Projet 3A MMF/SDI

Pipeline MLOps pour l'entraînement d'un classifier de document destiné à l'entraînement de LLM

Entreprise/Laboratoire : ILLUIN Technology
Adresse : 20 Place de la Défense, 92800 Puteaux
Encadrant : Rubenach Théo
Coordonnées de l'encadrant : theo.rubenach@illuin.tech
Coordonnées d'un contact relation entreprise/ecole : gautier.viaud@illuin.tech

1- Contexte/Orientation :

ILLUIN Technology est une jeune scale-up française spécialisée dans les solutions d'intelligence artificielle (IA) et les architectures data. Nous mettons à profit les avantages de ces technologies de pointe pour moderniser les systèmes d'information de nos clients et proposer des solutions digitales qui répondent à de nouveaux besoins stratégiques ou métiers. Au sein de l'entreprise, le pôle Data / ML Engineering s'occupe de développer les architectures et outils d'industrialisation de nos solutions d'IA. Vous serez accompagnés par ce pôle lors de votre projet.

Travaillant sur des modèles de data science de plus en plus pointus et nécessitant de plus en plus de ressources de calculs (à l'instar des Large Language Model), nous avons besoin d'architectures / outils spécifiques afin de développer (collecte de données, processing de données, entrainement / évaluation de modèle de data science), packager, orchestrer, déployer et monitorer nos solution d'IA. Pour cela, nous développons notre stack technologique de MLOps afin d'articuler tous ces besoins automatiquement de bout en bout. En particulier,

des plateformes de MLOps open-source comme ZenML, Kedro ou encore Flyte ou Cloud comme AWS SageMaker ou GCP Vertex AI nous permettent de créer ces pipelines afin de consolider et standardiser nos solutions de data science de l'acquisition des données jusqu'au déploiement en production de nos modèles.

Les technologies Cloud de MLOps étant de plus en plus attractives par leur capacité de scaling notamment, que ce soit pour les projets de nos clients ou nos besoins internes, nous avons besoin de mettre en place des pipele de MLOps pour le developpement end-to-end de nos projets d'IA.

Ce projet s'inscrit dans ce contexte et a pour objectif de créer une pipeline complète de MLOps sur ASW SageMaker / GCP VectexAI afin d'entraîner un modèle de classification de document pour collecter des documents de haute qualité pour le "late training" de LLM en R&D.

En effet, de nos jours, les meilleurs LLMs sont entraînés en plusieurs phases; après une première phase d'entraînement sur des données internet variées, on cherche à finaliser le pré-entraînement lors d'une phase courte d'"annealing" où on entraîne le modèle sur des données de très haute qualité (à haute valeure éducative, scientifique, mathématique). Pour constituer ce dataset de haute qualité, il est essentiel de trouver des données adaptées parmi les milliards de documents présents sur internet. On peut notamment entraîner des petits modèles de type classifier, pour détecter du contenu de haute qualité éducative (ou du contenu thématique comme maths, science, etc) et après les utiliser pour filtrer de très gros corpus.

2- Résultats attendus

Les objectifs poursuivis sont les suivants:

- Entraîner un classifier de document (exemple : https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier) pour collecter des données de haute qualité (domaine à choisir) pour la dernière phase d'entraînement de LLM ("annealing"). Cet entraînement peut se basser sur une dataset de ce type par exemple : https://huggingface.co/datasets/HuggingFaceFW/fineweb
- Mettre en place une pipeline de MLOps de l'acquisition et du preprocessing des données jusqu'au déploiement du modèle dans une infrastructure cloud. Cela peut notamment inclure :
    - l'acquisition et le versioning des données,
    - le pre-processing des données / features,
    - l'entraînement et l'évaluation du modèle avec de l'experiment tracking,
    - le packaging / versioning du modèle,
    - le serving et le déploiement du modèle dans une infrastructure cloud,
    - l'orchestration de réentrainement / redéploiement du modèle

3- Compétences à l'oeuvre et approches

- Challenges techniques :
    - Entraînement d'un modèle de classification de document à l'état de l'art pour préparer l'étape finale d'entraînement de LLM
    - Déploiement d'une pipeline de MLOps sur une architecture Cloud
    - Industrialisation d'une solution de data science
- Environnement technologique :
    - Plateforme Cloud de MLOps (AWSSageMaker, GCP Vertex AI)
    - Data science (Python, Hugging face)
    - Experiment tracker au choix (mlflow, W&B, Neptune, …)
    - Technologies / briques de MLOps au choix:
        - Data versioning (DVC, lakeFS, …)
        - Experiment Tracker (mlflow, W&B, …)
        - Model packaging / versioning / registry (BentoML, …)
        - Model serving (BentoML, Triton, …)
        - Orchestration (Airflow, Dagster, …)
        - Monitoring (Prometheus / Grafana, …)
    - Cloud provider au choix (GCP AWS, …)
    - Solution de conteneurisation / déploiement au choix (Kubernetes, …)
    - Optionnel : framework open-source de MLOps (ZenML, Flyte, …)

4- Planning dans les grandes lignes

Le planning pourra être adapté selon les besoins du projet, mais devrait comporter les phases suivantes:

- Design et documentation du modèle (utilisation, évaluation, domaine de document à classifier, évaluation de la mesure de qualité des documents) et de la pipeline de MLOps Cloud (objectif et lancement de la pipeline mise en place)
- Modèle de classification de qualité (domaine à définir au début du projet) de document entraîné
- Déploiement de la pipeline de MLOps

- Création de pipelines Kedro pour standardiser et industrialiser le modèle développé

5- Bibliographie

La bibliographie à l'état de l'art sera à établir dans la première phase du projet.

6- Format du livrable à préciser

Le livrable prendra la forme du code développé durant le projet, ainsi qu'éventuellement un rapport succinct résumant les recommandations concernant les technologies utilisées lors du projet.