Journal of The Institute of Electronics and Information Engineers Vol.54, NO.5, May 2017 https://doi.org/10.5573/ieie.2017.54.5.35
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2017-54-5-5

딥 residual network를 이용한 선생-학생 프레임워크에서 힌트-KD 학습 성능 분석

(Performance Analysis of Hint-KD Training Approach for the Teacher-Student Framework Using Deep Residual Networks)

배 지 훈*, 임 준 호**, 유 재 학*, 김 귀 훈*, 김 준 모**

(Ji-Hoon Bae, Junho Yim, Jaehak Yu, Kwihoon Kim, and Junmo Kim©)

요 약

본 논문에서는 지식추출(knowledge distillation) 및 지식전달(knowledge transfer)을 위하여 최근에 소개된 선생-학생 프레임워크 기반의 힌트(Hint)-knowledge distillation(KD) 학습기법에 대한 성능을 분석한다. 본 논문에서 고려하는 선생-학생 프레임워크는 현재 최신 딥러닝 모델로 각광받고 있는 딥 residual 네트워크를 이용한다. 따라서, 전 세계적으로 널리 사용되고 있는 오픈 딥러닝 프레임워크인 Caffe를 이용하여 학생모델의 인식 정확도 관점에서 힌트-KD 학습 시 선생모델의 완화상수 기반의 KD 정보 비중에 대한 영향을 살펴본다. 본 논문의 연구결과에 따르면 KD 정보 비중을 단조감소하는 경우보다 초기에 설정된 고정된 값으로 유지하는 것이 학생모델의 인식 정확도가 더 향상된다는 것을 알 수 있었다.

Abstract

In this paper, we analyze the performance of the recently introduced Hint-knowledge distillation (KD) training approach based on the teacher-student framework for knowledge distillation and knowledge transfer. As a deep neural network (DNN) considered in this paper, the deep residual network (ResNet), which is currently regarded as the latest DNN, is used for the teacher-student framework. Therefore, when implementing the Hint-KD training, we investigate the impact on the weight of KD information based on the soften factor in terms of classification accuracy using the widely used open deep learning frameworks, Caffe. As a results, it can be seen that the recognition accuracy of the student model is improved when the fixed value of the KD information is maintained rather than the gradual decrease of the KD information during training.

Keywords : Knowledge distillation, Hint training, Deep residual networks, Caffe

Ⅰ. 서 론

지난 몇 년 동안 컴퓨팅 처리 성능의 비약적인 발전과 더불어 빅데이터 시대가 도래함에 따라 기계학습을 이용한 지능적 시스템 개발 및 적용이 전 세계적으로 가속화되고 있고, 인간의 뇌를 모방한 딥러닝 기술이 크게 발전하고 있다. 특히, 딥러닝 기술은 영상, 비디오 등의 컴퓨터 비전 분야에서 탁월한 성능을 보여주고 있으며 현재 다양한 분야로의 적용범위가 확대되어 가고 있는 추세이다[1~3]. 하지만, 고사양의 딥러닝 모델은 일반적으로 수천만개 이상의 수많은 학습변수들로 이루어진 복잡한 네트워크 구조를 가지기 때문에, 이미 학습 완료된 복잡한 네트워크 구조로부터 다양한 응용분야로의 적용을 활성화하기 위해서는 효율적인 정보추출(knowledge distillation) 및 정보이전(knowledge transfer) 기술이 요구되고 있다.

* 정회원, 한국전자통신연구원 KSB융합연구단
(KSB Convergence Research Department, Electronics and Telecommunications Research Institute)
** 정회원, 한국과학기술원 전기 및 전자공학부
(Department of Electrical Engineering, Korea Advanced Institute of Science and Technology)
© Corresponding Author(E-mail : junmo.kim@kaist.ac.kr)
※ 이 논문은 2015년 정부(미래창조과학부)의 재원으로 국가과학기술연구회 융합연구단 사업(No. CRC-15-05-ETRI)의 지원을 받아 수행된 연구임
Received ; December 16, 2016 Revised ; March 27, 2017
Accepted ; April 5, 2017

