Hear You Say You: An Efficient Framework for Marine Mammal Sounds' Classification

Xiangrui Liu, Xiaoou Liu, Shan Du, Julian Cheng

The University of British Columbia, Canada
assa89456@gmail.com, xiaoou12@mail.ubc.ca, {shan.du, julian.cheng}@ubc.ca

Abstract

Marine mammals and their ecosystem face significant threats from, for example, military active sonar and marine transportation. To mitigate this harm, early detection and classification of marine mammals are essential. While recent efforts have utilized spectrogram analysis and machine learning techniques, there remain challenges in their efficiency. Therefore, we propose a novel knowledge distillation framework, named XCFSMN, for this problem. We construct a teacher model that fuses the features extracted from an X-vector extractor, a DenseNet and Cross-Covariance attended compact Feed-Forward Sequential Memory Network (cFSMN). The teacher model transfers knowledge to a simpler cFSMN model through a temperature-cooling strategy for efficient learning. Compared to multiple convolutional neural network backbones and transformers, the proposed framework achieves state-of-the-art efficiency and performance. The improved model size is approximately 20 times smaller and the inference time can be 10 times shorter without affecting the model's accuracy.

Introduction

Marine mammals serve as vital components within marine ecosystems (Bowen 1997), contributing significantly to their health and stability. The top predators, such as killer whales and certain dolphins, play an essential role in maintaining the balance of their prey populations. Furthermore, marine mammals serve as barometers of ecosystem well-being. However, the population of marine mammals has been declining sharply (Taylor et al. 2007). Human-induced factors such as habitat loss from construction, overfishing, and activities such as sonar and shipping contribute significantly to this decline (Buck and Clavert 2005). Military sonar, in particular, emits sound waves that can cause severe harm, and even fatalities, among marine mammals (Parsons 2017). Additionally, the frequency of active sonar disrupts the echolocation abilities crucial for whales' hunting, further compromising their survival. Urgent action is required to protect these marine creatures.

One way to protect marine mammals is to achieve advanced detection and classification so that ships can take a detour and the Navy can reschedule its exercises to avoid harming the protected marine mammals. Machine learning has been introduced to detect and classify marine mammals, and it can be used to address the problem in two ways. One approach is to analyze photos of marine mammals using machine learning models. For example, the Simple Linear Iterative Clustering technique was used to segment the aerial images of dugongs (Maire, Alvarez, and Hodgson 2015) and then train these images with convolutional neural networks (CNNs), obtaining an F1 score of 0.406. In addition to aerial images, an attempt has been made to compare three machine learning methods using photos of Commerson's dolphins (Pollicelli, Coscarella, and Delrieux 2020), and the F1 score of the best model is 0.833. Another approach classifies marine mammals using vocalizations recorded by passive acoustic monitors since marine mammals produce sounds of different frequencies and patterns. Taking advantage of the complex networks and weights trained with large datasets, different CNNs (Lu, Han, and Yu 2021; Zhong et al. 2020; Allen et al. 2021; Duan et al. 2022) have been used to build marine mammal sound classifiers to produce accurate predictions.

Indeed, machine learning has attracted interest because it can accurately identify the sounds of marine mammals without the presence of specialists. However, there is still room for improvement. First, existing work mainly focuses on the application of transfer learning using different pre-trained models. These models usually consist of complex and deep neural networks, which have the advantage of being accurate, but at the expense of efficiency. In other words, their training time and inference time can be shortened. Second, most existing deep learning marine mammal sound classification work is trained on private datasets, so the signal-to-noise ratio (SNR) and data size are unknown. The only known open-source dataset used was the Watkins Marine Mammals Sound Database (WMMSD)[1]. The experiment using WMMSD (Lu, Han, and Yu 2021) only tested the model on the best sections of the database, which are the high SNR audio sections. The noisy audio cuts were not explored; therefore, the robustness of the models is unknown. Therefore, our goal is to develop a computationally friendly and robust framework that efficiently recognizes the acoustic signals of marine mammals.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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

This paper proposes a knowledge distillation (KD) framework that is much more efficient than some well-known CNN-based models and transformers. We provide a comprehensive study that extends to a preliminary work based on a Cross-Covariance attended compact Sequential Memory Network (CC-FSMN) (Liu and Cheng 2023). Based on the CC-FSMN, we build a teacher network that embeds an X-vector extractor and a DenseNet network to better study the features and increase the accuracy of the system. We also replace the Dropout mechanism in the Cross-Covariance Attention (XCA) layer with a DropKey mechanism to improve the robustness of the model. The relatively large and complex teacher network effectively improves the CC-FSMN-based student model. Compared with the original CC-FSMN (Liu and Cheng 2023), XCFSMN has improved robustness, F1 score, and identical inference time. XCFSMN also has more comprehensive content and has gone through more experiments. To test the generalizability and performance of the system, we have included two datasets: WMMSD and the MobySound Database (Mellinger and Clark 2006). Besides CC-FSMN, we also compare the efficiency and performance of several CNNs (Lu, Han, and Yu 2021; Zhong et al. 2020; Allen et al. 2021; He et al. 2015) and transformers (Touvron, Cord, and Jégou 2022; Lee-Thorp et al. 2021; Zhang et al. 2021).

