Domestic Cat Sound Classification Using Transfer Learning

Yagya Raj Pandeya and Joonwhoan Lee

Department of Computer Science and Engineering, Chonbuk National University, Jeonju, Korea

Abstract

The domestic cat or house cats (*Felis catus*) are an ancient human pet animal that can deliver various alert message to human on environmental changes by its mysterious kinds of sounds generation capability. Cat sound classification using deep neural network had scarcity of labeled data, that impelled us to make CatSound dataset across 10 categories of sound. The dataset was even not enough to select data driven approach for end to end learning, so we choose transfer learning for feature extraction. Extracted feature are input to six various classifiers and ensemble techniques applied with predicted probabilities of all classifier results. The ensemble and data augmentation perform better in this research. Finally, various results are evaluated using confusion matrix and receiver operating characteristic curve.

Keywords: Labeled dataset, Transfer learning, Ensemble method, Data augmentation

1. Introduction

Any pet animal's sound can be very helpful for human beings in perspective to security or pre-prediction of natural disasters. The pet animal sound classification and recognition using deep learning techniques can be a growing area of research. The data driven approach for acoustic signal got high interest in recent year and some study also focus on animal sound classification [1, 2] and animal population identification based on their sound characteristics [3]. The marine mammal sound classification and its impact on marine life is studied on [4, 5]. Several study [6–12] are focused on bird sound identification, classification and challenges. The study in [13] perform the insect species classification based on their sound signals using machine learning techniques. The prediction of unusual animal sound behavior during earthquake and natural disaster is studied using machine learning techniques in [14]. Recently, the possibilities of transfer learning for bird sound identification using fundamental characteristics of music is explained in [15]. As far to our knowledge, there is not any detail study for pet animals sound, especially for domestic cat sound classification, even they are close friend of human beings.

In this paper our work particularly focuses on automatically classify unseen domestic cat sounds using transfer learning. This research was started with building a robust cat sound dataset that can able to handle the bio-diversity, species variation and age variation issues of domestic cat. In order to increase the generalization and robustness of deep neural network, we perform audio data augmentation as in [16]. The next step was feature extraction using pretrained neural network and successfully classify these features using popular machine

Received: May 29, 2018
 Revised : Jun. 16, 2018
 Accepted: Jun. 21, 2018

Correspondence to: Joonwhoan Lee
 (chlee@jbnu.ac.kr)
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

learning techniques. Finally, we make comparative analysis our results from various classifier and predict the miss-classified cat sounds. The rest of paper is organized as follows: Section 2 describes the domestic cat sound dataset preparation and its challenges. Section 3 is an overview of transfer learning and classification of extracted features. The results and discussion are included in Section 5, followed by the evaluation in Section 6. Finally, the conclusion and future works are mentioned in Section 6.

2. CatSound Dataset

Hearing is the second most important human sense after vision to recognize any animal. Automatic unseen cat sound recognition using deep neural network need large number of annotated data for successful training. It was one of great challenge for us to collect domestic cat sounds and separate the meaningful audio segment so that our network architectures could learn the proper semantics of audio. Finally, we were able to collect cat video and audio data files mostly from online source like YouTube (<https://www.youtube.com/>) and Flickr (<https://www.flickr.com/>). The categorization of these sounds was another big challenge because some sounds are very similar to other and cat can produce different sounds in very small-time difference. For example, a cat in angry mood may generates “growling,” “hissing,” or “nyaaan” sounds simultaneously. Semantic explanation of cat sounds in [17, 18] helps us to categories the various domestic cat sounds in desired classes. To preserve the semantic meaning of sounds, the segmented sound files may have varying length, which is like the same case described in [19] for music information retrieval problem. For example, the sound of cat in normal mood (“meow-meow”), defense in angry mood (“hissing”), kittens calling its mother (“pilling”), cat in pain (“miyoou”) may have shorter length sound but the cat in rest (“purring”), warning in angry mood (“growling”), mating (“gay-gay-gay”), fighting (“nyaaan”), angry (“momo-mooh”) and want to hunt (“trilling or chatting”) usually are more meaningful if they are analyzed in longer time duration. Even in same class of cat sound have varying length of data file because the bio-diversity widely changes across geographical location, cat species, age and mood of cat.

