HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

BrainLab of Seoul Tech National University

Department of Applied Artificial Intelligence

Jeong Sang-Yeop

BrAIn Lab.

Brain and Artificial Intelligence Lab.

I. Author

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Wei-Ning Hsu

PhD Student at Massachusetts Institute of Technology

125	18,766	9,159
Publications	Reads ⓘ	Citations

1439	2	10394
Cites in Papers	Cites in Patents	Full Text Views

Hubert: Self-supervised speech representation learning by masked prediction of hidden units WN Hsu, B Bolte, YHH Tsai, K Lakhotia, R Salakhutdinov, A Mohamed IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 3451-3460	3841	2021
Data2vec: A general framework for self-supervised learning in speech, vision and language A Baevski, WN Hsu, Q Xu, A Babu, J Gu, M Auli International conference on machine learning, 1298-1312	1065	2022
An unsupervised autoregressive model for speech representation learning YA Chung, WN Hsu, H Tang, J Glass INTERSPEECH	491	2019
Unsupervised learning of disentangled and interpretable representations from sequential data WN Hsu, Y Zhang, J Glass Thirty-first Conference on Neural Information Processing Systems (NeurIPS)	446	2017
On generative spoken language modeling from raw audio K Lakhotia, E Kharitonov, WN Hsu, Y Adi, A Polyak, B Bolte, TA Nguyen, ... Transactions of the Association for Computational Linguistics 9, 1336-1354	409	2021
Scaling speech technology to 1,000+ languages V Pratap, A Tjandra, B Shi, P Tomasello, A Babu, S Kundu, A Elkahky, ... Journal of Machine Learning Research 25 (97), 1-52	394	2024
Learning audio-visual speech representation by masked multimodal cluster prediction B Shi, WN Hsu, K Lakhotia, A Mohamed arXiv preprint arXiv:2201.02184	389	2022
Voicebox: Text-guided multilingual universal speech generation at scale M Le, A Vyas, B Shi, B Karrer, L Sari, R Moritz, M Williamson, V Manohar, ... Advances in neural information processing systems 36, 14005-14034	362	2023
Speech Resynthesis from Discrete Disentangled Self-Supervised Representations A Polyak, Y Adi, J Copet, E Kharitonov, K Lakhotia, WN Hsu, A Mohamed, ... INTERSPEECH	354	2021
Unsupervised speech recognition A Baevski, WN Hsu, A Conneau, M Auli Advances in Neural Information Processing Systems 34, 27826-27839	351	2021
Hierarchical generative modeling for controllable speech synthesis WN Hsu, Y Zhang, RJ Weiss, H Zen, Y Wu, Y Wang, Y Cao, Y Jia, Z Chen, ... Seventh International Conference on Learning Representations (ICLR)	334	2019
Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training WN Hsu, A Sriram, A Baevski, T Likhomanenko, Q Xu, V Pratap, J Kahn, ... INTERSPEECH	278	2021

II. Purpose

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Three problems from speech representation learning

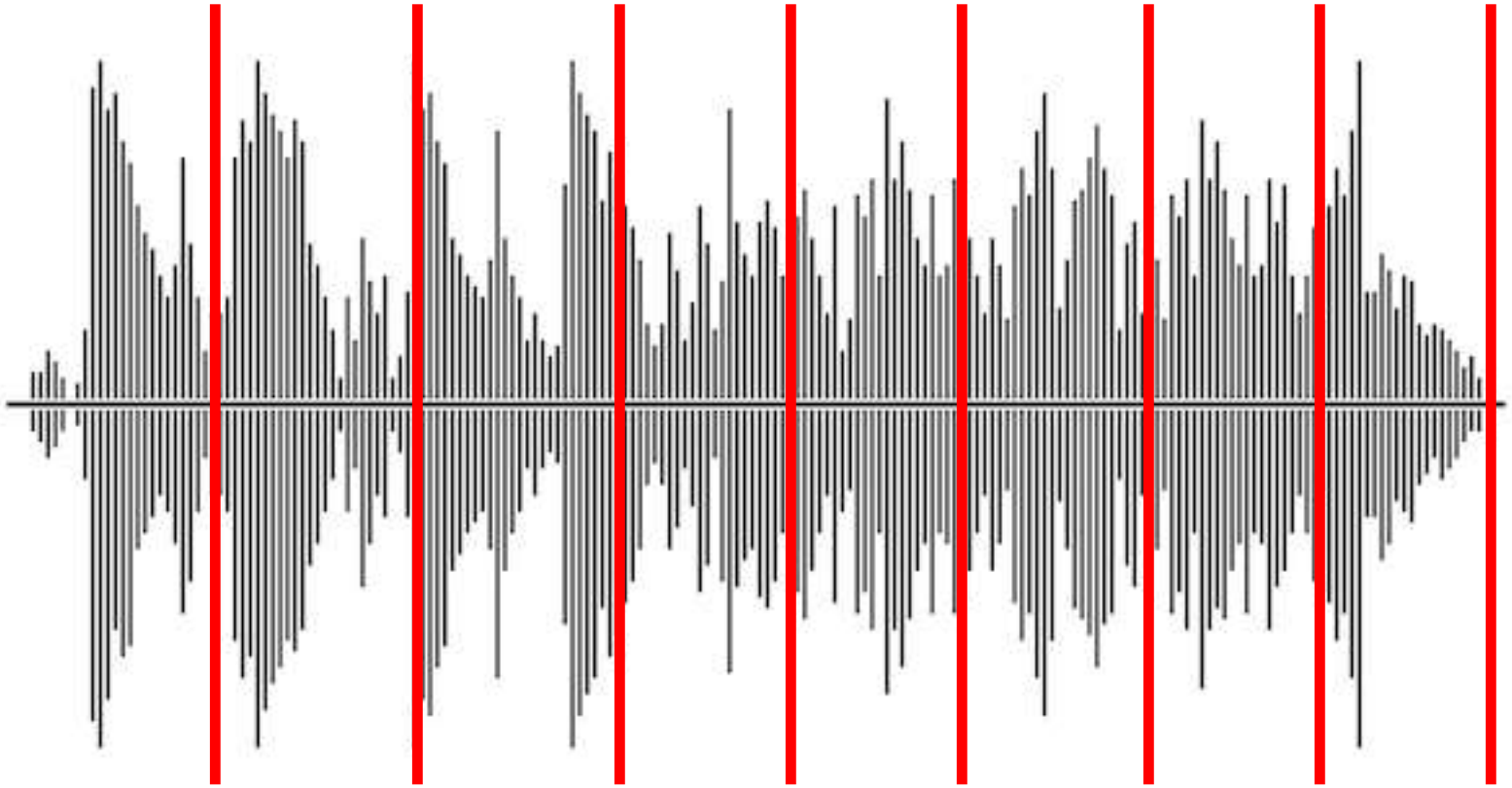
- Multiple sound units in each input utterance
- No lexicon of input sound units
- Sound units have variable lengths