We can summarize our contributions as follows:

- We propose a teacher model that fuses features extracted from DensetNet121, TDNN, and a Cross-Covariance attended cFSMN. Through extensive empirical evaluations and comparisons, we achieve state-of-the-art performance.
- We apply a temperature cooling strategy to the knowledge distillation approach to mitigate the gap caused by model size and architecture differences between the teacher and the student model, increasing the effectiveness of the process.
- Knowledge distillation produces a well-trained student model that is significantly smaller—about 25 times—than its teacher counterpart. This smaller student model boasts the fastest inference speed while preserving an accuracy level almost identical to that of the teacher model. The markedly reduced size of the student model renders it highly portable, simplifying deployment across various ships and platforms.

Related Work

Deep Learning Based Methods

With the advancement of deep learning, transfer learning based on CNNs is applied to classify marine mammals based on acoustic signals. The ability of CNN to discriminate spectro-temporal information from spectrograms makes it an ideal network for processing acoustic information. It has been widely used in acoustic classification tasks such as bioacoustic classification, environmental sound classification, and underwater sonar image classification (Duan et al. 2022). In addition to excelling in visual representation classification tasks, the shift towards deep learning is fueled by its superiority in automatic detection and classification systems. These systems, rooted in deep learning, surpass human analysis in terms of efficiency and effectiveness. For instance, various fine-tuned and pre-trained CNN backbones are used to build accurate marine mammals' sound classifiers (Lu, Han, and Yu 2021; Zhong et al. 2020; Allen et al. 2021). However, these works mainly focus on limited whale species and are relatively inefficient. A different and novel approach turns the classification task into a regression problem. Instead of a classifier, a regressor is built using YOLO to predict bounding boxes in the spectrograms (Duan et al. 2022).

Later, the advent of self-attention mechanisms and transformers (Vaswani et al. 2017) has significantly changed deep learning. The first transformer stacks self-attention layers and outperforms the best result of that time (Vaswani et al. 2017). The idea of transformers inspires researchers to develop more transformer architectures (Huang and Zhang 2022). For example, a newly proposed model, DeiT III (Touvron, Cord, and Jégou 2022), is a variant of ViT that incorporates a new data enhancement procedure that includes Gaussian blur, solarization, and grayscale, and it achieves a competitive performance in image classification. In fact, CNN-based and transformer models produce some decent classifiers, but they are computationally expensive. To speed up the full attention transformer, Longformer (Beltagy, Peters, and Cohan 2020) specifies the full self-attention matrix, making it a more efficient architecture for long sequence data. However, the transformers require large amounts of data and strong computing power to be effective. FNet (Lee-Thorp et al. 2021) is an attention-free transform architecture that replaces the attention layers with a Fourier mixing sublayer with a feedforward sublayer to increase the efficiency of the transform. However, the FNet transformer achieves improved efficiency at the expense of accuracy.

Knowledge Distillation

Large models have excellent accuracy, but they are computationally expensive and difficult to deploy on any kind of ship. Knowledge distillation (KD) is a model compression technique to obtain an accurate small model by transferring knowledge from large models (Mirzadeh et al. 2020). KD has been widely used in face recognition tasks. For example, ShrinkTeaNet (Duong et al. 2019) proposes an angular distillation loss that tries to minimize the angle between teacher and student embedding vectors. KD has also been introduced to solve translation tasks, two sequence-level KDs have been shown to be useful for translation (Kim and Rush 2016). The performance of the student model can be affected by the gap between the student and teacher models. To effectively transfer knowledge to a smaller model, an intermediate teacher-assistant model is proposed to bridge the gap (Mirzadeh et al. 2020). Another way to enhance the knowledge transfer process involves using a grouped KD loss, which consists of three parts, proposed to filter out knowledge that is not related to facial identities(Zhao et al. 2023). To the best of the authors' knowledge, knowledge distillation has yet been applied to recognizing and classifying marine mammals.

