레터논문 (Letter Paper)
방송공학회논문지 제28권 제3호, 2023년 5월 (JBE Vol.28, No.3, May 2023)
https://doi.org/10.5909/JBE.2023.28.3.333
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

지식 증류 기법을 사용한 트랜스포머 기반 초해상화 모델 경량화 연구

김 동 현[a], 이 동 훈[a], 김 아 로[a], Vani Priyanka Gali[a], 박 상 효[a]‡

A Study on Lightweight Transformer Based Super Resolution Model Using Knowledge Distillation

Dong-hyun Kim[a], Dong-hun Lee[a], Aro Kim[a], Vani Priyanka Gali[a], and Sang-hyo Park[a]‡

요 약

최근 자연어 처리에서 사용되던 트랜스포머 모델이 이미지 초해상화 분야에서도 적용되면서 좋은 성능을 보여주고 있다. 그러나 이러한 트랜스포머 기반 모델들은 복잡하고 많은 학습 파라미터를 가지고 있어 많은 하드웨어 자원을 요구하기 때문에 작은 모바일 기기에서는 사용하기 어렵다는 단점을 가지고 있다. 따라서 본 논문에서는 트랜스포머 기반 초해상화 모델의 크기를 효과적으로 줄일 수 있는 지식 증류 기법을 제안한다. 실험 결과 트랜스포머 블록의 개수를 줄인 학생 모델에서 제안 기법을 적용해 교사 모델과 비슷한 성능을 내거나 더 높일 수 있음을 확인하였다.

Abstract

Recently, the transformer model used in natural language processing is also applied to the image super resolution field, showing good performance. However, these transformer based models have a disadvantage that they are difficult to use in small mobile devices because they are complex and have many learning parameters and require high hardware resources. Therefore, in this paper, we propose a knowledge distillation technique that can effectively reduce the size of a transformer based super resolution model. As a result of the experiment, it was confirmed that by applying the proposed technique to the student model with reduced number of transformer blocks, performance similar to or higher than that of the teacher model could be obtained.

Keyword : Knowledge Distillation, Super Resolution, Transformer, Deep Learning, Image Processing

a) 경북대학교 IT대학 컴퓨터학부(School of Computer Science and Engineering, Kyungpook National University)
‡ Corresponding Author : 박상효(Sang-hyo Park)
E-mail: s.park@knu.ac.kr
Tel: +82-53-950-6373
ORCID: https://orcid.org/0000-0002-7282-7686

※ 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167169, 이동형 로봇 기반 실사 메타버스 실감형 비디오의 획득 및 처리 기술 개발).
※This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00167169, Development of Moving Robot-based Immersive Video Acquisition and Processing System in Metaverse).
· Manuscript April 7, 2023; Revised May 17, 2023; Accepted May 17, 2023.

Ⅰ. 서 론

딥러닝 모델의 등장으로 합성곱 신경망 기반의 네트워크를 이용해 실제 이미지와 시각적으로 구별할 수 없는 고해상도 이미지를 생성할 수 있게 되면서, 초해상화(Super Resolution)는 최근 몇 년 동안 이미지 처리 분야에서 활발한 연구 분야가 되었다. 한편 트랜스포머는 자연어 처리에서 우수한 성능을 보이고 있는데, 최근에는 이미지 처리 분야에도 적용돼 초해상화 분야에서도 좋은 성능을 보인다. 그러나 트랜스포머 기반 모델은 기존의 합성곱 신경망 기반 초해상화 모델보다 더 많은 학습 파라미터가 필요하다. 이러한 이유로 하드웨어 자원이 제한되고 배터리 수명이 우려되는 IoT 및 모바일 장치에서 사용하기 어렵다. 트랜스포머 기반 초해상화 모델의 경량화를 위해 자체 보정 효율적인 트랜스포머와 물리적으로 제한된 트랜스포머와 같은 다양한 방법[1][2]이 제안되고 있다. 본 논문에서는 트랜스포머 기반의 초해상화 모델에서 학습 파라미터 수를 줄이면서도 학생 모델이 교사 모델을 모방하여 최대한 성능을 유지하도록 하는 효과적인 지식 증류 기법을 제안한다.

