DistilHuBERT: Speech Representation Learning By layer-wise Distillation of Hidden-unit BERT

BrainLab of Seoul Tech National University

Department of Applied Artificial Intelligence

Jeong Sang-Yeop

I. Author

BrAIn Lab.
Brain and Artificial Intelligence Lab.

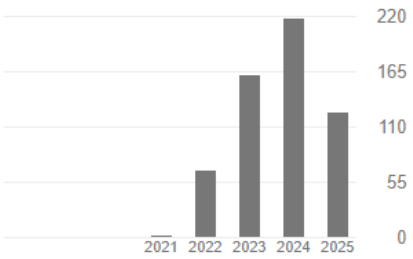
Heng-Jui Chang

Massachusetts Institute of Technology
mit.edu의 이메일 확인됨 - 홈페이지

Speech Processing Deep Learning

인용

	전체	2020년 이후
서지정보	575	575
h-index	11	11
i10-index	11	11



DistilHuBERT: Speech Representation Learning by Layer-wise Distillation of Hidden-unit BERT HJ Chang, S Yang, H Lee ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and …	198	2022
SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities HS Tsai, HJ Chang, WC Huang, Z Huang, K Lakhotia, S Yang, S Dong, … Proceedings of the 60th Annual Meeting of the Association for Computational …	121 *	2022
SpeechCLIP: Integrating Speech with Pre-Trained Vision and Language Model YJ Shih, HF Wang, HJ Chang, L Berry, H Lee, D Harwath 2022 IEEE Spoken Language Technology Workshop (SLT), 715-722	42	2023
Towards Lifelong Learning of End-to-end ASR HJ Chang, H Lee, L Lee Proc. Interspeech 2021	40	2021
A Large-scale Evaluation of Speech Foundation Models S Yang, HJ Chang, Z Huang, AT Liu, CI Lai, H Wu, J Shi, X Chang, … IEEE/ACM Transactions on Audio, Speech, and Language Processing 32, 2884-2899	39	2024
DinoSR: Self-Distillation and Online Clustering for Self-supervised Speech Representation Learning AH Liu, HJ Chang, M Auli, WN Hsu, J Glass Advances in Neural Information Processing Systems 36	28	2024
Self-supervised Fine-tuning for Improved Content Representations by Speaker-invariant Clustering HJ Chang, AH Liu, J Glass Proc. Interspeech 2023	25	2023
End-to-end Whispered Speech Recognition with Frequency-weighted Approaches and Pseudo Whisper Pre-training HJ Chang, AH Liu, H Lee, L Lee 2021 IEEE Spoken Language Technology Workshop (SLT)	21 *	2021

II. Problems

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Required Large Memory

High pre-train Cost

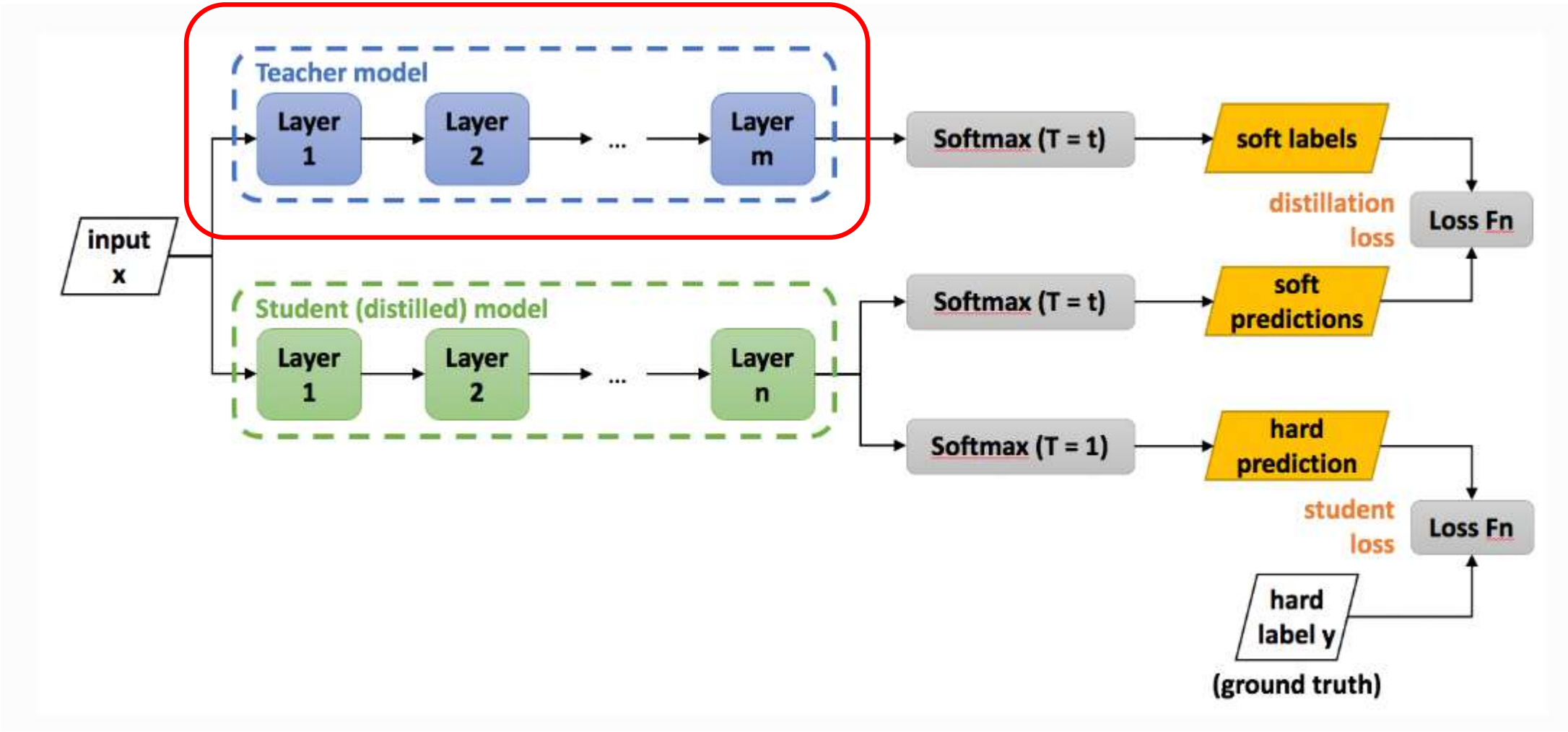
Hard to set on-device

Slow Inference speed

III. Background – Knowledge Distillation

BrAIn Lab. Brain and Artificial Intelligence Lab.

