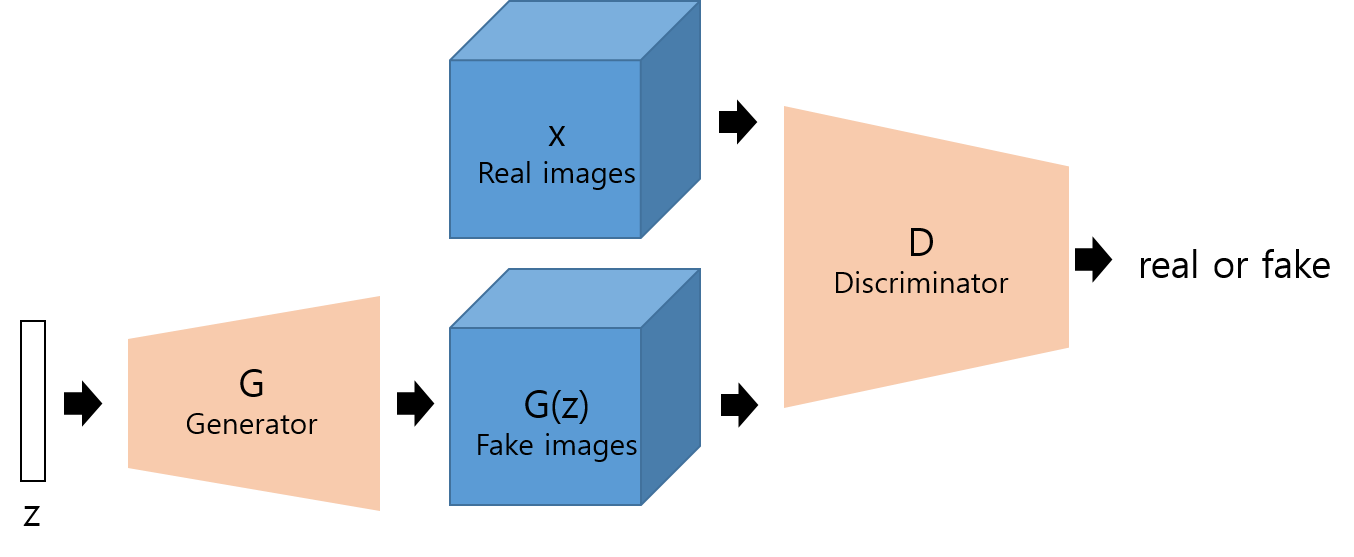
GAN(Generative Adversarial Networks)은 2014년 NIPS(Neural Information Processing Systems)에서 Ian Goodfellow가 제안한 이후 급속도로 많은 관심을 받게 되었는데, 그 중 이미지 생성에 대해 의미 있는 발전들이 진행되었다. DCGAN[3], EBGAN[4] 등 이미지 생성분야에서 GAN을 활용한 다양한 접근의 연구들이 발표되었으며, 이를 통해 보다 정교한 수준의 이미지 생성이 가능하게 되었다. 그러나 기술이 점차 발전하면서, 현실과 가상의 차이 구분이 어려울 정도로 정교한 이미지를 만들어내는 기술에 대해 우려를 표하는 시선도 적지 않다. 특히 국가 원수의 이미지를 조작하여 악용하는 경우, 국가 안보에 지대한 영향을 미칠 수 있기 때문에 이에 대한 대책이 필요하다. 우리는 이러한 문제의식에 따라 특정인물 이미지에 대한 진위 판별을 명확히 수행할 수 있는 분류기를 구축하고자 한다. 분류 모델은 이미지 분류에 탁월한 성능을 보이는 CNN 분류 학습기를 사용하기로 하였다.

특정인물의 이미지 진위 판별 분류기는 다음의 두가지 과제에 대한 접근이 필요하다. 첫째, 특정인물과 타인의 이미지 구별이 가능해야 한다. 둘째, 특정인물과 특정인물을 위조한 이미지를 구별해 내야한다. 따라서 분류기 학습을 위해서는 특정인물의 실제 이미지와 타인의 이미지, 특정인물을 위조한 이미지가 필요하다. 특정인물의 실제 이미지와 타인의 이미지는 쉽게 구할 수 있으나, 특정인물을 위조한 이미지는 확보에 한계가 있다. 이에 본 논문에서는 GAN학습기를 사용하여 특정인물을 위조한 이미지를 대량으로 생산하였고, 그 이미지들이 특정인물의 분류에 도움이 되는지 확인하고자 한다. 특히 특정인물에 대해 생산한 이미지의 품질 수준이 CNN 분류학습 결과에 영향을 미칠 것으로 판단하여 최근 발표된 GAN 모델들 중 성능이 뛰어난 BEGAN(Boundary Equilibrium Generative Adversarial Networks)[5] 모델을 사용하였다. CNN 분류학습으로는 학습 데이터 양의 한계를 고려하여 비교적 간단한 구조의 VGGNet[6]을 사용하였으며, 해당 분류기를 통해 특정인물의 실제 이미지를 타인 및 특정인물을 위조한 이미지와 구별하였다.

