View Reviews

Paper ID

5443

Paper Title

No Free Lunch from Audio Pretraining in Bioacoustics: A Benchmark Study of Embeddings

Track Name

Main Track

Reviewer #1

Questions

2. Relevance to IJCNN

Fair

3. Technical quality

Good

4. Novelty

Good

5. Quality of presentation

Good

9. Comments to Authors

This is an interesting paper about how the embedding space of pre-trained neural networks affects their clustering abilities and the effect of fine-tuning.
This research question is very relevant as it potentially reveals flaws of the current models due to training methodologies, network architectures etc. and could point to directions for future research.
The paper is well structured and has several informative illustrations. However, I believe the paper needs some more work to be ready for publishing in a conference. As I list below, I find several explanations insufficient and design choices not well explained.

Abstract:
- the distinction between embeddings and features is not clear
- "checking the embeddings" is too colloquial

I. Introduction
- any reference for defining embeddings as high-dimensional representations of audio features? audio feature vectors can also be high-dimensional. Maybe replace "audio features" by "audio signals"
- "Do we still need fine-tuning?" -> better integrate into text. Instead, better highlight what your driving research question is.

- "Fine tuned ... either" -> I would argue that most research in bioacoustic (e.g. AVESbio) involves pre-trained deep learning models which are fine-tuned for some downstream tasks
- I would remove the information "more than 0.8 accuracy" (it is not clear to the reader how many classes are considered here and hence what 0.8 actually means in terms of performance quality) and "less than 0.4" -> this is too specific and not required here to make your point
- after "Our contributions ... are as follows": form actual sentences.
- 1) "study" seems to vague, be more concrete: which properties do you study?
- 3)-5) are not methodological contributions (which you would usually list here) but rather main results (which you should discuss in the end of your paper in the conclusion section)
- Fig. 1 -> move "high-dimensional and latent" to text (focus in the caption only on aspects which explain the figure itself)

II.A.
- rephrase section title (the number of models is not important, rather "Model architectures" would be a better fit
- replace "earthspecies/aves" by actual URL (e.g. as footnote)

II.B
- "better performance" -> explain which metric you refer to

II.C
- "...is adjusted to meet" -> explain how this is done

III.A
- here you refer to fig. 2 -> make sure that figures appear in the same order as they are referenced in the main text
- is the NMI computed in the original (high-dimensional) embedding space or in the reduced embedding space (after t-SNE)?

III.B
- the section title is rather a statement -> consider rephrasing
- can you support the statement that the NMI value correlate with accuracy with some actual correlation values? (to show that this is a general observation across multiple datasets)
- can you give the reader any idea why the pre-trained models perform poor on the detection tasks here in this section?

III.C
- typo: "experimentat"

IV
- "three objectives" -> these 3 objectives need to be clearly formulated in the introduction part
- explain "overlooks background sounds" in the context of your study

Reviewer #2

Questions

2. Relevance to IJCNN
Good

3. Technical quality
Good

4. Novelty
Fair

5. Quality of presentation
Fair

9. Comments to Authors
This submission proposes to evaluate 11 deep learning modes with or without pretraining/fine-tuning on the bioacoustic dataset, BEANS. The model output embeddings are the main evaluation target. Several visualization methods and embedding quality measurement methods are applied, e.g., T-SNE, UMAP, mutual information, clustering quality, etc. My major concerns about this submission are on two folds: novelty and presentation quality. Some claimed contributions in this work are weak and not fairly proven.

Here are my questions:
1. One of the claimed contributions is "First study evaluates over ten DL models on same datasets". However, 1) in AVES, 8 models are evaluated on the BEANS, this work added two extra models, swin transformer and alexnet. The motivation of adding these two models is unclear. 2) in AVES, the model performances are estimated with the classification/detection accuracy, which is more related to the practical application, while the submission did not. The above two points weaken the novelty and contribution of this work.

2. One of the claimed contributions is "Finds fine-tuning is necessary for audio-pretrained models". Fine-tuning is necessary for any pretraining models, which has been proved in the filed of computer vision (SimCLR), language processing (BERT), audio pretraining (BEATs, FrameATST, M2D) and bioacoustic (AVES report the fine-tuned result). Compared to AVES, this work shows the linear-evaluation results of the pretrained models and the clustering visualization of AVES embeddings, which might not be considered as novel.

3. In III.B, the author shows "Audio-pretrained models are not the best" with the evidences that the image-pretrained ResNet NMI is better than audio-pretrained AVES in the detection task datasets. The claimed conclusion is not solid enough to me: 1) The comparison is not fair. Why not pretrain a ResNet with audio data and compare the image-pretrained and audio-pretrained ResNet to draw this conclusion? 2) NMI results is not solid. The author said "The clustering performance measured here aligns with accuracy [22]: the dog dataset, which has the highest accuracy, also exhibits the largest NMI". But in Table I, AVES and ResNet152 NMI results on `hiceas` are (0.145, 0.405), which are not positively related to the accuracy reported in [22] (0.629,0.273).

4. In III.D, the author claimed that removing unlabeled data will boost the model performance. This could be true if the unlabeled data is removed only in the training set. But in experiment, the unlabeled data in validation and test sets are also removed, which makes the `beans` column and `ours` columns are not evaluated from the same dataset, and makes the reported results are not fair for comparison.

5. The model name in Table 1 and other tables are not the same.

Reviewer #3

Questions

2. Relevance to IJCNN
Excellent

3. Technical quality
Very good

4. Novelty
Very good

5. Quality of presentation
Good

9. Comments to Authors
Here are some review comments based on the paper:

Strengths:
1. Clear Objective: The paper presents a clear comparison of the performance of various deep learning models in bioacoustic tasks, both with and without fine-tuning. It is a timely contribution as bioacoustics continues to grow as a monitoring tool for ecosystems.
2. Novelty and Relevance: The study tackles an important question—whether fine-tuning is necessary for pretrained models in bioacoustics. This question has broader relevance to the field of machine learning, where the generalization of pretrained

models across different domains is a key issue.
3. Empirical Analysis: The use of a benchmark study to assess the performance of 11 deep learning models is well-executed. The inclusion of clustering techniques for evaluating dimensionality reduction and feature separation is a strong approach to assess the quality of the embeddings.
4. Concrete Results: The findings of the study are clearly stated and useful for the field. The observation that fine-tuning models improves their performance, especially with fewer background sounds, is a valuable insight.

For Improvement:
Explanation of Why Some Models Fail: The study mentions that some models fail to separate background sounds from labeled sounds, but this could be expanded upon. Why does ResNet perform better in this case? Are there specific architectural or training characteristics that make ResNet better suited for this task? A more in-depth analysis could make the results more insightful.

Questions:
1. The study concludes that fine-tuning is necessary, but are there scenarios or particular models where fine-tuning might not significantly improve performance? Are there cases where it might even hinder performance due to overfitting?
2. Could the choice of background sounds in the fine-tuning process play a role in the model's generalization to new environments? Some clarification on this would be interesting.

Overall, this paper contributes valuable insights into the use of pretrained deep learning models for bioacoustic tasks.