Clémence Lebecque
Léonard Fromont
Marie-France Jallade

M2LI 2019-2020

Devoir de reconnaissance vocale

Pour ce projet, nous avons travaillé sur la différenciation automatique d'accents entre plusieurs pays francophones, à partir de fichiers audio. Cette tâche de classification est réalisée grâce à un réseau de neurones convolutif qui permet de faire de la classification d'images. Il apprend progressivement à distinguer les classes entre elles pour, à terme, associer un spectrogramme (la représentation en image d'un fichier son) à une classe plutôt qu'à une autre; la classe prédite est la classe à laquelle l'objet a le plus de chance (ici de probabilité) d'appartenir selon le réseau. Nous utilisons la bibliothèque Keras pour implémenter ce classifieur.

I - Présentation et pré traitement des données

Les données utilisées sont issues de Mozilla Common Voice dans la catégorie du français. Elles se trouvent dans un dossier zippé qui regroupe les enregistrements d'énoncés au format .mp3 ainsi que leurs métadonnées dans des fichiers .tsv. Nous travaillons sur les données "validées", ce qui signifie qu'elles ont été au préalable annotées avec le profil du locuteur (sexe, âge, origine), la phrase énoncée et un numéro d'identifiant, et que leur qualité est garantie.

Pour décider des pays sur lesquels travailler, nous avons compté combien de fichiers audios par pays étaient disponibles, afin de sélectionner les pays qui en contenaient le plus (voir annexe 1). Ainsi, nous avons utilisé le français parlé dans quatre pays: la France, la Suisse, la Belgique et le Canada. Pour ne pas travailler sur des données trop éparses et pour nous assurer que les distinctions apprises par le réseau ne seraient pas dues à la pluralité des profils (par exemple, nous ne voudrions pas que le réseau se focalise sur les différences acoustiques entre les hommes et les femmes ou bien entre les adolescents et les personnes âgées), nous avons décidé de cibler un profil de locuteur précis et majoritairement représenté dans nos données, à savoir les hommes entre 20 et 40 ans (voir annexe 2).

Nous avons ensuite dû choisir quel type de données nous souhaitions fournir à notre classifieur. Nous avons envisagé deux approches distinctes. Une approche assez linguistique aurait été de cibler un petit nombre de syllabes, voire de phonèmes qui nous paraissent révélateurs de différences entre les accents des quatre pays, et de les extraire des fichiers audio après les avoir alignés avec les phrases écrites. La deuxième possibilité était de prendre uniquement les fichiers son, sans prendre en compte leur transcription, de les découper en petits morceaux et de fournir ces extraits courts au réseau. Nous avons choisi d'emprunter cette deuxième approche car elle nécessite des prétraitements moins coûteux (pas d'alignement et de padding ou de redimensionnement nécessaires), mais surtout car elle permet d'avoir une quantité de données bien plus importante à fournir au réseau, ce qui est un grand avantage pour l'apprentissage. Notre intention était d'emprunter d'abord cette approche plus quantitative et, si les résultats obtenus n'étaient pas convaincants, de tester l'approche qualitative en alignant les fichiers et en extrayant des sons particuliers.

Les données ont donc suivi une chaîne de prétraitements avant d'être fournies au

réseau. Pour chacun des pays, nous avons sélectionné environ 2000 fichiers audios pour ne pas avoir de problème de déséquilibre entre les classes (c'est le nombre de fichiers disponibles pour la Suisse, qui a le moins de données parmis les quatre pays). Nous avons lancé le script Praat "extract_non_silent_parts_from_sound_files" qui nous était fourni sur l'ensemble de ces données, afin d'obtenir des fichiers audio ne contenant pas de silence. Puis nous avons découpé chacun de ces sons en extraits d'une seconde grâce à la librairie Pydub, et avons transformé ces morceaux d'une seconde en spectrogrammes. Ces traitements sont réalisés par la fonction "sound_to_spec".