Hidden-Unit BERT

- Offline clustering
- Applying the prediction loss over the masked regions only

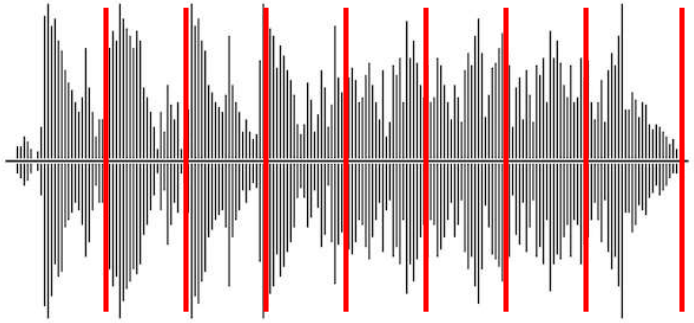
III. Backgrounds – Offline clustering(1)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

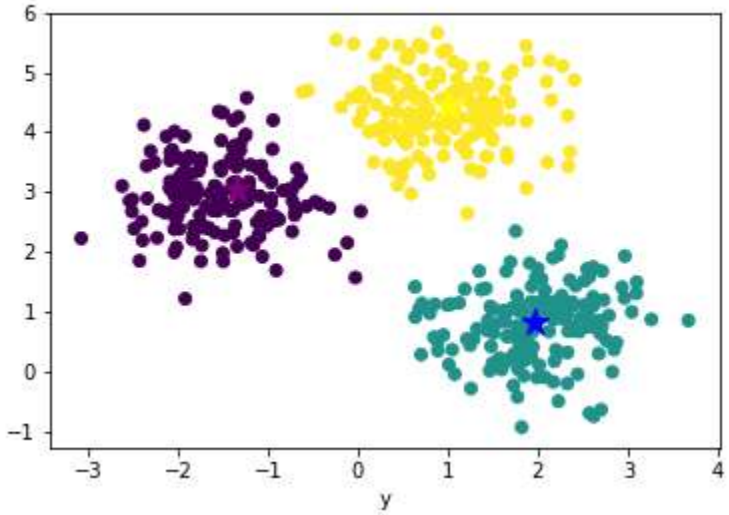


III. Backgrounds – Offline clustering(2)

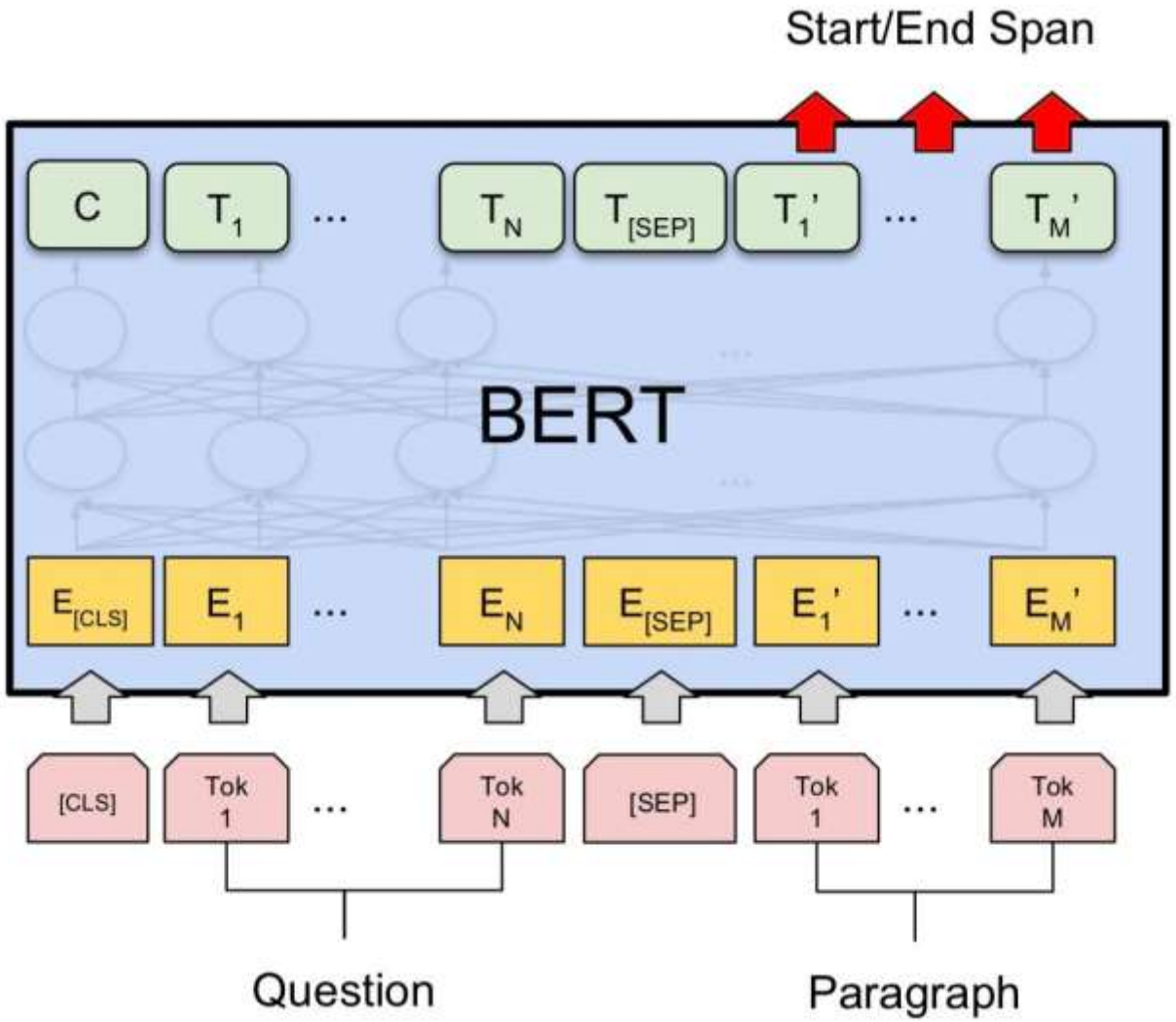
BrAIn Lab.
Brain and Artificial Intelligence Lab.