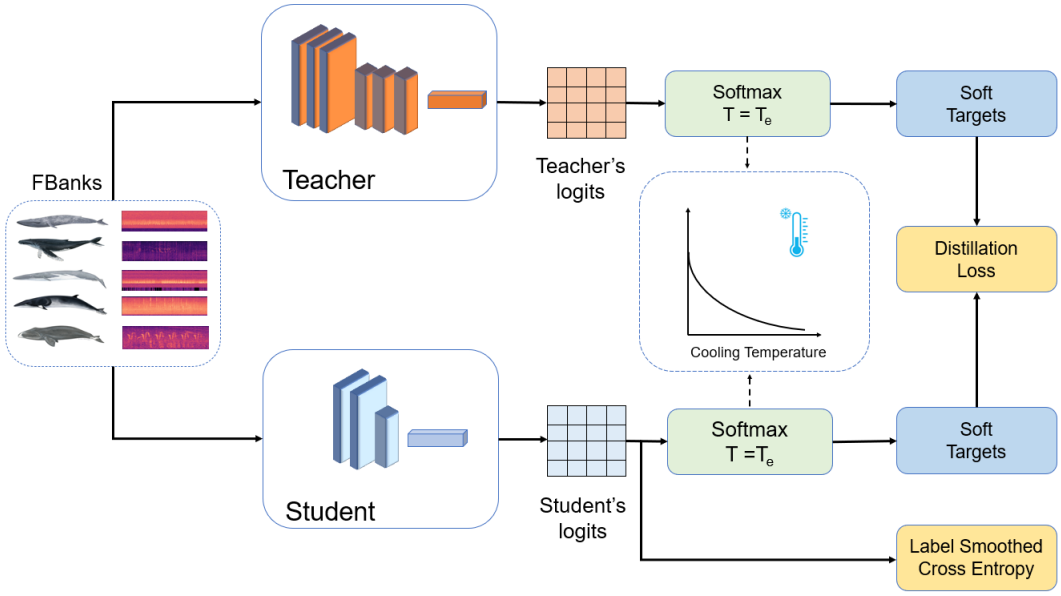

Figure 1: The overall flow of XCFSMN.

Methodology

The overall flow of the structure is shown in Fig 1. The framework consists of two networks: the teacher network, and the student network. The teacher network is a deep and complex neural network. In contrast, the student network is considerably smaller in scale. The teacher model will first be trained to obtain state-of-the-art accuracy. Then the pre-trained teacher model will transfer the knowledge to the student model to acquire and assimilate these learned features effectively.

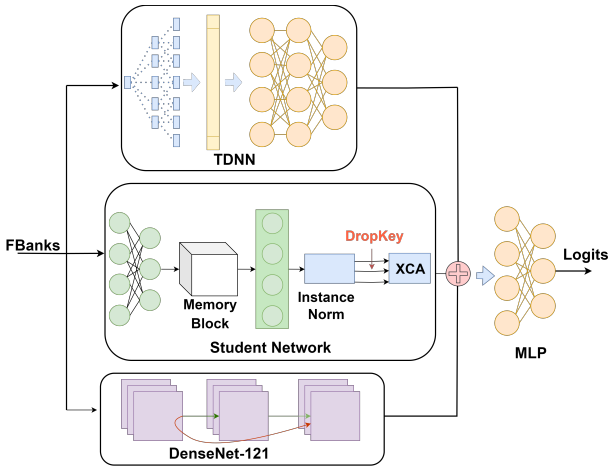

Figure 2: The structure of the teacher model.

Teacher Model

The teacher model consists of three major components: a pre-trained DenseNet network, a Time Delay Neural Network (TDNN), and an improved CC-FSMN. The pre-trained DenseNet121 extracts the spectral information of Log Mel-filter Banks (Fbanks), the TDNN extracts the animals' unique acoustic characteristics and the improved CC-FSMN learns the inter-frames context information. This architecture achieves state-of-the-art accuracy and transfers the knowledge to the student model.

Embedded X-vector X-vector (Snyder et al. 2018) is a type of feature embedding that is commonly used in speaker recognition tasks such as speaker verification, speaker identification, and speaker diarization. The X-vector is a fixed-length feature representation of variable-length speech segments, such as utterances, that can capture short-term and long-term variations in speaker characteristics. The X-vector can reduce intra-speaker variability and expand inter-speaker variability. Implementing the X-vector can help the model differentiate between cetaceans that produce sounds at similar frequencies.

The X-vector embedding model typically consists of five layers of Time-delay Neural Networks (TDNNs). The TDNNs extract relevant features from the speech signal. The resulting feature vectors are aggregated using a statistics pooling operation to produce a fixed-length X-vector. The X-vector is then used as input to a classifier, such as a support vector machine or neural networks, to perform the speaker recognition task.

Traditionally, the X-vector extractor undergoes isolated training, distinct from the main classifier. However, to enhance efficiency and reduce model size, we integrate the

X-vector model directly into the main model, embedding it seamlessly within the broader framework.

CC-FSMN The enhanced CC-FSMN has an identical architecture to the student model with a different number of neurons. The cFSMN (Zhang et al. 2016) layer processes the contextual information that links each frame in the data and is commonly used for automatic speech recognition. The XCA layers further strengthen the relationship between the frames. The justification for this step is that whales sing songs to communicate and the songs are like their languages. Different whale species have their unique songs so CC-FSMN can be used for the classification task. The details are explained in the next section.

