Réseaux de neurones pour la reconnaissance de l'oral : La reconnaissance du Beatbox

Jean Thomas Fidalgo Cardoso, Alexandre Nechab, Leo Rongieras, Marie Bauer

Colab CNN
Colab Wav2Vec2

Introduction

La reconnaissance de l'oral est un domaine de recherche en pleine expansion, avec des applications variées allant des assistants vocaux aux systèmes de transcription automatique. Parmi les défis spécifiques de ce domaine, la reconnaissance du beatbox se distingue par sa complexité et sa richesse sonore. Le beatbox, un art vocal imitant des instruments de percussion, présente des caractéristiques uniques qui rendent sa reconnaissance particulièrement intéressante et exigeante.

Ce projet se concentre sur la comparaison de deux méthodes distinctes pour la reconnaissance du beatbox : l'utilisation de réseaux de neurones convolutifs (CNN) et le modèle Wav2Vec. Les CNN, bien connus pour leur efficacité dans le traitement des images, ont également montré des résultats prometteurs dans le domaine de l'audio en raison de leur capacité à extraire des caractéristiques spatiales et temporelles. D'autre part, Wav2Vec, un modèle basé sur l'apprentissage non supervisé, a révolutionné la reconnaissance de l'oral en apprenant des représentations audio robustes directement à partir des formes d'onde brutes.

L'objectif de ce projet est de comparer les performances de ces deux approches en termes de précision, de robustesse et de capacité à généraliser sur des données de beatbox. En analysant les résultats obtenus, nous espérons non seulement identifier la méthode la plus efficace pour cette tâche spécifique, mais aussi contribuer à une meilleure compréhension des forces et des faiblesses de chaque approche dans le contexte plus large de la reconnaissance de l'oral.

Preparation du Dataset

Pour mener à bien notre projet de reconnaissance du beatbox, nous avons sélectionné le "Amateur Vocal Percussion Dataset" de Delgado (2019). Ce dataset est particulièrement adapté à notre étude en raison de sa richesse et de sa diversité. Il regroupe un total d'environ 10 000 utterances de beatbox, réparties dans 280 enregistrements audio. Les participants, au nombre de 28, sont des amateurs ayant peu ou pas d'expérience en beatbox, ce qui ajoute une dimension réaliste et accessible à notre analyse.

Le dataset est composé de quatre labels principaux : kick drum (kd), snare drum (sd), closed hi-hat (hhc) et opened hi-hat (hho). Ces sons de beatbox ont des caractéristiques linguistiques qui permettent de les distinguer. Le kick drum [p’], le snare drum [k’], le open hi-hat[ts:], et le closed hi-hat[ts’]. Ils se distinguent par leur lieu d'articulation: bilabial (kick), glottal(snare) et alvéolaire (hi-hat). Ils se distinguent aussi par la présence d’air pulmonaire ou pas, surtout pour différencier le open hi-hat et le closed hi-hat.

Ces labels sont représentés de manière assez équitable dans le dataset, ce qui nous permet de minimiser les biais potentiels et d'assurer une évaluation juste et équilibrée des performances de nos modèles. La préparation du dataset a impliqué plusieurs étapes :

Segmentation des Utterances : Chaque enregistrement audio a été segmenté en utterances individuelles, correspondant aux différents labels de beatbox. Cette segmentation permet de créer un ensemble de données structuré et facilement exploitable par les modèles car ils font maintenant moins d'une seconde et n'ont qu'un seul label. Nous avons ainsi également résolu le problème des enregistrements "impro" qui contenaient tous les labels.

Nettoyage des Données : Après l'examen du nouveau dataset créée, nous nous sommes rendu compte de deux coquilles. Une fois où le label a été laissé vide, et une fois où le label était 'pm', qui ne correspond à aucun label.

Division en Ensembles d'Entraînement, de Développement et de Test : Le dataset a été divisé en ensembles d'entraînement, de développement et de test, en veillant à ce que chaque ensemble contienne une représentation équilibrée des quatre labels. Cette division nous permettra d'entraîner nos modèles avec le jeu de train, de la validation ainsi que de test.

