Knowledge Distillation



Presented By
강정우

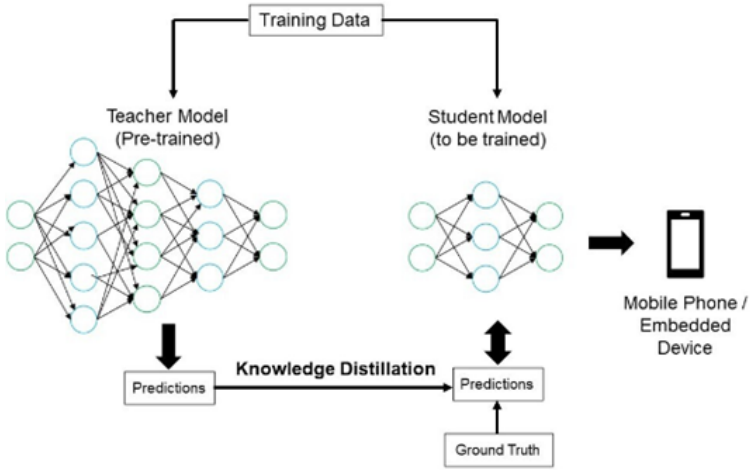
Outline

- Knowledge Distillation
- LG전자 과제
- Rethinking Feature-based Knowledge Distillation for Face Recognition
- 느낀점 및 추후계획

Knowledge Distillation

- Knowledge Distillation이란?
  - 우리말로 지식 증류로 큰 네트워크의 지식을 작은 네트워크에 전달하는 것
    - 큰 네트워크를 Teacher Network
    - 작은 네트워크를 Student Network
      - 지식 증류가 잘 된 작은 모델은 큰 모델의 성능과 비슷함
  - 작고 경량화 된 모델로 로봇, 휴대폰과 같은 디바이스에서 사용 가능
    - 메모리와 저장 공간을 효율적으로 사용



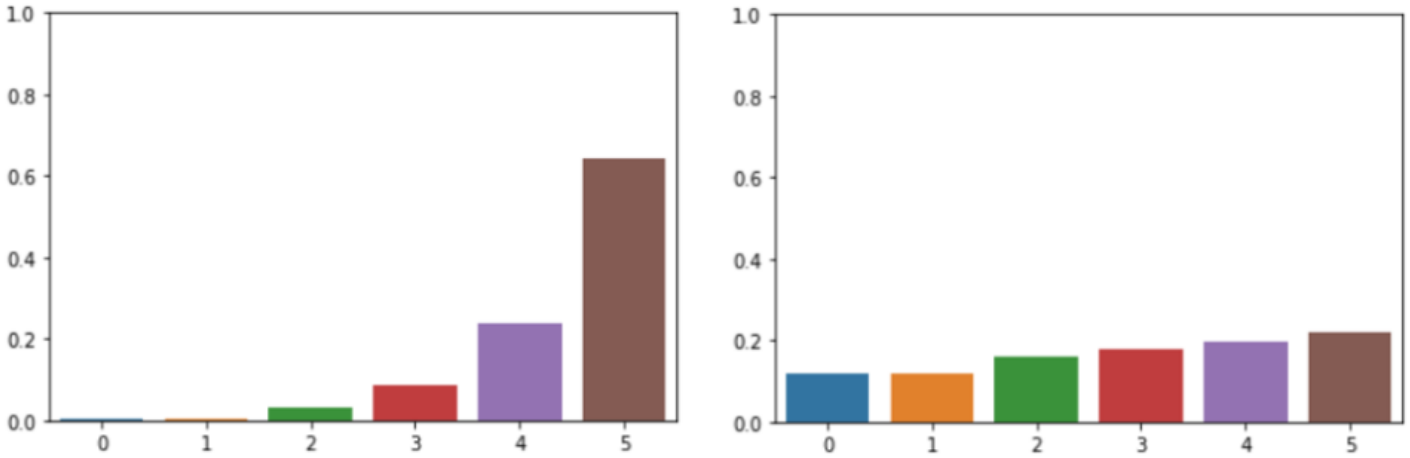


Knowledge Distillation

- Soft Label
  - 신경망의 마지막 Layer는 Softmax 함수를 통해 각 클래스의 확률값 출력
    - Knowledge Distillation은 Softmax 함수의 출력 값들을 모델의 지식으로 이용
      - T(온도)라는 설정을 통해 출력값의 분포를 Soft하게 만들 수 있음

