HYPERCONNECT

Temporal Knowledge Distillation for On-device Audio Classification

Kwanghee Choi*, Martin Kersner*, Jacob Morton*, Buru Chang
Hyperconnect Inc.

2022.04.18

CONTENT

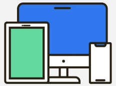
1. MOTIVATION

Summary
Our method

Our method can distill the temporal knowledge from attention weights of large transformer-based teacher models to on-device student models of various architectures.

Motivation #1
Why KD?

To improve the computationally restricted on-device models by transferring the knowledge of large models.



Motivation #2
Traditional KD

Temporal information, which is known to be beneficial in audio tasks, cannot be easily distilled when it is compressed into logits.

Motivation #3
Transformers KD

Where distilling self-attention maps will preserve temporal information, transformer-based KD methods are limited to those architectures only.

Our Goal

Our Target

Incorporate the **temporal knowledge** embedded in attention weights of large teacher models into on-device student models with various types of architectures.

Teacher Model

XLSR-wav2vec 2.0

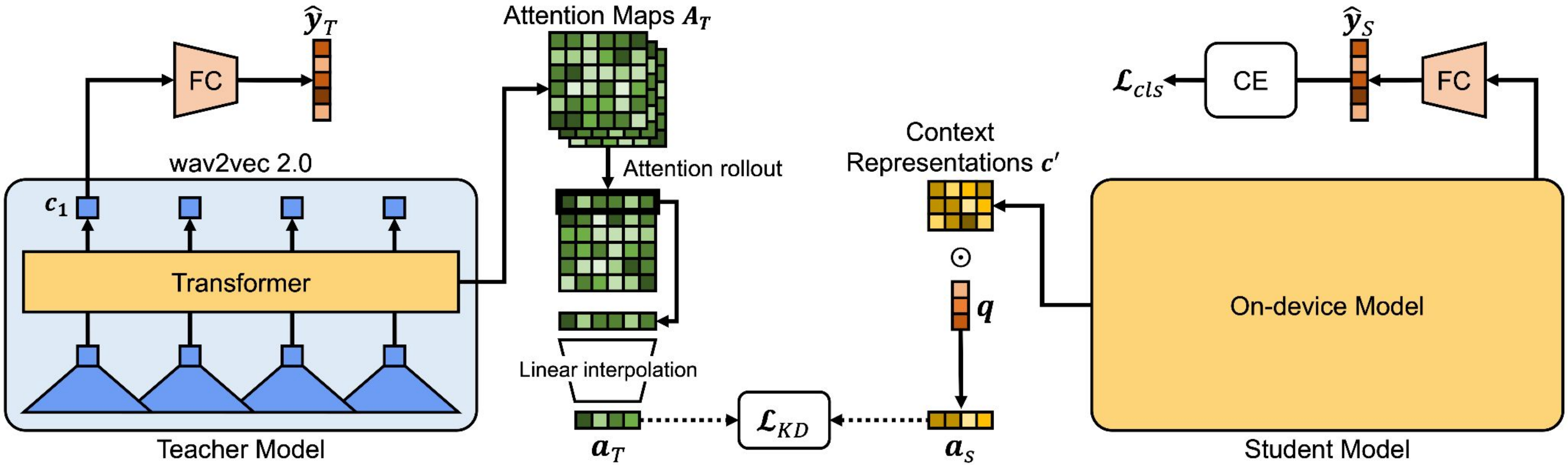
Large-scale transformer-based ASR model with state-of-the-art performance on multilingual ASR.

