Input Image

Activations from CNN (GoogLeNet)

Faster RCNN

Scene Detector

Object Dectector

Object Location

Input to Language Model