DenseNet121 The teacher model consists of three main parts, a pre-trained DenseNet network, a Time Delay Neural Network (TDNN), and an improved CC-FSMN. The pre-trained DenseNet121 extracts the spectral information of the log Mel-filter banks (Fbanks), the TDNN extracts the unique acoustic characteristics of the animals, and the enhanced CC-FSMN learns the inter-frame context information. This architecture achieves state-of-the-art accuracy and transfers the knowledge to the student model.

Student Model

The key to obtaining an efficient classifier is to simplify the architecture of the student model. Our student model is an improvement of CC-FSMN (Liu and Cheng 2023), the compact Feed-Forward Sequential Memory Network is still the main layer within the network. We apply an InstanceNorm layer after the cFSMN layer. The normalization process resolves the covariant shift that may occur in the network, which helps to speed up the training and stabilize the gradient descent process.

An improvement was also been applied to the Cross-Covariance Attention (XCA) layer. Conventionally, the XCA layer contains a Dropout (Srivastava et al. 2014) operation that randomly drops out neurons in a layer to prevent the network from putting too much weight on certain features. To avoid overfitting and improve the robustness of the system, We replace the Dropout layer in the XCA with a DropKey (Li et al. 2023) layer. Instead of blocking out neurons in the attention layer, DropKey drops the input Key units, which are a type of embedding vector. Dropping Keys before computing the attention matrix can penalize weight peaks and thus regularize the weights. Therefore, the DropKey mechanism has improved attention weights regularization, secures the patterns, and can avoid the local-bias problem. Practically, DropKey is done by randomly generating a mask matrix D, and each element in the matrix D has a chance of d turning negative infinity. The output of the attention with a DropKey mechanism is formulated as

$$Axc(K, Q) = Softmax(d_j + \frac{\hat{K}^T \hat{Q}}{\tau}), \quad (1)$$

where $d_j \sim Bernoulli(1 - dr)$, Q and K stand for query and key respectively.

To further reduce the size of the student model, we prune about 30% of the neurons from the cFSMN layer and the MLP layer using the global unstructured pruning (GUP) method. Instead of pruning an entire layer, GUP prunes neurons based on their importance scores, providing better granularity. GUP does not improve inference speed but it can compress a model well, making the model more adaptable to various capacity-limited devices.

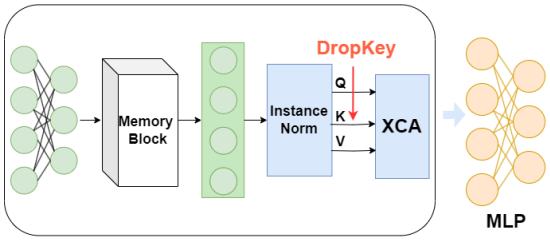

Figure 3: The structure of the student model.

Loss Function

The framework includes two loss functions: a Label Smooth Cross-Entropy Loss, and a Cooling Knowledge Distillation Loss. The Cross-Entropy Loss function is commonly used for classification tasks. The model predicts the probability for each label k. $p(k|x) = \frac{e^{z_k}}{\sum_{i=1}^{k} e^{z_i}}$, where z_i is the unnormalized log-probabilities. The Cross-Entropy loss function is then defined as $CE(i) = -\sum_{k=1}^{K} \log(p(k))q(k)$, and the $q(k)$ is the ground truth label distribution. The experiments adopt Label smoothing (Szegedy et al. 2016), which is a technique to regularize the model making it more robust by reducing the model's confidence. Label smoothing is achieved through the introduction of a smoothing term ϵ to the Cross-Entropy function turning it into a function $(1 - \epsilon) * CE(i) + \frac{\epsilon}{K}$. Hence, instead of predicting 1 for the correct class and 0 for others, the model predicts $1 - \epsilon$. The assessment of the student model relies, in part, on the utilization of the Label Smooth Cross-Entropy Loss.

The loss function employed for knowledge distillation is termed "Cooling Knowledge Distillation Loss." This strategy serves as a means to narrow the knowledge gap between the teacher and student models. In comparison to the conventional knowledge distillation loss, its formulation is as follows:

$$L_{KD} = -T^2 \sum_{c}^{i} \sigma_i(\frac{Z_t}{T}) \sigma_i(\frac{Z_s s}{T}). \quad (2)$$

