Audio Embeddings as Teachers for Music Classification

Yiwei Ding, Alexander Lerch

Georgia Tech | Center for Music Technology, College of Design

Introduction

Our paper targets at low-resource music classification with limited data and small models.

Different ways of knowledge transfer

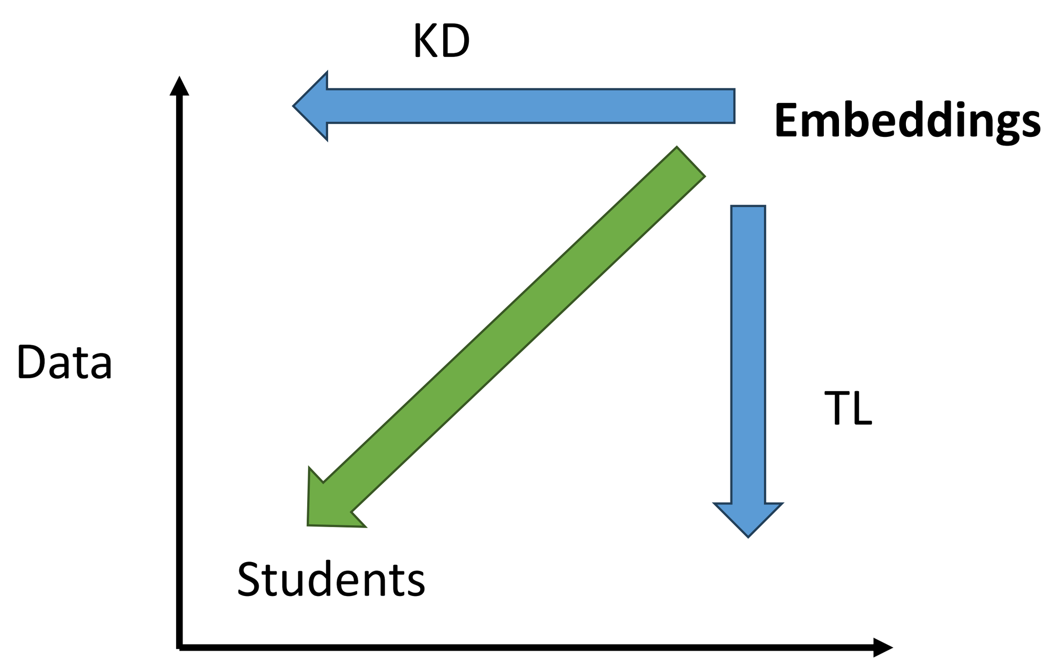

Figure: Different ways of knowledge transfer. Blue arrows indicate knowledge distillation or transfer learning. The green arrow indicate our approach.

