Localisation sonore par attention et apprentissage profond semi-supervisé

M. Moreaux[1,2,3] M. Garcia Ortiz[1] I. Ferrané[2] F. Lerasle[3]

[1] Softbank Robotics Europe - AI Lab, Paris
[2] IRIT, Université de Toulouse, CNRS, Toulouse
[3] LAAS, Université de Toulouse, CNRS, Toulouse

mmoreaux@laas.fr

1 Résumé

Afin d'interagir avec l'humain et son environnement, un robot de service doit pouvoir percevoir des informations visuelles et sonores de la scène qu'il observe ou à laquelle il participe. Il doit notamment être capable de repérer des éléments saillants dans les différents signaux captés : localisation spatiale dans une image ou temporelle dans un flux audio. L'aspect "datavore" des méthodes dites d'apprentissage profond, et le coût considérable de l'annotation des données, militent pour l'utilisation de méthodes semi-supervisées, capables d'une part d'extraire de l'information de manière supervisée, et d'autre part de prédire l'organisation spatiale ou temporelle des événements présents dans le signal traité. Dans le domaine de la vision, ce concept a été utilisé à plusieurs reprises pour effectuer de la localisation spatiale d'objet ou d'activité sur des images [1, 2, 3] à partir des signaux 2D bruts (pixels). Au niveau audio, la tendance consistant à s'affranchir des représentations bas niveau de type MFCC [4] a fait son apparition, permettant ainsi un traitement direct du signal audio brut [5, 6, 7, 8] et laissant aux réseaux de neurones la tâche d'extraire les caractéristiques représentatives optimales des signaux traités. Dans cet article, nous proposons un réseau convolutionnel, associé à un mécanisme d'attention, permettant l'exploitation du signal audio brut, afin non seulement de classifier, mais aussi de localiser temporellement un événement sonore présent dans le flux traité, et ce de manière semi-supervisé.

1.1 Etat de l'art

Plusieurs travaux ont proposé d'apprendre des réseaux de neurones à partir de sons bruts. Nous nous inspirons de Wavenet [5] générant de la musique, de SampleCNN [7] classifiant de la musique ou EnvNet [8] classifiant des sons environnementaux et proposons d'intégrer des mécanismes d'attention généralement utilisés en vision. Nous prenons exemple sur le "Global Average Pooling" (GAP) [9] permettant d'inférer une localisation spatiale en mode semi-supervisé [1, 3], en effet, une opération de moyennage couplée à un concept de 'Class Activation Mapping' (CAM) permet aux auteurs de localiser une classe d'objet ou d'activité dans une image et sur des "Gated Convolutional Layers" [10], aussi utilisées dans des contexte audio [5], que nous considérons ici comme un mécanisme d'attention. A notre connaissance, la localisation semi-supervisée de sons environnementaux par réseaux convolutionnels, centrale dans les travaux présentés ici, a été peu explorée.

2 Corpus audio utilisé

Le corpus ESC10 [11] conçu pour la détection d'événements sonores est composé de 10 classes correspondant à des sons environnementaux enregistrés en intérieur ou en extérieur : *chien, coq, vagues, pluie, feu de braise, horloge, hélicoptère, tronçonneuse, bébé qui pleure et éternuement*. Chaque classe est constituée de 40 enregistrements de 5 secondes. Comme dans les travaux de [11], nous avons appris notre système sur la base d'une validation croisée réalisée sur 5 parties, en ayant au préalable appliqué, à l'instar de [8], un prétraitement pour obtenir des enregistrements échantillonés à 16 Khz et codé sur 16 bits. Pour évaluer notre capacité à localiser temporellement un événement sonore, nous générons aléatoirement une base de test constituée de 1600 extraits de 8 secondes chacun. Chaque enregistrement comprend 8 secondes de signal tirés aléatoirement d'un bruit d'ambiance provenant d'un restaurant [1] sur lequel sont ajoutés deux signaux sonores, de 2.5 secondes chacun, aléatoirement sélectionné depuis la base de test ESC10 et aléatoirement positionnés en interdisant le recouvrement.

