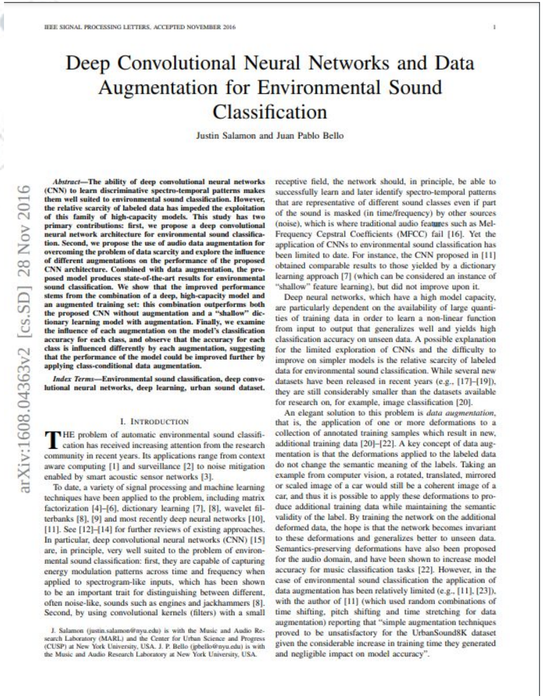
Projet DEEP LEARNING

Urban sound Classification

Sommaire :

1. L'article :



IEEE SIGNAL PROCESSING LETTERS, ACCEPTED NOVEMBER 2016 1

Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification

Justin Salamon and Juan Pablo Bello

arXiv:1608.04363v2 [cs.SD] 28 Nov 2016

***Abstract*—The ability of deep convolutional neural networks (CNN) to learn discriminative spectro-temporal patterns makes them well suited to environmental sound classification. However, the relative scarcity of labeled data has impeded the exploitation of this family of high-capacity models. This study has two primary contributions: first, we propose a deep convolutional neural network architecture for environmental sound classification. Second, we propose the use of audio data augmentation for overcoming the problem of data scarcity and explore the influence of different augmentations on the performance of the proposed CNN architecture. Combined with data augmentation, the proposed model produces state-of-the-art results for environmental sound classification. We show that the improved performance stems from the combination of a deep, high-capacity model and an augmented training set: this combination outperforms both the proposed CNN without augmentation and a "shallow" dictionary learning model with augmentation. Finally, we examine the influence of each augmentation on the model's classification accuracy for each class, and observe that the accuracy for each class is influenced differently by each augmentation, suggesting that the performance of the model could be improved further by applying class-conditional data augmentation.**

***Index Terms*—Environmental sound classification, deep convolutional neural networks, deep learning, urban sound dataset.**

I. Introduction

THE problem of automatic environmental sound classification has received increasing attention from the research community in recent years. Its applications range from context aware computing [1] and surveillance [2] to noise mitigation enabled by smart acoustic sensor networks [3].

To date, a variety of signal processing and machine learning techniques have been applied to the problem, including matrix factorization [4]–[6], dictionary learning [7], [8], wavelet filterbanks [8], [9] and most recently deep neural networks [10], [11]. See [12]–[14] for further reviews of existing approaches. In particular, deep convolutional neural networks (CNN) [15] are, in principle, very well suited to the problem of environmental sound classification: first, they are capable of capturing energy modulation patterns across time and frequency when applied to spectrogram-like inputs, which has been shown to be an important trait for distinguishing between different, often noise-like, sounds such as engines and jackhammers [8]. Second, by using convolutional kernels (filters) with a small receptive field, the network should, in principle, be able to successfully learn and later identify spectro-temporal patterns that are representative of different sound classes even if part of the sound is masked (in time/frequency) by other sources (noise), which is where traditional audio features such as Mel-Frequency Cepstral Coefficients (MFCC) fail [16]. Yet the application of CNNs to environmental sound classification has been limited to date. For instance, the CNN proposed in [11] obtained comparable results to those yielded by a dictionary learning approach [7] (which can be considered an instance of "shallow" feature learning), but did not improve upon it.

