KICSP 한국 융합신호처리학회

융합신호처리학회논문지 Vol. 24, No.1 : 21~26 March. 2023
DOI : 10.23087/jkicsp.2023.24.1.003

초해상화 모델 경량화를 위한 지식 증류 방법의 비교 연구

이여진[1], 박한훈[1*]
부경대학교 전자정보통신공학부

A Comparative Study of Knowledge Distillation Methods in Lightening a Super-Resolution Model

Yeojin Lee[1], Hanhoon Park[1*]
Division of Electronics and Communications Engineering, Pukyong National University

요 약 지식 증류는 깊은 모델의 지식을 가벼운 모델로 전달하는 모델 경량화 기술이다. 대부분의 지식 증류 방법들은 분류 모델을 위해 개발되었으며, 초해상화를 위한 지식 증류 연구는 거의 없었다. 본 논문에서는 다양한 지식 증류 방법들을 초해상화 모델에 적용하고 성능을 비교한다. 구체적으로, 초해상화 모델에 각 지식 증류 방법을 적용하기 위해 손실 함수를 수정하고, 각 지식 증류 방법을 사용하여 교사 모델을 약 27배 경량화한 학생 모델을 학습하여 2배 초해상화하는 실험을 진행하였다. 실험을 통해, 일부 지식 증류 방법은 초해상화 모델에 적용할 경우 유효하지 않음을 알 수 있었으며, 관계 기반 지식 증류 방법과 전통적인 지식 증류 방법을 결합했을 때 성능이 가장 높은 것을 확인하였다.

• 주제어 : 초해상화, 딥러닝, 모델 경량화, 지식 증류, 비교 연구

Abstract Knowledge distillation (KD) is a model lightening technology that transfers the knowledge of deep models to light models. Most KD methods have been developed for classification models, and there have been few KD studies in the field of super-resolution (SR). In this paper, various KD methods are applied to an SR model and their performance is compared. Specifically, we modified the loss function to apply each KD method to the SR model and conducted an experiment to learn a student model that was about 27 times lighter than the teacher model and to double the image resolution. Through the experiment, it was confirmed that some KD methods were not valid when applied to SR models, and that the performance was the highest when the relational KD and the traditional KD methods were combined.

• Key Words : Super-resolution, Deep learning, Model lightening, Knowledge distillation, Comparative study

Received 28 November 2022, Revised 26 December 2022, Accepted 10 January 2023

*** Corresponding Author** Hanhoon Park, Division of Electronics and Communications Engineering, Pukyong National University, 45, Yongso-ro, Nam-gu, Busan, Korea. E-mail: hanhoon.park@pknu.ac.kr

Ⅰ. 서론

초해상화(Super-resolution)란 저해상도(Low resolution) 영상을 고해상도(high resolution) 영상으로 변환하는 기술로서, 의료, 보안, 원격 탐사 등 다양한 분야에서 활용되고 있다. 최근에는 CNN(Convolutional Neural Network) 기반 딥러닝 기술의 발전과 함께 CNN을 이용한 딥러닝 기반 초해상화 방법에 대한 연구가 활발히 진행되고 있으며, 다양한 모델들이 제안되고 있다.

딥러닝 기반 초해상화 방법은 많은 메모리와 연산량을 요구하기 때문에 제약된 리소스를 가진 모바일, 사물인터넷 기기에 적용하기는 어렵다. 이런 한계를 극복하기 위해, 모델의 성능은 유지하며 크기를 줄이는 모델 경량화 연구가 활발히 진행되고 있다.

지식 증류(KD: Knowledge Distillation)는 지식을 전달하는 모델 경량화 기술로서, 기학습된 크고 깊은 모델의 지식을 작고 얕은 모델로 전달하는 방법을 말하며[1], 전달되는 지식의 종류나 전달 방법에 따라 다양한 지식 증류 방법들이 제안되고 있다. 그러나, 대부분의 지식 증류 방법들은 분류 모델을 경량화하기 위해 개발되어 왔으며, 초해상화 모델 경량화를 위한 지식 증류 연구는 거의 없다. 본 논문에서는 기존 분류 모델 경량화를 위해 개발된 다양한 지식 증류 방법들을 초해상화 모델에 적용하여 그 성능을 비교한다. 실험을 통해, 경량화된 초해상화 모델의 성능을 정량적으로 분석하여, 초해상화 모델 경량화에 가장 적합한 지식 증류 방법을 모색한다.

