Refine Myself by Teaching Myself
- FRSKD

2015104236
황 채 은

0. 조사해오기

- Confusion Matrix

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

- True Positive(TP) : 실제 True인 정답을 True라고 예측 (정답)
- False Positive(FP) : 실제 False인 정답을 True라고 예측 (오답)
- False Negative(FN) : 실제 True인 정답을 False라고 예측 (오답)
- True Negative(TN) : 실제 False인 정답을 False라고 예측 (정답)

https://sumniya.tistory.com/26

0. 조사해오기

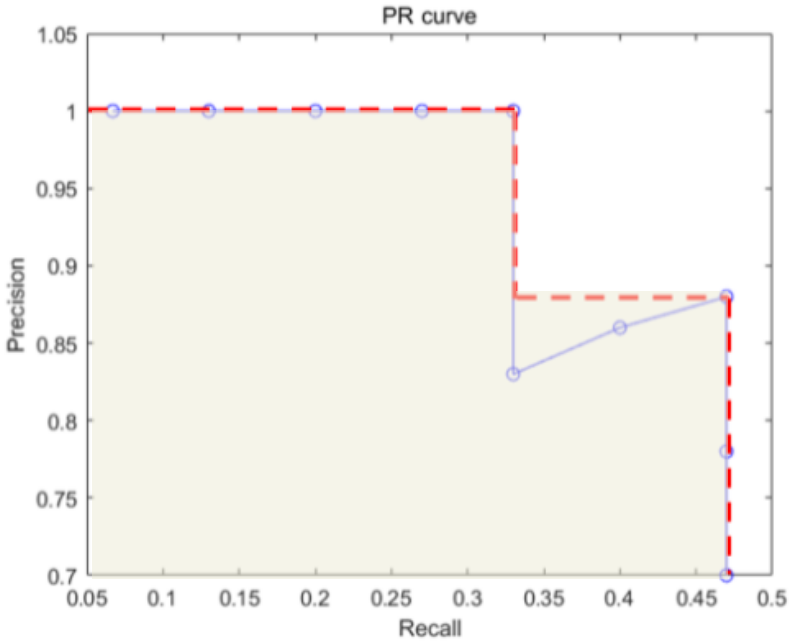
- Precision, Recall, Accuracy

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

- Prcecision : 정밀도 = $\frac{TP}{TP+FP}$, True로 분류된 것 중에 실제 True 인 비율
- Recall : 재현율 = $\frac{TP}{TP+FN}$, 실제 True인 것 중에 True라고 예측한 것의 비율
- Accuracy : 정확도 = $\frac{TP+TN}{TP+FN+FP+TN}$, 가장 직관적으로 모델의 성능을 평가할 수 있는 기준

0. 조사해오기

- AP : Average Precision. 물체 검출 알고리즘 성능을 평가하는 기준이 되며, Precision-recall 그래프에서 그래프 선 아래쪽의 면적으로 계산된다.



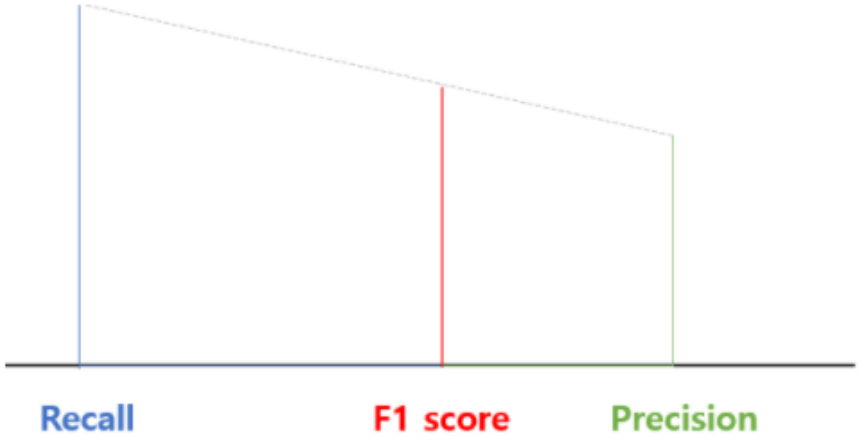
Precision-recall 그래프(= PR Curve)

https://bskyvision.com/465

0. 조사해오기

- F1 score : Precision과 Recall의 조화평균

$$(F1\text{-}score) = 2\times\frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\times\frac{Precision \times Recall}{Precision + Recall}$$

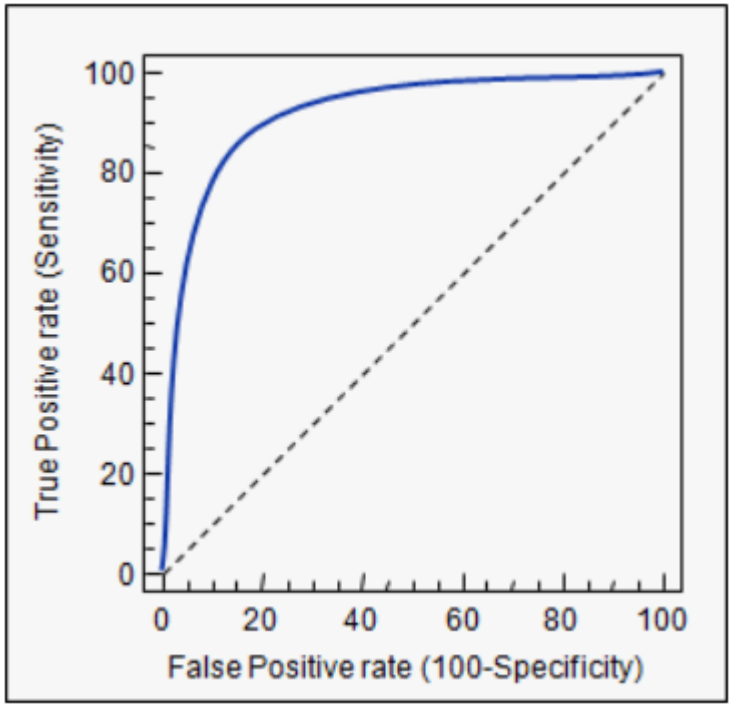


- 산술평균보다 좀 더 덜 치우친 평균을 구할 수 있음!

