DEEP EMBEDDED CLUSTERING FOR BIOACOUSTIC CLUSTERING OF MARINE MAMMAL VOCALIZATION

Ali Jahangirnezhad, Afra Mashhadi
Computing and Software Systems
University of Washington
Bothell, Washington, USA
{arshiaj,mashhadi}@uw.edu

ABSTRACT

With the decrease of hardware costs, stationary hydrophones are increasingly deployed in the marine environment to record animal vocalizations amidst ocean noise over an extended period of time. Bioacoustic data collected in this way is an important and practical source to study vocally active marine species and can make an important contribution to ecosystem monitoring. However, a main challenge of this data is the lack of annotation which many supervised neural network models rely on to learn to distinguish between noise and marine animal vocalizations. In this paper, we posit an unsupervised deep embedded clustering based on LSTM autoencoders, that aims to learn the representation of the input audio by minimizing the reconstruction loss and to simultaneously minimize a clustering loss through Kullback–Leibler divergence.

1 INTRODUCTION

Bioacoustic data is a valuable large-scale source of data that can help marine scientists gain information about vocally active marine species Norris et al. (1999); Cummings & Holliday (1985); Stafford et al. (1998); Abadi et al. (2018), and invaluable information for ecosystem monitoring. For example, in a recent study researchers have used AI with the acoustic data from a hydrophone dropped into the Santa Barbara Channel in order to identify blue, humpback and fin whales in near real-time. A surface buoy then transmits the data to scientists at Texas A&M Galveston for review and confirmation, so to avoid possible cargo ship collisions. Mehta et al. Mehta et al. (2020) also proposed a system that uses a Convolutional Neural Network algorithm on a two-category classification problem (presence/absence of a whale call).

One of the challenges with the bioacoustic data is the lack of annotated labels that is required by supervised deep neural networks. To account for the scarcity of the labels, auDeep Freitag et al. (2017) was proposed, which relies on a representation learning model based on Long-short term memory autoencoders (hereafter LSTM AE) in order to learn the feature representation of the audio input. Using auDeep, it is possible to train the model by optimizing representation loss and then use high dimensional clustering algorithms such as t-SNE Van der Maaten & Hinton (2008) to cluster the learned representations Schuller et al. (2019); Best et al. (2020).

Motivated by auDeep, we propose a deep embedded clustering model for LSTM AE. Deep clustering utilizes representation learning in order to learn features in an unsupervised setting. More specifically, by adding a clustering layer to the previously proposed LSTM AE, we are able to train a model that simultaneously minimizes both the reconstruction loss (Mean Squared Error) and a clustering loss, formulated as Kullback–Leibler divergence.

We validate our approach on the Orca Activity Sub Challenge. Our preliminary results show that by including a clustering layer and accounting for the clustering loss at the time of training, our results show a significant improvement compared to the non-clustering LSTM AE counterpart. Our initial experiments also demonstrate that a clustering weight (γ) between 0.5 and 1.5 gives the optimum performance in terms of accuracy.

2 RELATED WORK

The past works on analysing bioacoustic data can be grouped into two groups: [1]Traditional machine learning approaches, where various researchers have used traditional signal processing and speech recognition techniques, such as dynamic time warping, hidden Markov and Gaussian mixture models, as well as spectrogram correlation to develop algorithms in order to detect dolphin and whale vocalizations Brown et al. (2010); Brown & Smaragdis (2009); Schwock & Abadi (2020); Abadi (2018), [2]Linear techniques such as discriminant function analysis, random forest classifiers, decision tree classifications, and support vector machines applied in conjunction with mel-frequency spectrum coefficients for killer whale sound detection/classification. However, traditional machine-learning algorithms have been shown to perform worse than modern deep learning approaches, especially when the dataset contains a comprehensive amount of annotated data.