Ⅱ. 관련연구

2.1 초해상화

초해상화는 영상의 해상도를 높이기 위한 기술로, 기존 초해상화 방법들은 비교적 단순한 선형 매핑을 기반으로 하기 때문에 복잡하고 비선형적인 초해상화 모델을 구현하기 어려웠다. 이러한 문제를 해결하기 위해 최근 딥러닝 기술을 활용한 초해상화 방법들이 활발히 연구되고 있으며 다양한 구조나 형태의 CNN 기반 모델들이 제안되고 있다. SRCNN은 딥러닝 기술을 활용한 선구적인 초해상화 모델로서, 딥러닝 기술을 활용함으로써 고화질의 초해상화 영상 생성이 가능함을 보여주었다[2]. 이후 다양한 딥러닝 기반 초해상화 방법 및 모델이 제안되었으며, 그 중 EDSR (Enhanced Deep Super-Resolution)은 효율적인 네트워크 구성과 층 사이의 잔차(residual) 학습을 통해 성능을 크게 향상시켰다[3].

2.2 지식 증류

지식 증류는 기학습된 깊은 모델(교사 모델)의 지식을 전달하여 가벼운 모델(학생 모델)의 학습 능력을 향상시키는 기술을 말한다. 처음 소개된 지식 증류 방법은 Hinton 등에 의해 제안되었으며, 교사 모델의 소프트맥스(softmax) 층에서 출력된 확률분포 정보를 학생 모델로 전달하였다[1]. 이후, FitNets는 교사 모델의 마지막 층뿐만 아니라 중간층의 지식을 함께 전달되도록 하였으며[4], 학생 모델이 교사 모델의 주의집중 지도(attention map)를 모방하도록 하거나, 교사 모델 없이 학생 모델들이 서로 지식을 전달하면서 학습하거나, 다른 모델 없이 자가 증류(self-distillation)를 통해 학습하는 등 다양한 지식 증류 방법들이 제안되었다[5, 6, 7]. 그러나, 대부분의 지식 증류 방법들은 분류 모델을 경량화하기 위해 제안되었으며, 초해상화 분야에서 지식 증류를 사용한 모델 경량화 연구는 거의 없다.

지식 증류 방법은 크게 전달되는 지식의 종류와 증류 방법 등에 따라 분류할 수 있다. 지식의 종류에 따른 분류는 교사 모델의 마지막 층의 출력인 응답 기반 지식(response-based knowledge)을 전달하는 방법과 교사 모델의 중간층의 특징 지도와 같은 특징 기반 지식(feature-based knowledge)을 전달하는 방법, 특징 지도나 데이터 샘플, 출력 등의 상관관계인 관계 기반 지식(relation-based knowledge)을 전달하는 방법으로 나눌 수 있다. 그림 1은 지식 증류 방법에서 전달되는 지식의 종류를 보여준다. Hinton 등에 의해 제안된 방법이 응답 기반 지식을 전달하는 대표적인 예이다[1]. FitNets와 Fakd 등이 특징 기반 지식을 전달하는 방법이며[4, 5, 8], 대조 학습(contrastive learning)은 관계 기반 지식을 전달하는 방법에 속한다[9]. 증류 방법에 따른 분류는 오프라인 학습과 온라인 학습, 자가 학습(self-learning)으로 나뉜다. 오프라인 학습은 기학습된 교사 모델의 지식을 학생 모델 학습 시 전달하는 방법으로, 대부분의 초기 지식 증류 방법들이 여기에 속한다. 온라인 학습은 교사와 학생 모델이 함께 학습하며

지식을 전달하는 방식으로 end-to-end 구조이다. 온라인 학습의 예로 상호 증류(mutual distillation) 방법이 있다[6]. 자가 학습은 교사 모델 없이 학생 모델을 여러 개의 중간 단으로 구성하여 스스로 지식을 전달하는 방식으로, 자가 증류 방법이 대표적이다[7]. 그림 2는 오프라인, 온라인, 자가 학습 방법에서 지식의 전달 과정을 보여준다.

