FONDAZIONE BRUNO KESSLER

E3DA ENERGY EFFICIENT EMBEDDED DIGITAL ARCHITECTURES

tinyCLAP 👏: towards embedded foundational models

Francesco Paissan
Energy Efficient Embedded Digital Architectures
Fondazione Bruno Kessler

Foundational models

"any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks"

LLAMA: language model

Whisper: automatic speech recognition

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

CLIP: image retrieval

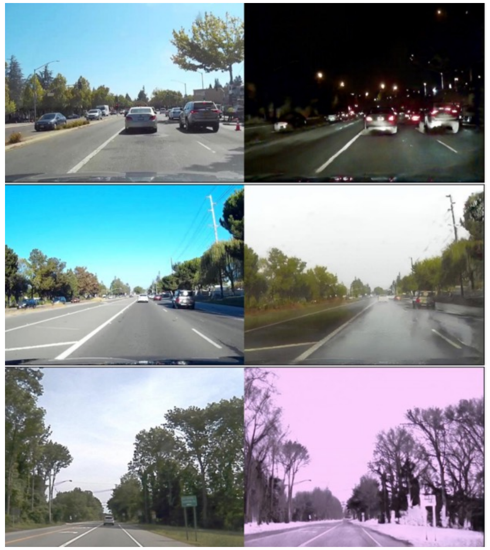
Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

"On the Opportunities and Risks of Foundation Models", Bommasani et al., 2021

Why is this relevant for tinyML?

- we all like a good model, regardless of the task;
- tinyML deals with dynamic environments;



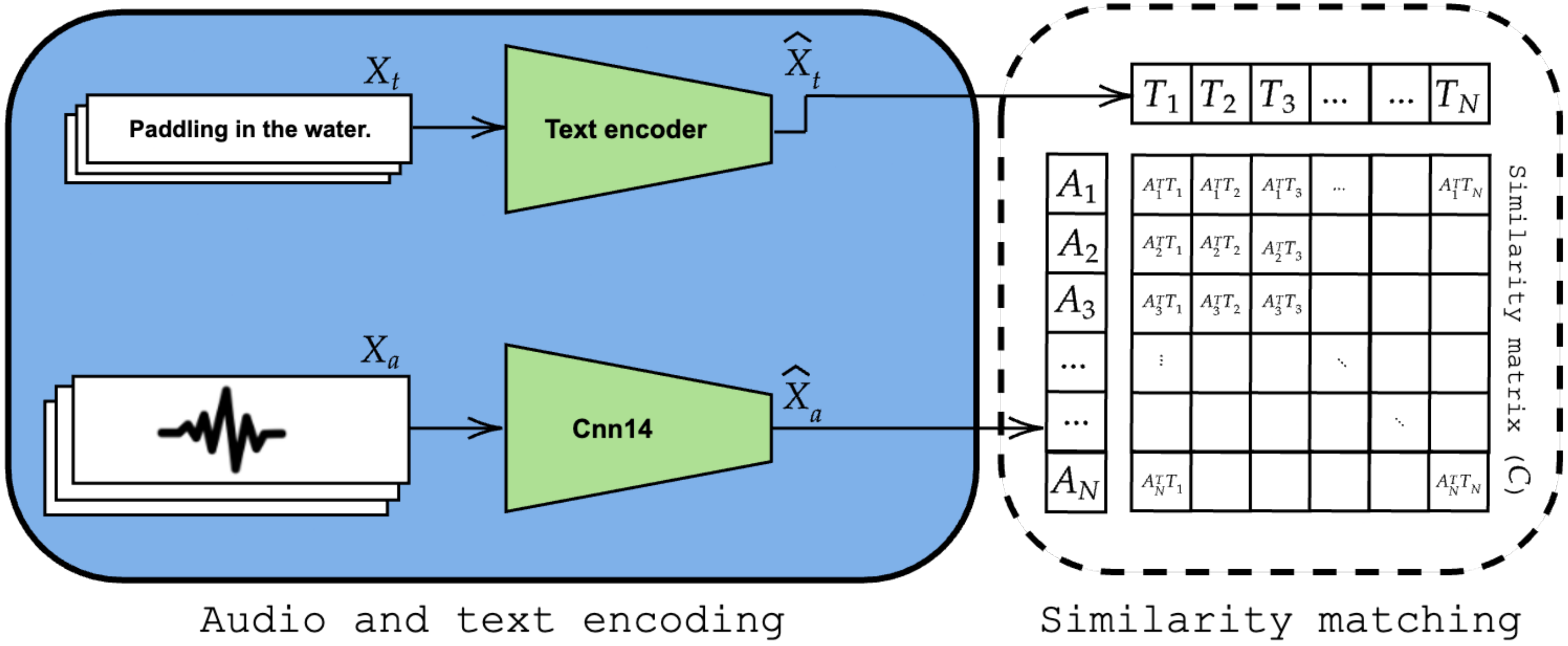
Domain adaptation

Class-incremental scenarios

"Universal Domain Adaptation from Foundation Models: A Baseline Study", Deng and Jia, 2023

