특성기여도 분석 방법을 이용한 자동 증강의 영향에 관한 연구

김민기, 김재일
경북대학교
alsrlehd2@knu.ac.kr, threeyears@gmail.com

A Study on the Effect of Auto Augmentation using Feature Attribution Methods

Mingi Kim, Jaeil Kim
Kyungpook National University, Korea

요 약

데이터 증강은 데이터의 변하지 않는 특징을 학습하는 것을 목표로 한다. 본 논문에서는 특성기여도 방법을 바탕으로 자동 증강이 어떤 특징을 학습하고 있는지 확인하고, 특성기여도 방법을 사용했을 때의 한계점을 제시한다.

Ⅰ. 서 론

심층신경망(Deep neural nets)은 많은 양의 데이터에서 복잡한 패턴을 학습하며, 다양한 임무(Task)에서 높은 성능을 달성하고 있다. 하지만, 의료영상을 활용한 진단 보조를 비롯하여 실제 현장에서는 많은 양의 데이터를 확보하기 어렵기 때문에, 적은 데이터로부터도 일반화 성능이 높은 모델을 만드는 것은 중요한 문제이다. 이 문제를 해결하기 위해, 데이터 증강(Data Augmentation)이 많이 활용되고 있다. 데이터 증강에서는 이동(Translation), 좌우 뒤집기(Horizontal Flipping), 회전(Rotation)이 많이 사용되며, 이들은 데이터의 변하지 않는(Invariant) 특징을 학습하는 것을 목표로 한다. 데이터 증강은 데이터의 양과 다양성 모두를 증가 시킬 수 있는 방법이고, 선행 연구들로부터 분류문제에서 일반화 성능을 높일 수 있다는 것이 밝혀졌다.

데이터 증강을 적용했던 기존 연구에서 데이터 증강에 대한 강도(Magnitude)를 실험적으로 설계(Design)하는 경우가 많았고, 비용이 많이 드는 작업으로 여겨졌다. 최근 자동 증강(Auto Augmentation), Fast Auto Augmentation 논문에서 데이터의 미니배치(mini-batch)마다 데이터 증강을 다르게 적용하고, 이들의 오차율을 최소화 하는 방향으로 데이터 증강을 찾는 연구가 진행되었다 [1,2].

본 연구에서는 특성기여도 분석 방법(Feature Attribution Methods)을 사용하여 자동 증강이 어떤 특징을 학습하고 있는지 시각적으로 확인하고, 한계점을 제시한다. 여기서 특성기여도 분석 방법이란 입력이미지의 어느 부분이 모델 예측결과에 큰 영향을 미치는지 시각적으로 확인하는 방법을 일컫는다.

Ⅱ. 본론

1. 데이터셋과 모델 구성

CIFAR10[9] 데이터셋을 사용하여 실험을 진행하였다. 성능의 최종 테스트에 사용할 테스트 데이터셋(Test Dataset)을 제외한 50000개의 데이터에서 7500개의 데이터를 검증 데이터셋(Validation Dataset)으로 사용하여 초모수조정(Hyper-Parameter Tuning)을 수행한다. 학습에는 WideResNet(depth=4, widen_factor=2) 모델[8]을 사용했다.

2. 자동 증강

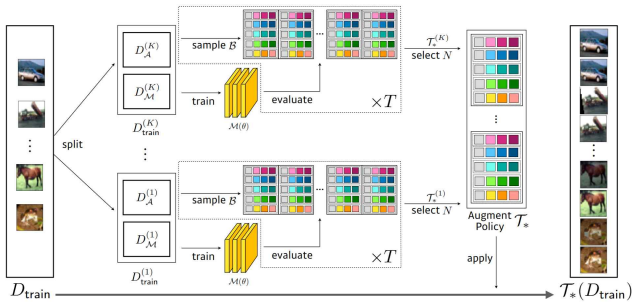

그림 1. 자동 증강 방법.

위 그림은 [2]에서 사용된 그림이다. 훈련 데이터셋(Train Dataset)을 K개의 구획(Fold)로 나누고, 각각의 구획에서 D_M과 D_A로 다시 나눈다. D_M을 이용하여 모델을 훈련 시킨 후, 초기 데이터 증강 T를 적용한 $T(D_A)$를 이용하여 오차율을 계산한다. 그리고 각 구획에서 가장 낮은 오차율을 가지는 N개의 데이터 증강 들을 모아 새로운 데이터 증강 T_*를 구성한다. 그리고 위의 과정을 반복하여 최종적인 데이터 증강 T_*를 확정한다.

