CoLab.IA

Plateforme expérimentale d'ingénierie pour le Deep Learning

Webinaire Kubernetes

Jocelyn DE GOËR, L. COURNEDE

UMR EPIA - CATI IMOTEP

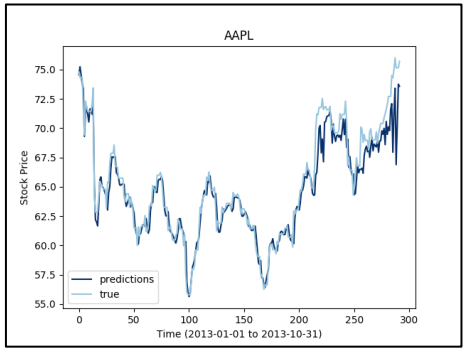
04 juillet 2023

Les réseaux de neurones artificiels

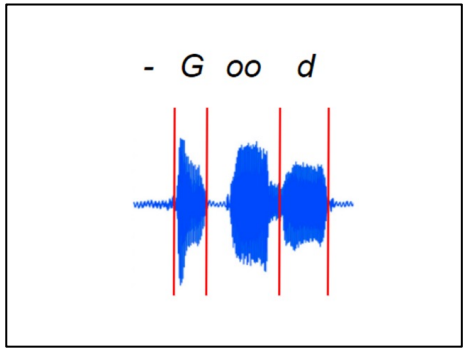
✓ **Utilisation dans de nombreux domaines :**

Analyse et génération de textes

Analyse de données Bio-Informatique

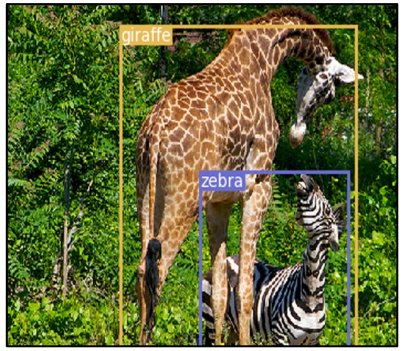


Analyse de séries temporelles



Analyse du son Reconnaissance vocale

Assistance au diagnostique médical



Analyse d'image ou de vidéos

Ressources de calcul pour le Deep Learning

- **Un développement lié à l'évolution de la puissance de calcul des machines**
  - **2008 : Début de l'ère du GPGPU (General-purpose processing on graphics processing units) : NVIDIA G80**
  - 2012 : Premier succès du DeepLearning : le réseau Alexnet remporte la compétition ILSVRC basé sur le DataSet d'images ImageNET
    - Puissance de calcul nécessaire : 0,0054 petaflop/s par jour
  - 2017 : Développement d'AlphaGo Zero
    - Puissance de calcul nécessaire : 1 860 petaflop/s par jour
  - 2022 : Développement du modèle GPT4
    - Puissance de calcul nécessaire : 6 000 petaflop/s par jour
      - 10 000 GPU NVIDIA A100

1 petaflops = 10^{15} flops

Ressources de calcul pour le Deep Learning

✓ Ressources de calcul

Station de travail « gaming »

- Entrainement de petits modèles
- Difficulté d'intégration dans un bureau
- Coût : 3 000€ à 10 000€

Plateforme INRAE CoLab.IA

- Offre intermédiaire

Centres de calcul

- 23 mésocentre proposent des ressources CPU+GPU : https://calcul.math.cnrs.fr/pages/mesocentres_en_france.html
- Permettent d'entrainer des modèles à partir de gros DataSet

- Délais d'accès
- Peuvent ne pas accepter tout les types de données (données médicales)

Plateforme CoLa.IA

- ✓ **Qu'est-ce que CoLab.IA ?**

  - ✓ **Plateforme expérimentale d'ingénierie dédiée au Deep Learning**

  - ✓ **Permettre aux équipes INRAE de pouvoir s'initier au Deep Learning**
    - ✓ Mise à disposition de ressources de calcul GPU, suffisamment dimensionnées pour la création et l'entrainement de réseaux de neurones

  - ✓ **Construire une animation communautaire**
    - ✓ Système participatif
    - ✓ Échange de savoir-faire techniques et méthodologiques
    - ✓ Organisation de formations en Deep Learning

Plateforme CoLa.IA

- ✓ **Deux modes de fonctionnement :**
  - ✓ **Accès communautaire :**
    - ✓ Ouvert à toutes les équipes INRAE souhaitant débuter une activité en Deep Learning
    - ✓ Accès aux ressources partagées avec une file d'attente et une limite de temps
  - ✓ **Accès prioritaire :**
    - ✓ Ouvert à toutes les équipes INRAE ayant fait l'acquisition de serveurs intégré à CoLab.IA
    - ✓ Utilisation des ressources acquises sans limitation de durée
    - ✓ Mise à disposition de la communauté des ressources inutilisées
    - ✓ Administration des machines par l'équipe technique
- ✓ **Publications et communications :**
  - ✓ Apparaitre dans la section « remerciements » des articles scientifiques résultants de l'utilisation de CoLab.IA
  - ✓ Apparaitre en tant qu'auteur des articles, lorsqu'un accompagnement méthodologique poussé a été réalisé par un des membres de l'équipe technique de CoLab.IA
  - ✓ Les données et scripts d'analyse restent la propriété des équipes utilisatrices

