지도 학습의 분류 작업에서 이미지 데이터의 정보 보호에 관한 연구

김세환, 이정우
서울대학교

{sehwankim, junglee}@snu.ac.kr

A Study to protect privacy of image data in Classification task of Supervised Learning

Sehwan Kim, Jungwoo Lee
Seoul National Univ.

요 약

지도 학습의 분류 작업(Classification task of Supervised Learning)은 이미지 데이터와 그에 해당하는 라벨이 학습 데이터로 주어졌을 때, 이를 학습하여 학습 데이터에 존재하지 않는 테스트 데이터에 대해 높은 정확도를 바탕으로 분류 작업을 수행하는 것을 목표로 한다. 하지만 기존의 분류 작업의 경우, 학습 과정에서 학습 데이터가 외부에 노출되는 상황이 발생할 수 있다. 이러한 취약점을 해결하기 위해, 분류 작업을 하기 전에 학습 데이터에 변형을 가해 인간의 눈에는 식별이 불가능하지만 인공신경망의 분류 능력에는 영향을 주지 않거나 최소화하는 학습 방법을 제안한다. 또한, 데이터 증강 기법과 하이퍼파라미터의 조정이 앞에서 제안한 학습 방법에 어떤 영향을 주는지 성능을 비교할 것이다.

Ⅰ. 서론

본 논문에서 우리는 학습 데이터에 변형을 가해 인간이 보기에는 형체를 알아볼 수 없지만 인공신경망의 분류 능력에는 영향을 주지 않는 학습 방법을 제안한다. 기존의 지도 학습(Supervised Learning)에서 분류 작업(Classification task)의 학습 과정에서는 학습 데이터가 분류 작업을 담당하는 인공신경망에 직접 들어가기 때문에 데이터가 외부에 노출될 수 있는 위험이 있다. 특히 학습 데이터가 의료 데이터와 같이 인간의 개인 정보가 담겨있는 데이터일 경우 큰 문제를 일으킬 수 있다. 이러한 취약점을 해결하기 위해, 인공신경망을 통해 학습 데이터 의 형태를 바꿔 인간의 눈으로는 알아볼 수 없게 만들되 기존의 분류 작업에는 영향이 없거나 거의 존재하지 않게 하는 방법을 제안한다. 이러한 학습 방법을 통해 지도 학습에서 분류 작업의 학습 데이터가 외부에 노출이 되더라도 인간의 눈으로는 알아볼 수 없는 형태이기 때문에 학습 데이터의 정보를 보호할 수 있다. 또한, 이 과정에서 데이터 증강 기법을 사용할 경우 더 높은 분류 정확도를 확인할 수 있다.

Ⅱ. 본론

본 논문에서는 원본 이미지 데이터를 인간의 눈에는 식별할 수 없게 변형시킨 뒤, 변형된 이미지 데이터를 바탕으로 분류 작업을 진행하는 학습 방법을 제안한다.

구조

그림 1 의 전체적인 구조도를 보면, 우선 원본 이미지 데이터가 입력으로 들어가 이미지 변형을 위한 인공신경망을 거쳐 변형된 이미지로 바뀌게 된다. 이 변형된 이미지를 다시 분류 작업을 위한 인공신경망에 넣어 분류 작업을 진행하게 한다. 학습과정에서 2 개의 손실함수가 존재하는데, 첫번째는 기능 재구성 손실함수(Feature Reconstruction Loss Function)이다. 식 1 이 이를 나타낸 식인데, C*H*W 는 네트워크의 k 번째 층의 기능맵(Feature map)의 크기를 나타낸다. 이 손실함수를 통해 원본 이미지 데이터와 변형된 이미지 데이터를 비교하여 손실을 구한다. 두번째 손실함수는 기존의 분류 작업에서 많이 사용되는 교차 엔트로피 손실함수 (Cross Entropy Loss Function)이다. 교차 엔트로피 손실함수는 예측된 라벨과 정답인 라벨의 값을 비교하여 손실을 구한다. 전체 손실은 기능 재구성 손실함수를 통해 구한 손실에 하이퍼파라미터인 Alpha 를 곱해 교차 엔트로피 손실함수를 통해 구한 손실에 더한 값이다. 이렇게 구한 전체 손실을 통해 전체 네트워크를 학습시킨다.

$$L_{feat}(x_i, \hat{x}_i) = \frac{1}{C_k H_k W_k} \|\phi_k(\hat{x}_i) - \phi_k(x_i)\|_2^2$$

