In-training quantization

Quantization is a fundamental step to optimize deep neural networks to fit memory and processing constraints. The most used technique is post-training quantization; however, it can considerably decrease accuracy.

In-training quantization is a novel technique that quantizes deep models during the training procedure and can lead to better performance in terms of accuracy.

This thesis aims to investigate and implement in-training quantization solutions, test them with state-of-the-art datasets such as ImageNet or Visual Wake Words, and deploy the resulting model in edge devices.