On the Efficacy of Knowledge Distillation

Abstract

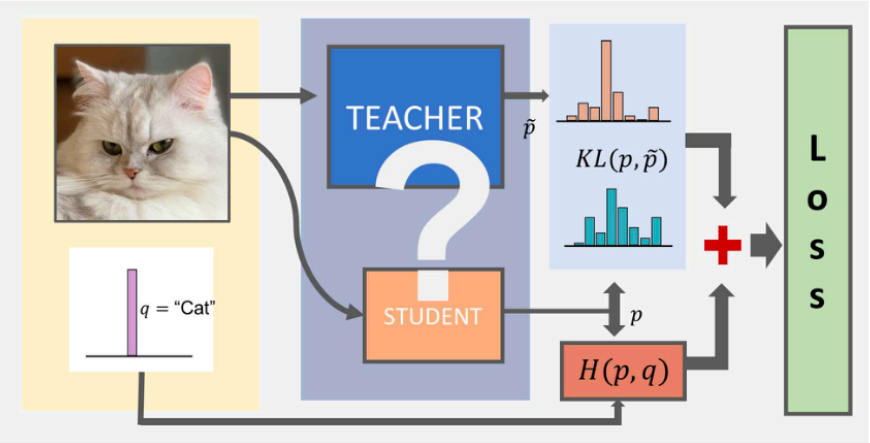

Figure 1. An illustration of standard knowledge distillation. Despite widespread use, an understanding of when the student can learn from the teacher is missing.

본 논문에서는 knowledge distillation의 효율성을 분석하고, 더 크거나 정확도가 높은 teacher 모델이 항상 좋은 teacher 모델이 되지는 않는다는 점을 발견했습니다. 특히 student 모델의 용량(capacity)이 teacher 모델과 큰 차이가 나면, student 모델이 teacher 모델을 효과적으로 모방하는 데 어려움이 있다고 주장합니다. 이러한 문제를 완화하기 위한 방법으로 teacher 모델의 학습을 완전히 진행하지 않고 조기에 중단(early stopping)하는 것이 보다 효율적임을 제안하고 있습니다.

Introduction

최근 다양한 이미지 인식 모델들이 발전했습니다. 이러한 모델들은 더 깊고 복잡한 구조를 통해 높은 성능을 얻었지만, 연산량이 많아 범용적인 응용 분야에서 활용하는 데 제약이 있습니다. 따라서 모델의 파라미터 수를 줄이는 pruning 기법이나 큰 모델에서 작은 모델로 지식을 이전하는 knowledge distillation(KD) 기법이 활발히 연구되고 있습니다.

지식 증류(KD)는 일반적으로 어떤 teacher 모델과 student 모델 사이에서도 잘 작동할 것이라고 예상되지만, 실제로는 잘 작동하지 않는 사례가 보고된 바 있습니다. 본 논문은 이러한 현상에 주목하여, 어떤 이유로 KD가 잘 작동하는 teacher-student 모델 조합과 잘 작동하지 않는 조합이 존재하는지에 대한 의문을 제기합니다. 구체적으로, 지식 증류가 잘 되는 특정한 teacher-student 모델의 조합이 있는지? 성능이 떨어지는 조합이 존재한다면 이를 개선할 수 있는 방법은 무엇인지? 라는 질문에 답하고자 합니다.

본 논문의 연구 결과에 따르면, Teacher 모델의 정확도(accuracy)는 student 모델의 성능을 예측하는 좋은 지표가 되지 못합니다. 오히려 teacher 모델의 용량(capacity)이 지나치게 크고 정확도가 높을수록 student 모델이 이를 효과적으로 모방하는 데 어려움이 있다고 주장합니다.

또한, 기존에 제안된 단계별 지식 증류(Sequential KD) 방법도 이러한 문제를 해결하는 데 별다른 효과가 없다는 점을 확인하였습니다.