The parameter L_{KD} is the cross-entropy between the teacher and student predictive probabilities $\sigma(Z_t)$ and $\sigma(Z_s)$. The T term denotes the temperature hyper-parameter. Increasing the T will increase the information received by the student model. Contrary to a common assumption, the effective pairing of a large pre-trained teacher model with a smaller student model is not a guaranteed outcome (Mirzadeh et al. 2020). Our proposed student model is approximately 18 times smaller than the teacher model. Hence, we believe that the conventional KD loss may not

be the best choice for this task. Moreover, most KD studies involve two networks with similar architecture but different layers and neurons. However, our proposed teacher model consists of a CNN backbone and a TDNN based on the student model. Hence, we propose to run a varying temperature strategy that reduces the knowledge gap caused by model size and architecture. The T is replaced with $T_c = T_{co}(1-\alpha)^n$ where T_{co} is the initial temperature, α is the decreasing factor and n is the number of decreases. Starting at $T_{co} = 20$, the temperature T_c of the KD loss will decrease exponentially as the training process progresses at a rate of $\alpha = 0.1$. The criterion for triggering the decrease is when the early stop counter reaches patience of 3. The reasoning behind this approach is as follows: initially employing a high temperature facilitates the transfer of a substantial volume of knowledge from the teacher model to the student model. Subsequently, as time progresses, the temperature gradually decreases. This deliberate reduction prompts the student model to gradually assimilate and comprehend features more independently. This is attributed to the inherent architectural distinctions between the student and teacher models. Consequently, this nuanced adjustment in the Knowledge Distillation process significantly enhances its effectiveness.

Experiments

Datasets

We employed two datasets to verify the robustness of the systems. WMMSD contains approximately 2000 sound recordings from more than 60 species of marine mammals. Spanning seven decades, these recordings potentially introduce challenges due to varying recording equipment and setups over time. MobySound is another database used in the experiments. The advantages of MobySound over the WMMSD are that the recordings in a sound set from one species were collected using a constant configuration. This ensures the consistency of the data features which makes feature extraction easier. In contrast, the variations in the sound collection configuration in the WMMSD allow us to test the robustness of the system and push the limits of what a classifier can achieve.

When cross-referenced, the two datasets have five classes in common; they are Finback Whales, Humpback Whales, Blue Whales, Minke Whales, and Bowhead Whales. The total recording lengths of the five classes are shown in Table 1. The sizes of the datasets show a significant difference. Furthermore, the classes are highly unbalanced which can generally lead to a biased model and poor generalization.

Class	WMMSD	MobySound
Finback Whale	3.9	21.0
Humpback Whale	1.98	2.12
Blue Whale	0.12	10.20
Minke Whale	0.05	3.70
Bowhead Whale	0.39	1.10

Table 1: Total Audio length of the five classes in hours.

Experiment Setup

The networks were created using Pytorch and trained using an A6000 GPU. The input shape of the features is 256x256, as this shape satisfies the input requirements of all the models used in the experiments. To generate a frame number of 256 for all classes, we chose an 8kHz sample rate for the FBanks extraction.

To verify the efficiency and performance of XCFSMN, we conducted comparative experiments with the following models: CC-FSMN (Liu and Cheng 2023), AlexNet used in (Lu, Han, and Yu 2021), ResNet50 used in (Allen et al. 2021), VGG16, DenseNet used in (Zhong et al. 2020), DeiT III (Touvron, Cord, and Jégou 2022), Longformer (Zhang et al. 2021), and FNet (Lee-Thorp et al. 2021). In addition to ResNet50, we also ran experiments on ResNet18, so that the network is tested in its smallest version for a comprehensive efficiency comparison.

Four sets of experiments were performed: a 12-class classification task using the WMMSD, a 5-class classification task for each of WMMSD and MobySound, and a 5-class classification task for a combined dataset of WMMSD and MobySound. All datasets were randomly split into a training set and a test set with a ratio of 0.3. The splitting process was performed at the file level before the feature extraction step for the 5-class classification task. First, we performed a 12-class classification task using the WMMSD. The subdataset contained the top 12 classes in descending order of sample size. The classes after the 12 classes have less than 100 samples, which is too few for meaningful training. This experiment serves as a validation of the models' capacity to effectively manage and process highly imbalanced multiclass data. Second, we trained and tested the models on the WMMSD and MobySound, respectively. These two experiments were performed to verify the ability of the models to train data from different sources. Then, the datasets were combined for training and testing to verify the robustness of the 5-class classifier, since the two datasets have different recording times, targets, and configurations.

To have a better assessment of the models, we applied a cross-validation of five splits. We split the training set into a training set and a validation set with five different random states. The results were then averaged to obtain the final results. All models were fine-tuned for the best comparison result.

Evaluation Metrics

To evaluate the efficiency of the models, we compared training time and inference time. Training time is the time taken for the model to complete 300 epochs or to trigger an early stop. Inference time refers to the time taken for the model to predict the classes of the fixed-size test set. A shorter training time or inference time indicates better efficiency. Also, inference time is more important than training time for the marine mammal classification tasks.

