North Atlantic Right Whale Call Detection with Convolutional Neural Networks

Evgeny Smirnov EVGENY.VERSUS.SMIRNOV@GMAIL.COM
Saint-Petersburg State University, Universitetskii prospekt 35, Petergof, Saint-Petersburg, Russia 198504

Abstract

Convolutional Neural Networks (CNN) have shown success in many image processing and speech recognition tasks. In this paper we propose to apply Convolutional Neural Network to the bioacoustic task of whale call detection. We trained a CNN to detect whale calls in 2-second audio clips in the Marinexplore and Cornell University Whale Detection Challenge. On the test data we achieved 2.4% error rate. Our best neural network consists of three convolutional layers followed by max-pooling layers, one fully-connected layer and final 2-way softmax layer. We used maxout hidden units with dropout to improve accuracy and reduce overfitting.

1. Introduction

The North Atlantic right whale, *Eubalaena glacialis*, is in danger of extinction (Kraus et al., 2005). One of the main threats to whale survival is high human activity in the areas of their migration. One third of all right whale mortalities are caused by collisions with ships and entanglement in fishing gear.

One way to reduce whale mortality is monitoring for the occurrences of whales by detecting their sounds on data recordings (Spaulding et al., 2009). Right whale species produce many different sounds, but most frequent and distinct one is a contact call ("up-call"). Automatic detection of such calls became a popular method of detecting right whales, and now there is a need for good algorithms of call detection in raw audio data.

There are already several different approaches to this task (Mellinger & Clark, 2000), (Urazghildiiev & Clark, 2006), (Urazghildiiev & Clark, 2007), (Urazghildiiev et al., 2009), (Dugan et al., 2010a). One of them is neural network approach (Dugan et al., 2010b). In this paper we try to improve it by using state-of-the-art type of neural networks (Convolutional Neural Networks with maxout hidden units) in the Marinexplore and Cornell University Whale Detection Challenge [1].

2. Dataset

The Marinexplore and Cornell University Whale Detection Challenge team provided us with a dataset of 30,000 training samples and 54,503 testing samples. Each sample is a 2-second .aiff sound clip with a sample rate of 2 kHz. Dataset contains mixture of right whale calls, non-biological noise and other sounds. The task was to create an algorithm for detecting right whale calls and to beat the existing whale detection algorithm of Cornell University.

For our experiments we compute Mel-frequency cepstral coefficients (MFCCs) along with their first and second temporal derivatives, and Fourier-transform-based filter-banks for all sound clips.

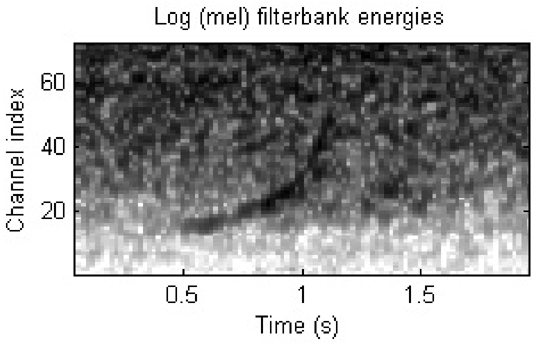

Figure 1. Example of filter-bank representation of a sound clip, containing right whale call

ICML 2013 Workshop on Machine Learning for Bioacoustics, Atlanta, Georgia, USA, 2013. Copyright 2013 by the author(s).

[1] http://www.kaggle.com/c/whale-detection-challenge

MFCCs were calculated with Hamming window, frame length of 25 ms and frame shift of 10 ms, for whole 2-second sound clip, so there were 2010 total input values (MFCCs with first and second derivatives) for each example. Filter-banks were calculated in range of 50 - 650 Hz, and include 72 coefficients, distributed on mel scale, for each of the 97 time steps.

3. Model

We used two different types of neural networks in our experiments: fully-connected Neural Network (NN) with sigmoid hidden units (Rumelhart et al., 1986), and Convolutional Neural Network (CNN) (LeCun et al., 1998) with maxout hidden units (Goodfellow et al., 2013).