The pie chart in Figure 1 illustrates cat sound dataset where each class are named according to the sound of cat in various moods. The number of cat sound files in each class are nearly 300, that is 10% of total data files, so we termed this CatSound dataset as a balanced dataset. CatSound dataset contain more

Figure 1. CatSound dataset class representation.

than three hours of domestic cat sounds files with across 10 sound classes.

3. Method

This section covers feature extraction of CatSound dataset using transfer learning and classification of these features using various classifiers in Python library (<http://scikit-learn.org>).

3.1 Preparation and Audio Augmentation

Mel-spectrograms are extracted from CatSound data in real-time on the GPU using Kapre [20]. The input has a single channel, 96-mel bins, and 1,360 temporal frames as described in [21]. To increase the data size and more generalize the cat sound feature, we use audio data augmentation as [16] by random selection of time stretching, pitch shifting, dynamic range compression, and insertion of noise. We augment our original dataset using LibROSA [22] making one to three augmented clones or copies of each original data files, named as *1x_Aug*, *2x_Aug*, and *3x_Aug* dataset, respectively. These four-set data (one original plus three augmented) used in our study for comparatively observe the classifier results.

3.2 Transfer Learning

We use pre-trained five-layer convolutional neural network (CNN) of [21] as a feature extractor for CatSound dataset. This network was trained with Million Song dataset [23] and secure very good accuracy for music classification and regression. The cat sounds are very different from the studio recorded music because of the frequency variation, signal-to-noise ratio, sudden

Figure 2. Overview of feature extraction from CNN network. From each layer of CNN, the globally averaged 32-dimensional features are concatenated into one feature vector and fed into various classifier. The predicted probability of each classifier is ensembled for final prediction result.

change in pitch and magnitude, frequent interruption of environmental sounds, bio-diversity of species, aging and animal mood. Even with these issues, the lack of enough labeled data made us to use pre-trained network as source network for transfer learning.

The experiment result shows that even the cat sound and music characteristics are very different, but every audio signal have some similar feature that are very helpful for small dataset if transfer learning is performed properly. The feature extraction and classification using pre-trained CNN architecture can be shown in Figure 2. The source network had total 32 feature in each layer and the global average pooling for each layer map them into 32-dimensional feature vector. Finally, we concatenate each layer feature and got 160-dimensional feature vector. These feature vectors are the input for our classifiers and voting the predicted probability for final ensembled result.

3.3 Cat Sound Classifiers

The comparative study of various classifier is mandatory to analyses diversified feature of cat sound. We select six popular classifiers to classify our dataset features that are briefly described. Random forest (RF) [24] classifier uses majority voting scheme (Bagging [25] or Boosting [26]) to make final prediction from a prespecified number of decision trees. K-nearest neighbor

(KNN) [27] classifier find the class to which an unknown object belongs, by using majority voting of KNNs. Extremely randomized trees (or Extra Tree) [28] classifier tries to find optimal cut-point on entire given feature randomly. The linear discriminant analysis (LDA) [29] and quadratic discriminant analysis (QDA) [30] classifier are types of Bayesian classifiers that tries to maximize the component axis for class-separation. Support vector machine (SVM) [31] classifier with a radial basis function (RBF) is used to classify our extracted features non-linearly. The 10-fold cross-validation was performed during training the SVM. The cross-validation theory explains better performance on increasing fold until certain limit, but it also increases the computational cost. The research [32] suggest us to use 10-fold cross-validation for better performance of model's hyperparameters in supervised learning. The comparative results of these classifiers are mentioned on Section 5.