Instead of using accuracy, the F1 score can provide a more accurate measure of the models when the datasets are highly

Model	# Params (M)	WMMSD-12		WMMSD-5		MobySound-5		Merged-5	
		TT	IT	TT	IT	TT	IT	TT	IT
AlexNet	57.0	5	0.4	2	**0.1**	17	**0.2**	2	0.4
ResNet18	11	5.5	0.9	15	0.2	5	0.6	5	0.8
ResNet50	23.5	13	2.7	15	0.6	7	1.6	6.3	2.2
VGG16	134.0	33	5.9	10	1.0	15	2.7	17	3.7
DenseNet	7.0	27	3.5	3	0.7	18	2.0	31	2.6
DeiT III	21.6	32	16.3	6	4.7	52	11.5	23	2.6
Longformer	16.9	47	8.4	14	11.0	12	4.6	19	6.6
FNet	**3.1**	**4.5**	0.7	1.3	1.2	5	3.2	**0.33**	0.5
CC-FSMN	4.5	5	**0.32**	**0.7**	**0.1**	**3**	**0.2**	2	**0.3**
Teacher (ours)	81.3	12	10.8	3	1.2	22	10.3	5	4
Student (ours)	3.15	5.8	**0.32**	1	**0.1**	6.5	**0.2**	12	**0.3**

Table 2: Comparative results of the models on efficiency. 'TT' denotes the training time in minutes and 'IT' denotes the inference time in seconds.

Model	WMMSD-12			WMMSD-5			MobySound-5			Merged-5		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
AlexNet	0.594	0.735	0.612	0.636	0.607	0.607	0.865	0.884	0.863	0.867	0.800	0.782
ResNet18	0.782	0.846	0.806	0.791	0.616	0.646	0.950	0.873	0.880	0.934	0.809	0.840
ResNet50	0.787	0.850	0.809	0.791	0.616	0.646	0.928	0.845	0.846	0.904	0.828	0.842
VGG16	0.762	0.842	0.789	0.891	0.823	0.783	0.860	0.833	0.827	0.784	0.801	0.790
DenseNet	0.705	0.810	0.735	0.842	0.819	0.728	0.957	0.897	0.911	0.916	0.830	0.869
DeiT III	0.569	0.660	0.586	0.781	0.704	0.660	0.892	0.860	0.896	0.928	0.740	0.787
Longformer	0.237	0.274	0.215	0.472	0.247	0.275	0.657	0.705	0.645	0.668	0.594	0.627
FNet	0.120	0.225	0.120	0.318	0.430	0.228	0.275	0.561	0.283	0.261	0.296	0.222
CC-FSMN	0.785	0.828	0.787	0.791	0.616	0.646	0.970	0.898	0.922	0.893	0.817	0.828
Teacher (ours)	**0.855**	**0.858**	**0.853**	**0.946**	**0.828**	**0.827**	**0.981**	**0.991**	**0.986**	**0.976**	**0.890**	**0.925**
Student (ours)	0.816	0.837	0.820	0.900	0.824	0.791	0.970	0.977	0.973	0.944	0.890	0.910

Table 3: Comparative results of the models on performance. "Prec." denotes precision and "Rec." denotes recall.

imbalanced. The Macro F1 score is typically used to evaluate the performance of the models and it is the average of the F1 scores of each class. F1 score is calculated through

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (3)$$

Precision in Equation (3) is defined as the percentage of the correct labels out of all the predicted positives. A model with high recall indicates that the model can find all the positive cases with some negative cases identified as positive. Since the goal is to detect and protect all types of marine mammals, a high recall is important.

Results Comparison

The efficiency results are shown in Table 2 and the performance results are shown in Table 3. Of the 11 models tested, FNet is the smallest in model size followed by the proposed student model which is only approximately 1.6% larger. CC-FSMN (Liu and Cheng 2023) and the FNet (Lee-Thorp et al. 2021) show excellent training speed in all the experiments. The teacher model takes much longer to train as expected. The student model is slightly slower in training speed. CC-FSMN and the student model have identical infer speeds and are dominant over other models in all experiments. This indicates that our proposed framework and CC-FSMN achieve state-of-the-art efficiency.

Table 3 provides clear evidence of the teacher model's consistent and exceptional performance, displaying superior precision, recall, and F1 Score consistently throughout all experiments. Implementing the temperature cooling knowledge distillation process, the student model has significant improvements over the CC-FSMN. There is a minimum improvement of 0.064 in the absolute F1 Score and a maximum of 0.145. Although the student model cannot obtain as high an F1 Score as the teacher model, it still outperforms all other models since it has almost the smallest size.