Ⅱ. 제안 방법

본 논문에서는 트랜스포머 기반 초해상화 모델로 SwinIR[3]을 사용하였고 그림 1과 같이 학생, 교사 모델을 구성하였다. 교사와 학생 모델은 트랜스포머 블록 개수를 제외하고는 동일한 모델이며 교사 모델은 트랜스포머 블록 4개, 학생 모델은 트랜스포머 블록 1~3개로 구성된다.

교사 모델과 학생 모델 간의 지식 증류를 하는 과정에서 교사 모델의 출력 결과와 트랜스포머 블록의 특징에 대한 지식을 학생 모델에게 전달해주기 위해 학생 모델의 전체 손실(total loss)을 학생 손실인 $L_{student}$와 증류 손실인 L_{KD}의 합으로 계산하여 아래의 수식 1과 같이 나타낸다.

$$L_{total} = \alpha * L_{student} + (1-\alpha) * L_{KD} \quad (1)$$

여기서 α는 모델 학습 도중 전체 손실에서 학생 손실과 증류 손실의 비율을 조절하는 역할로 α가 커질수록 증류 손실보다 학생 손실의 영향이 커진다. 본 논문에서는 α를 1/4, 2/4, 3/4으로 실험해 본 결과 α가 3/4일 때 PSNR 결과가 가장 좋아 해당 값을 α로 결정하였다.

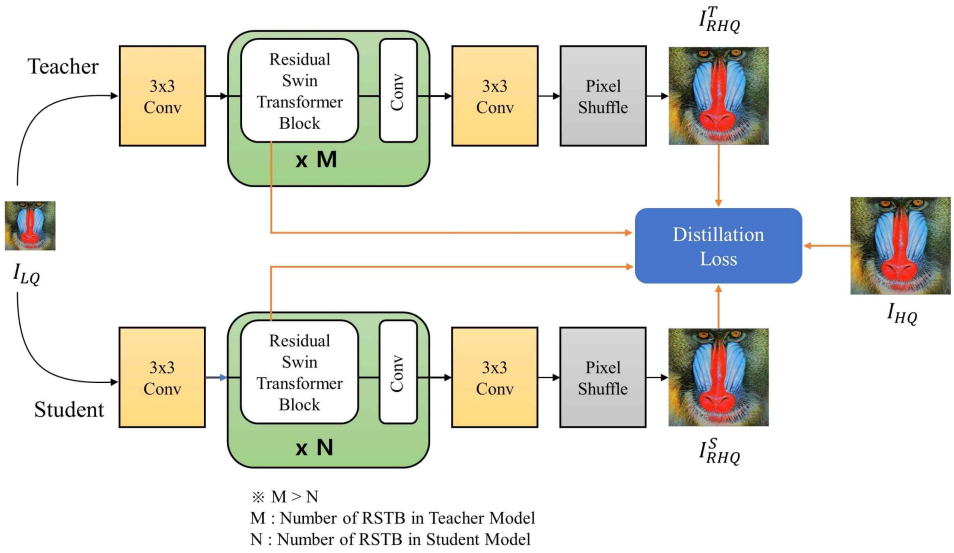

그림 1. 교사, 학생 모델의 구조와 지식 증류 과정
Fig. 1. Structure of Teacher, Student Model and Knowledge Distillation process

$L_{student}$는 학생 모델에서 생성된 이미지 I^S_{RHQ}와 고해상도 원본 이미지 I_{HQ} 사이에서 차이 값의 절댓값 합(L1 loss)으로 초해상화 모델에서 자주 사용되는 손실함수를 사용하여 수식 (2)와 같이 표현되고, 증류 손실 L_{KD}는 아래의 수식 (3)처럼 2개 항의 합으로 정의된다.