최근, 이미 학습 완료된 복잡한 네트워크로부터 유용한 정보를 추출하고 상대적으로 저사양의 딥러닝 모델로 그 정보를 이전하는 연구가 진행되어 왔다[4~5]. 이미 학습 완료된 복잡한 네트워크를 선생모델(teacher network)로 상기 선생모델로부터 정보를 이전받아 학습을 수행하는 저사양 네트워크를 학생모델(student network)로 정의하고 상기 선생모델의 확률기반의 소프트맥스 출력층(softmax output layer)에서 출력된 값을 완화상수(soften factor)로 완화하여 이를 이용한 비용함수를 기존의 비용함수에 포함하여 학생모델을 학습시키는 방법, 즉 knowledge distillation(KD) 학습방법이 처음 소개되었다[4]. [5]에서는 선생모델의 힌트층(hint layer)을 이용한 가중치(weight) 최적화를 기존 KD 학습에 포함하여 힌트학습과 KD 학습을 차례대로 수행하여 기존 KD 학습 성능을 향상시키는 힌트-KD 학습기법을 제안하였다. 기존 연구결과에 따르면 저사양의 학생모델이 전통적인 역전파(backpropagation) 학습기법으로 학습한 결과보다 선생모델로부터 상기 힌트정보 및 KD 정보들을 이전받아 학습한 결과가 더 우수함을 보여주었다.

본 연구에서는 최신 딥러닝 모델로 각광받고 있는 residual network(ResNet)[6~7]을 적용한 선생-학생 프레임워크에서의 힌트학습과 KD 학습 성능들을 분석한다. 기존 힌트-KD 학습은 많은 특징맵(feature map) 개수를 가지는 폭은 넓고 층수는 작은 maxout[8] 선생모델과 폭은 좁고 오히려 층수는 더 많은 VGG[9] 구조와 유사한 학생모델을 이용한 반면, 본 연구에서 고려하는 선생모델과 학생모델은 모두 최신 딥러닝 모델인 ResNet으로 폭은 서로 같은 크기를 가지고 단지 선생모델이 학생모델 보다 층이 더 깊은 네트워크 구조를 고려하였다. 학습성능을 검증하기 위하여 현재까지 전 세계적으로 널리 활용되고 있는 오픈 딥러닝 프레임워크인 Caffe[10]를 이용하였다. 특히, 힌트-KD 학습 시 선생모델의 완화상수 기반의 비용함수에서 KD 정보 비중에 대한 영향을 살펴본다.

Ⅱ. 본 론

1. 힌트 학습 알고리즘

여기에 본 절에서는 ResNet에 적용하기 위한 힌트 학습 알고리즘[5]에 대하여 먼저 간략히 기술한다. 먼저, 이미 학습 완료된 선생모델의 중간층인 힌트층에서의 출력값과 학습하고자 하는 학생모델의 중간층인 가이드층(guided layer)에서의 출력값을 이용하여 다음 식 (1)과 같이 L2 비용을 계산한다[5].

$$C_{HT}(\widehat{W}_g, \widehat{W}_r) = \frac{1}{2} \| f_h(X; W_h) - f_r(f_g(X; W_g); W_r) \|^2 \quad (1)$$

여기서, f_h 와 f_g 및 f_r은 입력층으로부터 힌트층까지의 가중치에 해당하는 W_h로부터 활성함수(activation function)들을 통과한 출력을, 입력층으로부터 가이드층까지의 가중치에 해당하는 W_g로부터 활성함수들을 통과한 출력을, W_r 가중치를 가지는 회귀함수(regressor function)의 출력을 각각 나타낸다. 이때, 회귀함수는 상기 힌트층의 특징맵의 크기 및 개수와 가이드 층의 특징맵의 크기 및 개수가 서로 다를 경우 이들을 일치하도록 중간에 삽입하는 회귀연산자를 의미한다. 본 연구에서는 회귀함수 연산에 따른 계산의 복잡성을 줄이기 위하여 층수는 서로 다른 반면 특징맵 크기는 서로 같은 ResNet 구조를 고려한다. 따라서, 식 (1)에서 회귀함수를 제외하여 다음 식과 같이 힌트학습을 위한 L2 비용을 계산한다.

$$C_{HT}(\widehat{W}_g) = \frac{1}{2} \| f_h(X; W_h) - f_g(X; W_g) \|^2 \quad (2)$$

식 (2)로부터 추출된 학생모델에서 가이트층까지의 $\widehat{W}_g$를 활용하여 다음 Ⅱ.2절의 힌트-KD 학습을 수행한다.