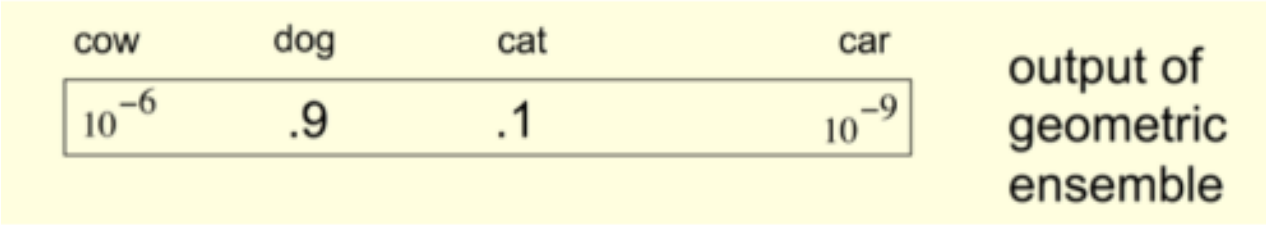
$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

  - Temperature에 따른 Softmax 변화
    - T가 높아지면 출력값의 분포를 Soft하게 만들 수 있음



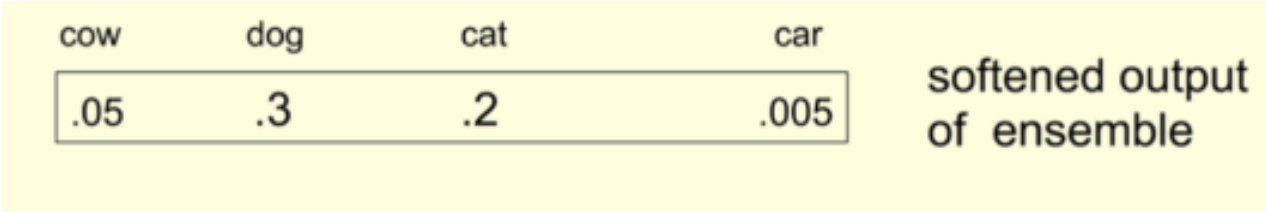
Knowledge Distillation

- Soft Label
  - Soft하게 된 출력값들이 모델이 가진 지식이 됨
    - 기존 Softmax를 통해 출력 된 값들은 너무 작아 모델에 반영 어려움
      - ∴ 정보가 많이 없음