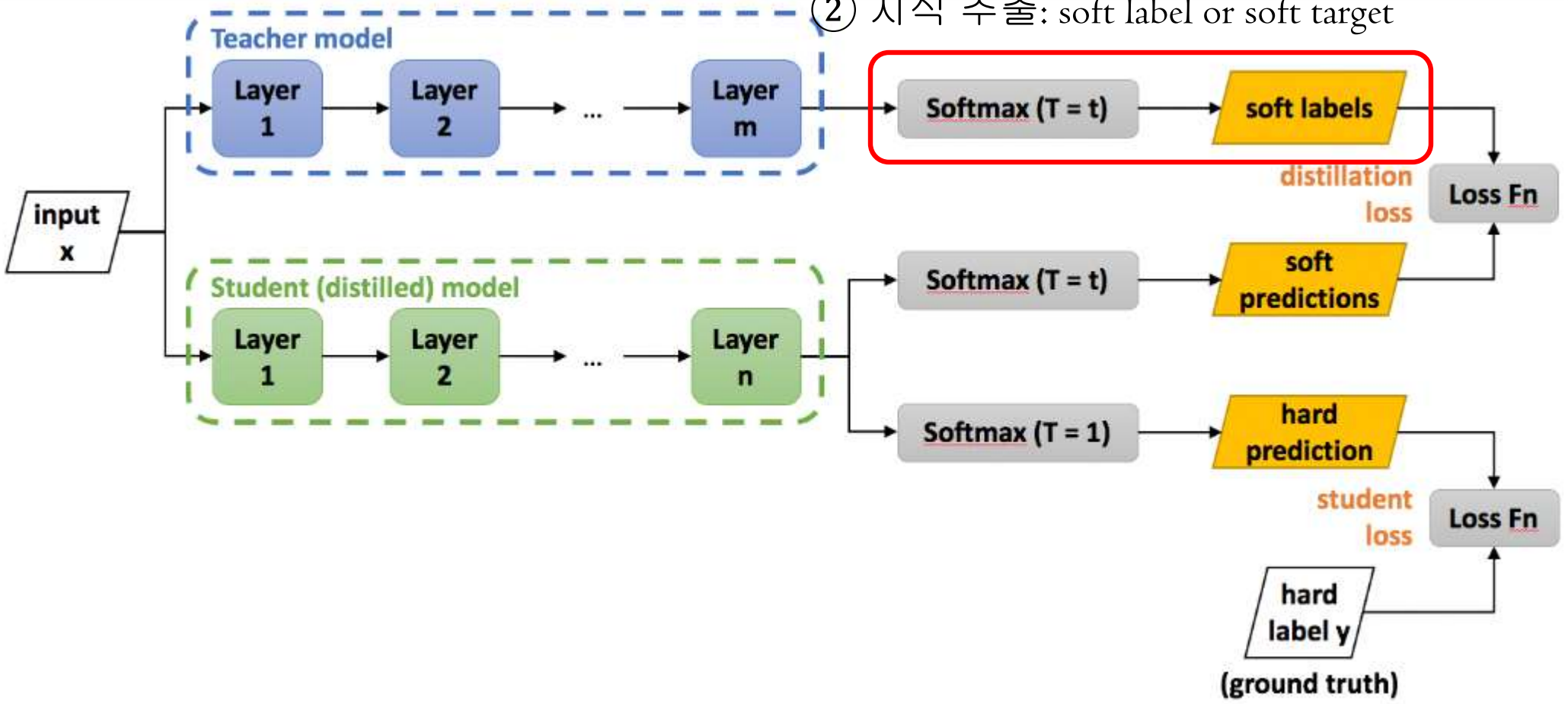
① 대규모 데이터로 Teacher model 학습



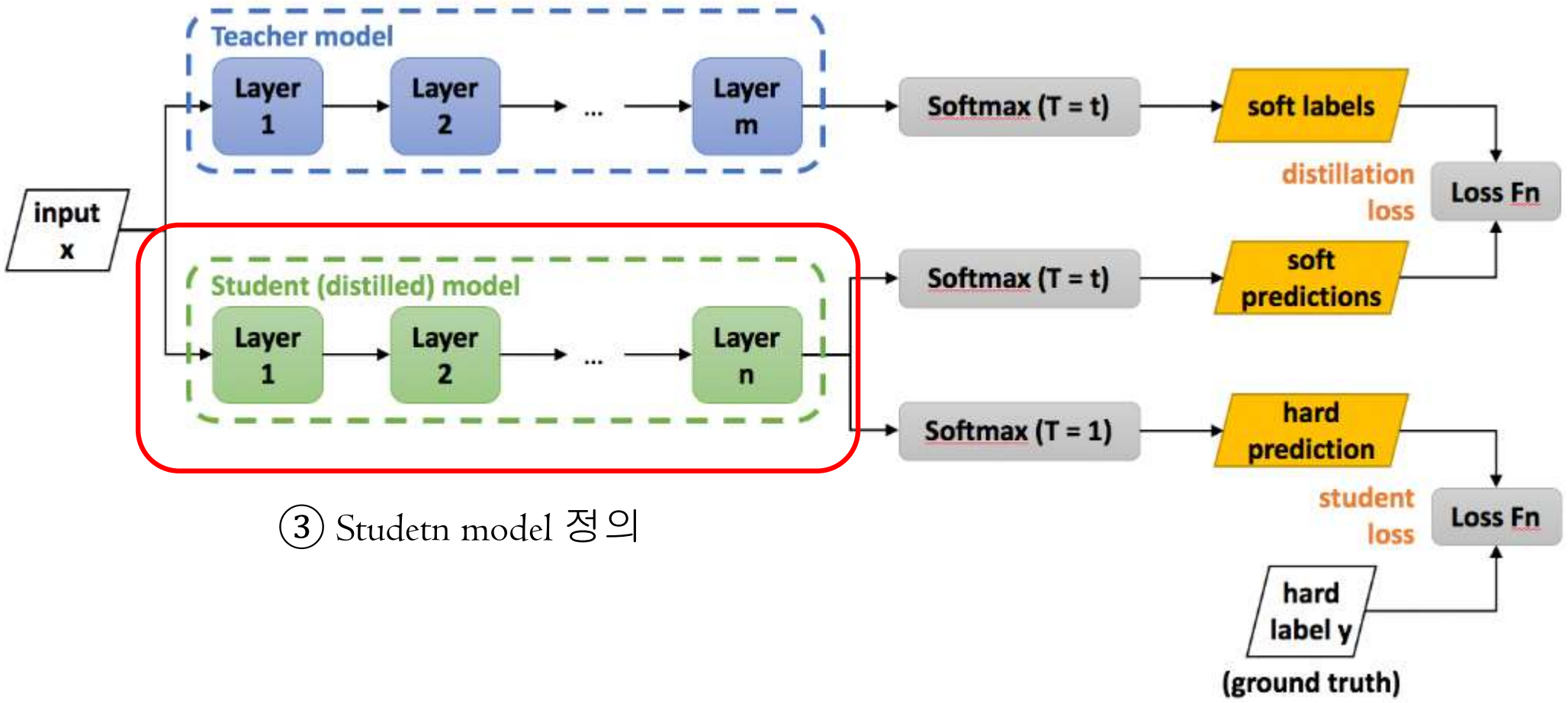
III. Background – Knowledge Distillation

BrAIn Lab.
Brain and Artificial Intelligence Lab.