2. 힌트-KD 학습 알고리즘

본 절에서는 Ⅱ.1절에서 기술한 식 (2)의 힌트학습으로부터 추출된 가중치 정보와 선생모델의 완화상수 기반의 KD 정보 모두를 이용하는 힌트-KD 학습 방법에 대하여 기술한다. 먼저, 힌트-KD 학습 전 학생모델의 전체 가중치 초기화를 다음 식 (3)과 같이 수행한다.

$$W_S = [\widehat{W}_g; W_s^1] \quad (3)$$

여기서, W_s^1는 학생모델에서 가이드층으로부터 최종 출력층까지의 랜덤값을 가지는 가중치를 의미한다. 즉, 학생모델에서 입력층으로부터 가이드층까지의 가중치는 식 (2)의 이전 힌트학습에서 구한 $\widehat{W}_g$를 이용하고 그 나머지인 가이드층으로부터 최종 출력층까지의 가중치인 W_s^1는 랜덤값으로 초기화하여 W_S 를 구성한다.

다음으로, 선생모델의 최종 출력층에서의 소프트맥스

출력값을 τ로 완화하고 학습하고자 하는 학생모델의 최종 출력값 또한 τ 로 완화하여 다음 식 (4)와 같이 비용함수를 구성하여 KD 학습을 수행한다[4~5].

$$C_{KD}(\widehat{W_s}) = E(Y_{true}, P_T) + \lambda E(P_T, P_S) \quad (4)$$

여기서, E는 크로스 엔트로피(cross-entropy)를, λ는 다중 비용함수 제어상수를, Y_{true}는 실제 레이블(true label)을 가지는 실측 데이터(ground truth)를, $P_T = softmax(a_T/\tau)$, $P_S = softmax(a_S/\tau)$, a_T는 선생모델에서 소프트맥스 출력 전의 입력벡터를, a_S는 학생모델에서 소프트맥스 출력 전의 입력벡터를, τ는 완화상수를 각각 나타낸다.

따라서, 식 (2)를 이용하여 선생모델에서 입력층으로부터 힌트층까지의 가중치인 W_h를 닮아가도록 학생모델의 $\widehat{W_g}$를 구하는 힌트학습을 먼저 수행하고, 식 (3)을 이용하여 $\widehat{W_g}$로부터 초기 가중치인 W_S를 구성한 다음, 식 (4)를 이용하여 힌트정보와 KD정보를 모두 이용하는 힌트-KD 학습을 최종 수행한다.

Ⅲ. 실험 결과

본 장에서는 최신 딥러닝 모델인 ResNet에 대하여 Ⅱ장에서 기술한 힌트-KD 학습 시 λ값에 대한 영향을 살펴본다. 본 실험에서 고려한 선생모델 및 학생모델은 [6]의 구조(층수 = 6m + 2, m = 1, 2, …)에 따라 모두 동일한 특징맵 크기를 가지고 층수가 서로 다른 ResNet을 사용하였으며, 초기 학습율(learning rate)은 0.1로 시작하여 32000 훈련횟수(iteration)에서 0.01로 1/10로 줄인 다음, 다시 48000 훈련횟수에서 0.001로 다시 1/10로 줄인 다음, 64000 훈련횟수에서 학습을 최종 종료하였다. [5]의 기준을 고려하여 선생모델 및 학생모델에 대한 힌트층 및 가이드층은 중간층을 선택하였다.

