Sound Classification and Localization in Service Robots with Attention Mechanisms

Matteo Bodini
Department of Computer Science
Università degli Studi di Milano
Via Celoria 18, 20133 Milano (Italy)
matteo.bodini@unimi.it

Abstract—Human-machine interaction is calling for a sophisticated understanding of subjects' behavior performed by smartphones, home automation and entertainment devices, and many service robots. During an interaction with human beings in their environment, a service robot has to be capable to perceive and process visual and sound information of the scene that he observes. To capture salient elements in such different signals many semi-supervised deep learning methods have been proposed. In this article, it is proposed a new convolutional neural network, endowed with a mechanism of attention in order not only to classify, but also to localize temporally a sound event, and in a semi-supervised way.

Index Terms—service robots, convolutional neural networks, deep learning, semi-supervised learning, audio pattern recognition

I. INTRODUCTION

Human-machine interaction in the Internet era - where smartphones, automotive, home automation and entertainment are natural players within the heterogeneous digital ecosystem - is calling for sophisticated understanding of subjects' behavior. Interactions are likely to become more personalized, relevant and authentic. For instance, consumers will expect more than the ability to be authenticated for their mobile devices; they will count on existing applications (e.g., Facebook, Snapchat, Twitter) to make sense of the imagery and multimedia content for them. Social signal processing and affective computing are fast growing research fields that aim at bridging such cognitive-emotional gap [1], [6], [25].

In the realm of Human-machine interaction we are in touch also with many sophisticated service robots. Generally, the latter are system-based autonomous and adaptable interfaces that interact, communicate and deliver service to a human, where robots are widely seen as machines able of carrying out complex series of actions [20], [26]. They are capable of autonomous decision making based on the data they receive by various sensors and other sources (i.e. the sense-think-act paradigm) and adapt to the situation, thus they can learn from previous episodes. In order to interact with the human and his environment, a service robot must be able to perceive visual and sound information of the scene that he observes or in which he participates. In particular, it must be able to identify salient elements in the different signals captured: spatial location in an image, or temporal in an audio stream. Considering the plethora of the so-called deep learning methods, and the considerable cost of the annotation of the data, often researchers tend to make use of semi-supervised methods [7], able on the one hand to extract information in a way supervised, and secondly to predict the spatial or temporal organization of the events present in the processed signal. In the field of vision, this concept has been used repeatedly to perform object or activity spatial location on images [2]-[4], [17], [27] from raw 2D signals (pixels). At the audio level, the tendency to dispense with MFCC low level representations [8], [16] has appeared, allowing direct processing of the raw audio signal [12], [13], [22], [23] and leaving the neural networks the task of extracting the optimum representative characteristics of the processed signals. In this article, it is proposed a convolutional neural network, associated with a mechanism of attention, allowing the exploitation of the raw audio signal, in order not only to classify, but also to localize temporally a sound event present in the flow treated, and in a semi-supervised way.

Several studies have been proposed to learn neural networks from raw sounds. This work is inspired by the remarkable works of Wavenet [23] generating music, SampleCNN [13] classifying music or EnvNet [22] classifying environmental sounds and here it is proposed to integrate attention mechanisms generally used in vision. Take the example of the Global Average Pooling (GAP) [14] which permits to infer a spatial location in semi-supervised mode [27], indeed, an averaging operation coupled to a concept of Class Activation Mapping (CAM) allow the authors to locate an object or activity class in an image and also Gated Convolutional Layers [11], [24], used in audio context [23], which is considered here as a mechanism attention. To our knowledge, the semi-supervised localization of environmental sounds via convolutional networks, central to the works presented here, has been little explored.

II. THE USED AUDIO DATASET

The ESC [18] dataset (Dataset for Environmental Sound Classification) is a collection of short environmental recordings available in a unified format (5-second-long clips, 44.1 kHz, single channel, Ogg Vorbis compressed @ 192 kbit/s). All clips have been extracted from public field recordings available through the Freesound.org project. The dataset is available under the terms of the Creative Commons, and consists of three parts:
ESC-US: an unlabeled dataset of 250,000 environmental recordings (5-second-long clips), suitable for unsupervised pre-training.

ESC-50: a labeled set of 2000 environmental recordings (50 classes, 40 clips per class).

ESC-10: a labeled set of 400 environmental recordings (10 classes, 40 clips per class. This is a subset of ESC-50 - created initially as a proof-of-concept/standardized selection of easy recordings). The classes correspond to environmental sounds recorded indoor or outdoor of dog, rooster, waves, rain, embers, clock, helicopter, chainsaw, baby who cries, and sneezes. Each class consists of 40 records of 5 seconds.

As in the work of [18], the convolutional neural network (CNN) that will be presented is learned on the basis of cross-validation carried out on 5 folds, having previously applied, like in [22], a preprocessing to obtain sample 16 Khz and coded on 16 bits. To evaluate the ability of the CNN to temporally locate a sound event, it is randomly generated a test database consisting of 1600 extracts of 8 seconds each. Each recording consists of 8 seconds of randomly selected signal from a restaurant noise on which are added two beeps of 2.5 seconds each, randomly selected from the ESC10 test base and randomly positioned, avoiding overlapping.

III. PRESENTATION OF THE SYSTEM

The system, presented in Fig. 1, is based on the EnvNet network [22], to which are added specific elements to the semi-supervised localization: two systems of attention, as well as a naive mechanism of semi-supervised localization. The systems of attention are organized by layers as follows:

- Mechanism of attention on the lower layers of the network: the first attention mechanism underlying the network is to introduce a multiplicative constraint on a convolutive layer of EnvNet (Fig. 1). This constraint is implemented by multiplying a standard layer of the original network Conv4 (whose activations are activation functions of type ReLu) by a new layer Conv4 Att whose sigmoidal activations act as gates regulating the signal of the Relu. This method, first initiated on image data by [24], was applied in an audio context by [23].