More recently, Grill & Schlüter (2017) adopted feed forward convolutional neural networks (CNNs) trained on mel-scaled log-magnitude spectrograms in a bird audio detection challenge. Google AI Harvey et al. (2018) Perception has recently successfully trained a convolutional neural network (CNN) in detecting humpback whale calls from over 15 years of underwater recordings captured at several locations in the Pacific. In 2020, Best et al. (2020) proposed a CNN based model that was trained on 11,509 killer whale (Orcinus orca) signals and 34,848 noise segments. The resulting toolkit ORCA-SPOT was tested on a large-scale bioacoustic repository – the Orchive – comprising roughly 19,000 hours of killer whale underwater recordings. Similarly, Mehta et al. (2020) proposed a system that uses a Convolutional Neural Network algorithm on a two-category classification problem (presence/absence of a whale call). They then visualized the predictions and asked expert annotators for their input so to verify and correct the wrong predictions. This feedback loop enabled supplementing the existing training dataset with additional annotations, and also engaged the users in the annotation process. They showed that the active learning system improves the performance of the model as they presented that the f1-score increased from 0.83 to 0.84 with 50 new annotations corresponding to 3% increase of the labeled dataset. In contrast to the above works, our approach is designed for unsupervised representation learning, allowing us to use a large volume of unlabeled data to reduce the reconstruction loss and the clustering loss simultaneously.

3 DEEP EMBEDDED CLUSTERING BASED ON LSTM AUTOENCODER

A Deep Embedded Clustering network (such as those proposed by Xie et al. (2016) and Guo et al. (2017)) is composed of two main components: an AutoEncoder (AE) which is used in order to learn the hidden representation of the data, and a clustering layer that is used to group the embedded points together. Similar to other variations of autoencoders, we construct a LSTM AE which is composed of an encoder part $f_\omega(.)$ and a decoder $g_{\omega'}(.)$ respectively. The LSTM AE aims to find a code for each input sample by minimizing the mean squared errors (MSE) between its input and output over all samples such that $x' = g_\omega(f_\omega(x))$.

The embedded layer corresponds to the latent features (also known as the code). The decoding parts of the network are the mirror construction of the described part in which the embedded features are transformed back to the original input. The objective of the LSTM AE part is to minimize the reconstruction loss denoted as L_r and is measured as mean squared error $L_r = \frac{1}{n}\sum_{i=1}^{n} \|G_{\omega'}(F_\omega(x_i)) - x_i\|$, where n is the number of audio clips in the dataset, and x_i is the ith input. Figure 1 shows spectrograms of three audio clips alongside their respective reconstructed version.

In addition to reducing the reconstruction loss, the Deep Embedded Clustering networks aim to minimize a clustering loss function. This is done by creating a clustering layer connected to the embedded layer of the LSTM AE. The clustering layer maps each learned representation (z_i) of the input sound x_i into a soft label. The clustering loss L_c is then defined as Kullback-Leibler divergence (KL divergence) between the distribution of soft labels and a predefined target distribution.

$$L_c = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{1}$$

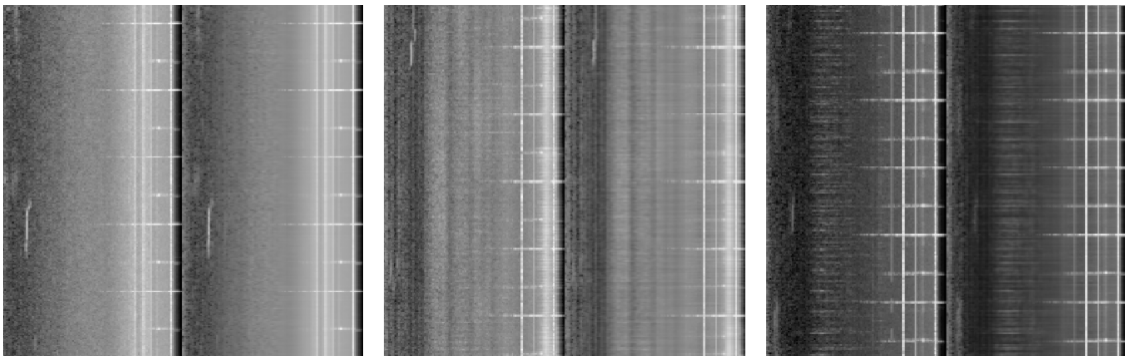

Figure 1: Three pairs of spectrograms from the original audio clip inputted to the model (left of each pair) and the reconstructed audio clip (right of each pair)

where q_{ij} is the similarity between embedded point z_i and cluster center μ_j measured by Student's t-distribution (2), and the target distribution p_{ij} is formulated as in (3).

