AN AUGMENTED NEURAL NETWORK FOR THE DCASE2019 URBAN SOUND TAGGING CHALLENGE

Technical Report

Ferran Orga[1], *Joan Serrà*[2], *Carlos Segura Perales*[2], *Francesc Alías*[1], *Rosa Ma. Alsina-Pagès*[1]

[1] GTM - Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull (URL),
C/Quatre Camins, 30, 08022 Barcelona (Spain)
{ferran.orga, francesc.alias, rosamaria.alsina}@salle.url.edu
[2] Telefónica Research, Barcelona (Spain)
{joan.serra, carlos.seguraperales}@telefonica.com

ABSTRACT

The Sounds of New York City (SONYC) research project aims to mitigate urban noise pollution in the context of a megacity. This project has deployed 50 different sensors in various areas of the New York City installed back in 2015 to monitor the overall sound pressure level. However, this is not enough to determine the noise sources, needed to detect noise code violations. Within the Task 5 of DCASE2019 challenge, an urban sound tagging challenge is proposed where the participants are asked to develop a machine listening system that distinguishes between 23 sources of noise pollution. The system is asked to predict whether the source is present or absent in 10-second scenes recorded by the SONYC. Moreover, annotations are also provided at a higher level, classifying the 23 fine labels in 8 coarser labels. In this report, the authors present a machine listening approach based on an augmented neural network where both coarse and fine-level annotations are used to predict the event presence in the same network. This approach obtains a classification accuracy on the validation split of 87% at the coarse level and 92% at the fine level.

Index Terms— urban sound tagging, DCASE, soundscape, classification, machine listening

1. INTRODUCTION

The fifth edition of the IEEE AASP Detection and Classification of Acoustic Scenes and Events (DCASE2019)[1] takes place in 30 March - 31 July 2019. In this edition, five tasks have been proposed: *i)* acoustic scene classification, *ii)* audio tagging with noisy labels and minimal supervision, *iii)* sound event localization and detection, *iv)* sound event detection in domestic environments, and *v)* urban sound tagging; being the latter the one addressed in this report.

The Urban Sound Tagging (UST) task asks the participants to predict whether each of the 23 predefined noise sources are present or absent in a 10-second scene recorded by the Sounds of New York City (SONYC) acoustic sensor network [1]. Another annotation is also provided, grouping the 23 noise sources in 8 higher-level labels. The participants are asked to predict the presence probability of coarse and fine-labelled events using a multi-label approach in two separate rankings (”coarse-grained” and ”fine-grained” predictions).

In this paper, the authors present a deep Convolutional Neural Network (CNN) using data augmentation techniques meant to predict the presence probability for each noise source in the given 10-s audio snippets. The presented system in this paper used a basic network with several data augmentation techniques using both coarse and fine-level annotations to predict the presence probability at both levels in the same network.

In Section 2, a description of the task and the given dataset are described. In Section 3, our approach to address the problem is described, detailing both the used model and the applied data augmentation techniques. Finally, in Section 4, the prediction results of the system are explained.

2. TASK DESCRIPTION

Machine listening systems are being used in other in other contexts to monitor noise using sensor networks with the aim of mitigating noise pollution [2, 3, 4]. In this task the SONYC [5] team provides a realistic use case for the development and evaluation of innovative machine listening systems.

In the challenge, a total of 23 noise ”tags”, each of them referring to a noise source, being many of them the cause of noise complaints in New York City. The primary goal of the task is to determine the probability of presence of each one of the 23 sources of noise. However, this is an ambitious goal, as these tags differentiate between very similar sounds, *e.g.* motorcycle, car and truck engines, these are called ”fine-grained” annotations. A secondary goal is defined in the UST task, which consists in predicting between 8 coarser tags, *e.g.* all sounds originated by an engine would be labelled in the ”engine” category, these are called ”coarse-grained” annotations. Also, 6 other labels tagged as ”other/unknown” are added in the fine labels belonging to 6 different coarser categories. In Figure 1 the labelling taxonomy is detailed at both levels.

