Reviewer Comments

Classifying Marine Mammals Using Convolutional Neural Networks

Summary: This paper presents a deep learning-based framework for classifying individual whales and dolphins from the Happywhale dataset using a CNN with an EfficientNetB5 backbone and Elastic ArcFace loss. The model employs diverse data augmentation strategies, transfer learning, KNN for inference refinement, and ensemble techniques to improve identification accuracy. The best performing model achieves 88% accuracy based on MAP@5. The study addresses practical challenges in marine mammal identification and has implications for conservation through automation of a traditionally manual process.

Major Issues:

1. The choice of EfficientNetB5 over other EfficientNet variants or recent transformer-based models (e.g., Vision Transformers or ConvNeXt) is not critically justified. A comparative baseline or ablation is missing.

- Added subsections "Architecture Selection Results" and "Architecture Selection Rationale", which compares EfficientNetB5 with other model backbones.

2. The Happywhale dataset is inherently imbalanced. However, there is no mention of techniques like class reweighting, focal loss, oversampling, or data balancing strategies being used.

- Added subsections "Class Imbalance Analysis" and "Imbalance Mitigation strategy" to describe increasing minorities by performing image augmentations, and we explained why oversampling was avoided.

3. The role of the KNN classifier at inference time is described vaguely. There's no justification for the choice of k, nor clarity on how it integrates with the CNN feature embeddings.

- Subsection "K-Nearest-Neighbor Head" lists the steps taken to integrate KNN, and the "Hyper-parameter selection for k" subsection justifies choosing k = 50 using grid search. Threshold selection is also explained in "Distance-based thresholds for new_individual."

4. MAP@5 is the only evaluation metric. Precision, recall, confusion matrix, or top-1 accuracy should be reported to provide a more holistic view of model performance.

- This could not be accomplished because labels are only provided for training data. Labels used to generate MAP@5 scores for the Happywhale competition are not provided.

5. Accuracy improvements are shown, but no standard deviations, confidence intervals, or statistical tests (e.g., t-tests or Wilcoxon) are provided to assess model reliability across folds.

- Cross-validation metrics now include standard deviation, confidence intervals, and more in the "Cross-Validation Performance and Reliability" subsection.

6. It's not explicit whether the split between training, validation, and test data was stratified to prevent data leakage (e.g., same individual across sets).

- The "Data Splitting and Leakage Prevention" subsection describes that the competition enforces a separation between the train/test folders and that no images are repeated in either folder.

7. There's no use of interpretability tools like Grad-CAM or saliency maps to verify what parts of the image the model uses to make predictions—important for conservation tools.

- Subsection "Model Interpretability with Grad-CAM" includes a few Grad-CAM heatmaps, and a short description of what is being displayed.

8. There's no experimental evidence or ablation showing how Elastic ArcFace outperforms

ArcFace in this particular application.

- The manuscript now does not mention Elastic ArcFace at all. Only ArcFace and ArcMargin are used, compared, and discussed.

9. The "blend of models" section lacks specific weights, number of models involved, validation strategy for weight optimization, and blending method used.

- Sections "Model Selection Criteria", "Weight Optimization", and the blend equation now detail the number of models (30), weight vector, and ensemble methodology.

10. The manuscript mentions that augmentation simulates real-world imaging variance, but no quantitative analysis of model robustness to low-quality images (blurry, low-light) is done.

- Subsection "Model Robustness to Image Quality Degradation" introduces Gaussian blur and brightness reduction and reports the changes in MAP@5 as a result.

11. It's unclear how multi-sample dropout and learning rate scheduling interplay. More structured presentation of training loops, epochs, and validation checkpoints would help.

- The learning rate schedule is plotted and described. Training epochs and batch sizes are described in "Training Strategy".

12. Given this is a research manuscript, it should include or link to a GitHub repository to allow reproducibility of results.

- In subsection "Ensemble and Blending" the link is included. (https://github.com/adityapkatre/Research)

13. Figures are listed (e.g., Figure 1, Figure 2), but they're either not clearly captioned or integrated poorly. Graphs like learning rate schedules should have axis labels and units.

- All figures have captions. The learning rate schedule graph has axis labels.

14. "Argumentation" is repeatedly used incorrectly instead of "augmentation," affecting technical accuracy and clarity.

- All instances of "argumentation" have been corrected to "augmentation."

15. The performance gap between ResNet + ArcMargin (48%) and EffNetB5 + ArcFace (86%) is substantial. An in-depth discussion of why this leap occurs is missing.

