**Rapport sur l’annotation et entraînement**

**Projet** : Détection automatique du passage des publicités dans des fichiers audios

**Introduction**

L'automatisation de l'annotation et de la segmentation des données audio est une étape cruciale pour gérer efficacement de grands volumes de données, en particulier dans des domaines comme la détection de publicités dans des enregistrements audio. Ce rapport décrit les étapes nécessaires pour automatiser ce processus en utilisant des modèles de réseaux de neurones tels que les Transformers et les CNN (Convolutional Neural Networks).

**Préparation des Données**

La préparation des données est une phase essentielle qui conditionne la qualité de l'entraînement des modèles.

**Collecte des Segments**

La première étape consiste à diviser les fichiers audios en segments plus petits. Ces segments permettent d'analyser les portions contenant des publicités et celles qui n'en contiennent pas. Cette division est importante car elle permet de mieux gérer les variations dans la durée des publicités et facilite leur traitement par les modèles d'apprentissage automatique. Pour assurer une analyse robuste, il est recommandé de disposer d'environ 1000 à 5000 segments audio. Ce volume doit inclure une proportion équilibrée de segments contenant des publicités et de segments sans publicités.

**Annotation Manuelle**

Après la collecte des segments, une étape d'annotation manuelle est nécessaire pour établir une base de données de référence. Cela implique d'étiqueter un sous-ensemble de segments audio, en les classifiant comme "publicité" ou "non-publicité". Cette annotation manuelle est essentielle car elle sert de base pour entraîner et évaluer les modèles de machine learning. Il est conseillé d'étiqueter entre 100 et 200 segments pour créer un ensemble d'exemples diversifié.

**Entraînement du Modèle**

Une fois les données préparées, l'étape suivante est l'entraînement des modèles. Deux types de modèles sont particulièrement recommandés pour ce projet : les Transformers et les CNN.

**Utilisation des Modèles Pré-Entraînés**

Pour les Transformers, il est recommandé d'utiliser le modèle facebook/wav2vec2-base-960h de Hugging Face. Ce modèle est conçu pour la reconnaissance vocale, mais il peut être fine-tuné pour la classification audio, et notamment pour la détection des publicités. Quant aux CNN, des modèles comme **YamNet** ou **VGGish** sont appropriés pour la classification audio. Vous pouvez également créer une architecture CNN personnalisée si nécessaire pour répondre aux besoins spécifiques de ce projet.

**Quantité de Données pour l’Entraînement**

Pour un fine-tuning efficace du modèle Transformer, un minimum de 1000 à 3000 segments audio est recommandé. Plus le nombre de segments est élevé, meilleures seront les performances du modèle. Pour le modèle CNN, la même quantité de données est recommandée, avec la possibilité d’augmenter ce nombre pour améliorer la précision et la robustesse du modèle.

**Segmentation et Prédiction**

Une fois le modèle entraîné, il peut être utilisé pour la segmentation et la prédiction sur de nouveaux fichiers audio.

**Segmentation de l’Audio**

La segmentation est réalisée en appliquant une technique avec une fenêtre glissante pour diviser les fichiers audios longs en segments plus courts. Cette méthode facilite l'analyse détaillée de chaque segment et améliore la précision des prédictions, permettant ainsi au modèle de mieux identifier les segments contenant des publicités.

**Analyse des Résultats**

Après la segmentation, le modèle entraîné est utilisé pour prédire les étiquettes des segments audio. Ces résultats permettent de classifier automatiquement les segments comme publicités ou non-publicités. Pour évaluer les performances du modèle, il est recommandé de tester le modèle sur un ensemble de validation représentant environ 20% des données annotées. Cette évaluation permet d'ajuster les paramètres du modèle si nécessaire pour obtenir des résultats optimaux.

**Exigences Matérielles et Capacités**

Les exigences matérielles varient selon le type de modèle utilisé.

**Transformer (Wav2Vec2)**

Les modèles Transformers, comme Wav2Vec2, sont très performants pour les tâches de reconnaissance vocale et de classification audio, mais nécessitent une puissance de calcul significative. Un GPU performant, tel qu'un **NVIDIA RTX 3090** ou **A100**, est recommandé pour un entraînement efficace. En termes de capacité, un minimum de 16 Go de mémoire GPU est nécessaire, accompagné de 32 Go ou plus de RAM, ainsi qu’un espace de stockage de 500 Go à 1 To pour gérer les ensembles de données volumineux.

**CNN (Convolutional Neural Network)**

Les modèles CNN sont plus légers en termes de ressources comparés aux Transformers. Ils peuvent être entraînés sur des machines avec des ressources moins élevées, bien qu'un GPU soit recommandé pour accélérer l'entraînement. Pour ce type de modèle, 8 Go de mémoire GPU et 16 Go ou plus de RAM suffisent. Un espace de stockage de 250 Go à 500 Go est également recommandé, en fonction de la taille des données audio.

**Conclusion**

L'automatisation de l'annotation et de la segmentation des données audio est réalisable en utilisant des modèles de deep learning tels que Wav2Vec2 et CNN. Chaque modèle présente des avantages en fonction des ressources disponibles et des exigences spécifiques du projet. Les Transformers nécessitent des machines plus puissantes mais offrent une précision élevée, tandis que les CNN sont plus accessibles en termes de ressources tout en offrant de bonnes performances. Ce rapport fournit les bases pour choisir et mettre en œuvre la meilleure approche en fonction des contraintes matérielles et des besoins du projet.