3.4 Ensemble Classifiers

Ensemble is a well-known machine learning technique that combines the prediction power of different classifiers and make the overall system more robust, as described in [33]. It can be done by majority voting or averaging the predicted probabilities of all classifiers of Section 3.3. We select averaging methods for ensemble our six classifiers described in Section 3.3. The ensemble technique is beneficial on all individual classifiers in all datasets (original plus augmented) and evaluation metrics described in Section 4. Table 1 and Figure 3 show the boost of system performance using ensemble on classifier perfected probability.

4. Evaluation

In evaluation the original and augmented CatSound dataset, the CNN features are divided into 90% of data for training our classifiers and remaining 10% for test these classifier predications. We find the influence of augmentation on our dataset boost classifier result and reduce the rate of confusion. To evaluate the performance of classifiers described in Section 3.3, accuracy, F-score (or F1-score) [34] and area under the receiver operating characteristic curve (ROC-AUC) scores [35] FVF are used. The accuracy metric refers to the percentage of correctly classified unknown data samples and F-score compute harmonic mean between precision and recall. The ROC-AUC scores are the prediction values of each curve in ROC curve [34] discussed in Section 5.

Table 1. Best performance of classifier in *3x_Aug* dataset using three metrics: accuracy, F1-score and AUC score

Classifiers	Accuracy (%)	F1-Score	AUC score
RF	78.99	0.79	0.978
KNN	79.07	0.79	0.884
Extra Trees	77.30	0.77	0.977
LDA	73.67	0.74	0.967
QDA	80.76	0.81	0.974
SVM	78.57	0.78	0.978
Ensemble	87.76	0.88	0.990
RF	78.99	0.79	0.978

Receiver Operating Characteristic (ROC) Curve of Domestic Cat **Figure 3.** ROC curve of the best performing classifiers with *3x_Aug* datasets.

5. Results and Discussion

The best performing result using *3x_Aug* dataset illustrates in Table 1. The experiment result shows that the ensemble classifier performs best in CNN feature with *3x_Aug* dataset and success to achieve more than 87% of accuracy.

The evaluation of classifiers results with and without augmentation scheme is presented in Figure 4. The illustration shows the boost of performance of classifiers after augmentation in different scale using on various evaluation metrics.

The confusion matrix is tabular representation of test result that enable us to further understand the strength and weakness of the network model and classifier. Figure 5 shows confusion matrix of our best classifier, QDA, in one times augmented dataset. Analysis of each confusion matrix represented by our six classi-

Figure 4. The average of accuracy, F1-score and area under curve comparison of our classifiers (six classifier and one ensemble) with original and augmented dataset. **Figure 5.** Confusion matrix of the best performing ensemble classifier with *3x_Aug* dataset.

fiers, we reach in some conclusions. The Resting (“Purring”) and Defense (“Hissing”) sounds are unique to other cat sound in CatSound dataset so these classes are relatively less confusing. In the other hand, the classes MotherCall (“pilling”) and HuntingMind (“trilling or chatting”), Happy (“meow-meow”) and Paining (“miyoou”), and Angry (“momo-mooh”) and Mating (“gay-gay-gay”) have some similarity in sound so we find these classes are very confusing for our all classifiers.

ROC curve is a graphical plot to illustrates the diagnostic ability of classifiers using two parameters (true positive rate [TPR] and false positive rate [FPR]) at various decision thresh-

old. Each threshold value produces a different point in ROC space and the classifier produces positive result if its output is above that threshold. ROC curves are typically for binary classification so we binarize the output for our 10-class classification problem. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The ROC curve in Figure 3 showing six classifier curves and one averaging curve of all classifiers. The ROC-AUC scores of each curve are also illustrated in the same ROC curve.

6. Conclusion and Future Work

Automatic bio-diversity estimation using animal sounds will be future area of research, but the data-driven approach need large number of labeled data. One way to full fill that scarcity can be the use of transfer learning. Even the source network trained with music data, the transfer learning will still beneficial for small animal sound dataset. The data augmentation, and majority voting (bagging or boosting) method can boost the system performance even using small sized animal sound dataset.