기존 Softmax 출력값

    - Soft하게 된 출력 된 값들은 학습에 더 잘 반영할 수 있음
      - ∴ 더 많은 정보 전달



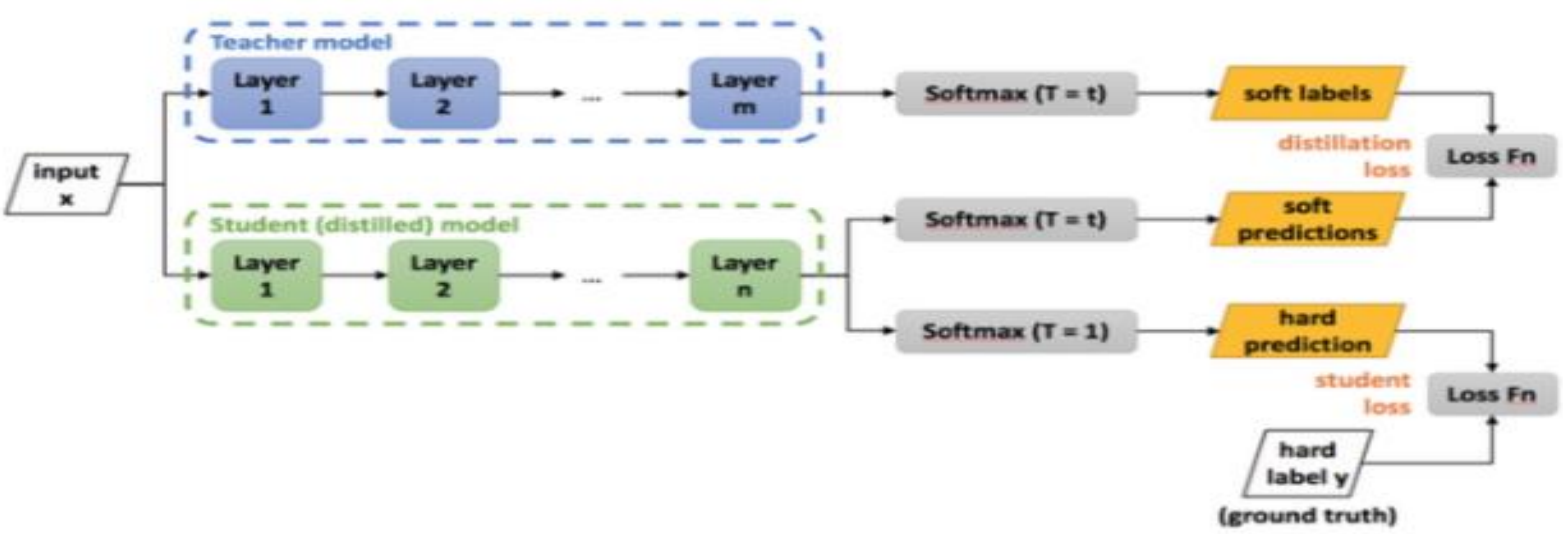
Soft된 Softmax 출력값

Knowledge Distillation

- Distillation loss
  - 작은 모델이 큰 모델의 출력과 비교하여 학습하는 손실 함수
    - L은 손실함수, S는 student model, T는 teacher model을 의미
    - (x, y)는 하나의 이미지와 라벨, θ는 모델의 학습 파라미터, τ는 temperature를 의미

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

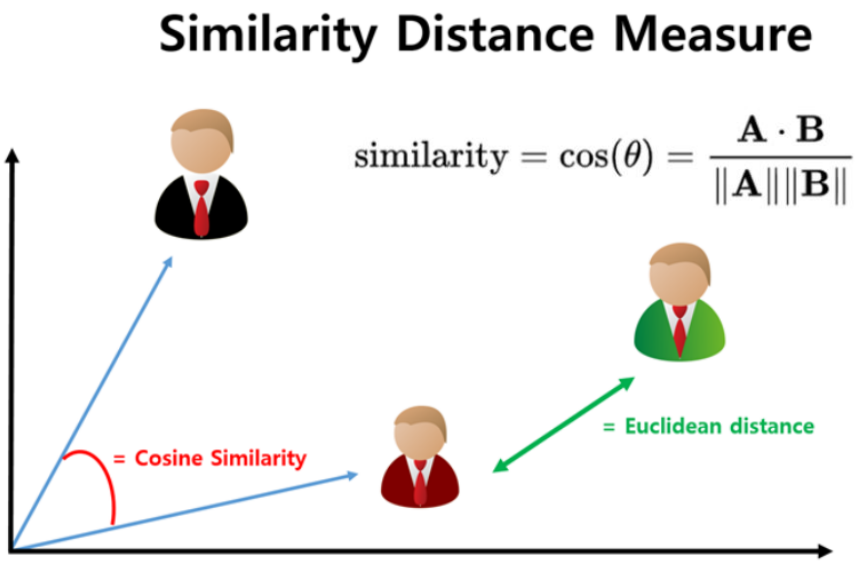
  - Distillation loss와 student loss를 더하여 Total loss값을 최소화 하는 방향으로 학습 진행



LG전자 과제

- Teacher model에서 정보를 얻는 방식
  - Response-Based Knowledge: 최종만 빼내는 경우
  - Feature-Based Knowledge: 중간 Layer의 결과를 빼내어 student에게 넘겨주는 경우

- LG전자 과제
  - 과제명: 홈로봇 환경을 고려한 야간 사람 인식 기술 개발 및 주간 얼굴 인식 성능 개선
    - 주간 얼굴인식 기술 개발 및 성능개선
      - 두 모델이 출력하는 Feature를 비교하여 얼굴인식 진행중



LG전자 과제

- Feature Only Distillation
  - 얼굴의 특성을 나타내는 임베딩 간의 거리를 최소화하기 위해 MSE 손실 사용

$$L_{emb} = L_{feat}(\boldsymbol{f}_s, \boldsymbol{f}_t) = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{\boldsymbol{f}_s^i}{\|\boldsymbol{f}_s^i\|_2} - \frac{\boldsymbol{f}_t^i}{\|\boldsymbol{f}_t^i\|_2} \right\|_2^2$$

  - f_s와 f_t는 각각 학생 벡터와 선생 벡터
  - N은 Batch size

- Feature Only Distillation Code

```
def distillation_loss(student_output, teacher_output, teacher_norm, student_norm):

    distill_mse = F.mse_loss(student_output/student_norm, teacher_output/teacher_norm)

    return distill_mse
```

  - student_output, student_norm은 학생 모델 feature와 크기
  - teacher_output, teacher_norm은 선생 모델 feature와 크기



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethinking Feature-based Knowledge Distillation for Face Recognition