1. CIFAR10을 이용한 힌트-KD 학습 결과

먼저, 딥러닝 훈련 및 테스트 데이터셋으로 널리 사용되는 32×32 CIFAR10 영상 데이터에[11] 대하여 λ값에 대한 최신 ResNet 모델 기반 선생-학생 프레임워크의 힌트-KD 학습 성능에 대하여 살펴본다. 실험조건으로 92.08%의 인식 성능을 가지는 {16,32,64} 특징맵 개수의 26층 ResNet 선생모델과 8층 ResNet 학생모델에 대하여, 식 (2)의 힌트학습 시 초기 학습율은 0.0001로 시작하여 25000 훈련횟수에서 0.00001로 1/10로 줄인 다음 35000 훈련횟수에서 힌트학습을 최종 종료하였다. 이후, 식 (3)을 이용하여 힌트학습으로부터 구한 동일한 가중치와 나머지 층에서의 서로 다른 랜덤값으로 구성된 가중치를 초기화하여 학생모델을 3개로 복사한 후 (학생모델1, 학생모델2, 학생모델3) 식 (4)의 힌트-KD 학습을 수행한다. 이때, 완화상수는 [5]에서 제시된 것과 같이 τ=3으로 설정하였다. 다음의 표 1은 여러 개의 λ 값에 대하여 50000개의 훈련데이터를 [6]에서 제시된 데이터 전처리를 통하여 128배치 크기로 힌트-KD 학습을 수행한 후, 10000개의 테스트 데이터에 대한 인식율을 보여준다. 이때, λ는 학습 시 고정된 값으로 유지하였다. 표 1의 결과와 같이 선생모델의 힌트정보와 완화상수 τ를 이용한 KD 정보 모두를 고려하는 것이 상기 두 정보를 이용하지 않은 기존 학생 모델의 인식 성능(λ=0,힌트(X))뿐만 아니라 선생모델의 힌트정보만을 이용하여 학습한 경우(λ=0,힌트(O)) 보다 더 좋은 인식 성능을 보여주는 것을 관찰할 수 있다. 또한, λ=6인 경우, 선생모델의 힌트정보 없이 단지 KD 학습[4]만 수행한 인식 성능이(λ=6,힌트(X)) 힌트 및 KD 정보 모두 이용하는 경우보다(λ=6,힌트(O)) 성능이 열화되는 것을 관찰할 수 있다.

표 1. λ 값에 대한 인식율 [%] 비교 (CIFAR10)
Table1. Recognition rate [%] as to λ (CIFAR10).

정확도[%]	학생모델1	학생모델2	학생모델3	평균값
λ=0,힌트(X)	88.03	88.01	88.24	**88.09**
λ=0,힌트(O)	88.73	88.71	88.48	**88.64**
λ=1,힌트(O)	88.64	88.72	88.6	88.65
λ=2,힌트(O)	88.59	88.6	89.04	88.74
λ=3,힌트(O)	88.92	88.74	88.69	88.79
λ=4,힌트(O)	88.92	88.74	88.69	88.78
λ=5,힌트(O)	88.95	88.83	88.84	88.87
λ=6,힌트(O)	88.74	88.94	89.05	**88.91**
λ=6,힌트(X)	88.95	88.53	88.5	***88.66**

다음의 그림 1은 8층 ResNet 학생모델과 다른 층수들을 가지는 선생모델들에 대하여 λ 값에 대한 인식 성능 결과를 보여준다. 실험조건은 위의 표 1과 동일하게 설정하였고 모든 λ에 대하여 힌트정보를 모두 포함시켰다. 그림의 결과와 같이 선생모델의 힌트정보와 완화상수 τ를 이용한 KD 정보 모두를 이용하는 것이 학생모델의 인식 정확도가 향상된다는 것을 관찰할 수 있다. 다음의 표 2는 [5]의 연구와 같이 힌트학습 후 KD 학습 시 λ값을 각 초기값에서 1로 단조감소하였을 때의 인식율을 보여준다. 이때, 상기 표 1 및 그림 1의 각 선생모

델에 대하여 인식율이 가장 높은 λ에 대하여 단조감소를 적용하였다. 표 2의 결과에서와 같이 λ값을 고정한 경우의 평균 인식율이 단조감소하는 경우보다 더 좋은 것을 관찰할 수 있다.

