H2O LLM Studio

H2O LLM STUDIO	**A framework and no-code GUI for fine-tuning state-of-the-art large language models (0.3.0)**	
	Système d'exploitation :	Linux Ubuntu
	Développé par :	H2O.ai
Apache License, Version 2.0	Personne de contact :	katy.fokou@smals.be

Fonctionnalités

H2O LLM Studio est une plateforme no-code et open-source dédiée au réglage fin des *large language models* (LLM). H2O LLM Studio dispose pour cela d'une interface graphique facile d'utilisation où les hyperparamètres nécessaires à l'entrainement du modèle peuvent être sélectionnés. Comme modèles de départ (backbone), des modèles open-source tels que Llama2, Falcon, openLlama sont proposés en standard par l'application. Il est également possible d'utiliser les modèles disponibles sur la plateforme HuggingFace comme point de départ. La plateforme est principalement composée d'une interface de chargement et visualisation des sets de données ainsi que d'une interface d'entrainement qui permet la configuration de nombreux hyperparamètres relatifs à la tokenisation, l'architecture, la validation, la prédiction, etc. On dispose en outre des fonctionnalités suivantes :

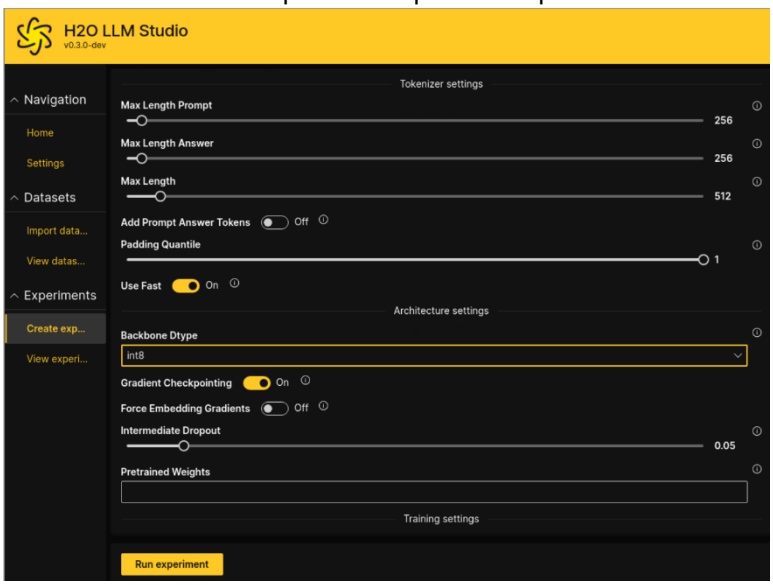
- L'intégration aux plateformes Neptune (outil de tracking) et HuggingFace (plateforme collaborative pour les modèles de machine learning). Les modèles entrainés peuvent être facilement publiés sur HuggingFace en un clique.
- L'évaluation et le monitoring de l'entrainement. Des métriques d'évaluation de modèles sont disponibles et une interface graphique permet de monitorer l'évolution des performances du modèle pendant le training. Cette interface permet également de comparer plusieurs expériences (instances d'entrainement de modèle) entre-elles.
- "Chat with the model". Pour chaque modèle entrainé, une interface chatbot est disponible dans l'application afin de tester directement les performances du modèle en mode conversationnel.

Conclusions & Recommandations

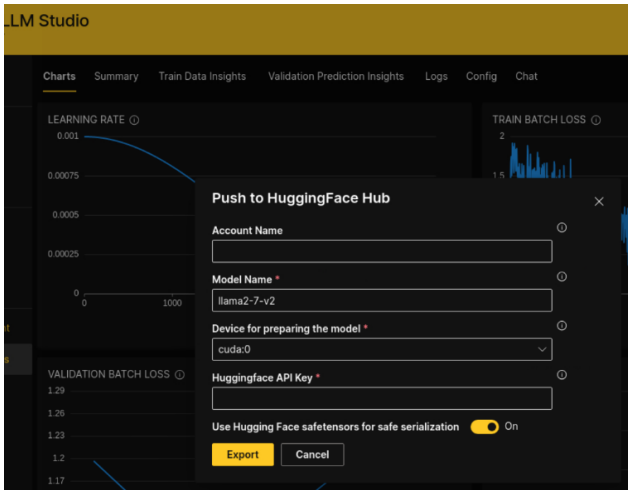
H2O LLM Studio est un des rares outils qui permet d'affiner des LLMs sans devoir écrire une ligne de code. Cependant, il est fortement recommander de satisfaire les prérequis en termes de hardware pour une installation sans accrocs et un entraînement optimal.

Testen & Resultaten

Pour nos tests, nous avons choisi de raffiner un modèle pour la recherche d'information dans des documents relatifs à la sécurité sociale (question answering). Pour cela nous avons généré un set de données d'environ 800 paires de question-réponse en format csv. L'interface est facile d'utilisation, les données sont chargées en deux cliques et l'entrainement d'un modèle se fait par le simple réglage des hyperparamètres dans une fenêtre. Cependant, nous nous sommes très vite heurtés aux limitations de notre hardware peu adapté à l'entrainement de LLMs qui nécessite une grande puissance de calcul. Certains modèles et certaines configurations d'hyperparamètres (ex. utilisation du type float pour les paramètres du réseau) n'ont pu être testés, ce qui a pour conséquence l'entrainement de modèles sous optimum. Pour la plupart de nos expériences, nous avons reçu des messages de type OOM (out of memory) et l'entrainement a dû être abandonné. Ces messages ainsi que les logs détaillés disponibles dans l'application permettent une bonne analyse des problèmes et contiennent des indications sur la façon de régler les hyperparamètres pour résoudre les problèmes de mémoire. Une fois entrainé, le modèle final est envoyé vers la plateforme HuggingFace ou exporté en local.



Nous avons raffiné le modèle Llama2-7b et utilisé l'interface Ollama pour faire tourner le modèle localement. L'application QA proprement dite a été construite avec la librairie Langchain. Les résultats de nos tests QA ont démontré que les performances d'un modèle open-source peuvent être améliorées avec peu d'effort en utilisant un outil « no code » tel que H2O LLM Studio néanmoins celles-ci n'excèdent pas les performances du modèle GPT4.



Conditions d'utilisation & Budget

H2O LLM Studio est un outil disponible en open source sous licence Apache et peut s'installer localement. Son installation requiert une machine Ubuntu ayant au moins une GPU récente et 128 GB de RAM. H2O LLM Studio tourne dans un environnement Python.