이에 본 논문에서는 teacher 모델 학습을 완전히 진행하지 않고 조기에 중단(Early stopping) 하는 방식으로 teacher 모델을 정규화하고, student 모델의 학습 또한 완전한 수렴(convergence)에 가까워졌을 때 멈추는 방법을 제안합니다. 이를 통해 student 모델이 teacher 모델을 보다 효과적으로 모방하여 성능이 향상될 수 있음을 실험적으로 증명했습니다.

> 즉, 본 논문은 기존의 knowledge distillation 방식이 적용되지 않는 문제점을 teacher model - student model 사이의 용량(capacity) 등 으로 설명하였고, 이를 해결하기 위해서 teacher model과 student model의 early stopping을 제안하였습니다.

Background: Knowledge distillation

본 논문에서는 기본적인 knoledge distillation에 대해서 설명하고 있습니다. 기본적인 knowledge distillation의 경우 teacher 모델의 logit 값이 T라는 하이퍼 파라미터를 사용하여 soft label 분포를 만들게 됩니다. T = 1은 일반적인 softmax이며, T가 클수록 출력 logit 확률이 균등해지게 됩니다.

$$\tilde{p}_t^k(x) = \frac{e^{s_t^k(x)/\tau}}{\sum_j e^{s_t^j(x)/\tau}}$$

그리고 기본적인 라벨로 부터 나온 손실과 사용하기 위해서 아래와 같은 수식을 통해서 최종 손시을 구할 수 있습니다.

$$L = \alpha L_{cls} + (1 - \alpha) L_{KD}$$

그리고 soft label을 사용한 손실의 경우, teacher model과 student model의 cross entropy 값을 통해 구할 수 있고, 이를 수식화 하면 아래와 같습니다.

$$L_{KD} = -\tau^2 \sum_k \tilde{p}_t^k(x) \log \tilde{p}_s^k(x)$$

그리고 일반적으로 $\tau \in \{3, 4, 5\},\ \alpha = 0.9$ 다음과 같은 하이퍼 파라미터 조합이 주로 사용된다고 합니다.

Result

Bigger models are not better teacher

본 논문에서는 항상 정확도가 높고, 큰 모델이 좋은 teacher 모델이 아님을 실험을 통해서 증명합니다.

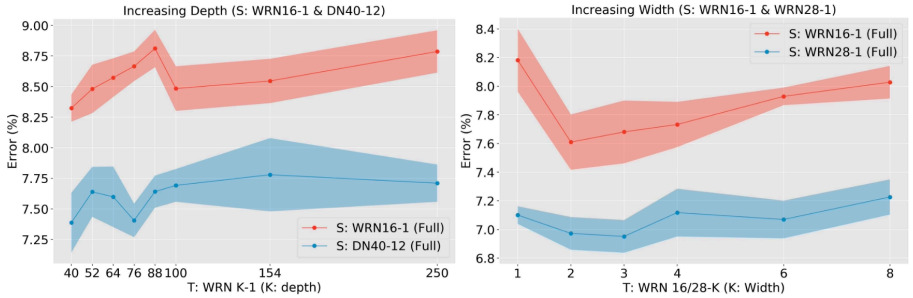

Figure 2. The error plot of student networks distilled from different teachers on CIFAR10. WideResNet [25] 16-1 (**left/red**, **right/red**), 28-1 (**right/blue**)), and DenseNet [13] 40-12 (**left/blue**) were used as the student networks. Increasing teacher capacity (depth: **left**, width: **right**) and thus accuracy does not necessarily increase the accuracy of the student network, indicating that the accuracy of the teacher network alone is not a valid metric to knowledge distillation.

CIFAAR10 데이터를 통한 실험 결과

Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

Table 1. Top-1 error rate for various teachers for a ResNet18 student on ImageNet. The first row corresponds to training from scratch.

해당 그래프에서 빨간색과 파란샌 그래프는 서로 다른 Teacher 모델로 부터 학습된 동일한 구조의 student 모델입니다. 왼쪽은 깊이를 오른 쪽은 너비를 키워가며 실험한 결과 입니다. 그래프와 같이 teacher 모델의 크기가 커질 수록 student 모델의 성능 또한 올라가다가, 특정 시점에서 student 모델의 성능이 하락함을 확인할 수 있습니다. 이를 통해서 항상 거대하고 정확도가 높은 모델이 teacher 모델로서 좋은 역할을 하지는 않음을 입증합니다.