https://bskyvision.com/465

0. 조사해오기

- ROC Curve : Receiver Operating Characteristic Curve.
  - X축 : Fallout = $\frac{FP}{TN+FP}$, 실제 False인 것 중에 True라고 예측한 비율
  - Y축 : Recall = $\frac{TP}{TP+FN}$



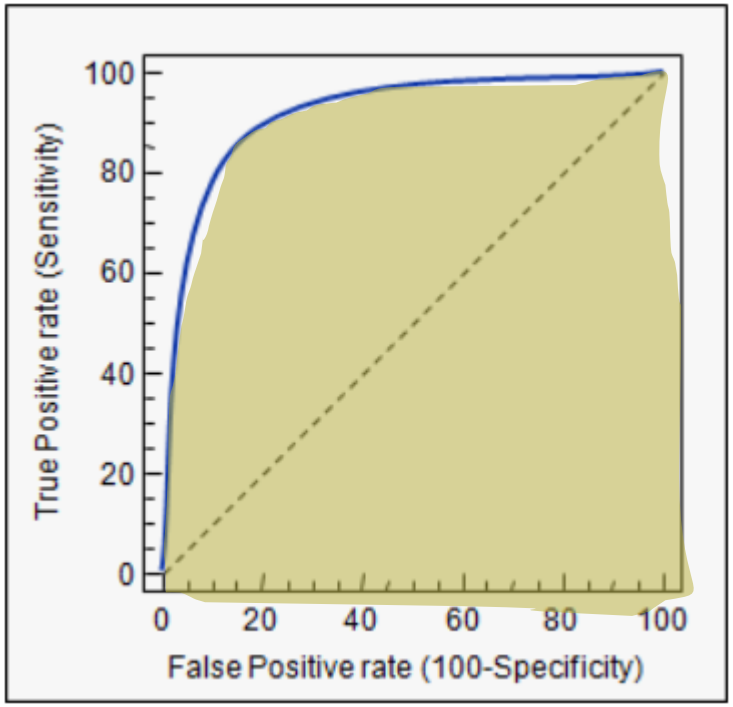
ROC Curve 그래프

Curve가
왼쪽 위 모서리에
가까울수록
모델의 성능이 좋다고
평가한다.

즉,
Recall이 크고
Fall-out이 작은
모형이 좋은 모형
(y=x 그래프보다
상단에 위치)

0. 조사해오기

- AUC : Area Under Curve. ROC 그래프 아래의 면적값을 의미.
  - 최댓값은 1
  - 좋은 모델일수록 1에 가까운 값이 나온다.(Fall-out에 비해 Recall값이 클수록)



ROC Curve 그래프

Refine Myself

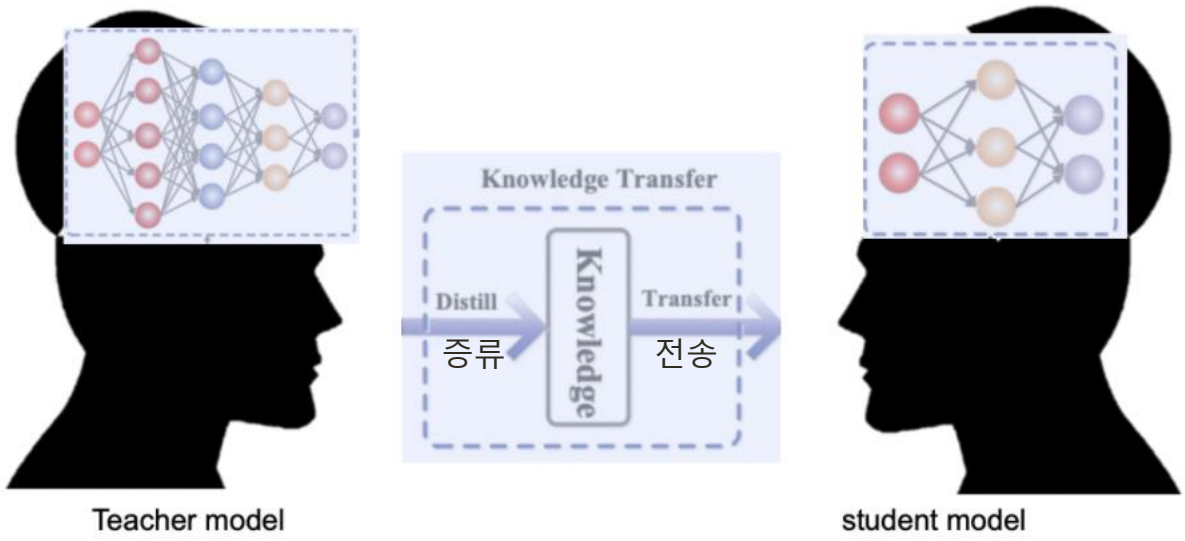
스스로를 다듬는다

by Teaching Myself

스스로를 가르침으로써

Abstract - 초록

- Knowledge Distillation : 지식 증류

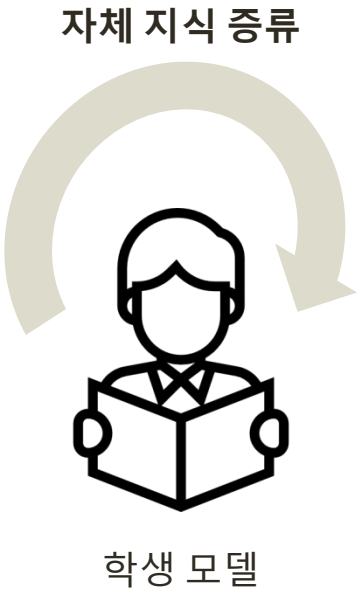


지식을 증류시켜 전달함으로써
모델의 크기는 줄이되, 중요한 정보는 남긴다

https://velog.io/@dldydldy75/%EC%A7%80%EC%8B%9D-%EC%A6%9D%EB%A5%98-Knowledge-Distillation