Enfin, nous avons converti les spectrogrammes en array Numpy et les avons redimensionnés pour les préparer à l'entrée dans le réseau. Ces traitements sont effectués grâce à la fonction "load_data_images " qui associe également à chaque image son label de référence (un entier entre 0 et 3 pour les quatre pays, encodé dans un matrice binaire). Ainsi, en appliquant cette fonction sur les extraits d'une seconde des quatre pays, on obtient l'ensemble d'apprentissage et l'ensemble de test à fournir à notre modèle (séparation 80% pour l'apprentissage et 20% pour le test).

II - Présentation du modèle

Notre modèle est un réseau convolutif classique contenant 4 types de couches présentés ci-après, en respectant leur ordre d'apparition dans l'architecture du réseau. D'abord, les deux couches de convolutions (avec fonction d'activation Relu), chacune étant suivi de sa couche de Pooling. Le réseau prend en entrée une image de taille 28x28 dans sa première couche de convolution, cette dernière contient 32 filtres de taille 5x5, contre 64 dans la deuxième. Ensuite vient une couche Flatten transformant les tenseurs en vecteurs, suivie enfin d'une couche Dense avec une fonction softmax. Cette couche finale, qui représente la sortie du réseau, renvoie les probabilités qu'un spectrogramme donné appartienne à chacune des classes.

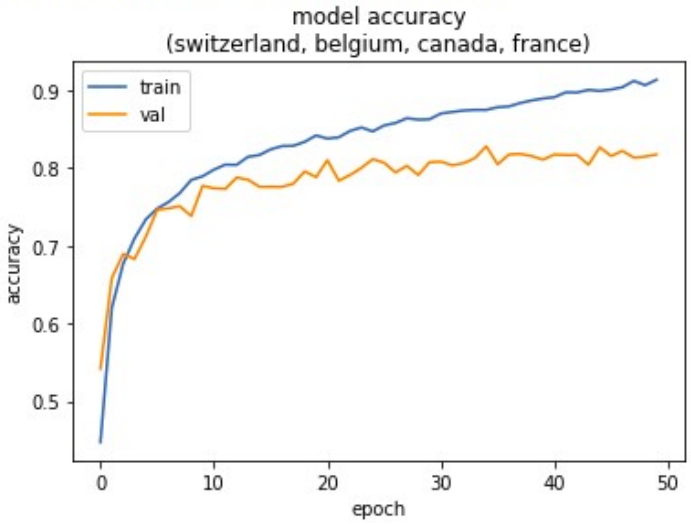
Nous avons opté pour l'optimiseur Adam, ainsi que l'entropie croisée comme fonction de perte. De plus, nous avons utilisé certains "callbacks" de Keras afin que le modèle final conservé soit celui ayant eu la perte la plus faible sur les données de validation. Pour cela, nous avons mis à True l'attribut "save_best_only" de ModelCheckpoint, et utilisé EarlyStopping qui arrête l'apprentissage une fois que la perte sur les données de validation ne diminue plus.

III - Analyse des résultats

Une fois que l'entraînement du modèle arrive à son terme et que le modèle le plus performant a été trouvé, ce dernier est utilisé afin de prédire les classes des données du test. Comme nous pouvons l'observer sur le graphique de gauche ci-dessous, la précision sur les données du test est très moyenne (63%). C'est pourquoi nous avons voulu tester notre classifieur avec un nombre de classes réduit, afin de voir s'il se comporte mieux. Nous avons entraîné ce classifieur-là sur les trois pays les plus dotés (la Belgique, le Canada et la France); nous pouvons observer que la précision sur les données de test est nettement meilleure avec un bond de 11 points de plus (74%).

