Data camp : project description

Lu Lin and Olivier coudray

January 21, 2018

Motivations

Le Deep Learning est aujourd'hui appliqué et commercialisé dans plusieurs domaines, notamment en Computer-Vision, NLP (*Natural Language Processing*), *Speech Recognition*, *Reinforcement Learning*...etc. Récemment, Google (Youtube) a encore innové en matière de *Speech Recognition*. Le site propose en effet pour certaines vidéos, un sous-titrage automatique (par *speech recognition*). La machine est aussi capable d'identifier la source sonore (e.g le son de la musique, le bruit du métro, ...etc). En nous inspirant des procédés mis en oeuvre dans le cadre de ces applications, nous voulons travailler d'abord sur un sous problème de *Speech Recognition* proposé comme compétition Kaggle – *'TensorFlow Speech Recognition Challenge'* (avec le prix 25,000 dollars)[1].

Description du problème

Nous allons commencer par la compétition Kaggle mentionnée ci-dessus. Il s'agit, à partir d'un dataset conséquent (environ 65 000 données) de pouvoir reconnaître des commandes simples (parmi 30 mots seulement). Les données sont disponibles sous un format brut (fichiers *.wav*). Chaque enregistrement dure une seconde et correspond à un mot. L'enjeu ici est donc double : il s'agira d'une part de nous familiariser avec des méthodes de preprocessing adaptées aux fichiers audio puis de travailler sur les algorithmes de machine learning afin d'obtenir la meilleure précision possible.

Si cet objectif est atteint suffisamment rapidement, nous pourrons nous tourner vers des problèmes plus complexes.

Plan du projet

En fonction de la difficulté du problème, nous ajusterons le projet au cours du temps. Nous commencerons par traiter le problème décrit ci-dessus. Puis, si les résultats sont satisfaisants et que le temps nous le permet, nous augmenteront la difficulté du problème pour nous intéresser à de véritables problématiques de *speech recognition* (dans un ordre de difficultée estimée croissante):

- Chatbot/assistant vocal (extension directe de 'Command recognition')
- *Speech recognition* (transformer la parole vocale nette en texte, puis avec des bruits)
- Voiceprint I (construction de l'identité vocale d'humain)
- Voiceprint II (identification de la source sonore. e.g. le son de la musique, le bruit du métro, du trains ...etc) (**manque datasets** ?)
- Reading system (speak like a humain) (Utiliser peut-être GAN)

Pour ces problématiques les données Kaggle ne seront plus suffisantes. Nous pourrons utiliser des bandes sonores de films associées à des sous-titres (youtube par exemple).

Les objectifs listés ci-dessus sont, pour certains (voire beaucoup), assez ambitieux. Aussi, nous les traiterons uniquement si le temps nous le permet et par ordre de difficulté croissante.

[1] La compétition est malheureusement déjà terminé depuis le 16 janvier 2018.