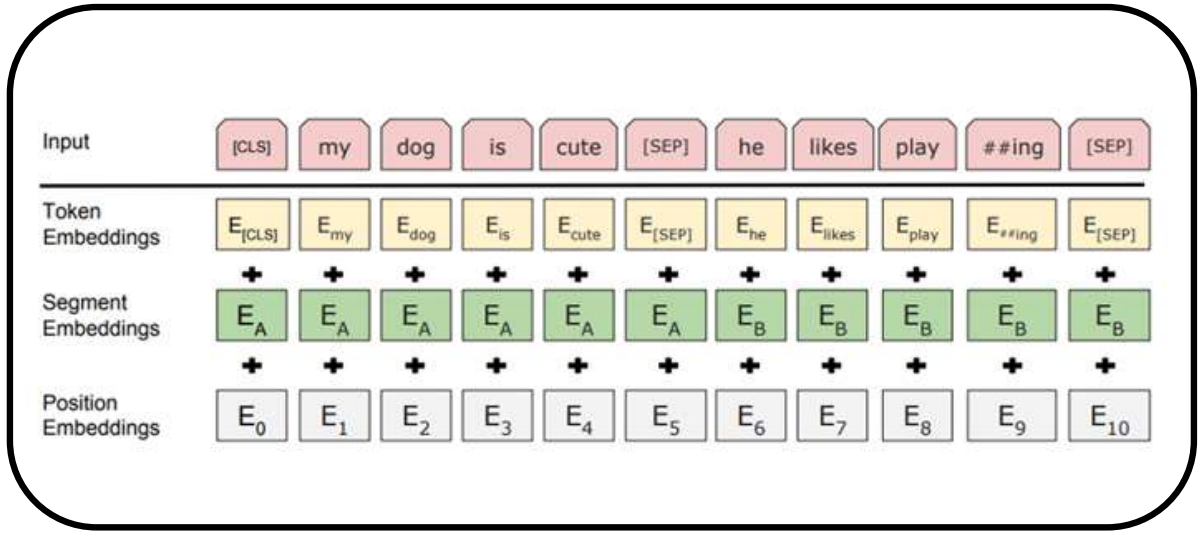
< k-means clustering >



IV. Backgrounds – BERT(1)



IV. Backgrounds – BERT(2)



Word piece Embedding

- 자주 등장하는 단어와 자주 등장하지 않는 단어를 구분하여 임베딩

Segment Embedding

- 각 문장을 구분하기 위해서 [SEP] 토큰을 기준으로 임베딩

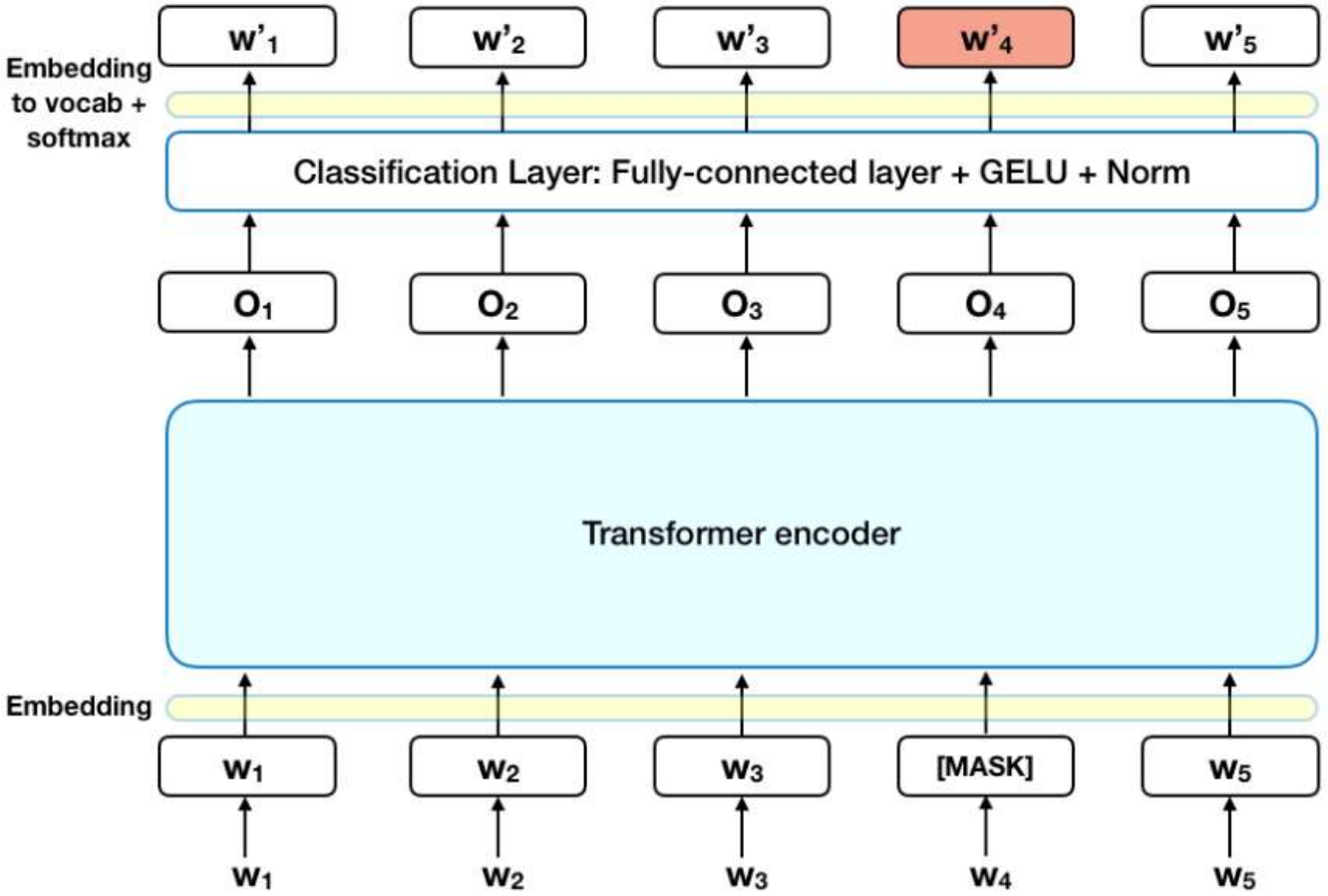
