

From TVM to Deep Learning: Cross-Framework-Device-One-Shot-Inference-Solution

The main aim of the project is to develop a cross-framework-device inference solution for deep learning based on TVM. This solution supports receiving pre-trained deep learning models from cloud centres and performing inference these models on edge devices, which enables the process of training-inference of deep learning models to be deployed smoothly. The project architecture [Figure 1] consists of four essential steps, exporting pre-trained models' configuration files and weights, downloading these files and importing to TVM stack, performing inference in TVM, getting output and processing based on task type. In the first step, users will be able to build and train deep learning models in the cloud centre/PC or even get pre-trained models directly. Then they need to export the models as configuration files and weights. In second steps, this solution provides methods to download the files to edge devices (Linux), reload model and import to TVM stack. For the next step, this solution performs the model's inference based on TVM's API. In the final step, this solution gets the output from TVM and process output according to task type, for image recognition, this solution will display the content in pictures and the percentages of accuracy, for object detection, this solution will display the classifications, the confidence levels and bounding boxes.

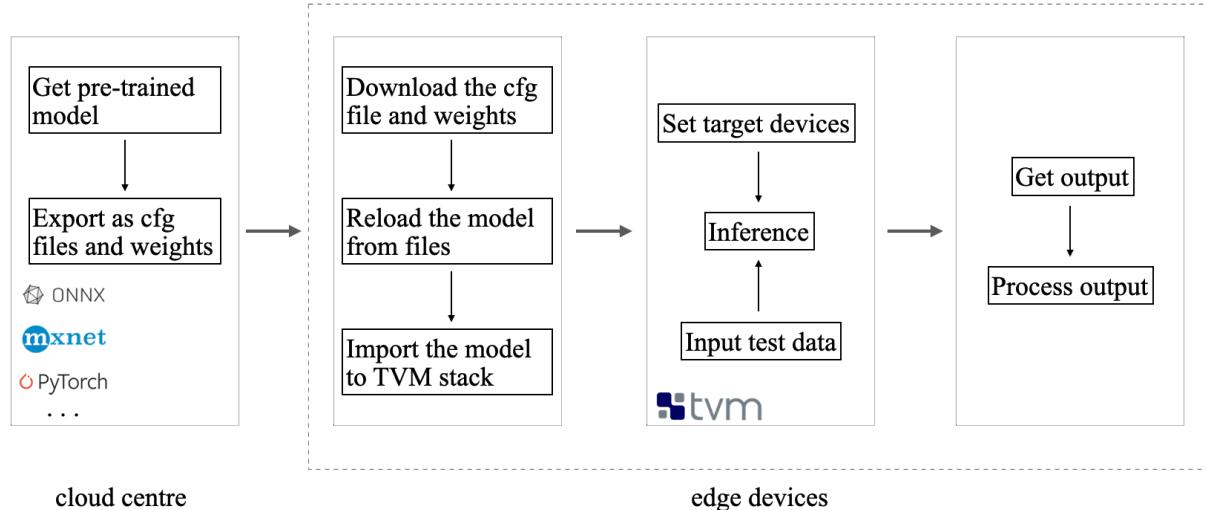


Figure 1: Main Architecture