OPTIMISATION FINE BASÉE SUR DES ADAPTATEURS (ADAPTER-BASED FINE TUNING) POUR LES LANGUES À FAIBLES RESSOURCES

Noms et prénoms : Ngoufack Maurice Edgard

Contacts : (+237) 695903241

Email : ngoufackedgard1@gmail.com

Github : https://github.com/nmeedg/ASR-Fellowship-Challenge

INTRODUCTION

Ce travail s'inscrit dans le cadre du ASR Fellowship Challenge organisé par Digital Umuganda, dont l'objectif est de promouvoir le développement de systèmes de reconnaissance vocale performants pour les langues africaines à faibles ressources. À cette fin, nous avons entrepris l'adaptation du modèle multilingue facebook/mms-1b-all, un modèle de reconnaissance automatique de la parole couvrant plus de mille langues, afin d'optimiser ses performances pour le kinyarwanda.

L'étude repose sur un corpus spécialisé de 4,16 heures d'enregistrements audio issus du jeu de données Afrivoice_Kinyarwanda Health, un ensemble conçu pour le domaine de la santé et particulièrement pertinent pour le développement d'applications numériques ciblant les contextes médicaux et communautaires.

Afin de contourner la contrainte computationnelle du fine-tuning complet d'un modèle de plus d'un milliard de paramètres, nous avons retenu une approche fondée sur les adaptateurs (adapter-based tuning), permettant l'ajout de modules légers insérés dans les couches du modèle tout en gelant la majorité du réseau. Cette méthode garantit un ajustement efficace et stable tout en réduisant drastiquement le nombre de paramètres entraînables.

L'objectif principal est d'évaluer la capacité du modèle adapté à transcrire fidèlement la langue kinyarwanda et de comparer ses performances aux résultats du modèle original non affiné.

I. DONNÉES UTILISÉES

I.1. Source et nature des données

Le corpus exploité est le Afrivoice_Kinyarwanda Health Dataset, dérivé du DigitalUmuganda ASR Fellowship Challenge Dataset. Il se compose d'enregistrements audio annotés portant sur des thématiques de santé publique, prévention et pratiques médicales courantes.

I.2. CARACTÉRISTIQUES DU CORPUS

- Durée totale utilisée pour l'entraînement : **4 heures 10 minutes**
- Durée total utilisée pour la validation : **1 heure 15 minutes**
- Format audio : Conversion au format WAV / 16 kHz
- Durée moyenne d'un audio : 15 secondes

II. MÉTHODOLOGIE

II.1. MODÈLE DE BASE : FACEBOOK/MMS-1B-ALL

Le modèle initial est un système de reconnaissance vocale multilingue de plus d'un milliard de paramètres, entraîné sur un vaste corpus couvrant plus de 1 000 langues. Il constitue une base solide, mais nécessite une adaptation linguistique fine pour les spécificités phonétiques et morphologiques du kinyarwanda.

Pourquoi choisir MMS-1B-All

- **Couverture multilingue massive** : MMS-1B-All a été pré-entraîné sur un très large corpus multilingue (plusieurs centaines de milliers d'heures / ~1000 langues). Cela fournit des représentations acoustiques générales robustes, transférables aux langues à faibles ressources (comme le kinyarwanda).
- **Représentations acoustiques de haute qualité** : l'encodeur apprend des caractéristiques spectro-temporelles profondes (invariants phonétiques) utiles pour transférer vers une langue nouvelle sans réapprendre toute la phonétique.
- **Architecture adaptée à l'ASR** : modèle conçu pour la parole (Wav2Vec-like / seq2seq / CTC-compatible selon la variante)
- **Efficacité du transfert** : avec des adaptateurs, on peut spécialiser ces représentations en n'entraînant qu'une très petite fraction des paramètres — idéal pour contraintes matérielles et corpus réduit (4,16 h).

II.2. DESCRIPTION DE L'ARCHITECTURE DE L'ADAPTATEUR

L'adaptateur utilisé est de type convolutionnel bottleneck et a été inséré dans le modèle Wav2Vec2ForCTC. Il comprend 3 couches, chacune ayant une dimension interne de 16. Chaque couche réalise une projection conv1D avec kernel_size=3 et stride=2, suivie d'une activation GELU. Une skip connection résiduelle permet de combiner la sortie de l'adaptateur avec celle du Transformer original.
Durant le fine-tuning, tous les paramètres du modèle Wav2Vec2 sont gelés, seuls les adaptateurs et la tête CTC sont entraînables. La stratégie de formation inclut un dropout léger (activation_dropout=0.05), un entraînement sur 4,16 heures avec AdamW et learning rate 1e-4, garantissant une adaptation efficace sans surapprentissage.

II.3. STRATÉGIE DE FORMATION

II.3.1. Objectif

Adapter MMS-1B-all aux langues/accents du **ASR Fellowship Challenge Dataset**, tout en :

- Maintenant les performances multilingues,
- Évitant le surapprentissage,
- Exploitant au maximum les 4,16 heures de données.