Contrastive Language-Audio Pretraining

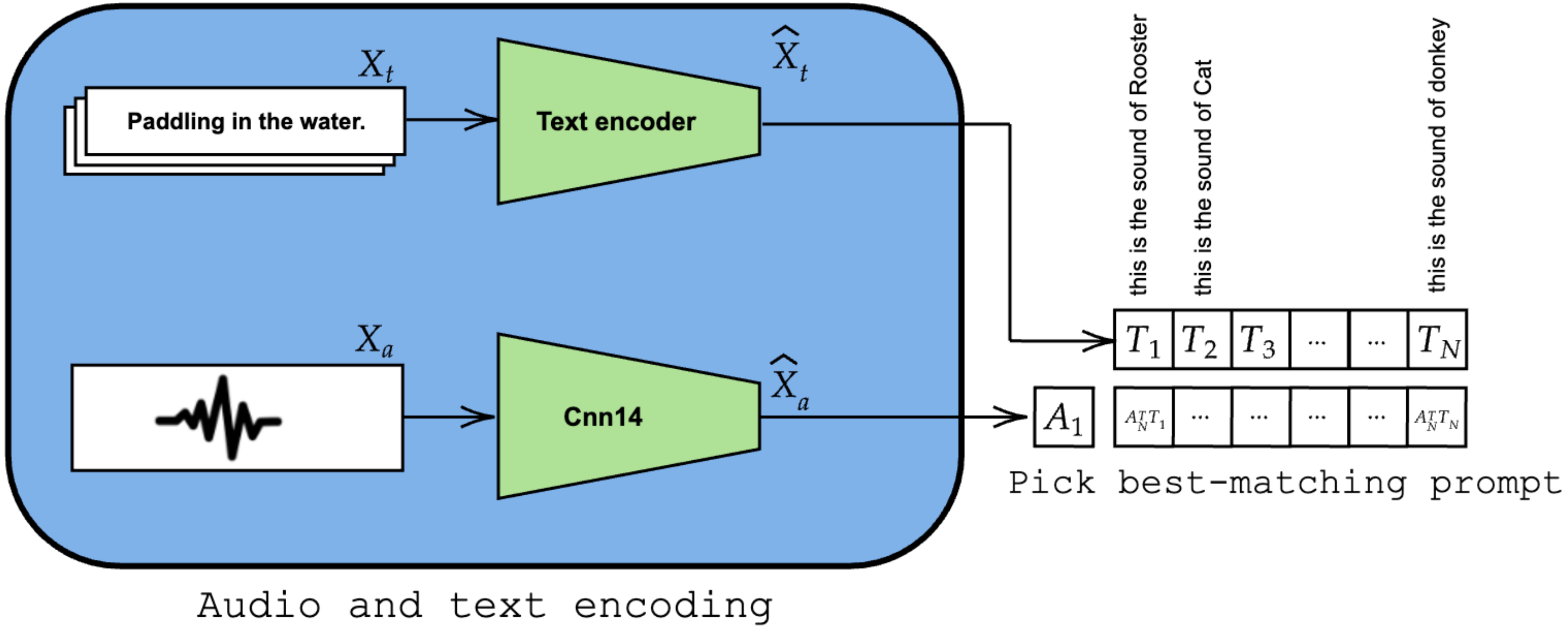
- learns a similarity score between audio and text modalities;
- enables zero-shot classification;



"CLAP: Learning Audio Concepts From Natural Language Supervision", Elizalde et al. 2022