Abstract - 초록

- Self Knowledge Distillation : 자가 지식 증류



데이터 증강 기반 접근법

보조 네트워크 기반 접근법

Abstract - 초록

데이터 증강 기반 접근법

- 증강 과정에서 지역 정보 소실
- 다양한 비전 작업에의 적용 가능성이 낮음
- 정교한 특징 맵을 받지 못함

보조 네트워크 기반 접근법

- 제안 ▶새로운 자기 지식 증류 방법 필요

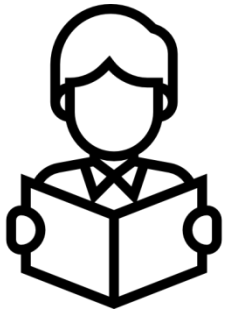
자기 지식 증류를 통한 특징 개선
(FRSKD, Feature Refinement via Self-Knowledge Distillation)

Abstract - 초록

자기 지식 증류를 통한 특징 개선
(FRSKD, Feature Refinement via Self-Knowledge Distillation)

보조 자기 교사 네트워크

정제된 지식

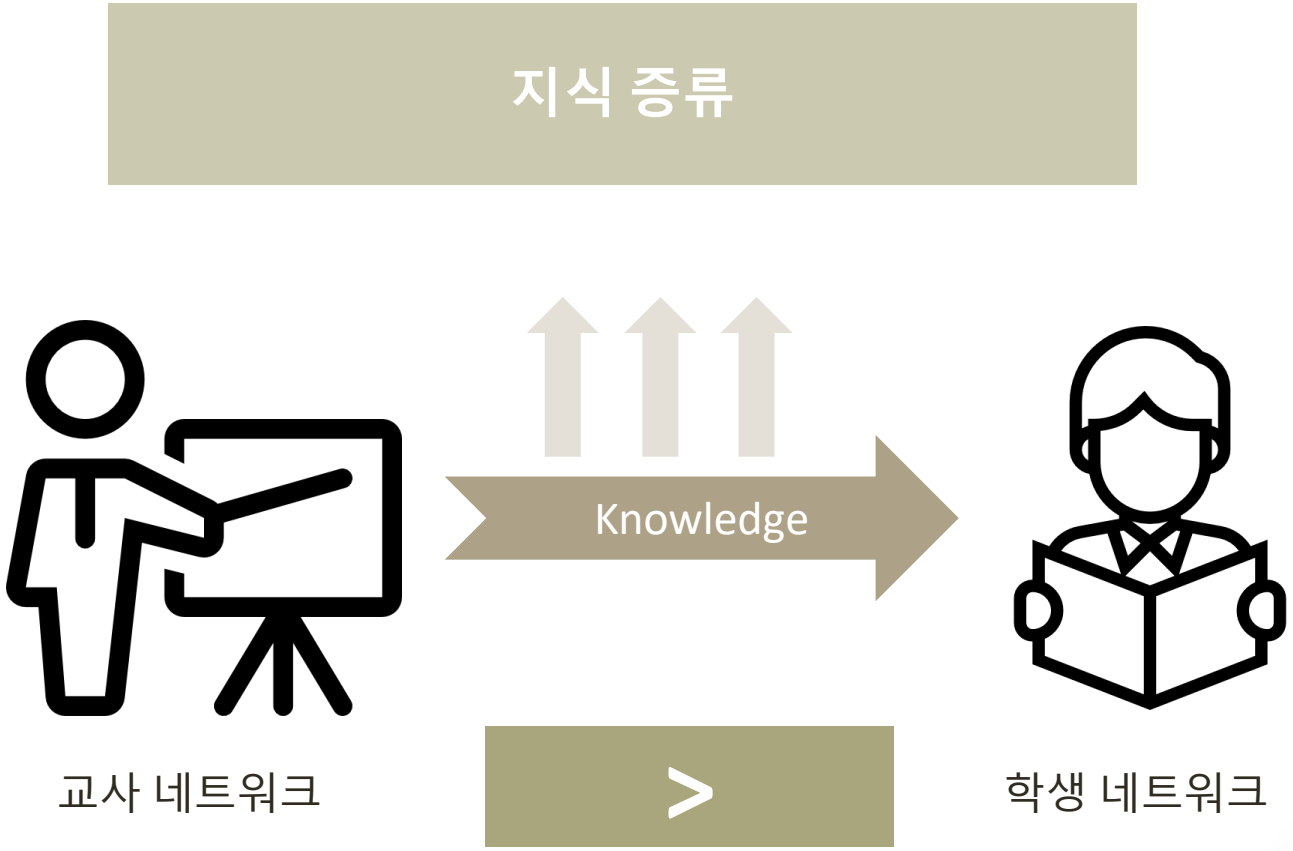


분류기 네트워크

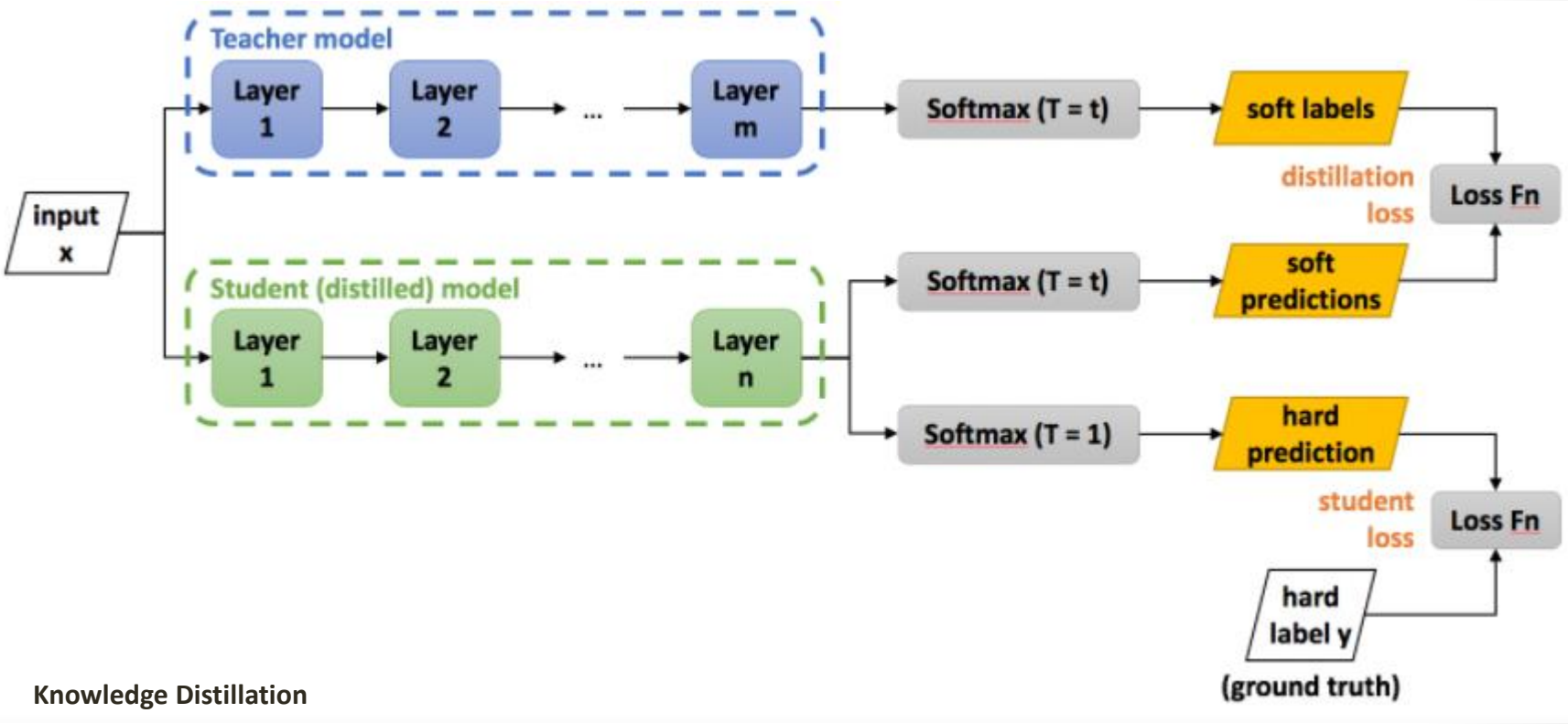
- 자체 지식 증류를 위해 소프트 라벨과 형상도 증류 모두 활용 가능