Position Embedding

- 입력 토큰의 위치정보를 임베딩

IV. Backgrounds – BERT(3)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

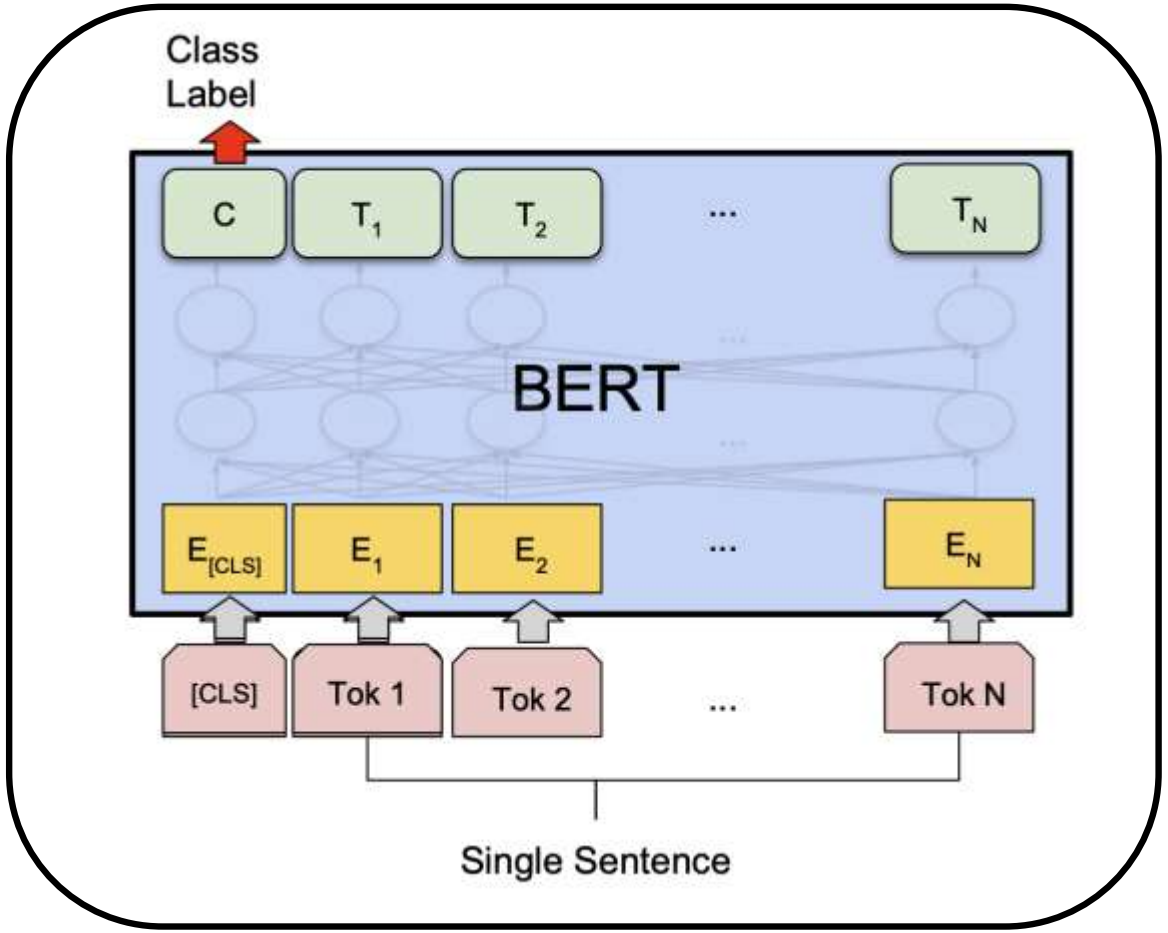
< Pre-Training >



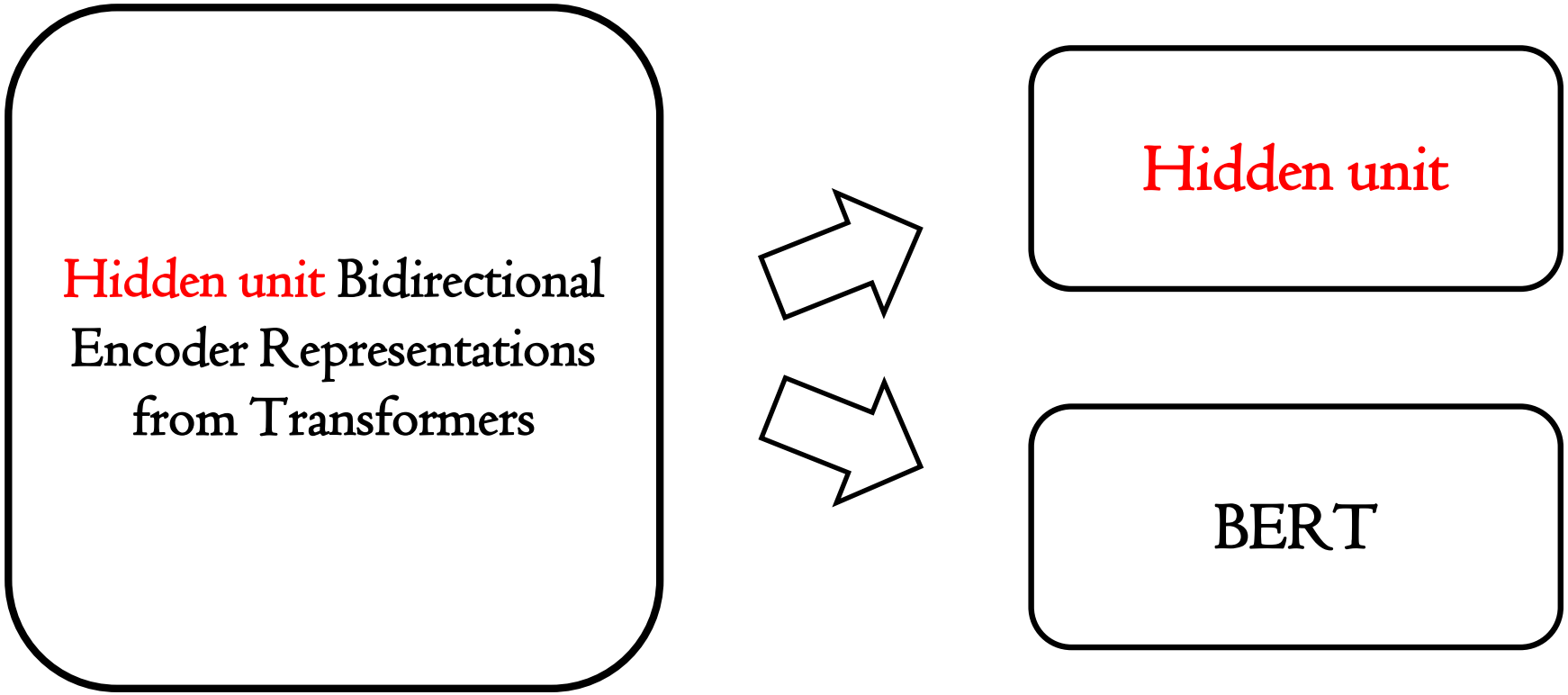
IV. Backgrounds - BERT(3)

BrAIn Lab.
Brain and Artificial Intelligence Lab.

