DATALAB GROUPE

DataLab Groupe Crédit Agricole

Stages 2024

Présentation du service

Crédit Agricole, Agir chaque jour dans l'intérêt de nos clients et de la société

Au sein du pôle **Innovation & Transformation Digitale (ITD)**, la **Direction Data Groupe** a pour ambition de maximiser la contribution de la **Data et de l'Intelligence Artificielle** au fonctionnement du Crédit Agricole. Elle s'appuie pour cela sur la fonction de Chief Data Officer Groupe et le DataLab Groupe, pôle de référence en **conception interne de solutions Data & IA innovantes et industrielles** en partenariat avec les Caisses régionales, filiales et métiers de Crédit Agricole SA.

DATALAB GROUPE

Le **DataLab Groupe** dispose de toutes les **compétences Data** coopérant au sein de **Squads pluridisciplinaires** selon une **méthode interne d'inspiration Agile**:

- **Data & AI Engineering** visant à préparer les données, définir les architectures, infrastructures et « packager » les solutions qui y seront déployées pour intégration dans le SI,
- **Data Science Analytique et Sémantique** pour concevoir des algorithmes d'Intelligence Artificielle basés sur l'open source exploitant respectivement des données structurées (tabulaires) et des données non structurée (texte, image, voix, vidéos) afin de répondre aux besoins exprimés par les métiers des entités du Groupe,
- **Gestion de projets** qui avec l'ensemble des partenaires et équipes techniques du DataLab Groupe, permet d'identifier et étudier les opportunités, cadrer les projets et en coordonner la réalisation.

Déroulement des stages

Les stages se dérouleront sous **l'encadrement d'experts IA**, au sein d'un **squad pluridisciplinaire** ayant comme **référent fonctionnel un chef de projet**, dans l'objectif de livrer des fonctionnalités intégrables dans des solutions en production, dans un contexte **industriel** et selon la méthode Projet du DataLab Groupe qui fait l'objet d'une **certification**.

Les étapes clés du stage sont les suivantes :

1. Veille bibliographique sur la problématique;
2. Sélection et implémentation des approches les plus adaptées à la problématique ;
3. Réalisation d'une étude comparative sur des données internes et externes;
4. Intégration des développements dans les produits du DataLab Groupe ;
5. Publication scientifique si les travaux aboutissent à de nouvelles approches plus performantes que l'état de l'art.

Le stagiaire aura accès à une **infrastructure de calcul GPU puissante**, ainsi qu'à un **environnement d'engineering industriel à l'état de l'art**.

Les modèles seront évalués sur des **corpus internes** (annotés si beosin) ainsi que des corpus externes (open-data), et seront intégrés dans **les produits et services IA en production**.

Des **interactions** fréquentes avec l'ensemble des équipes **data science et engineering** et des **experts métier** du Groupe auront lieu.

DATALAB GROUPE

Liste des stages 2024

#1 Suggestions de réponses aux emails par des IA génératives

#2 Confidentialité des données dans les IA génératives textuelles

#3 LLMOps : Cycle industriel de prompt engineering

#4 Optimisation des IA génératives opensources

#5 Industrialisation de processus DataOps

#6 Cybersecurité et fuite de données

#7 Optimisation de chaines de traitements analytiques sur GPU

#8 Risques climatiques

#1 Suggestions de réponses aux emails par des IA génératives

#LLM #NLP #deep learning #Traitement des emails

Contexte du stage :

Les conseillers des agences Crédit Agricole reçoivent de nombreux e-mails de clients et doivent rédiger des réponses pertinentes et bien formulées. Afin de gagner du temps, une **Intelligence Artificielle de suggestion de réponse** vise à préparer la rédaction de la réponse afin que le conseiller la valide, au besoin après l'avoir modifiée, puis l'envoie à son client.

Avec l'avènement des **grands modèles de langage (LLM)** popularisés par ChatGPT [1], de nouveaux modèles d'IA générative textuelle pré-entrainés [2][3][4] sont régulièrement mis à disposition de la communauté scientifique.

Ceux-ci ont démontré des performances encourageantes en matière de « zero-shot learning » [2], c'est-à-dire la capacité à réaliser des tâches sans avoir été spécifiquement entraînés pour celles-ci. Cette propriété est intéressante pour la **génération de contenu**, car elle permet d'exploiter les connaissances générales acquises par les LLM lors de leur pré-entraînement. Différentes approches de « fine-tuning » visant à spécialiser les LLM ont émergé [5][6] afin d'augmenter les performances sur un domaine de connaissances spécifique. Elles prennent en compte la taille de ces modèles (plusieurs milliards de paramètres) en proposant des solutions frugales.