본 논문에서는 이러한 현상이 일어나는 한가지 문제로 Teacher 모델의 정확도가 높아 출력 확률 분포가 one-hot-encoding에 가까운 분포를 갖게 되면서 Student 모델이 학습할 정보가 줄어든다고 주장합니다. 그래서 이를 해결하기 위해서 Temperature를 높혀서 진행 했을 때도 결과가 변경되지 않았다고 합니다.

Analyzing student and teacher capacity

본 논문에서는 위와 같은 문제의 원인을 파악하기 위해서 2가지 접근을 사용합니다.

1. Student 모델이 Teacher 모델을 잘 모방하지만, 이러한 모방이 Student 모델의 성능 향상에 도움이 되지 않는다.
2. Student 모델이 Teacher 모델을 모방하지 못한다.

Student	Teacher	KD Error (%,Train)	KD Error (%,Test)
WRN28-1	WRN28-3	0.23	4.05
	WRN28-4	0.25	4.53
	WRN28-6	0.23	4.54
	WRN28-8	0.31	4.81
WRN16-1	WRN16-3	1.70	6.32
	WRN16-4	1.69	6.52
	WRN16-6	1.94	6.91
	WRN16-8	1.69	7.01

Table 2. KD error on CIFAR10 for multiple teachers and students. The supplementary shows similar results from teachers with increasing depth.

위의 테이블에서는 Student 모델과 Teacher 모델이 예측한 정답이 다른 비율을 보여주고 있습니다.

그리고 실험 결과 모두 Teacher Model의 크기가 가장 큰 경우 Student 모델의 손실이 가장 높게 나왔습니다. 이러한 실험 결과를 통해서 Teacher 모델의 용량에 비해 Student 모델의 용량이 충분히 크지 못하다면, Teacher 모델이 출력하는 출력 분포를 표현하는 결과를 얻지 못한다고 주장합니다.

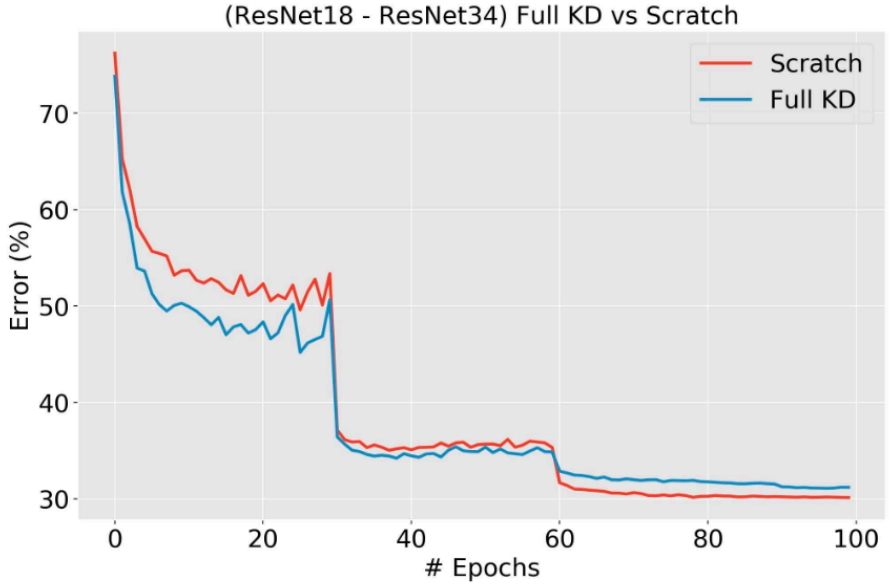

Figure 3. Imagenet result of error plot of full knowledge distillation and training from scratch. In the figure the student is trained with ResNet34. Knowledge distillation helps initially but starts to hurt accuracy later in training. The same behavior occurs in the plots with different teachers (more plots in Supplementary).