CNN

Training

Nous avons choisi l'accuracy comme métrique pour tester le modèle car les classes sont équilibrées et car nous savons que le modèle ne doit prévoir qu'une seule classe pour chaque exemple. Les résultats se sont montrés très bons assez rapidement, à l'aide des hyperparamètres suivants :

-epochs = 20
-batch_size = 16
-nb_filter_conv1 = 32
-nb_filter_conv2 = 64
-kernel_size = (3,3)
-dropout_conv = 0.25
-dropout_dense = 0.5
-optimize r= adam

Une fois les 20 epochs passées, l'accuracy était déjà au-delà de 0.95.

Recherche d'hyper paramètres

Afin d'améliorer nos résultats, nous décidons d'optimiser 6 hyperparamètres : les nombres de filtres des deux convolutions, la taille du kernel, la taille et densité du dropout, ainsi que l'optimiseur. Une fois l'apprentissage terminé, nous avons pu constater que la taille du kernel était l'hyperparamètre le plus décisif sur l'accuracy du dev set. Cependant, en entraînant

à nouveau le modèle avec ces nouveaux hyperparamètres, l'accuracy n'a pas réellement augmenté. En effet, le modèle apprend déjà très bien et très vite.

Tester avec d'autres donnés

Nous avons testé le modèle avec des données de beatbox faites par deux participants (un amateur et un confirmé) et le modèle atteint 97% d' accuracy. Ainsi le modèle généralise bien et il a appris à reconnaître les caractéristiques linguistiques (lieu d'articulation, présence d'air pulmonaire) pour distinguer le kick, le snare, le opened hi-hat et le closed hi-hat.

Wav2Vec

Preprocessing

Pour préparer les données audio en vue de leur utilisation dans notre modèle de classification, nous devons configurer les ressources Wav2Vec2 pertinentes en fonction de notre cas d'usage. Étant donné que nous classifions des sons de beatbox et non une langue parlée, nous avons choisi d'utiliser le modèle facebook/wav2vec2-base.

Afin de gérer les représentations contextuelles pour des longueurs d'audio variables, nous avons adopté une stratégie de fusion (mode de pooling) pour concaténer les représentations 3D en représentations 2D. Nous avons opté pour le mode de pooling "mean" (moyenne), qui permet de réduire la dimensionnalité des représentations tout en conservant les informations essentielles. En outre, nous avons fixé la fréquence d'échantillonnage cible à 16.000 Hz. Cette fréquence d'échantillonnage est couramment utilisée dans les applications de traitement de l'audio et offre un bon compromis entre la qualité de l'audio et la complexité computationnelle.

Training

Pour garantir des résultats optimaux, nous avons effectué une recherche aléatoire (random search) afin de déterminer les meilleurs hyperparamètres pour notre modèle. Cette méthode nous a permis d'explorer un large éventail de configurations et de sélectionner celles qui offrent les meilleures performances. Nous avons entraîné notre modèle sur 4 essais de random search. L'espace de variation des hyperparamètres sont les suivants :

- Learning rate : [1e-5 ; 5e-5]
- La taille du batch : [4, 8, 16]

Les hyperparamètres qui nous ont donné les meilleurs résultats sont les suivants :

Lr = 1.16e-05

Taille du batch : 8

Le modèle est entraîné sur 10 epoch et atteint 0.89 d'accuracy

Epoch	Training Loss	Validation Loss	Accuracy
1	1.717600	0.800727	0.645897
2	1.349800	0.613275	0.773556
3	1.102600	0.546252	0.807497
4	0.973400	0.504766	0.826241
5	0.691500	0.495098	0.816109
6	0.651300	0.454537	0.848531
7	0.495000	0.425347	0.864742
8	0.644100	0.411594	0.870821
9	0.747300	0.400082	0.881459

Conclusion

Notre projet démontre qu'il est possible de réaliser un modèle de classification de percussions de beatbox à l'aide de deux architectures différentes. Ces deux expériences ont montré des résultats très satisfaisants, nécessitant assez peu de données d'apprentissage. D'autres expériences sont également réalisables dans le domaine de l'ASR et du beatbox, par exemple comparer un modèle entraîné sur de la parole et un modèle entraîné sur des données musicales. De plus, il peut être intéressant d'utiliser l'ASR pour le beatbox en dehors de la classification de sons, comme l'analyse des sons et la recherche de features.

Références

Delgado, A. (2019). Amateur Vocal Percussion Dataset [Base de données]. Dans *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.3245959

Jana, S. (2023, November 2). A Deep Learning-Based Beatbox sound recognition method. *Medium*. https://medium.com/@subjana/a-deep-learning-based-beatbox-sound-recognition-method-1635f6ae396d