- Mechanism of attention on the high layers of the network: the second attention mechanism introduced into our network adds a so-called Global Average Pooling layer. This technique, first proposed for spatial localization, is also suitable for temporal location of sound events in the audio signal. Eventually, the last convolutional layer of EnvNet, which is convolved in time, is connected to a new convolution layer containing as many neurons as classes (10 neurons for 10 classes with ESC-10) then goes through a layer of Global Average Pooling and softmax resulting in 10 predictions. The final model is noted Attentive-Net.

For time localization, it is designed a strategy based on a set of heuristics. In order to extract the localization of the sound events present in a sound recording of several seconds we:

- define an activation bitmask (MBA) based on a threshold value \( \tau = 150\% \) of the maximal activation of the GAP neurons (Fig. 2, the upper part) when the input audio signal corresponds to silence. \( MBA_i = 1 \) if \( GAP_i > \tau \), 0 otherwise.

- Filter the activation mask so that the activations are equal to 1 if they are close to at least two other positive activations in a centered window of 10 activations, and to 0 otherwise.

- Use the CAM method on each of the positive windows of the filtered activation mask to predict what the nature of the sound is.

IV. EXPERIMENTAL RESULTS

In this section it is described the experimental settings and the results obtained. AttentiveNet is trained with the same settings of EnvNet: cross-entropy criterion is used with momentum stochastic gradient descent (momentum SGD). Training is terminated after 150 epochs. It is used a learning rate of \( 10^{-2} \) for the first 80 epochs, \( 10^{-3} \) for the next 20 epochs, \( 10^{-4} \) for the next 20 epochs, and \( 10^{-5} \) for the last 30 epochs. The weights of AttentiveNet are initialized randomly. A percentage of 50% of dropout [21] is applied to the fully.
connected layers to prevent overfitting. In addition, batch normalization \[10\] is applied to all the convolutional layers to accelerate the learning.

Let us now first assess the performance of attention mechanisms. The ESC-10 corpus described in Sec II is used to train AttentiveNet in a supervised manner. Same increased data as in the work of [22] is used, modulo the time extracted by recording, from 1.5 seconds to 2.5 seconds. This modification is motivated by the insertion of the mechanisms of attention. By evaluating our error rate in the same way as [22] we observe that the modifications introduced through AttentiveNet do not degrade the performances of EnvNet. Indeed, AttentiveNet achieves a slightly lower error rate than EnvNet from 11.05% for EnvNet to 10.3% for AttentiveNet.

Considering the Location system performance, the corpus generated randomly described in Sec. 2 is used to evaluate the performance of the semi-supervised localization method presented in Sec. 3. Qualitatively, it is immediate to observe in the lower part of Fig. 2 that the system is capable, despite the presence of background noise, of locating the sound events on which it has been trained. The model has a precision of 78.89%, a false negative ratio of 10.97%, a ratio of true positives of 44.80%, a ratio of false positives of 11.99%, and a ratio of true positives of 32.24%. Confusion matrix in Fig. 3 tells us that the system can discern and locate sounds. It is observed that the most difficult classes to predict (sneeze and clock or tick) are those containing the most silence in ESC.

V. CONCLUSIONS

Service robots will have important implications for individual user experience. It is important to stress that, in the future, virtually all service robots will be connected and embedded into a bigger system (e.g. via knowledge bases and cloud-based systems). That is, in addition to their local input channels (e.g. cameras, microphones, and sensors) they can access data from a wide range of other sources including the internet, the collective organizational knowledgebase and its customer relationship management (CRM) system which contains customer background, preference and transaction data. Combined
with biometrics (e.g. facial and voice recognition systems [5], [9]), a service robot will be able to identify a user and provide highly customized and personalized service on scale at negligible marginal cost [9]. Thus, to such service robots, audio pattern recognition will become highly important, and in a wide view, as they are intelligent systems, they can greatly benefit from an ability to perceive the human environment and its major actors, allowing them to better understand what is happening around them. The proposed network, AttentiveNet, is useful to this aim and it is capable of classifying sound events with efficiency comparable to the original EnvNet model on the ESC-10 database. It is shown that in addition to the classification, our network modification allowed us to temporally locate sound events with high accuracy.

Although sufficient for a proof of concept, the database used is quite limited. We therefore see several possible extensions to this preliminary work. In order to test our approach more systematically, it is planned to evaluate recognition and location using different levels and sources of noise, and to in-cresce the difficulty of the task by applying our approach to the ESC-50 database, containing 50 classes of sound events. Further, other datasets can be used. One of great interest is UrbanSound8K [19]: a collection of 8732 short (less than 4 seconds) excerpts of various urban sound sources (air conditioner, car horn, playing children, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music) prearranged into 10 folds. If the dataset size is too small to use for deep learning, modifications will be made to the original data for instance at negligible marginal cost [9]. Thus, to such service robots, provide highly customized and personalized service on scale at negligible marginal cost [9].

Our ultimate goal is to be able to adapt this method to real conditions of use, as actual deep learning techniques suffer when data are captured in uncontrolled conditions, entailing variations and occlusions [15]. This will help in integration in a scenario of interaction between a robot and a human being.

ACKNOWLEDGMENTS
I would like to thank the two anonymous peer reviewers who have taken the time to both quickly and thoroughly review this manuscript.

REFERENCES