Analyse de performance de TensorFlow

ASR 2019-2020

TENSORFLOW PROFILING

Auteurs

Réalisé par :
Mehdi Messaoudi
Fatma Bouzghaia

Encadré par:
François Trahay

Contexte, Problématique et Objectifs

Contexte

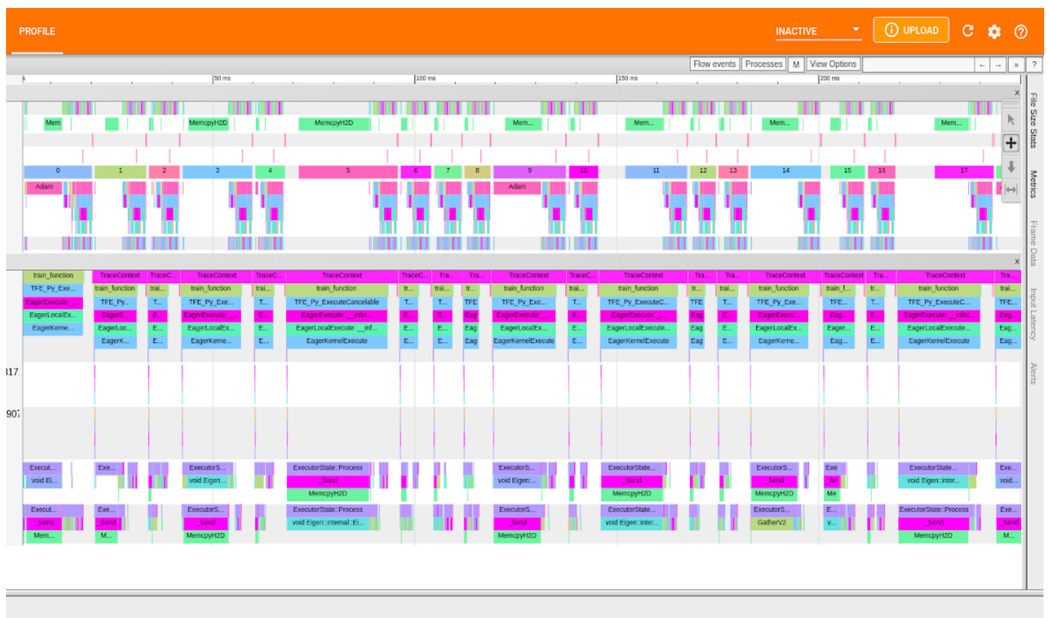
- TensorFlow est une bibliothèque de logiciels open source pour le calcul numérique à l'aide de graphiques de flux de données.
- Nous voulons toujours que nos modèles TF s'entraînent plus rapidement: Optimisation de code et favorisation de l'utilisation de la GPU..

Problématique

- Il est possible que même après avoir accéléré les calculs, le modèle présente des inefficacités dans la pipeline.
- Dans de tels cas, il est très difficile de déboguer son code, ou même de dire ce qui cloche.
- L'utilisation des outils d'analyse de performance comme Tensorboard et NVIDIA Profiler s'impose.

Objectifs

- Etudier les outils d'analyse existants de TensorFlow
- Mise en place de deux modèles Tensorflow utilisant la dataset CIFAR-10, basés respectivement sur Le modèle CNN simple et le modèle Resnet.
- Appliquer différentes optimisations sur les deux modèles dans le but de voir leur impact sur l'utilisation de la GPU par TensorFlow.



Tensorboard et NVIDIA Profiler

Tensorboard

- Surveille la formation du modèle.
- Stocke le temps nécessaire à l'exécution des opérations.
- Stocke le temps nécessaire à l'exécution des étapes du modèle.
- Note des informations concernant l'utilisation des ressources coûteuses (CPU et GPU)

NVIDIA Profiler

- Fournit des informations de profilage pour les applications CUDA.
- Montre l'activité du CPU et du GPU qui s'est produite au fil du temps.
- Guide le développeur en adoptant une approche étape par étape pour comprendre les principaux limiteurs de performance.

Technologies phares

Modèles et application des optimisations

Modèles basés sur la dataset CIFAR-10

Modèle CNN simple

Un réseau de neurones convolutifs simple qui est formé par un empilement de couches de traitement : Une couche de convolution, Max pooling Layers, Flatten, ReLU et Dense.

Modèle RESNET-50

ResNet-50 est un réseau résiduel profond. Le «50» fait référence au nombre de couches dont il dispose. Il s'agit d'une sous-classe de réseaux de neurones convolutifs, ResNet étant le plus couramment utilisé pour la classification d'images.

Application des optimisations

Mise en place d'un code fonctionnel de ces 2 modèles ensuite application de différentes optimisations basées sur les observations dégagées à partir de Tensorboard et de NVIDIA Profiler. Parmi les optimisations effectuées on cite:

- Variation du batch size.
- Utilisation du la GPU en mode privée.
- Utilisation de la précision mixte.
- Prefetching data…