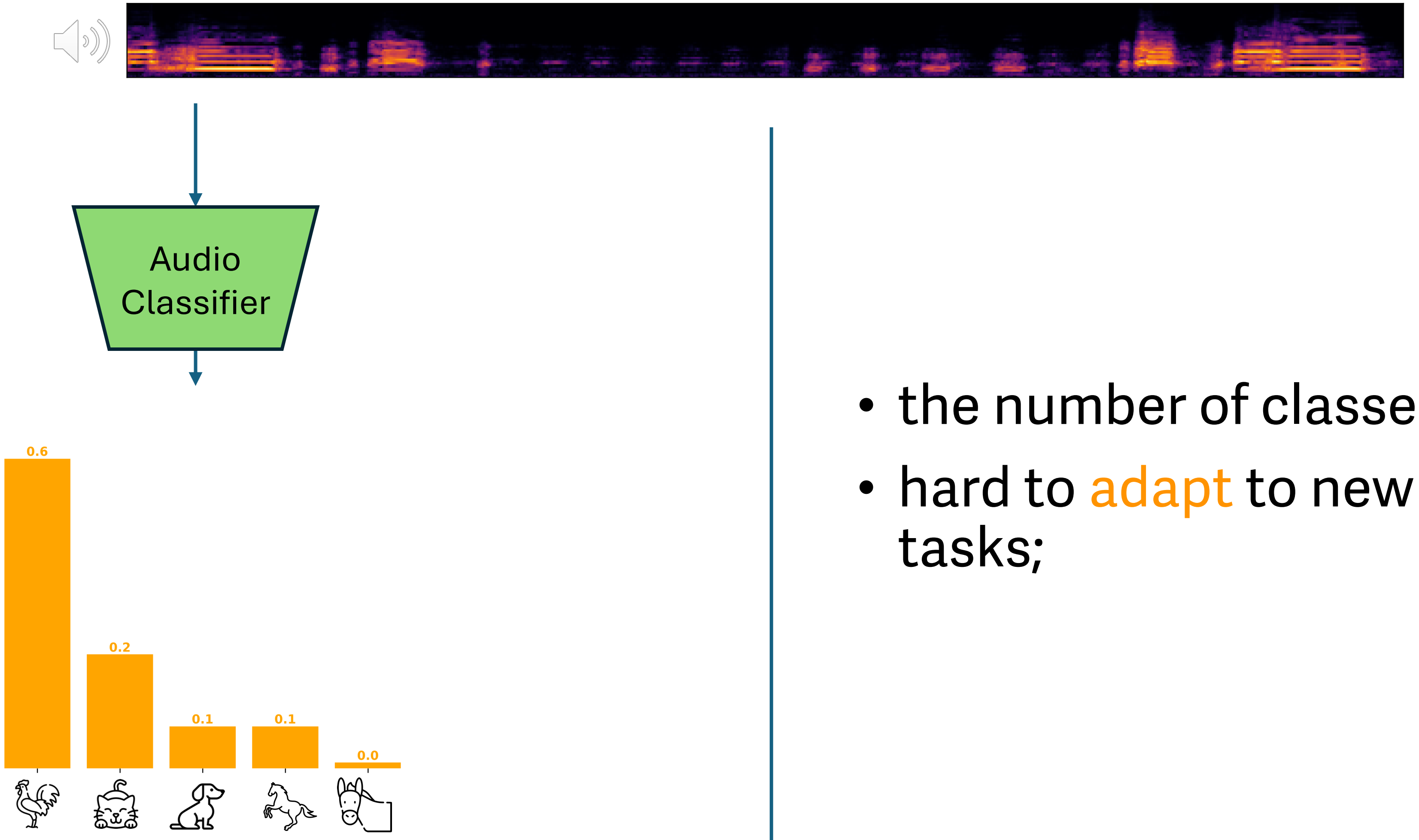
CLAP: how to do classification

- to each audio input, assign the most similar text prompt;
- CLAP encoders can be finetuned to boost (supervised) performance;



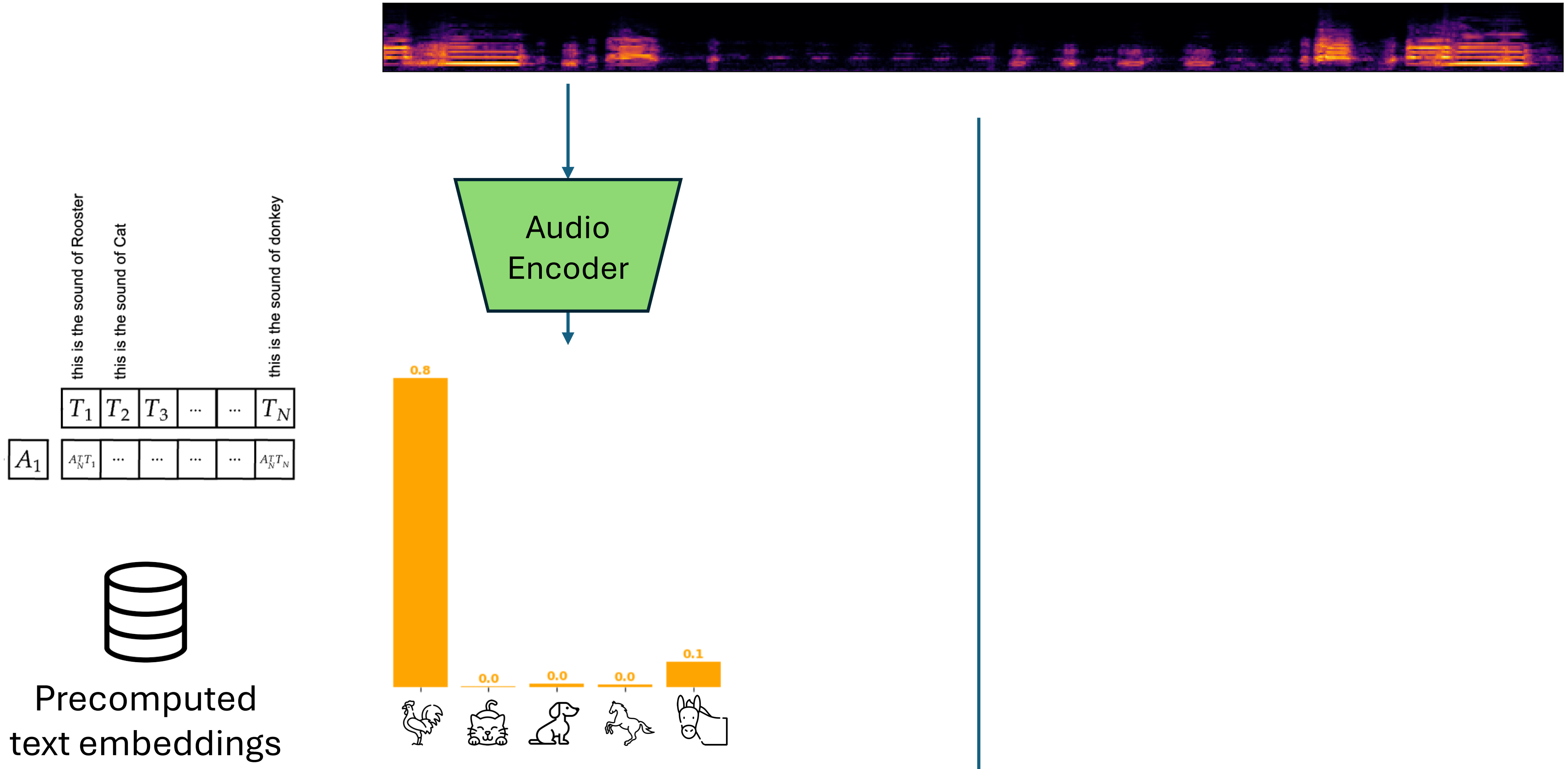
"CLAP: Learning Audio Concepts From Natural Language Supervision", Elizalde et al. 2022