$$q_{ij} = \frac{\sum_j 1 + ||z_i - \mu_j||^2}{1 + ||z_i - \mu_j||^2} \tag{2}$$

$$pij = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \tag{3}$$

The end to end network objective is to reduce a loss function that is a weighted sum of the reconstruction loss (L_r) and the clustering loss (L_c).

$$L = \alpha L_r + \gamma L_c \tag{4}$$

In this setting, a network with $\gamma = 0,\ \alpha = 1$ results in a LSTM AE network similar to the previous model proposed by auDeep Freitag et al. (2017), and a network with $\gamma = 1,\ \alpha = 0$ would only function as a clustering network. For the purposes of this paper, we will hereafter assume $\alpha = 1$, and change γ in order to study its effects.

4 Experimental Evaluation

Dataset: To evaluate our model we use the Orca Activity (OA) Sub-Challenge Dataset as collected by the DeepAL Fieldwork Data project[1]. This data was collected on a 15-meter research trimaran in 2017 and 2018 in Northern British Columbia. For the OA sub-challenge, we use a sub-sample that amounts to a total duration of 4.6 hours (sound files: range 0.3-5.0 s; mean duration 1.23 ± 0.96 s). The two classes to be told apart are Noise and Orca sounds.

Experiment Setup: Audio clips are cut or padded with silence in order to achieve a fixed length data set. This fixed length is calculated in a way that minimizes the amount of data loss across the data set. Mel Spectrograms with magnitude 2 and 128 bands are then extracted from the audio clips, and their time-frequency data is stored.
The encoder network comprises one LSTM layer and two back to back Dense layers with dropouts, at the end of which a feature vector is produced. Then the network is divided to two branches: the clustering layer and the decoder layers. The decoder layers are a mirror of the encoder layers. Figure 2 shows the overall architecture of our model.

[1] https://lme.tf.fau.de/dataset/deepal-fieldwork-data-2017-2018-dlfd/

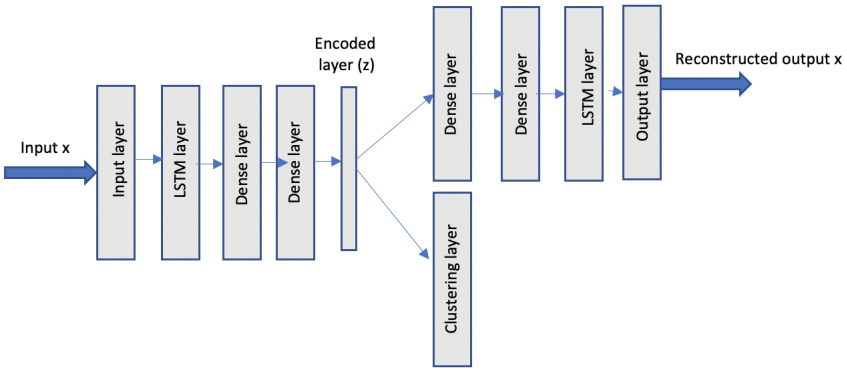

Figure 2: Our model representing the LSTM encoder and decoder and the appended clustering layer.

5 PRELIMINARY RESULTS

In this section we present our preliminary results of our model when trained with learning rate of 0.001, momentum of 0.99 on standard Adam Optimizer parameters. We evaluate our model by measuring clustering accuracy (ACC) and Normalized Mutual Information (NMI) which are widely used in unsupervised learning scenarios. We first measure the training Accuracy and NMI for when the clustering and reconstruction loss are both regarded equally ($\alpha = 1, \gamma = 1$) and compare it to the state-of-the-art models which rely only on the reconstruction loss. We observed a significant (30%) increase in accuracy for the training and a 10% increase for the test set. To understand the impact of γ on the clustering performance, Figure 3 presents our model accuracy and NMI for variable γ. As we can see the increase in the γ initially leads to a better performance, but then it plateaus. The figure also presents the NMI values for various γ. shortcoming in future.