② 지식 추출: soft label or soft target



III. Background – Knowledge Distillation

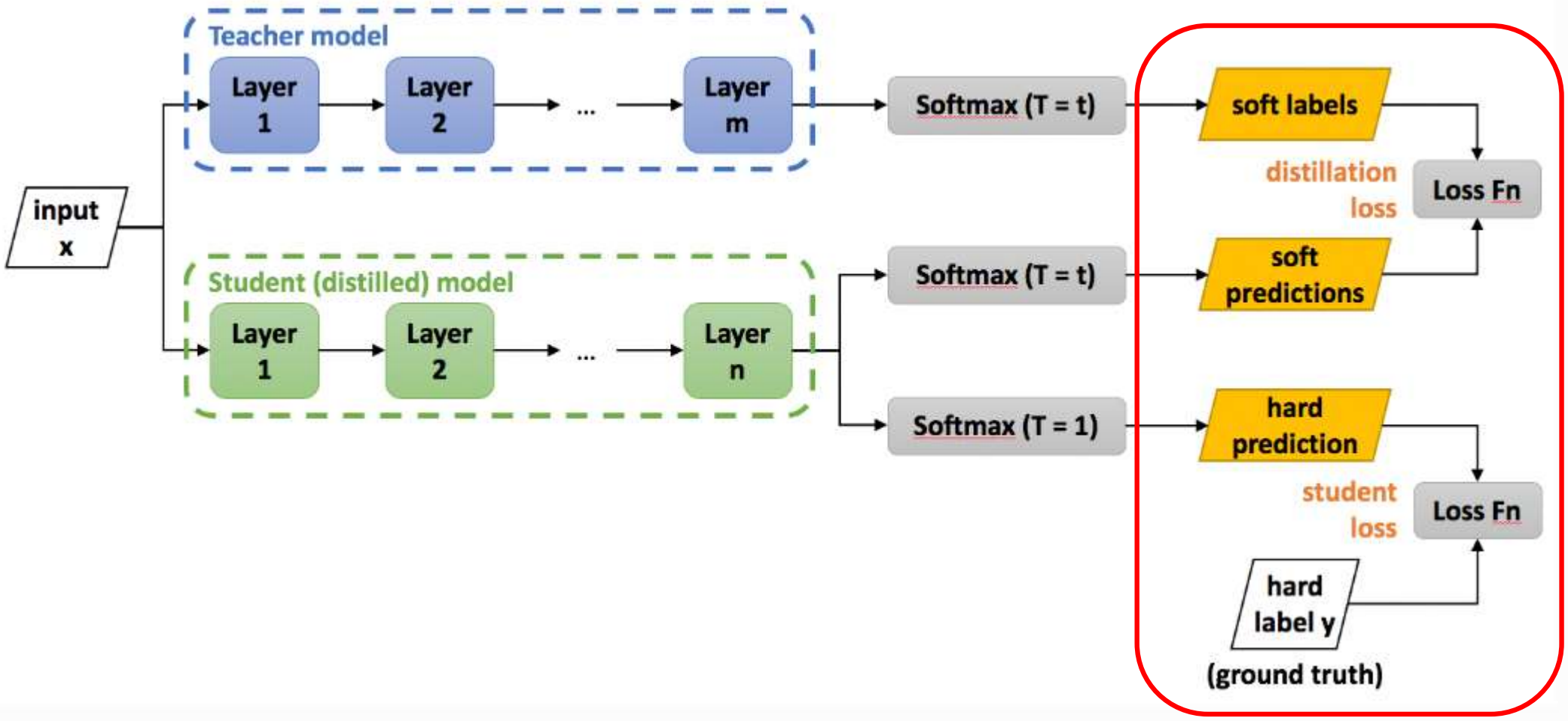


③ Studetn model 정의

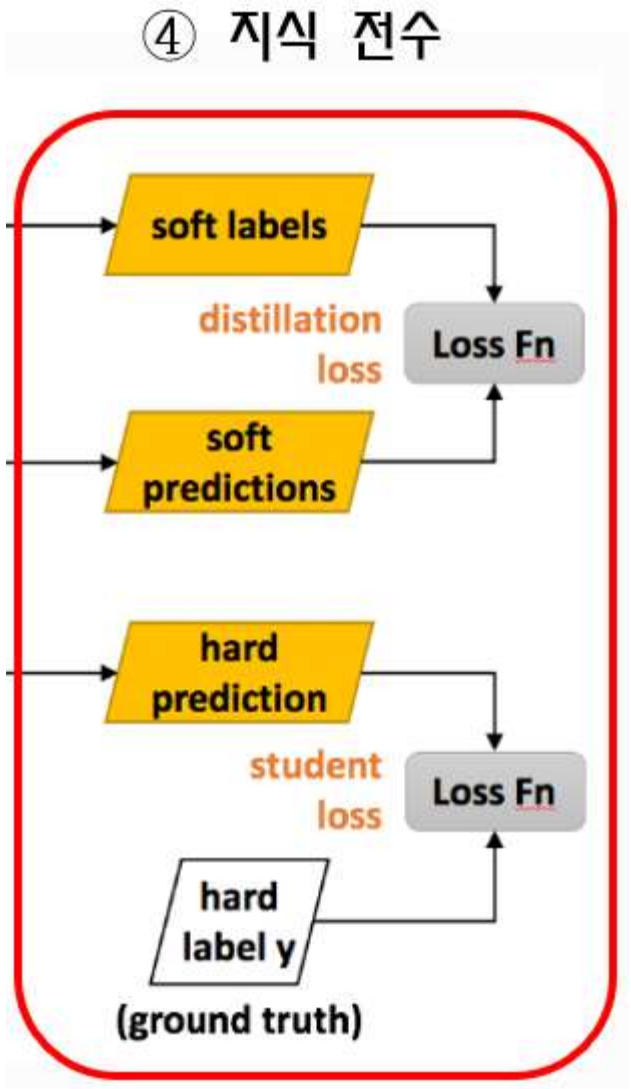
III. Background – Knowledge Distillation

BrAIn Lab.
Brain and Artificial Intelligence Lab.

④ 지식 전수



III. Background – Knowledge Distillation



Distillation Loss

- Student Model Prediction(Soft Predictions)이 Teacher model의 Soft Target(Soft Labels)와 유사해지도록 하는 손실

Student Loss

- Student Model Prediction이 Hard Target(Hard Label)과 일치하도록 하는 손실

III. Background – Knowledge Distillation

BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$\begin{pmatrix} Bear \\ Cat \\ Dog \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} Bear \\ Cat \\ Dog \end{pmatrix} = \begin{pmatrix} 0.05 \\ 0.75 \\ 0.2 \end{pmatrix}$$

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \quad \underset{with\ T}{Softmax(z_i)} = \frac{e^{z_i/T}}{\sum_{j=1}^{K} e^{z_j/T}}$$

$$Softmax\begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix} = \begin{pmatrix} 0.000335 \\ 0.000911 \\ 0.998754 \end{pmatrix}, \quad \underset{T=3}{Softmax}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.059 \\ 0.083 \\ 0.857 \end{pmatrix}$$

정보 손실
최소화!

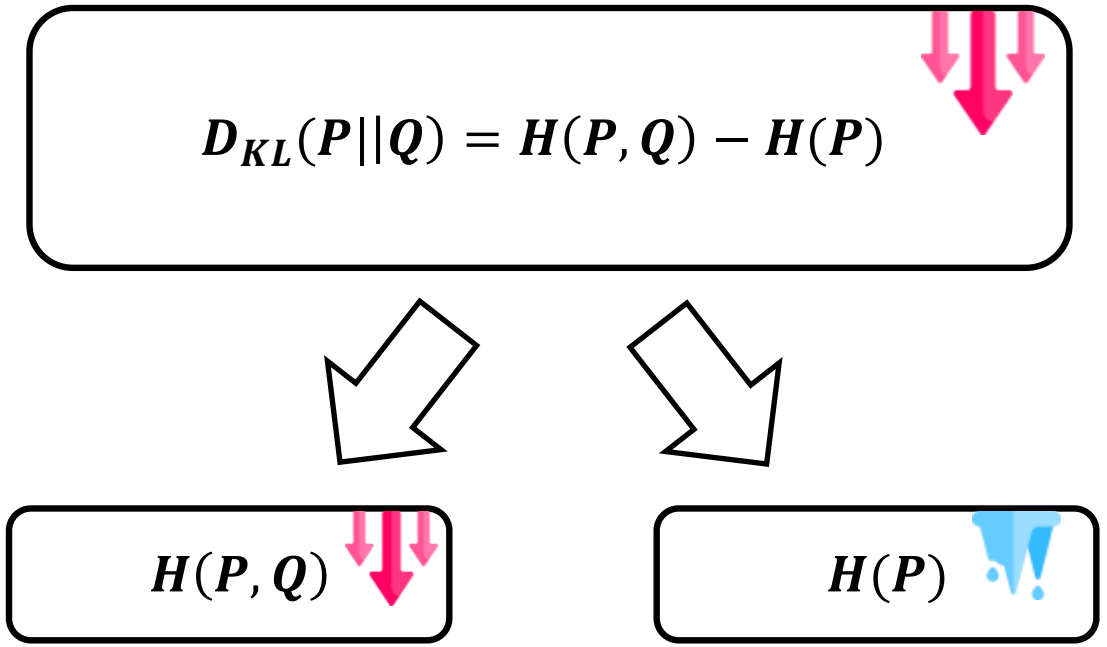
III. Background – Knowledge Distillation

BrAIn Lab.
Brain and Artificial Intelligence Lab.