- 지역 정보의 보존을 강조하는 분류 및 의미 분할에 적용 가능
- 다양한 곳에서 입증된 효과
- 공개된 코드 (https://github.com/MingiJi/FRSKD)

1. Introduction - 소개

- 장치에서의 제한된 자원 ▶ 모델의 압축이 중요



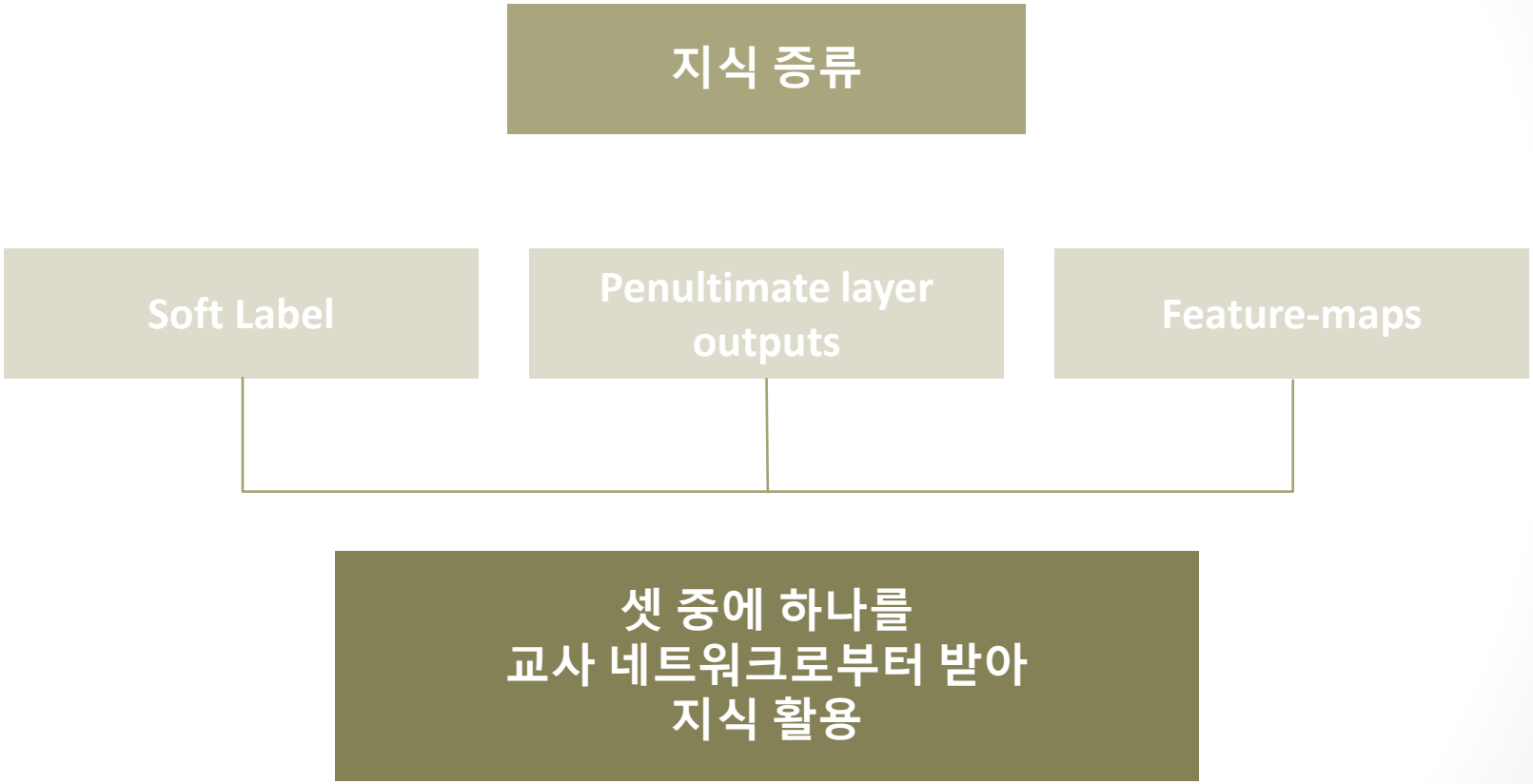
1. Introduction -소개



Knowledge Distillation

https://intellabs.github.io/distiller/knowledge_distillation.html

1. Introduction - 소개



1. Introduction - 소개

- Soft Label : Task는 신경망의 마지막 softmax 레이어를 통해 각 클래스의 확률값을 낸다.
  - i번째 클래스에 대한 확률값(q_i)

$$q_i = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

Soften

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

여기서 T는 온도(Temperature)

