Knowledge Distillation From Transformers for Low-Complexity Acoustic Scene Classification

Florian Schmid, Shahed Masoudian, Khaled Koutini and Gerhard Widmer

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

System for ASC Task of the DCASE'22 Challenge [1]

Key Ingredients

- Main Difficulties
  1. Low-complexity Constraints
  2. Cross-Device Generalization

- Tackle 1. with Transformer-to-CNN Cross-Model **Knowledge Distillation**
- Tackle 2. with **Freq-MixStyle [2]**

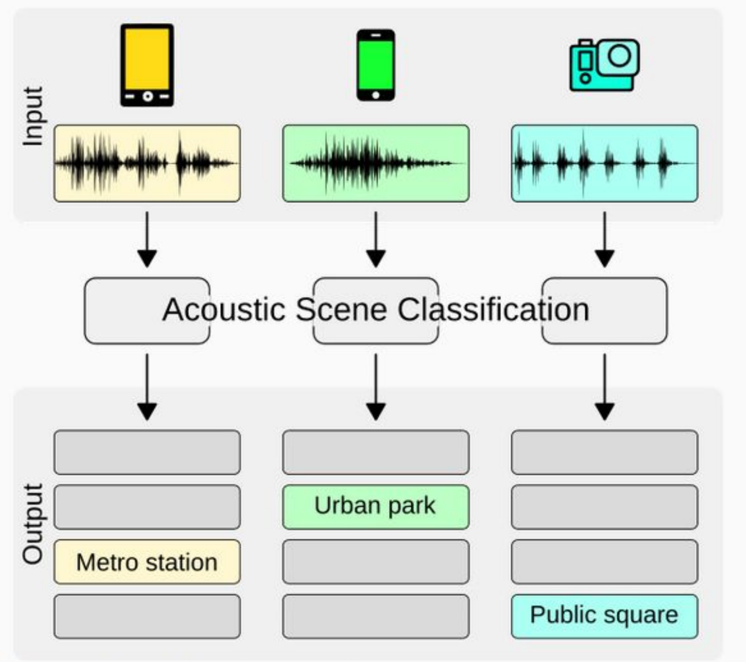
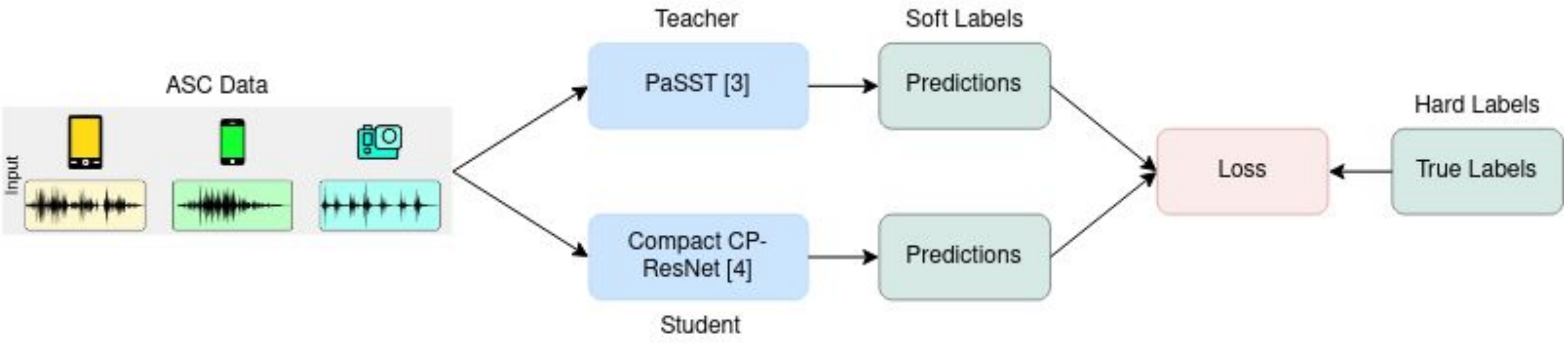


Figure 1: Overview of acoustic scene classification system.

