합성 이미지에 대한 기존 머신 러닝 이미지 분류 모델의 성능 비교

*정윤진, **한지형
서울과학기술대학교 컴퓨터공학과
*yjj_@seoultech.ac.kr, **jhhan@seoultech.ac.kr

Comparison of Machine Learning Models for Image Classification on Composite Images

*Jeong, YoonJin, **Han, Ji-Hyeong
Department of Computer Science and Engineering
Seoul National University of Science and Technology

요 약

증강현실은 현실 공간에 가상의 객체를 합성한 영상을 생성하는 기술이다. 증강현실 기술에 대한 지속적인 수요 증가와 기술 발전이 이루어져 왔으며, 앞으로 사용자에게 현실을 기반으로 생성된 이질감이 느껴지지 않는 정교한 영상을 제공할 수 있으리라 기대할 수 있다. 본 논문에서는 증강현실 기술로 생성된 합성 영상이 정교한 영상임을 판단할 수 있는 객관적인 기준을 마련하기 위해 기존의 머신 러닝 기반의 이미지 분류 모델들로 합성 이미지 예측에 대한 실험을 진행하고 그 결과를 비교한다.

1. 서론

증강현실(Augmented reality, AR)은 현실의 공간을 기반으로 컴퓨터 그래픽 기술을 통해 가상의 객체를 합성하여 현실에서 확장된 새로운 영상을 생성하는 기술이다. 코로나 19 이후 언택트(Untact) 시대를 맞이하며 비대면 서비스에 대한 수요가 증가함에 따라 증강현실 기술에 대한 수요도 증가하고 있다. 이에 과학기술정보통신부도 원격회의협업·교육·유통 등 VRAR 을 활용한 비대면 서비스 개발에 80 억을 지원하였으며, 정보통신산업진흥원은 22 년 AR 세계 디바이스 시장규모에 대해 170.2 억 달러로 전망하고 있다 [1]. 증강현실 기술은 교육, 의료, 국방, 게임, 콘텐츠 제작 등 다양한 분야에 활용되며, 사용자에게 현실에서 확장된 경험을 제공하기 위해 계속해서 발전하고 있다. 증강현실 기술로 생성된 영상이 점점 현실과 가깝게 정교해지리라 기대할 수 있다. 증강 현실 기술로 생성된 영상의 정교한 수준에 대해서 휴리스틱(heuristics)하게 판단할 수 있지만, 머신 러닝 기술을 사용하여 판단할 수도 있다. 합성 이미지인지 객관적으로 판단하는 기준이 존재한다면, 현재 AR 기술 수준을 판단하고 더욱 발전시키는데 도움이 될 수 있다.

본 논문에서는 기존의 이미지 분류를 위한 머신 러닝 모델이 합성 이미지인지 판단할 수 있는가에 대해 실험하고 그 결과들을 비교한다. 여러 모델로 실험함으로써 어떤 환경에서 합성 이미지 분류를 잘 하는지 파악한다. 합성 이미지 분류는 이진 분류로 파라미터 수가 너무 많은 모델은 오히려 과적합 문제가 발생할 수 있다. 그러나 단순한 이진 분류와는 다르게 뚜렷한 특징이 없어 정교한 패턴 분석이 필요하기 때문에 오히려 많은 파라미터 수를 요구할 수 있다. 실험을 통해 어떤 모델이 합성 이미지 분류에서 좋은 성능을 보이는지 연구한다.

본 논문의 구성은 다음과 같다. 2 절에서는 전체 실험 프로세스와 실험환경 및 데이터셋을 설명한다. 3 절에서는 실험결과를 비교하고, 마지막으로 4 절에서는 결과에 따른 결론을 맺는다.

2. 합성 이미지 분류

전체 실험은 기존의 이미지 분류 모델로 데이터를 학습하는 것으로 구성된다. 실험에 사용되는 이미지 분류 모델은 VGG[2], GoogleNet[3], ResNet[4], SENet[5]이다. VGG 는 모든 Convolution 레이어에 3×3 크기의 필터를 사용했으며, 본 논문의 실험에서는 19 개의 레이어를 쌓은 VGG19 를 사용한다. GoogleNet 은 다양한 필터 크기를 사용하는 것이 특징이며, Vanishing Gradient problem 을 해결하기 위해 Auxiliary loss 를 구해 총 loss 를 계산하는데 이용한다. ResNet 은 Residual Block 을 도입해 Vanishing Gradient problem 을 해결했으며, 깊은 네트워크가 가지는 많은 파라미터의 부담을 줄이기 위해 Bottleneck 구조를 채택했다. SENet 은 Convolution layer 에서 생성된 feature map 에 대해서 Global Average Pooling(GAP) 레이어를 사용하여 Squeeze 를 수행한다. 그 다음 Excitation 과정에서 채널 간 의존성을 파악하기 위해 fully connected layer 와 비선형 함수를 사용한다. SENet 은 다른 모델과 결합해 사용될 수 있어 본 논문의 실험에서는 ResNet 과 결합해 학습했다.