Jingzhi Li[†,*,1], Zidong Guo[*,1], Hui Li[1], Seungju Han[2], Ji-won Baek[2], Min Yang[1], Ran Yang[1], Sungjoo Suh[2]

[1]Samsung R&D Institute China Xi'an (SRCX)

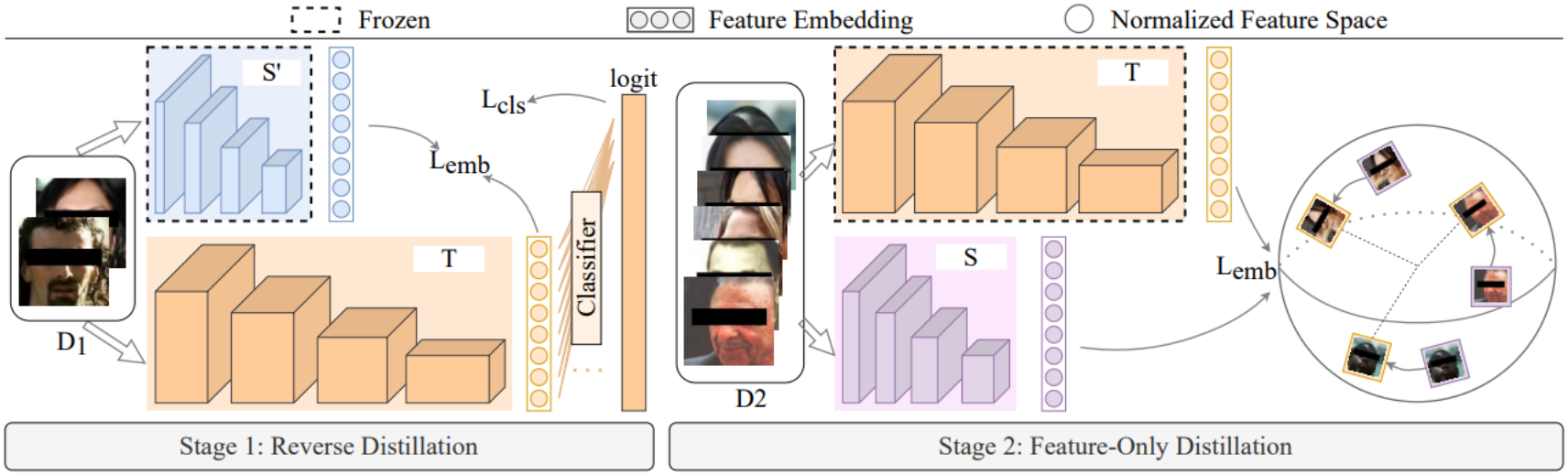
[2]Samsung Advanced Institute of Technology (SAIT), South Korea

jingzhi.li, zidong.guo, hui01.li, sj75.han, jw0328.baek, min16.yang, ran01.yang, sungjoo.suh@samsung.com

- 용어 정의
  - Intrinsic Gap
    - 선생 모델과 학생 모델의 Gap
  - Intrinsic Dimension
    - 실질적으로 쓰이는 Dimension
  - ID Supervison
    - ID label, Face label
  - Capacity Gap
    - 모델 네트워크끼리의 Gap

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Reverse distillation empowered Feature-Only (ReFO) 모델 구조