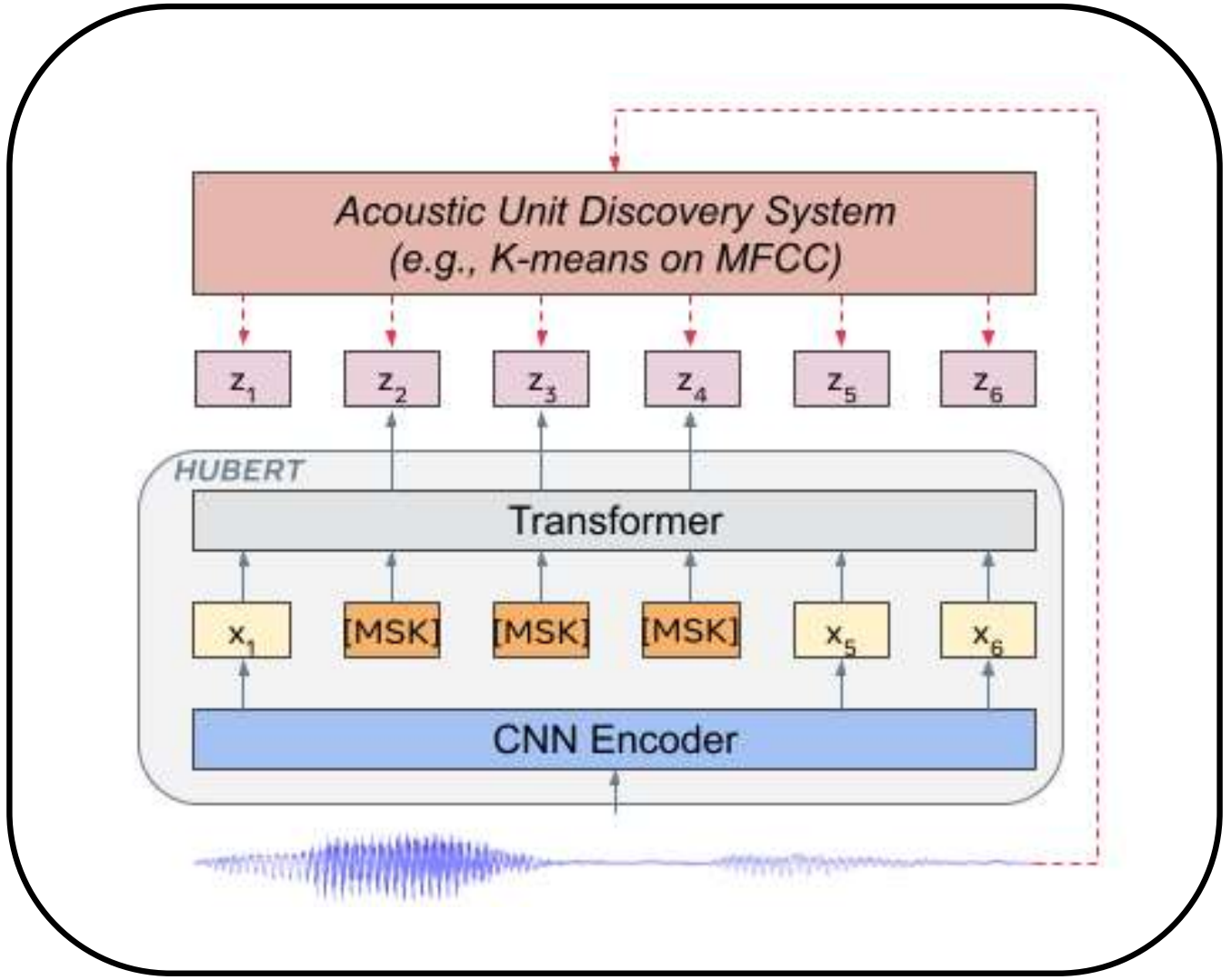
< Fine-Tuning >



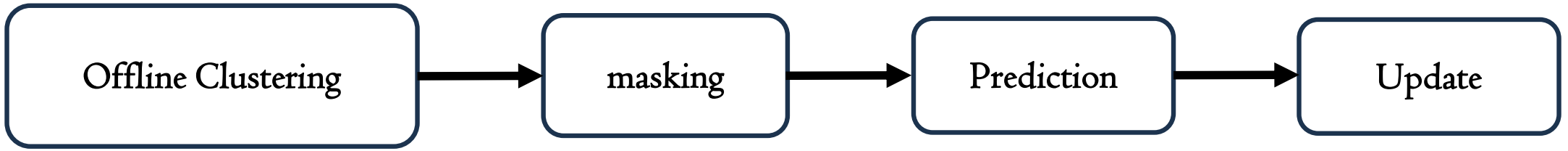
V. HuBERT(PreTraining) – The Name



V. HuBERT(PreTraining) - PreTraining



V. HuBERT(PreTraining) – Overview



V. HuBERT(PreTraining) – Loss Functions

BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t \mid \tilde{X}, t)$$

$$L_u(f; X, M, Z) = \sum_{t \notin M} \log p_f(z_t \mid \tilde{X}, t)$$

V. HuBERT(PreTraining) – Loss Functions

BrAIn Lab.
Brain and Artificial Intelligence Lab.

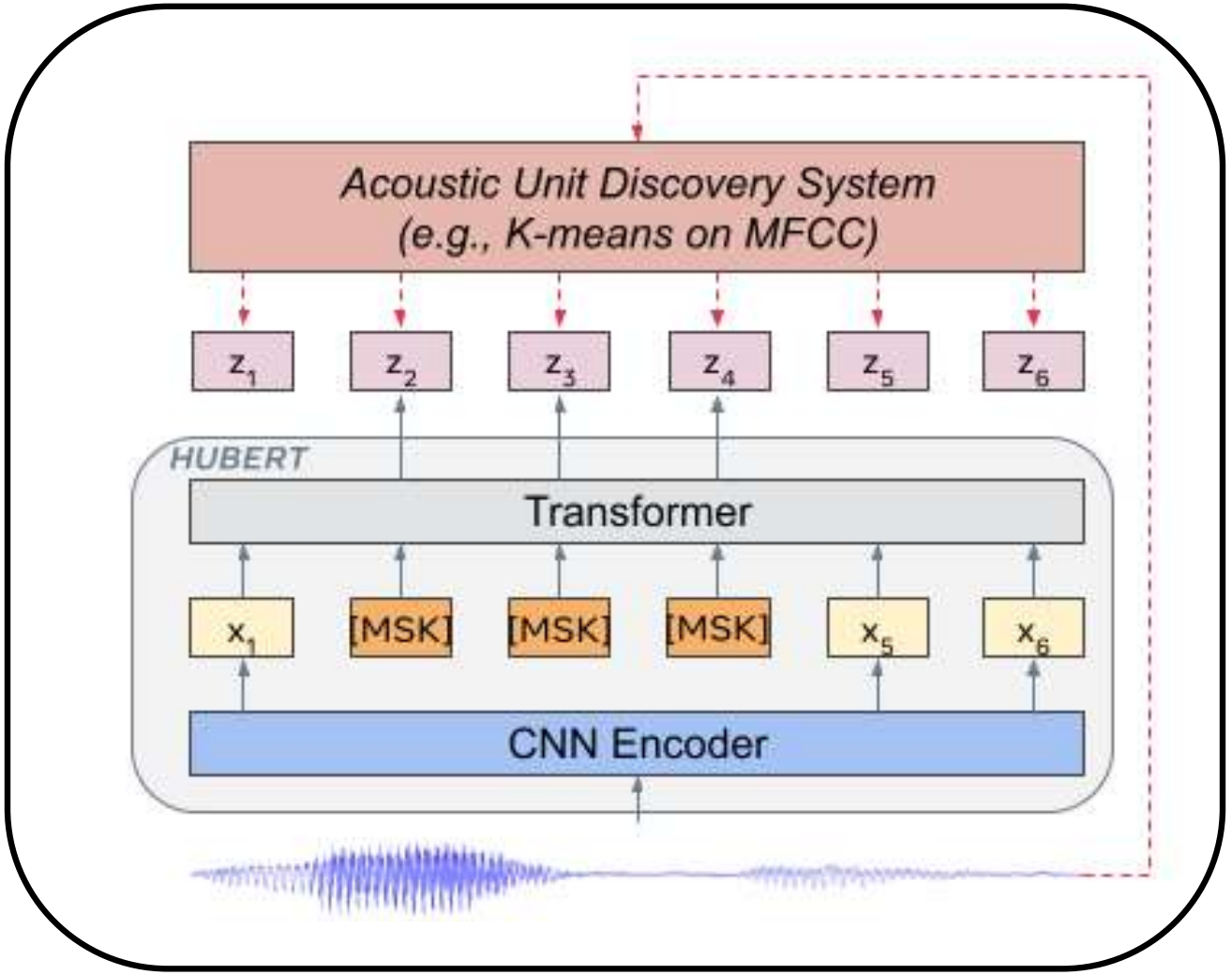
$$\mathrm{p_f}\left(\mathrm{c}\middle|\tilde{X}, t\right) = \frac{\exp(sim(Ao_t, e_c)/\tau)}{\sum_{c'=1}^{C} \exp(sim(Ao_t, e_{c'})/\tau)}$$