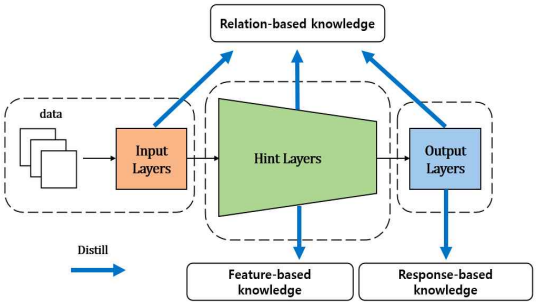


Fig. 1. Type of knowledge transferred in knowledge distillation methods.

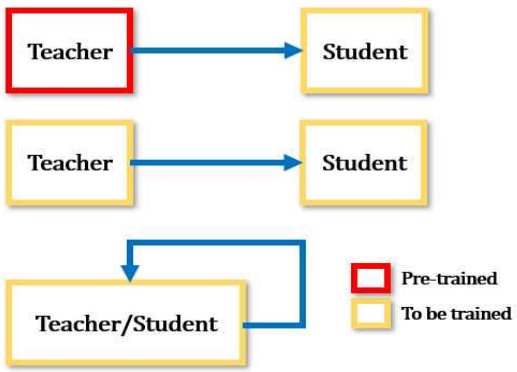


Fig. 2. Process of transferring knowledge in online, offline, and self-learning.

Ⅲ. 비교 연구 대상

본 논문에서 비교를 위해 사용된 지식 증류 방법은 다음과 같다.

3.1 전통적인 지식 증류(HKD)

전통적인 지식 증류(HKD)는 깊은 교사 모델의 소프트맥스(softmax) 층에서 출력된 확률분포 정보를 완화하여 얕은 학생 모델로 전달하는 방식으로, 학생 모델은 교사 모델 출력과의 차이를 최소화하도록 학습된다[1]. 학생 모델과 교사 모델의 차이 L_{TS}는 쿨백-라이블러 발산(Kullback-Leibler divergence)을 사용하여 다음과 같이 계산된다.

$$L_{TS} = \sum_{x_i \in X} l(f_T(x_i), f_S(x_i)) \quad (1)$$

여기서, f_T와 f_S은 교사와 학생 모델의 소트프맥스 출력을 뜻하고 l은 쿨백-라이블러 발산을 나타낸다.

3.2 대조 학습 기반 지식 증류(CKD)

대조 학습은 특징에 대한 일반화를 최대화하고, 적은 데이터로 학습 효율을 향상하는 방법으로, 긍정적(positive) 샘플과 부정적(negative) 샘플을 구분하여 각 샘플에 대한 유사성(similarity)과 비유사성(dissimilarity)을 학습한다. 즉, 긍정적 샘플과의 유사성을 최대화하는 것과 동시에 부정적 샘플과의 유사성은 최소화하게 된다.

대조 학습 기반 지식 증류(CKD)는 대조 학습을 사용하여 지식 증류를 시도한 방법으로, 초해상화 모델에 적용하여 효과적으로 경량화된 학생 모델을 생성할 수 있었다[9]. CKD에서 사용된 손실 함수 L_{CS}은 학습 배치 내 i 번째 저해상도 영상에 대해, 교사 모델이 생성한 초해상화 영상을 긍정적 샘플로 간주하고, i 번째 저해상도 영상을 제외한 나머지 영상으로부터 학생 모델이 생성한 초해상화 영상들을 부정적 샘플로 간주하여 다음과 같이 계산된다.

$$L_{CS} = \sum_{i=1}^{N} \frac{d(SR_i^S, SR_i^T)}{\sum_{k=1,k\neq i}^{N} d(SR_i^S, SR_k^S)} \quad (2)$$

여기서, N은 배치 크기를 의미하며, SR^T, SR^S는 각각 교사 모델과 학생 모델의 초해상화 영상을 나타낸다. d는 평균 제곱 오차(mean squared error)를 나타낸다.

