레이어 프루닝을 이용한 생성적 적대 신경망 모델 경량화 및 성능 분석 연구

김동휘, 박상효, *배병준, *조숙희
경북대학교, *한국전자통신연구원
paka96@knu.ac.kr, s.park@knu.ac.kr, *1080i@etri.re.kr, *shee@etri.re.kr

Optimization And Performance Analysis Via GAN Model Layer Pruning

Dong-hwi Kim Sang-hyo Park *Byeong-jun Bae *Suk-hee Cho
Kyungpook National University *ETRI

요 약

딥 러닝 모델 사용에 있어서, 일반적인 사용자가 이용할 수 있는 하드웨어 리소스는 제한적이기 때문에 기존 모델을 경량화 할 수 있는 프루닝 방법을 통해 제한적인 리소스를 효과적으로 활용할 수 있도록 한다. 그 방법으로, 여러 딥 러닝 모델들 중 비교적 파라미터 수가 많은 것으로 알려진 GAN 아키텍처에 네트워크 프루닝을 적용함으로써 비교적 무거운 모델을 적은 파라미터를 통해 학습할 수 있는 방법을 제시한다. 또한, 본 논문을 통해 기존의 SRGAN 논문에서 가장 효과적인 결과로 제시했던 16 개의 residual block 의 개수를 실제로 줄여 봄으로써 기존 논문에서 제시했던 결과와의 차이에 대해 서술한다.

Keyword: Deep learning, generative adversarial networks, layer pruning, optimizing, lightweight

1. 서론

인공지능 및 딥 러닝 분야에서 활발한 연구가 진행중인 생성적 적대 신경망 모델(Generative Adversarial Networks, GAN)은 여러 분야에서 적용하기 위한 노력을 하고 있다. 그 중, 이미지 처리 분야인 초 해상화 분야에서도 과거에 사용된 보간법을 대체하여 생성적 적대 신경망을 통해 이미지 혹은 동영상 컨텐츠의 화질 향상 기법을 해결하려 노력을 하고 있는데 [1, 2], 하지만, 생성적 적대 신경망의 경우 기존의 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network, CNN)를 이용한 초 해상화 기법인 SRCNN [4] 보다 비교적 신경망이 깊고 더 많은 학습 파라미터가 존재하는데, 레이어 프루닝을 통해 모델 자체를 경량화 시킴으로써 학습 파라미터를 줄이고, 그에 대한 결과론 적인 차이를 서술한다.

본 실험의 내용으로는기존의 SRGAN [1] 에서 사용되는 Residual block 의 개수의 경우, 16 개로 설정했을 때의 훈련 시간과, 속도를 계산했을 때 가장 효율적이라 나타냈기에, 이를 비교 대상 모델로 설정하여 Residual block 의 개수를 달리 설정하여 기존 모델과의 성능의 변화에 대해 기술한다.

2. 모델 경량화 연구 및 결과

모델 구현은 TensorFlow-keras 를 이용했다. 기존의 SRGAN [4] 모델은 ImageNet [4] 을 사용하여 실험을 진행했으나, 본 실험에서는 리소스의 한계가 있어 DIV2K 데이터셋을 사용하여 모델 생성 및 검증과정을 거쳤다. 또한, DIV2K_train_HR 데이터셋 [6] 을 이용하여 훈련을 실시했으며, 타겟 데이터셋은 bicubic 보간법으로 생성된 DIV2K_train_LR_bicubic_X4 를 이용하여 실험을 진행했다.

표 1 데이터 셋

훈련 데이터 셋	DIV2K_train_HR [6]

타겟 데이터 셋	DIV2K_train_LR_bicubic_X4
검증 데이터 셋	DIV2K_valid_HR

표 2 프루닝 결과

	Trainable Parameters	PSNR(↑)	SSIM(↑)
SRGAN (Residual block 16)	1,543,074	25.28	0.6446
Model#1 (Residual block 12)	1,238,943	25.88	0.7043
Model#2 (Residual block 8)	934,827	26.06	0.7076
Model#3 (Residual block 4)	630,711	24.62	0.5651
Model#4 (Residual block 2)	478,653	24.33	0.5187

본 논문에서 인용한 SRGAN [1] 본문에서는 Residual block 개수를 0 개부터 25 개까지 점차 개수를 늘렸을 때 성능이 점진적으로 좋아지는 결과를 나타냈으나, 실험 결과, Residual block 을 75% 줄였을 때와 같이 극단적으로 줄였을 때 성능이 1dB 이상 차이나는 결과를 보이고 있다.

3. 결론

본 논문은 하드웨어 추가를 통한 성능 향상이라는 딥 러닝 모델의 한계점에 집중하기 보다는 프루닝을 통하여 일반적인 이용자의 제한적인 리소스 측면을 고려하여, 레이어를 줄여 봄으로써 효과적으로 성능을 향상시키는 것이 목적이다. 점진적으로 성능이 향상되는 결과를 보이는 본래의 SRGAN 논문의 실험과는 달리, 절반까지 줄였을 때의 결과는 본 모델의 결과보다 오히려 결과가 더 좋게 나오는 결과도 있었다. 그러나, 극단적으로 Residual block 수를 줄인 모델의 경우, Image quality assessment (IQA) 의 결과가 크게 하락하게 되었다. 본 논문의 결과를 통하여, 원래 모델보다 비교적 적은 개수의 Residual block 을 사용하더라도 효과적인 결과를 나타낼 수 있음을 입증하였다.

Acknowledgment

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00087, SD/HD 급 저화질 미디어의 고품질 변환 기술 개발)

참 고 문 헌 (References)

[1] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681-4690

[2] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295-307, 2016.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017. 3, 6

[4] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In International Conference on Learning Representations, 2019c.

[5] Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2017) 126-135