In future our research can be extended for better results by increasing the labeled dataset or selection of pre-trained network trained on animal sounds for transfer learning because the music data are very different from animal sounds.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

The research leading to these result, authors would like to thank Korean Ministry of Education for funding. The National Research Foundation of Korea (NRF) support us under Basic Science Research Program (NRF-2015R1D1A1A01058062).

References

- [1] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder, "Discrimination and retrieval of animal sounds," in *Proceedings of the 12th IEEE International Multi-Media Modelling Conference*, Beijing, China, 2006. <https://doi.org/10.1109/MMMC.2006.1651344>
- [2] S. Gunasekaran and K. Revathy, "Content-based classification and retrieval of wild animal sounds using feature selection algorithm," in *Proceedings of the 2nd International Conference on Machine Learning and Computing*, Bangalore, India, 2010, pp. 272-275. <https://doi.org/10.1109/ICMLC.2010.11>
- [3] N. Raju, S. Mathini, T.L. Priya, P. Preethi, and M. Chandrasekar, "Identifying the population of animals through pitch, formant, short time energy: a sound analysis," in *Proceedings of the International Conference on Computing, Electronics and Electrical Technologies*, Kumaracoil, India, 2012, pp. 704-709.
- [4] S. Zaugg, M. Schaar, L. Houegnigan, C. Gervaise and M. Andre Real-time acoustic classification of sperm whale clicks and shipping impulses from deep-sea observatories," *Applied Acoustics*, vol. 71, no. 11, pp. 1011-1019, 2010. <https://doi.org/10.1016/j.apacoust.2010.05.005>
- [5] F. R. Gonzalez-Hernandez, L. P. Sanchez-Fernandez, S. Suarez-Guerra, and L. A. Sanchez-Perez, "Marine mammal sound classification based on a parallel recognition model and octave analysis," *Applied Acoustics*, vol. 119, pp. 17-28, 2017. <https://doi.org/10.1016/j.apacoust.2016.11.016>
- [6] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K. H. Tauchert, and K. H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustics monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524-1534, 2010. <https://doi.org/10.1016/j.patrec.2009.09.014>
- [7] X. Zhang and Y. Li, "Adaptive energy detection for bird sound detection in complex environments," *Neurocomputing*, vol. 155, pp. 108-116, 2015. <https://doi.org/10.1016/j.neucom.2014.12.042>
- [8] S. Rassak, M. Nachamai, and A. K. Murthy, "Survey study on the methods of bird vocalization classification," in *Proceedings of IEEE International Conference on Current Trends in Advanced Computing*, Bangalore, India, 2016, pp. 1-8. <https://doi.org/10.1109/ICCTAC.2016.7567337>
- [9] D. Stowell, E. Benetos, and L. F. Gill, "On-bird sound recordings: automatic acoustic recognition of activities and contexts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1193-1206, 2017. <https://doi.org/10.1109/TASLP.2017.2690565>