Student Model

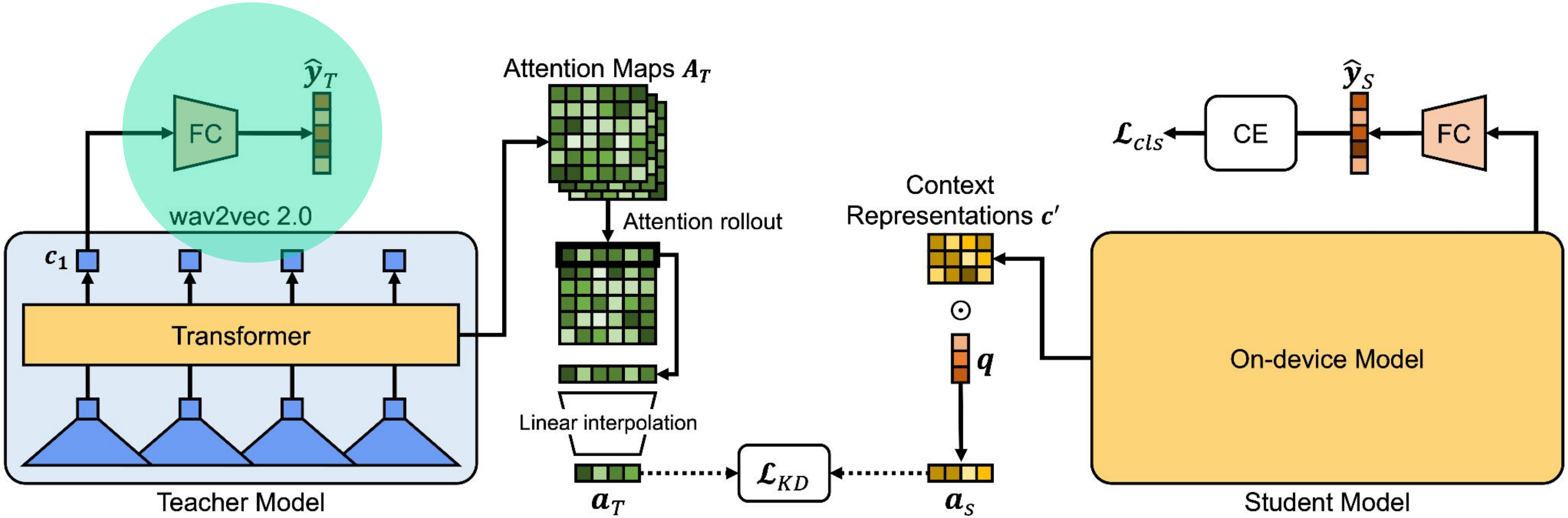
Five on-device audio classification models

RNN-based (LSTM-P), CNN-based (TC-ResNet), using both CNN and RNN (CRNN), containing attention mechanism (Att-RNN), and a multi-head variant (MHAtt-RNN)