식 1. 기능 재구성 손실함수

실험 환경 및 결과

우선, 데이터셋으로 45000 장의 학습 데이터와 5000 장의 검증 데이터, 10000 장의 테스트 데이터로 이루어진 CIFAR-10 을 사용했다. 또, 데이터 증강 기법으로는 32 x 32 크기의 이미지로 무작위로 자르는 기법, 0.5 의 확률로 좌우 반전, 0.5 의 확률로 상하 뒤집기를 사용했다. 이미지 변형을 위한 인공신경망으로는 U-Net 을, 분류 작업을 위한 인공신경망으로는 ResNet-20 을 사용했다. 전체 네트워크를 학습시킬 때, 가중치 감쇠(Weight decay)는 0.0005, 운동량(Momentum)은 0.9 인 SGD(Stochastic Gradient Descent)를 사용하여 200 에폭동안 학습시켰다. 또한 학습률은 초기에는 0.1 로 주어지고 60, 120, 160 에폭마다 0.2 를 곱했다. 배치 크기로 128 을 사용했다.

그림 2 를 보면, 원본 데이터와 이미지 변형을 위한 인공신경망을 거쳐 나온 변형된 데이터를 시각적으로 확인할 수 있다. 원본 데이터를 알아볼 수 없을 만큼 데이터를 성공적으로 변형했음을 확인할 수 있다. 또한, 변형된 데이터로 분류 작업의 성능이 잘 나와야 하는 것이 핵심인데, 표 1 을 보면, 분류 작업의 성능 역시 이미지 변형을 거치지 않은 경우(Alpha = 0)과 비교해 매우 작은 크기만 감소했음을 확인할 수 있다. 또한 데이터 증강 기법의 사용 유무에 따른 성능 비교도 하였는데, 데이터 증강 기법(RandomCrop, RandomHorizontoalFlip, RandomVerticalFlip)을 사용한 경우 각각의 Alpha 값에 대해 약 5~7 %p 의 증가를 확인할 수 있다.

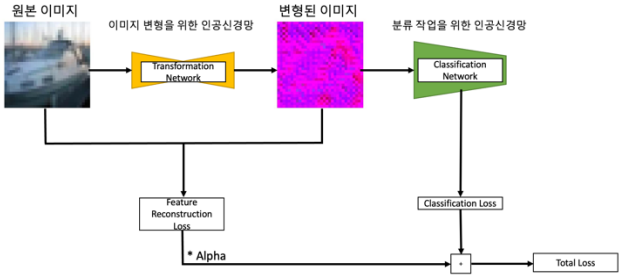

그림 1. 전체적인 구조도

그림 2. 원본 데이터와 인공신경망을 통해 변형된 데이터 비교

Accuracy (%)	Alpha	0	1000	2500	5000	10000
Augmentation Use	Yes	93.38	91.91	91.90	93.35	92.40
	No	86.40	86.98	84.75	87.90	87.41

표 1. 하이퍼파라미터 (Alpha)와 데이터 증강 기법의 사용 유무에 따른 분류 정확도

Ⅲ. 결론

본 논문에서 우리는 학습 데이터에 변형을 가해 인간의 눈에는 식별 불가능하지만 인공신경망의 분류 능력에는 영향을 주지 않거나 최소화하는 학습 방법을 제안했다. 또한, 이 과정에서 데이터 증강 기법이 분류 작업의 성능을 올리는데 큰 영향을 끼침을 확인했다. 우리의 방법을 통해 기존의 분류 작업에 비해 성능 감소를 최소화 시키면서 이미지 데이터의 정보를 보호할 수 있다.

ACKNOWLEDGEMENT

This work is in part supported by Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-00106 (30%), 2021-0-00180 (40%), 2021-0-02068 (30%)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21-plus.

참 고 문 헌

[1] H. Ito, Y. Kinoshita, M. Aprilpyone and H. Kiya, "Image to Perturbation: An Image Transformation Network for Generating Visually Protected Images for Privacy-Preserving Deep Neural Networks," in *IEEE Access*, vol. 9, pp. 64629-64638, 2021.

[2] W. Sirichotedumrong, Y. Kinoshita and H. Kiya, "Pixel-Based Image Encryption Without Key Management for Privacy-Preserving Deep Neural Networks," in *IEEE Access*, vol. 7, pp. 177844-177855, 2019.

[3] M. Tanaka, "Learnable Image Encryption," *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 1-2, 2018.

[4] Madono, Koki, et al. "Block-wise scrambled image recognition using adaptation network." *arXiv preprint arXiv:2001.07761* (2020).

[5] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention.* Springer, Cham, 2015.