Multimodal prototypical approach for unsupervised sound classification

Saksham Singh Kushwaha[1,2], Magdalena Fuentes[2,3]

[1]Courant Institute of Mathematical Sciences, New York University, NY, USA

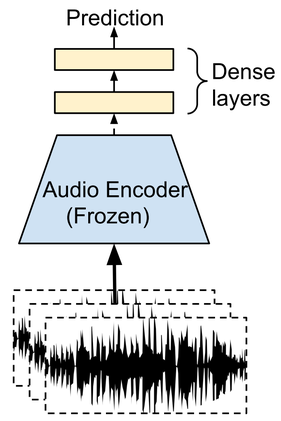
[2]Music and Audio Research Lab, New York University, NY, USA

[3]IDM, New York University, NY, USA

CODE

INTRODUCTION

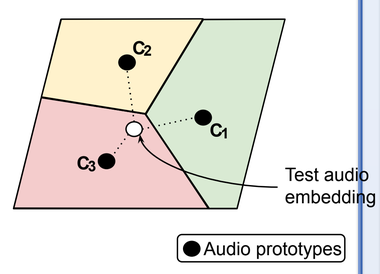
- Environmental sound recognition has several applications in public health and industry.
- Typical sound recognition system consists of supervised models with fixed class vocabulary.
- Target class vocabulary is not known a priori or changes dynamically for many real-world applications like assistive devices



Supervised Learning

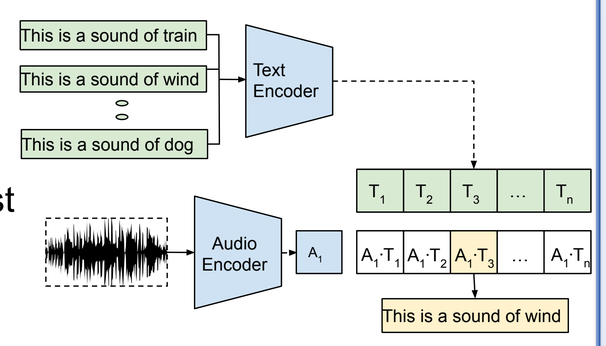
Previous approaches to make these systems more adaptable

- **PROTOTYPICAL FEW-SHOT APPROACH**
  - Create centres/prototypes using few annotated data
  - Predicted class is the nearest prototype

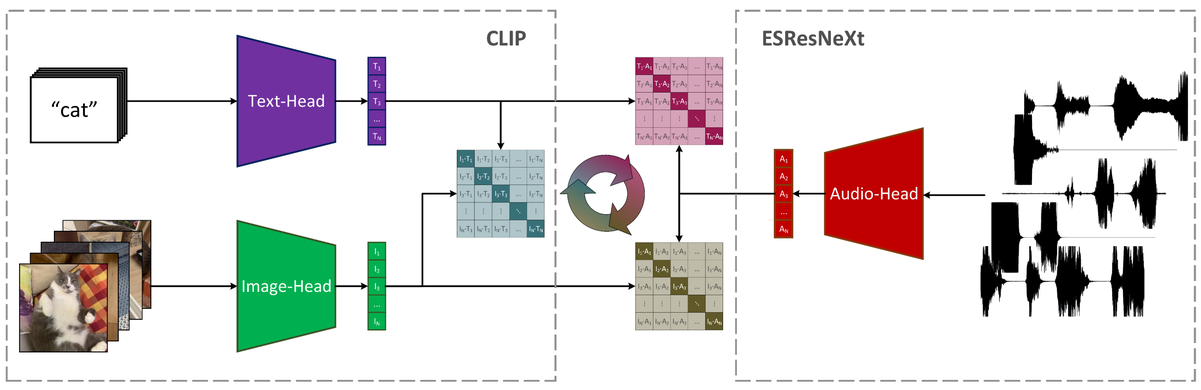


Prototypical

- **ZERO-SHOT(ZS) LEARNING**
  - Leverage shared embedding space of large multimodal models
    Eg. AudioClip[1], LAION-CLAP[2]
  - Predicted class is the label with highest cosine similarity