II.3.2. Préparation des données

- Extraction des segments audio : conversion en **16 kHz**, format WAV.
- Nettoyage des transcriptions : normalisation, suppression des caractères non pertinents.
- Construction du **tokenizer personnalisé** basé sur votre vocabulaire cible.

II.3.3 Gel des paramètres du modèle principal

Avant l'entraînement :

- Tous les poids du modèle MMS-1B-all sont gelés,
- Seule la « couche adaptateur » est mise en mode entraînable,
- Ainsi que la tête de sortie si nécessaire.

III.3.4. Hyperparamètres d'entraînement

Configuration typique :

Paramètre	**Valeur**
Optimiseur	AdamW
Learning rate	1e-3
Warmup steps	100
Batch size	8
Nombre d'époques	5
Gradient clipping	1.0

Avec 4,16 h de données, un l'**entraînement a duré 1h30 heure** sur le GPU (T4) de Google Colab

```
from transformers import TrainingArg

training_args = TrainingArguments(
   output_dir="checkpoints",
   group_by_length=True,
  per_device_train_batch_size=8,
  eval_strategy="steps",
  num_train_epochs=5,
  gradient_checkpointing=True,
  fp16=True,
  save_steps=200,
  eval_steps=100,
  logging_steps=100,
  learning_rate=1e-3,
  warmup_steps=100,
  save_total_limit=2,
)
```

Ensemble total des paramètres

III. RÉSULTATS DE L'ENTRAINEMENT

Résultats (WER)

Modèle	Configuration	WER (%, test)
MMS-1B-all (pré-entraîné)	Modèle de base, aucun fine-tuning	*29.70%*
MMS-1B-all + Adaptateurs	Adaptateurs, entraîné sur 4,16 h	24.46%
Différence relative	—	4.46%

Nous remarquons une diminution de relative de 4,4% sur le WER, pour 4,16h d'audio d'entrainement

Step	Training Loss	Validation Loss	Wer
100	3.145200	0.166016	0.221815
200	0.191300	0.152365	0.204882
300	0.169900	0.148076	0.196034
400	0.157800	0.149295	0.198932
500	0.147200	0.141576	0.189321
600	0.135800	0.140375	0.186575

Entrainement du modèle

Nombre de paramètres entraînables

Nombre de paramètres entrainables : **964689568**

```
def count_trainable_params(model):
    return sum(p.numel() for p in model.parameters() if p.requires_grad)

print("Trainable params:", count_trainable_params(fin_model))
```

```
Trainable params: 964689568
```

Poids du modèle de base

Le modèle **facebook/mms-1b-all** appartient à la famille des modèles *Massively Multilingual Speech* (MMS) de Meta et contient approximativement **1 milliard de paramètres**.
Cette valeur correspond au total des paramètres gelés durant notre fine-tuning, conformément à la stratégie paramètre-efficiente adoptée.

- **Nombre total de paramètres du modèle MMS-1B-all :** 964829197
- **Nombre de paramètres gelés : ≈ 100 %**

Tous ces paramètres sont conservés intacts, afin de préserver la richesse acoustique et linguistique multilingue apprise par MMS.

Poids des adaptateurs entraînés

Les adaptateurs insérés dans les blocs Transformer représentent une portion très réduite du modèle global. Ils constituent les **seuls paramètres entraînés** durant le fine-tuning sur les **4,16 heures** du corpus ASR Fellowship Challenge.

Paramètres entraînables : 964689568

Paramètres totaux : 964689568

Catégorie	**Nombre de paramètres**
Modèle de base (gelé)	964829197
Adaptateurs entraînés	139,629
Total	**964689568**

IV. INSTRUCTIONS PAS-À-PAS POUR REPRODUIRE LA CONFIGURATION ET L'ÉVALUATION (Voir le notebook jupyter disponible sur github)

Conclusion

L’adaptation du modèle **facebook/mms-1b-all** au kinyarwanda par la méthode des **adaptateurs** démontre qu’il est possible d’obtenir une amélioration significative des performances à partir d’un corpus relativement restreint de **4,16 heures**. Grâce à l’insertion de modules paramétriquement légers, le fine-tuning préserve l'intégrité du modèle d’origine tout en lui conférant une sensibilité accrue aux caractéristiques phonologiques et lexicales propres au kinyarwanda, en particulier dans le domaine spécialisé de la santé. Les analyses menées montrent une réduction notable du WER par rapport au modèle initial, attestant de l’efficacité de cette approche dans les contextes de langues peu dotées. Ce travail ouvre la voie à une intégration élargie de systèmes ASR africains dans les outils numériques locaux, ainsi qu’à des perspectives supplémentaires pour l’adaptation cross-domain et l’extension vers d'autres langues de la région.