JOHANNES KEPLER UNIVERSITY LINZ

Knowledge Distillation From Transformers
Setup

- **Teacher**: Patchout FaSt Spectrogram Transformer (PaSST) [3]
- **Student**: Compact Version of CP-ResNet [4]
- **Loss**: weighted combination of Hard Label and Distillation Loss



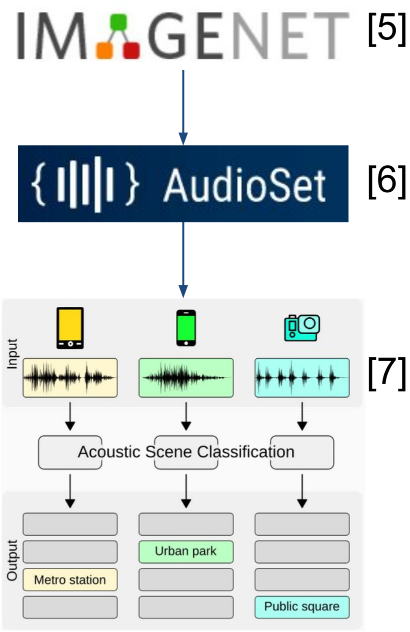
JOHANNES KEPLER UNIVERSITY LINZ

Knowledge Distillation From Transformers

Teacher and Student

- **Teacher**: PaSST [3]
  - Audio Spectrogram Transformer
  - well-performing but large and complex
- **Student**: Compact CP_ResNet [4]
  - receptive field regularized
  - reduced width and depth
  - grouping
  - (frequency-)pooling

PaSST Training Steps

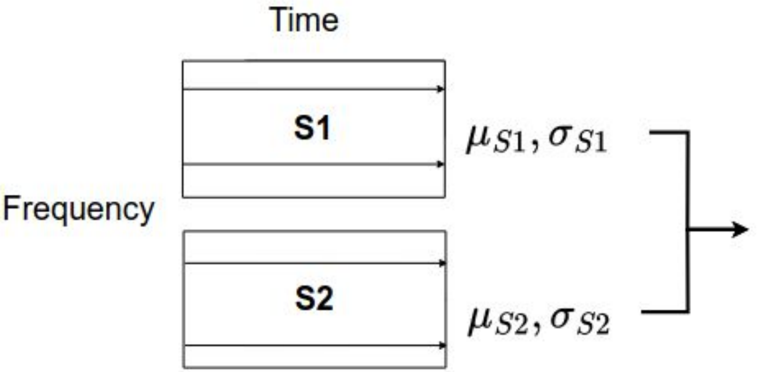


Freq-MixStyle [2]

Method

- Method to improve **Cross-Device Generalization** by mixing frequency statistics
- Vary device-style, retain scene label

$$\text{Freq} - \text{MixStyle(S1)} = \sigma_{mix} * \frac{S1 - \mu_{S1}}{\sigma_{S1}} + \mu_{mix}$$



$$\sigma_{mix} = \lambda\sigma_{S1} + (1-\lambda)\sigma_{S2}$$
$$\mu_{mix} = \lambda\mu_{S1} + (1-\lambda)\mu_{S2}$$

Results
Key Findings

- Transformer-to-CNN Knowledge Distillation improves performance on ASC substantially
- Having a better teacher (KD variations) improves results slightly
- Freq-MixStyle improves generalization to unseen devices for student and teacher
- Ensembling PaSST models trained with different Freq-MixStyle configurations improves results on unseen devices

References

[1] I. Martìn-Moratò, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, “Low complexity acoustic scene classification in dcase 2022 challenge,” 2022.

[2] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, “Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification,” in Interspeech, 2022.

[3] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, “Efficient Training of Audio Transformers with Patchout,” in Interspeech, 2022.

[4] K. Koutini, H. Eghbal-zadeh, and G. Widmer, “Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks,” IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in CVPR, 2009.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in ICASSP, 2017.

[7] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions,” in DCASE 2020 Workshop, 2020, pp. 56–60.

Appendix

Results
Student Model

Method	Configuration					Test Accuracy (%)				Log Loss
	Mixup	Freq-MixStyle	Temp	Teach. Type	AS	Real	Sim	Unseen	Overall	Overall
Student Baseline	✗	✗	-	No	✗	61.97	50.10	40.71	50.92	1.5822
	✓	✗	-	No	✗	62.70	52.48	42.99	52.72	1.4161
	✗	✓	-	No	✗	63.89	56.00	49.98	56.62	1.2344
KD Baseline	✗	✗	H	Single	✗	66.21	57.35	50.14	57.89	1.1316
	✓	✗	H	Single	✗	66.43	58.31	51.32	58.68	1.1063
	✗	✓	L	Single	✗	64.36	58.36	55.12	59.28	1.1431
KD Ensemble	✓	✗	H	Ensemble	✗	66.30	58.65	52.06	59.00	1.0888
	✗	✓	L	Ensemble	✗	64.74	58.59	55.14	59.49	1.1322
KD Superior Teacher	✓	✗	H	Superior	✗	66.53	58.54	51.89	58.98	1.1033
	✗	✓	L	Superior	✗	64.73	58.60	**55.15**	59.49	1.1313
KD Audioset	✓	✗	M	Single	✓	**66.54**	59.09	52.49	59.37	1.0906
	✗	✓	M	Single	✓	64.99	58.50	54.43	59.30	1.0939
	✓	✗	M	Ensemble	✓	66.35	**59.95**	52.99	**59.76**	**1.0794**