특히 ImageNet에 대해서는 단순히 student 모델을 학습 시키는 것보다 KD 방식으로 학습 시키는 것이 안좋은 성능을 보여주었습니다. 초기에는 그래도 빠르게 수렴하는 듯 하지만, 후반으로 갈 수록 단순히 Student 모델을 학습하는 것 보다 수렴 속도가 늦어지게 됨을 확인할 수 있습니다.

이에 본 논문에서는 용량이 작은 Student 모델의 경우 hard label 손실과 soft label 손실을 동시에 최소화 할 수 있는 표현력을 갖을 수 없어서 Underfitting 된 상태라고 주장합니다.

이러한 주장을 증명하기 위해서 초반에는 KD 방식으로 학습 하고 이후에는 cross-entropy loss만을 최소화 하도록 학습하도록 하였습니다. 그리고 이렇게 KD → CorssEntropy 로 학습하는 방식을 " Early stopped knowledge distilltaion(ESKD)" 라고 제안합니다. 그리고 기존의 KD는 full KD 라고 명명합니다.

Teacher	Top-1 Error (%, Test)	CE (Train)	KD (Train)	KD (Test)
ResNet18	30.57	0.146	2.916	3.358
ResNet18 (ES KD)	29.01	0.123	2.234	2.491
ResNet34	30.79	0.145	1.357	1.503
ResNet34 (ES KD)	29.16	0.123	2.359	2.582
ResNet50	30.95	0.146	1.553	1.721
ResNet50 (ES KD)	29.35	0.124	2.659	2.940

Table 3. Early-stopping the knowledge distillation can prevent the student from degrading its classification performance on ImageNet.

3가지 실험 결과 모든 부분에서 ESKD 방식으로 학습한 모델의 성능이 더 좋음을 확인할 수 있었습니다. 추가적으로 CE 손실과 KD 손실이 반비례 관계를 가지고 있음 또한 알 수 있었습니다.

이렇게 간단하게 early stopping 방식만 사용해도 지식 증류의 원래 목적을 달성 할 수 있다고 주장합니다.

그래도 여전히 ESKD방식으로 학습한다고 해도 정확도가 높고, 용량이 큰 Teacher 모델일수록 Student 모델의 성능이 좋아진다는 가설은 틀리다고 주장합니다. 이에 Student 모델의 성능을 결정짓는 것은 Student 모델의 용량 (Capcaity) 라고 주장합니다.

The efficacy of repeated knowledge distillation

만일 Teacher 모델과 Student 모델의 용량 차이가 매우 크다면, 중단 크기의 model을 통해서 단계적으로 지식을 증류하는 방법이 제안되어왔습니다. 이러한 방법은 Student 모델이 충분히 Teacher 모델의 표현력을 학습할 만큼 용량이 크다는 조건도 만족하게 됩니다.

본 논문에서는 이러한 과정을 증명하기 위해서 실험을 진행하였고, 이러한 과정에는 몇가지 문제가 있다고 주장합니다.

1. 일부 모델의 경우 순차적으로 KD 한 마지막 모델의 경우 단순히 데이터로 학습한 모델보다 성능이 낮다.
2. 순차적 KD로 학습한 모델은 단일 모델보다는 성능이 좋지만, 같은 수의 단일 모델을 앙상블 한 것 보다는 성능이 낮다 → KD 말고 단순히 앙상블 한게 더 성능이 좋다.

[Table 5 : 5단계를 거친 Student 모델의 성능이 단순히 데이터로 학습한 모델보다 성능이 낮다]

Teacher Training	Teacher Error (%)	Student Error (%)
Scratch	5.34	7.61 (7.68 ± 0.259)
5 KD iterations	4.89	7.79 (7.67 ± 0.19)

Table 5. Sequential knowledge distillation does not make better teachers even when it improves accuracy. The student is WRN16-1, which achieves an error of (8.759 ± 0.129) when trained from scratch. The teacher is WRN16-3.

[Table6 : large → mid → small 순서의 단계적 증류와 large → small의 증류는 크게 차이가 없다 = 순서적으로 증류하는 방식이 효과가 없다.]