Objectifs du stage :

Dans ce contexte, le stage "Génération de suggestions de réponses aux emails" vise à tirer parti des performances en zero/few-shot des LLM pour la génération automatique de réponses d'emails tout en surmontant les défis liés à leur coût computationnel [7]. Disposant d'un large historique d'emails, des approches de fine-tuning seront également étudiées pour générer des réponses plus précises et plus spécifiques aux problématiques du secteur bancaire et de la relation client.

Ce stage constituera une extension d'un projet en cours de traitement d'emails.

Ressources bibliographiques :

[1] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

[2] OpenAI. (2023). GPT-4 Technical Report. https://arxiv.org/abs/2303.08774

[3] S. Zhang et al. (2022). OPT: Open Pre-Trained Transformer Language Models.

[4] BigScience. (2023). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

[5] X. Liu et al. (2022). P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks.

[6] E. Hu et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models.

[7] D. Tim et al. (2022). LLM.Int8(): 8-Bit Matrix Multiplication for Transformers at Scale.

Pour postuler :

STAGE - Assistant Data Scientist – Génération de suggestions de réponse aux e-mails H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

#2 Confidentialité des données dans les IA génératives textuelles

#LLM #NLP #DEEP LEARNING #CONFIDENTIALITE

Contexte du stage :

Les modèles de langage génératifs, tels que les **grands modèles de langage** (LLM), ont récemment connu un essor considérable dans de nombreux domaines, tels que la traduction, la génération et la synthèse de texte, ou encore la réponse aux questions. Ces modèles sont entraînés sur des ensembles de données textuelles très volumineux, ce qui leur permet de générer des textes cohérents et pertinents. Cependant, il a été démontré que ces modèles de langage génératifs sont **vulnérables à des attaques** permettant de reconstituer une partie des données sur lesquelles ils ont été entraînés [1,2,3,7]. Cette vulnérabilité soulève des préoccupations majeures en matière de confidentialité des données, car elle peut entraîner la divulgation involontaire d'informations sensibles, telles que des informations personnelles identifiables ou des secrets d'entreprise.

Les problèmes de confidentialité des données dans les LLMs, incluent les fuites de données, les attaques d'ingénierie inverse, les attaques par modèle et les problèmes liés aux biais et à la discrimination. Bien que certaines études aient été menées sur ces questions, la problématique de la mémorisation des données par les modèles de langage est encore récente, et de nombreux aspects restent à clarifier et à formaliser [4].

En particulier, l'application de techniques de **confidentialité différentielle** (DP) est souvent recommandée [5,6], mais il n'y a pas encore d'étude approfondie sur l'impact de la DP sur les modèles de langage génératifs, en termes de performance et de confidentialité.

Objectifs du stage :

Le stage proposé vise dans un premier temps à étudier les vulnérabilités des modèles de langage génératifs, finetunés sur des données confidentielles, **aux attaques de reconstruction des données d'entraînement** ; et dans un second temps à explorer les techniques de **défense** pour protéger ces données.

Les objectifs du stage incluent l'analyse des attaques existantes, l'évaluation de leur efficacité, la proposition de nouvelles approches pour renforcer la sécurité des modèles et la mise en œuvre de protocoles d'évaluation pour mesurer les compromis entre la performance du modèle et la confidentialité des données.

Ressources bibliographiques :

[1] Carlini et al, 2023, Quantifying memorization across neural language models.
[2] Li et al, 2023, Multi-step Jailbreaking Privacy Attacks on ChatGPT.
[3] Yu et al, 2023, Bag of Tricks for Training Data Extraction from Language Models.
[4] El Mhamdi et al, 2022, On the Impossible Safety of Large AI Models.
[5] Behnia et al, 2022, Privately Fine-Tuning Large Language Models with Differential Privacy.
[6] Majmudar et a, 2022, Differentially Private Decoding in Large Language Models.
[7] Carlini et al, 2020, Extracting Training Data from Large Language Models.

Pour postuler :

STAGE - Assistant Data Scientist – Confidentialité des IA Génératives textuelles H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

#3 LLMOps : Cycle industriel de prompt engineering

#LLM #LLMOps #PROMPT #VERSIONNING

Contexte du stage :

Avec l'avènement des **grands modèles de langage** (LLM) popularisés par ChatGPT, de nouveaux modèles d'IA générative textuelle pré-entrainés sont régulièrement mis à disposition.