Two datasets are provided for the task, one for development, including a training split and a validation split, and another for evaluation, which is required to rank the submitted systems. The systems proposed by the participants will be evaluated in two different rankings, one for the coarse labels and another for the fine labels. The classification metric used to evaluate and rank the participants is the Area Under the Precision-Recall Curve (AUPRC) [6]. To compute this curve, a threshold of confidence for every tag in every snippet

[1] http://dcase.community/

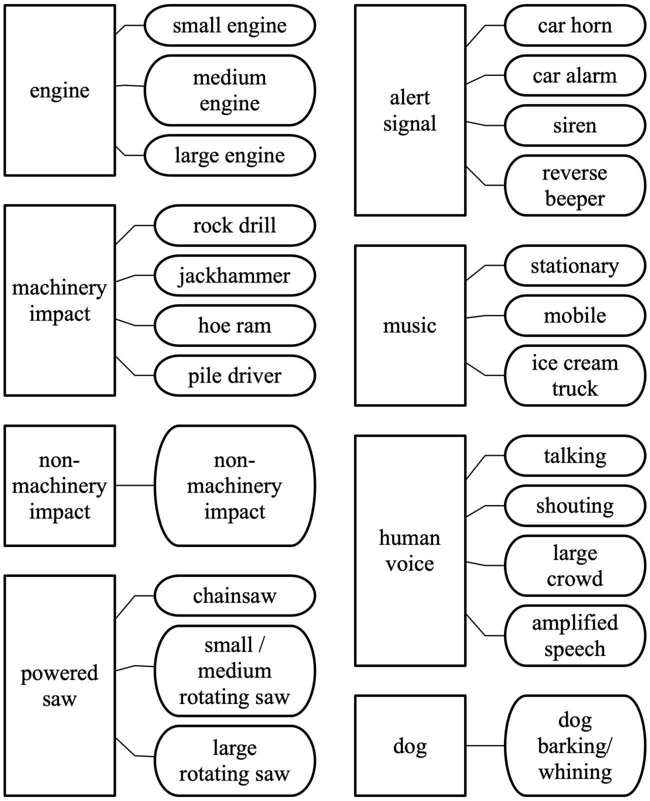

Figure 1: Hierarchical taxonomy of urban sound tags in the DCASE2019 Urban Sound Tagging task. Rectangular and round boxes respectively denote coarse and fine tags[3].

is fixed, resulting in a one-hot encoding of predicted tags. Then, the total number of True Positives (TP), False Positives (FP), and False Negatives (FN) is computed between prediction and consensus ground truth over the entire evaluation dataset. The predictions of the system should be submitted in a Comma-Separated Values (CSV) file, where the filename of the audio in the evaluation dataset indexed with the coarse-level and fine-level predictions.

2.1. Dataset

The organizers provide two datasets, one for development and the other for evaluation. Both the datasets contain recording excerpts of 10 seconds, the development ones include annotations and the evaluation ones do not only include the raw audio. The development dataset contains a *train* split of 2351 recordings and a *validate* split of 443 recordings, making a total of 23510 seconds for the training and 4430 seconds for the validation split. The annotations in the development dataset include:

- **Split**: Train or validate
- **Sensor ID**: The ID of the sensor the recording is from. These have been anonymized to have no relation to geolocation.
- **Audio Filename**: The filename of the audio recording
- **Annotator ID**: The anonymous ID of the annotator.
- **Fine-level Presence**: Columns of this form indicate the presence of fine-level class: 1 if present, 0 if not present.

