Découverte non supervisée des unités sonores élémentaires d'un langage parlé

Stage M2 Recherche

Encadrants :
T. Artières – QARMA@LIS - https://pageperso.lis-lab.fr/thierry.artieres/
R. Marxer - DYNI@LIS - http://www.ricardmarxer.com/

Durée : 5 à 6 mois (Avril - Septembre)

Rémunération : Tarif stage (non encore acquise)

Sujet :

La segmentation et labellisation des données séquentielles, une tâche complexe qui vise à simultanément identifier les différentes parties d'une séquence et à leur attribuer des catégories, est un problème récurrent dans des nombreuses disciplines. La transcription de la parole pour les interfaces homme-machine ou le découpage en sous-tâches pour le hierarchical reinforcement learning en sont des exemples. L'apprentissage profond est spécialement adapté pour des données structurées et séquentielles en particulier et a en effet fait sauter de nombreux verrous pour des données du type parole, vidéo, ou texte. Néanmoins les méthodes de *deep learning* requièrent un grand nombre de données labellisées ce qui restreint son domaine d'application. Même si des variantes semi-supervisées ou basées sur l'apprentissage actif peuvent réduire significativement la quantité de données requise, le besoin d'input expert rend moins générique la solution tout en apportant des biais humains. Concernant les données séquentielles, les autres approches classiques de *machine learning* comme les CRF, les HMMs ne sont pas non plus très bien adaptées à l'apprentissage avec peu de données ni au mode semi-supervisé, et surtout n'égalent pas l'expressivité apportée par les réseaux neuronaux profonds.

Ce stage vise à apporter des solutions neuronales efficaces pour la segmentation non supervisée de signaux de parole d'un langage inconnu, c'est à dire à l'apprentissage en mode non supervisé des unités élémentaires de la langue. Cette thématique est récente et difficile, elle fait notamment l'objet d'une série de compétitions depuis quelques années [Zerospeech]. Diverses approches ont été proposées dans les éditions précédentes de la compétition (en 2015 et 2017) pour des tâches similaires quoiqu'un peu différentes. Certaines des méthodes utilisées jusqu'ici exploitent de modèles neuronaux de type autoencoders ou variational

autencoders [Glarner et al., 2018] pour apprendre des représentations compactes des sons élémentaires (généralement au niveau sous phonétique) [van den Oord et a., 2017] . Une discrétisation de cet espace permet alors d'inférer à la fois la nature des phonèmes de la langue et la segmentation de signaux dans ces phonèmes découverts.

Toutefois les travaux précédents sont encore limités dans le sens où ils ne prennent pas explicitement en compte la nature langagière du signal de parole, i.e. les autoencodeurs sont appris sur des sons courts mais sans prise en compte des sons les précédant ou les suivant. Par ailleurs les langages, quels qu'ils soient présentent des régularités statistiques notamment en termes de distribution (fréquence relatives d'apparition des phonèmes) et de perplexité (nombre de phonèmes alternatifs possibles après un phonème donné). Nous proposons d'explorer des méthodes étendant l'état de l'art dans ces deux directions, en cherchant à intégrer un apprentissage à la SkipGram de [Mikolov et al., 2013] dans l'apprentissage d'autoencodeurs type VQ-VAE ou à intégrer dans ce type de modèles des contraintes sur les distributions obtenues à l'aide de stratégies adversarial popularisées ces dernières années. La méthode sera testée sur quelques jeux des données (synthétique et réels) et les résultats seront comparées aux alignements basées sur des annotations humaines.

Références

[Zerospeech] http://www.zerospeech.com/

[Mikolov et al., 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean, Distributed Representations of Words and Phrases and their Compositionality. *in Neural Information Processing Systems* NIPS 2013: 3111-3119.

[van den Oord et al., 2017] van den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." *Advances in Neural Information Processing Systems(NIPS)*. 2017.

[Glarner et al., 2018] Glarner, T., Hanebrink, P., Ebbers, J., Haeb-Umbach, R., Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery. *in Proceedings of Interspeech 2018*, 2688-2692, DOI: 10.21437/Interspeech.2018-2148.