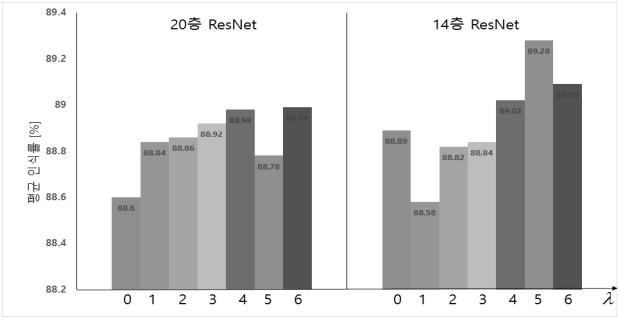

그림 1. 서로 다른 층수를 가지는 선생모델에 대하여 λ 값에 대한 인식율 [%] 비교 (CIFAR10)

Fig. 1. Recognition rate [%] as to λ when teacher models with different layers are considered (CIFAR10).

표 2. 단조감소하는 λ값에 대한 인식율 [%] (CIFAR10)

Table2. Recognition rate [%] as to annealing λ (CIFAR10).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	*평균값
λ=6→1, 힌트(O), 26층 선생모델	88.88	88.82	88.7	88.8 / (**88.91**)
λ=6→1, 힌트(O), 20층 선생모델	88.62	88.71	88.79	88.71 / (**88.99**)
λ=5→1, 힌트(O), 14층 선생모델	88.99	88.6	88.79	88.79 / (**89.28**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

다음의 표 3은 [4]의 힌트없이 KD정보만을 이용하는 학습방법에 대하여 표 2와 같이 λ값을 각 초기값에서 1로 단조감소하였을 때의 인식율을 보여준다. 표 3의 결과 또한 λ값을 줄이지 않고 고정하는 것이 인식 정확도 측면에서 더 유리하다는 것을 알 수 있다.

표 3. 단조감소하는 λ값에 대한 인식율 [%] (CIFAR10)

Table3. Recognition rate [%] as to annealing λ (CIFAR10).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	*평균값
λ=6→1, 힌트(X), 26층 선생모델	88.04	87.92	88.47	88.14 / (**88.66**)
λ=6→1, 힌트(X), 20층 선생모델	87.7	88.21	88.05	87.99 / (**88.52**)
λ=5→1, 힌트(X), 14층 선생모델	88.26	88.42	88.11	88.26 / (**88.41**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

2. CIFAR100을 이용한 힌트-KD 학습 결과

본 절에서는 CIFAR10 영상 데이터와 함께 널리 활용되고 있는 32×32 CIFAR100 영상 데이터에[11] 대하여 λ값에 대한 힌트-KD 학습 성능에 대하여 살펴본다. 실험조건은 Ⅲ.1절의 CIFAR10 경우와 동일하고, 여러 실험 환경을 반영하기 위하여 학습 시 훈련 데이터에 대한 전처리는 수행하지 않았다. 또한, CIFAR100 데이터 구성은 CIFAR10에 비하여 클래스가 100개로 늘어나는 반면 각 클래스당 훈련 데이터 개수는 그만큼 줄어들기 때문에, ResNet 모델에서 특징맵 개수를 Ⅲ.1절의 ResNet 모델보다 4배 증가시켰다 ({64,128,256}). 다음의 표 4 및 표 5는 63.66%의 인식 성능을 가지는 26층 ResNet 선생모델 및 61.14%의 인식 성능을 가지는 14층 ResNet 선생모델에 대하여 λ에 따른 8층 ResNet 학생모델에 대한 인식 정확도 성능들을 각각 보여준다. CIFAR10 데이터셋의 경우와 마찬가지로 CIFAR100 데이터셋 경우에 대해서도 선생모델의 힌트정보와 완화상수 τ를 이용한 KD 정보 모두를 이용하는 것이 학생모델의 인식 정확도가 향상된다는 것을 관찰할 수 있다.

표 4. 26층 선생모델에 대하여 λ 값에 대한 인식율 [%] 비교 (CIFAR100).

Table4. Recognition rate [%] as to λ for 26-layer teacher model (CIFAR100).