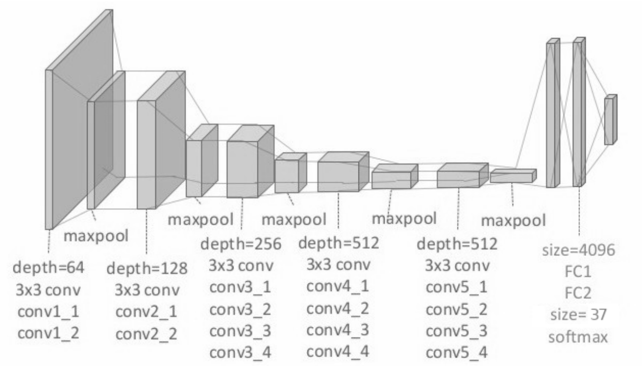

Figure 2: The VGG-19 model has a total of 19 layers, including convolutions and fully-connected.

- **Coarse-level Presence**: Columns of this form indicate the presence of a coarse-level class: 1 if present, 0 if not present.
- **Fine-level Proximity**: Columns of this form indicate the proximity of a fine-level class according to the annotators. If present, it has three levels: *near*, *far*, *notsure*.

The evaluation dataset includes 204 audios of 10 s each without annotations. The participants in the UST are asked to predict the probability of each coarse and/or fine-level class and to save the filenames with the probabilities in a CSV file.

3. CNN-BASED APPROACH AND MODEL ARCHITECTURE

The authors propose a deep convolutional neural network using data augmentation techniques (de-emphasis, compression and mix-up) based on the Visual Geometry Group (VGG) architecture [7].

Firstly, the 10-s audios are transformed on-the-fly using a de-emphasis[4] and a compression filter in random factors to generate different similar audios at every training epoch. After that, the audio is mixed with another random 10-s audio of the training split using a random value to mix up both audios in different proportions. The hot-encoding labels are also added up using the same factor and a threshold at a hearing level is applied to get the binary one-hot encoding of the audio again.

After that, the Mel spectrogram [8] of the audio is computed with 128 filters for the whole 10-s audio snippet. This is the input of the VGG-based model, a deep CNN that uses small 3x3 convolutional filters.

Both coarse and fine-grained annotations are used as input classes for the mode. In order to predict both labellings, it computes one loss function for each level, in order to train both coarse and fine-levels equally. The model, based on the VGG-19 architecture [7] is depicted in Figure 2 where the layers of the model are detailed in order, number and size. Also, the model output size is 37, belonging to 8 coarse labels and 29 fine labels (23 defined and 6 other/unknown). The loss function of the model is a Sigmoid combined with the Binary Cross-Entropy to obtain more numerical stability [5]. The model is trained with an Stochastic Gradient De-

[3] http://dcase.community/challenge2019/task-urban-sound-tagging

[4] http://www.fon.hum.uva.nl/praat/manual/Sound__Filter__de-emphasis____.html

[5] https://pytorch.org/docs/stable/nn.html#bcewithlogitsloss

Layer (type)	Output Shape	Param. Number
InstanceNorm2d-1	[-1, 1, 128, 1001]	0
MelSpectrogram-2	[-1, 1, 128, 1001]	0
InstanceNorm2d-3	[-1, 1, 128, 1001]	0
Conv2d-4	[-1, 64, 128, 1001]	640
BatchNorm2d-5	[-1, 64, 128, 1001]	128
ReLU-6	[-1, 64, 128, 1001]	0
MaxPool2d-7	[-1, 64, 64, 500]	0
Conv2d-8	[-1, 128, 64, 500]	73,856
BatchNorm2d-9	[-1, 128, 64, 500]	256
ReLU-10	[-1, 128, 64, 500]	0
MaxPool2d-11	[-1, 128, 32, 250]	0
Conv2d-12	[-1, 256, 32, 250]	295,168
BatchNorm2d-13	[-1, 256, 32, 250]	512
ReLU-14	[-1, 256, 32, 250]	0
Conv2d-15	[-1, 256, 32, 250]	590,08
BatchNorm2d-16	[-1, 256, 32, 250]	512
ReLU-17	[-1, 256, 32, 250]	0
MaxPool2d-18	[-1, 256, 16, 125]	0
Conv2d-19	[-1, 512, 16, 125]	1,180,160
BatchNorm2d-20	[-1, 512, 16, 125]	1,024
ReLU-21	[-1, 512, 16, 125]	0
Conv2d-22	[-1, 512, 16, 125]	2,359,808
BatchNorm2d-23	[-1, 512, 16, 125]	1,024
ReLU-24	[-1, 512, 16, 125]	0
MaxPool2d-25	[-1, 512, 8, 62]	0
Conv2d-26	[-1, 512, 8, 62]	2,359,808
BatchNorm2d-27	[-1, 512, 8, 62]	1,024
ReLU-28	[-1, 512, 8, 62]	0
Conv2d-29	[-1, 512, 8, 62]	2,359,808
BatchNorm2d-30	[-1, 512, 8, 62]	1,024
ReLU-31	[-1, 512, 8, 62]	0
MaxPool2d-32	[-1, 512, 4, 31]	0
Linear-33	[-1, 4096]	260,050,944
ReLU-34	[-1, 4096]	0
Dropout-35	[-1, 4096]	0
Linear-36	[-1, 4096]	16,781,312
ReLU-37	[-1, 4096]	0
Dropout-38	[-1, 4096]	0
Linear-39	[-1, 37]	151,589
VGG-40	[-1, 37]	0

