Exploring LAM redundancy in bird sound classification

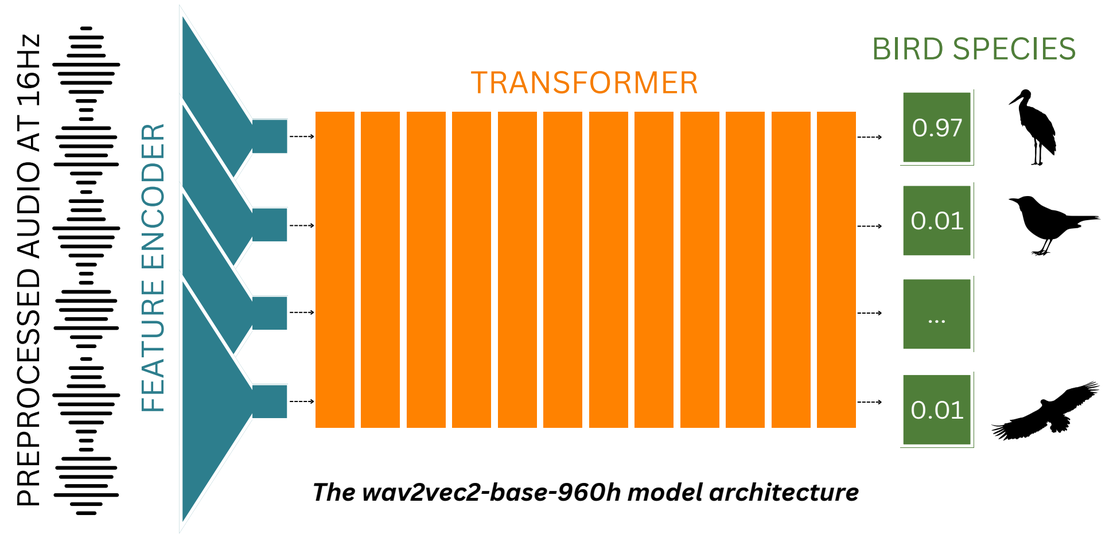
Marco Andreis (s243116)[1], Rasmus Porsgaard (s203953)[1] and Csilla Duray (s233903)[2]
1 Technical University of Denmark, 2 University of Copenhagen

Introduction

The field of bioacoustics has evolved rapidly in recent years with the emergence of advanced deep learning methods. Transformer models are now applied for in-depth analysis of animal sounds, for example to monitor biodiversity [1]. Yet, studies reveal that such large audio models (LAMs) can be significantly redundant having layers that don't necessarily contribute to their performance [2,3]. In this poster, we explore this redundancy using the wav2vec2 base model [4] as a case study by finetuning it for the classification of birds songs on 2,161 labelled examples. We show that it is possible to preserve performance while reducing model complexity and inference time, through pruning both pre and post training, and also a comparison to a simple CNN.

Models and Architecture

- **wav2vec2-base-960h:**
  - **Feature encoder:** 7 convolutional layers extract continuous feature representations from raw sound waves.
  - **Transformer:** 12 blocks composed by 8 attention heads of one attention layer and two linear layers each to learn meaningful representations.
- **CNN:**
  - **2 convolution layers + pooling**
  - **2 fully connected layers**



The wav2vec2-base-960h model architecture

Reducing redundant layers in finetuned models

- **Pruning** was initially used in the forms of back and forward pruning, as well as pruning individual layers.
- **Optimal pruned layers** were found by a trial and error approach based on results from the previous combinations.

Methods

- Mutual k-nearest neighbors

$$S_{\mathrm{kNN}}(i,j) = \frac{1}{n}\sum_{l=1}^{n}\left(\frac{1}{k}\left|\mathcal{N}_k\left(A_{l,\cdot}^{(i)}\right) \cap \mathcal{N}_k\left(A_{l,\cdot}^{(j)}\right)\right|\right)$$

- Cosine similarity

$$S_{\cos}(i,j) = \frac{1}{n}\sum_{l=1}^{n}\frac{\left(A_{l,\cdot}^{(i)}\right)^{\top} A_{l,\cdot}^{(j)}}{\|A_{l,\cdot}^{(i)}\|\|A_{l,\cdot}^{(j)}\|}$$

- Centered kernel alignment

$$S_{\mathrm{CKA}}(i,j) = \frac{\left\|\left(A^{(i)}\right)^{\top} A^{(j)}\right\|_F^2}{\left\|\left(A^{(i)}\right)^{\top} A^{(i)}\right\|_F \left\|\left(A^{(j)}\right)^{\top} A^{(j)}\right\|_F}$$