Dans le but d'assurer des résultats **pertinents et stables** pour les applications spécifiques de ces LLM, une attention particulière est portée **à l'optimisation et à la sécurisation des processus de création, d'optimisation et de sélection des prompts**.

Objectifs du stage :

L'objectif de ce stage est **d'industrialiser de manière complète et générique le processus de prompting** d'un LLM afin de l'intégrer aux chaînes de fabrication industrielle des cas d'usages.

Pour postuler :

STAGE - Assistant IA Engineer – Cycle de développement de prompts en IA Générative H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

DATALAB GROUPE

#4 Optimisation des IA génératives opensources

#LLM #OPTIMISATION #RSE

Contexte du stage :

Les **grands modèles de langage** (LLM) ont révolutionné diverses applications dans le domaine du traitement du langage naturel, telles que la traduction, la génération, la synthèse de texte ou encore la réponse aux questions. Cependant, ces modèles présentent des défis majeurs en termes de taille [1], de coûts d'utilisation et de consommation d'énergie.

Afin de réduire la consommation d'énergie et de faciliter le déploiement de ces modèles sur des appareils à ressources limitées, plusieurs techniques ont été proposées dans la littérature pour **optimiser l'usage des LLMs** [2,3].

Parmi ces techniques figurent la quantification des poids et des activations des modèles [4,5], la distillation des connaissances, la factorisation de matrices des poids. La quantification consiste à réduire la précision des poids et des activations des modèles en utilisant un nombre réduit de bits, ce qui permet de réduire la taille du modèle et d'accélérer l'inférence tout en maintenant des performances acceptables. La distillation des connaissances [6], quant à elle, vise à transférer les connaissances d'un modèle de grande taille (enseignant) à un modèle plus petit (étudiant) en ajustant les sorties de l'étudiant pour qu'elles correspondent à celles de l'enseignant.

Récemment, Intel a proposé une approche prometteuse appelée SmoothQuant pour la quantification des modèles de langage à grande échelle, qui a permis de réduire de moitié la taille des modèles tels que LLaMA et OPT avec une perte de précision négligeable [7]. Cette approche se base sur une quantification non uniforme des poids et des activations des modèles, en utilisant une fonction de quantification lisse pour réduire les erreurs de quantification et améliorer la précision du modèle quantifié.

Objectifs du stage :

Le DataLab Groupe, dont le **processus de fabrication d'IA est labélisé RSE**, met au cœur de ses préoccupations la nécessité de développer des IA dont le coût financier et l'empreinte carbone sont raisonnés. En particulier, le fait de privilégier l'inférence de modèles tels que les **LLM sur des infrastructures CPU** déjà existantes contribue grandement à ces objectifs.

Le stage proposé vise à :

- Identifier, comprendre finement et comparer les différentes techniques d'état de l'art pour réduire la taille et la consommation d'énergie des LLM,
- Optimiser le compromis entre performances et sobriété pur une inférence sur CPU.

Quelques ressources :

[1] Bondarenko et al. "Understanding and overcoming the challenges of efficient transformer quantization" ACL 2021.

[2] Dettmers et al. "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale" NeuriPS 2022

[3] Zadeh et al. "Gobo: Quantizing attention-based NLP models for low latency and energy efficient inference." MICRO 2020

[4] Shen, Dong & Ye, et al. "Q-BERT: Hessian based ultra low precision quantization of BERT" AAAI 2020.

[5] Yao et al. "ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers" arXiv preprint arXiv:2206.01861 (2022).

[6] Hinton et al. "Distilling the Knowledge in a Neural Network" NIPS 2014.

[7] Xiao & Lin "SmoothQuant: Accelerated sparse neural training: A provable and efficient method to find N:M transposable masks."

Pour postuler :

STAGE - Assistant Data Scientist – Frugalité des LLM - Optimisation pour une inférence sur CPU H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

DATALAB GROUPE

#5 Industrialisation de processus DataOps

#DATAOPS #VERSIONNING

Contexte du stage :

Le **DataOps** est un cadre, inspiré du DevOps, qui permet d'industrialiser et optimiser le processus de gestion du cycle de vie de la donnée. Il introduit des pratiques agiles dans le développement des traitements de données afin que les équipes responsables des données et les utilisateurs finaux travaillent ensemble plus efficacement.

Objectifs du stage :

Un chantier a été démarré au DataLab Groupe pour couvrir une pratique clé du DataOps : **le versioning des données**. L'objectif de ce stage est de poursuivre ce chantier sur environnement de type **cluster GPU** en utilisant le service de stockage **S3** ainsi que l'outil de versionning **DVC**.