Table 1: VGG-based model architecture

scent (SGD) optimizer [9] that drops the learning rate periodically given a patience period where the loss is not reduced.

Finally, the evaluation dataset is given to the model and a sigmoid function is applied to obtain the probability of the predictions. These are saved into a CSV file to collect the results for the challenge rankings.

Concerning the implemented approach, a summary of our model is given in Table 1, where the parameters used in each layer are detailed, following the same structure as Figure 2. 286,208,677 trainable parameters are used by the model occupying a total of 1091.90 MB, and a total estimated GPU memory of 1500 MB is used to train the model. The source of the project is available online [6].

4. RESULTS

The results of the model have been evaluated on the *validate* split of the development dataset for the ”coarse-grained” and the ”fine-grained” challenge with the following values:

- AUPRC Coarse: 78.3%
- AUPRC Fine: 58.2%
- Accuracy Coarse: 86.7%
- Accuracy Fine: 92.0%

5. ACKNOWLEDGMENT

Ferran Orga thanks the support of Telefónica Research for the given facilities and the European Social Fund and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement of the Catalan Government for the pre-doctoral FI grant No. 2019_FI_B2_00168. Francesc Alías and Rosa Ma Alsina-Pagès would like to thank the Obra Social La Caixa for supporting this research through grants ref. 2018-URL-IR2nQ-029 and ref. 2018-URL-IR2nQ-038, respectively.

6. REFERENCES

[1] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, “The life of a new york city noise sensor network,” *Sensors*, vol. 19, no. 6, p. 1415, 2019.

[2] F. Mietlicki, C. Mietlicki, and M. Sineau, “An innovative approach for long-term environmental noise measurement: Rumeur network in the paris region,” in *Proceedings of EuroNoise 2015*. Maastrich, Netherlands: EAA-NAG-ABAV, 31 May - 3 June 2015, pp. 2309 – 2314.

[3] L. Nencini, P. De Rosa, E. Ascari, B. Vinci, and N. Alexeeva, “SENSEable Pisa: A wireless sensor network for real-time noise mapping,” in *Proceedings of the EuroNoise 2012*, Prague, Czech Republic, 10 - 12 June 2012, pp. 10–13.

[4] X. Sevillano, J. C. Socoró, F. Alías, P. Bellucci, L. Peruzzi, S. Radaelli, P. Coppi, L. Nencini, A. Cerniglia, A. Bisceglie, R. Benocci, and G. Zambon, “DYNAMAP Development of low cost sensors networks for real time noise mapping,” *Noise Mapping*, vol. 3, pp. 172–189, May 2016.

[5] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, “Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution,” *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, Feb 2019.

[6] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[8] P. Mermelstein, “Distance measures for speech recognition, psychological and instrumental,” *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

[9] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.

[6] https://bitbucket.org/ferranorga/dcase19