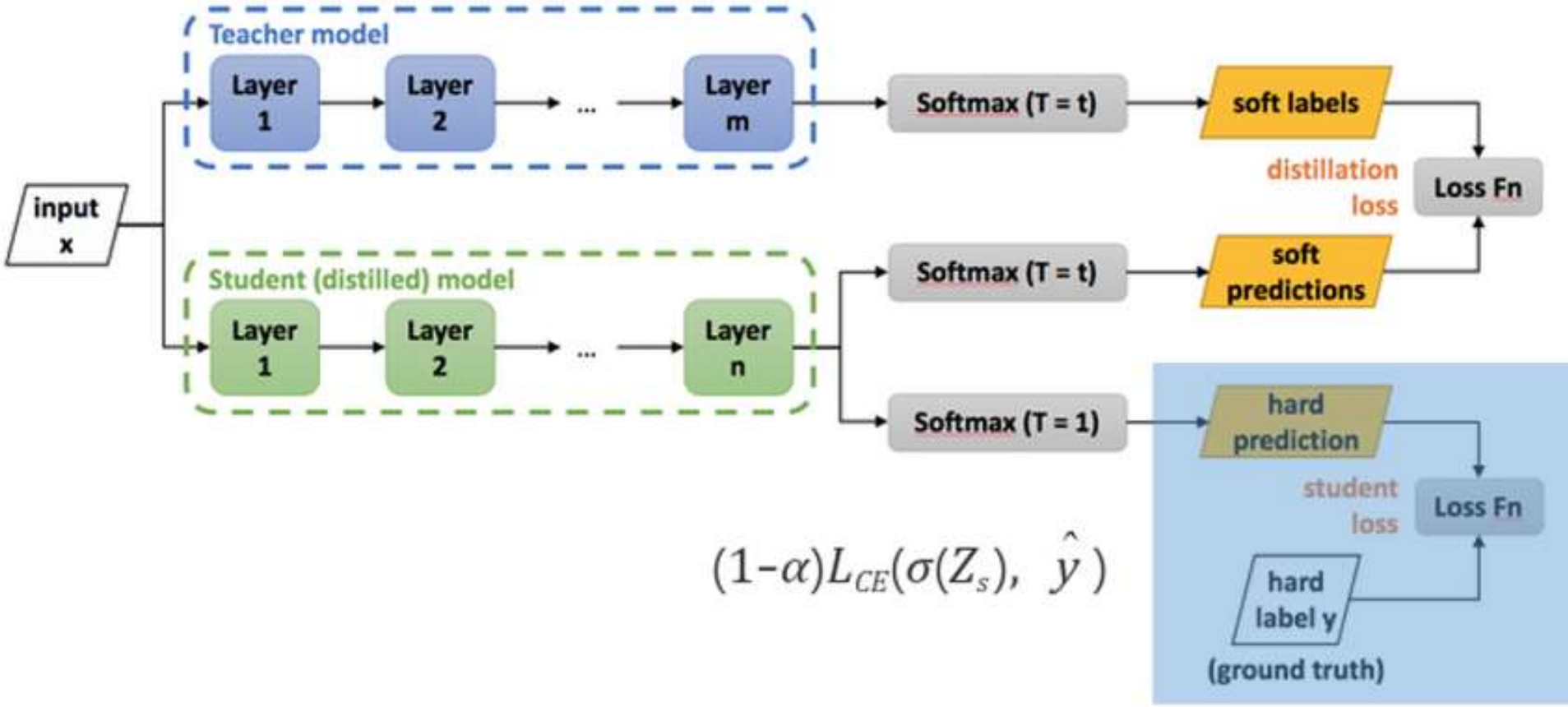
$$\boldsymbol{Total\ loss = \alpha \cdot L_{CE}\left(y, \sigma(Z_s)\right) + (1\ - \alpha) \cdot T^2 \cdot D_{KL}(\sigma\left(\frac{Z_t}{T}\right) || \sigma\left(\frac{Z_s}{T}\right))}$$

$$\boldsymbol{Total\ loss = \alpha \cdot L_{CE}(\sigma(Z_s), y\ ) + (1 - \alpha) \cdot T^2 \cdot L_{CE}\ (\sigma\left(\frac{Z_t}{T}\right), \sigma\left(\frac{Z_s}{T}\right))}$$

