WIRTSCHAFTS INFORMATIK

Masterthesis (6 Monate)

TECHNISCHE UNIVERSITÄT DARMSTADT

Requirement specification for Deep-Learning pipeline deployment on edge

For deploying our deep learning solutions developed in open-source frameworks (PyTorch, OpenCV, etc.) on edge, we need high performance software servers to handle the inference requests. We develop deep learning software solutions with a high inference speed on development hardware but struggle with high performance deployment on edge.

For this thesis, you continue an existing project where we have tested TorchServe. However, TorchServe does not meet our requirements for performant, hardware optimized multi model serving and sequential pipelines. Our next look is towards Nvidia Triton, as it is native to Nvidia Jetson (our deployment hardware) since August 2021. Our software solutions can be sequential multi model pipelines but also with parallel parts.

As an example use case we have implemented a visual inspection solution using deep learning frameworks (PyTorch) and Python and all its corresponding software/code (Exported model, python code for handling). We would like to test specific deep learning models that we have implemented on dummy data with many possible setup changes. Many different testing setups and the influence of each setup change on performance results needs to be evaluated:

1) Testing of different models: UNet, MaskRCNN, FasterRCNN, YoloV4, …
2) Testing of different parameters / setups: Image resolution change, floating point precision, python handling libraries
3) Testing of different deployment frameworks
4) Testing of different hardware
5) Optional: Additional tests using Tensorflow models instead of pytorch

Specifics of the external work

We are looking for an interested person to thoroughly evaluate and compare deployment options on specialized hardware like Nvidia Jetson, providing performance examples on some of our given solutions.

During the work, you will be employed by Heraeus for the duration of this work. You will have access to infrastructure and data and work with experts from the areas of production, data science and digitalization as well as with various executives and stakeholders. If you apply for this job, we ask you to provide us with a short motivation letter, your CV and your current reference.

Requirements

Knowledge in Python, Numpy, OpenCV, Docker, and Linux/Ubuntu. Ideal: Basic deep learning knowledge like pytorch, ONNX, GPU-hardware.

Supervisor / Coach

Start possible immediately. If you are interested, please contact:

Timo Koppe
timo.koppe@tu-darmstadt.de