Plateforme CoLa.IA

✓ Services proposés

- ✓ **Environnements Jupyter**
  - ✓ Environnements construits sur mesure
  - ✓ Notebooks pour l'exécution de code Python ou R
  - ✓ Terminal BASH

- ✓ **Mise en place de services en ligne accessibles via des API (En construction)**
  - ✓ **Retranscription textuelle d'enregistrements audios**
    - ✓ Pré-traitement audio ave FFMPEG
    - ✓ Retranscription textuelle via le réseau de neurone OpenAI Whisper
      - ✓ Formats : TXT, CSV, STR...
    - ✓ Reconnaissance des locuteurs

  - ✓ **Exploitation de LLM (Large Language Model) OpenSource**
    - ✓ Développement de Chatbots spécialisés sur un domaine particulier
      - ✓ Embeddings à partir de documents PDF
      - ✓ Finetunning à partir d'un corpus de texte important

Plateforme CoLa.IA

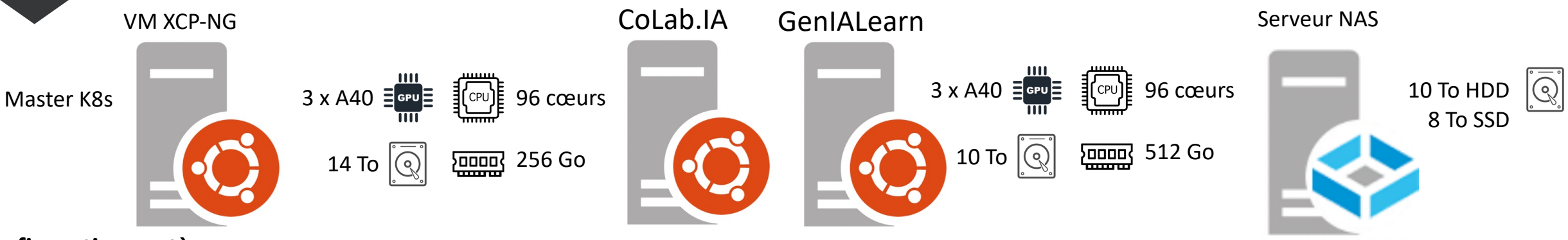
- **Besoins croissants de puissance de calcul GPU**

  - **Entrainement d'un réseau de neurones de type YoloV5 pour l'analyse d'images :**
    - 40 000 images
    - 6 classes d'objets
    - 40 millions de paramètres
    - Temps de calcul avec un GPU A40 : 5 jours

  - **Retranscription textuelle audio avec Whisper**
    - Avec le modèle Large de Whisper
    - Pour 30 min d'audio il faut 15 min de traitement
      - FFMPEG + Whisper + détection des locuteurs

  - **Interrogation d'un modèle LLM pour une tâche de Chatbot**
    - LLM 40 milliards de paramètre
    - VRAM nécessaire : 50Go

Plateforme CoLa.IA



Configuration système

- ✓ Système d'exploitation : Ubuntu 20.04 LTS
- ✓ Version de CUDA : 11,4

Kubernetes (MicroK8s)