The CNNs have similar performance except for AlexNet, which shows significant drops in the 12-class WMMSD and the merged dataset. It is evident that all the transformers perform poorly. The poor performance can be attributed to the small dataset size as the vision transformers require large datasets to demonstrate their strength for image classification tasks (Strudel et al. 2021). Deit can hold its own in the 5-class experiments but fails for the 12-class because the 12-class WMMSD has a really limited sample size for the minority class. Longformer performs poorly since it was designed for long-sequence languages. Furthermore, the replacement of the attention mechanism with the Fourier Mixing Layer in FNet restricts the model from capturing long-sequence context information, reducing the model's ability to classify the signal.

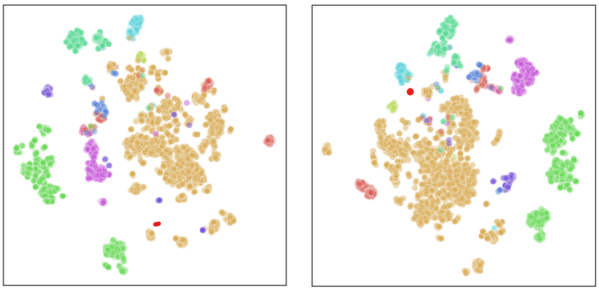

Figure 4: Visualization of features using a t-SNE plot for the 12-class WMMSD problem.

To better illustrate the rankings of the model, we performed The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). 0.1 is assigned to the weights of precision, recall, and training time; 0.35 is assigned to the F1 score and inference time since fast and accurate identification of mammals is prioritized. The average performance scores indicate that the student model is the best choice followed by CC-FSMN and then ResNet18.

In addition, we compared the results using the image approach in (Pollicelli, Coscarella, and Delrieux 2020) and (Maire, Alvarez, and Hodgson 2015). The mean F1 score achieved by the model trained using aerial images in (Maire, Alvarez, and Hodgson 2015) is 0.4059 and the classification using Commerson's dolphin photos in (Pollicelli, Coscarella, and Delrieux 2020) is 0.883. XCFSMN has a better F1 score and can identify multiple species. Therefore, marine mammal classification based on acoustic signals with XCFSMN is better than the same classification problem based on image processing with deep learning.

Ablation Study

We performed an ablation study for the student model shown in Table 4 by comparing efficiency and performance using the 5-class merged dataset. The first row shows the results for a CC-FSMN (Liu and Cheng 2023) and behaves as a baseline. Adding Instance Normalization to the output of cFSMN increases the precision of the model and is able to maintain the recall. Replacing Dropout with DropKey increases the precision more but decreases the recall. The combination of Instance Normalization and DropKey boosts the student model's precision without affecting its recall. Finally, the introduction of temperature-cooling KD increases the recall and the F1 score to reach the state-of-the-art.

The second ablation study on the KD temperature is shown in Table 5. Experiments were conducted with different fixed temperatures ranging from 1 to 40 and the best one was selected for comparison. We introduced a temperature-increasing strategy that contrasts with our primary approach. Our findings suggest that maintaining a fixed temperature doesn't maximize the advantages of knowledge distillation (KD). By dynamically adjusting the temperature parameter—either increasing or decreasing—during the KD process, we observed enhanced effectiveness in the student model's acquisition of knowledge from the teacher. Interestingly, both warming and cooling strategies demonstrate comparable training times, yet the cooling approach notably enhances recall and F1 score in our experiments.

Sturcture						
InsNorm	DropKey	KD	IT	Prec.	Rec.	F1
×	×	×	0.32	0.893	0.817	0.828
✓	×	×	0.32	0.939	0.814	0.839
×	✓	×	0.38	0.954	0.805	0.844
✓	✓	×	0.38	0.951	0.817	0.850
✓	✓	✓	0.33	0.944	0.890	0.910

Table 4: Results of the ablation study on the student model. The denotations are the same as in Tables 2 and 3.

Temperature Strategy	TT	IT	Prec.	Rec.	F1
Fix T = 10	1.5	0.3	0.918	0.836	0.854
Warming	3.5	0.3	0.943	0.859	0.887
Cooling	3.5	0.3	0.944	0.890	0.910

Table 5: Results of the ablation study on the KD temperature strategy. The denotations are the same as in Tables 2 and 3.

To provide a more intuitive result, we generated the t-distributed stochastic neighbor embedding (t-SNE) to visualize the high-dimensional information output layer neurons. Fig 4 shows that the teacher model has the best ability to isolate different classes. The student model isolates class 0 and class 6 better than CC-FSMN. Therefore, we can claim that the XCFSMN framework can better capture the small differences between classes and is more robust than CC-FSMN.

