Proposition de thèse en informatique à l'IRIT, Toulouse
2017

Titre du sujet : réseaux de neurones profonds pour l'apprentissage de concepts audio haut-niveau

Domaine : apprentissage automatique, traitement du signal

Mots clés : *deep learning*, réseaux de neurones profonds, audio

Laboratoire d'accueil : IRIT, Université Toulouse III - Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse

Contact : Thomas Pellegrini, thomas.pellegrini@irit.fr, https://www.irit.fr/~Thomas.Pellegrini/

Les humains développent dès leur enfance des representations abstraites de concepts avec un fort pouvoir de généralisation. Ce processus s'appelle apprentissage de concepts, *concept learning,* aussi connu sous le nom d'apprentissage de categories, *category learning* [1]. Le sujet de cette thèse aborde la question suivante : **est-ce que les machines sont capables d'apprendre des concepts et de construire leurs propres schémas représentatifs, à partir de signaux sonores ?** On s'intéressera donc à des concepts liés à l'audio, entre autres les chants d'oiseaux [2]. Si nous entendons un oiseau d'une espèce donnée pour la première fois, nous sommes tout à fait en mesure de reconnaître qu'il s'agit d'un oiseau même si nous ne connaissons pas son espèce précisément. En apprentissage automatique, cela se révèle plus compliqué : il est possible d'entraîner des modèles, comme par exemple des réseaux de neurones profonds de type convolutifs pour détecter les chants d'oiseaux. Cela nécessite un grand corpus d'enregistrements sonores contenant des oiseaux annotés manuellement pour disposer d'une « vérité terrain » qui permet d'entraîner un modèle (apprentissage dit supervisé). Cependant, malgré de grandes quantités de données d'apprentissage, il est très probable que le modèle échoue à détecter le chant d'un oiseau qu'il n'aura pas vu dans la base d'apprentissage (manque de pouvoir de généralisation).

Dans cette thèse, il s'agira de mener des recherches sur **l'apprentissage de concepts sonores haut-niveau** pour améliorer le pouvoir de généralisation des modèles, dans le cadre du *deep learning,* plus précisément à l'aide de réseaux de neurones profonds (DNNs et/ou CNNs). Nous nous intéresserons à des données annotées manuellement de manière « grossière », c'est-à-dire des données annotées avec une seule étiquette par document sonore. Concernant l'exemple des oiseaux, l'étiquette serait dans ce cas une étiquette binaire disant si un ou des chants d'oiseaux sont audibles dans chaque enregistrement à disposition, sans préciser où se trouvent ces chants d'oiseaux précisément dans les fichiers.

L'objectif de la thèse sera double :

- Analyser les représentations apprises par un DNN (CNN) pour caractériser ce que signifie le concept sonore visé pour le modèle entraîné sur des données grossièrement annotées,
- Mettre en œuvre des approches de localisation temporelle des concepts dans les enregistrements à disposition, *i.e.* rendre possible une annotation précise des fichiers sonores de manière automatique, à partir des DNNs entraînés sur des données grossièrement annotées.

Pour répondre à ces objectifs, une piste consiste à tenter d'identifier les zones « saillantes » des documents sonores, saillantes au sens de zones qui ont été utilisées par le réseau de neurones pour qu'il prenne sa décision de classification [3].

Bibliographie récente de l'équipe en lien avec le sujet

Pellegrini, T., & Mouysset, S. (2016). Inferring Phonemic Classes from CNN Activation Maps Using Clustering Techniques. In Proc. INTERSPEECH, San Franscico, 2016, pp. 1290-1294.

C. Manenti, T. Pellegrini, J. Pinquier, CNN-Based Phone Segmentation Experiments in a Less-Represented Language, in Proc. INTERSPEECH, San Francisco, Sept. 2016

T. Pellegrini, V. Barriere. Time-continuous estimation of emotion in music with recurrent neural networks, in Proc. Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Sept. 2015

T. Pellegrini. Comparing SVM, softmax, and shallow neural networks for eating condition classification, in Proc. INTERSPEECH, Dresden, Sept. 2015

Bibliographie générale

[1] Bruner, J., Goodnow, J. J., & Austin, G. A. (1967). A study of thinking. New York: Science Editions

[2] Stowell, D., Wood, M., Stylianou, Y., & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. *arXiv preprint arXiv:1608.03417*

[3] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning Deep Features for Discriminative Localization. arXiv preprint arXiv:1512.04150