III. Background – Knowledge Distillation

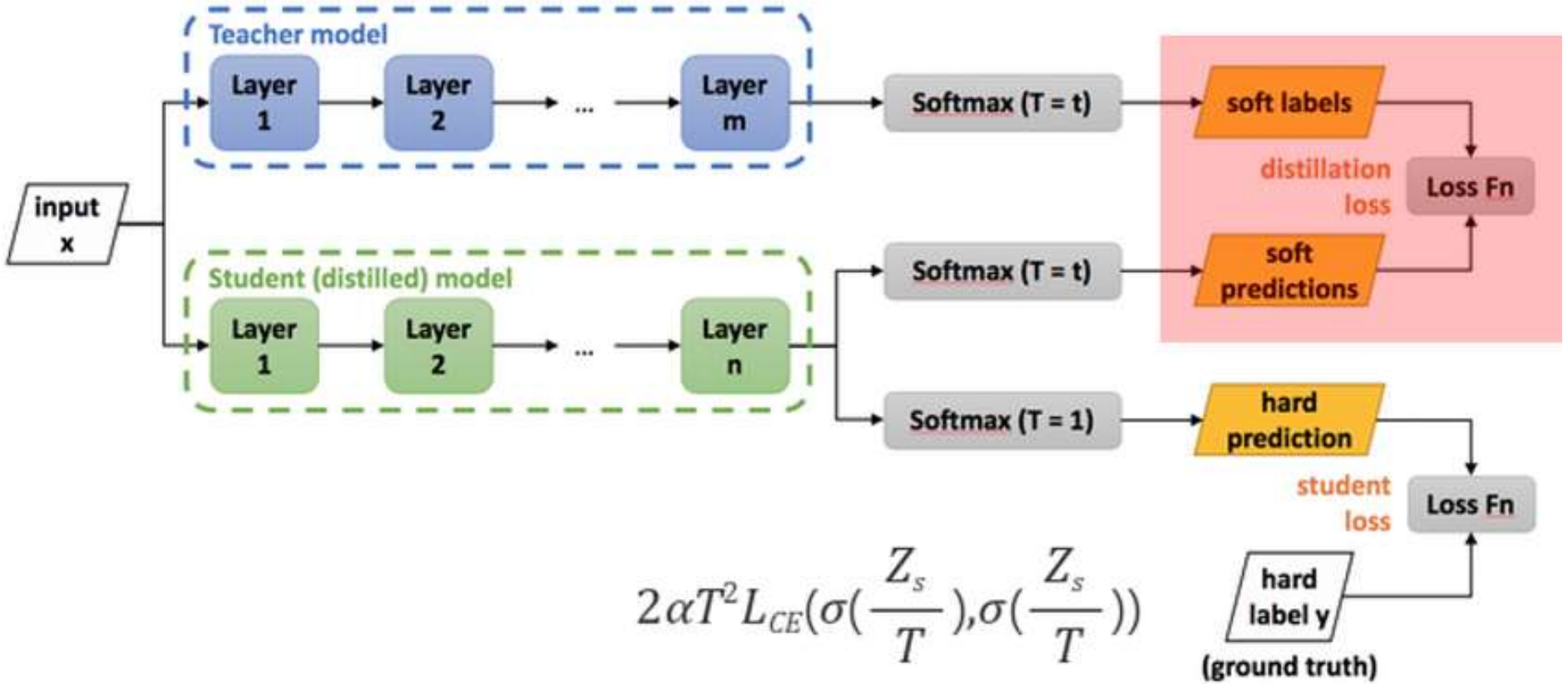


III. Background – Knowledge Distillation



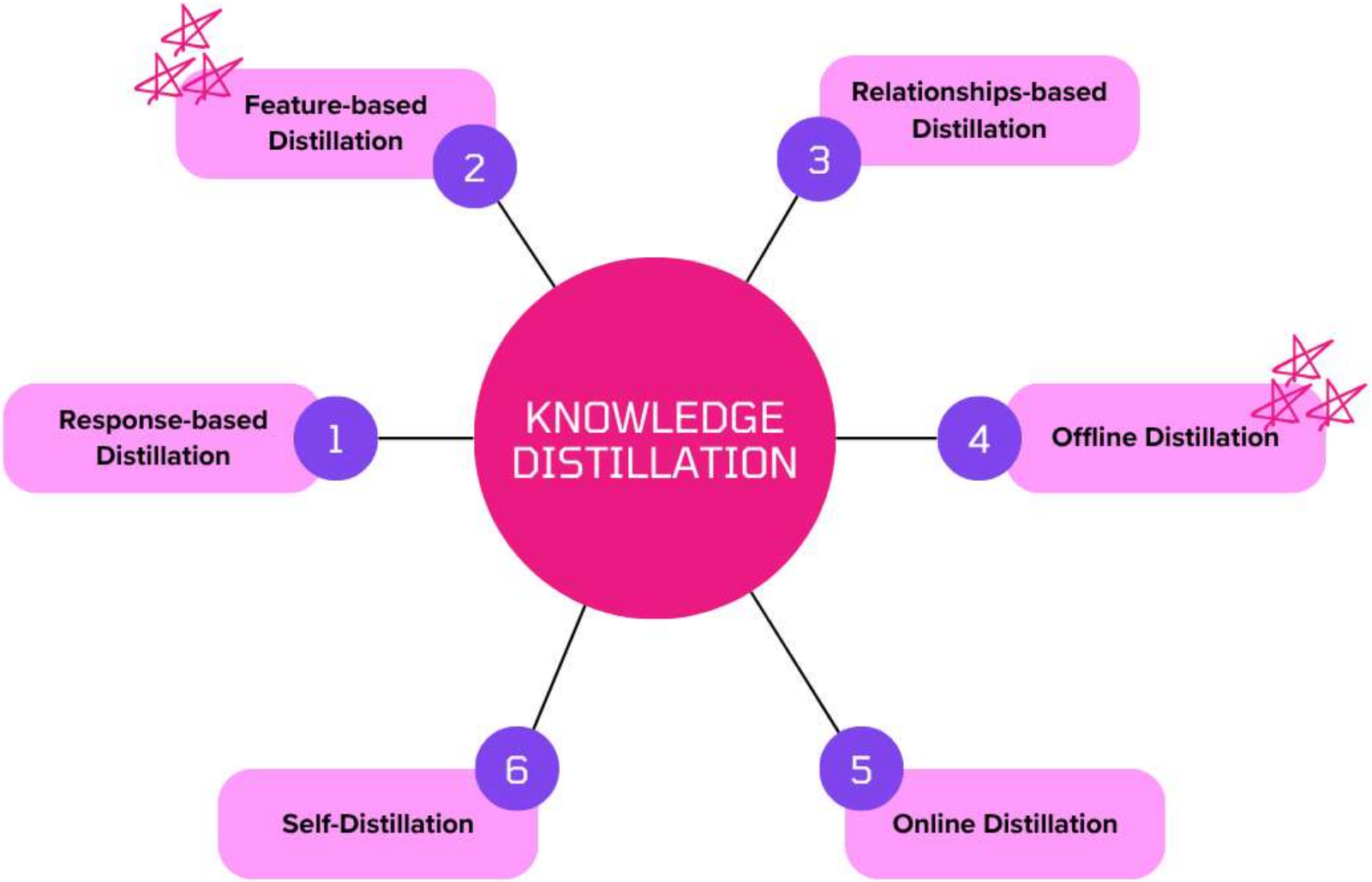
$$(1-\alpha)L_{CE}(\sigma(Z_s),\ \hat{y})$$

III. Background – Knowledge Distillation



$$2\alpha T^2 L_{CE}(\sigma(\frac{Z_s}{T}),\sigma(\frac{Z_s}{T}))$$

III. Knowledge Distillation Types



III. Knowledge Distillation Types

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Feature-based Distillation

- Student model이 Teacher model의 중간 계층, 피처 맵 등을 통해서 학습하는 방법

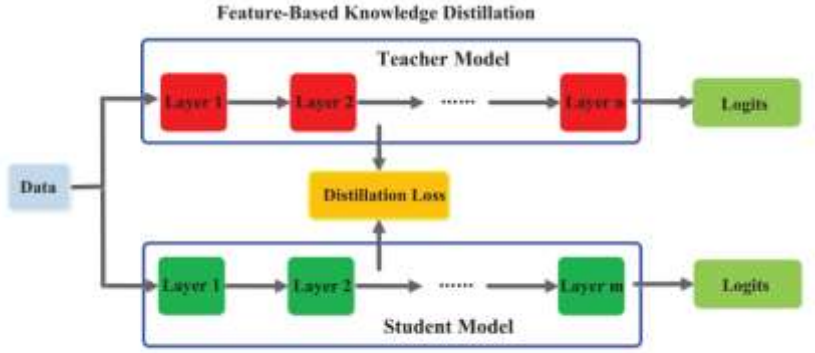


Fig. 6 The generic feature-based knowledge distillation.

Relation-based Distillation

- Student model이 Teacher model 내부의 여러 부분들 사이의 관계를 모방하는 법을 배우는 방법

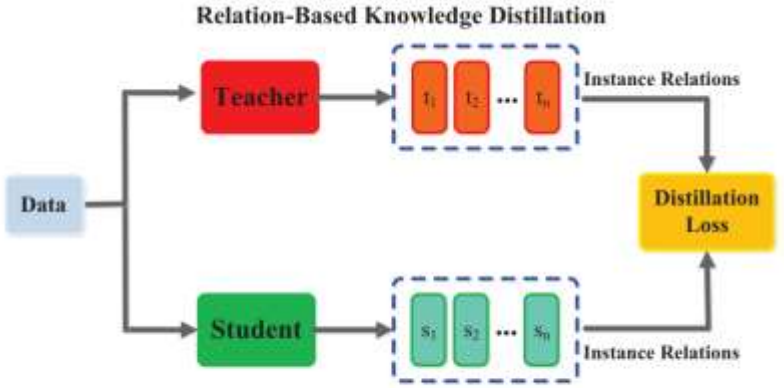


Fig. 7 The generic instance relation-based knowledge distillation.