1. https://freesound.org/people/MrCream/sounds/78378/

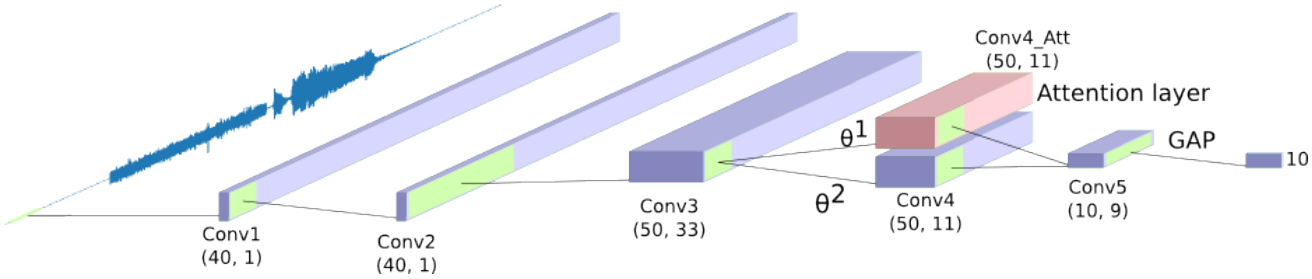

FIGURE 1 – Envnet-Att-GAP, 2.5 sec en entrée.

3 Présentation de notre système

Notre système, présenté en Figure 1, est basé sur le réseau Envnet [8], auquel nous ajoutons des éléments propres à la localisation semi-supervisée : deux systèmes d'attention, ainsi qu'un mécanisme naïf de localisation semi-supervisée.

Mécanisme d'attention sur les couches basses du réseau. Le premier mécanisme d'attention sous-jacent au réseau consiste à introduire une contrainte multiplicative sur une couche convolutive d'Envnet (Figure 1). Cette contrainte est implémentée en multipliant une couche standard du réseau original *conv4* (dont les activations sont des fonctions d'activation de type "ReLu") par une nouvelle couche *conv4_att* dont les activations sigmoïdales agissent comme des portes régulant le signal du "Relu". Cette méthode, initiée d'abord sur des données image par [10] fut appliquée dans un contexte audio par [5].

Mécanisme d'attention sur les couches hautes du réseau. Le second mécanisme d'attention introduit dans notre réseau ajoute une couche dite de "Global Average Pooling". Cette technique, proposée d'abord pour de la localisation spatiale, est également appropriée pour de la localisation temporelle d'événements sonores dans le signal audio. De manière équivalente à [3], la dernière couche convolutionelle d'Envnet, qui est convoluée dans le temps, est connectée à une nouvelle couche de convolution contenant autant de neurones que de classes (10 neurones pour 10 classes avec ESC-10) puis passe par une couche de "Global Average Pooling" et de "softmax" résultant en 10 prédictions. Le modèle final est noté Envnet-Att-GAP.

Localisation temporelle. Nous proposons une stratégie basée sur un ensemble d'heuristiques, afin d'extraire la localisation des événements sonores présents dans un enregistrement sonores de plusieurs secondes depuis la couche de GAP :

1. Nous définissons un masque binaire d'activation (MBA) basé sur une valeur de seuil $\tau = 150\%$ de l'activation maximale des neurones du GAP (Fig. 2(b)) lorsque le signal audio d'entrée correspond à du silence. $\text{MBA}_i = 1$ si $\text{GAP}_i > s$ sinon 0
2. Ce masque binaire d'activation est filtré de telle sorte que les activations soient égales à 1 si elles sont voisines d'au moins deux autres activations positives dans une fenêtre centrée de 10 activations, et à 0 sinon.
3. Enfin, la méthode CAM est utilisée sur chacune des fenêtres positives du masque binaire d'activation filtré pour prédire quelle est la nature du son qui s'y trouve.

4 Résultats

Performance des mécanismes d'attention. Le corpus ESC-10 décrit en Sec. 2 est utilisé pour entraîner Envnet-Att-GAP de manière supervisée. Nous utilisons les mêmes augmentations de données que dans les travaux de [8], modulo la durée extraite par enregistrement, passant de 1.5 secondes à 2.5 secondes. Cette modification est motivé par l'insertion des mécanismes d'attention.En évaluant notre taux d'erreur de la même manière que [8] nous observons que les modifications introduites au travers d'Envet-Att-GAP ne dégradent pas les performances d'Envnet. En effet, Envnet-Att-Gap atteint un taux d'erreur légèrement inférieur à Envnet nous faisant passer de 11.05% pour Envnet à 10.3 pour Envnet-Att-Gap%.