- Reverse distillation empowered Feature-Only (ReFO) 알고리즘

Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

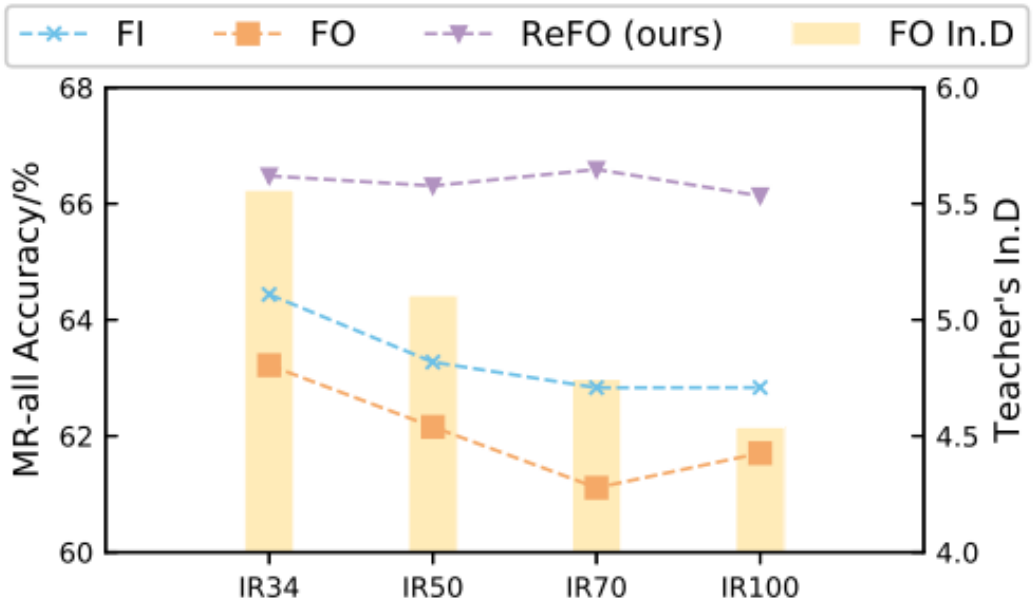
Reverse Distillation

Feature-Only Distillation

CVSP

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 1: Reverse Distillation
  - 일반적인 Knowledge Distillation은 선생에서 학생으로 지식 증류
  - Reverse Distillation은 학생이 선생에게 지식 증류
    - Reverse Distillation을 통해 선생의 Intrinsic dimension을 학생에게 맞춤
      - 선생과 학생의 Intrinsic Gap이 좁아짐

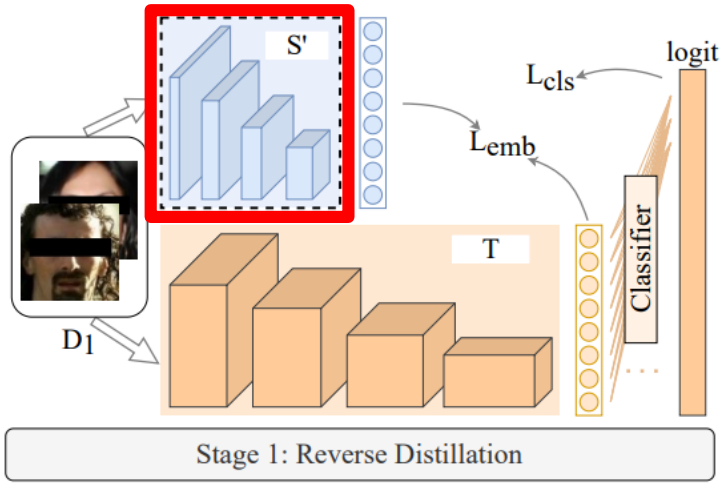


Resnet 18(S) is distilled by four different teacher

- ✓ **FI: Feature based with ID supervision**
- ✓ **FO: Feature Only Distillation**
- ✓ **ReFO: Our Proposed Method**
- ✓ **In.D: Intrinsic Dimension**

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 1: Reverse Distillation - ReFO 알고리즘1



Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

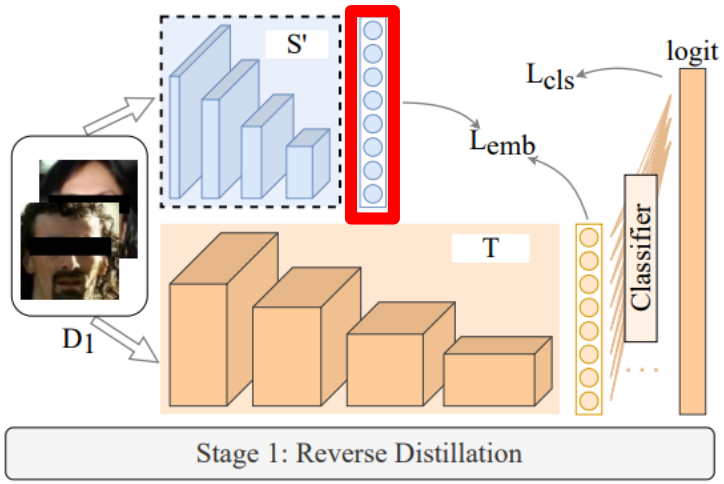
5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

- S' 모델을 데이터셋 D1 사용하여 학습하고, classification loss 구함
  - S'은 initial student

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 1: Reverse Distillation - ReFO 알고리즘2



Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

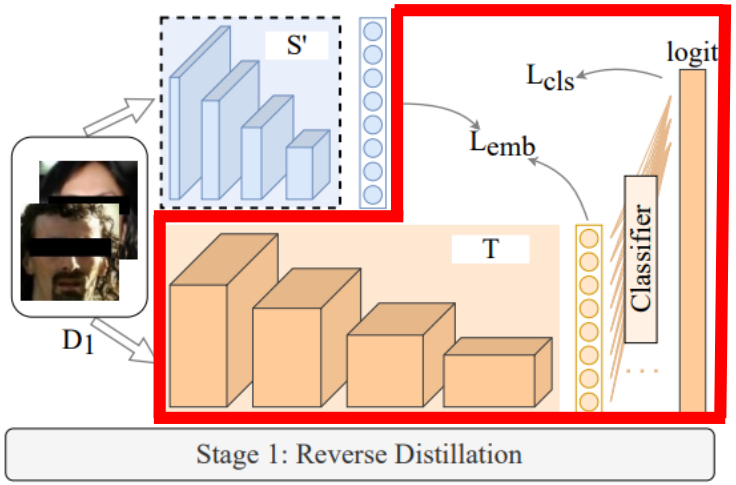
5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

  - 데이터셋 D1으로 훈련된 S' 모델에서 Feature 추출

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 1: Reverse Distillation - ReFO 알고리즘3



Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

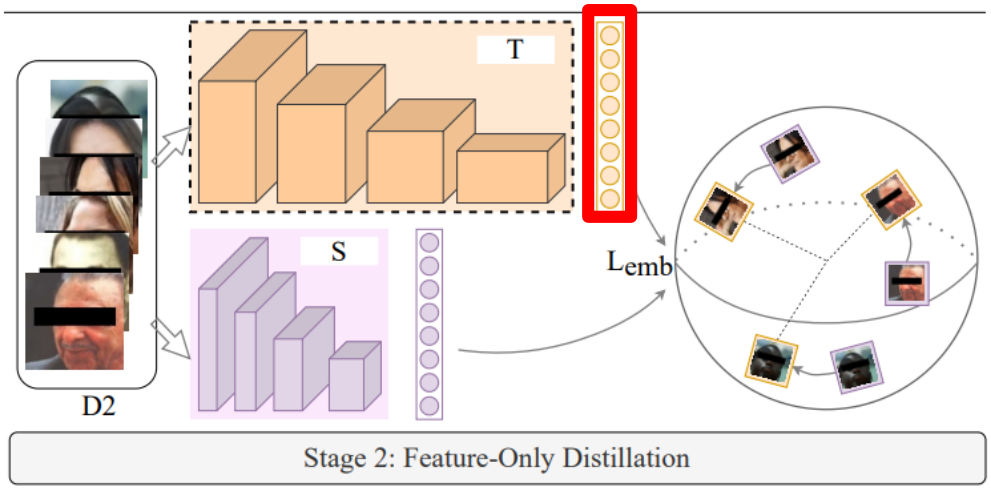
5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

- T모델을 데이터셋 D1를 사용하여 학습하고, classification loss 구함
- Step2에서 구한 feature와 feature only distillation 진행

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 2: Feature-Only Distillation - ReFO 알고리즘4



Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

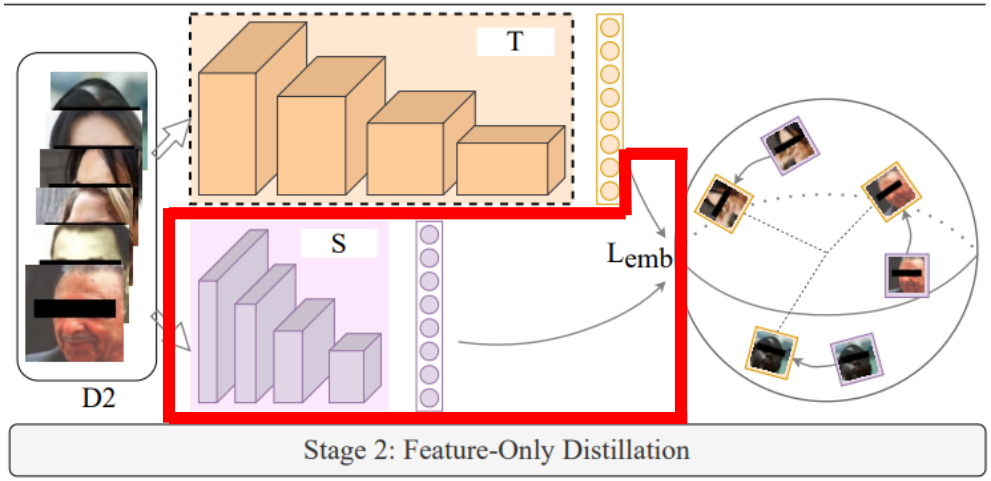
5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

- 데이터셋 D2로 훈련된 모델 T에서 Feature 추출

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 2: Feature-Only Distillation - ReFO 알고리즘5



Algorithm 1 ***ReFO*** Knowledge Distillation

1: Train a student model S' on dataset D_1 with standard classification loss for face recognition.

$$L = L_{cls}.$$

2: Obtain features of model S' on dataset D_1.

