Question Answering avec Transformers

Projet de fouille de données et traitement automatique du langage

Exploration comparative de modèles Transformers (DistilBERT, BERT-base, RoBERTa-base) sur la tâche de Question Answering extractif avec le dataset SQuAD.

Problématique et objectifs du projet

01

Question Answering extractif

Mettre en œuvre un système capable d'extraire une réponse précise depuis un texte de contexte en réponse à une question donnée.

02

Fine-tuning de modèles Transformers

Adapter et entraîner plusieurs modèles Transformer pré-entraînés pour la tâche de question answering extractif.

03

Analyse comparative

Comparer les modèles en termes de performances (Exact Match, F1, Precision, Recall, AUC) et de temps d'inférence afin d'analyser les compromis entre précision et coût de calcul.

04

Interface utilisateur

Développer une interface web permettant de tester les modèles fine-tunés sur des contextes et questions définis par l'utilisateur.

Dataset SQuAD et prétraitement

Le dataset SQuAD

SQuAD (Stanford Question Answering Dataset) fournit des triplets composés d'un contexte textuel, d'une question et d'une réponse extraite du contexte sous forme de span.

Préprocessing des données

- Tokenisation des textes à l'aide du tokenizer associé au modèle Transformer
- Découpage des contextes trop longs en segments à l'aide d'une technique de stride
- Alignement des positions de début et de fin des réponses avec les tokens générés
- Conversion en tenseurs PyTorch pour l'entraînement

Modèles Transformers étudiés

DistilBERT

Spécificités : Version allégée de BERT obtenue par distillation

Avantages : Temps d'inférence plus faible et consommation mémoire réduite

Compromis : Performances légèrement inférieures à celles des modèles plus lourds

BERT-base

Spécificités : Architecture Transformer de référence (12 couches)

Avantages : Bon compromis entre performances et coût de calcul

Positionnement : Utilisé comme modèle de comparaison principal dans ce projet

RoBERTa-base

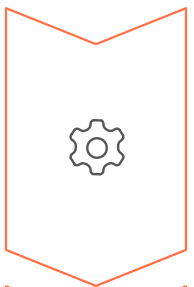
Spécificités : Architecture proche de BERT-base avec une stratégie de pré-entraînement différente

Avantages : Entraîné sur davantage de données et sans tâche Next Sentence Prediction

Résultat : Performances généralement supérieures à BERT-base au prix d'un temps de calcul plus élevé

Une architecture de pipeline identique garantit une comparaison équitable entre les trois modèles.

Entraînement et évaluation des modèles



Configuration

- Utilisation de l'API `Trainer` de Hugging Face pour assurer un entraînement reproductible
- Gestion des hyperparamètres, des logs et de la sauvegarde des modèles

Optimisation temps

- Entraînement réalisé sur un sous-ensemble du dataset SQuAD
- Objectif : réduire le temps de calcul tout en conservant une comparaison équitable entre les modèles

Métriques

- Exact Match (EM) et F1-score pour le question answering extractif
- Precision, Recall, ROC et AUC obtenues via une reformulation du problème en classification binaire
- Mesure du temps moyen d'inférence par question

Résultats principaux de l'analyse comparative

Résultats détaillés par modèle

BERT-base obtient les meilleures performances globales (EM = 16.4 %, F1 = 24 %), ce qui en fait le modèle de référence le plus solide.

DistilBERT est nettement plus rapide (≈16 ms vs 29 ms pour BERT), mais avec des performances plus faibles, illustrant le compromis vitesse / qualité.

RoBERTa-base obtient des performances très faibles sur ce sous-ensemble (EM = 0.2 %, F1 = 3.6 %), probablement en raison de la taille réduite du dataset et d'une convergence insuffisante.

Conclusion

Ces résultats montrent qu'un modèle plus complexe n'est pas nécessairement plus performant lorsque les données sont limitées.

Le choix du modèle dépend des contraintes spécifiques du projet :

- Pour les performances maximales : BERT-base
- Pour la rapidité d'inférence : DistilBERT
- RoBERTa-base nécessiterait plus de données d'entraînement pour exprimer son potentiel

Application et déploiement de la solution

Backend : FastAPI

- API REST permettant de charger les modèles fine-tunés
- Génération de réponses à partir d'un contexte et d'une question

Frontend : Streamlit

- Interface web permettant à l'utilisateur de saisir un contexte et une question
- Sélection du modèle de question answering à utiliser

Choix utilisateur

- Sélection dynamique entre DistilBERT, BERT-base et RoBERTa-base
- Comparaison directe du comportement des modèles sur un même exemple

Déploiement cloud

- Application déployée sur Hugging Face Spaces
- Accès public pour tester les modèles via l'interface

Conclusion et perspectives

Conclusion

Ce projet a permis de mettre en œuvre un pipeline complet de question answering extractif basé sur des modèles Transformers.

La comparaison entre DistilBERT, BERT-base et RoBERTa-base met en évidence un compromis clair entre performances et coût de calcul.

Les résultats obtenus sont cohérents avec les différences d'architecture entre les modèles étudiés.

Perspectives d'amélioration

- Entraîner les modèles sur l'intégralité du dataset SQuAD afin d'améliorer les performances
- Explorer des architectures plus récentes ou plus larges (BERT-large, DeBERTa)
- Étendre l'étude à d'autres jeux de données, comme SQuAD 2.0
- Adapter le pipeline pour le traitement de contextes plus longs