T 증가 → Soft

T 감소 → Hard

1. Introduction - 소개

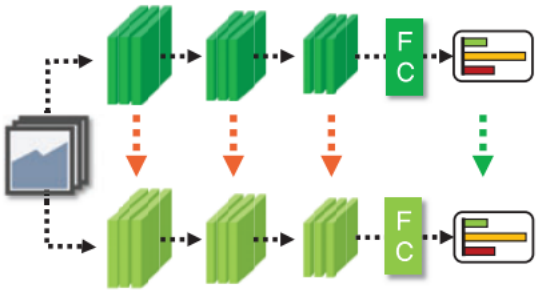
손실함수(L) 을 통해 학습시킴

$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x,\theta_S,\tau), T(x,\theta_T,\tau)) + \lambda L_{CE}(\hat{y}_S, y)$$

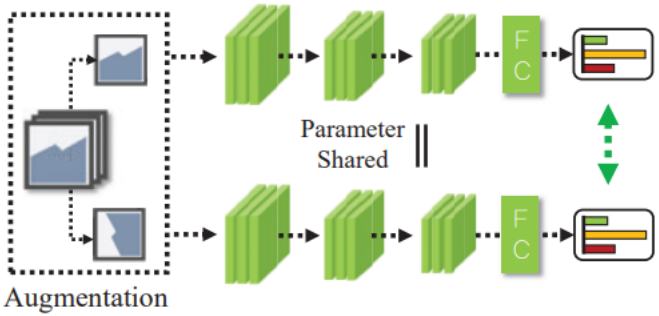
L = 손실함수
S = Student model
T = Teacher model
(x,y) = 하나의 이미지와 그 레이블
Θ = 모델의 학습 파라미터
τ = temperature

https://baeseongsu.github.io/posts/knowledge-distillation/

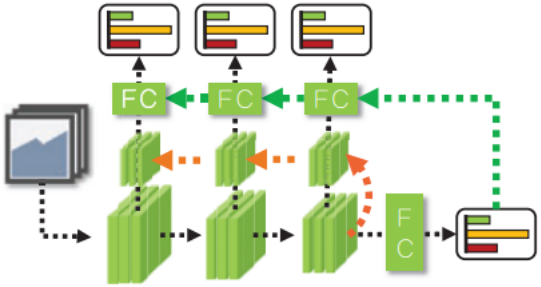
1. Introduction - 소개



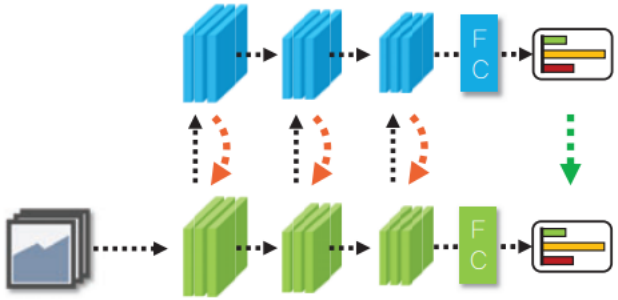
(a) Knowledge Distillation



(b) Self-Knowledge Distillation via Data-augmentation

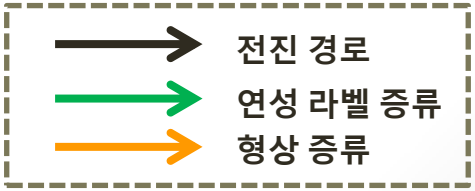


(c) Self-Knowledge Distillation via Auxiliary Classifiers



(d) Feature Refinement via Self-Knowledge Distillation

[다양한 증류 방법 비교]



1. Introduction - 소개

2. Related Work – 관련 작업

- Knowledge Distillation(지식 증류)
  - 목표 : 사전 훈련된 복잡한 네트워크(교사 네트워크)에 대한 지식을 전달하여 더 단순한 네트워크(학생 네트워크)를 효과적으로 훈련시키는 것

  - 한계점
    - 1) 지식 증류는 복잡한 교사 모델의 사전 훈련을 필요로 함
    - 2) 교사 네트워크가 변하면 동일한 학생 네트워크라도 다른 성과를 냄

2. Related Work – 관련 작업

- Self-Knowledge Distillation(자가지식 증류)
  - : 교사 네트워크 없이 자신의 지식을 활용하여 학생 네트워크 훈련의 효과를 향상시킨다.

  - 1) 일부는 자가지식 증류를 위해 보조 네트워크 활용
    - Ex) BYOT의 분류기 네트워크 도입
      : 추정값과 실제 감시에 대한 공동 감독으로 훈련
    - Ex) ONE의 추가 분기 활용
      : 모델 매개변수와 중간 계층의 추정 특징을 다양화

  - 2) 데이터 확대 사용
    - Ex) DDGSD : 다르게 증강된 instance 제공 > 일관된 예측 유도
    - Ex) CSKD : 정규화 목적으로 동일한 클래스에 속하는 다른 instance들의 Logit 사용 > 동일한 클래스에 대해 유사한 결과 예측

제안 : 단일 Instance에서 정교한 기능 맵을 생성하는 셀프 교사 네트워크 작업

2. Related Work – 관련 작업

- Feature Networks

논문에서 제안한
보조 셀프 교사 네트워크

- 객체 감지 분야에서 사용되는 Feature Network 에서 개발
- 지식 증류 목적에 다중 스케일 기능을 처리하는 네트워크를 조정
  ▶ 정교한 특징 맵 생성
- 분류 작업에 적합하도록 BiFPN구조에서 변경된 보조 셀프 교사 네트워크

