A Highly Efficient Marine Mammals Classifier Based on a Cross-Covariance Attended Compact Feed-Forward Sequential Memory Network (Student Abstract)

Xiangrui Liu, Julian Cheng
University of British Columbia, Kelowna, British Columbia, Canada
assa8945@student.ubc.ca, julian.cheng@ubc.ca

Abstract

Military active sonar and marine transportation are detrimental to the livelihood of marine mammals and the ecosystem. Early detection and classification of marine mammals using machine learning can help humans to mitigate the harm to marine mammals. This paper proposes a cross-covariance attended compact Feed-Forward Sequential Memory Network (CC-FSMN). The proposed framework shows improved efficiency over multiple convolutional neural network (CNN) backbones. It also maintains a relatively decent performance.

Introduction

Marine mammals play an essential role in maintaining the ecosystem of the Arctic. Humans' activities including the use of active sonar and vessels are part of the reasons that led to the decrease in the marine mammals' population. One direction to mitigate the impacts of active sonar and shipping is to achieve early detection and classification of marine mammals. Hydrophones are easy to be deployed on vessels to collect real-time acoustic signals emitted by marine mammals. The acoustic signals can be used to detect and classify mammals.

Deep learning has attracted interest as it can identify marine mammals fast and accurately without the presence of experts. Despite the popularity of attention mechanisms in computer vision, the adaptation of attention mechanisms and the Feed-Forward Neural Network (FNN) are barely found in the classification of marine mammals. Many researchers proposed to have CNN as the backbone network. Lu et al. (Lu, Han, and Yu 2021) used a fine-tuned AlexNet to train a classifier that differentiates three marine mammals' sounds. Zhong et al. (Zhong et al. 2020) ensembled multiple fine-tuned CNNs to detect Beluga whales' sounds collected in Alaska and achieved an accuracy of 96.3%. More, Allen et al. (Allen et al. 2021) picked ResNet50 as the backbone network and trained a classifier that can identify Humpback Whales' songs accurately.

Indeed, CNN-based models produced some decent classifiers. However, CNN models are computationally costly. It requires high-performance hardware, and it generally takes more time to train a CNN model. Further, the number of classes involved in the above work is limited. There are far more species of marine mammals.

This paper proposes a cross-covariance attended compact Feed-Forward Sequential Memory Network (CC-FSMN) that is much more efficient than some well-known CNN models and can distinguish thirty-two marine mammals. We compared the efficiency and the performance of different popular CNN backbones that include AlexNet (Lu, Han, and Yu 2021), VGG16 (Zhong et al. 2020), DenseNet (Zhong et al. 2020) and ResNet50 (Allen et al. 2021). Our preliminary results suggest our framework outperforms them in efficiency.

Methodology

The proposed framework is shown in Fig 1. First, compute the mel-scale frequency cepstral coefficients (MFCCs) and segment them into smaller aligned segments. Second, apply cepstral mean and variance normalization to the data. Lastly, the normalized and aligned MFCCs segments will be used for model training. The different layers in the cc-FSMN model are connected using a cascade structure.

Compact Feed-Forward Sequential Memory Network (cFSMN) FSMN (Zhang et al. 2018) is a new type of FNN with memory blocks that encodes the context information, allowing the transmission of long-term dependency in sequential data without the need for recurrent feedback. The

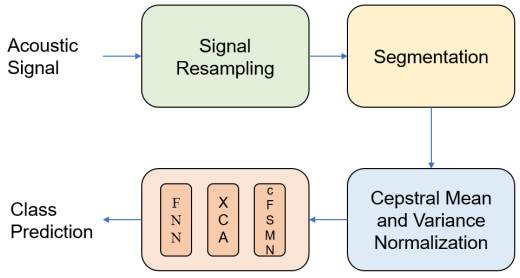

Figure 1: The overall structure of the framework

cFSMN (Zhang et al. 2018) is a variant of FSMN architecture. With reference to the structure of cFSMN in the supplementary material, the information from a non-linear hidden layer is projected to a linear layer and then added to a memory block. This compact form of FSMN makes cFSMN highly efficient in model training.