JOHANNES KEPLER UNIVERSITY LINZ

Results

PaSST Downstream Training on *TAU Urban Acoustic Scenes 2022 dev. dataset [7]*

Method	**Real**	**Sim**	**Unseen**	**Overall**
PaSST Baseline	67.63	57.66	56.11	60.46
+ Mixup	**67.85**	58.45	57.10	61.13
+ Freq-MixStyle	67.68	**58.97**	**58.22**	**61.64**
Ensemble	68.62	60.11	59.73	62.82

Knowledge Distillation From Transformers

KD Variations

- PaSST Ensemble: Five PaSST [3] models trained with different Freq-MixStyle [2] configurations
- Distillation on Out-Of-Domain-Dataset: KD on AudioSet [6]
- Predictions on extended audio sequences: reassembled 10-second audio snippets

JOHANNES KEPLER UNIVERSITY LINZ

Knowledge Distillation From Transformers

Student Architecture: Compact CP_ResNet [4]

WIDTH	GROUPING	BLOCK	CONFIG
W		INPUT	5×5, P
W	1	R	3×3, 1×1, P_f
W	1	R	3×3, 3×3, P_f
		LINEAR	$W \rightarrow 2\dot{W}$
$2 \times W$	2	R	3×3, 3×3
		LINEAR	$2W \rightarrow 4\dot{W}$
$4 \times W - C$	1	R	3×3, 1×1
CLASSIFIER $4 \times W - C \rightarrow 10$ CLASSES			
GLOBAL MEAN POOLING			

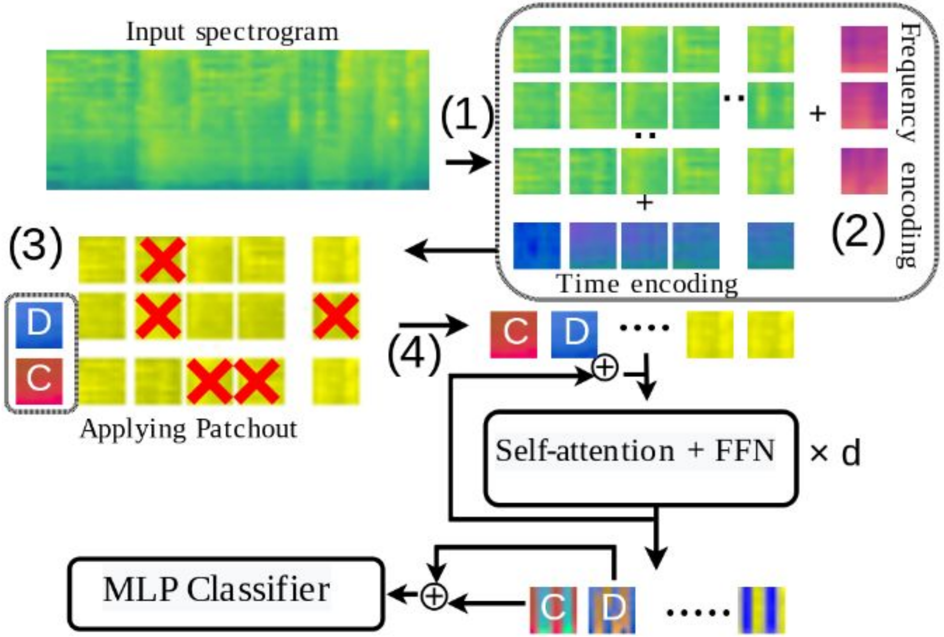
P: 2×2 MAX POOLING.

P_f: 2×1 MAX POOLING OVER THE FREQUENCY DIMENSION.

R: RESIDUAL, THE INPUT IS ADDED TO THE OUTPUT

Knowledge Distillation From Transformers

Patchout faSt Spectrogram Transformer (PaSST) [3]



Knowledge Distillation From Transformers

Student Loss Calculation

$$Loss = CE(y, \delta(z_S)) + \lambda \mathrm{KL}(\delta(z_T/\tau)||\delta(z_S/\tau))$$

z_T... Teacher Logits

z_S... Student Logits

τ... Temperature

δ... Softmax Activation

CE... Cross-Entropy Loss

KL... Kullback-Leibler Divergence

λ... Distillation Loss Weight

JOHANNES KEPLER UNIVERSITY LINZ

Freq-MixStyle [1]
Intuition

- Frequency Fingerprints across Labels more stable than across Devices