정확도[%]	학생모델1	학생모델2	학생모델3	평균값
λ=0,힌트(X)	56.51	54.89	55.88	**55.76**
λ=0,힌트(O)	56.9	55.94	57.27	**56.70**
λ=1,힌트(O)	59.28	58.84	59.47	59.2
λ=2,힌트(O)	59.09	59.3	59.22	59.2
λ=3,힌트(O)	59.78	59.76	59.14	**59.56**
λ=3,힌트(X)	57.22	58.17	57.84	***57.74**
λ=4,힌트(O)	59.68	59.46	59.15	59.43
λ=5,힌트(O)	59.11	59.89	59.0	59.33
λ=6,힌트(O)	59.55	59.4	59.24	59.40

표 5. 14층 선생모델에 대하여 λ 값에 대한 인식율 [%] 비교 (CIFAR100)

Table5. Recognition rate [%] as to λ for 14-layer teacher model (CIFAR100).

정확도[%]	학생모델1	학생모델2	학생모델3	평균값
λ=0,힌트(X)	56.51	54.89	55.88	**55.76**
λ=0,힌트(O)	55.84	56.71	58.47	**57.0**
λ=1,힌트(O)	58.74	58.49	58.25	58.49
λ=2,힌트(O)	58.54	59.15	58.03	58.57
λ=3,힌트(O)	58.73	58.86	58.7	58.76
λ=4,힌트(O)	58.93	59.04	58.88	**58.95**
λ=4,힌트(X)	57.87	57.58	57.98	***57.81**
λ=5,힌트(O)	58.61	58.22	58.61	58.48
λ=6,힌트(O)	58.58	58.57	59.15	58.77

또한, 표 6 및 표 7의 결과와 같이 CIFAR100 데이터셋에 대해서도 KD 정보만을 이용하는 경우(표 6)와 힌트 및 KD 정보 모두를 이용하는 경우(표 7)에 대하여 λ값을 단조감소하는 것보다 초기의 고정된 값으로 유지하는 것이 학생 모델의 인식 정확도 측면에서 더 유리하다는 것을 관찰할 수 있다. 이때, 상기 표 4 및 표 5의 각 선생모델에 대하여 인식 정확도가 가장 높은 λ에 대하여 단조감소를 적용하였다.

표 6. 단조감소하는 λ값에 대한 인식율 [%] (CIFAR100)
Table6. Recognition rate [%] as to annealing λ (CIFAR100).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	평균값
λ=3→1, 힌트(X), 26층 선생모델	57.09	57.77	57.65	57.5 / (**57.74**)
λ=4→1, 힌트(X), 14층 선생모델	57.47	57.32	57.64	57.48 / (**57.81**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

표 7. 단조감소하는 λ값에 대한 인식율 [%] (CIFAR100)
Table7. Recognition rate [%] as to annealing λ (CIFAR100).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	평균값
λ=3→1, 힌트(O), 26층 선생모델	59.25	59.35	59.78	59.46 / (**59.56**)
λ=4→1, 힌트(O), 14층 선생모델	58.43	58.81	58.15	58.46 / (**58.95**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

3. MNIST를 이용한 힌트-KD 학습 결과

앞의 실험 결과에서 살펴본 KD 정보에 대한 영향을 좀더 검증하기 위하여 영상인식에 널리 활용되고 있는 28×28 MNIST 필기체 영상 데이터[12]를 이용하여 ResNet 모델을 이용한 선생-학생 프레임워크 기반 힌트-KD 학습 성능에 대하여 살펴본다. 실험조건으로 0.39% 인식 오차율(100-인식율[%])을 가지는 {16,32,64} 특징맵 개수의 32층 ResNet 선생모델과 8층 ResNet 학생모델을 이용하였으며, 식 (2)의 힌트학습 시 초기 학습율은 0.0001로 시작하여 20000 훈련횟수에서 0.00001로 1/10로 줄인 다음 25000 훈련횟수에서 힌트학습을 종료하였다. 다음으로, 식 (4)의 힌트-KD 학습 시 초기 학습율은 0.1로 시작하여 18000 훈련횟수에서 0.01로 1/10로 줄인 다음, 다시 27000 훈련횟수에서 0.001로 다시 1/10로 줄인 다음, 36000 훈련횟수에서 학습을 최종 종료하였다. 이때, 데이터 전처리가 없는 64배치 크기의 훈련 데이터를 이용하였다. 표 8의 결과에서와 같이 CIFAR10/100 데이터셋들의 경우와 마찬가지로 선생모델의 힌트 및 KD 정보 모두를 이용하는 것이 학생모델의 인식 정확도가 향상된다는 것을 관찰할 수 있다.