Conclusion and Future work

This paper proposed the XCFSMN framework, which trains a small and fast student model using a pre-trained teacher model with a temperature-cooling strategy. XCFSMN outperforms CNN backbones and transformers in efficiency and performance for marine mammals classification. XCFSMN also demonstrates that the acoustic approach has better accuracy than the approach based on images. XC-FSMN has significant improvements in performance and similar efficiency compared to CC-FSMN. The pruned network makes it easier to deploy on different devices. To further improve the classifier, one can work on the preprocessing of the data instead of the architecture of the model. In the future, we aim to employ a diffusion model (Ho, Jain, and Abbeel 2020) to denoise the data and generate data to balance the dataset. In addition, XCFSMN is not limited to marine mammals' sound classification. The idea of XCFSMN can be further applied to acoustic signal classification tasks such as speech sentiment recognition and speaker identification.

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.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Bowen, W. D. 1997. Role of Marine Mammals In Aquatic Ecosystems. *Marine Ecology Progress Series*, 158: 267–274.

Buck, E. H.; and Clavert, K. 2005. Active Military Sonar and Marine Mammals: Events and References References.

Duan, D.; gang Lü, L.; Jiang, Y.; Liu, Z.; Yang, C.; Guo, J.; and Wang, X. 2022. Real-time Identification of Marine Mammal Calls Based on Convolutional Neural Networks. *Applied Acoustics*, 192.

Duong, C. N.; Luu, K.; Quach, K. G.; and Le, N. 2019. ShrinkTeaNet: Million-scale Lightweight Face Recognition via Shrinking Teacher-Student Networks. In *Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition (CVPR)*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Conference on Neural Information Processing Systems (NeurIPS)*, 6840–6851.

Huang, M.; and Zhang, L. 2022. Atrous Pyramid Transformer with Spectral Convolution for Image Inpainting. 4674–4683. Association for Computing Machinery (ACM). ISBN 9781450392037.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. *Empirical Methods in Natural Language Processing*.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontanon, S. 2021. FNet: Mixing Tokens with Fourier Transforms. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Li, B.; Hu, Y.; Nie, X.; Han, C.; Jiang, X.; Guo, T.; and Liu, L. 2023. DropKey. In *Computer Vision and Pattern Recognition (CVPR)*.

Liu, X.; and Cheng, J. 2023. A Highly Efficient Marine Mammals Classifier Based on a Cross-Covariance Attended Compact Feed-Forward Sequential Memory Network (Student Abstract). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Lu, T.; Han, B.; and Yu, F. 2021. Detection and Classification of Marine Mammal Sounds Using AlexNet With Transfer Learning. *Ecological Informatics*, 62.

Maire, F.; Alvarez, L. M.; and Hodgson, A. 2015. Automating Marine Mammal Detection in Aerial Images Captured During Wildlife Surveys A Deep Learning Approach. In *28th Australasian Joint Conference on Artificial Intelligence*, 379–385.

Mellinger, D. K.; and Clark, C. W. 2006. MobySound: A Reference Archive For Studying Automatic Recognition of Marine Mammal sounds. *Applied Acoustics*, 67: 1226–1242.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; Ghasemzadeh, H.; and Shaw, D. 2020. Improved Knowledge Distillation via Teacher Assistant. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Parsons, E. C. 2017. Impacts of Navy Sonar on Whales and Dolphins: Now Beyond A Smoking Gun? *Frontiers in Marine Science*, 4: 295–306.

Pollicelli, D.; Coscarella, M.; and Delrieux, C. 2020. RoI Detection and Segmentation Algorithms for Marine Mammals Photo-identification. *Ecological Informatics*, 56.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. volume 2018-April, 5329–5333. Institute of Electrical and Electronics Engineers Inc. ISBN 9781538646588.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15: 1929–1958.

Strudel, R.; Garcia, R.; Inria, I. L.; and Inria, C. S. 2021. Segmenter: Transformer for Semantic Segmentation. In *International Conference on Computer Vision*, 7242–7252.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; and Shlens, J. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taylor, B. L.; Martinez, M.; Gerrodette, T.; Barlow, J.; and Hrovat, Y. N. 2007. Lessons From Monitoring Trends in Abundance of Marine Mammals. *Marine Mammal Science*, 23: 157–175.

Touvron, H.; Cord, M.; and Jégou, H. 2022. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision*, 516–533.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Zhang, P.; Dai, X.; Yang, J.; Xiao, B.; Yuan, L.; Zhang, L.; and Gao, J. 2021. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. *International Conference on Computer Vision (ICCV)*.

Zhang, S.; Jiang, H.; Xiong, S.; Wei, S.; and Dai, L. 2016. Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition. 3389–3393.

Zhao, W.; Zhu, X.; Guo, K.; Zhang, X.-Y.; and Lei, Z. 2023. Grouped Knowledge Distillation for Deep Face Recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Zhong, M.; Castellote, M.; Dodhia, R.; Ferres, J. L.; Keogh, M.; and Brewer, A. 2020. Beluga Whale Acoustic Signal Classification Using Deep Learning Neural Network Models. *The Journal of the Acoustical Society of America*, 147: 1834–1841.