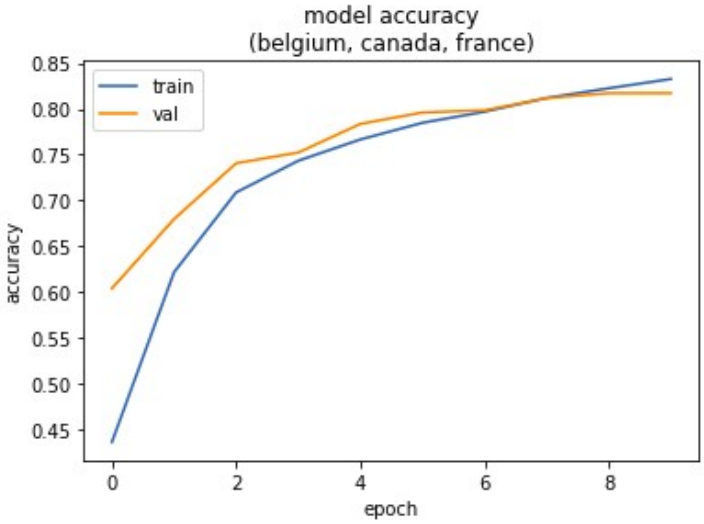
Étant donné que les résultats de cette approche quantitative nous semblent convaincants, nous n'avons pas eu recours à l'approche qualitative qui aurait consisté à extraire et donner au réseau des données contenant des sons spécifiques.

Test accurracy: 0.6271097046413502



Courbe d'apprentissage d'un modèle à 4 classes

Test accurracy: 0.7421972534332085



Courbe d'apprentissage d'un modèle à 3 classes

Enfin, afin de mieux interpréter ce que le modèle a appris, et ainsi tirer de potentielles conclusions linguistiques portant sur les différences phonétiques entre les français parlés au sein des différents pays, nous avions dans l'idée d'afficher les cartes d'activations à chaque étape du modèle à trois classes. Malheureusement, nous n'y sommes parvenus qu'à moitié: nous n'avons pas réussi à générer celles de la couche de sortie (la sortie softmax provoque un problème de dimensions), et nous n'avons pas pu en tirer de conclusions linguistiques. Néanmoins, après avoir agrandi les images générées, nous pouvons noter un comportement intéressant: les deux premières couches (la première couche de convolution et sa couche de pooling) semblent avoir des cartes d'activations similaires voire identiques pour les trois exemples étudiés; de légères différences apparaissent au bout de la troisième couche, et deviennent plus claires lors de la quatrième couche. Cela peut être expliqué par le fait que le réseau, avec cette couche de pooling, se focalise sur les légères différences qu'il avait cernées lors de la couche précédente. Cela a été observé sur les cartes d'activations qui se trouvent en annexe.

Annexe A

accent	sex	age	Compter - accent
belgium	female	fifties	5
		twenties	8
	male	fifties	155
		fourties	554
		seventies	29
		sixties	52
		teens	6
		thirties	3298
		twenties	608
canada	female	fourties	45
		thirties	866
	male	fifties	15
		fourties	712
		sixties	4
		teens	56
		thirties	973
		twenties	685
france	female	fifties	69
		fourties	3504
		sixties	488
		teens	286
		thirties	1052
		twenties	4607
	male	fifties	7065
		fourties	20234
		seventies	15
		sixties	1915
		teens	4371
		thirties	25920
		twenties	26294
		(vide)	1969
switzerland	male	fifties	29
		fourties	9
		teens	27
		thirties	51
		twenties	1392
	other	(vide)	607
(vide)	(vide)	(vide)	
Total Résultat			**107975**

Vue d'ensemble des profils de locuteurs par pays

accent	Compter - accent
cameroon	5
mayotte	6
st_pierre_et_miquelon	6
tunisia	6
cote_d_ivoire	8
martinique	11
guadeloupe	14
madagascar	15
portugal	17
french_guiana	18
senegal	22
other	43
italy	50
united_kingdom	56
united_states	56
germany	62
algeria	71
netherlands	82
monaco	104
benin	183
reunion	407
switzerland	2115
canada	3356
belgium	4740
france	98983
Total Résultat	**110436**

Nombre de fichiers audio par pays

Annexe B

BELGIQUE CANADA FRANCE

Cartes d'activations des couches de convolution et de pooling d'un modèle entraîné sur 3 classes