BiFPN
하향식 및 상향식 네트워크 사용
▶ 효율적 인 네트워크 구조

3. Method – 방법

- Self-Teacher Network
  - 주요 목적 : 분류기 네트워크를 위한 정교한 기능 맵과 소프트 레이블 제공

BiFPN 구조 수정

횡방향 컨볼루션 레이어

$$L_i = Conv(F_i; d_i)$$

- Con$v$는 출력 치수가 d_i인 convolution operation

하향식경로 & 상향식 경로

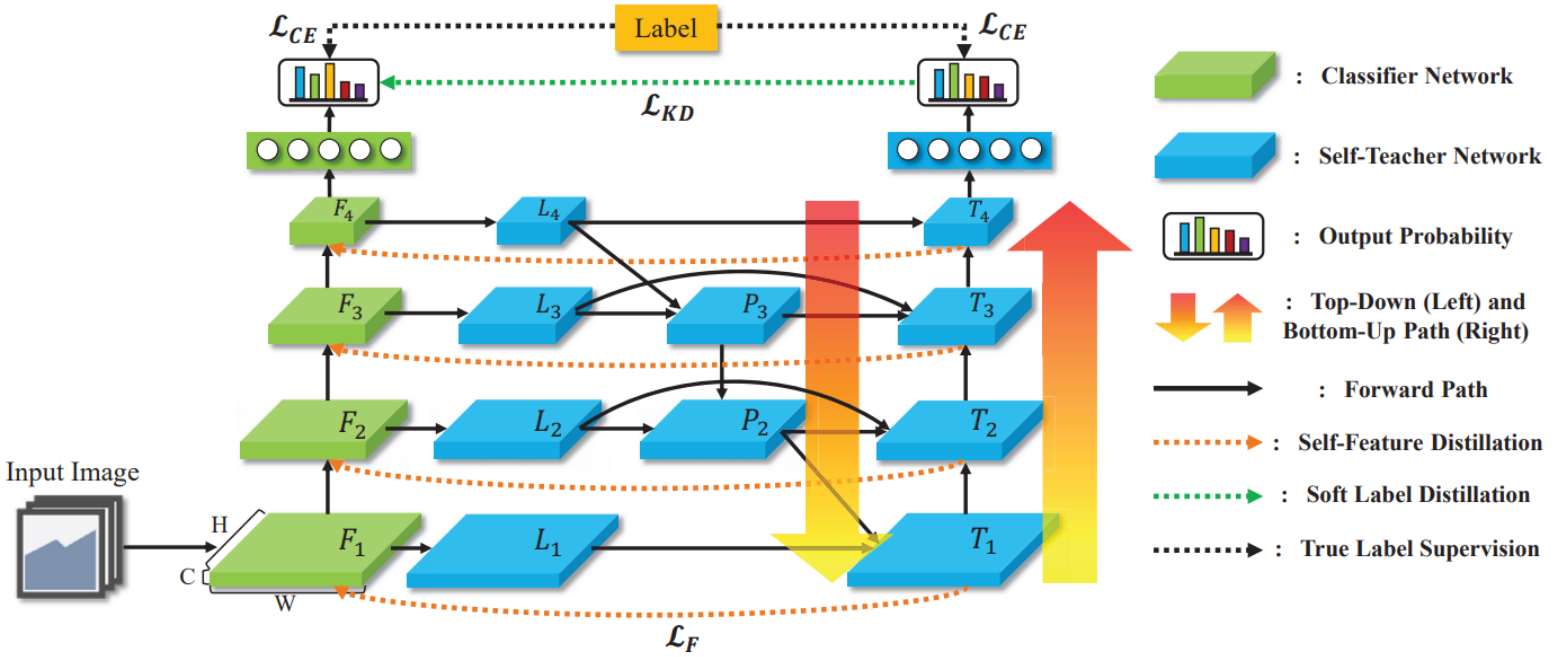
$$P_i = Conv(w_{i,1}^P \cdot L_i + w_{i,2}^P \cdot Resize(P_{i+1}); d_i)$$

$$T_i = Conv(w_{i,1}^T \cdot L_i + w_{i,2}^T \cdot P_i + w_{i,3}^T \cdot Resize(T_{i-1}); d_i)$$

- $\mathcal{P}_i$는 하향식 경로의 i번째 층
- $\mathcal{T}_i$는 상향식 경로의 i번째 층

3. Method – 방법

Feature Refinement
Self-Knowledge Distillation
(FRSKD)



[FRSKD의 개요]

3. Method – 방법

- Self-Feature Distillation : 자기 형상 증류

$$\mathcal{L}_F(T, F; \theta_c, \theta_t) = \Sigma_{i=1}^{n} ||\phi(T_i) - \phi(F_i)||_2$$

- $\mathcal{L}_{\mathcal{F}}$ = 형상 증류 손실
- ϕ = combination of channel-wise pooling function with $\mathcal{L}_2$ normalization
- θ_c = 분류기 네트워크의 parameter

$$\mathcal{L}_{KD}(\boldsymbol{x}; \theta_c, \theta_t, K) = D_{KL}(\text{softmax}(\frac{f_c(\boldsymbol{x}; \theta_c)}{K}) || \text{softmax}(\frac{f_t(\boldsymbol{x}; \theta_t)}{K})$$

- f_c = 분류기 네트워크
- $\mathcal{K}$ = 온도 스케일링 매개 변수

손실함수
: 통합하여 최적화

3. Method – 방법

최적화된 목표 함수

$$\begin{aligned}&\mathcal{L}_{FRSKD}(\boldsymbol{x}, y; \theta_c, \theta_t, K)\\&= \mathcal{L}_{CE}(\boldsymbol{x}, y; \theta_c) + \mathcal{L}_{CE}(\boldsymbol{x}, y; \theta_t)\\&\quad + \alpha \cdot \mathcal{L}_{KD}(\boldsymbol{x}; \theta_c, \theta_t, K) + \beta \cdot \mathcal{L}_{F}(T, F; \theta_c, \theta_t)\end{aligned}$$