Training Procedure	Large Error (%)	Medium Error (%)	Small Error (%)
Large→Med.→ Small	4.41	4.80	8.04 (7.99 ± 0.24)
Med. → Small	-	5.34	**7.614** ($\mathbf{7.68 \pm 0.26}$)
Large → Small	4.41	-	7.98 (8.03 ± 0.14)

Table 6. Using sequential knowledge distillation to distill from a large model (WRN16-8) to a medium model (WRN16-3), and from the latter to a small model (WRN16-1) does not help. The optimal approach still is to distill directly from the medium model to the small model, even though the teacher in this case has lower accuracy.

이러한 실험 결과는 ImageNet에서도 동일하게 나타났고, 오히려 작은 모델 → 더 작은 모델의 성능이 가장 좋았다고 합니다. 이에 이러한 Sequential KD 방식이 그다지 효과가 있지 않다고 주장합니다.

Early-stopped teachers make better teachers

앞선 실험을 통해서 순차적 KD 방식이 효과적이지 않음을 확인하였습니다. 이러한 문제는 결국 Teacher 모델의 표현력이 Student 모델의 표현력으로는 모방할 수 없는 차원 혹은 표현 공간에 놓여있기 때문입니다.

하지만 Student 모델에 최적화된 Teacher 모델을 찾는 것은 자원이 많이 필요하기에 본 논문에서는 Teacher 모델을 Regularization을 적용하는 방법을 제안합니다. 그 중 Teacher의 학습을 조기에 종료하는 Early stopping 방식을 제안합니다. 즉, 거대한 데이터의 표현력을 조금만 학습한다면 충분히 작은 모델에서도 이를 넓은 범위에서 모바할 수 있다는 개념입니다. 그리고 본 논문에서는 1/3 ~ 1/4 정도의 에폭만 학습하면 된다고 주장합니다.