Pour postuler :

STAGE - Assistant Data Ingénieur - Industrialisation du processus de DataOps H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

DATALAB GROUPE

#6 Cybersecurité et fuite de données

#CYBERSECURITE #SOC #REINFORCEMENT LEARNING

Contexte du stage :

La détection des fuites de données est nécessaire pour assurer la **sécurité des données des clients et des collaborateurs** du Groupe. Elle doit être réalisée au plus près du démarrage de l'attaque afin de limiter les impacts. Plusieurs contraintes rendent cette tâche complexe, principalement la **volumétrie** des données à analyser et le **perpétuel changement des méthodes** utilisées par les attaquants pour exfiltrer la donnée.

Objectifs du stage :

Dans ce stage, nous souhaitons étudier des nouvelles approches de détection de fuite de données à base de modèles par **Reinforcement learning** ([1], [2], [3]) pour répondre à ces contraintes.

Comme les alertes remontées par les systèmes actuels sont analysées par les experts métiers pour évaluer leur pertinence et leur incidence, l'objectif serait de prendre en compte ces retours métiers dans le processus d'apprentissage par ces nouveaux modèles.

Quelques ressources :

[1] Tekkali, C.G., Natarajan, K. RDQN: ensemble of deep neural network with reinforcement learning in classification based on rough set theory for digital transactional fraud detection. Complex Intell. Syst. (2023)

[2] Nguyen, Thanh & Janapa Reddi, Vijay. Deep Reinforcement Learning for Cyber Security. IEEE Transactions on Neural Networks and Learning Systems. (2021)

[3] A. E. Bouchti, A. Chakroun, H. Abbar and C. Okar. Fraud detection in banking using deep reinforcement learning. Seventh International Conference on Innovative Computing Technology (INTECH). (2017)

Pour postuler :

STAGE - Assistant Data Scientist – Détection de la fuite de données avec du Reinforcement Learning H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

#7 Optimisation de chaines de traitements analytiques sur GPU

#DATA ENGINEERING #GPU #REINFORCEMENT LEARNING

Contexte du stage :

L'optimisation des chaînes de traitement est essentielle pour accélérer les usages Data et IA. Elle vise à améliorer :

- La **rapidité** d'exécution au regard des ressources de calcul disponibles et de leur coût financier et environnemental,
- La **scalabilité** afin de s'adapter de manière fluide à des volumes de données plus importants,
- La **simplicité** de prise en mains et d'utilisation par les experts Data pour une meilleure exploitation.

Actuellement, la Squad IA Analytique travaille sur des clusters **Spark** qui ont prouvé leur efficacité en termes de montée en charge. Toutefois, afin de surmonter les limites de Spark, telles que la complexité d'utilisation, des fonctionnalités parfois restrictives et l'incompatibilité avec certains algorithmes, nous cherchons des alternatives.

Objectifs du stage :

Dans ce stage, un premier travail de veille sera entrepris pour explorer ces alternatives, dont des méthodes distribuées ou non, basées sur l'utilisation de **cluster de GPU.**

Ensuite, une méthodologie rigoureuse de comparaison et d'analyse sera mise en place pour évaluer ces options par rapport à l'existant, permettant ainsi de sélectionner les solutions les plus appropriées pour une utilisation effective en production sur les projets data et IA actuels et futurs.

Pour postuler :

STAGE - Assistant Data & AI Engineer - Optimisation des chaînes de traitement sur GPU H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

DATALAB GROUPE

#8 Risques climatiques

#OPEN DATA #RSE #CLIMAT

Contexte du stage :

De par notre engagement sociétal, nous nous intéressons aux **impacts du changement climatique sur les territoires de nos Caisses régionales** afin d'aider à prévoir et anticiper les risques.

Des premiers travaux du DataLab Groupe ont permis de mettre en lumière les risques sur **l'agriculture**. Nous voulons aujourd'hui aller plus loin dans cette étude notamment sur les **risques immobiliers et l'habitabilité des territoires** exposés aux aléas climatiques.

Objectifs du stage :

Ce stage devra donc :

- Faire émerger des données **Open Data** permettant de traiter ces aspects, valider leur pertinence et leur qualité et les mettre à disposition sur l'**OpenDataMart** (magasin de données Open Data mis à disposition de l'ensemble des entités du Groupe),
- Réaliser la préparation de ces données (calcul d'indicateurs…) et les exploiter pour de la **Data Vizualisation**.

Pour postuler :

STAGE - Assistant Data Analyste – Risques Climatiques H/F | Stage | Montrouge | France | Crédit Agricole Carrières (groupecreditagricole.jobs)

DATALAB GROUPE