Performance du système de localisation. Le corpus généré aléatoirement décrit en Sec. 2 est utilisé pour évaluer les performances de la méthode de localisation semi-supervisée présentée en Sec. 3. Qualitativement, nous observons dans la Fig. 2(a) que le système est capable, malgré la présence du bruit de fond, de localiser les événements sonores sur lesquels il a été entraîné. Le modèle a une précision de 78.89%, un ratio de faux negatifs de 10.97%, un ratio de vrais positifs de 44.80%, un ratio de faux positifs de 11.99%, et un ratio de vrais negatifs de 32.24%. La matrice de confusion en Fig. 2(c) nous indique que le système arrive a discerner les sons et a les localiser. Nous constatons que les classes les plus difficiles à prédire (*éternuer* et *horloge ou tic-tac*) sont celles contenant le plus de silence dans ESC.

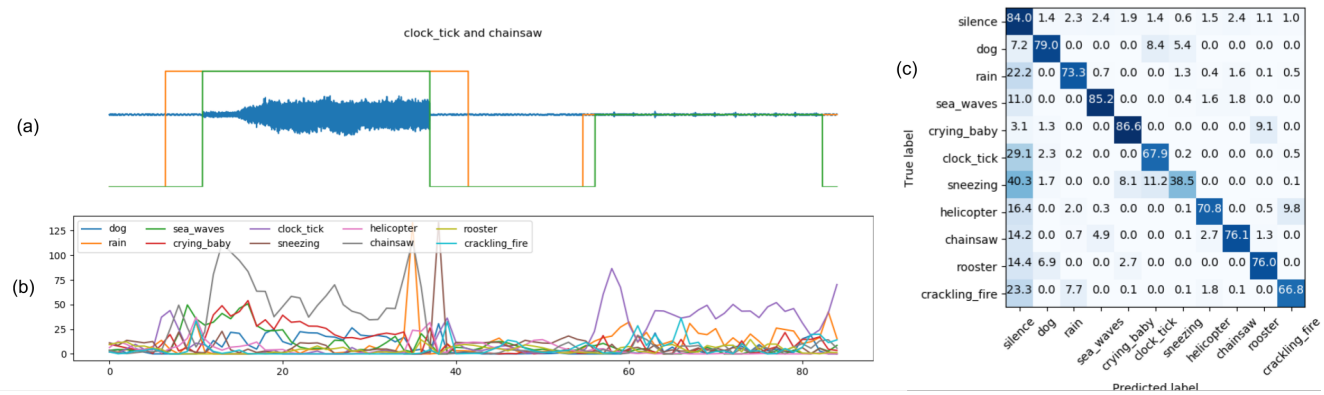

FIGURE 2 – Localisation temporelle d'événements sonores : (a) son généré aléatoirement avec la méthode décrite en section 2 (en vert) et la prédiction faite par le système Envnet-Att-GAP (en orange), (b) activations des neurones par classe provenant de la couche GAP, (c) matrice de confusion des 10 classes plus du 'silence' à interpréter comme 'pas de classes'.

5 Conclusion

Le réseau proposé, Envnet-Att-GAP, est capable de classifier les événements sonores avec une efficacité comparable au modèle original Envnet sur la base de données ESC-10. Nous avons montré qu'en plus de la classification, notre modification du réseau nous permettait de localiser temporellement des événements sonores avec une précision élevée. Bien que suffisante pour une preuve de concept, la base de données utilisée est assez limitée. Nous voyons donc plusieurs extensions possibles à ce travail préliminaire. Afin de tester notre approche de manière plus systématique, nous prévoyons d'évaluer la reconnaissance et la localisation en utilisant différents niveaux et sources de bruits, et d'augmenter la difficulté de la tache en appliquant notre approche à la base de données ESC-50, contenant 50 classes d'événements sonores. Notre but à terme est de pouvoir adapter cette méthode à des conditions réelles d'utilisation, et de pouvoir l'intégrer dans un scénario d'interaction entre un robot et un être humain.

Références

[1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016.

[2] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free ?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[3] Marc Moreaux, Natalia Lyubova, Isabelle Ferrané, and Frederic Lerasle. Mind the regularized gap, for human action classification and semi-supervised localization based on visual saliency. *VISAPP*, 2018.

[4] Beth Logan et al. Mel frequency cepstral coefficients for music modeling.

[5] Aaron Van Den Oord, Sander Dieleman, and et al. Zen. Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499*, 2016.

[6] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv :1703.01789*, 2017.

[7] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Samplecnn : End-to-end deep convolutional neural networks using very small filters for music classification. *Applied Sciences*, 8(1) :150, 2018.

[8] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv :1711.10282*, 2017.

[9] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv :1312.4400*, 2013.

[10] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[11] Karol J Piczak. Esc : Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM, 2015.