##### 이론적 배경

1. GAN(Generative Adversarial Networks)

GAN은 label이 주어지지 않는 비지도학습(Unsupervised learning)의 대표적인 알고리즘으로 두 개의 신경망 모델을 학습한다. 생성 모델 G는 임 의의 z를 이용하여 가짜 데이터 G(z) 생성해내고, 판별 모델 D는 G를 통해 생성된 가짜 데이터 G(z)와 진짜 데이터 x를 진짜인지 가짜인지 판별한다.



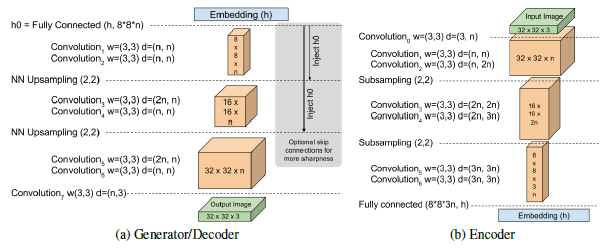
*그림 1 - GAN의 구조*

생성 모델은 진짜 같은 가짜 데이터를 만들어 내는 것, 판별 모델은 진짜 같은 가짜 데이터와 진짜 데이터를 구분하는 것을 목적으로 한다. 생성적 적대 신경망이라는 이름에 맞게 두 신경망 모델은 경쟁을 통해 학습하고 결과를 만들어 낸다.

2. BEGAN(Boundary Equilibrium GAN)

GAN을 이용하면 진짜 같은 가짜 이미지를 무한정으로 만들어 낼 수 있다. 그러나 기존의 GAN에는 아직 풀지 못한 숙제들이 많이 남아있다. 두 가지 모델에 대한 min-max problem을 풀어야 하므로 불안정한 구조를 띈다. 또한 적절한 hyper-parameter 선택이 매우 중요하며 생성되는 이미지의 다양성을 조절하기 힘들다. 무엇보다 생성되는 이미지의 품질이 좋지 않다. BEGAN은 GAN의 단점을 해결하기 위해 다양한 기법을 적용하였다. 먼저 판별 모델에서 sample 간의 직접적인 비교를 통한 학습이 아닌 auto-encoder 개념을 도입시켜 auto-encoder의 loss간 분포 차이를 이용하였다. 이 때 Wasserstein distance를 이용하여 더욱 안정적인 학습을 할 수 있게 만들었다.

를 pixel-wise auto-encoder를 훈련하면서 생긴 loss라 하면 식 (1)과 같이 정리할 수 있다.

 (1)

*그림 2 - BEGAN의 네트워크 구조*

(는 auto-encoder 함수, 는 target norm, 는 차원의 sample)

를 실제 sample x의 auto-encoder loss 의 분포, 를의 auto-encoder loss 의 분포라고 하고 를 각각의 평균이라고 하자. Jensen’s inequality를 이용하면 Wasserstein distance의 lower bound는 식(2)와 같다.

(2)

라 두면 최소화하려는 각 모델의 loss , 는 식 (3)과 같다.

(3)

또한 생성 모델과 구별 모델에 평형 개념을 도입시켰다. 이 평형과 관련된 hyper-parameter를 이용하여 이미지의 다양성을 조절할 수 있게 하였다.

3.VGGNet

2014년 ILSVRC(ImageNet Large-Scale Visual Recognition Challenge)에서 top-5 test error 7.3%를 기록한 VGGNet은 현재까지 이미지 분류를 위한 방법으로 많이 사용되고 있다. VGGNet은 가장 단순한 구조인 3x3 convolution layer를 쌓아서 사용하는 특징을 가지고 있으며, 이로 인해 parameter의 수를 줄이고 학습 속도를 빠르게 한다. 또한, layer의 수가 많아질수록 non-linearity가 더 증가하여 더 유용한 feature를 추출하는데 좋다고 알려져 있다. ILSVRC-2012의 데이터를 이용하였을 때 layer의 수를 다르게 하여 성능을 비교한 결과, 16-layer가 넘어가면 성능의 개선의 효과가 미미하므로 일반적으로 16-layer의 VGGNet을 많이 이용한다.