표 8. 32층 선생모델에 대하여 λ 값에 대한 인식 오차율 [%] 비교(MNIST)
Table8. Recognition error rate [%] as to λ for 32-layer teacher model (MNIST).

정확도[%]	학생모델1	학생모델2	학생모델3	평균값
λ=0,힌트(X)	0.52	0.52	0.52	**0.52**
λ=0,힌트(O)	0.51	0.54	0.55	**0.533**
λ=3,힌트(O)	0.55	0.52	0.49	0.52
λ=4,힌트(O)	0.44	0.46	0.48	0.46
λ=5,힌트(O)	0.43	0.46	0.44	**0.443**
λ=5,힌트(X)	0.55	0.52	0.43	***0.5**
λ=6,힌트(O)	0.49	0.43	0.53	0.483

다음의 표 9 및 표 10의 결과들은 KD 정보만을 이용하는 경우(표 9)와 힌트 및 KD 정보 모두를 이용하는 경우(표 10)에 대하여 각각 λ값을 단조감소하였을 때의 인식 오차율을 보여준다. MNIST 데이터셋에 대해서도 λ값을 고정된 값으로 유지하지 않고 감소하는 경우가 인식 성능 향상에 더 불리하다는 것을 알 수 있다. 즉, 본 연구에서 수행한 여러 실험결과들을 바탕으로 λ값을 단조감소하는 경우 대비, 실측 데이터 정보(true label) 비중보다 선생모델의 완화상수를 고려한 KD 정보의 비중을 높게 설정하여 유지하는 것이 학생모델 학습 시 더 유리하다는 것을 알 수 있었다.

표 9. 단조감소하는 λ값에 대한 인식 오차율 [%] (MNIST)
Table9. Recognition error rate [%] as to annealing λ (MNIST).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	평균값
λ=5→1, 힌트(X), 32층 선생모델	0.59	0.57	0.65	0.603 / (**0.5**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

표 10. 단조감소하는 λ값에 대한 인식 오차율 [%] (MNIST)
Table10. Recognition error rate [%] as to annealing λ (MNIST).

정확도[%]	학생 모델1	학생 모델2	학생 모델3	평균값
λ=5→1, 힌트(O), 32층 선생모델	0.52	0.54	0.49	0.516 / (**0.443**)

* 평균값: λ값 단조감소 적용 시 평균인식율 / (고정된 λ값 적용 시 평균 인식율)

Ⅳ. 결 론

본 논문에서는 최신 딥러닝 모델에서 각광받고 있는 ResNet을 이용한 선생-학생 프레임워크에 대하여 Caffe 딥러닝 오픈 프레임워크를 활용하여 기존 힌트-KD 학습 기법을 적용한 인식 정확도 성능을 분석하였다. [5]의 결과와 달리 높은 층수를 가지는 ResNet 선생모델에 대하여 학습 시 λ값을 초기 설정값에서 1로 단조감소하는 경우보다 초기의 고정된 λ값을 사용하여 선생모델의 완화상수를 고려한 소프트맥스 출력 정보를 높은 비중으로 고정적으로 유지하는 경우가 더 유리하다는 것을 관찰할 수 있었다. 따라서, 본 연구에 따르면, 최신 딥러닝 모델인 ResNet을 이용한 선생-학생 프레임워크에 대하여 힌트정보와 KD 정보 모두를 이용하는 것이 힌트 혹은 KD 정보만을 이용하여 학습한 경우보다 더 좋은 인식정확도를 얻을 수 있었다. 하지만, 기존 연구결과와 달리 KD 정보를 학습 시 단조감소하지 않고 높은 비중으로 유지하는 것이 선생모델이 힌트 및 KD 정보를 좀 더 효율적으로 학생모델에 반영할 수 있음을 실험결과를 토대로 관찰할 수 있었다.

REFERENCES

[1] Y.-T. Park, “A comparative study of image recognition by neural network classifier and linear tree classifier”, Journal of The Institute of Electronics and Information Engineers-B, vol. 31, no. 5, pp. 141-148, 1994.