- α와 β = 초모수

4. Experiments

소프트 라벨 전용 증류(FRSKD\F) 활용

정제된 Feature map 및 소프트 라벨 증류(FRSKD)를 이용한 LF RSKD 최적화

자체 지식 증류법 SLASD+를 이용한 데이터 확대 첨부

4. Experiments

- 6가지 자체 증류 방법 : 7개의 기준선 생성
  - **ONE[43]**은 소프트 라벨로 추가 분지에 대한 앙상블 예측을 이용한다.
  - **DDGSD[32]**는 단일 인스턴스의 서로 다른 왜곡 버전을 생성하고 DDGSD 트레인은 왜곡 데이터에 대한 일관된 예측을 산출한다.
  - **BYOT[39]**는 중간 레이어의 출력을 활용하는 보조 분류기를 적용하고, BYOT는 예측 로짓 또는 형상 지도와 같은 네트워크 자체로부터의 신호와 지상 실측 라벨에 의해 보조 분류기를 훈련시킨다.
  - **SAD [10]**은 네트워크 자체에서 레이어별 주의 증류에 의한 차선 감지에 초점을 맞춘다.
  - **CS-KD[35]**는 소프트 라벨과 동일한 등급 내의 다른 인스턴스 예측을 활용하여 동일한 등급에 대해 일관된 예측을 강제한다.
  - **SLA-SD [18]**는 라벨 증대를 활용하여 원래의 분류 작업과 자체 감독 작업을 공동으로 수행하는 네트워크를 훈련시킨다. SLA-SD는 집계된 예측을 소프트 레이블로 활용합니다.

4. Experiments

Methods	CIFAR100		TinyImageNet	
	WRN-16-2	ResNet18	WRN-16-2	ResNet18
Baseline	70.42±0.08	73.80±0.60	51.05±0.20	54.60±0.33
ONE	73.01±0.23	76.67±0.66	52.10±0.20	57.53±0.39
DDGSD	71.96±0.05	76.61±0.47	51.07±0.24	56.46±0.24
BYOT	70.22±0.26	76.68±0.07	50.33±0.03	56.61±0.30
SAD	70.31±0.45	74.65±0.33	51.26±0.39	54.45±0.06
CS-KD	71.79±0.68	77.19±0.05	50.08±0.18	56.46±0.10
SLA-SD	73.00±0.45	77.52±0.30	50.77±0.33	58.48±0.44
FRSKD\F	73.12±0.06	77.64±0.12	52.91±0.30	59.50±0.15
FRSKD	73.27±0.45	77.71±0.14	**53.08±0.33**	59.61±0.31
FRSKD+SLA	**75.43±0.21**	**82.04±0.16**	51.83±0.37	**63.58±0.04**

Table 1 : CIFAR-100과 TinyImageNet의 성능 비교

실험은 세 번 반복 / 마지막 실험의 정확도에 대한 평균과 표준편차 보고
성능이 가장 좋은 모델은 **굵은체** / 차선 모델은 밑줄

4. Experiments

Methods	CUB200	MIT67	Dogs	Stanford40
Baseline	51.72±1.17	55.00±0.97	63.38±0.04	42.97±0.66
ONE	54.71±0.42	56.77±0.76	65.39±0.59	45.35±0.53
DDGSD	58.49±0.55	59.00±0.77	69.00±0.28	45.81±1.79
BYOT	58.66±0.51	58.41±0.71	68.82±0.15	48.51±1.02
SAD	52.76±0.57	54.48±1.30	63.17±0.56	43.52±0.06
CS-KD	64.34±0.08	57.36±0.37	68.91±0.40	47.23±0.22
SLA-SD	56.17±0.71	61.57±1.06	67.30±0.21	54.07±0.38
FRSKD\F	62.29±1.65	61.32±0.67	69.48±0.84	53.16±0.44
FRSKD	65.39±0.13	61.74±0.67	70.77±0.20	56.00±1.19
FRSKD+SLA	**67.80±1.24**	**66.04±0.31**	**72.48±0.34**	**61.96±0.57**

Table 2 : FGVR에 대한 성능 비교

실험은 세 번 반복 / 마지막 실험의 정확도에 대한 평균과 표준편차 보고
성능이 가장 좋은 모델은 **굵은체** / 차선 모델은 밑줄

4. Experiments

Model	Method	Top-1	Top-5
ResNet18	Baseline	69.76	89.08
	FRSKD	**70.17**	**89.78**
ResNet34	Baseline	73.31	91.42
	FRSKD	**73.75**	**92.11**

Table 3 : ImageNet의 성능 비교

성능이 가장 좋은 모델은 **굵은체**

4. Experiments

- 의미론적인 부분

Model	Method	mIOU
EfficientDet-d0	Baseline	79.07
	FRSKD	**80.55**
EfficientDet-d1	Baseline	81.95
	FRSKD	**83.88**

Table 4 : 의미론적 세분화 작업에 대한 성능 비교

성능이 가장 좋은 모델은 **굵은체**

FRSKD가 자체 교사 네트워크의 자가 지식 증류를 활용하여 모델의 성능을 향상시키고 있음을 보여줌

4. Experiments – 추가 분석

Method	CIFAR-100	TinyImageNet	CUB200	MIT67	Dogs	Stanford40
Baseline	73.80±0.60	54.60±0.33	51.72±1.17	55.00±0.97	63.38±0.04	42.97±0.66
Fit+SKD	77.03±0.05	59.06±0.20	<u>61.05±1.05</u>	<u>57.69±0.28</u>	<u>67.50±0.32</u>	<u>51.66±1.32</u>
OD+SKD	<u>77.12±0.09</u>	<u>59.14±0.20</u>	57.44±0.92	54.83±2.63	66.51±0.87	49.09±0.47
FRSKD	**77.71±0.14**	**59.61±0.31**	**65.39±0.13**	**61.74±0.67**	**70.77±0.20**	**56.00±1.19**