*그림 3 - 16layers의 VGGNet 구조*

4. 전이학습

적은 양의 데이터로 딥러닝 모델을 학습시킬 경우 모델 학습에 여러가지 어려움이 발생한다. 딥러닝 모델은 가중치 수가 많기 때문에, 부족한 정보를 사용하여 학습을 진행하는 경우 neural network가 overfitting될 수 있기 때문이다. 이러한 경우 전이학습이 대안이 될 수 있다. 전이학습은 대량의 데이터셋으로 이미 학습된 모델의 가중치를 새로운 데이터셋의 학습에 사용하여 모델의 성능을 개선시키는 방법이다. 이미지 학습에 대한 예시로는, 공개된 VGGNet을 소량의 나뭇잎 영상만 있는 상황에 전이하여 실용적인 나뭇잎 인식 프로그램을 만들어 앱 스토어에 공개한 사례 등이 있다.

##### 실험

1. 데이터 셋

웹 크롤링을 통해 특정인물의 이미지를 1250장 수집하였다. 이후 상하 좌우 5픽셀씩 이동을 하는 방식으로 데이터 증폭을 진행하여 10000장의 데이터 셋을 확보하였다. 또한 전이학습을 위하여 202,599장의 CelebA 데이터셋을 활용하였다.



*그림 4 – 특정인물 데이터셋*



*그림 5 – CelebA 데이터셋*

2. 실험 과정

CelebA 데이터셋을 이용하여 BEGAN 모델을 학습시킨다. 이후 특정인물의 이미지에 대하여 BEGAN 모델을 학습시키는데, CelebA 데이터셋을 통해 학습했던 가중치를 이용하여 전이학습을 진행한다. 이를 통해 특정인물과 닮은 위조 이미지를 무한정으로 생성해 낼 수 있게 된다.

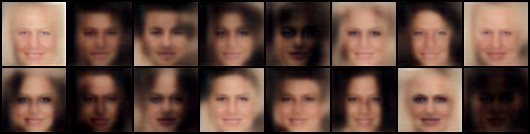
BEGAN을 통해 생성해낸 위조 이미지가 특정인물의 진위 판별에 도움이 되는 지 판단하기 위하여 두가지 VGG16 비교 모델을 만든다. 특정인물의 이미지(10000장)와 CelebA 이미지(10000장)를 분류하는 모델 하나, 특정인물의 이미지(10000장)와 CelebA 이미지(5000장)+BEGAN 위조 생성 이미지(5000장)를 분류하는 모델 하나이다. 이 두가지 모델을 만들 때에도 전이학습을 이용하는데 keras 라이브러리의 사전 학습 된 VGG16의 가중치를 이용한다. 테스트 데이터셋으로는 학습에 사용하지 않은 특정인물의 이미지 100장과 특정인물이 아닌 이미지 100장(특정인물과 닮은 사람의 이미지 50장+해당인물을 타인과 합성한 이미지 50장)을 사용하였다. 마지막으로 각 모델에 대하여 30번의 실험을 통해 결과를 확인하였다.

*표 1 - 각 모델의 hyper-parameter*

|  |  |  |
| --- | --- | --- |
|  | BEGAN | VGG16 |
| optimizer | Adam | Adam |
| Beta1(momentum term of adam) | 0.5 | 0.9 |
| Beta2(step size term of adam) | 0.999 | 0.999 |
| Batch size | 16 | 20 |
| learning rate | - | 0.001 |
| D\_learning rate | 0.00002 | - |
| G\_learning rate | 0.00002 | - |
| Gamma | 0.7 | - |

3. 실험 결과

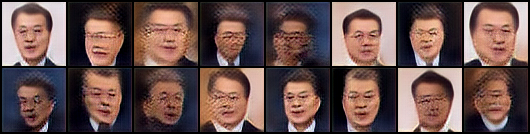
BEGAN을 활용하여 위조 이미지를 생성하였다. CelebA 데이터셋을 이용하여 5000번 학습 시킨 가중치를 전이학습을 통해 활용하고, 특정인물에 대하여 20000번의 학습을 통해 생성 모델을 완성시켰다.



*그림 6 – CelebA 5000번 학습 후 생성된 이미지*



*그림 7 – 특정인물 5000번 학습 후 생성된 이미지*



*그림 8 – 특정인물 10000번 학습 후 생성된 이미지*



*그림 9 – 특정인물 20000번 학습 후 생성된 이미지*

학습이 진행되면서 더욱 선명한 이미지를 추출하는 것을 확인할 수 있었다. 다음으로 생성 모델이 단순하게 이미지를 기억했다는 것이 아님을 증명하기 위해 z의 변화에 따라 부드러운 이미지 변화가 나타나는 것을 [그림 10]를 통해 확인하였다.



*그림 10 - 위조 이미지 둘 사이의 선형 보간*

20000번 학습시킨 후 특정인물에 대한 위조 사진을 5000개 생성해냈고, 이를 VGG16 모델에 사용하였다. 각 모델은 2000번씩 학습을 하였고, 일반적인 결과를 도출하기 위하여 30번 반복실험 후 평균값을 계산하였다. VGG16을 통한 각 모델의 분류 결과는 [표 2],[표 3]과 같다.