- [10] Z. Zhao, S. Zhang, Z. Xu, K. Bellisario, N. Dai, H. Omrani, and B. C. Pijanowski, "Automated bird acoustic event detection and robust species classification," *Ecological Informatics*, vol. 39, pp. 99-108, 2017. <https://doi.org/10.1016/j.ecoinf.2017.04.003>
- [11] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, 2017, pp. 141-145. <https://doi.org/10.1109/ICASSP.2017.7952134>
- [12] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin Bird detection in audio: a survey and a challenge," in *Proceedings of IEEE 26th International Workshop on Machine Learning for Signal Processing*, Vietri sul Mare, Italy, 2016, pp. 1-6. <https://doi.org/10.1109/MLSP.2016.7738875>
- [13] J. J. Noda, C. M. Travieso, D. Sanchez-Rodriguez, M. K. Dutta, and A. Singh, "Using bioacoustic signals and support vector machine for automatic classification of insects," in *Proceedings of IEEE 3rd International Conference on Signal Processing and Integrated Networks*, Noida, India, 2016, pp. 656-659. <https://doi.org/10.1109/SPIN.2016.7566778>
- [14] W. Astuti, A. M. Aibinu, M. J. E Salami, R. Akmelawati, and A. G. A Muthalif, "Animal sound activity detection using multi-class support vector machines," in *Proceedings of IEEE 4th International Conference on Mechatronics*, Kuala Lumpur, Malaysia, 2011, pp. 1-5. <https://doi.org/10.1109/ICOM.2011.5937122>
- [15] S. Ntalampiras, "Bird species identification via transfer learning from music genres," *Ecological Informatics*, vol. 44, pp. 76-81, 2018. <https://doi.org/10.1016/j.ecoinf.2018.01.006>
- [16] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *The IEEE Signal Processing Letters*, 2016. <https://doi.org/10.1109/LSP.2017.2657381>
- [17] Wikipedia, "Cat communication," Available <https://en.wikipedia.org/wiki/Catcommunication>
- [18] L. Moss, "Cat sounds and what they mean," 2013, Available <https://www.mnn.com/family/pets/stories/cat-sounds-and-what-they-mean>
- [19] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," 2018, Available <https://arxiv.org/abs/1709.04396>
- [20] K. Choi, D. Joo, and J. Kim, "Kapre: on-GPU audio preprocessing layers for a quick implementation of deep neural network models with keras," 2017, Available <https://arxiv.org/abs/1706.05781>
- [21] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017, pp. 141-149.
- [22] "LibROSA v0.6.1," 2018, Available <https://librosa.github.io/librosa/>
- [23] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval*, Miami, FL, 2011, pp. 591-596.
- [24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. <https://doi.org/10.1023/A:1010933404324>
- [25] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996. <https://doi.org/10.1007/BF00058655>
- [26] L. Breiman, "Bias, variance, and arcing classifiers," University of California, Berkeley, CA, 1996.
- [27] J. Laaksonen and E. Oja Classification with learning k-nearest neighbors," in *Proceedings of IEEE International Conference on Neural Networks*, Washington, DC, 1996, pp. 1480-1483. <https://doi.org/10.1109/ICNN.1996.549118>
- [28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006. <https://doi.org/10.1007/s10994-006-6226-1>
- [29] Wikipedia, "Linear discriminant analysis," Available <https://en.wikipedia.org/wiki/Lineardiscriminantanalysis>
- [30] S. Srivastava, M. R. Gupta, and B. A. Frigiyik, "Bayesian quadratic discriminant analysis," *Machine Learning Research*, vol. 8, pp. 1277-1305, 2007.

- [31] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. <https://doi.org/10.1007/BF00994018>
- [32] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp. 1137-1143.
- [33] H. Bahuleyan, "Music genre classification using machine learning techniques," 2018, Available <https://arxiv.org/abs/1804.01149>
- [34] Y. Sasaki, "The truth of the F-measure," University of Manchester, Manchester, UK, 2007.
- [35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. <https://doi.org/10.1016/j.patrec.2005.10.010>

Yagya Raj Pandeya was born in Dadeldhura, Nepal in 1988. He receives the B.E. and M.E. degree in Computer Engineering from the Pokhara University of Nepal, in 2010 and 2013, respectively. He was Head of Depart-

ment of Computer Engineering in NAST College in Dhangadhi, Nepal. He join Ministry of Home Affairs Nepal in 2015 to 2017. Mr. Yagya is currently a Ph.D. fellow at Fuzzy Logic and Artificial Intelligence Laboratory in Chonbuk National University, Korea.

E-mail: yagyapandeya@gmail.com

Joonwhoan Lee received his BS degree in Electronic Engineering from the University of Hanyang, Korea in 1980. He received his MS degree in Electrical and Electronics Engineering from KAIST, Korea in 1982, and the Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA in 1990. He is currently a Professor in Department of Computer Engineering, Chonbuk National University, Korea. His research interests include image and audio processing, computer vision, emotion engineering etc.

E-mail: chlee@jbnu.ac.kr