- ✓ Serveur : 1.23
- ✓ Client : 1.23

Registry

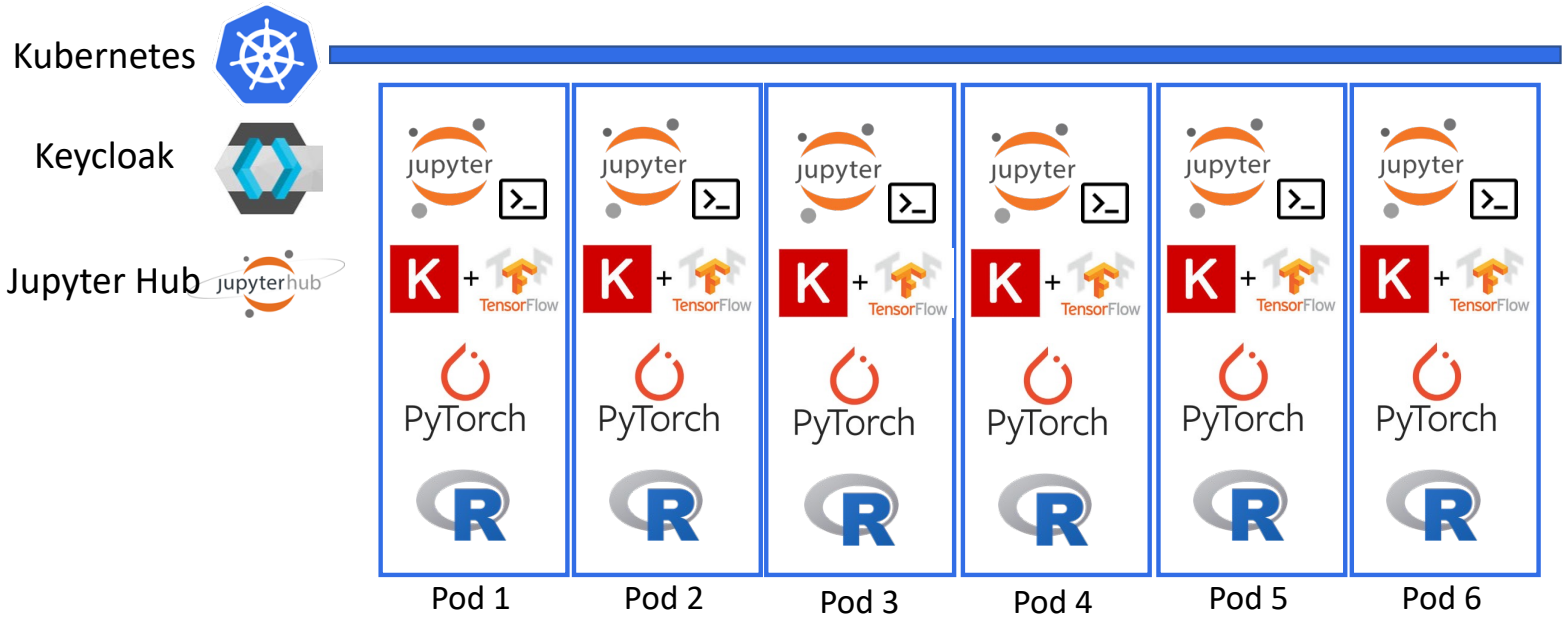
- ✓ Docker-Registry 2
- ✓ Docker-Registry UI 2.4

Environnements Jupyter

- ✓ Jupyter Hub 3.0

Authentification

- ✓ Keycloack 20



Plateforme CoLa.IA

- ✓ **Utilisateurs réguliers de CoLab.IA**
  - ✓ 30 comptes utilisateur

- ✓ **Unités :**
  - ✓ **UMR EPIA :** J. de Goër, D. Abrial, Y. Frendo
  - ✓ **Plateforme GAMAE** : F. Johany
  - ✓ **IGEPP :** N. Parisey
  - ✓ **LPGP :** J. Bugeon
  - ✓ **UR ASSET :** M. Bonneau

- ✓ **Équipes des projet GenIALearn**
  - ✓ **UMR GABI** :
    - ✓ **Équipe BIGE-IBISC** : E. Barrey, A. Ricard
    - ✓ **Éguipe G2B** : P. Croiseau, T. Tribout, B. Castro Dias Cuyabano
  - ✓ **MIA Paris** : J. Chiquet, J. Kwon, T. Mary-Huard

Plateforme CoLa.IA

- **COPIL CoLab.IA :**
  - **CATI IMOTEP** : Jocelyn DE GOËR, Nicolas PARISEY, Thierry HOCH et Hervé RICHARD
  - **CATI SICPA** : Bernard BENET, Bernadette URBAN, François LAPERRUQUE, Yann LABRUNE
  - **Direction des Systèmes d'Information** : Éric MALDONADO
  - **Plateforme MIGALE** : Valentin LOUX

- **Équipe technique :**
  - Jocelyn DE GOËR, Yann FRENDO et Laurent COURNEDE

- **Financements :**
  - 2021 : 20k€ - AAP DipSO SAPI 2021
  - 2021 : 20k€ - Projet GenIALearn AAP DigitBIO 2021
  - 2022 : 800€ - Projet FAVEC – AAP DipSO SAPI 2022
  - 2022 : 1,3k€ - Projet GAME-PLAAI : AAP DipSO SAPI 2022
  - 2023 : 10k€ - UMR EPIA
  - 2023 : En cours de demande : 15k€ (projet GenIALearn, DeepPheno – AAP DIGIT-BIO)

- **Tarifs du marché MatInfo5 HP :**
  - **20k€ :** serveur de calcul 96 cœurs, 512Go de RAM, 1 GPU A100 80Go de VRAM
  - **10k€ :** GPU A100 supplémentaire (3 GPU par serveur)

CoLab.IA

Plateforme expérimentale d'ingénierie pour le Deep Learning

Webinaire Kubernetes

Jocelyn DE GOËR

UMR EPIA - CATI IMOTEP

04 juillet 2023