Deep neural networks, which have a high model capacity, are particularly dependent on the availability of large quantities of training data in order to learn a non-linear function from input to output that generalizes well and yields high classification accuracy on unseen data. A possible explanation for the limited exploration of CNNs and the difficulty to improve on simpler models is the relative scarcity of labeled data for environmental sound classification. While several new datasets have been released in recent years (e.g., [17]–[19]), they are still considerably smaller than the datasets available for research on, for example, image classification [20].

An elegant solution to this problem is *data augmentation*, that is, the application of one or more deformations to a collection of annotated training samples which result in new, additional training data [20]–[22]. A key concept of data augmentation is that the deformations applied to the labeled data do not change the semantic meaning of the labels. Taking an example from computer vision, a rotated, translated, mirrored or scaled image of a car would still be a coherent image of a car, and thus it is possible to apply these deformations to produce additional training data while maintaining the semantic validity of the label. By training the network on the additional deformed data, the hope is that the network becomes invariant to these deformations and generalizes better to unseen data. Semantics-preserving deformations have also been proposed for the audio domain, and have been shown to increase model accuracy for music classification tasks [22]. However, in the case of environmental sound classification the application of data augmentation has been relatively limited (e.g., [11], [23]), with the author of [11] (which used random combinations of time shifting, pitch shifting and time stretching for data augmentation) reporting that "simple augmentation techniques proved to be unsatisfactory for the UrbanSound8K dataset given the considerable increase in training time they generated and negligible impact on model accuracy".

J. Salamon (justin.salamon@nyu.edu) is with the Music and Audio Research Laboratory (MARL) and the Center for Urban Science and Progress (CUSP) at New York University, USA. J. P. Bello (jpbello@nyu.edu) is with the Music and Audio Research Laboratory at New York University, USA.

- **Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification**

- Auteur : Justin Salamon and Juan Pablo Bello
- Cité 916 fois
- L'interet de l'article
  - CNN pour sons urbains
  - Audios augmentés

2. Dataset :

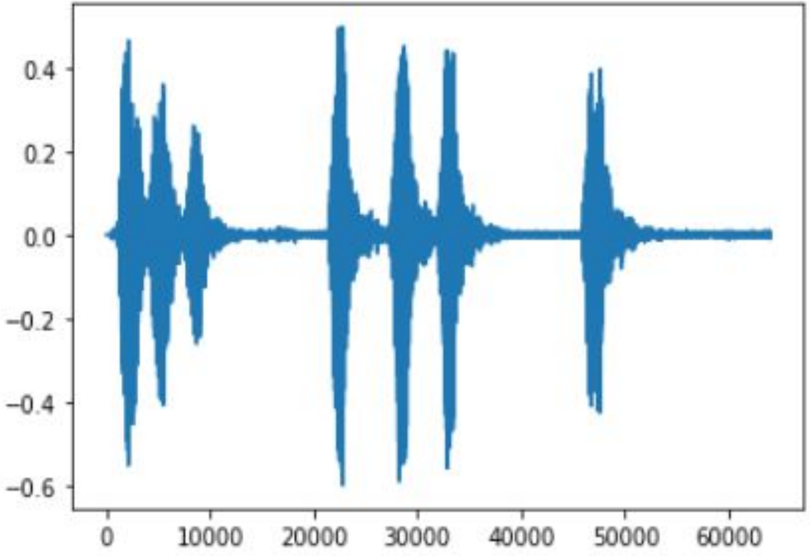
- **UrbanSound8K**
- 8732 sons urbains de moins de 4 secondes
- 10 classes : air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music
- Separés en 10 fold

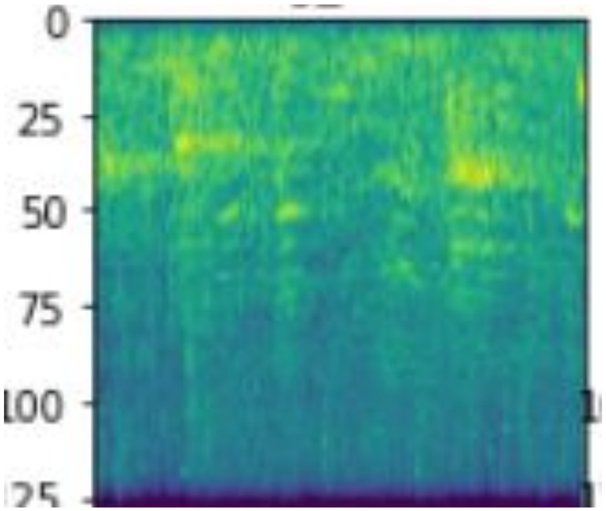
3.

