Rapport de Projet : Classification de News via Mixture of Experts (MoE)

Votre Nom

15 janvier 2026

1 Introduction et Revue du Concept de MoE

L'objectif de ce projet est d'explorer l'architecture *Mixture of Experts* (MoE) pour la classification de textes. Les modèles MoE reposent sur le principe de "diviser pour régner" : au lieu d'utiliser un seul réseau dense, on utilise plusieurs sous-réseaux spécialisés (**Experts**) et un mécanisme de routage (**Gating Network**) qui décide quel expert doit traiter quel échantillon.

2 Description du Modèle

2.1 Architecture

Notre modèle, `MoEBertModel`, utilise **BERT** (*bert-base-uncased*) comme extracteur de caractéristiques (backbone). La sortie du *pooler* de BERT alimente :

— Un réseau de routage (couche linéaire) qui génère des poids pour chaque expert.
— Une liste de N experts (`TextExpert`), chacun étant un réseau de neurones *feed-forward* avec normalisation et dropout.

2.2 Choix de Conception

— **Soft Routing** : Nous utilisons un routage "doux" via une fonction Softmax pour permettre un apprentissage stable de tous les experts simultanément.
— **Exploration** : Ajout d'un bruit gaussien sur les scores du router durant l'entraînement pour éviter l'effondrement précoce sur un seul expert.

3 Protocole Expérimental et Hyperparamètres

L'entraînement a été effectué sur le dataset AG News.

Hyperparamètre	Valeur
Batch Size	64
Nombre d'Époques	5
Learning Rate	3×10^{-5}
Nombre d'Experts	8
Architecture Backbone	BERT Base

Table 1 – Configuration des hyperparamètres

4 Résultats Expérimentaux et Interprétation

Le modèle atteint une précision globale de **87 %**.

— **Performance par classe** : La classe 2 montre la meilleure performance avec un F1-score de 0.95.

— **Usage des experts** : L'analyse thermique (heatmap) montre une spécialisation des experts. Certains experts traitent des caractéristiques transversales tandis que d'autres se spécialisent sur des thématiques précises.

5 Discussion Critique

5.1 Forces

Capacité de spécialisation et maintien d'une haute précision sur des classes complexes grâce à la modularité des experts.

5.2 Limites

Risque d'**Expert Collapse** (effondrement) où seuls quelques experts sont activés, rendant le reste de la capacité du modèle inutile si le router n'est pas bien régularisé.

5.3 Perspectives

Utiliser des fonctions de perte de *Load Balancing* plus agressives ou explorer le **Top-k routing** pour réduire le coût computationnel à l'inférence.