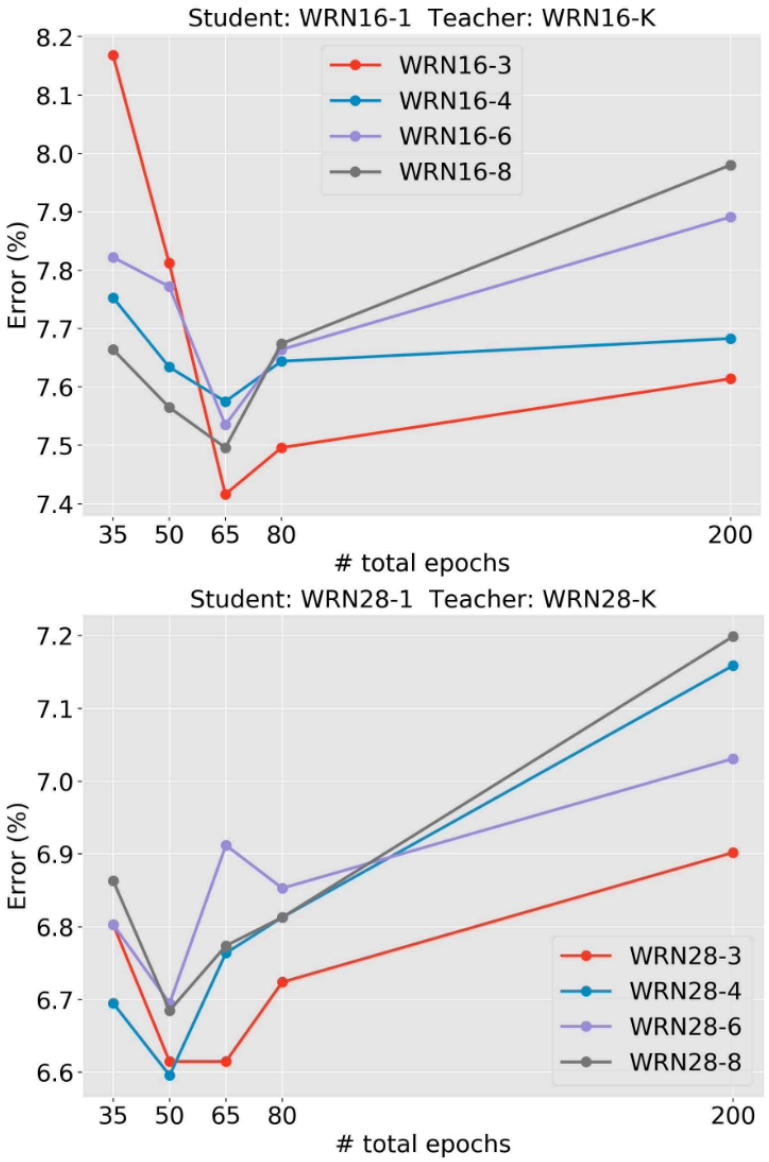

Figure 4. CIFAR10 result to examine the effectiveness of the knowledge distillation with early-stopped teachers. For both student types (WRN16-1 and WRN28-1), there are clear "sweet spots" which optimize the performance of student network.

CIFARA10에 대한 실험 결과 다음과 같이 특정 epoch 까지만 Teacher를 모방하도록 한 후 이후에는 단순히 데이터를 통해서 학습하는 방법이 효과적임을 확인할 수 있습니다 .그리고 각 모델별 Teacher 모델을 얼마나 학습해야하는지도 명확하게 드러나게 됩니다. 이러한 결과는 단순히 Early stopping을 사용함으로 써 성능 향상을 보임을 입증한 실험입니다.

추가로 Teacher모델의 내부 표현력도 모방하는 Attention Transfer 방식에서 Early stopping 방식은 더욱 효과적임을 제안합니다.

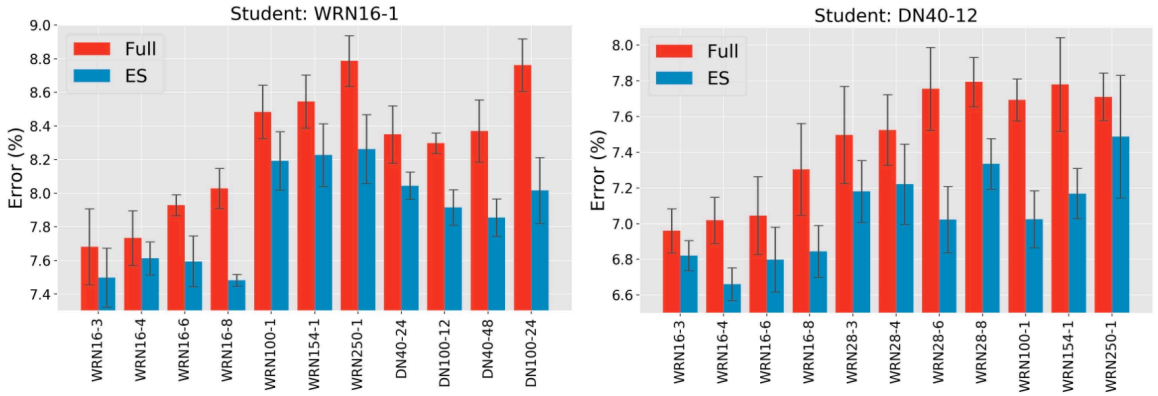

Figure 5. In all different student and teacher configurations, students trained from early-stopped teacher consistently outperforms those trained from regular teacher. DenseNet40-12 and WideResNet16-1 were used as student network and DenseNet and WideResNet models with varying width and depth were used as teachers (x-axis) [13, 25]. More results are in Supplementary.

