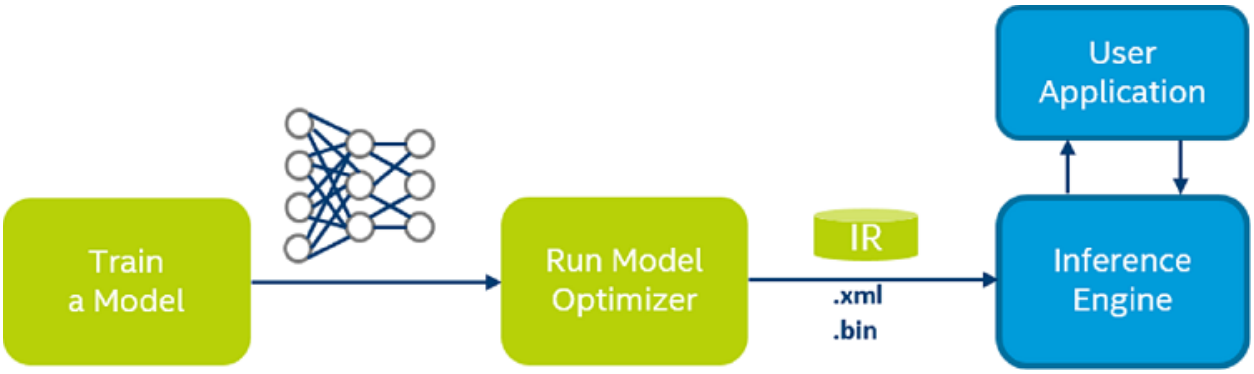
Model Optimizer

Model Optimizer is a cross-platform command-line tool that facilitates the transition between the training and deployment environment, performs static model analysis, and adjusts deep learning models for optimal execution on end-point target devices.

Model Optimizer process assumes you have a network model trained using a supported deep learning framework. The scheme below illustrates the typical workflow for deploying a trained deep learning model:



Model Optimizer produces an Intermediate Representation (IR) of the network, which can be read, loaded, and inferred with the Inference Engine. The Inference Engine API offers a unified API across a number of supported Intel® platforms. The Intermediate Representation is a pair of files describing the model:

- .xml - Describes the network topology
- .bin - Contains the weights and biases binary data.

While running Model Optimizer you do not need to consider what target device you wish to use, the same output of the MO can be used in all targets. Model optimizer performs a generic optimization on the imput model while inference engine performs hardware specific optimization.

Inference Engine

Inference Engine is a runtime that delivers a unified API to integrate the inference with application logic:

- Takes as input the model. The model presented in the specific form of Intermediate Representation (IR) produced by Model Optimizer.
- Optimizes inference execution for target hardware.
- Delivers inference solution with reduced footprint on embedded inference platforms.

The Inference Engine supports inference of multiple image classification networks, including AlexNet, GoogLeNet, VGG and ResNet families of networks, fully convolutional networks like FCN8 used for image segmentation, and object detection networks like Faster R-CNN.

For the full list of supported hardware, refer to the Supported Devices section.

The Inference Engine package contains headers, runtime libraries, and sample console applications demonstrating how you can use the Inference Engine in your applications.