Expériences et résultats

3.1 : Reproduction de l'expérience de l'article.

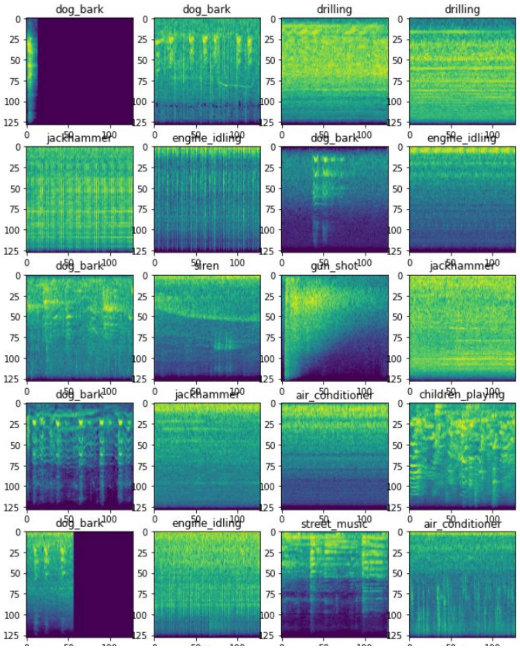
- .WAV → MelSpectrogram





3.1 : Reproduction de l'expérience de l'article.

- Les différents mel obtenus selon les classes



3.1 : Reproduction de l'expérience de l'article.

