Explaining Custom Layers

- **The process behind converting custom layers:**
  Converting a custom layer is dependent on the types of Model, that is either Tensorflow Models or Caffe Models.
  There are two options in converting the custom layers.
  1. Register the custom layers as extensions to the Model Optimizer.
  2. For Caffe you register the custom layers as Custom and use the system Caffe to calculate the output shape of each Custom Layer while for TensorFlow you have to replace a subgraph with a known subgraph
- **Some of the potential reasons for handling custom layers:**
  Before a model can be used with Inference Engine, it has to be converted into its Intermediate Representation. This is done using Model Optimizer. Model Optimizer can only work with some known predefined layers before building the model's internal representation, If the model contains some unknown layers, those are referred to custom layers. And these layers have to be converted to what the Inference Engine will be able to work with.

Comparing Model Performance

Performance of the model was carried out using C++ version of OpenVINO benchmark app Asynchronous inference, 1 inference requests using 1 streams for CPU, limits: 60000 ms duration

Model Performance after conversion

	FP16		**FP32**	
Model	Latency	FPS	Latency	FPS
ssd_mobilenet_v2_coco	68.55 ms	14.46 FPS	67.90 ms	14.57 FPS
ssdlite_mobilenet_v2_coco	30.28 ms	32.79 FPS	30.05 ms	33.11 FPS

	FP16		**FP32**	
Model	Count	Duration	Count	Duration
ssd_mobilenet_v2_coco	68.55 ms	14.46 FPS	67.90 ms	14.57 FPS

ssdlite_mobilenet_v2_coco	30.28 ms	32.79 FPS	30.05 ms	33.11 FPS

Model performance before conversion:

Model	Speed (ms)	COCO mAP[^1]
ssd_mobilenet_v2_coco	31	22
ssdlite_mobilenet_v2_coco	27	22

Source: TensorFlow model zoo

Model Size:

	Before Conversion	After Conversion	
Model		**FP16**	**FP32**
ssd_mobilenet_v2_coco	66.5MB	31.2MB	64.2MB
ssdlite_mobilenet_v2_coco	19.0MB	8.5MB	17.1MB

Assess Model Use Cases

1. The model can be used in counting the numbers of people present before and after an incidence: This will help the appropriate bodies involved in assessing the scene carry out a detailed forensic report of what happened and perhaps the potential cause if perhaps it's caused by a person.
2. The model can be used in monitoring numbers of people visiting a recreational center and the average time they spend: This model will help the management of the

recreational center analyze what seems to engage visitors and if they don't stay long, it can give them an indication why they don't stay long.

3. The model can be used in counting numbers of pedestrians: This is useful especially in a traffic control system which needs to make a decision of when to stop the traffic to allow pedestrians to cross the road.

Assess Effects on End User Needs

Lighting, model accuracy, and camera focal length/image size have different effects on a deployed edge model. The potential effects of each of these are:

1. The performance will be optimal if the resources of the end-user device are okay. Or
2. Intensive device resource usage if these parameters are too high for the end-user device resources.