V. HuBERT(PreTraining) – Loss Functions

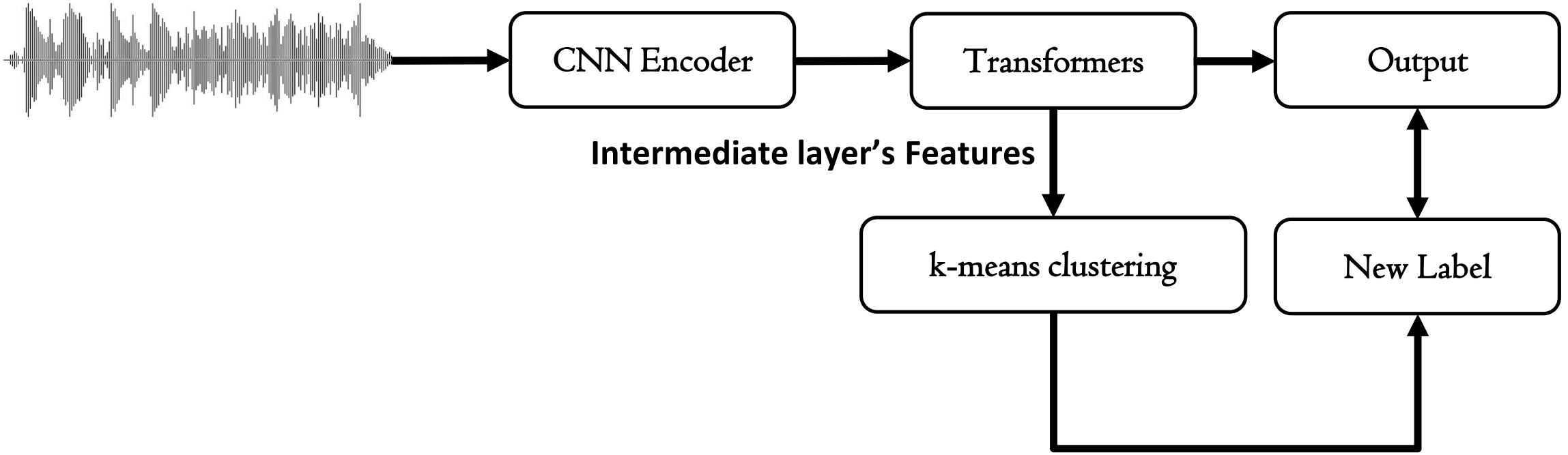
BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$L = \alpha L_m + (1 - \alpha) L_u$$

V. HuBERT(PreTraining) – 1st Iterations



V. HuBERT(PreTraining) – After 1st Iterations

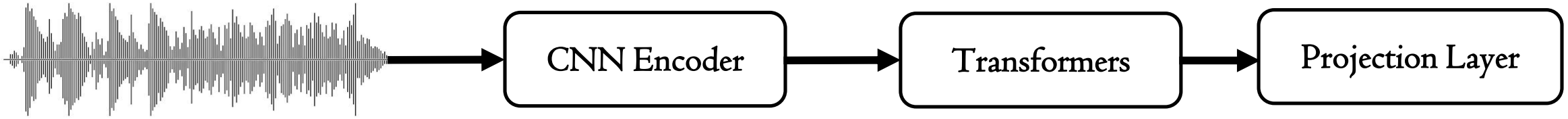


V. HuBERT(Model) – Architecture Summary

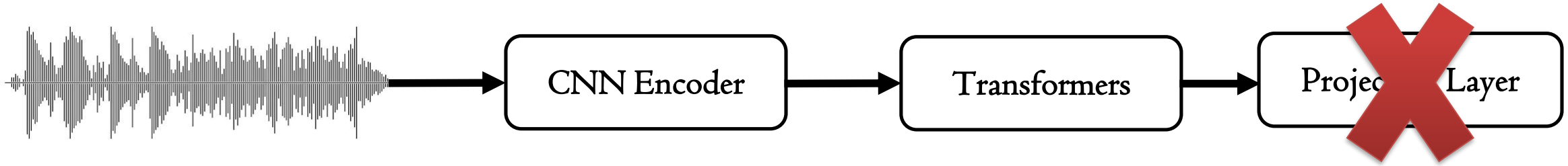
BrAIn Lab.
Brain and Artificial Intelligence Lab.

		BASE	LARGE	X-LARGE
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width	10, 3, 3, 3, 3, 2, 2		
	channel	512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	8	16	16
Projection	dim.	256	768	1024
Num. of Params		95M	317M	964M