$$L_{student} = \| I_{HQ} - I^S_{RHQ} \|_1 \quad (2)$$

$$L_{KD} = \| I^T_{RHQ} - I^S_{RHQ} \|_1 + \| F^T_1 - F^S_1 \|^2 \quad (3)$$

$$F_1 = H_{RSTB_1}(F_0) \quad (4)$$

먼저 교사 모델에서 생성된 이미지 I^T_{RHQ}와 학생 모델에서 생성된 이미지 I^S_{RHQ}사이에서 차이 값의 절댓값 합(L1 loss)으로 계산한 결과를 증류 손실에 포함한다. 교사 모델의 출력값과 학생 모델의 출력값 사이의 차이가 최소화되게 학습하도록 손실함수를 구성해 학생 모델이 교사 모델이 생성한 이미지를 모방할 수 있도록 한다. 그리고 교사 모델의 첫 번째 트랜스포머 블록의 출력 결과인 F^T_1와 학생 모델의 첫 번째 트랜스포머 블록의 출력의 결과인 F^S_1의 평균 제곱 오차(MSE)를 증류 손실에 포함한다. 수식 (4)는 수식 (3)의 교사, 학생 모델에서 첫 번째 트랜스포머 블록의 출력 결과인 F^T_1, F^S_1를 나타내는 수식으로 그림 1에서 첫 번째로 있는 3x3 합성곱 층의 출력값 F_0를 첫 번째 트랜스포머 블록 H_{RSTB_1}에 입력으로 넣어 결괏값 F_1을 얻는다. 교사, 학생 모델을 구성하고 있는 여러 트랜스포머 블록 중에서 첫 번째 트랜스포머 블록 결과만을 이용하여 학생 모델이 교사 모델의 깊은 정보는 모방하지 않고 얕은 정보만 모방하도록 손실함수를 구성한다. 그리고 미리 사전 학습된 교사 모델의 가중치를 고정하고 학생 모델의 손실함수를 수식 (1) 과같이 변경 후 학생 모델을 학습시킨다.

Ⅲ. 모델 실험

1. 실험 환경

실험을 위해 학습데이터로 DIV2K[4], Flickr2K 데이터 세트를 사용하였고 테스트 데이터로는 Set5[5], Set14[6], Manga109[7], Urban[8], BSD100[9]을 사용하였다. 학습 데이터 세트들은 고해상도 이미지인 원본 이미지와 Bicubic 보간법으로 1/2배 다운 샘플링된 저해상도 이미지 쌍으로 구성되어 있다. 실험은 Window 10에서 RTX 3060Ti 1대를 사용해 학습률은 2e-4, 150000 이터레이션으로 진행하였다.

교사 모델은 기존 SwinIR 모델 중 트랜스포머 블록이 4개인 Lightweight_sr 모델의 구조를 사용하였고 학생 모델은 교사 모델에서 트랜스포머 블록을 줄여 트랜스포머 블록이 1, 2, 3개인 구조를 사용하였다. 그 외 임베딩 차원이나 Head의 개수는 동일하다. 교사 모델의 파라미터 개수는 910K개, 학생의 파라미터 개수는 각각 258K, 475K, 692K개로 교사 모델의 28%, 52%, 76% 크기로 경량화하였다.

2. 실험 결과

모델 평가 방법으로는 PSNR과 SSIM 평가지표를 사용

표 1. 지식 증류 기법을 적용한 다양한 학생 모델과 교사 모델의 PSNR, SSIM 결과
Table. 1. PSNR and SSIM results of various student and teacher models using knowledge distillation

	Teacher(910K)		Student1(258K)		Student2(475K)		Student3(692K)	
	PSNR	SSIM	PSNR(↓)	SSIM(↓)	PSNR(↓)	SSIM(↓)	PSNR(↑)	SSIM(↑)
Set5	35.7226	0.9430	34.8933	0.9384	35.6556	0.9428	35.7301	0.9430
Set14	31.3508	0.8914	30.6187	0.8851	31.3103	0.8921	31.3909	0.8925
Manga109	36.5240	0.9694	35.0840	0.9633	36.4240	0.9695	36.5537	0.9698
Urban	29.8622	0.9231	28.0036	0.8992	29.6572	0.9205	29.8678	0.9230
BSD100	30.7345	0.8887	30.1377	0.8813	30.6932	0.8883	30.7436	0.8888
Average	32.8388	0.9231	31.7474	0.9134	32.7480	0.9226	32.8572	0.9234

