Sujet de Projet : Fine-tuning de Modèles de Question-Answering sur le SQuAD Dataset avec Déploiement Web

Hanane Azzag / Bilal Faye

B.U.T 3 Apprentissage : 2023-2024

Introduction :

Le Question-Answering (QA) est une tâche clé en traitement automatique du langage naturel (NLP) où un modèle doit répondre à une question posée en se basant sur un contexte donné. Dans ce projet, nous proposons de fine-tuner différents modèles de QA sur le célèbre jeu de données Stanford Question Answering Dataset (SQuAD). Nous explorerons plusieurs architectures de modèles pré-entraînés, et comparerons leurs performances.

Objectifs du Projet :

1. Fine-tuner au moins trois modèles de QA parmi une sélection de modèles pré-entraînés (voir la liste des modèles dans la section lien).

2. Justifier et comprendre l'architecture des modèles sélectionnés et les différences entre eux.

3. Comparer les performances des modèles fine-tunés sur des métriques d'évaluation standard telles que F1-score, Exact Match (EM), Precision, Recall, AUC, ROC curve et temps d'inférence etc.

4. Développer une interface utilisateur conviviale permettant aux utilisateurs de poser des questions sur un contexte donné en utilisant le modèle fine-tuné.

5. Déployer l'application et le modèle fine-tuné sur Hugging Face Spaces pour un accès facile.

Détails du Projet :

1. Collecte et Prétraitement des Données : Télécharger le jeu de données SQuAD et prétraiter les données en extrayant les contextes, les questions et les réponses pour former des exemples d'entraînement.

2. Fine-tuning des Modèles : Utiliser la bibliothèque Hugging Face Transformers pour fine-tuner les modèles sélectionnés sur le jeu de données SQuAD. Vous pouvez utiliser tensorflow/keras ou pytorch.

3. Évaluation des Modèles : Évaluer les modèles fine-tunés en utilisant des métriques standard comme F1-score, Exact Match (EM), Accuracy, Precision, Recall, AUC, ROC curve et temps d'inférence etc sur des ensembles de validation choisis (comprendre les avantages et inconvénients de chaque métrique).

4. Développement de l'Interface Utilisateur : Développer une interface utilisateur interactive en utilisant **FastAPI** ou **Flask** pour le backend et **Streamlit** pour le frontend. Cette interface permettra aux utilisateurs de poser des questions sur un contexte donné en utilisant le modèle fine-tuné. L'utilisateur doit avoir la possibilité de charger un ou des fichiers comme contexte, ou écrire le contexte directement sur l'interface.

5. Déploiement : Déployer l'application et le modèle fine-tuné sur Hugging Face Spaces pour permettre un accès public.

Améliorations Possibles :

1. Exploration de différentes techniques de prétraitement de texte pour améliorer les performances des modèles fine-tunés.

2. Utilisation de techniques d'augmentation de données pour enrichir le jeu de données d'entraînement et améliorer la généralisation des modèles.

3. Intégration de techniques d'interprétabilité pour comprendre les raisonnements des modèles lors de la prédiction des réponses.

4. Ajout de fonctionnalités supplémentaires à l'interface utilisateur telles que la sauvegarde des résultats, la comparaison de différents modèles, etc.

Ce projet fournira une expérience pratique dans le fine-tuning de modèles NLP, l'évaluation de leur performance et le déploiement d'applications web interactives pour une utilisation publique.

Liens:

Jeu de données: SQuAD

FastAPI, Flask et Streamlit: full stack notebook

Exemple de code pour le fine-tuning: Hugging face example

Modèles: ALBERT, BART, BERT, BigBird, BigBird-Pegasus, BLOOM, CamemBERT, CANINE, ConvBERT, Data2VecText, DeBERTa, DeBERTa-v2, DistilBERT, ELECTRA, ERNIE, ErnieM, Falcon, FlauBERT, FNet, Funnel Transformer, OpenAI GPT-2, GPT Neo, GPT NeoX, GPT-J, I-BERT, LayoutLMv2, LayoutLMv3, LED, LiLT, LLaMA, Longformer, LUKE, LXMERT, MarkupLM, mBART, MEGA, Megatron-BERT, MobileBERT, MPNet, MPT, MRA, MT5, MVP, Nezha, Nyströmformer, OPT, QDQBert, Reformer, RemBERT, RoBERTa, RoBERTa-PreLayerNorm, RoCBert, RoFormer, Splinter, SqueezeBERT, T5, UMT5, XLM, XLM-RoBERTa, XLM-RoBERTa-XL, XLNet, X-MOD, YOSO