각각의 모델은 동일한 학습 데이터셋에 대해서 독립적으로 학습을 진행하며, 학습이 완료되면 동일한 테스트 데이터셋으로 모델이 합성 이미지라고 예측하였는지에 대해 정확도로 평가한다. 기존 모델이 가지는 클래스 수를 데이터셋에 맞게 수정하기 위해 새로운 Fully connected layer 을 추가하지 않고 모델이 가진 기존의 마지막 Fully connected layer 의 출력 수를 수정했으며, 각각의 모델에 대해서 사전 훈련된 모델의 가중치를 사용하지 않았다

데이터셋은 데이터를 공유하는 사이트 캐글(Kaggle)에서 합성 이미지와 관련된 데이터셋으로 제공하는 'Real and Fake Face Detection'이다. 이 데이터셋은 합성되지 않은 사진과 합성한 얼굴 이미지가 포함되어 있다. 실험을 위해서 2,041 개의 데이터셋에 대해서 9 대 1 의 비율로 학습 데이터셋과 테스트 데이터셋을 나누었다.

3. 실험결과

Real and Fake Face Detection 데이터셋으로 기존 모델을 학습한 결과는 표 1 과 같다. 먼저 전반적으로 상당히 많은 레이어(파라미터)를 가진 ResNet152 와 SENet152 보다 훨씬 적은 레이어(파라미터)를 가진 VGG19 와 GoogleNet 이 더 좋은 성능을 보였다. 그리고 GoogleNet 의 정확도가 60.78%으로 가장 좋은 성능이 나왔다. 실제 이미지와 합성 이미지를 예측한 각각의 정확도 또한 65.74%와 55.21%로, 실제 이미지에 대한 정확도는 가장 높으며 합성 이미지에 대한 정확도는 다른 모델들과 비교했을 때 나쁘지 않은 결과를 보인다. 한편 ResNet152 는 정확도가 53.92%로 가장 낮지만, 합성 이미지에 대한 정확도는 다른 모델에 비해 가장 높다.

표 1. 합성 이미지 분류 결과

Model	Total accuracy(%)	Accuracy of real(%)	Accuracy of fake(%)
VGG19_bn	59.31	62.04	56.25
GoogleNet	60.78	65.74	55.21
ResNet152	53.92	50.93	57.29
SENet152	57.84	60.19	55.21

레이어가 적은 모델들이 좋은 성능을 보였기 때문에, ResNet 과 SENet 구조를 가지면서 레이어가 적은 모델로 추가적인 실험을 진행했다. 그 결과는 표 2 와 같다. ResNet 은 더 적은 레이어를 가질 때 더 좋은 성능을 보였지만, SENet34 는 그렇지 않았다. 그리고 ResNet34 의 경우 실제 이미지와 합성 이미지에 대한 정확도가 58.33%로, 각각의 이미지에 대해서 안정적으로 예측했다.

표 2. ResNet18, 34 와 SENet18, 34 모델 학습 결과

Layer	Model	Total accuracy(%)	Accuracy of real(%)	Accuracy of fake(%)
18	ResNet18	57.84	60.19	55.21
	SENet18	54.9	59.26	50.0
34	ResNet34	58.33	58.33	58.33
	SENet34	51.96	57.41	45.83

4. 결론

대부분의 모델이 실제 이미지와 합성 이미지를 50%보다 높은 확률로 예측하였다. 보다 많은 데이터로 더 여러 번 학습을 거치면, 기존의 이미지 내에 존재하는 객체에 따라 분류하는 모델들이 합성된 이미지도 분류할 수 있다고 기대할 수 있다. 결과적으로 증강현실 기술로 만들어진 영상에 대해서 실제 현실과 이질감이 느껴지는지 혹은 현실과 매우 비슷한 지 판단할 수 있을 것이다. 또한 뚜렷한 특징이 적어 정교한 패턴 분석이 필요한 합성 이미지를 분류해야 하지만, 실제 이미지인지 합성 이미지인지 두 가지만 구분해내면 되기 때문에 복잡하고 깊은 레이어는 불필요하다는 것을 알 수 있었다.

Acknowledgement

This work was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2018R1C1B6007230).

참고문헌

[1] 이혁준, VR·AR 디바이스 동향 및 시사점, 정보통신산업진흥원, 2020.

[2] Simonyan, K. and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv, 2015.

[3] Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, “Going deeper with convolutions”, In CVPR, 2015.

[4] He, K., X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, In CVPR, 2016.

[5] Hu, J., L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, arXiv, 2017.