3.1. Fully-connected Neural Network

This kind of neural networks was already used for whale call detection (Dugan et al., 2010b). We tried to improve its performance by using larger neural network and new regularization technique called “dropout” (Hinton et al., 2012). We tried several architectures with different parameters, and our best one consisted of 2010 units in input layer, 2000 sigmoid units in first and second hidden layers, and 2-way softmax layer. We used MFCC-based vector as input, and trained neural network for 500 epochs with backpropagation with batch size of 100, starting learning rate of 1 (reduced linearly for 300 epochs to the final value of 0.01) and dropout fraction of 0.5 for both hidden layers.

3.2. Convolutional Neural Network

After using fully-connected neural networks, we decided to try Convolutional Neural Network (CNN), other type of neural network, which uses some extra concepts like local filters, max-pooling and weight sharing (LeCun & Bengio, 1995). Convolutional Neural Networks already demonstrated good performance in several speech- and music-related tasks (Dieleman et al., 2011) (Abdel-Hamid et al., 2012), so they seem to perform well with sound and can be useful in bioacoustic tasks too.

Main difference between CNN and fully-connected NN is that CNN is aware of 2D structure of the input data. It can be very helpful if there are some local correlations between spatially adjacent input values. In image recognition tasks CNN uses local receptive fields to extract local features like oriented edges and corners, and then combine them in higher layers to get more complex features. Since in our whale detection task we have 2D filter-bank input data, which contains local correlations between energy values both in time and frequency domain, we can use CNN in image-like manner.

For preventing overfitting and for using highly-optimized implementation of 2D-convolution (cuda-convnet[2], made by Alex Krizhevsky), we cropped out three overlapping square patches of size 72 x 72 from our 72 x 97 filter-bank input data. Due to the lack of time, memory and fast GPU, we rescaled 72 x 72 patches to the size of 36 x 36.

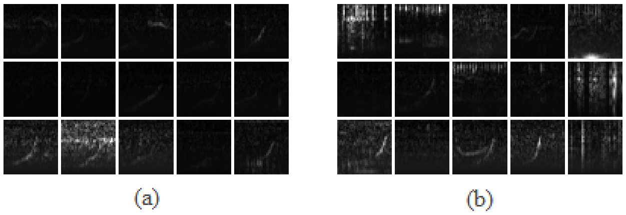

Figure 2. Examples of filter-bank-based patches (a) with right whale call, (b) without right whale call

We used recently proposed maxout units (Goodfellow et al., 2013) as hidden units. Given an input $x \in \mathbb{R}^d$, a maxout hidden layer implements the function

$$h_i(x) = \max_{j \in [1,k]} z_{ij}$$

where

$$z_{ij} = x^T W_{\dots ij} + b_{ij}$$

for learned parameters $W \in \mathbb{R}^{d \times m \times k}$ and $b \in \mathbb{R}^{m \times k}$. In the context of convolutional networks, a maxout feature map can be constructed by taking the maximum across k affine feature maps. A single maxout unit can be interpreted as making a piecewise linear approximation to an arbitrary convex function. So, training algorithm learns not just the relationship between hidden units, but also the activation function of each hidden unit.

Due to the lack of time, we didn’t perform proper hyper-parameter search, and just used the same parameters, as in (Goodfellow et al., 2013) for MNIST and CIFAR-10 datasets. Our first CNN architecture consisted of 36 x 36 input layer, three convolutional layers, followed by max-pooling layers, and final 2-way softmax layer. First and second convolutional layers had 48 kernels of size 8 x 8, followed by max-pooling with pool size of 4 x 4. Third layer had 24 kernels

[2] https://code.google.com/p/cuda-convnet/

of size 5 x 5 and followed by max-pooling with pool size of 2 x 2. Learning rate at the start was 0.05, and then decreased by dividing by 1.00004 after each epoch. Dropout was used on the first convolutional layer, with dropout rate of 0.8. At the testing time, when all of the 36 x 36 patches were already classified, we averaged the results for each three patches, cropped from single testing 2-second sample.

Our second CNN architecture consisted of three convolutional layers, followed by max-pooling layers, one fully-connected layer with maxout hidden units and final 2-way softmax layer. First convolutional layer had 48 kernels of size 8 x 8 followed by max-pooling with pool size of 4 x 4. Second convolutional layer had 128 kernels of size 8 x 8 followed by max-pooling with pool size of 4 x 4. Third layer had 128 kernels of size 5 x 5 followed by max-pooling with pool size of 2 x 2. Fourth layer was fully-connected and had 240 maxout hidden units. Learning rate at the start was 0.1, and then decreased by dividing by 1.00004 after each epoch. Dropout was used on the first convolutional layer, with dropout rate of 0.8.