Table 5 : FRSKD 형상 증류법에 따른 성능 비교

Fit+SKD의 형상 증류법은 FitNet
OD+SKD는 오버홀 증류법
FRSKD는 주의력 전달법

ResNet18은 분류기 네트워크

성능이 가장 좋은 모델은 **굵은체** / 차선 모델은 <u>밑줄</u>

4. Experiments – 추가 분석

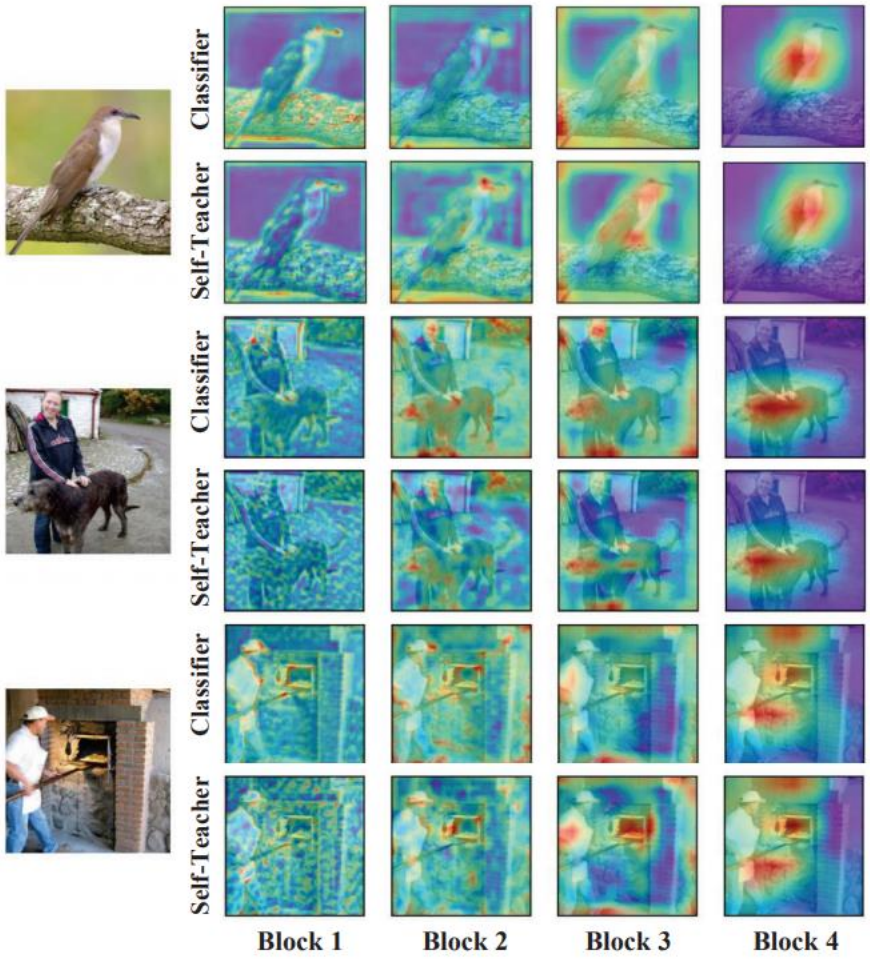


Figure 3 : 분류기 네트워크와 셀프 교사 네트워크 간의 블록별 주의 맵 비교

4. Experiments – 추가 분석

Type	#channel	Parameters	FLOPs	CIFAR-100
BiFPN	128	× 0.30	× 0.67	72.64±0.12
BiFPN	256	× 0.97	× 2.38	73.54±0.41
BiFPNc	128	× 0.19	× 0.21	71.70±0.19
BiFPNc	256	× 0.59	× 0.68	73.27±0.45

Table 6 : 셀프 교사 네트워크 구조 간의 성능 및 효율성 비교

WRN-16-2는 CIFAR-100의 분류기 네트워크

FRSKD는 분류기 네트워크를 중복하여 사용하는
데이터 증강 기반 자체 지식 증류 방법보다 더 효율적이다.

4. Experiments – 추가 분석

Method	CIFAR-100	TinyImageNet	CUB200	MIT67	Dogs	Stanford40
Baseline	73.80±0.60	54.60±0.33	51.72±1.17	55.00±0.97	63.38±0.04	42.97±0.66
FitNet	76.65±0.25	59.38±0.10	58.97±0.07	59.15±0.41	67.18±0.10	46.64±0.24
ATT	77.16±0.15	**59.83±0.28**	59.21±0.34	59.33±0.22	67.54±0.18	47.04±0.17
Overhaul	74.59±0.32	59.50±0.09	58.82±0.12	58.81±0.58	66.43±0.08	47.06±0.26
FRSKD	**77.71±0.14**	59.61±0.31	**65.39±0.13**	**61.74±0.67**	**70.77±0.20**	**56.00±1.19**

Table 7 : 지식 증류에 대한 성능 비교

ResNet18은 분류기 네트워크

성능이 가장 좋은 모델은 **굵은체** / 차선 모델은 밑줄

FRSKD가 대부분의 데이터 세트 에서
사전 교육을 받은 교사 네트워크를 사용하여 실험된 지식 증류 방법을 능가한다.

4. Experiments – 추가 분석

Method	CIFAR-100	TinyImageNet	CUB200	MIT67	Dogs	Stanford40
Baseline	73.80±0.60	54.60±0.33	51.72±1.17	55.00±0.97	66.38±0.04	42.97±0.66
Mixup	76.26±0.41	56.28±0.24	57.60±0.42	56.77±1.45	65.96±0.03	47.15±0.60
FRSKD + Mixup	78.74±0.19	60.30±0.38	**67.98±0.58**	62.11±0.81	71.64±0.29	**56.50±0.36**
CutMix	79.23±0.23	58.97±0.29	51.54±1.12	60.87±0.30	67.71±0.14	46.90±0.29
FRSKD + CutMix	**80.49±0.05**	**61.92±0.11**	65.92±0.59	**66.19±0.49**	**72.81±0.23**	55.75±0.43

Table 8 : FRSKD를 이용한 데이터 확대 방법의 성능

ResNet18은 분류기 네트워크

성능이 가장 좋은 모델은 **굵은체** / 차선 모델은 밑줄

FRSKD가 데이터 증대와 함께 사용될 때 성능이 크게 향상되었다.

5. Conclusion

- 하향식 및 상향식 경로를 가진 자가 지식 증류를 위한 특수 신경망 구조를 제시.
  - 이러한 경로를 추가하면, 분류기 네트워크에 정교한 형상 맵과 그 소프트 레이블을 제공할 것으로 예상.

- FRSKD는 분류 및 의미 분할의 비전 작업에 자가 지식 증류를 적용할 수 있다.

- 성능은 정량적으로 확인되었으며, 다양한 절제 연구를 통해 작업 매커니즘의 효율성을 보여준다.

THANK YOU!