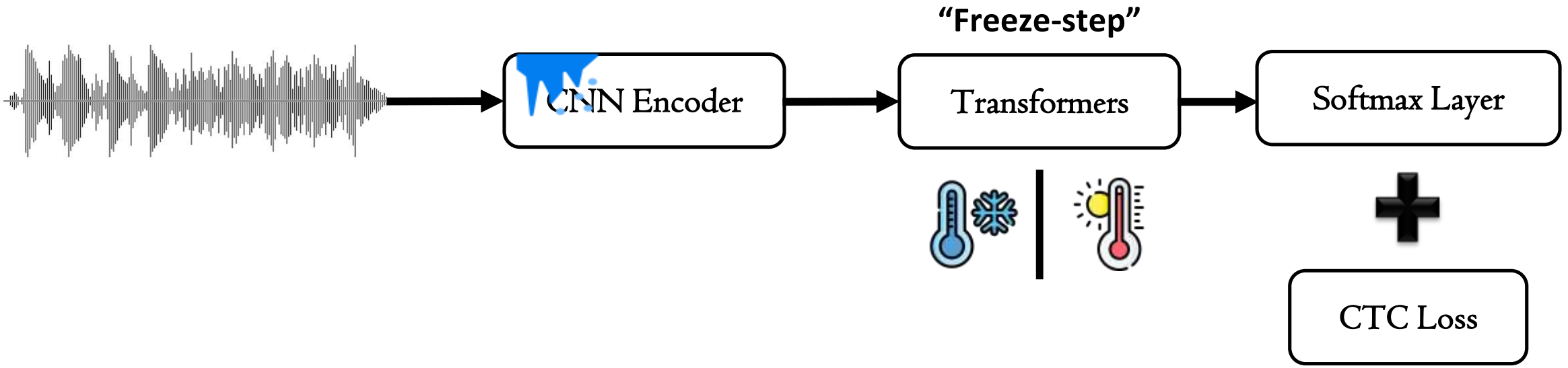
VI. HuBERT(FineTuning) - ASR



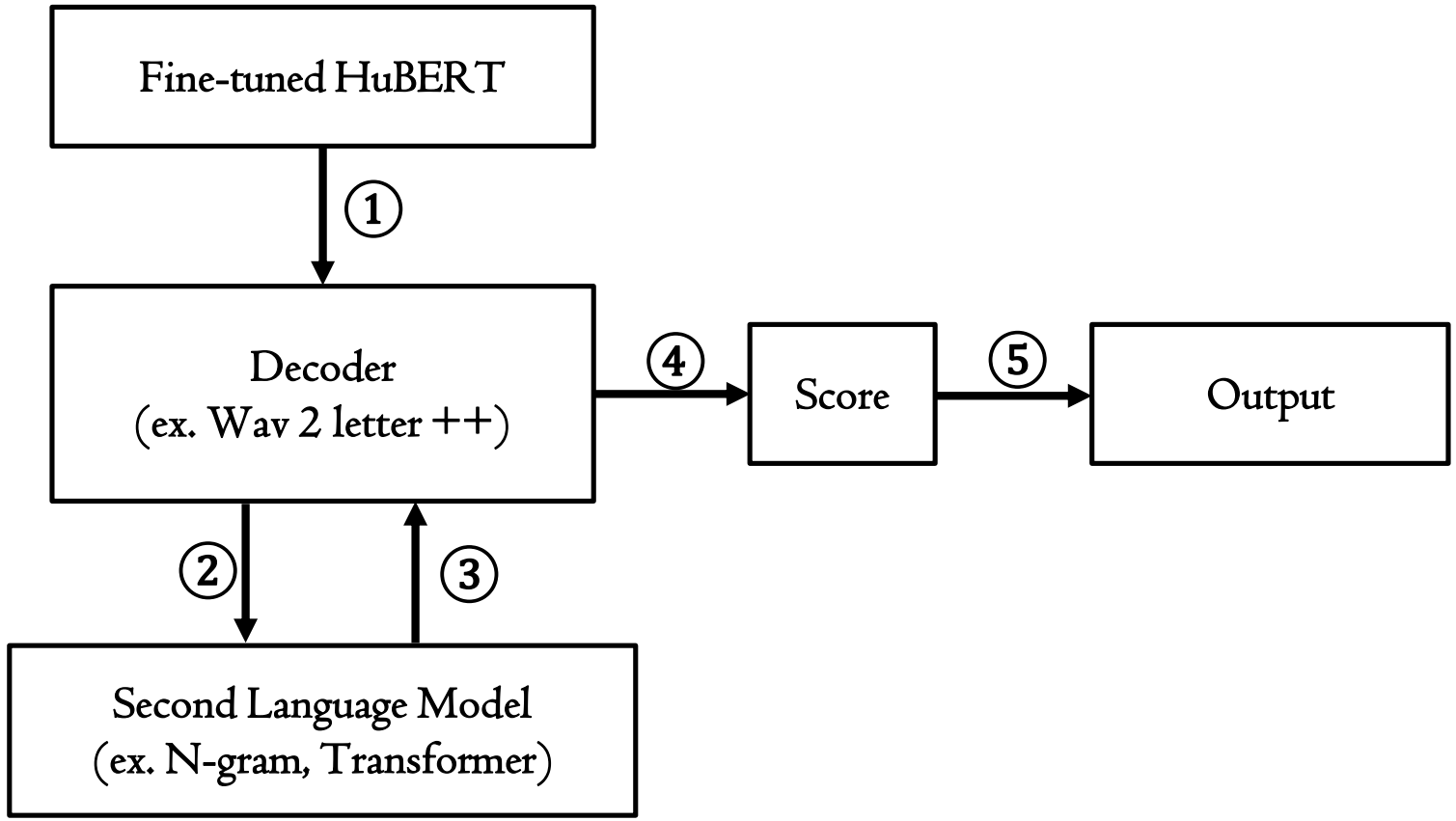
VI. HuBERT(FineTuning) - ASR



VI. HuBERT(FineTuning) - ASR



VII. HuBERT(Decoding) – Process



VII. HuBERT(Decoding) – Score Function

BrAIn Lab.
Brain and Artificial Intelligence Lab.

$$\log P_{CTC}(Y|X) + w_1 log P_{LM}(Y) + w_2|Y|$$

VIII. HuBERT(Result) - SOTA

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		10-min labeled				
DiscreteBERT [51]	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE [6]	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE [6]	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE	LL-60k	Transformer	4.3	7.0	4.7	7.6
HUBERT X-LARGE	LL-60k	Transformer	4.4	6.1	4.6	6.8
		1-hour labeled				
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT [51]	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE [6]	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE	LL-60k	Transformer	2.6	4.9	2.9	5.4
HUBERT X-LARGE	LL-60k	Transformer	2.6	4.2	2.8	4.8
		10-hour labeled				
SlimIPL [54]	LS-960	4-gram + Transformer	5.3	7.9	5.5	9.0
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.4	13.3
DiscreteBERT [51]	LS-960	4-gram	5.3	13.2	5.9	14.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	3.8	9.1	4.3	9.5
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.4	4.8	2.6	4.9
HUBERT BASE	LS-960	4-gram	3.9	9.0	4.3	9.4
HUBERT LARGE	LL-60k	Transformer	2.2	4.3	2.4	4.6
HUBERT X-LARGE	LL-60k	Transformer	2.1	3.6	2.3	4.0
		100-hour labeled				
IPL [12]	LL-60k	4-gram + Transformer	3.19	6.14	3.72	7.11
SlimIPL [54]	LS-860	4-gram + Transformer	2.2	4.6	2.7	5.2
Noisy Student [61]	LS-860	LSTM	3.9	8.8	4.2	8.6
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.0	12.1
DiscreteBERT [51]	LS-960	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	2.7	7.9	3.4	8.0
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.9	4.0	2.0	4.0
HUBERT BASE	LS-960	4-gram	2.7	7.8	3.4	8.1
HUBERT LARGE	LL-60k	Transformer	1.8	3.7	2.1	3.9
HUBERT X-LARGE	LL-60k	Transformer	1.7	3.0	1.9	3.5

VIII. HuBERT(Result) – Self Training

BrAIn Lab. Brain and Artificial Intelligence Lab.

