Onnx Instructions

ONNX

ONNX (The Open Neural Network Exchange) is an AI ecosystem that created standards for representing machine learning algorithms. These standards allow ONNX machine learning algorithms to be run on a wide variety of machines (unfortunately excluding Apple).

ONNX supports a wide variety of frameworks, including Matlab, Keras, TensorFlow, and PyTorch. For a full list and detailed instructions on converting to and from ONNX models, see this page: https://onnx.ai/supported-tools.html

Instructions for converting a TensorFlow model to ONNX:

1. Install ONNX convertor:
   pip install git+https://github.com/onnx/tensorflow-onnx
2. To convert a model from TensorFlow to ONNX, run the following.
   python -m tf2onnx.convert --saved-model tensorflow-model-path --output model.onnx
3. The model is now ready to be run on ONNX Runtime.

ONNX Runtime

ONNX does not itself perform inference. To do that, you need to install a runtime that can run ONNX. Use the following website for installation and operation instructions: https://onnxruntime.ai/index.html#getStartedTable

Using this, you can perform inference up to 17x faster. ONNX Runtime also allows you to accelerate PyTorch training (only PyTorch so far) up to 40% faster.