Réseaux de neurones pour la détection d'évènements sonores

Maël Alet

Décembre 2021

Table des matières

1 Objectif

Les réseaux de neurones faits pour la détection d'évènements sonores (« Sound Event Detection » ou « SED ») peuvent être utilisés pour détecter des évènements comme des clics de souris et des frappes au clavier.

2 Fonctionnement

On utilise un réseau de neurones déjà entraîné qui prend en entrée un enregistrement audio et donne en sortie une probabilité pour chaque évènement que le réseau est capable de détecter (comme « clic de souris », « frappe au clavier », etc...). Si la probabilité pour un évènement est haute le réseau est plutôt sûr que l'évènement est présent dans l'enregistrement. Si la probabilité est faible, le réseau est au contraire plutôt sûr que l'évènement n'y est pas.

Dans notre cas, si un évènement comme « clic » ou « frappe au clavier » est détecté au-delà d'une certaine probabilité alors on considère l'enregistrement comme étant défectueux.

2.1 Avantages

L'avantage est que si le réseau est vraiment performant, un seul réseau peut permettre de détecter tous les défauts présents et dans un enregistrement donné dire quels défauts sont présents ou pas. Certains réseaux sont également capables d'indiquer à quel endroit de l'enregistrement le défaut est présent, permettant de savoir plus rapidement s'il y a réellement un défaut ou s'il s'agit d'un faux positif.

2.2 Inconvénients

L'inconvénient est que la plupart des réseaux de neurones sont plutôt lourds en termes de puissance de calcul et de RAM nécessaire.

3 Performance

On teste les performances sur le réseau AST (Audio Spectrogram Transformer) qui est un des meilleurs réseaux disponibles en termes de performance sur le jeu de donnée AudioSet.

3.1 Temps de traitement

En traitant les fichiers un à un, on obtient un temps de traitement moyen de 6,22 s/fichier avec des fichiers extraits de Lingua Libre ayant une durée moyenne de 1,779 s.

En utilisant les mêmes fichiers, si l'on traite plusieurs fichiers simultanément on observe qu'à mesure que l'on augmente le nombre de fichiers traités

simultanément, le temps de traitement par fichier diminue. En traitant 4 fichiers simultanément, on obtient un temps de traitement moyen de 5,33 s/fichier. Soit une augmentation de performance de 13-14% mais le temps de traitement par fichier reste tout de même très long.

Il n'est pas possible de traiter plus de 4 fichiers simultanément pour les raisons invoquées ci-dessous dans la partie Mémoire RAM.

3.2 Mémoire RAM

De manière générale, l'utilisation de RAM est d'environ 3 Gio. Cependant, l'utilisation de la mémoire RAM est étrangement très variable.

En traitant un seul fichier à la fois, on ne dépasse vraisemblablement jamais les 3 Gio d'utilisation de RAM. En revanche en traitant 4 fichiers à la fois, si la plupart du temps on reste également en dessous de 3 Gio, on observe parfois lors de certains lancements (toujours sur les mêmes fichiers) une utilisation de RAM jusqu'à 6 Gio.

À partir de 5 fichiers, la quantité de RAM nécessaire pour traiter les fichiers augmente très brutalement de manière inexpliquée ($>$12 Gio pour 5 fichiers contre généralement $<$3 Gio pour 4 fichiers). Ce phénomène empêche de traiter plus de 4 fichiers en même temps et donc de réduire le temps de traitement par fichier.

3.3 Autres réseaux

Il est peut-être possible d'obtenir de meilleure performance en termes de temps de traitement et d'utilisation de la RAM en utilisant d'autres réseaux. En effet, les auteurs de l'architecture PaSST prétendent obtenir de meilleurs résultats que AST tout en réduisant de manière significative la puissance nécessaire pour exécuter le réseau par rapport à AST.

4 Précision

Quelques tests ont été effectués sur la détection de clics de souris. Les résultats sont décevants, la probabilité assignée à la classe « clic » sur des enregistrements avec des clics très facilement audibles est très faible. Au contraire la probabilité assignée à certains enregistrements sans aucun clic est parfois relativement élevée. De plus, il ne semble pas y avoir assez d'écart entre la probabilité moyenne affectée aux enregistrements sans clic et la probabilité moyenne affectée aux enregistrements avec clics pour permettre de différencier les deux groupes.

En testant également un fichier contenant un léger grésillement, le réseau AST ne détecte aucun label avec une probabilité $>$10% autre que le label « Speech ». Les labels détectés avec une probabilité $>$1% ne sont pas présents dans l'enregistrement et ne représentent pas des catégories de bruits qu'il serait intéressant de détecter.

5 Conclusion

Cette méthode n'a pas été retenue car elle nous paraissait trop gourmande en ressources et ne semble également pas capable de détecter le genre de défauts que l'on recherche sur les enregistrements de Lingua Libre.