Methods

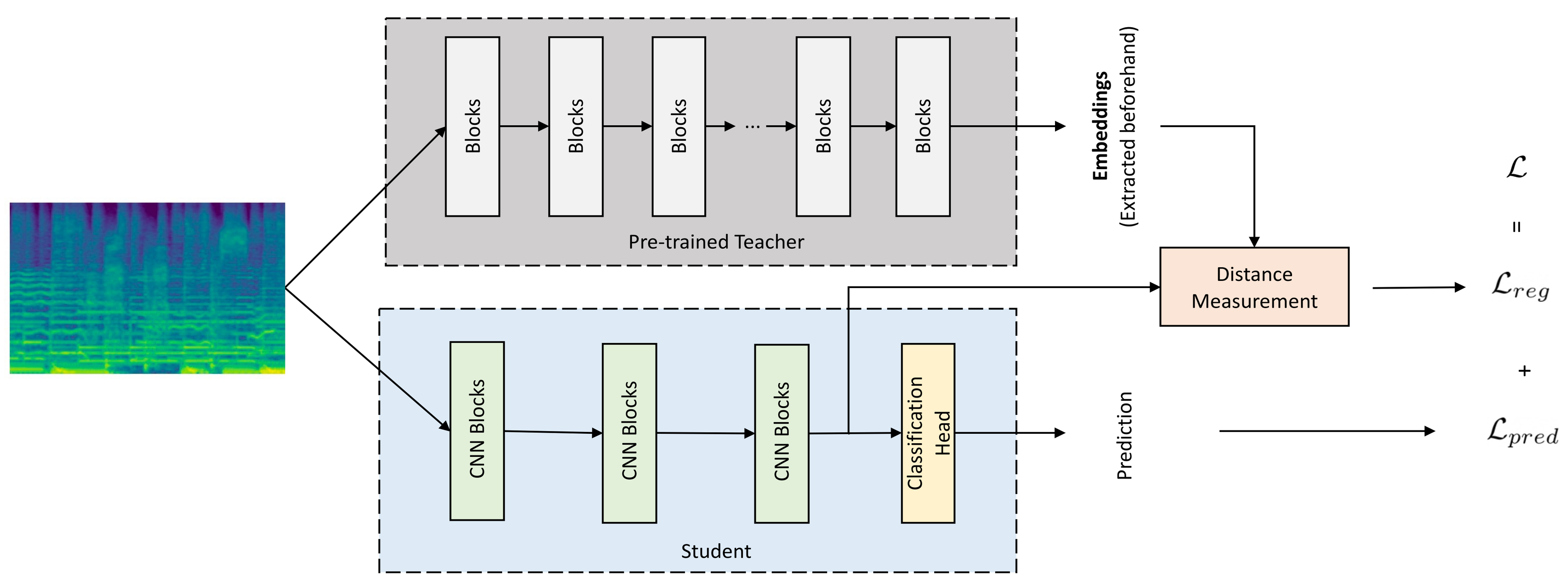

Figure: Overall pipeline of using audio embeddings as teachers. During inference , only the bottom part in blue is used.

During training

- Weighted loss:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{reg}}$$

- Stage of regularization: penultimate layer only or all stages
- Distance measures
  1) cosine distance difference
  2) distance correlation

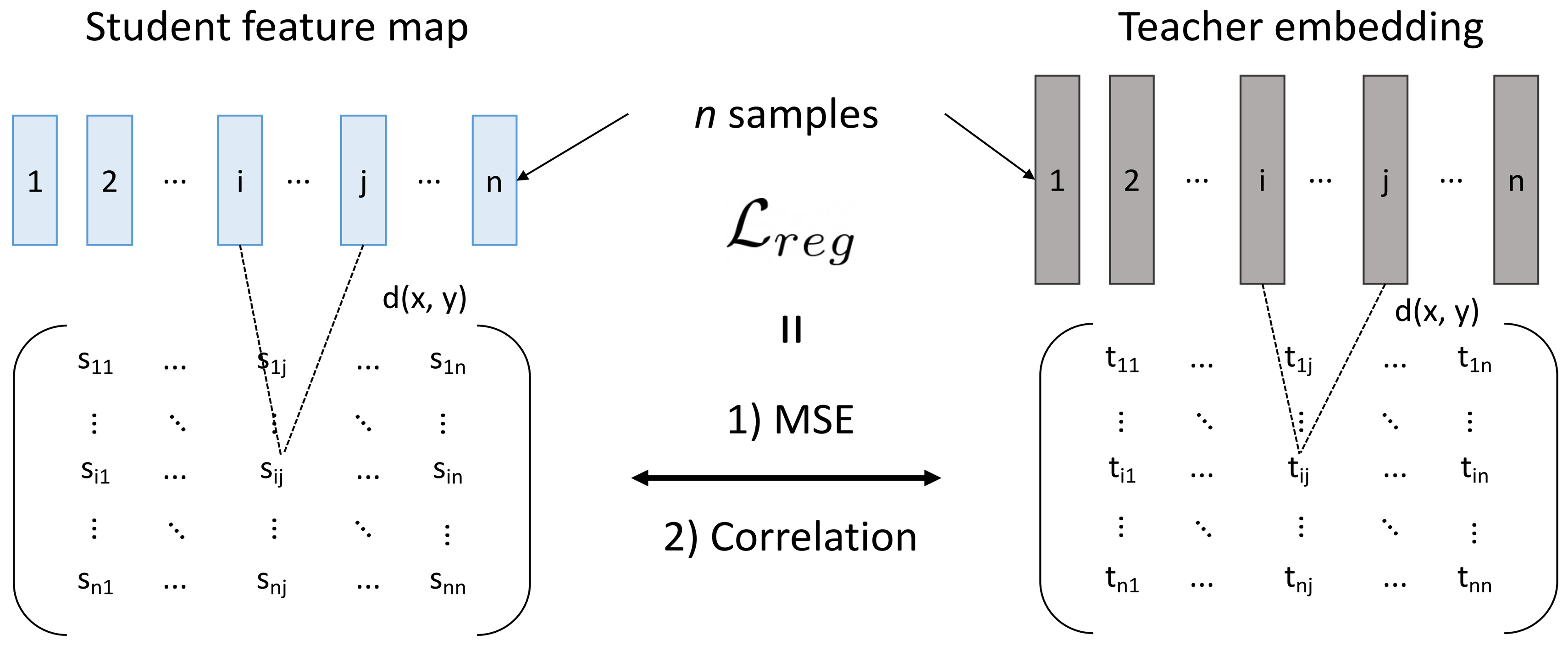

Figure: Illustration of distance measures.