*표 2 – 특정인물 vs CelebA 모델의 교차표*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | 예측값 | | 합계 |
| 특정인물 | 특정인물  이 아님 |
| 실제값 | 특정인물 | 99.1 | 0.9 | 100 |
| 특정인물이 아님 | 62.8 | 37.2 | 100 |

*표 3 – 특정인물 vs (CelebA+위조) 모델의 교차표*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | 예측값 | | 합계 |
| 특정인물 | 특정인물이 아님 |
| 실제값 | 특정인물 | 99.2 | 0.8 | 100 |
| 특정인물이 아님 | 49.5 | 50.5 | 100 |

[표 2]는 특정인물의 이미지 10000장과 CelebA 데이터셋 10000장을 사용하여 만든 분류모델에 테스트 데이터셋을 넣은 결과이고, [표 3]은 특정인물의 이미지 10000장과 CelebA 데이터셋 5000장+BEGAN 위조 생성 이미지 5000장을 사용하여 만든 분류모델에 테스트 데이터셋을 넣은 결과이다. 본 연구의 목적은 특정인물이 아닌 이미지를 특정인물이 아니라고 잘 분류하는 것이다. CelebA만 사용하여 만든 모델의 경우 특정인물이 아닌 이미지를 특정인물이 아니라고 평균 37.2% 분류하였고, 표준편차는 9.8%이다. CelebA와 BEGAN 위조생성 이미지를 모두 사용하여 만든 모델의 경우 평균 50.5%, 표준편차는 8.2%이다. 위의 결과에 따라 분류모델 구축에 GAN으로 생성한 위조 데이터를 사용한 경우에 분류 성능이 향상됨을 확인하였다.

##### 결론

특정인물의 이미지에 대한 진위 판별기를 이미지 분류에 특화된 VGG16 모델을 이용하여 만들었다. 또한 BEGAN 모델을 통해 실제 이미지와 비슷한 위조 이미지를 생성해냈다. 생성된 위조 이미지를 이용하여 판별기를 학습한 결과 특정인물의 이미지 판별 성능을 향상시킨다는 결과를 보였다. 이를 통해 분류의 관점에서 GAN을 활용할 수 있음을 확인할 수 있었고, 다음과 같은 보완을 수행하면 더 큰 성능 향상을 기대할 수 있을 것이다. 첫째, 특정인물의 이미지 품질을 향상시켜야 한다. CelebA 데이터셋의 경우 눈의 위치, 얼굴의 크기 등 특징이 대부분 일치하였다. 하지만 본 연구에서 추출한 이미지의 경우 웹 크롤링을 통해 다양한 각도에서 촬영된 이미지를 사용하였기 때문에 이러한 특징들의 위치를 섬세하게 조절하지 못하였다. 둘째, 학습에 사용된 CelebA 데이터셋과 판별하려는 특정인물간 특징 차이를 줄여야 한다. CelebA 데이터셋은 세계 유명인사들의 이미지로써 주로 서양인들의 사진으로 구성되어 있기 때문에 한국인의 특징이 많이 누락된다. 따라서 이 연구에서 사용한 특정인물의 이미지 특징과 많은 차이가 존재하게 되었고, CNN 분류기가 판별을 하는데 정확한 기준점이 생기지 못했다. 더불어 학습 데이터셋을 충분히 확보한다면 더욱 큰 효과를 기대할 수 있다.

#### 참고문헌

[1] Ian J. Goodfellow. (2014). “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661v1*

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. (2017). “Generalization and equilibrium in generative adversarial nets (gans),” *arXiv preprint arXiv:1703.00573*

[3] Alec Redford, Luke Metz, Soumith Chintala. (2016). “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv preprint arXiv:1511.06434v2*

[4] Zhao, Junbo, Michael Mathieu, and Yann LeCun. (2016). “Energy-based generative adversarial network,” *arXiv preprint arXiv:1609.03126*

[5] David Berthelot, Thomas Schumm, Luke Metz. (2017). “BEGAN: Boundary Equilibrium Generative Adversarial Networks,” *arXiv preprint arXiv:1703.10717*

[6] Karen Simonyan, Andrew Zisserman. (2014). “Very Deep Convolutional Networks for Large-scale Image Recognition,” *arXiv preprint arXiv:1409.1556*

[7] Martin Arjovsky, Soumith Chintala, Léon Bottou. (2017). “Wasserstein GAN,” *arXiv preprint arXiv:1701.07875v3*

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang. (2015). “Deep learning face attributes in the wild,”  *arXiv:1411.7766v3*