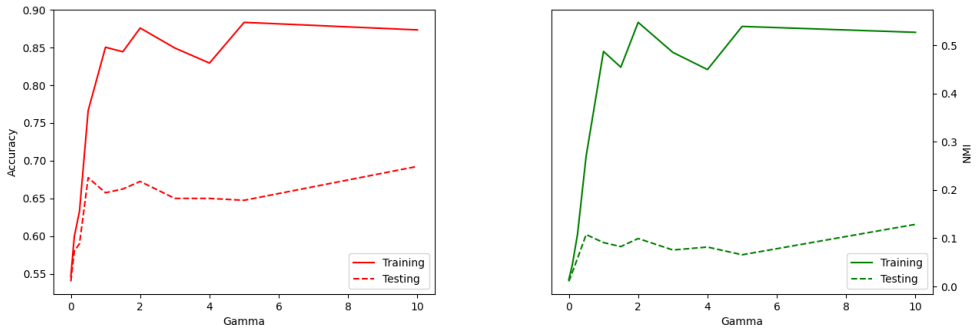

Figure 3: Accuracy and NMI for a varying clustering weight γ, where $\gamma = 0$ corresponds to the state of the art models based on the LSTM AE and without KL divergence optimization.

Figure 4 shows how different γ values affect the reconstruction loss from the AE, and clustering loss from the clustering layer. Our experiments show that higher γ values result in a lower clustering loss and faster convergence of the clustering loss, while keeping the AE loss as low as before.

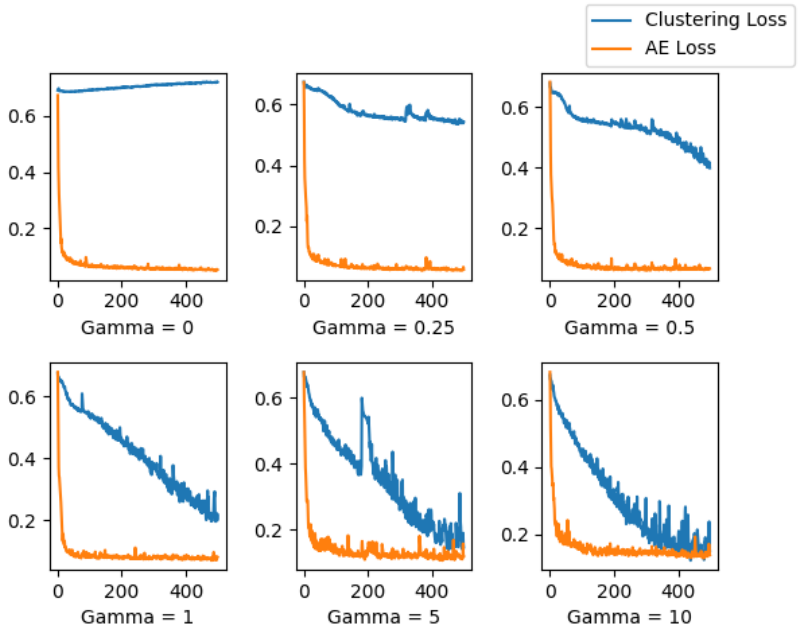

Figure 4: Clustering and Auto-Encoder loss for different γ values

In general, as it can be seen in Figure 4, higher values of γ result in a much faster decrease in the clustering loss. However, it is also important to pay attention to the possible distortion of our autoencoder, specially when γ values are huge ($\gamma >> 10$). On the other hand, higher gamma values put the network in danger of over-fitting. Based on our preliminary experiments and observations, $\gamma = 1$ provided the best overall results for both the autoencoder and the clustering layer. With this setting, the network gives the same weight to clustering as autoencoding. This observation may however be very different for different data sets, different number of clusters, etc.

Unlike accuracy, we find that our model suffers from low NMI, but we also observed the same pattern of increase in NMI as we observed with pure accuracy, and that is a steady increase with higher values of γ. Even though our preliminary results do not reflect the effects of larger γ as of yet, we predict a higher NMI and accuracy value based on the trends observed in our experiments. It is however imperative that we have in mind the fact that all of our experiments are done with only 2 clusters (namely noises and Orca sounds clusters).