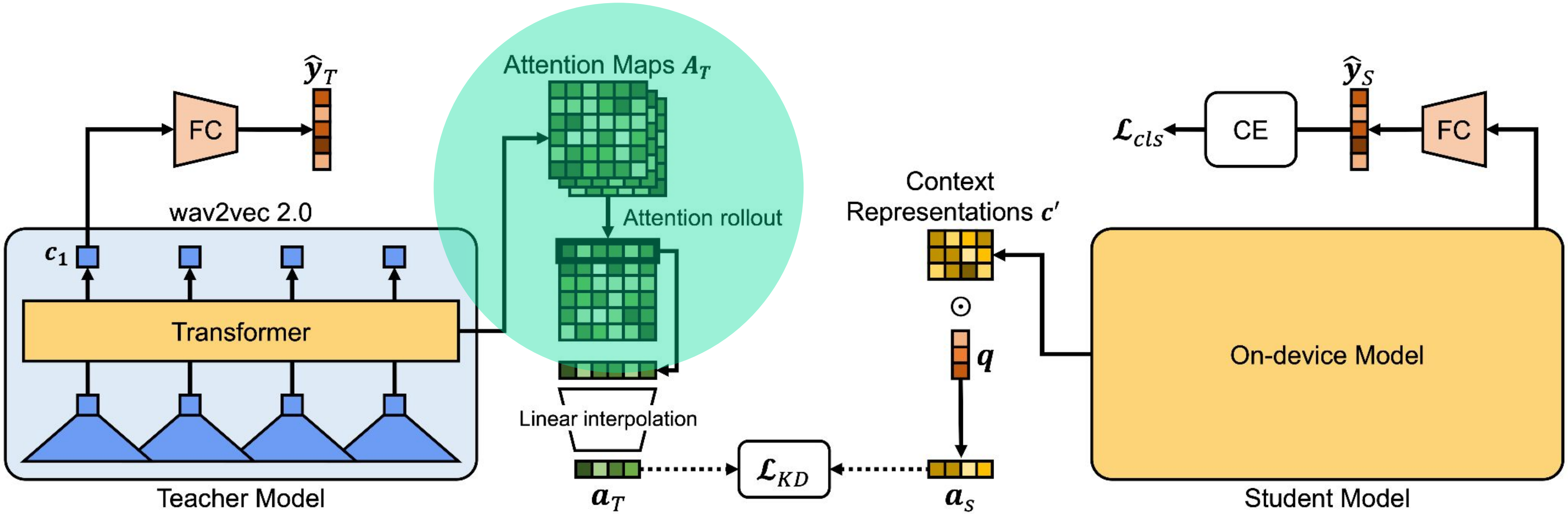
Brief Overview



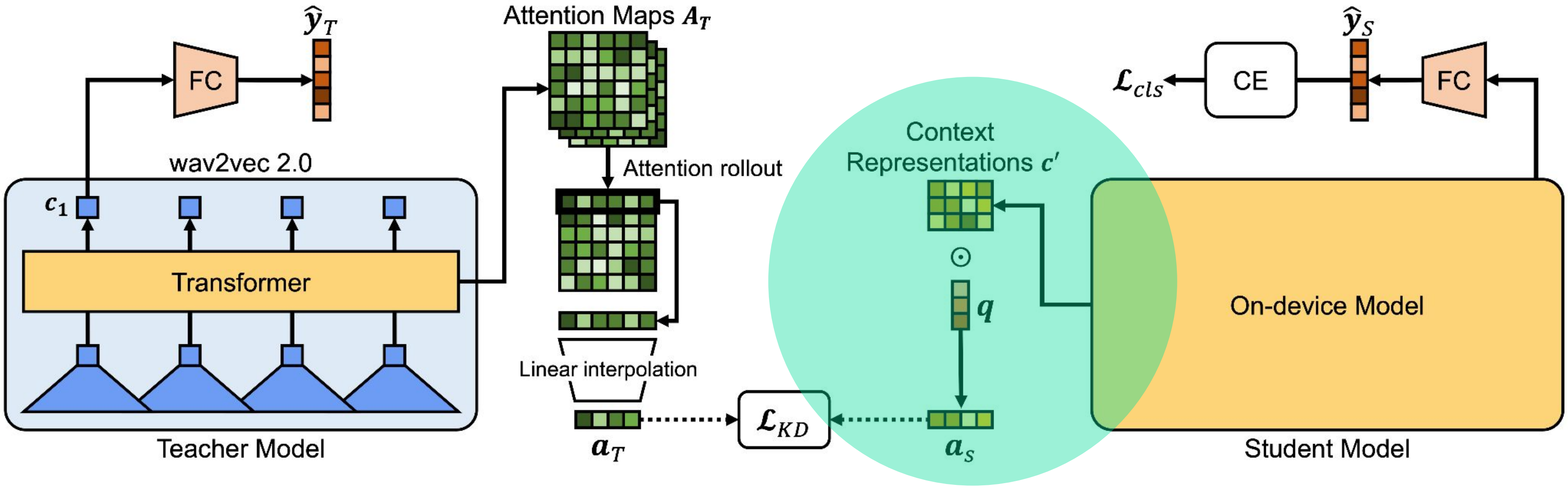
Training the Teacher Model



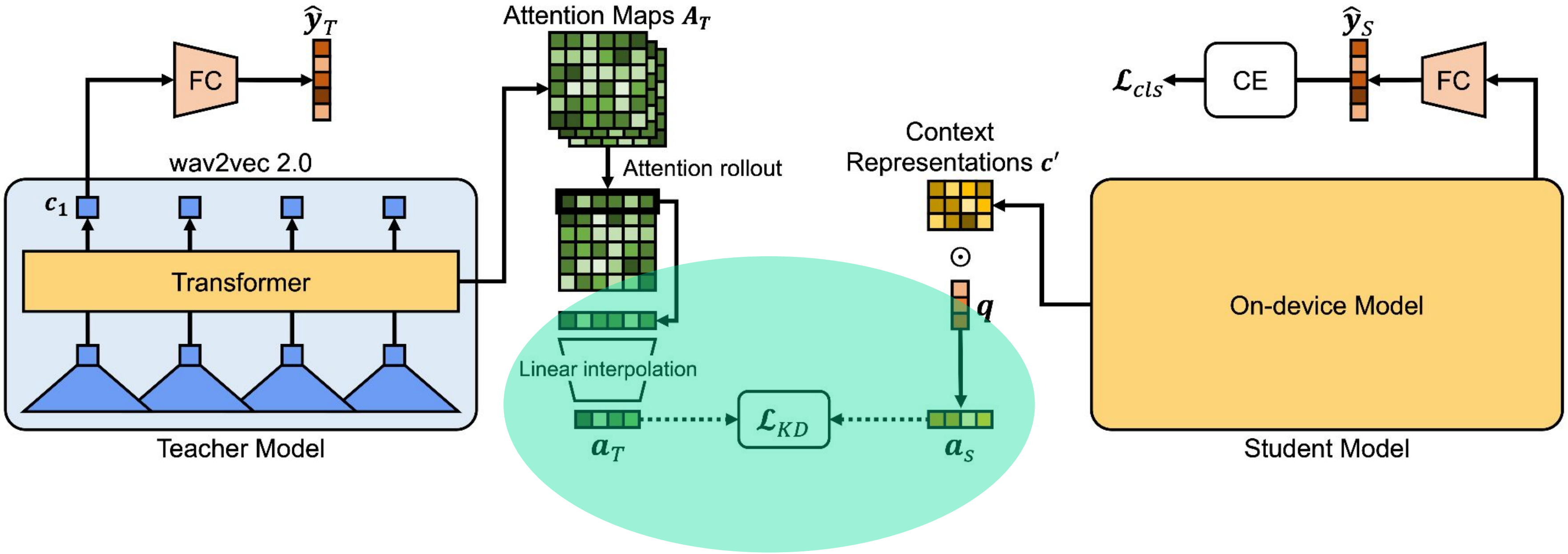
Extracting the Teacher Attention Maps



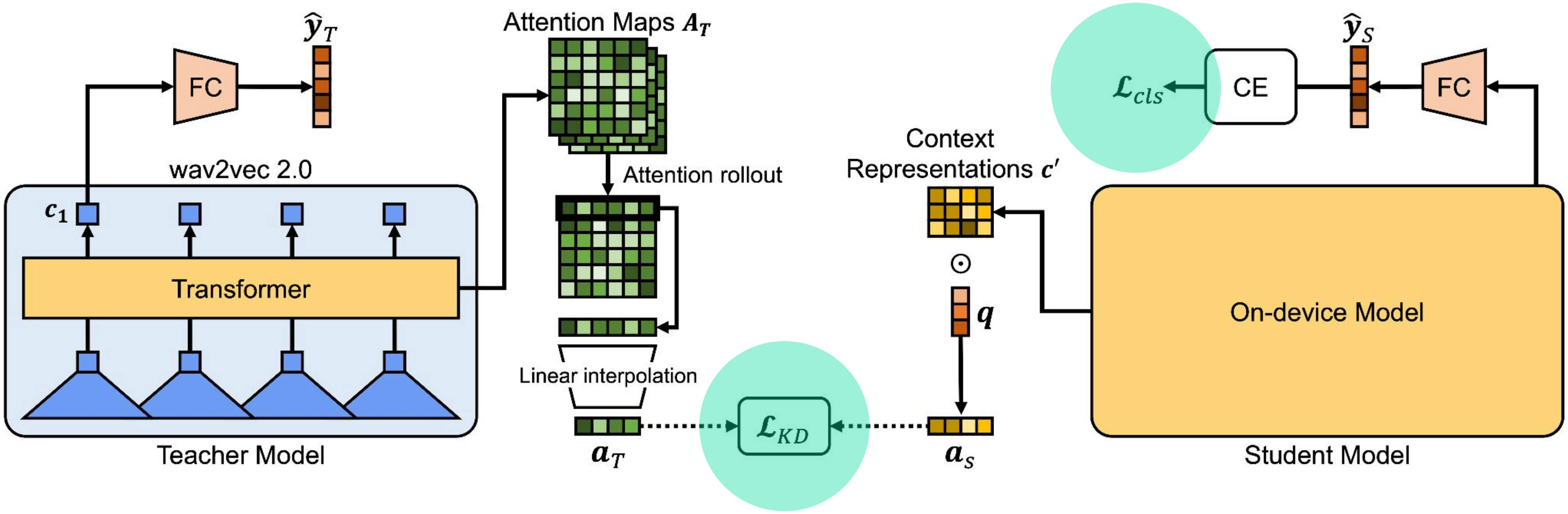
Extracting the Student Attention Maps



Minimizing the Distance between Attention Maps



Final Loss



$\hat{y}_T$
FC
Attention Maps A_T
wav2vec 2.0
c_1
Transformer
Attention rollout
Linear interpolation
a_T
$\mathcal{L}_{KD}$
Teacher Model
$\mathcal{L}_{cls}$
CE
$\hat{y}_S$
FC
Context Representations c'
$\odot$
q
a_S
On-device Model
Student Model