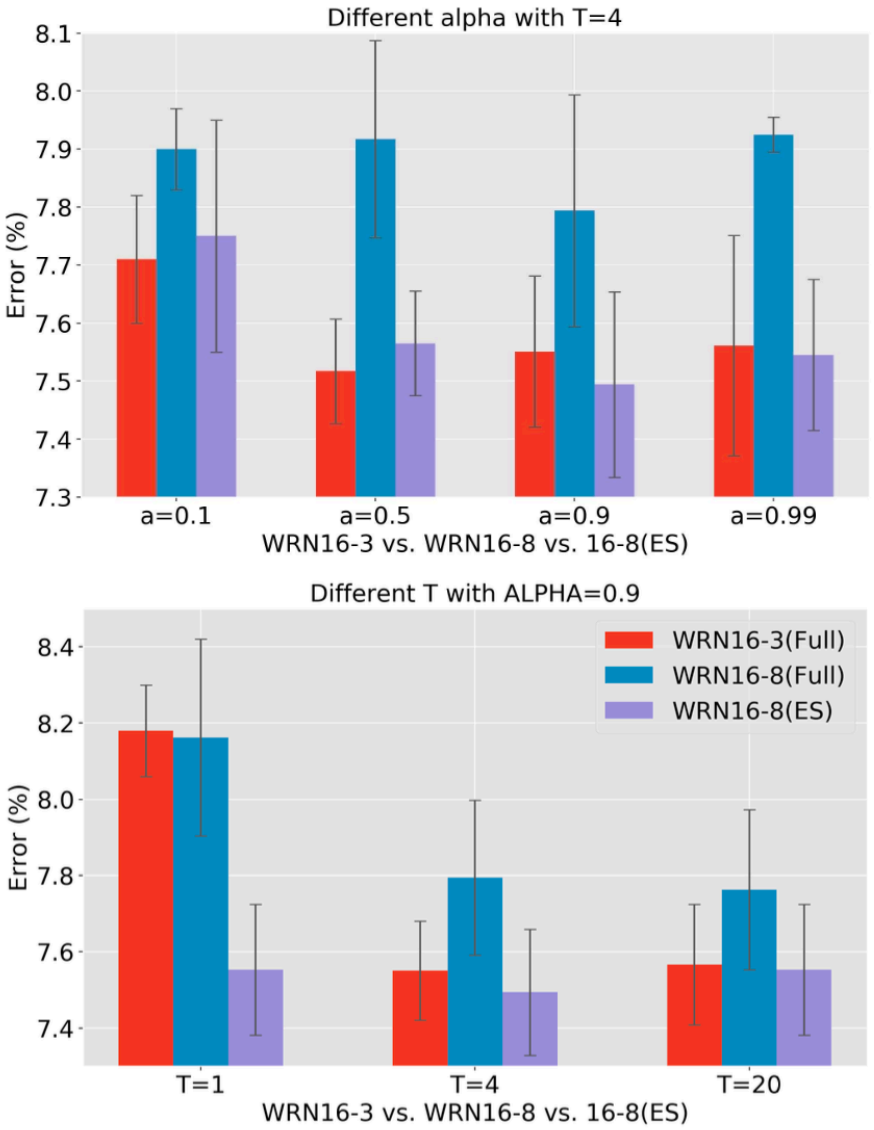

Figure 6. CIFAR10 result to examine that the effectiveness of using early-stopped network as a teacher is consistent to different hyperparameter settings. WRN16-8 (ES) has the same/better error compared to the optimal teacher WRN16-3. WRN16-1 is chosen as the student network.

그리고 다양한 실험을 통해서 하이퍼 파라미터, 다른 모델 등에서도 ESKD 방법이 유효함을 입증하였습니다.

Conclusion

본 논문에서는 KD와 관련된 포괄적인 실험을 진행하였고, KD방식은 만능이 아니며, Student 모델의 용량을 문제로 주장하였습니다. 그리고 Student 모델과 맞는 Teacher 모델을 찾는것 보다, Teacher 모델을 정규화 하는 Early stopping 을 적용한 ESKD를 통해서 이러한 문제를 해소할 수 있다고 주장합니다.

추가로 이러한 미묘한 차이와 동작 조건들에 대한 연구가 많이 필요하다고 주장합니다.

내 생각

결국 Student 모델의 용량에 딷라서 수용 가능한 표현력의 한계가 있기에 Early stopping 이 효과가 있는것 같다. 그리고 순차적으로 작은 모델을 학습 시키는 과정에 있어서 student 모델이 계속 비슷한 확률 분포를 학습하는데, 이때 라벨 스무딩이나 T를 바꿔가며 진행한다면 효과가 그래도 있지 않을까? 라는 의문이 든다.