Traditional supervised classification

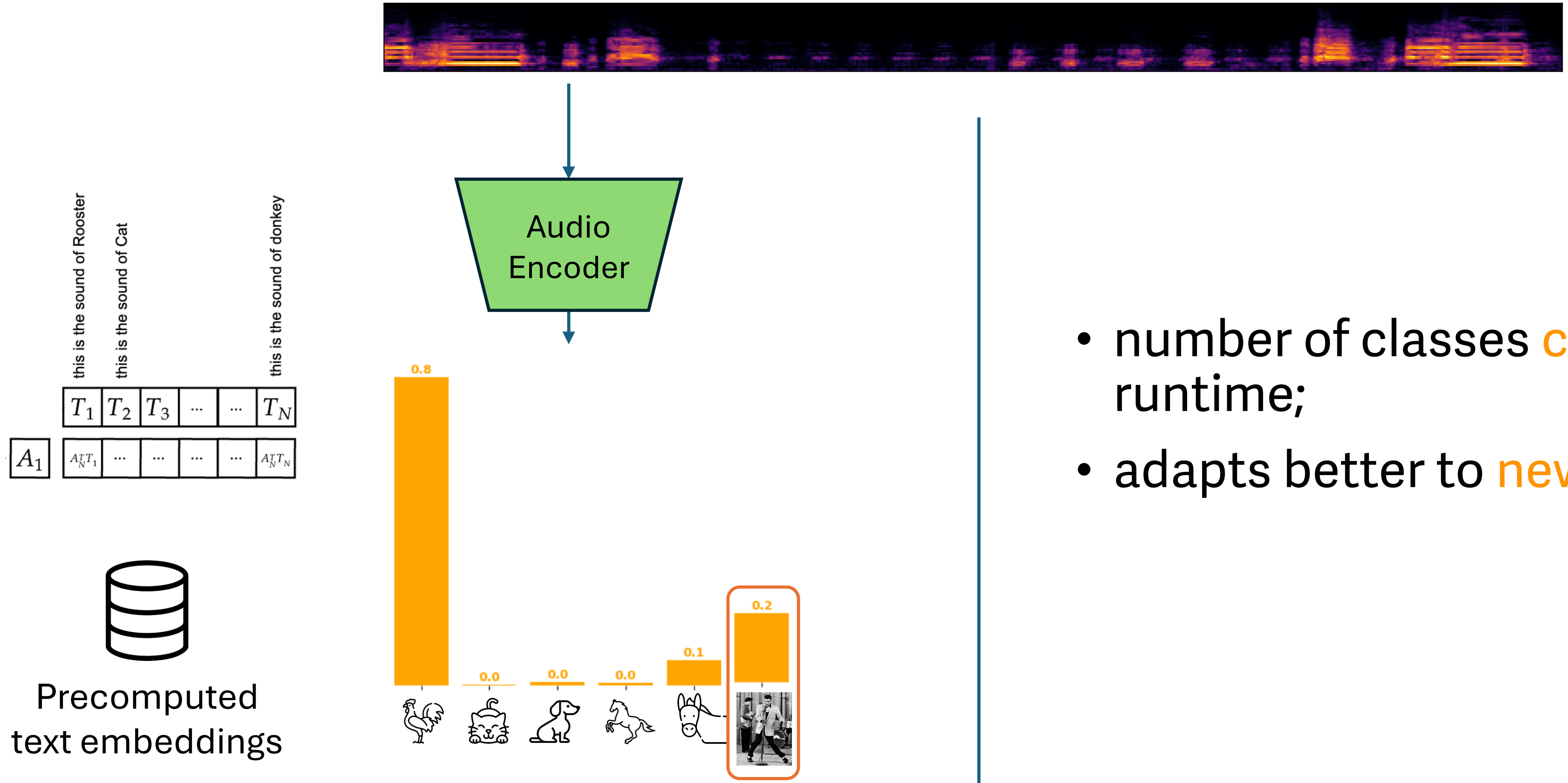


- the number of classes is fixed;
- hard to adapt to new domains and tasks;