3.3 관계 학습 기반 지식 증류(RKD)

관계 기반 지식 증류(RKD)는 교사 모델과 학생 모델의 출력 사이의 유기적인 관계 구조를 지식으로 전달, 학습하는 방법으로, 관계 구조는 출력 샘플 사이의 거리나 각도를 계산하여 얻어진다[10]. 예를 들어, 교사 모델과 학생 모델에 대해, 각각 출력 샘플 사이의 유클리디안 거리(Euclidean distance)를 계산하고, 이를 완화된 평균 제곱 오차(Huber loss)를 통해 최소화하여 샘플 사이의 유기적인 관계 지도를 학습한다. RKD에서 사용된 손실 함수 L_{RS}는 다음과 같다.

$$L_{RS} = \sum_{(x_i,x_j)\in X^2} l_\delta(\psi_D(t_i,t_j),\psi_D(s_i,s_j)) \quad (3)$$

여기서, ψ_D는 유클리디안 거리를 의미하고 l_δ은 완화된 평균 제곱 오차를 의미한다. t와 s는 교사 모델과 학생 모델의 출력($s=f_S(x)$, $t=f_T(x)$)을 의미한다.

분류 모델에 적용한 기존 연구에서 RKD만을 사용할 경우 성능이 좋지 못했으며, RKD와 HKD를 함께 사용할 경우 기존 지식 증류 방법보다 우수한 성능을 보였다[10]. 이를 참고로 하여, 본 논문에서도 RKD와 HKD를 결합한 방법을 추가적으로 비교한다.

3.4 재학습 기반 지식 증류(RVKD)

재학습(review) 기반 지식 증류(RVKD)는 특징 기반 지식을 전달하는 방법으로, 같은 층 사이의 지식을 전달하는 기존 지식 증류 방법들과 달리 이전 층들의 특징 지도를 ABF(Attention Based Fusion) 알고리즘을 통해 융합하여 손실을 계산하고 최소화하도록 학습한다[11]. 이를 통해, 각 층은 이전 층들의 정보를 함께 학습함으로써, 더욱 풍부한 지식을 습득할 수 있다. RVKD의 학습 손실 L_{RVS}은 연산량을 줄이기 위해 논문에서 제안된 HCL(Hierarchical Context Loss)을 사용하여 계산한다.

Ⅳ. 실험 방법 및 환경

본 논문에서는 Ⅲ 장에서 설명한 지식 증류 방법(HKD, CKD, RKD, RKD+HKD, RVKD)을 초해상화 모델인 EDSR에 적용하여 성능을 비교하는 실험을 진행하였다[3].

초해상화 모델에 지식 증류를 적용하기 위해, 분류 모델과 다르게 손실을 계산하기 위한 입력 데이터와 함수를 바꿔야 할 필요가 있다. 우선, 손실 함수 L_{DS}는 ground-truth인 고해상도 영상과 학생이 생성한 초해상화 영상 사이의 평균 제곱 오차를 나타낸다. 다음으로, 식 (1)에서 l은 평균 제곱 오차로 대체하였다. 각 지식 증류 방법에서 손실 함수의 비율은 기존 논문을 따라 조정하였으며 각각의 식은 다음과 같다.

$$L_{HKD} = L_{DS} + L_{TS} \quad (4)$$

$$L_{CKD} = L_{DS} + 10L_{CS} \quad (5)$$

$$L_{RKD} = L_{DS} + 10^3 L_{RS} \quad (6)$$

$$L_{RKD+HKD} = L_{DS} + 10^3 L_{RS} + L_{TS} \quad (7)$$

$$L_{RVKD} = L_{DS} + 10^2 L_{RVS} \quad (8)$$

Fig. 3. Datasets used in our experiments.

Table 1. Network parameters of teacher and student models used in our experiments.