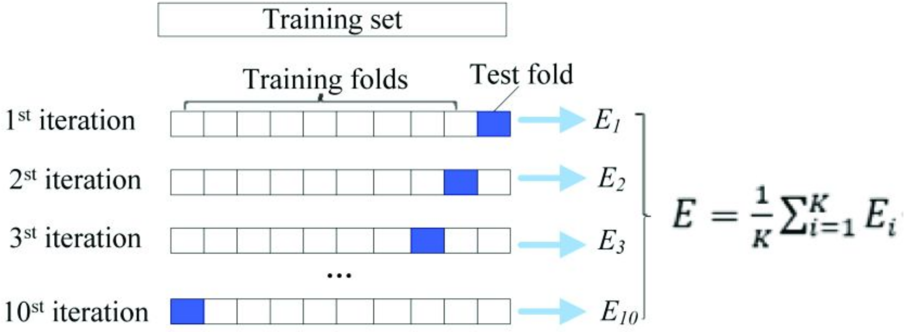
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 124, 124, 24)      624
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 41, 41, 24)        0
_________________________________________________________________
activation (Activation)      (None, 41, 41, 24)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 38, 38, 36)        13860
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 19, 19, 36)        0
_________________________________________________________________
activation_1 (Activation)    (None, 19, 19, 36)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 17, 17, 48)        15600
_________________________________________________________________
activation_2 (Activation)    (None, 17, 17, 48)        0
_________________________________________________________________
global_average_pooling2d (Gl (None, 48)                0
_________________________________________________________________
dense (Dense)                (None, 60)                2940
_________________________________________________________________
dropout (Dropout)            (None, 60)                0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                610
=================================================================
Total params: 33,634
Trainable params: 33,634
Non-trainable params: 0
_________________________________________________________________
```

- CNN
- 3 Couches de Conv2D
- 2 couches Dense

- 10-fold cross validation

Résultat du 10-fold cross validation : 100 epochs

Fold	1	2	3	4	5	6	7	8	9	10	Moyenne
Accuracy	75,48	74.01	66.59	74.54	84.08	72.90	70.40	71.83	79.41	78.73	74.79
Loss	0.90	0.82	1.01	0.87	0.63	0.89	0.88	1.08	0.79	0.75	0.77



3.2 : Data augmentation

- **PitchShift** : +/- 4 demi tons
- **GaussianNoise** : Ajout bruit gaussien
- **TimeStretch** : changement durée/tempo du son (entre x0,8/x1,25)

Augmentation	PitchShift	GaussianNoise	TimeStretch
Classe	Air_conditioner Car_horn Engine_idling Gun_shot Jackhammer	Car_horn Children_playing Gun_shot Jackhammer	Car_horn Gun_shot Jackhammer

- Avant augmentation : 8732 sons
- Après augmentation : 15771 sons

Résultat : 100 epochs avec data augmentation

Fold	1	2	3	4	5	6	7	8	9	Moyenne
Accuracy	74.88	73.42	66.37	73.83	81.62	69.98	71.47	67.12	78.19	77.06
Loss	1.04	0.92	1.15	0.95	0.73	1.02	0.93	0.93	0.78	0.93

- Avant augmentation : 74.79 % acc
- Après augmentation : 77.06 % acc

3.3 : Influence des 10 fold

- Réalisation des expériences sans séparation des données en fold et sans DA.
  - 80% de train
  - 10% de validation
  - 10% de test

No Fold	Test	Train
Accuracy	0.75	0.90

Sans augmentation

(8732 sons)

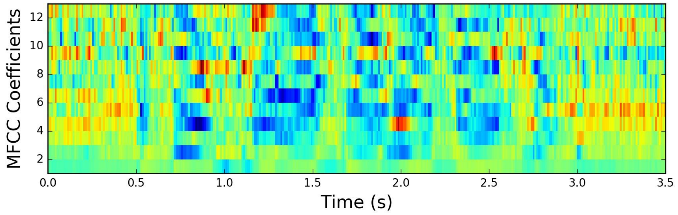
No Fold	Test	Train
Accuracy	0.73	0.88

Avec augmentation

(15771 sons)

3.4 : Tout autre représentation de l’audio

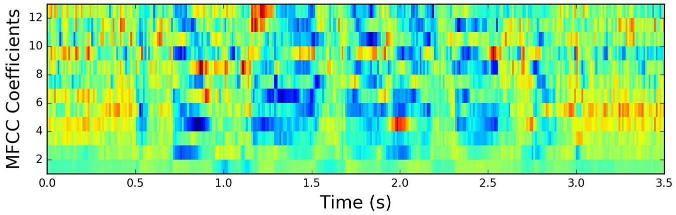
- MFCC Mel-Frequency Cepstral Coefficients

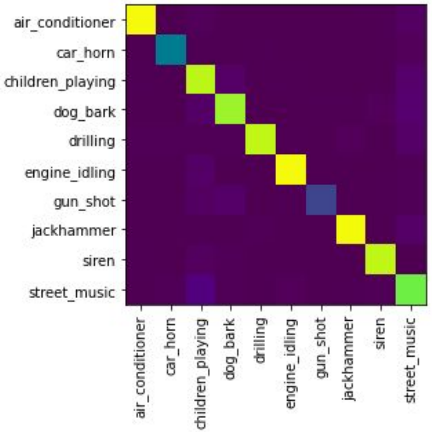


Fold	1	2	3	4	5	6	7	8	9	10	Moyenne
Accuracy	0.48	0.51	0.46	0.47	0.54	0.47	0.58	0.50	0.59	0.6	0.52
Loss	4.2	2.7	2.3	2.04	1.7	2.13	1.28	2.19	1.9	1.49	2.193

3.4 : Tout autre représentation de l'audio

- MFCC Mel-Frequency Cepstral Coefficients - sans fold





	Accuracy
train	0.92
test	0.88

Conclusion

Pistes d'améliorations :

- ❖ Leaf by Google
- ❖ yamNet

Merci de votre attention!

Des questions?

Maxence David

Ambre Baffert

Lisa Casino