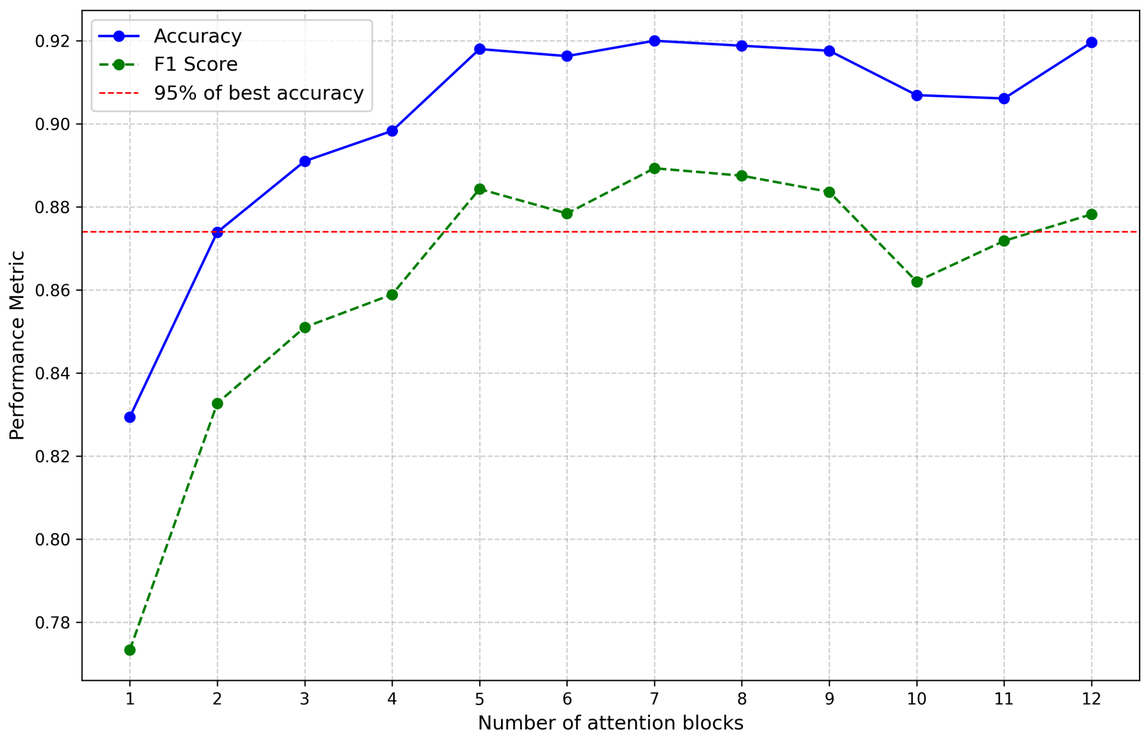
- Convexity:

Let (V, E) be a graph and $A \subseteq V$. A is convex if for all pairs $x, y \in A$, there exists a shortest path $P = (x = v_0, v_1, v_2, \ldots, v_{n-1}, y = v_n)$ and $\forall i \in \{0, \ldots, n\} : v_i \in A$.

Layers - forward pruning												Accuracy
1	2	3	4	5	6	7	8	9	10	11	12	
												0.9118
	■											0.8040
	■	■										0.6440
	■	■	■									0.4240
Layers - backward pruning												
											■	0.8960
										■	■	0.8760
									■	■	■	0.8200
								■	■	■	■	0.3120
Layers - middle pruning												
				■	■	■						0.7269
				■	■	■	■					0.7976
						■	■	■				0.8457
							■	■	■			0.8584
					■	■	■					0.6747
			■	■	■							0.6004

Layers - single layer pruning												Accuracy
1	2	3	4	5	6	7	8	9	10	11	12	
												0.9118
■												0.8040
	■											0.7960
		■										0.8240
			■									0.8760
				■								0.8680
					■							0.8960
						■						0.9080
							■					0.9040
								■				0.9000
									■			0.9000
											■	0.8960
Best pruning results												**Accuracy**
1	2	3	4	5	6	7	8	9	10	11	12	
							■		■	■		0.8898
				■			■		■	■		0.8314

Pruning the base model



Results

Model	Accuracy	F1-score	Number of parameters
Wav2Vec2-Base	0.0331	0.0038	94,581,426
Feature extractor	0.0261	0.0157	9,526,962
Wav2Vec2-Finetuned 7 blocks	**0.9200**	**0.8893**	**59,142,066**
Wav2Vec2-Finetuned 12 blocks	0.9118	0.8724	94,581,426
Wav2Vec2-Finetuned Pruned	**0.8898**	**0.8132**	**66,229,938**
CNN	0.6771	0.6747	4,760,914

References

1. Rasmussen, J. H., Stowell, D,, Briefer, E. F. (2024). Sound evidence for biodiversity monitoring. Science 385,138-140.DOI:10.1126/science.adh2716
2. Dorszewski, T., Tětková, L., & Hansen, L. K. (2024a). Convexity-based Pruning of Speech Representation Models. arXiv:2408.11858. https://arxiv.org/pdf/2408.11858
3. Dorszewski, T., Jacobsen, A. K., Tětková, L., & Hansen, L. K. (2024b). How Redundant Is the Transformer Stack in Speech Representation Models? arXiv:2409.16302. https://arxiv.org/pdf/2409.16302
4. Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. https://doi.org/10.48550/arXiv.2006.11477