Zero-shot classification using CLAP



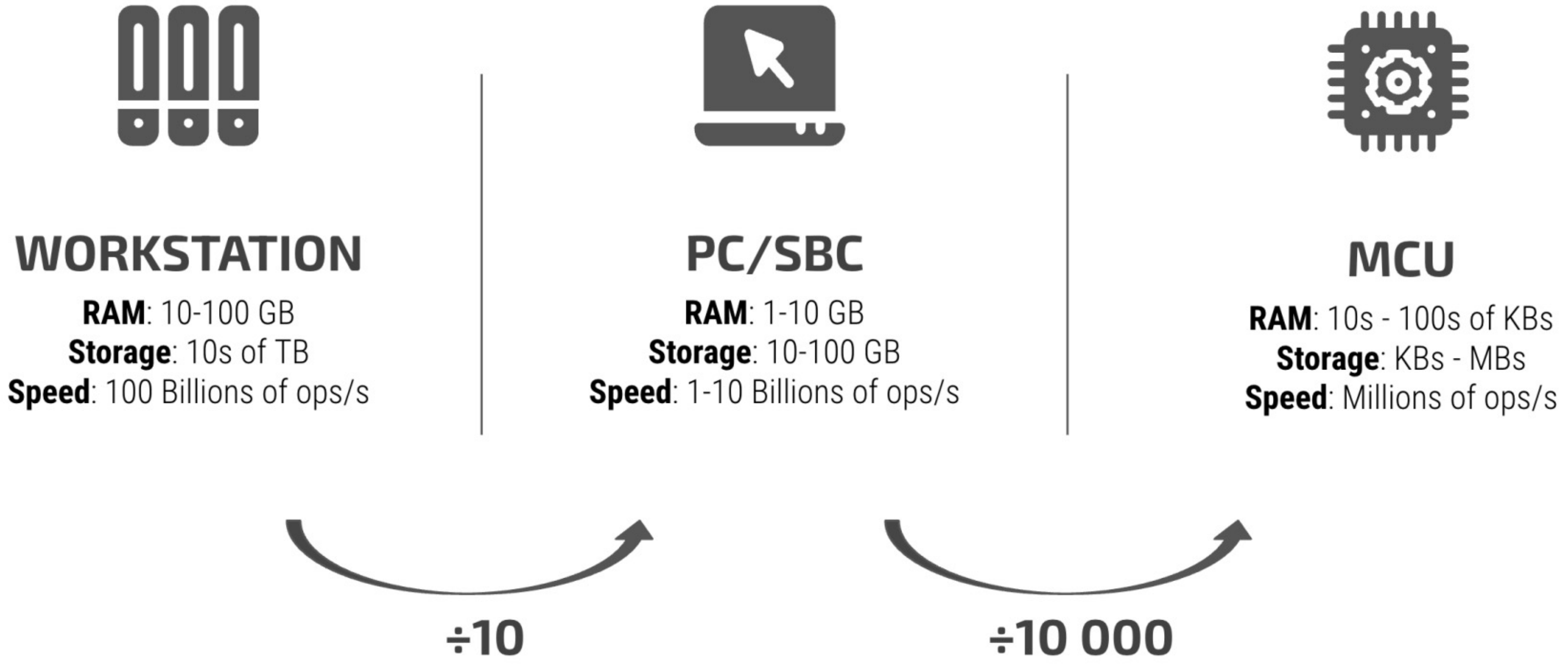
Zero-shot classification using CLAP



- number of classes can increase at runtime;
- adapts better to new domains;

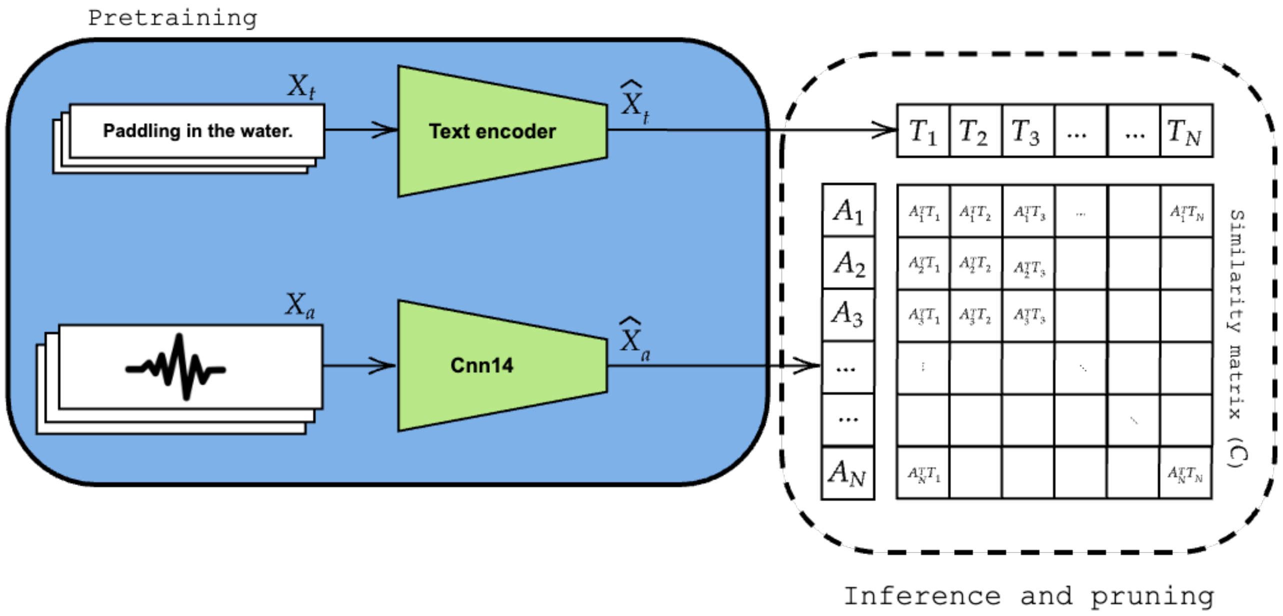
Challenges of bringing FM to the edge

- generally small models underfit big datasets;
- big models do not fit the requirements of edge processing;



