ML-Group

Gruppe A-5: Computer Vision for Edge devices

Brief Description: Training, evaluation, and deployment of visual models on edge devices: The subprojects are a) Data preparation and selection of pre-trained models b) Training the selected models on the corresponding edge devices and choosing the best model c) Deploying the trained model on the corresponding edge devices.

Framework: Pytorch (https://pytorch.org/docs/stable/index.html)

Flower (https://github.com/adap/flower).

Orga

17:25 - 17:50 (s.t) à if needed, we could select this time slot for students to coordinate, plan and split their subtasks and align progress.

Group (0 / 10) –> we set the maximum number of the group to be 10.

Detailed Description

Recent advances in artificial intelligence (AI), edge computing, and the adoption of Internet of Things (IoT) devices have collectively opened up opportunities for edge AI [1]. Edge AI refers to deploying AI applications across devices throughout the physical world. It is termed "edge AI" because AI computations are performed near the user, at the network's edge, close to where the data is generated, rather than in centralized cloud computing facilities or private data centers. With the internet's global reach, the network edge can virtually be anywhere. It might be in retail stores, factories, hospitals, or devices around us like traffic signals, road traffic flow, and self-service supermarkets [2]. Businesses across all sectors are seeking to increase automation to improve processes, efficiency, and safety. To aid them, computer vision systems need to identify patterns and perform tasks repetitively and safely. However, the world is unstructured, and the range of tasks performed by humans spans an infinite variety of environments that cannot be fully described in programs and rules.

The advancements in edge AI have paved the way for devices to operate with "intelligence" and imitate human cognition, regardless of their location. AI-powered devices are capable of learning to perform diverse tasks in various situations.

Why do we need to deploy AI models on the edge and what will we do in this project?

Deploying AI models entirely in centralized clouds or corporate data centers is either impractical or costly, due to issues related to latency, bandwidth, and privacy [3].

- **Intelligent Edge Devices**: AI-powered edge devices are more powerful and flexible than traditional devices because traditional devices can only react to anticipated inputs. In contrast, learning-based devices are trained not on how to answer a specific question but on how to respond to a series of questions. Without AI, traditional applications couldn't possibly handle the infinitely varied inputs like text, speech, or video. In this project, we will focus on visual models, using datasets such as CIFAR [4].

- **In-time Response**: Edge AI enables real-time responses to user needs by analyzing data locally, rather than transfer data to the clouds with remote communication delays. In this project, we will train and evaluate models on low-power edge devices, such as Jetson Nano [5].
- **Cost Reduction**: By bringing processing power closer to the edge, applications require less internet bandwidth, significantly reducing network costs.
- **Increased Privacy**: Edge AI can process data locally without exposing it to the cloud, greatly enhancing privacy protection. Even if some data is uploaded for training purposes, it can be anonymized to protect user identities. By protecting privacy, edge AI simplifies challenges related to data compliance. In this project, we will explore training models on several edge devices in the context of federated learning.
- **High Availability**: Decentralization and offline capabilities make edge AI more robust, as processing data does not rely on the support from computing power from the centralized cloud. This brings higher availability and reliability for critical missions and production-level AI applications. In this project, we will deploy the trained models on the corresponding edge devices.
- **Continuous Improvement**: AI models become more accurate with training on more target data. When edge AI applications encounter the scenarios they can't process accurately or confidently, if needed there is an option to upload the data anonymously so that cloud models can be retrained and learn from it. Thus, the longer a model operates at the edge (closer to the target data), the more accurate it becomes.

In conclusion, during this project, we will use lightweight edge devices to simulate this process, training, evaluating, and deploying models on edge devices in conjunction with federated learning.

[1] Situnayake, Daniel, and Jenny Plunkett. *AI at the Edge.* " O'Reilly Media, Inc.", 2023.

[2] Wang, Xiaofei, et al. *Edge AI: Convergence of edge computing and artificial intelligence*. Singapore: Springer, 2020.

[3] https://www.thedigitalspeaker.com/edge-ai-bridging-innovation-privacy-efficiency/

[4] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[5] https://developer.nvidia.com/embedded/jetson-modules