Systems for comparison

Baseline: student without regularization

$\text{Teacher}_{\text{LR}}$: teacher embeddings + logistic regression

KD: traditional knowledge distillation with soft targets

$\text{EAsT}_{\text{Cos-Diff}}$: using cosine distance difference to compute loss

$\text{EAsT}_{\text{Final}}$ and EAsT_{All} : using distance correlation to compute loss

EAsT_{KD} : combine our method with KD

Experimental Setup

We test the effectiveness of our methods on two different tasks.

- Musical Instrument Classification with OpenMIC
  - Baseline model: ResNet with regularized receptive field (CP-ResNet)
  - Evaluated by mean Average Precision (mAP) and macro F1-score
- Music Auto-Tagging with MagnaTagATune
  - Baseline model: Mobile FCN
  - Evaluated by mean Average Precision (mAP) and ROC-AUC

Four pre-trained embeddings are used: VGGish, OpenL3, PaSST and PANNs.

Results

OpenMIC	VGGish		OpenL3		PaSST		PANNs	
	mAP	F1	mAP	F1	mAP	F1	mAP	F1
CP ResNet	mAP = .819 / F1 = .809							
$\text{Teacher}_{\text{LR}}$	.803	.799	.803	.798	**.858**	**.837**	.853	**.834**
KD	.829	.820	.823	.813	.851	.834	.848	.823
$\text{EAsT}_{\text{Cos-Diff}}$	.838	.824	**.838**	.820	.837	.822	.836	.814
$\text{EAsT}_{\text{Final}}$	**.842**	**.828**	.835	**.822**	.847	.830	.849	.828
EAsT_{All}	.836	.823	.835	**.822**	.845	.827	.845	.827
EAsT_{KD}	.836	.825	.836	.821	.852	.834	**.857**	.831

MTAT	VGGish		OpenL3		PaSST		PANNs	
	mAP	AUC	mAP	AUC	mAP	AUC	mAP	AUC
Mobile FCN	mAP = .437 / AUC = .905							
$\text{Teacher}_{\text{LR}}$	.433	.903	.403	.890	**.473**	**.917**	**.460**	.911
KD	.447	.911	.439	.907	.454	.912	.448	.909
$\text{EAsT}_{\text{Cos-Diff}}$	.446	.906	.438	.907	.453	.912	.453	.911
$\text{EAsT}_{\text{Final}}$	.454	**.912**	.447	.910	.459	.912	.449	.909
EAsT_{All}	**.455**	.911	**.452**	**.911**	.458	.913	.457	.911
EAsT_{KD}	.441	.908	.437	.904	.461	.915	.459	**.912**

Table: Results on the OpenMIC dataset (top) and MagnaTagATune dataset (bottom). Best performances are in bold, and best results excluding the teachers are underlined.

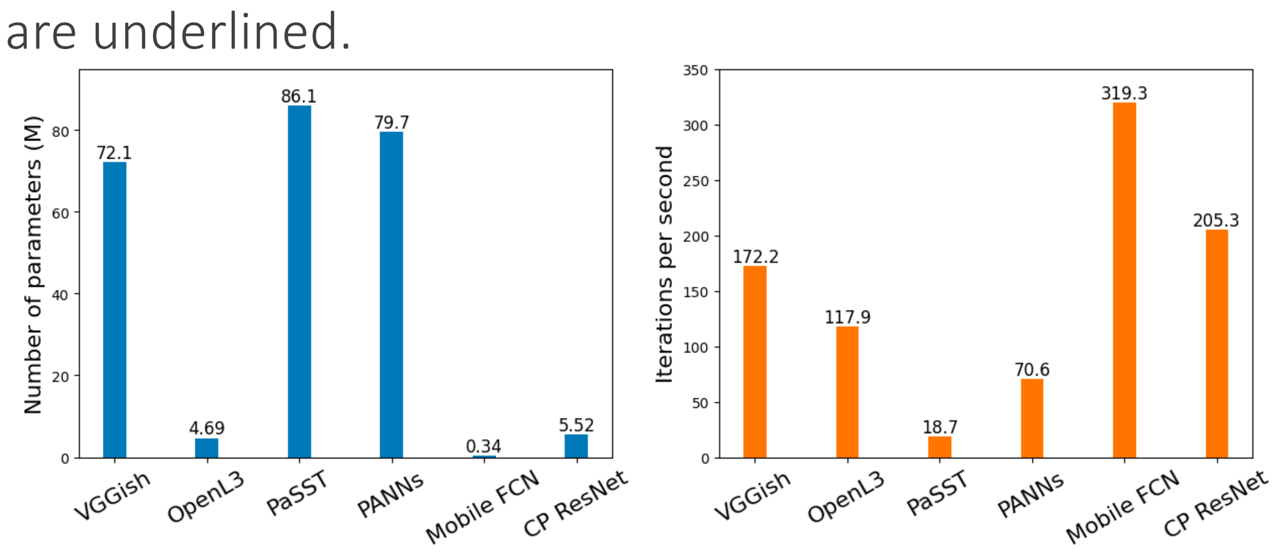

Fig: Comparison of the model complexity.