tinyCLAP to the rescue

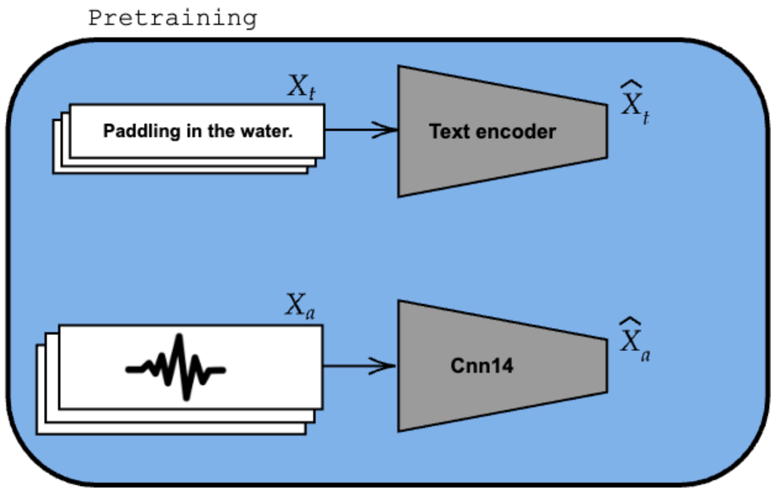
- tackles underfitting by distilling from zero-shot classifiers;
- prunes the latent representations to increase control on requirements;



"tinyCLAP: distilling contrastive language-audio pretrained models", Paissan, Farella, 2024

tinyCLAP to the rescue

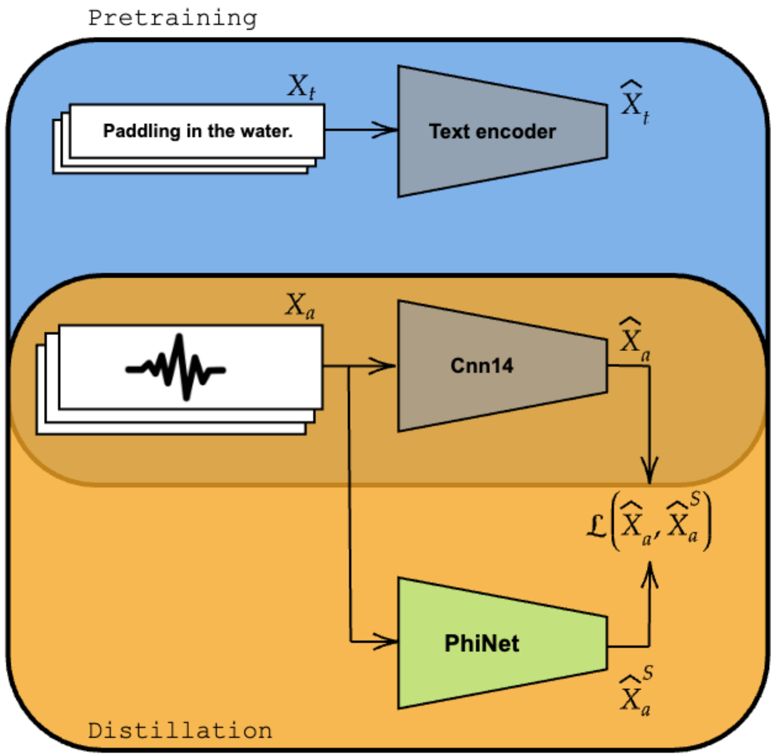
- tackles underfitting by distilling from zero-shot classifiers;
- prunes the latent representations to increase control on requirements;



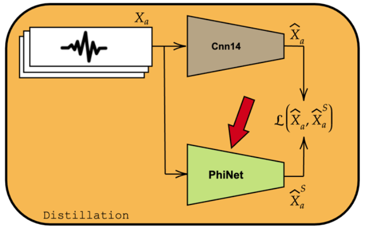
"tinyCLAP: distilling contrastive language-audio pretrained models", Paissan, Farella, 2024

tinyCLAP to the rescue

- tackles underfitting by distilling from zero-shot classifiers;
- prunes the latent representations to increase control on requirements;

