Projet de fin de module – Mixture of Experts (MoE)

Master 2 Informatique – Parcours Intelligence Artificielle
Cours : *Architectures de réseaux de neurones avancées*

Contexte

Les *Mixture of Experts* (MoE) constituent une approche moderne pour augmenter la capacité des modèles neuronaux tout en limitant leur coût d'inférence. L'idée est de combiner plusieurs sous-modèles (*experts*) et un réseau de *gating* qui décide dynamiquement quels experts doivent être activés pour chaque entrée.

Cette approche, dite de *computation conditionnelle*, permet de n'activer qu'une partie du modèle à chaque passage. Elle est à la base de modèles récents tels que **Switch Transformer** (Google, 2021), **GLaM** (Google Research, 2022) ou **Mixtral** (Mistral, 2024).

L'objectif du projet est d'implémenter, d'entraîner et d'analyser un modèle de type Mixture of Experts.

Objectifs pédagogiques

À l'issue de ce projet, vous devez être capables de :

- Comprendre le principe de la *conditional computation* et du *routing*.
- Concevoir et implémenter un modèle modulaire combinant plusieurs experts et un réseau de gating.
- Expérimenter sur un jeu de données réel et comparer les performances avec un modèle dense.
- Analyser la spécialisation des experts et l'effet du gating sur la performance et la généralisation.

Spécifications minimales

1. Structure du modèle

Le modèle doit comporter :

- Un ensemble d'experts $E_1, E_2, \ldots, E_N$ (par exemple, des MLPs ou petits CNNs).
- Un réseau de *gating* $g(x)$ produisant une distribution sur les experts.
- Une combinaison des sorties selon :

$$y = \sum_i g_i(x)\, E_i(x)$$

  pour un routage **soft**, ou une sélection discrète des meilleurs experts pour un routage **hard**.

Vous êtes libres de choisir le nombre d'experts, la taille des couches et la stratégie de routage.

2. Données et tâche

Choisissez une tâche adaptée à votre architecture :

Domaine	Dataset suggéré	Tâche
Vision	MNIST / FashionMNIST / CIFAR-10	Classification
NLP (optionnel)	IMDB / AG News	Analyse de sentiment / Catégorisation
Tabulaire	Jeux UCI (Wine, Iris, etc.)	Régression / Classification

3. Expérimentations attendues

- Comparer votre MoE à un réseau dense de capacité équivalente.
- Étudier l'impact :
  - du nombre d'experts ;
  - du type de gating (soft vs hard) ;
  - de la régularisation (dropout, entropie du gating, etc.).
- Visualiser :
  - la distribution d'activation des experts ;
  - l'évolution des pertes de chaque expert ;
  - la répartition des données traitées par chaque expert.

Livrables attendus

Code

Organisation suggérée :

```
project_moe/
|
|-- moe_model.py         # définition du modèle (experts + gating)
|-- train.py             # script d'entraînement principal
|-- experiments.ipynb    # analyses et visualisations
|-- report.pdf           # rapport synthétique
```

Rapport (4–6 pages)

Le rapport devra comporter :

- une introduction et une brève revue du concept de MoE ;
- une description de votre modèle (architecture, choix de conception) ;
- le protocole expérimental et les hyperparamètres utilisés ;
- les résultats expérimentaux et leur interprétation ;
- une discussion critique (forces, limites, perspectives).

Critères d'évaluation

Critère	Description	Poids
Implémentation correcte	Code fonctionnel, modulaire et respectant le principe du MoE	30%
Rigueur expérimentale	Comparaisons, reproductibilité, clarté du protocole	25%
Analyse critique	Interprétation des résultats, visualisation du gating	20%
Originalité	Extensions, variantes ou analyses supplémentaires	15%
Présentation	Clarté du rapport, qualité des figures et rédaction	10%

Pistes d'extension (optionnelles)

— Routage top-k avec **Gumbel-Softmax** ou **Straight-Through Estimator**.
— Régularisation de sparsité sur le gating.
— Visualisation du routage par expert (heatmap expert $\leftrightarrow$ classes).
— Insertion d'une couche MoE dans un petit Transformer.

Références conseillées

— Jacobs et al., *Adaptive Mixtures of Local Experts*, Neural Computation, 1991.
— Shazeer et al., *Outrageously Large Neural Networks : The Sparsely-Gated Mixture-of-Experts Layer*, ICLR 2017.
— Fedus et al., *Switch Transformers : Scaling to Trillion Parameter Models*, arXiv :2101.03961.
— Touvron et al., *Mixtral of Experts*, 2024.

Organisation

— Travail en binôme recommandé.
— Durée estimée : 3 à 4 semaines.
— Rendu : code + rapport + mini-présentation orale (10 minutes).

Bonne chance et bon travail !