	#filters	#resblocks	#params
Teacher	256	32	40.7M
Student	64	16	1.5M

Table 2. Mean PSNR results of EDSR with various knowledge distillation methods when scale factor = 2.

	Teacher	Student					
		w/o KD	HKD	CKD	RKD	RVKD	RKD+HKD
Set5	38.193	37.832	**37.852**	**37.879**	**37.868**	**37.832**	**37.846**
Set14	33.948	33.402	33.355	**33.408**	33.368	33.346	**33.419**
BSD100	32.32	32.028	**32.06**	**32.049**	32.019	32.02	**32.061**
Urban100	32.967	31.594	**31.671**	**31.604**	31.486	31.569	**31.671**

비교 실험을 위한 학습 데이터 셋은 DIV2K, 검증 데이터 셋은 Set5, Set14, BSD100, Urban100을 사용하였다(그림 3 참조)[12-16]. 학습 및 검증을 위한 저해상도 영상은 각 데이터 셋에 포함된 고해상도 영상을 바이큐빅 보간법(bicubic interpolation)을 사용하여 다운샘플링(down-sampling) 하여 생성하였다. 표 1은 실험에서 사용된 초해상화 모델인 EDSR의 교사 모델과 학생 모델 구성을 나타낸 것이다. 교사 모델은 32개의 resblock과 256개의 filter로 구성되고 학생 모델은 16개의 resblock과 64개의 filter로 구성되어, 학생 모델의 학습 파라미터의 수는 대략 27배 줄어들었다. 모델 구현은 Pytorch(버전 1.11.0)를 사용하였으며, 모델 학습은 i7-8700 3.20GHz CPU와 16GB RAM, NVIDIA RTX 3060 GPU를 가진 PC에서 진행되었다. Adam optimizer (β_1 = 0.9, β_2 = 0.999)를 사용하였으며, epoch 수는 100, 학습률(learning rate)은 10-4, 배치 크기(batch size)는 8로 설정되었다. 2배 초해상화하는 실험을 진행했으며, 초해상화 영상 결과의 화질 평가를 위해 정량적 화질 측정 지표인 PSNR(Peak Signal-to-Noise Ratio)를 측정하였다.

Ⅴ. 실험 결과 및 분석

표 2는 각 지식 증류 방법에 따른 EDSR 모델의 2배 초해상화 결과 영상의 평균 PSNR 값을 보여준다. 빨간색으로 표시한 값은 각 데이터 셋에 대해 가장 높은 PSNR 값을 뜻하며, 파란색은 다음으로 높은 값을 나타낸다. 볼드체는 증류하지 않은 학생 모델(w/o KD)보다 높은 값을 표시한 것이다. 즉, 볼드체로 표시되지 않은 것은 지식 증류를 수행함으로써 오히려 성능이 떨어진 경우를 의미한다. 실험 결과, 단독으로 사용할 경우 CKD의 성능이 가장 우수했으며, HKD가 그 다음이었다. HKD의 경우 가장 전통적이고 간단한 방법임에도 불구하고 다양한 형태의 많은 지식을 전달하는 방법들에 비해 높은 성능을 보였다. RKD나 RVKD는 Set5 데이터 셋을 제외하곤 지식 증류로 인해 학생 모델의 성능이 떨어졌다. 즉, RKD나 RVKD는 초해상화 모델을 경량화하는 데는 유효하지 않았다. RVKD는 분류 모델을 경량화하는 실험 결과에서는 매우 우수한 성능을 가지는 것으로 보고되었으나 초해상화 모델에 대해서는 그렇지 않았다. 그러나, RKD의 경우 분류 모델에 적용한 기존 연구 결과에서와 마찬가지로 HKD를 함께 사용하였을 때 더 높은 성능을 보였으며, CKD를 포함한 다른 지식 증류 방법보다 높은 성능을 보였다. 결과적으로, 본 논문의 비교 실험에서 RKD+KD가 교사 모델(Teacher)에 근접하는 가장 높은 PSNR 값을 가졌으며, 다음으로 CKD가 높았다. 또한, 두 방법만이 모든 데이터 셋에 대해 증류하지 않은 학생 모델(w/o KD)보다 높은 PSNR 결과를 가졌다.