- Subsection "Model Performance Gap" explains the key reasons for the jump.

16. There is no analysis of misclassified samples or types of species/images that are hardest to predict. This is essential for identifying model limitations.

- Subsection "Misclassification Analysis" discusses common failure modes, including image quality and visually similar species. A Grad-CAM for one of these is also included.

17. The manuscript doesn't discuss how this model might be integrated into a conservationist's workflow or as a field tool.

- Subsection "Practical Applications" describes potential integration in mobile apps for public engagement.

18. The use of Google Cloud TPUs is mentioned, but no runtime benchmarks, cost analysis, or comparison to GPU training is provided. This limits the understanding of scalability.

- Subsection "Runtime Benchmarks" compares TPU with GPU training times and describes the Kaggle free tier.

19. The abstract and results conflate "accuracy" with Mean Average Precision at 5 (MAP@5), a competition-specific metric. This is misleading, as MAP@5 evaluates ranked predictions, not classification accuracy. For example, the claim of "88% accuracy" is incorrect; the model achieved 0.88 MAP@5.

- The abstract and results have been updated to refer to "MAP@5" instead of "accuracy" when reporting the 0.88 score.

20. The table shows a progression from 0.10 to 0.88, but the metrics are labeled as "Accuracy (%)", which is misleading because the competition uses MAP@5. This is a critical error. The results section also includes code snippets and figures, but the figures (like Figure 1) are referred to but not included in the text, which is confusing.

- The table caption now labels these values as “MAP@5 Scores,” not “Accuracy.”

21. Author should provide comparison to state-of-the-art methods on the Happywhale dataset. How does 0.88 MAP@5 rank against existing solutions?

- Subsection “Comparison with other Happywhale Leaderboard Models” compares this work with other MAP@5 scores on the leaderboard.

22. No details on class imbalance (critical for a dataset with 15,587 individuals). Were techniques like oversampling or weighted loss used?

- See “Class Imbalance Analysis” and “Imbalance Mitigation Strategy."Augmentations and cross-validation were applied to account for the class imbalances.

23. The role of KNN during inference is unclear. How many neighbors? How are cosine similarity thresholds determined?

- Covered in subsections “K-Nearest-Neighbor Head”, “Hyper-parameter selection for k”, and “Distance-based thresholds for new_individual.” Details on k = 50, cosine distance threshold of 0.62, and other methods are explained.

24. The blending equation lacks context. How were weights wiwi optimized? What criteria governed model selection for the ensemble?

- Subsections “Model Selection Criteria” and “Weight Optimization” explain the coordinate ascent procedure and selection process for final weights used in the ensemble.

25. Discuss dataset limitations, computational requirements, and ecological applicability.

- Subsections “Model and Dataset Limitations” and “Practical Applications”have been created.

26. Contrast results with prior work, explicitly state limitations, and clarify conservation impact.

- Subsections “Comparison with Prior Work” and compares this study with other works and “Model and Dataset Limitations” explains biases in the dataset.

27. Include a detailed error analysis section that examines common failure modes. Provide examples of misclassified images and discuss patterns in the errors.

- Subsection “Misclassification Analysis” discusses errors and a Grad-CAM visual for evidence.

Minor Issues:

1. Several citations lack consistent formatting some inline, others with parentheses.

- Citation style is now consistent.

2. Collect more literature.

- Additional literature has been added, and the citation format is now correct.

3. Punctuation errors in long compound sentences (e.g., missing commas).

- Grammar and punctuation has been revised at several parts of the manuscript.

4. Use consistent image size notation—either “380x380” or “(380, 380),” not both.

- Image size is now always written as 380x380.

5. Revisit the section on ArcFace, some repetition can be removed for conciseness.

- Redundancy in the ArcFace explanation has been reduced.

6. Equations for MAP@5 should use standard LaTeX formatting or be visualized for clarity.

- MAP@5 are all using standard LaTeX formatting.

7. Define all acronyms at first use (e.g., MAP@5, TPU, CNN).

- All acronyms are now defined at first use.

8. Some subheadings lack parallel structure (e.g., “Discussion” vs “Blend of Models”).

- Subheadings have been revised for consistency.

9. Unclear what “Model Architectures and Code Excerpts” are doing in the results section.

- Subsection “Model Architectures and Code Excerpts” is in the results section, and has a contextual explanation.

10. Overuse of passive voice reduces clarity in methodological descriptions.

- Use of passive voice has been minimized.

11. Some figures are overlapping the text.

- None of the figures are overlapping text in the PDF.