III. Knowledge Distillation Types

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Offline Distillation

- Teacher model을 먼저 훈련시킨 다음, Student model을 가르침

Online Distillation

- Teacher model, Student model을 함께 훈련

Self Distillation

- Student model이 자기 자신으로부터 스스로 배움

III. Knowledge Distillation - Problems

BrAIn Lab.
Brain and Artificial Intelligence Lab.

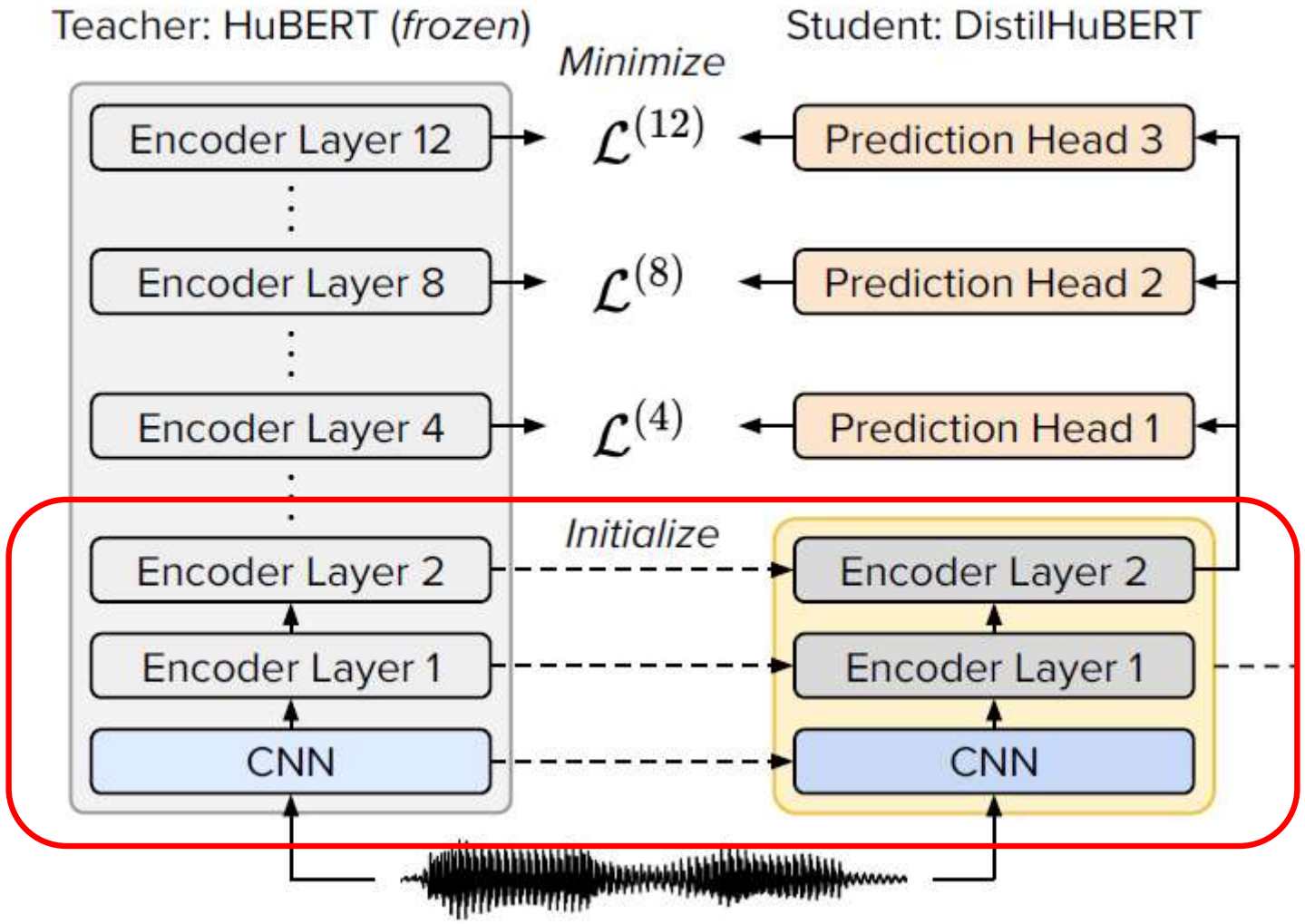
Problem I

- Student model이 Teacher model의 여러 hidden layer로부터 학습하려면 student model이 충분히 깊은 구조여야 한다는 것

