Modèles de Langage de Grande Taille (LLMs) et Applications Avancées

TP : Pré-entraînement et Fine-Tuning d'un Modèle de Langage

Objectifs

1. Comprendre les bases du pré-entraînement d'un modèle simple.
2. Réaliser le fine-tuning d'un modèle pré-entraîné (BERT) pour une tâche de classification de texte.
3. Manipuler des données textuelles réelles et entraîner un modèle sur celles-ci.

Partie 1 : Pré-entraînement (Simple Exemple)

Introduction

Simuler un **pré-entraînement simplifié** en entraînant un petit modèle pour prédire des mots masqués dans des phrases.

1. **Créer un dataset de texte**
   - Fournir un corpus simple (ou en générer un) contenant des phrases courtes.
   - Exemple de corpus :

     ```
     Le chat dort sur le tapis.
     Le chien joue dans le jardin.
     La voiture est rouge.
     ```

2. **Prétraiter les données**
   - **Tokenisation** : Découper les phrases en mots.
   - **Masquage de mots** : Masquer un mot aléatoire dans chaque phrase pour le prédire.
     - Exemple : "Le chat dort [MASK] le tapis."
3. **Créer un modèle simple** : Construisez un modèle capable de prédire un mot masqué dans une phrase.

   1. **Ajouter une couche d'embedding** (utilisez `nn.Embedding`) : Transformez chaque mot en un vecteur numérique.
   2. **Ajouter une couche dense** (utilisez `nn.Linear`) : Trouvez des relations complexes entre les mots.
   3. **Ajouter une softmax** (utilisez `nn.Softmax`) : Transformez les sorties en probabilités pour prédire le mot masqué.

   Structure minimale du modèle :

   Entrée : Indices des mots de la phrase.
   Sortie : Probabilités pour chaque mot du vocabulaire.

4. **Entraîner le modèle**
   - Tâche : Prédire les mots masqués à partir des contextes.
   - Fonction de perte : Cross-Entropy Loss.
   - $d_{Embedding} = 10$