4. Results

Table 1. Test set AUC performance of different whale call detection methods

Method	AUC
NN with sigmoid units (this paper)	0.954
First CNN with maxout units (this paper)	0.971
Second CNN with maxout units (this paper)	0.976
Gradient Boosting Classifier	0.984
Cornell University Benchmark	0.721

Our best fully-connected neural network got Area under the ROC curve (AUC) performance of 0.954, our best CNN with maxout units got AUC performance of 0.976. Cornell University algorithm before the challenge got AUC performance of 0.721. Winner team of the Marinexplore and Cornell University Whale Detection Challenge used two averaged gradient boosting classifiers with complex feature engineering, and got AUC performance of 0.984.

5. Discussion

Our results show that fully-connected and convolutional neural networks are capable of achieving good performance in whale call detection task. We also used new type of hidden units - maxout units - and show that they can perform well in audio processing tasks. Our results can be easily improved with more careful parameter tuning, using better GPU and training for longer time. Also it must be useful to pre-train neural network on unlabeled data with some unsupervised feature learning model like CDBN (Lee et al., 2009).

References

Abdel-Hamid, O., Mohamed, A., Jiang, Hui, and Penn, G. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4277–4280, 2012.

Dieleman, Sander, Brakel, Philé mon, and Schrauwen, Benjamin. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th international society for music information retrieval conference : Proc. ISMIR 2011*, pp. 669–674. University of Miami, 2011.

Dugan, Peter J., Rice, Aaron N., Urazghildiiev, Ildar R., and Clark, Christopher W. North Atlantic right whale acoustic signal processing: Part II. improved decision architecture for auto-detection using multi-classifier combination methodology. In *Systems, Applications and Technology Conference IEEE Long Island*, 2010a.

Dugan, P.J., Rice, A.N., Urazghildiiev, I.R., and Clark, C.W. North atlantic right whale acoustic signal processing: Part i. comparison of machine learning recognition algorithms. In *Applications and Technology Conference (LISAT), 2010 Long Island Systems*, pp. 1–6, 2010b.

Goodfellow, Ian J, Warde-Farley, David, Mirza, Mehdi, Courville, Aaron, and Bengio, Yoshua. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

Hinton, Geoffrey E., Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

Kraus, Scott D., Brown, Moira W., Caswell, Hal, Clark, Christopher W., Fujiwara, Masami, Hamilton, Philip K., Kenney, Robert D., Knowlton, Amy R., Landry, Scott, Mayo, Charles A., McLellan, William A., Moore, Michael J., Nowacek, Douglas P., Pabst, D. Ann, Read, Andrew J., and Rolland, Rosalind M. North atlantic right whales in crisis. *Science*, 309(5734):561–562, 2005.

LeCun, Yann and Bengio, Yoshua. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, Honglak, Pham, Peter T., Largman, Yan, and Ng, Andrew Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS*, pp. 1096–1104, 2009.

Mellinger, David K. and Clark, Christopher W. Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, 107(6):3518–3529, 2000.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Spaulding, Eric, Robbins, Matt, Calupca, Tom, Clark, Christopher, Tremblay, Tremblay, Waack, Amanda, Warde, Ann, Kemp, John, and Newhall, Kristopher. An autonomous, near-real-time buoy system for automatic detection of North Atlantic right whale calls. *The Journal of the Acoustical Society of America*, 125(4):2615, 2009.

Urazghildiiev, Ildar R. and Clark, Christopher W. Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test. *The Journal of the Acoustical Society of America*, 120(4):1956–1963, 2006.

Urazghildiiev, Ildar R. and Clark, Christopher W. Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics. *The Journal of the Acoustical Society of America*, 122(2):769–776, 2007.

Urazghildiiev, I.R., Clark, C.W., Krein, T.P., and Parks, S.E. Detection and recognition of north atlantic right whale contact calls in the presence of ambient noise. *Oceanic Engineering, IEEE Journal of*, 34(3):358–368, 2009. ISSN 0364-9059.