3: Train tailored teacher model T on dataset D_1 with classification loss and embedding distillation loss using features from step 2.

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}).$$

4: Obtain features of model T on dataset D_2.

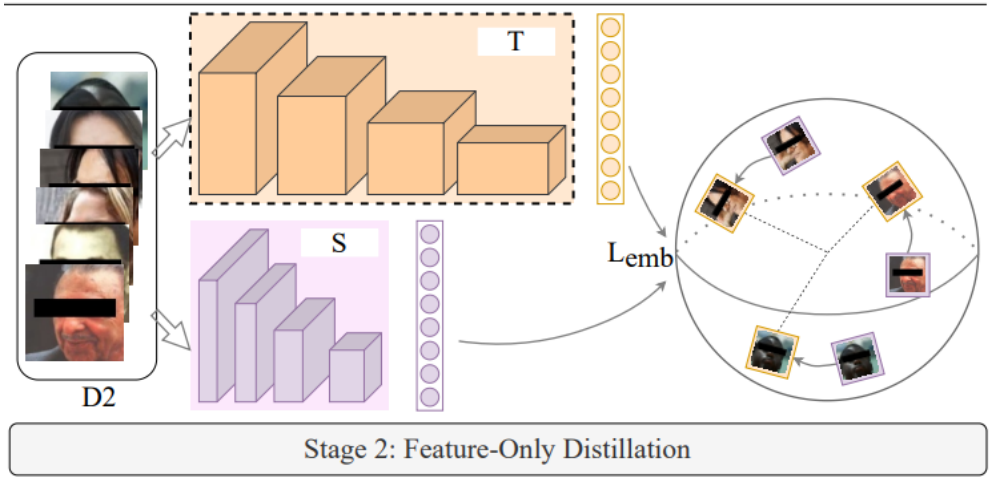
5: Train final student model S on dataset D_2 with embedding distillation loss using features from step 4.

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t).$$

- S모델을 데이터셋 D2를 사용하여 학습
- Step4에서 구한 feature와 feature only distillation 진행

Rethinking Feature-based Knowledge Distillation for Face Recognition

- Stage 2: Feature-Only Distillation



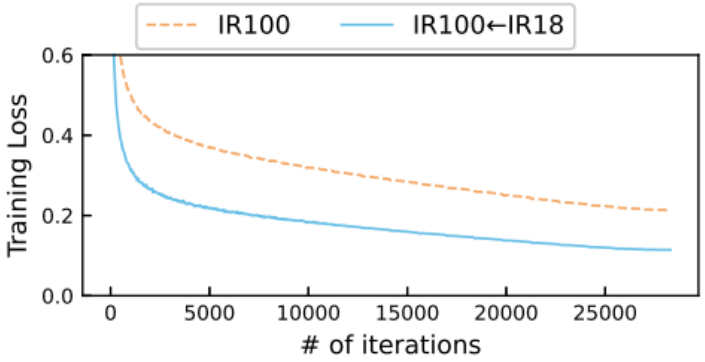
  - Classifier가 없는 이유
    - ID Supervision 없이 Feature Distillation을 진행
      - Unlabeled dataset 접근 가능

Rethinking Feature-based Knowledge Distillation for Face Recognition

- 실험 결과
  - Intrinsic Gap을 좁게하는 ReFO의 성능이 더 좋음

Student	Intrinsic Gap		MR-all/%	
	FO	ReFO	FO	ReFO
MFN	4.10	3.75	53.86	57.27
MNv2	3.53	3.08	58.33	63.04
IR18	2.25	2.03	61.70	66.13
IR34	1.02	0.97	73.17	75.07

  - IR100에 비해 IR100 ← IR18이 더 빠르게 Train Loss 감소
    - IR100 ← IR18: IR100 선생에 Distillation 된 모델
    - IR100: Original KD