Cross-Covariance Attention In this framework, XCA (El-Nouby et al. 2021) is applied to the network to increase the robustness of the network and avoid the use of expensive computation of quadratic attention map. It is an attention-mapping mechanism proposed for an image transformer that allows an image classifier to handle images of different resolutions. Instead of a full pairwise interaction between the tokens by self-attention, XCA derives the attention map from a cross-covariance matrix from the projections of tokens. We believe it can be beneficial in the robustness of our classifier against noises. The detailed structure of XCA can be found in the supplement material.

Experiments and Results

Experiment Setup The experiment was done using the best audio cuts from the Watkins Marine Mammal Sound Database (WMMSD)[1]. The dataset contained 1694 sound cuts of 32 different species. This group of sound cuts is selected due to its high signal-to-noise ratio (SNR). Furthermore, it was the dataset used in Tao's experiment (Lu, Han, and Yu 2021) which is going to be one of the comparison experiments. In addition, an earlystop mechanism with 10 patiences applied. All the models were fine-tuned for single validation due to time limitations.

Experimental Results and Discussion As Table 1 shows that our framework takes the least amount of time to finish training. As compared to VGG16, CC-FSMN is 23 times faster. More, CC-FSMN has more parameters than a pure cFSMN model yet it has reduced training time. This can be attributed to the earlystop mechanism. The attention mapping in CC-FSMN speeds up the gradient descent process. The CC-FSMN framework reaches its global minima faster. Generally, the CC-FSMN has significantly reduced training time and inference time over the CNN backbones.

The performance results of the models are shown in table 2. The F1 score of the CC-FSMN surpasses other models. VGG16 and DenseNet also have F1 scores higher than 0.900 but they are much less efficient. Moreover, AlexNet shows similar training time but its performance is not as promising as the CC-FSMN. In this series of experiments, the CC-FSMN shows the best efficiency and performance.

The confusion matrix generated from the CC-FSMN in the supplementary material indicates that the model has difficulty distinguishing sperm whales and short-finned pilot whales. These two classes show confusion for all other models. Therefore, the problem may be a result of the feature itself rather than the architecture of the networks. The model shows a decent performance in other classes despite the dataset being highly unbalanced.

[1]https://cis.whoi.edu/science/B/whalesounds/index.cfm

Network	# of parameters	Training time	Inference time
AlexNet	61,100,840	20 minutes	5.4 seconds
ResNet50	23,567,328	135 minutes	14.2 seconds
VGG16	138,487,496	371 minutes	11.2 seconds
DenseNet	8,005,448	269 minutes	33.9 seconds
cFSMN	8,211,744	16 minutes	3.9 seconds
CC-FSMN	8,474,152	11 minutes	3.3 seconds

Table 1: Efficiency results

Network	Precision	Recall	F1 score
AlexNet	0.880	0.919	0.880
ResNet50	0.821	0.868	0.811
VGG16	0.918	0.932	0.916
DenseNet	0.912	0.940	0.917
cFSMN	0.897	0.934	0.908
CC-FSMN	0.929	0.957	0.931

Table 2: Performance results

Conclusion and Future Work

This paper proposed the CC-FSMN framework that significantly outperforms CNN backbones in efficiency for marine mammals classification. In the future, the experiment will include the noisier data from the WMMSD to study the performance of the framework when the input contains noise. Further, we will investigate the effects of X-vector embedding(Snyder et al. 2018) on the confusing two classes to further improve the framework.

References

Allen, A. N.; Harvey, M.; Harrell, L.; Jansen, A.; Merkens, K. P.; Wall, C. C.; Cattiau, J.; and Oleson, E. M. 2021. A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, 8.

El-Nouby, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; and Jegou, H. 2021. XCiT: Cross-Covariance Image Transformers.

Lu, T.; Han, B.; and Yu, F. 2021. Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecological Informatics*, 62.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Zhang, S.; Lei, M.; Yan, Z.; and Dai, L. 2018. Deep-FSMN for Large Vocabulary Continuous Speech Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Zhong, M.; LeBien, J.; Campos-Cerqueira, M.; Dodhia, R.; Ferres, J. L.; Velev, J. P.; and Aide, T. M. 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics*, 166.