Contribution ID: **62** Type: **Poster only**

Optimizing Multi-Model Compression for Resource-Constrained Devices (Poster Upload)

Monday 23 September 2024 17:13 (1 minute)

We address the challenge of compressing a sequence of models for deployment on computing- and memory-constrained devices. This task differs from single model compression, as the decision to apply compression schemes either independently or jointly across all sub-networks introduces a new degree of freedom. We evaluate the performance of pruning and quantization techniques for model compression in the context of a prototypical image restoration and object detection multi-model system. We propose an adaptation of Quantization Aware Training (QAT) and pruning techniques, where the multi-model system is fine-tuned as a single unit with an adapted loss function, rather than applying these techniques to each model individually.

What of the following keywords match your abstract best?

Other

Please tick if you are a PhD student and wish to take part to the poster prize competition!

Author: PRADO, João Luís

Co-authors: Dr MAKAREM, Laleh (Logitech); Dr SALZMANN, Mathieu (EPFL)

Presenter: PRADO, João Luís

Session Classification: Flash talks / poster session