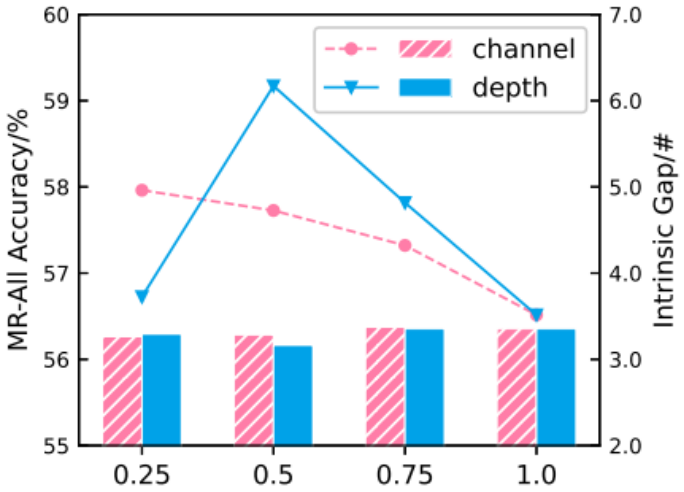
(a) IResNet18

Rethinking Feature-based Knowledge Distillation for Face Recognition

- 실험 결과
  - 기존의 선생보다 Reverse Distillation 된 선생의 성능이 더 좋음
    - A ← B: A는 B에 의해 Reverse Distialltion

Teacher	Student			
	MFN	MNv2	IR18	IR34
IR100	53.86	58.32	61.70	73.16
IR100←MFN	57.27	63.22	66.58	76.18
IR100←IR18	56.53	63.57	66.13	76.01

  - 모델의 Depth가 제일 작다고해서 좋은 성능을 내는 것은 아님
    - S_d가 0.25일때 오히려 성능이 감소



Rethinking Feature-based Knowledge Distillation for Face Recognition

- 실험 결과
  - 이전에 제시된 SOTA Method들과의 비교

Method	L_{cls}	LFW	CFP-FP	AgeDB	IJB-C*		MegaFace		ICCV21-MFR*		
					$1e-4$	$1e-5$	Id(R)	Ver(R)	MR-all	Children	Mask
IR50 (teacher)	✓	99.80	97.63	97.92	96.05	93.96	98.14	98.34	75.48	49.41	54.50
MFN (student)	✓	99.52	91.66	95.82	92.16	85.83	90.91	92.71	53.43	24.71	27.90
KD ('15) [17]	✓	99.50	91.71	95.93	86.96	69.98	90.40	92.00	50.77	26.36	25.74
FitNet ('15) [38]	✓	99.47	91.30	96.18	91.73	86.07	91.16	92.34	54.46	26.62	28.47
DarkRank ('18) [6]	✓	99.55	91.84	95.60	92.15	86.28	90.76	92.41	56.82	28.84	30.07
SP ('19) [41]	✓	99.53	92.33	96.17	91.79	87.22	91.25	92.41	54.44	26.63	29.75
CCKD ('19) [35]	✓	99.47	91.90	95.83	91.73	85.75	91.17	92.76	55.64	27.65	30.22
RKD ('19) [33]	✓	99.58	92.13	96.18	89.36	81.88	91.44	92.92	53.92	27.91	27.94
ShrinkTeaNet ('19) [11]	✓	99.47	91.97	96.00	91.50	86.23	90.73	92.32	55.28	27.73	30.24
MarginKD ('21) [8]	✓	*99.61*	92.01	*96.55*	91.02	83.39	91.70	92.96	50.73	25.14	28.54
EKD ('22) [20]	✓	99.60	94.33	96.48	92.28	86.47	91.02	93.08	56.60	28.95	*32.14*
SFTN* ('21) [32]	✓	99.48	92.77	96.30	90.96	82.67	91.69	93.38	55.50	28.51	29.66
SH-KD* ('22) [3]	✓	99.47	*94.67*	96.53	91.75	85.76	**92.51**	**93.93**	*57.69*	30.15	32.01
ReFO (ours)	×	99.55	94.51	**96.92**	*92.23*	*87.55*	92.38	*93.80*	56.63	**33.36**	31.88
ReFO+ (ours)	×	**99.65**	**94.77**	96.42	**92.41**	**87.80**	*92.41*	93.75	**59.17**	*32.80*	**32.24**
ReFO+ (UD) (ours)	×	99.67	95.61	97.07	93.51	89.41	93.23	94.16	63.32	33.21	37.72

느낀점 및 추후 계획

- 태어나서 처음 읽은 논문
  - 많은 논문을 읽으며 연습 필요
- 수학이 거의 없이 아이디어와 실험으로 CVPR 게재
  - 수학 Base가 많이 약한 나도 가능성이 있음
- 해당 논문을 더 꼼꼼하게 읽고 LG과제 접목
  - 비판적 사고를 통해 연구가 더 될게 있는지 찾기