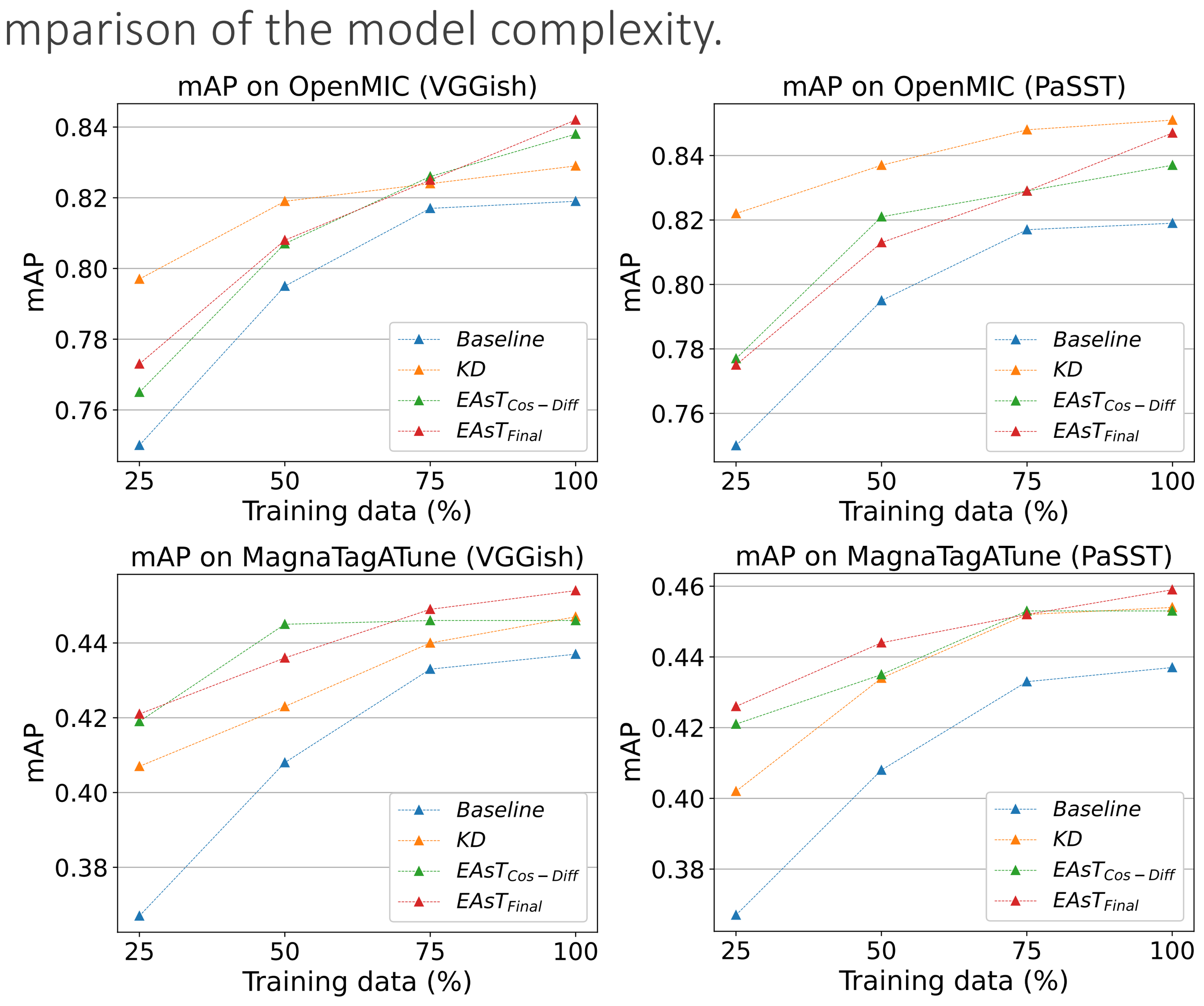

Fig: Results with limited training data on OpenMIC and MagnaTagATune

References

Y.-N. Hung and A. Lerch, "Feature-Informed Embedding Space Regularization for Audio Classification," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 419–423.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

CONTACT

Yiwei Ding
Music Informatics Group
Center for Music Technology
yding402@gatech.edu