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		Superivsed				
Conformer L [62]	-	LSTM	-	-	1.9	3.9
		Self-Training				
IPL [12]	LL-60k	4-gram + Transformer	1.85	3.26	2.10	4.01
Noisy Student [61]	LV-60k	LSTM	1.6	3.4	1.7	3.4
		Pre-Training				
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.6	3.0	1.8	3.3
pre-trained Conformer XXL [40]	LL-60k	LSTM	1.5	3.0	1.5	3.1
		Pre-Training + Self-Training				
wav2vec 2.0 + self-training [63]	LL-60k	Transformer	1.1	2.7	1.5	3.1
pre-trained Conformer XXL + Noisy Student [40]	LL-60k	LSTM	1.3	2.6	1.4	2.6
		This work (Pre-Training)				
HUBERT LARGE	LL-60k	Transformer	1.5	3.0	1.9	3.3
HUBERT X-LARGE	LL-60k	Transformer	1.5	2.5	1.8	2.9

VIII. HuBERT(Result) – PNMI

BrAIn Lab.
Brain and Artificial Intelligence Lab.

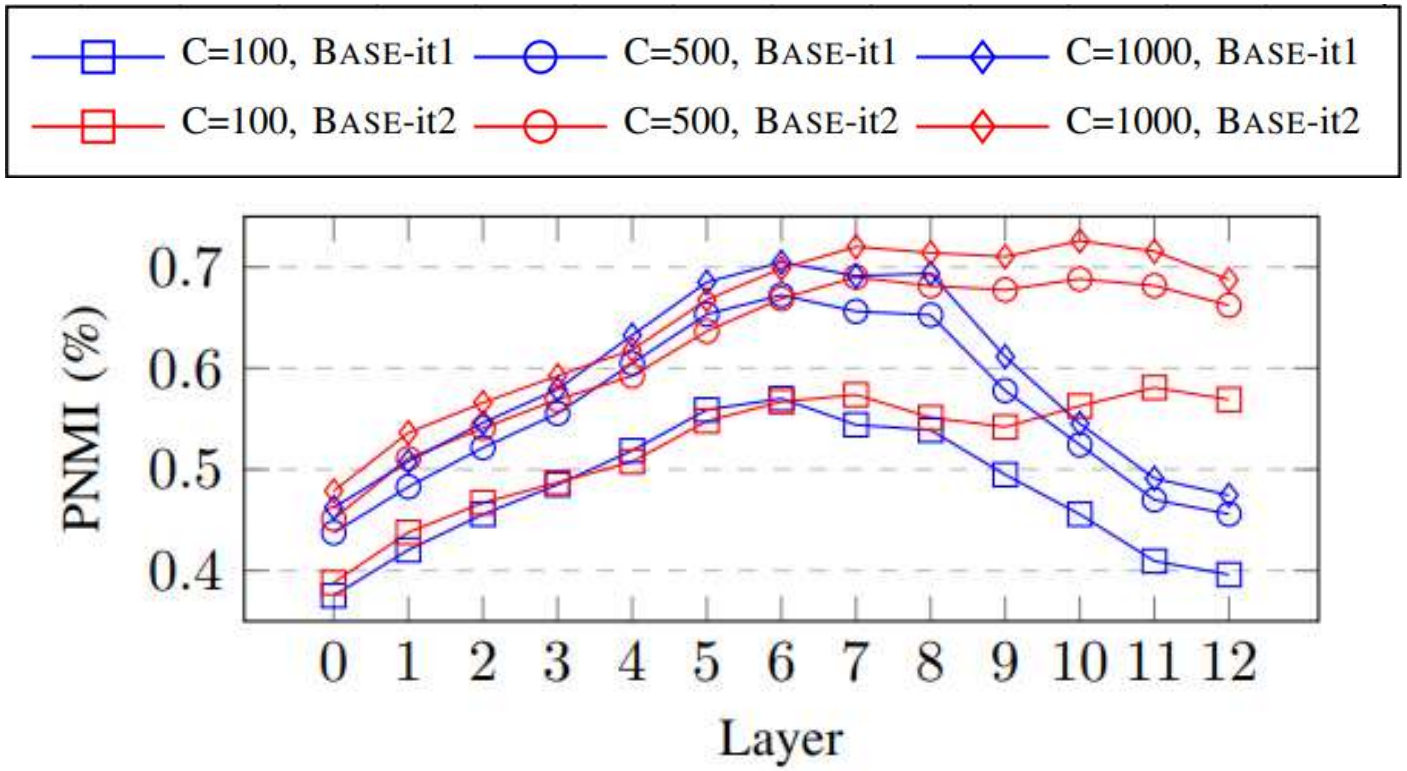
$$\frac{I(y;z)}{H(y)} = 1 - \frac{H(y|z)}{H(y)}$$

VIII. HuBERT(Result) – different feature

BrAIn Lab.
Brain and Artificial Intelligence Lab.

feature	C	PNMI (mean ± std) with K-means Training Size = 1h	10h	100h
MFCC	100	0.251 ± 0.001	0.253 ± 0.001	0.253 ± 0.001
	500	0.283 ± 0.001	0.285 ± 0.000	0.287 ± 0.001
BASE-it1-L6	100	0.563 ± 0.012	0.561 ± 0.012	0.575 ± 0.008
	500	0.680 ± 0.005	0.684 ± 0.003	0.686 ± 0.004

VIII. HuBERT(Result) – Quality



VIII. HuBERT(Result) – Effect of training objective

BrAIn Lab. Brain and Artificial Intelligence Lab.

teacher	C	PNMI	dev-other WER (%) $\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.0$
Chenone (supervised top-line)	8976	0.809	10.38	9.16	9.79
K-means on MFCC	50	0.227	18.68	31.07	94.60
	100	0.243	17.86	29.57	96.37
	500	0.276	18.40	33.42	97.66
K-means on BASE-it1-layer6	500	0.637	11.91	13.47	23.29
K-means on BASE-it2-layer9	500	0.704	10.75	11.59	13.79

VIII. HuBERT(Result) – Pre Training steps

BrAIn Lab.
Brain and Artificial Intelligence Lab.

teacher	C	dev-other WER (%) steps=100k	250k	400k	800k
K-means	50	18.68	13.65	12.40	11.82
	100	17.86	12.97	12.32	11.68
[51]	13.5k	26.6			

VIII. HuBERT(Result) – masking probability, batch size