Performance Comparison

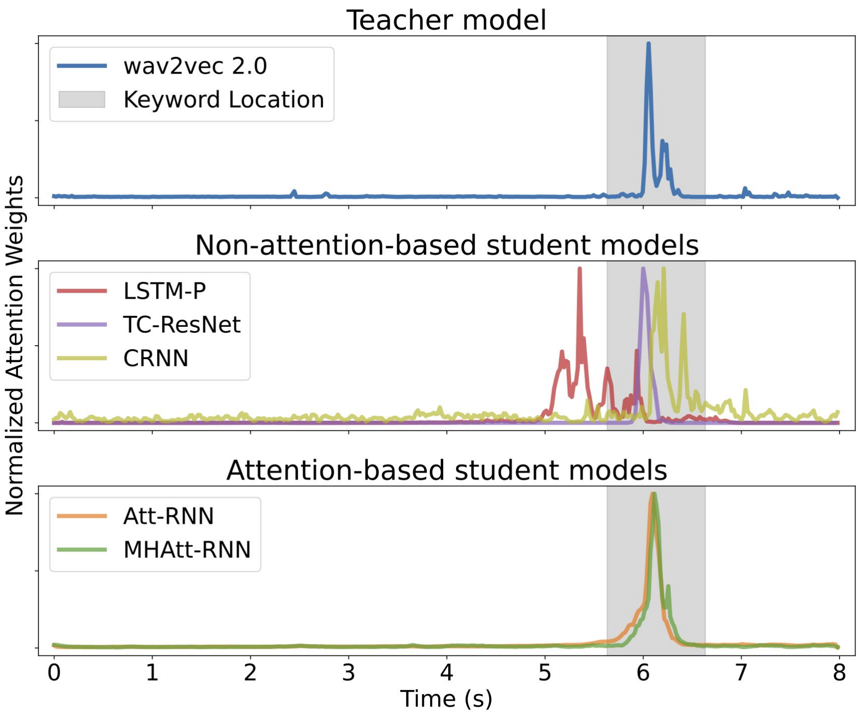
FSD50K

Model	wav2vec 2.0	LSTM-P	TC-ResNet	CRNN	Att-RNN	MHAtt-RNN
w/o KD	0.5498	0.1141	0.1814	0.2789	0.2856	0.2647
w/ KD	N/A	**0.1300**	**0.1951**	**0.3053**	**0.3471**	**0.3317**

Table 1:mAP performance comparison on the FSD50K dataset, a real-world multi-label audio event detection dataset, with and without applying our KD loss.

Comparing Attention Maps

Noisy Speech Commands v2



Performance Comparison

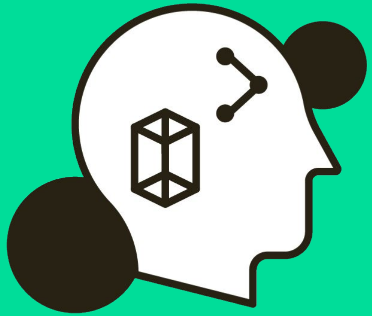
Noisy Speech Commands v2

L	Model	wav2vec 2.0	LSTM-P	TC-ResNet	CRNN	Att-RNN	MHAtt-RNN
2s	w/o KD	90.59	88.73	87.77	89.96	89.88	89.75
	w/ KD	N/A	**89.31**	**88.08**	**90.06**	**91.67**	**91.75**
4s	w/o KD	91.22	85.19	87.60	89.69	90.65	91.19
	w/ KD	N/A	**89.08**	**88.33**	**90.21**	**91.98**	**92.12**
6s	w/o KD	90.93	45.27	86.00	88.58	90.88	90.58
	w/ KD	N/A	**85.58**	**86.85**	**89.88**	**91.19**	**91.67**
8s	w/o KD	90.95	78.44	77.81	88.94	88.81	88.33
	w/ KD	N/A	**82.19**	**85.79**	**89.79**	**90.98**	**91.79**

Table 2:Test accuracy (%) performance comparison on the Noisy Speech Commands v2 dataset.

Conclusions and Takeaways

1. One can distill knowledge between different architectures.
2. Attention matters for audio classification tasks.
3. Large transformer models can automatically attend to important locations.



Thank You For Listening!

Temporal Knowledge Distillation for On-device Audio Classification

Contacts

Kwanghee Choi kwanghee.choi@hpcnt.com
juice500@sogang.ac.kr
Martin Kersner kersner@hpcnt.com
Jacob Morton jake.m@hpcnt.com
Buru Chang buru.chang@hpcnt.com

Links

https://arxiv.org/abs/2110.14131
https://hyperconnect.github.io/2022/03/29/temporal-kd-ondevice-audio.html