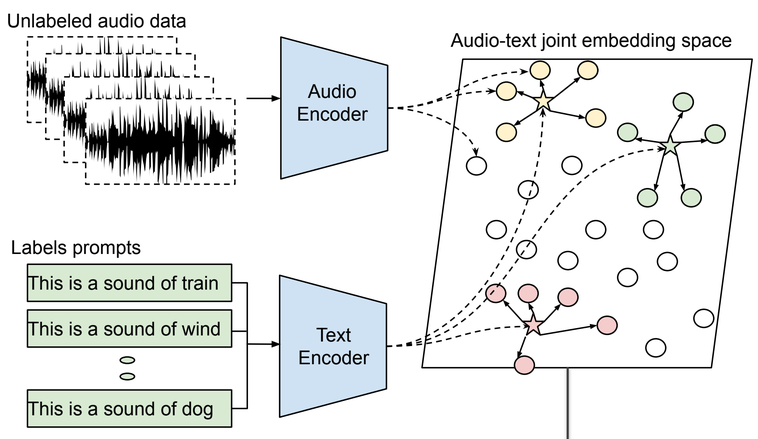
Zero-shot approach



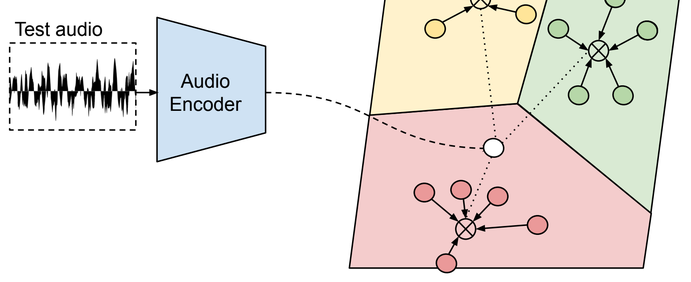
Joint training schema for Audio-Clip[1]

OUR APPROACH

- Unsupervised audio prototype creation using text
  - Find top-k nearest unlabelled audios for the interested text labels
  - Centroids of these audios are used as prototypes for corresponding class



- Classifying unseen audio samples
  - Predicted label is the nearest prototype

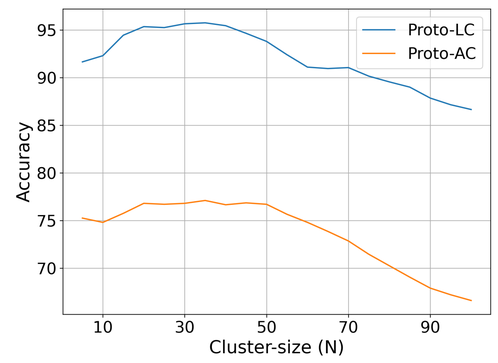


EXPERIMENTS

- Prompt Selection on ESC-50 [3]

	AudioClip	LAION-CLAP	Proto-AC	Proto-LC
{Class label}	0.67	0.83	0.77	0.94
I can hear {class label}	0.68	0.86	0.72	0.92
This is an audio of {class label}	**0.69**	0.90	0.72	**0.96**
This is {class label}	0.68	0.88	**0.78**	0.95
This is a sound of {class label}	0.67	**0.92**	0.70	**0.96**

- Cluster-size vs. Accuracy on ESC-50 [3]



- Comparison with other approaches

	ESC-50 (acc)		US8k (acc)		FSD50K (mAP)	
	Zero Shot	Supervised	Zero Shot	Supervised	Zero Shot	Supervised
Wav2Clip	0.41	0.86	0.40	0.81	0.03	0.43
AudioClip	0.68	0.88	0.62	0.86	0.20	0.50
CLAP	0.83	**0.97**	**0.73**	0.88	0.30	0.59
LAION-CLAP	0.91	0.96	0.72	**0.89**	0.22	0.61
Proto-AC(Ours)	0.78	0.82	0.71	0.77	0.40	0.48
Proto-LC(Ours)	**0.96**	**0.97**	**0.73**	0.83	**0.52**	**0.65**

CONCLUSION

- **Effect of prompt:**
  - Existing models are sensitive to the input prompts
  - AudioClip is most robust maybe due to extra image modality
- **Impact of the cluster size:**
  - Reasonable size to ensure the quality of prototype
  - Found optimal cluster size=35, used it for all other datasets
- **Performance comparison:**
  - Our approach improves upon ZS approaches by ~12%
  - Our ZS approach performs comparable to supervised approaches
  - Prototypes aligns text label embeddings to audio embedding space

REFERENCES

[1] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 976–980.

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," arXiv preprint arXiv:2211.06687, 2022.

[3] K. J. Piczak, "Esc: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018