This model works with the weighted summation of two different losses, and this makes the model very prone to getting stuck at a local minima. This is in part because of the fact that both loss values change in an iteration, but their weighted summation may stay relatively the same. Based on our preliminary results, higher learning rates converge very quickly, but fail to converge, while a small learning rate gets stuck at a local minima in a few epochs. We found that the best results are achieved by using a variable learning rate.

6 Conclusion

Using LSTM networks parallel to clustering can lead to increased accuracy in unsupervised learning from time-series data. The clustering layer proposed in this paper not only tries to label the unlabeled data, but it also affects the autoencoder so that the features that affect the clustering most are the ones focused on during encoding. This leads to a low dimensional feature vector which captures the most important elements of a time-series data with respect to the clusters.

As the model has produced acceptable results in its preliminary training, we plan on fine tuning the hyper parameters to the specific task of identifying marine mammals. We are also

working on implementing an active learning environment in collaboration with the Megaptera (2021) online game, where users are prompted to label an audio file, in order to refine our model. Afterwards, the model will be tested with a real time stream of hydrophone recordings provided by the Ocean Observatories Initiative (OOI (2021)) in order to detect mammal presence in the vicinity of a hydrophone, in real time.

REFERENCES

Shima Abadi. Using machine learning in ocean noise analysis during marine seismic reflection surveys. *The Journal of the Acoustical Society of America*, 144(3):1744–1744, 2018.

Shima Abadi, Derek Flett, Ryan Berge, Jeremy DeHaan, Virdie Guy, Urooj Qureshi, and Michael Cook. Studying underwater sound level caused by bridge traffic in lake washington. *The Journal of the Acoustical Society of America*, 144(3):1808–1808, 2018.

Paul Best, Maxence Ferrari, Marion Poupard, Sébastien Paris, Ricard Marxer, Helena Symonds, Paul Spong, and Hervé Glotin. Deep learning and domain transfer for orca vocalization detection. In *International joint conference on neural networks*, 2020.

Judith C Brown and Paris Smaragdis. Hidden markov and gaussian mixture models for automatic call classification. *The Journal of the Acoustical Society of America*, 125(6):EL221–EL224, 2009.

Judith C Brown, Paris Smaragdis, and Anna Nousek-McGregor. Automatic identification of individual killer whales. *The Journal of the Acoustical Society of America*, 128(3):EL93–EL98, 2010.

William C Cummings and DV Holliday. Passive acoustic location of bowhead whales in a population census off point barrow, alaska. *The Journal of the Acoustical Society of America*, 78(4):1163–1169, 1985.

Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1):6340–6344, 2017.

Thomas Grill and Jan Schlüter. Two convolutional neural networks for bird detection in audio signals. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1764–1768. IEEE, 2017.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pp. 1753–1759, 2017.

Matt Harvey et al. Acoustic detection of humpback whales using a convolutional neural network. *Google AI Blog*, 2018.

Megaptera, 2021. URL `http://megaptera.swipesforscience.org`.

Kunal B Mehta, Jorge Rodriguez Saltijeral, Jesse Lopez, Abhishek Singh, Valentina Staneva, Scott Veirs, and Val Veirs. Active listening and learning for orca sound detection. *The Journal of the Acoustical Society of America*, 148(4):2728–2728, 2020.

Thomas F Norris, Mark Mc Donald, and Jay Barlow. Acoustic detections of singing humpback whales (megaptera novaeangliae) in the eastern north pacific during their northbound migration. *The Journal of the Acoustical Society of America*, 106(1):506–514, 1999.

OOI. Ocean observatories initiative, 2021. URL `https://oceanobservatories.org/`.

Björn Schuller, Anton Batliner, Christian Bergler, Florian B Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, et al. The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. *International Symposium on Computer Architecture (ISCA)*, 2019.

Felix Schwock and Shima Abadi. Statistical analysis and modeling of wind-generated ocean noise in the northeast pacific ocean. *The Journal of the Acoustical Society of America*, 148(4):2688–2688, 2020.

Kathleen M Stafford, Christopher G Fox, and David S Clark. Long-range acoustic detection and localization of blue whale calls in the northeast pacific ocean. *The Journal of the Acoustical Society of America*, 104(6):3616–3625, 1998.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487, 2016.