Réduction de l'empreinte carbone des modèles d'apprentissage automatique

Palyart, Marc, marc.palyart@malt.fr, Malt, Bordeaux (Orateur)

Thématique : Thème 7 : Sobriété numérique

Résumé : *La consommation énergétique est la principale source anthropique d'émissions des gaz à effet de serre qui sont à l'origine du réchauffement climatique. Or, avec leur utilisation et leur complexité toujours croissantes, les modèles d'apprentissage automatique consomment de plus en plus d'énergie. Après un rapide tour des outils permettant d'évaluer l'impact de ces modèles, nous aborderons dans cette présentation plusieurs familles de bonnes pratiques permettant de réduire l'impact de l'apprentissage et l'utilisation de ces modèles.*

Mots clés : *empreinte carbone, apprentissage automatique, optimisation*

1. Introduction

La consommation énergétique est la principale source anthropique d'émissions des gaz à effet de serre qui sont à l'origine du réchauffement climatique. Or, avec leur utilisation et leur complexité toujours croissantes, les modèles d'apprentissage automatique consomment de plus en plus d'énergie.

2. Méthodologie

L'approche détaillée dans cette présentation se décompose en trois parties. Dans un premier nous évoquerons différentes bibliothèques permettant d'évaluer l'impact en termes d'équivalent C02. Nous verrons aussi qu'il est possible de se baser sur d'autres mesures proxy que ces estimations.

Dans un second temps, nous nous attarderons sur l'optimisation de l'entraînement avec notamment les techniques permettant d'éviter de nouveaux entraînements (modèles pré-entraînés spécifiques ou approches zero-shot learning) ou d'en limiter leur durée (approches par fine tuning/transfert learning). Nous discuterons également de l'optimisation des cycles d'apprentissage grâce à la surveillance des performances du modèle. Nous conclurons cette partie en évoquant un certain nombre de bonnes pratiques généralistes: recherche plus efficace d'hyper-paramètres, choix d'architecture matérielle et du datacenter, …

Dans une troisième et dernière partie nous aborderons l'optimisation de l'inférence des modèles d'apprentissage automatique. Nous commencerons par présenter les runtimes permettant d'accélérer la phase de prétraitement (Numba, Polars, CuPy) et la phase d'inférence pure (Torchscript/ONNX/TensorRT). Nous expliquerons ensuite plusieurs techniques permettant de réduire la taille d'un modèle: distillation, quantization et pruning. Comme pour la section sur l'entraînement nous terminerons en évoquant un certain nombre de bonnes pratiques généralistes: importance des mises à jour, mise en cache, réduction des descripteurs, choix d'architecture matérielle, …

3. Originalité / perspective

La plupart des présentations sur le sujet restent assez théoriques et abstraites. A contrario cette présentation se veut pratique en permettant la découverte de certaines techniques existantes facilitant ainsi une mise en application plus rapide.

A propos de Malt:

Malt est une marketplace européenne où plus de 430 000 consultants freelances mettent leurs compétences et leurs expertises au service des entreprises qui recherchent des talents externes pour accélérer leur projets.

Marc Palyart, Staff Data Scientist à Malt

Docteur en informatique, cela fait plus de 10 ans que je travaille dans la data. Avec un mix d’expériences en recherche universitaire et dans l’industrie, j'ai eu la chance de collaborer avec des personnes fabuleuses. Je suis actuellement lead de la partie search et matching chez Malt, la marketplace de freelances.