Ⅵ. 결론

본 논문에서는 초해상화 모델인 EDSR에 다양한 지식 증류 방법을 적용하여 성능을 비교, 분석하는 실험을 진행하였다. 이를 위해, 각 방법의 손실 함수를 초해상화 모델에 적용하기 위해 변경하였으며, 교사 모델에 비해 약 27배 경량화된 학생 모델을 학습하여 2배 초해상화하는 실험을 진행하였다. 실험 결과, RKD와 HKD를 결합한 방법이 가장 높은 성능을 보이는 것을 확인하였다.

추후, 상호 증류나 자가 증류 등의 방법과도 비교, 분석하는 실험을 진행할 예정이다.

ACKNOWLEDGMENTS

본 연구는 산업통상자원부와 한국산업기술진흥원의 “지역혁신클러스터육성사업(R&D, P0004797)”으로 수행된 연구결과 입니다.

REFERENCES

[1] G. Hinton, et al., “Distilling the knowledge in a neural network,” Proc. of NIPS, 2014.

[2] C. Ledig, et al., “Photo-realistic single image super-resolution using a generative adversarial network,” Proc. of CVPR, pp. 105-114, 2017.

[3] B. Lim, et al., “Enhanced deep residual networks for single image super-resolution,” Proc. of CVPRW, pp. 136-144, 2017.

[4] A. Romero, et al., “FitNets: hints for thin deep nets,” Proc. of ICLR, 2015.

[5] N. Komodakis and S. Zagoruyko, “Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer,” Proc. of ICLR, 2017.

[6] Y. Zhang, et al., “Deep mutual learning,” Proc. of CVPR, pp. 4320-4328, 2018.

[7] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: improve the performance of convolutional neural networks via self distillation,” Proc. of ICCV, pp. 3713-3722, 2019.

[8] Z. He, et al., “Fakd: feature-affinity based knowledge distillation for efficient image super-resolution,” Proc. of ICIP, pp. 518-522, 2020.

[9] H. Moon, et al., “Compression of super-resolution model using contrastive learning,” Proc. of KIBME Summer Conference, pp. 557-559, 2022.

[10] W. Park, et al., “Relational knowledge distillation,” Proc. of CVPR, pp. 3967-3976, 2019.

[11] P. Chen, et al., “Distilling knowledge via knowledge review,” Proc. of CVPR, 2021.

[12] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: dataset and study,” Proc. of CVPRW, pp. 126-135, 2017.

[13] M. Bevilacqua, et al., “Low complexity single image super-resolution based on nonnegative neighbor embedding,” Proc. of BMVC, 2012.

[14] R. Zeyde, et al., “On single image scale-up using sparse-representations,” Proc. of Int. Conf. on Curves and Surfaces, pp. 711-730, 2010.

[15] D. Martin, et al., “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” Proc. of ICCV, vol. 2, pp. 416-423, 2001.

[16] J.-B. Huang, et al., “Single image super-resolution from transformed self-exemplars,” Proc. of CVPR, pp. 5197-5206, 2015.

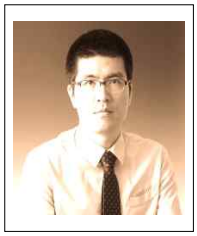
저자 소개

이 여 진 (Yeojin Lee)

2021년 3월~현재 : 부경대학교 전자공학과 학부생

관심분야 : 지식 증류, SLAM

박 한 훈 (Hanhoon Park)



2000년 2월 : 한양대학교 전자통신전파공학과(공학사)

2002년 2월 : 한양대학교 전자통신전파공학과(공학석사)

2007년 8월 : 한양대학교 전자통신전파공학과(공학박사)

2012년 3월~현재 : 부경대학교 전자공학과 교수

관심분야 : 증강현실, 인간컴퓨터상호작용, 컴퓨터비전/그래픽스, 딥러닝 응용