3. 특성기여도 분석 방법의 구성

자동 증강이 어떤 특징을 학습하는지 시각적으로 확인하기 위하여, Grad-Cam[3], Smooth-grad[4], Integrated Gradient[5], Guided Backpropagation[6] 방법을 사용했다.

4. 실험결과

CIFAR10 데이터셋으로 Fast Auto Augmentation 방법을 사용하여 데이터 증강을 수행했다. 모델을 훈련시킨 결과 테스트 셋에서의 상위 1개 오차율(Top1 error)은 3.62%로 나타났고, Fast Auto Augmentation 방법을 사용하지 않았을 때의 상위 1개 오차율은 11.13%로 나타났다.

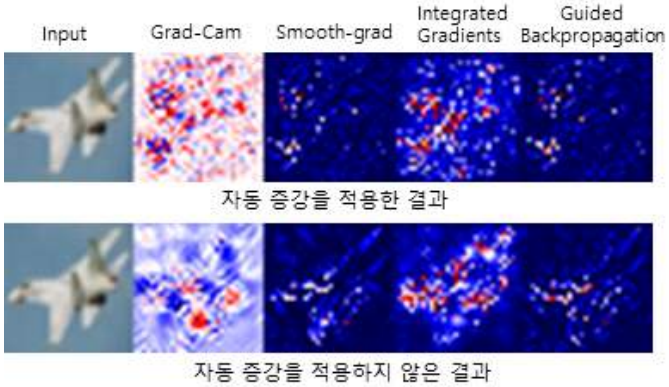

그림 2. 특성기여도 분석 방법을 이용한 자동 증강 효과 분석.

자동데이터 증강을 적용하지 않았을 때, 물체를 구별할 수 있는 특징을 강조하고, 자동데이터 증강을 적용했을 때, 물체의 특징 주변 배경을 강조하는 실험결과를 얻을 수 있다.

5. 실험결과 분석

자동 증강이 학습과정에서 배운 물체의 고유한 특징이 특성기여도 분석 방법을 통해 나타날 것이라 가정했었지만, 물체 주변을 강조하는 결과를 얻을 수 있었다. 위 결과를 얻게 된 원인을 세 가지로 추정할 수 있다. 첫 번째는 자동 증강을 통해 물체의 고유한 특징과 물체가 가지고 있는 맥락(Context)을 함께 학습했다는 것이다. 두 번째는 모델이 미니배치 내부의 검증 데이터셋의 오차율을 줄이는 방향으로 학습하는 특징과 인간이 중요하게 생각하는 특징이 다르다는 것이다. 세 번째는 특성기여도 분석 방법으로 데이터 증강이 학습할 때 배우는 특징을 정확하게 시각화 할 수 없다는 것이다. [7]에서 제시된 것처럼 특성기여도 분석방법은 적대적 공격(Adversarial Attack)에 취약하고, 모델의 입장에서 데이터 증강은 적대적 공격이라고 생각될 수 있다.

Ⅲ. 결론

자동 증강을 적용하여 모델을 학습시킬 때, 일반화 성능이 증가하는 것은 이미지를 구별할 수 있는 고유한 특징을 학습하기 때문이라고 생각된다. 이를 관찰하기 위해, 자동 증강을 적용한 모델에서 입력이미지의 기여도를 특성기여도 분석 방법을 통해 나타내어 보았다. 자동 증강을 적용하지 않은 모델은 물체의 구별할 수 있는 특징을 나타내는 반면, 자동 증강을 적용한 모델은 주변 배경을 강조하는 것이 관찰되었다. 하지만, 실험결과 분석에서 서술한 것처럼 다양한 해석이 있을 수 있고, 다른 실험조건(데이터셋과 모델의 변화)에서 실험을 수행하지 않았기 때문에 자동 증강이 추출하는 특징을 정확하게 관찰하기 위해서는 추가적인 연구가 필요할 것으로 사료된다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(“2020R1I1A3074639”)

References

[1] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

[2] Sungbim Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. arXiv:1905.00397 [cs.LG], 2019.

[3] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[4] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017.

[5] Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi. Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365, 2017.

[6] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv 2014

[7] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. arXiv preprint arXiv:1710.10547, 2017.

[8] S. Zagoruyko and N. Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.

[9] A. Krizhevsky. Learning multiple layers of features from tiny images. Tech Report, 2009.