[2] S. Hong, W. Im, J. Park, and H.-S. Yang, “Deep CNN-based person identification using facial and clothing features”, in Proc. of Summer Conference on Institute of Electronics and Information Engineers (IEIE), pp. 2204-2207, June, 2016.

[3] Y. Shin, J.-H. Park, S. Shin, G. Lim, S. Song, C. Lee, and J.-M. Chung, “Improvement of image classification in augmented reality based on deep learning”, in Proc. of Summer Conference on Institute of Electronics and Information Engineers (IEIE), pp. 1771-1773, June, 2016.

[4] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural netwok”, arXiv prreprint arXiv:1503.02531, pp. 1-19, 2015.

[5] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets”, in Proc. of 5th International Conference on Learning Representations (ICLR), pp. 1-13, San Diego, May 7-9, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-12, Las Vegas, June 26-July 1, 2016.

[7] A. Veit, M. Wilber, and S. Belongie, “Residual networks are exponential ensembles of relatively shallow networks”, arXivpreprint arXiv:1605.06431, pp. 1-12, 2016.

[8] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks”, arXiv:1302.4389, pp. 1-9, 2013.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, in Proc. of 5th International Conference on Learning Representations (ICLR), pp. 1-14, San Diego, May 7-9, 2015.

[10] [Online available] “Caffe, deep learning frame work”, http://caffe.berkeleyvision.org/

[11] [Online available] “CIFAR-10 and CIFAR-100 datasets”, https://www.cs.toronto.edu/~kriz/cifar.html

[12] [Online available] “MNIST dataset”, http://yann.lecun.com/exdb/mnist

저 자 소 개

배 지 훈(정회원)
2000년 2월 경북대학교 전자공학과 학사 졸업.
2002년 2월 포항공과대학교 전자전기공학과 석사 졸업.
2016년 2월 포항공과대학교 전자전기공학과 박사 졸업.
2002년 1월～현재 한국전자통신연구원 책임연구원
<주관심분야 : 딥러닝 및 기계학습, 레이더 영상 신호처리, HF/UHF RFID 시스템, RFID 디지털 모뎀, 배열안테나 빔형성, 최적화 기법>

임 준 호(정회원)
2012년 8월 한국과학기술원 전기 및 전자공학부 학사 졸업.
2015년 2월 한국과학기술원 전기 및 전자공학부 석사 졸업.
2015년 3월～현재 한국과학기술원 전기 및 전자공학부 박사 과정
<주관심분야 : 딥러닝 및 기계학습, Computer Vision>

유 재 학(정회원)
2001년 건국대학교 전산과학과 학사 졸업.
2003년 고려대학교 전산학과 석사 졸업.
2010년 고려대학교 전산학과 박사 졸업.
2006년 3월～2008년 2월 고려대학교 컴퓨터정보학과 초빙전임강사.
2010년～현재 한국전자통신연구원 선임연구원
<주관심분야 : 기계학습, 딥러닝, 데이터 마이닝, 시맨틱 IoT/IoE 플랫폼, 네트워크 마이닝, 침입탐지>

김 귀 훈(정회원)
1998년 2월 한국과학기술원 전기 및 전자공학과 학사 졸업.
2000년 2월 한국과학기술원 전기 및 전자공학과 석사 졸업.
2013년 2월 한국과학기술원 전기 및 전자공학과 박사 수료.
2000년 2월～2005년 6월 LG데이콤 전임연구원
2005년 7월～현재 한국전자통신연구원 책임연구원
<주관심분야 : 머신러닝, 딥러닝, 강화학습, GAN, IoT, 5G이동통신, 컴퓨터, 신호처리, IPTV, 스마트 TV, 디지털홀로그래피>

김 준 모(정회원)
1998년 8월 서울대학교 전기공학부 학사 졸업.
2000년 8월 MIT EECS 석사 졸업.
2005년 2월 MIT EECS 박사 졸업.
2005년 5월～2009년 6월 삼성종합기술원 전문연구원.
2009년 7월～2016년 8월 한국과학기술원 전기 및 전자공학부 조교수
2016년 9월～현재 한국과학기술원 전기 및 전자공학부 부교수
<주관심분야 : 딥러닝 및 기계학습, Statistical Signal Processing, Image Processing & Computer Vision, Information Theory>