하였다. 표 1은 교사 모델과 지식 증류 기법을 적용한 학생 모델의 RGB 채널 PSNR, SSIM 결괏값과 교사 모델과의 성능차이를 화살표르 나타내고 있다. Student1 모델의 평균 PSNR과 SSIM은 Teacher에 비해 1.0914dB, 0.0097만큼 감소하였고 Student2 모델은 0.0908dB, 0.0005만큼 감소함을 확인할 수 있다. Student1 모델은 파라미터 개수를 교사 모델의 28%로 줄여서 지식 증류를 기법을 적용했음에도 성능이 많이 줄어들었음을 알 수 있고, Student2 모델은 파라미터 수를 52%로 줄였지만 Teacher 모델 보다 성능이 비슷하거나 약간 줄어들었음을 확인할 수 있다. 여기서 주목할 점은 Student3 모델의 평균 PSNR과 SSIM이 Teacher 모델보다 0.0184dB, 0.0003만큼 증가하였다는 것이다. student 3 모델은 파라미터수가 Teacher 모델의76%로 앞선 학생 모델들에 비해 조금 감소 되었고, 지식 증류 기법을 통해 Teacher 모델의 얕은 정보와 출력 값을 학습에 이용하면서 Teacher 모델보다 더 좋은 성능을 낼 수 있었다.

Ⅳ. 결 론

본 논문에서는 트랜스포머 기반 초해상화 모델에서의 지식 증류 기법을 제시하고 다양한 트랜스포머 블록 개수를 가진 학생 모델에 해당 기법을 적용해보면서 모델에 대한 성능을 평가하였다. 학생모델의 파라미터 수를 교사 모델 파라미터수의 28%, 52%로 줄였을 때는 평균 PSNR, SSIM 값이 약간 감소하였고, 76%로 줄였을 때는 오히려 평균 PSNR, SSIM 값이 향상되었다. 이를 통해 트랜스포머 내에서 단순히 블록 수를 줄이는 것만으로도 교사 모델에 비해 유사한 성능을 내거나 더 높일 수 있는 모델을 만들 수 있다는 것을 확인하였다. 후속 연구로 SwinIR 모델 외에 다양한 트랜스포머 기반 초해상화 모델에도 지식 증류 기법을 적용해보면서 보다 발전된 기법을 찾고자 한다.

참 고 문 헌 (References)

[1] Zou, W., Ye, T., Zheng, W., Zhang, Y., Chen, L., & Wu, Y., "Self-calibrated efficient transformer for lightweight super-resolution." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 930-939 2022.
doi: https://doi.org/10.1109/CVPRW56347.2022.00107

[2] Wang, X., Zhu, S., Guo, Y., Han, P., Wang, Y., Wei, Z., & Jin, X., "TransFlowNet: A physics-constrained Transformer framework for spatio-temporal super-resolution of flow simulations." *Journal of Computational Science*, 65, 101906, 2022.
doi: https://doi.org/10.1016/j.jocs.2022.101906

[3] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. "Swinir: Image restoration using swin transformer." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833-1844. 2021.
doi: https://doi.org/10.1109/ICCVW54120.2021.00210

[4] E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, pp. 1122-1131, 2017.
doi: https://doi.org/10.1109/CVPRW.2017.150

[5] Bevilacqua, M., Roumy, A., Guillemot, C., & Alberi-Morel, M. L., "Low-complexity single-image super-resolution based on nonnegative neighbor embedding." In *Proceedings of the 23rd British Machine Vision Conference (BMVC). BMVA Press*, 135.1-135.10. 2012.
doi: https://doi.org/10.5244/C.26.135

[6] Zeyde, R., Elad, M., & Protter, M., "On single image scale-up using sparse-representations." In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pp. 711-730. Springer Berlin Heidelberg, 2012.
doi: https://doi.org/10.1007/978-3-642-27413-8_47

[7] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., & Aizawa, K., "Sketch-based manga retrieval using manga109 dataset." *Multimedia Tools and Applications*, 76, 21811-21838.2017
doi: https://doi.org/10.1007/s11042-016-4020-z

[8] Huang, J. B., Singh, A., & Ahuja, N., "Single image super-resolution from transformed self-exemplars." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 5197-5206. 2015.
doi: https://doi.org/10.1109/cvpr.2015.7299156

[9] Martin, D., Fowlkes, C., Tal, D., & Malik, J., "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In *Proceedings Eighth IEEE International Conference on Computer Vision*. ICCV 2001, vol. 2, pp. 416-423. IEEE, 2001.
doi: https://doi.org/10.1109/iccv.2001.937655