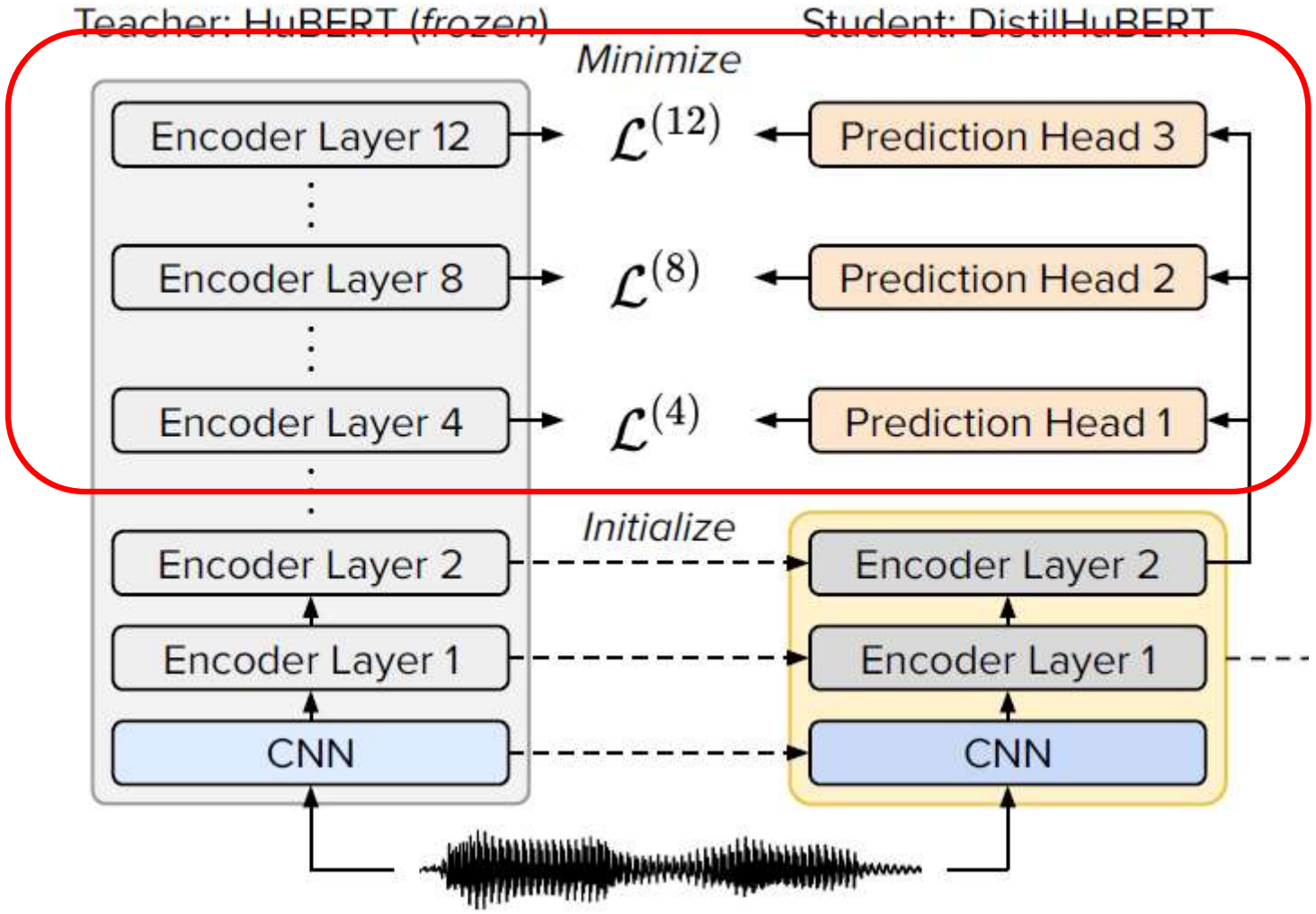
Problem 2

- 최종 출력 계층만 사용한다고 하면, 최종 출력 계층의 출력이 항상 풍부한 정보를 제공하지 않을 수도 있다는 점.
- Wav2vec 2.0은 음운 정보를 중간 계층에 저장하므로, 마지막 계층만 학습해서는 얻는 이점이 적음

IV. PreTraining - Initialization



IV. PreTraining – Train & Update



IV. PreTraining – Objective Function(1)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$L^{(l)} = L_{l1}^{(l)} + \lambda L_{cos}^{(l)} = \sum_{t=1}^{T} [\frac{1}{D} \left|\left| h_t^{(l)} - \hat{h}_t^l \right|\right|_1 - \lambda log\sigma(cos(h_t^{(l)}, \hat{h}_t^l))]$$

IV. PreTraining – Objective Function(2)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

LI 거리 최소화

- Student model의 예측 벡터가 Teacher model의 벡터와 '값' 측면에서 비슷해짐.

코사인 유사도 최대화

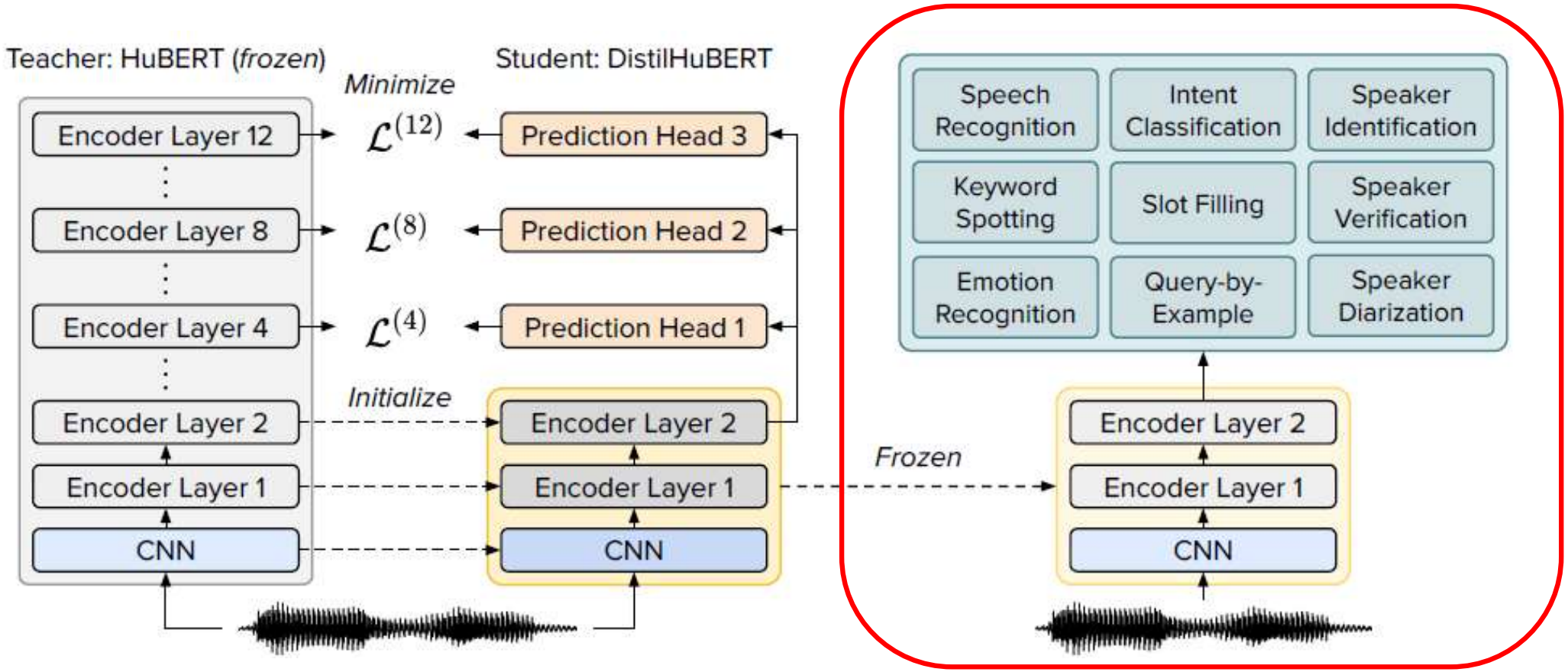
- 두 벡터 간의 코사인 유사도를 최대화
- 이는 Student model의 예측 벡터가 Teacher model의 벡터와 '방향' 측면에서 최대한 일치하도록 함.

IV. PreTraining – Objective Function(3)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$L = L^{(4)} + L^{(8)} + L^{(12)}$$

V. Downstream



VI. Results - SUPERB Bench Mark

BrAIn Lab. Brain and Artificial Intelligence Lab.

Method	# param. Millions	PR PER↓	KS Acc↑	IC Acc↑	SID Acc↑	ER Acc↑	ASR (+LM) WER↓	QbE MTWV↑	SF F1↑ / CER↓	ASV EER↓	SD DER↓	Rank↓
(I) Baselines [19]												
FBANK	0	82.01	8.63	9.10	8.5E-4	35.39	23.18 / 15.21	0.0058	69.64 / 52.94	9.56	10.05	8.8
TERA	21.33	49.17	89.48	58.42	57.57	56.27	18.17 / 12.16	0.0013	67.50 / 54.17	15.89	9.96	7.9
wav2vec	32.54	31.58	95.59	84.92	56.56	59.79	15.86 / 11.00	0.0485	76.37 / 43.71	7.99	9.90	6.5
DeCoAR 2.0	85.12	14.93	94.48	90.80	74.42	62.47	13.02 / 9.07	0.0406	83.28 / 34.73	7.16	6.59	4.1
HuBERT (teacher)	94.68	5.41	96.30	98.34	81.42	64.92	6.42 / 4.79	0.0736	88.53 / 25.20	5.11	5.88	1.0
(II) Distillation Baselines												
predict last layer	24.67	16.71	96.07	95.57	43.44	62.81	13.67 / 9.43	0.0423	81.52 / 37.31	7.29	6.45	4.6
predict w/ hidden	23.49	17.43	95.46	94.91	54.90	61.49	14.95 / 10.27	0.0529	83.73 / 35.14	7.41	6.86	5.0
(III) Proposed												
w/ prediction heads	27.03	14.35	96.20	94.17	72.83	62.73	13.26 / 8.99	0.0451	83.14 / 35.53	6.85	5.97	3.3
DistilHuBERT	23.49	16.27	95.98	94.99	73.54	63.02	13.37 / 9.21	0.0511	82.57 / 35.59	8.55	6.19	3.8