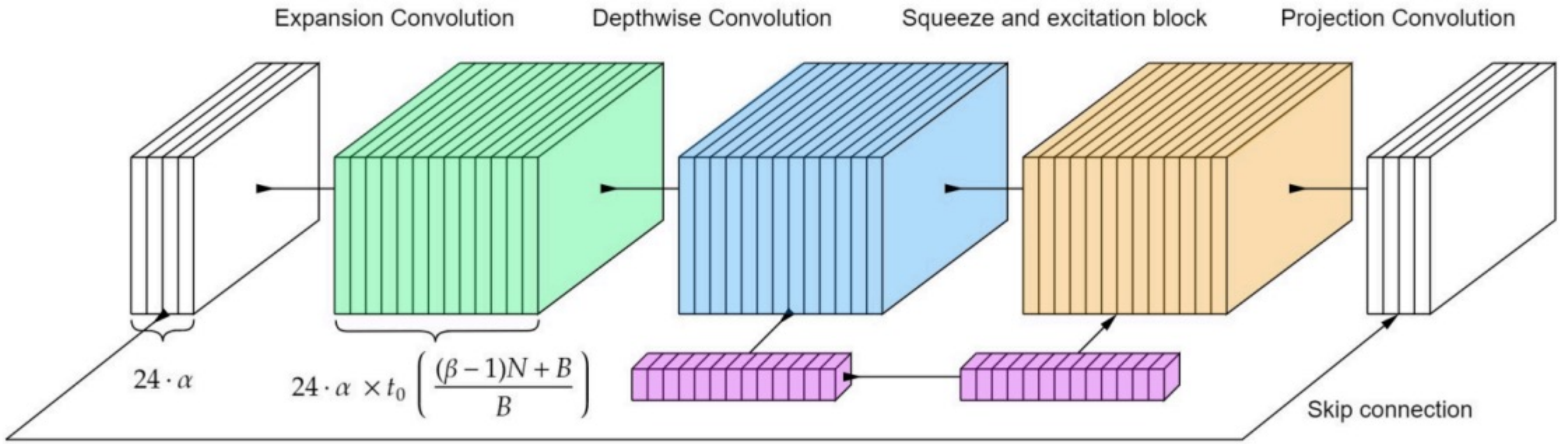


"tinyCLAP: distilling contrastive language-audio pretrained models", Paissan, Farella, 2024



PhiNet

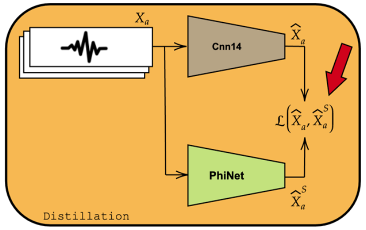
- edge-oriented NN: a sequence of scalable inverted residual blocks;
- adapts to changing computational constraints using hardware-aware scaling;




```
from micromind.networks import PhiNet
```

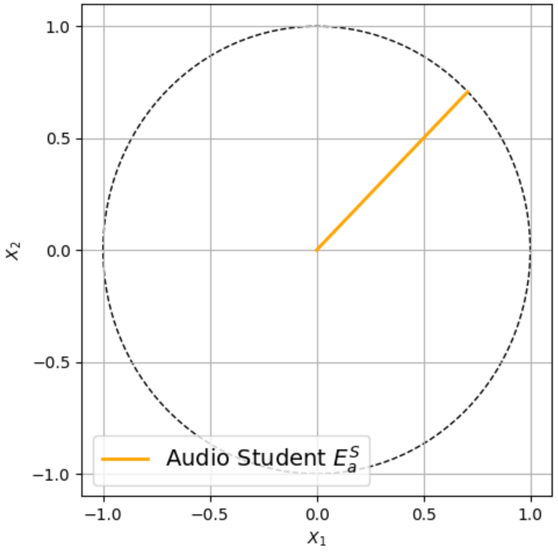

"PhiNet: a scalable backbone for low-power AI at the edge", Paissan, Ancilotto, Farella. 2021

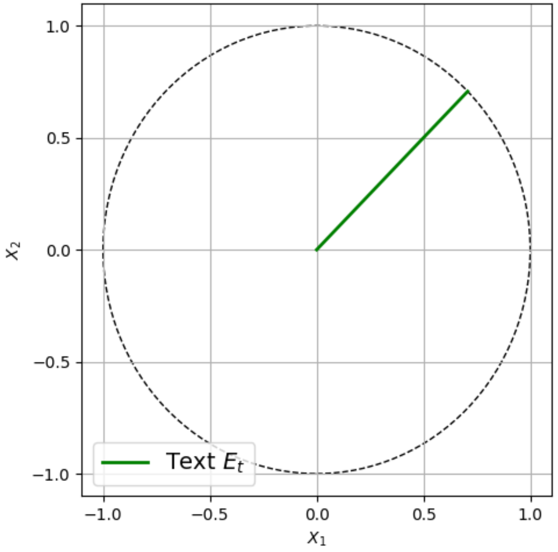
Which loss function is best?

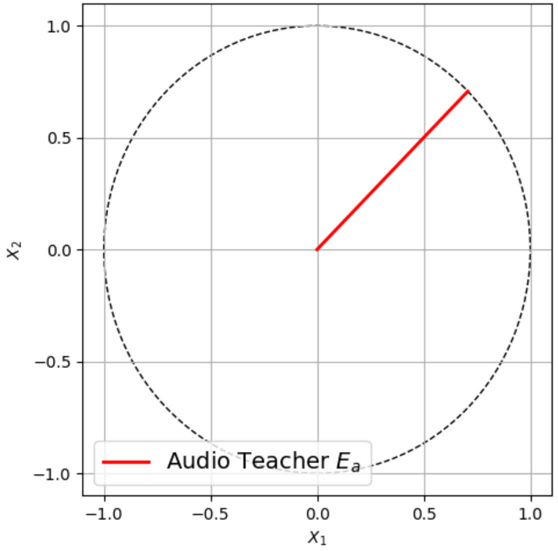


- Symmetric CE, L2 norm, cosine distance... a lot of options;
- what if text is not available?

$$\cos(\mathbf{E}_a, \mathbf{E}_t) = 1 \Leftrightarrow \cos(\mathbf{E}_a^S, \mathbf{E}_t) = 1.$$







"tinyCLAP: distilling contrastive language-audio pretrained models", Paissan, Farella, 2024

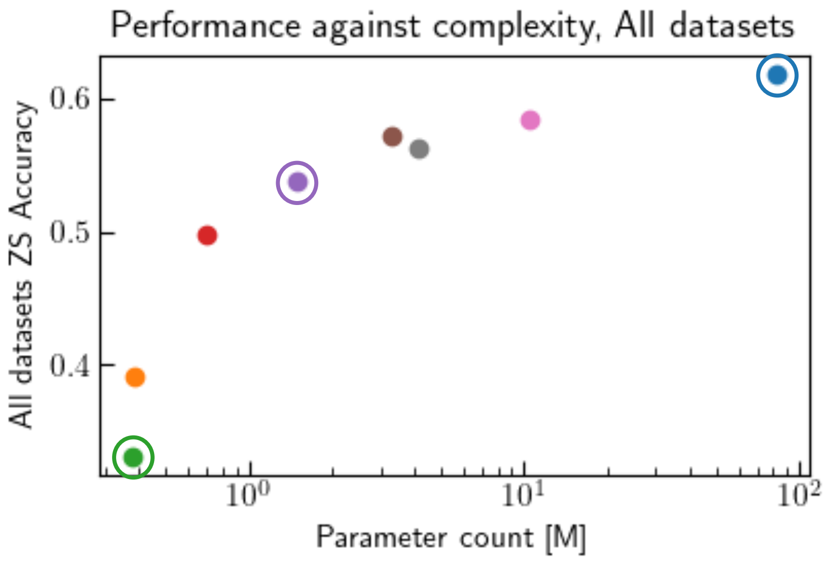
Validating the loss function

- experimental proof on self-distillation task;

						ZS Accuracy (%)		
Student model	α	β	t_0	N	**Params [M]**	TUT17	US8k	ESC50
CNN14	/	/	/	/	82.8	28.9	72.1	82.3
CNN14-CLAP	/	/	/	/	82.8	29.6	73.2	82.9

This distillation strategy works!

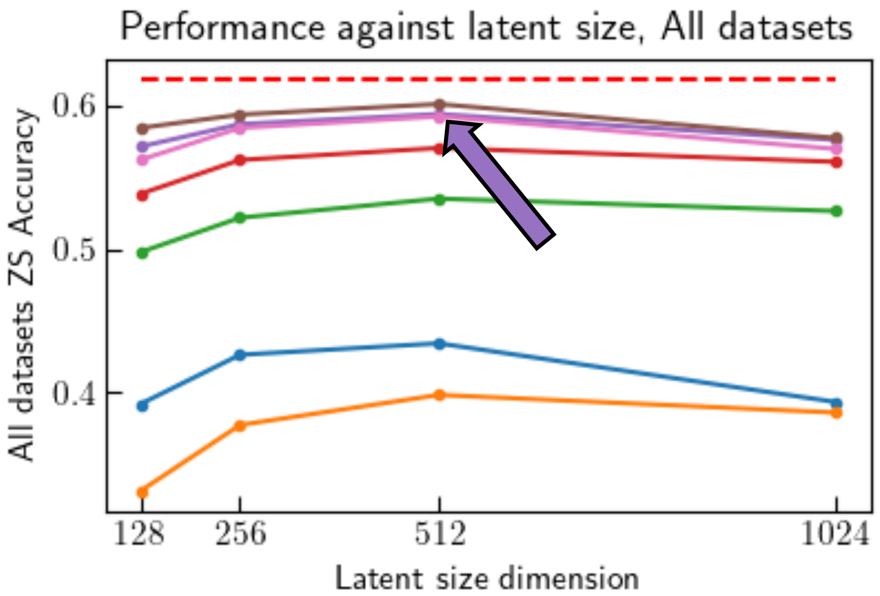
Efficient Zero-Shot classifiers



Student model	α	β	t_0	N	Params [M]	ZS Accuracy (%) TUT17	US8k	ESC50
PhiNet_1	3.00	0.75	6	7	7.0	25.2	68.3	77.4
PhiNet_2	3.00	0.75	6	9	13.0	26.4	69.7	77.2
PhiNet_3	3.00	0.75	4	7	6.2	26.1	70.3	76.5
PhiNet_4	1.50	0.75	6	7	4.4	27.5	67.9	73.0
PhiNet_5	0.75	0.75	4	7	3.5	26.7	65.2	66.1
PhiNet_6	0.75	0.75	4	4	3.2	22.1	51.8	41.9
PhiNet_7	0.75	0.75	6	4	3.3	22.3	51.6	44.1
CNN14	/	/	/	/	82.8	28.9	72.1	82.3
CNN14-CLAP	/	/	/	/	82.8	29.6	73.2	82.9

The impact of latent size

- changing the latent dimension does not decrease performance monotonically;
- enables fine-grained design based on computational requirements;



6M parameters, negligible performance drop wrt baseline

Conclusion

- we present a technique to learn efficient CLAP models without text supervision;
- we present the first efficient zero-shot audio classifier;
- scaling (down) foundational models can push the frontiers of what we can achieve at the edge.

tinyCLAP webpage and demo
https://fpaissan.github.io/tinyclapweb