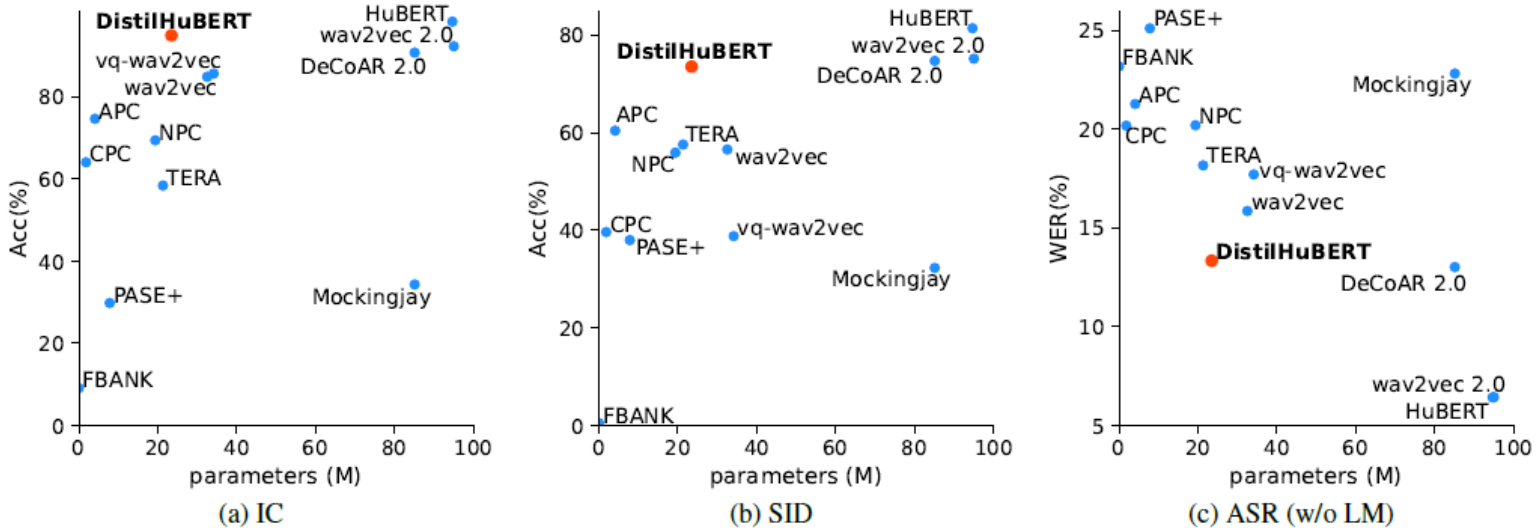


Fig. 2. Comparisons of pre-trained model size vs. performance on three tasks: IC, SID, and ASR.

Metrics

(1): 음소 인식 (Phoneme Recognition, PR)

(2): 키워드 감지 (Keyword Spotting, KS)

(3): 의도 분류 (Intent Classification, IC)

(4): 화자 식별 (Speaker Identification, SID)

(5): 감정 인식 (Emotion Recognition, ER)

(6): 예시 기반 음성 용어 탐지 (Query-by-Example Spoken Term Detection, QbE)

(7): 슬롯 채우기 (Slot Filling, SF)

(8): 화자 검증 (Automatic Speaker Verification, ASV)

(9): 화자 분리 (Speaker Diarization, DZ)

VI. Results - Model Size & Inference Speed

BrAIn Lab. Brain and Artificial Intelligence Lab.

Table 2. Comparison of different SSL model sizes and inference time. The inference was performed on 4 CPUs extracting all features from the LibriSpeech dev-clean set with a batch size of one. Results were averaged over three runs.

	# param.	Inf. time
Model	Millions	seconds
HuBERT	94.68 (100%)	992 (1.00X)
DeCoAR 2.0	85.12 (90%)	924 (1.07X)
DistilHuBERT	23.49 (25%)	574 (1.73X)

VI. Results – Ablation study(1)

BrAIn Lab. Brain and Artificial Intelligence Lab.

Table 3. Ablation study of the techniques used.

	IC	SID	ASR
Method	Acc↑	Acc↑	WER↓
DistilHuBERT	94.99	73.54	13.34
w/o $\mathcal{L}_{\text{cos}}$	93.36	72.54	13.74
w/o teacher init	94.20	73.68	13.41

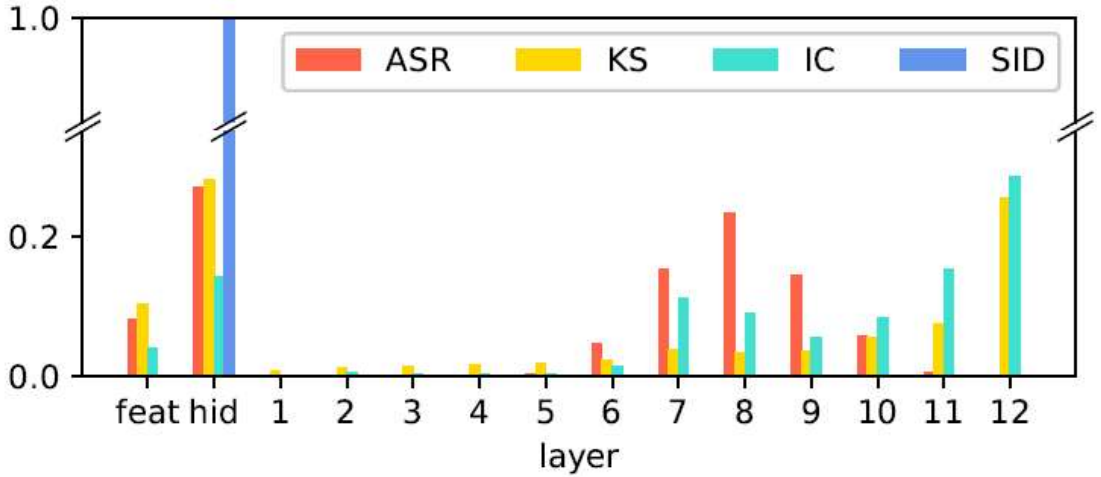


Fig. 3. Weights of the feature summarization in several SUPERB tasks after training on DistilHuBERT representations (predicting all hidden layers in HuBERT). `feat` and `hid` respectively denotes the CNN's and transformer's outputs.

VI. Results – Ablation study(2)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Table 4. Results of predicting different HuBERT's hidden layers.

	IC	SID	ASR
Predicted Layers	Acc↑	Acc↑	WER↓
4, 8, 12	94.99	73.54	13.34
4	79.09	76.85	14.90
8	96.89	61.52	12.28
12	95.52	65.14	13.48
4, 8	91.67	75.36	13.73
4, 12	91.88	74.48	14.03
8, 12	96.34	65.30	13.03

VI. Results – Ablation study(3)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Table 5. Pre-training DistilHuBERT with different datasets.

	Data Size	IC	SID	ASR
Dataset	hours	Acc↑	Acc↑	WER↓
LibriSpeech	960	94.99	73.54	13.34
LibriSpeech	100	93.17	69.46	14.77
WSJ	81	90.22	64.14	15.59
AISHELL-1	150	87.29	67.65	16.42