Homework 4

Problem 2

a. Here, we use get_gaussian_scoremap() instead of one-hot encoding because each object can have more than one affordances. One-hot encoding means that the robot's action is not probabilistic, but 100% certain about the grasp.

b. self.aug_pipeline = iaa.Sometimes(0.7, iaa.Affine(
translate_percent="x": (-0.2, 0.2), "y": (-0.2, 0.2),
rotate=(-angle_delta/2,angle_delta/2),
))

This pipeline applies affine transformation on 70% of all rgb images. Each translation is defined to be:
translate/move x and y coordinates by -20% to 20%;
rotate image by 22.5 degrees, ranges = [-12.5, 12.5].

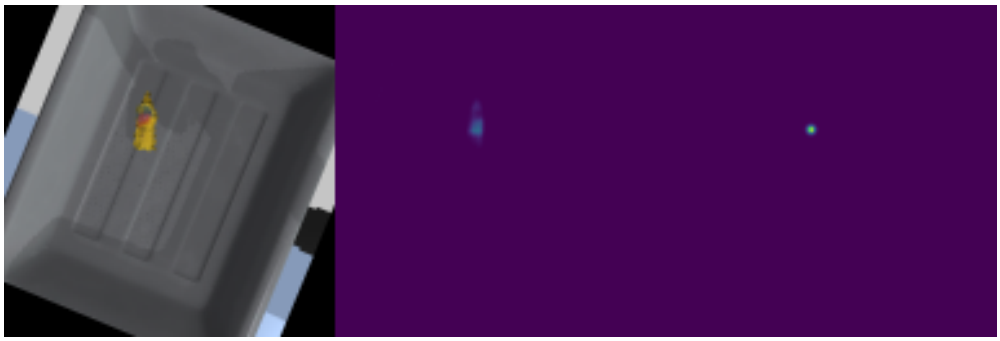
c. nn.BCEWithLogitsLoss is a loss function that combines a Sigmoid layer and the BCELoss. A Sigmoid layer is usually the last layer of many networks, as it's a Sigmoid function that converts the model's output into a probability, ranges between 0 and 1, adding non-linearity and normalize the output. BCEWithLogitsLoss is more numerically stable than applying BCELoss then followed by a Sigmoid layer separately, as we can apply the logsumexp trick in BCEWithLogitsLoss. See a more mathematical explanation here and here. In addition, normalization helps with accuracy if the values have a large range.

d. Start epoch 100
step 693 training loss 0.0011736743617802858
Epoch (100 / 101)

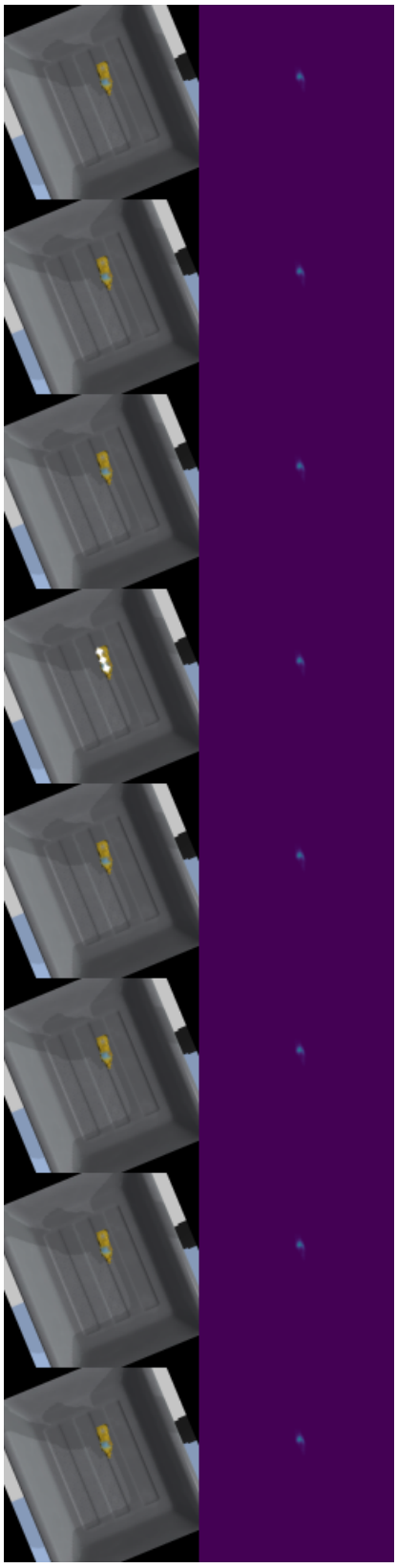
Train loss: 0.0011
Test loss: 0.0011

Start epoch 101
step 700 training loss 0.0010870592668652534
Epoch (101 / 101)

Train loss: 0.0011
Test loss: 0.0011



f. success rate = 73.33%



See video here on YouTube

g. 2 objects are left in the bin.
The method is sample efficient because of transitional equivalence. It argues that the position of the object (detection target) should not be fixed in order for the CNN to detect it. We don't need to feed the CNN all possible images of the objects in all possible positions/angles because the gripper will follow the object, using Gaussian map to locate the output. Here, all weights are considered the same, meaning that an image and all its transitions are equivalent, so they should have equivalent translation in outputs. For example, a drill that is rotated and moved to another coordinate in the box should still be a drill object, and the gripper should adjust/perform equivalent translation on its grasp angle and coordinates. Transitional equivalence improves the CNN in generalization abilities, when the data set size is limited.

See more here.

See video here on YouTube

Problem 3

a. Start epoch 200
step 1393 training loss 0.015717506408691406
Epoch (200 / 201)

Train loss: 0.0134
Test loss: 0.0180

Start epoch 201
step 1400 training loss 0.011761422269046307
Epoch (201 / 201)

Train loss: 0.0108
Test loss: 0.0179

b. success rate = 13.33%

See video here on YouTube

c. Only one item was successfully moved to the other bin. 14 items left in the first bin.
Action regression performs worse than visual affordance because the former doesn't use Gaussian prediction to normalize and produce probabilistic output. Regression is will very likely to mis-classify due to very limited data set, which significantly impacts the performance because visual affordance (sample efficient) is not used.

See video here on YouTube

Problem 4

1. Modifications made:
I noticed that in producing the labels in problem 1, it was difficult to record a

good grasp angle, as it increments by 22.5 degrees. I changed the increment to 180/16 = 11.25 degrees, then relabelled and retrained the CNN.

2. Hypothesis explained:
   With finer grasp angles to train the model on, the robot should have more precise grasp, especially on smaller objects with irregular shapes, such as clamp, and larger objects that are in tricky positions, such as in the bin corners.

   Hypothesis: enabling finer and smaller grasp angles in label production will improve robot grasp success rate.

3. Results:
   Train and test losses:
   Start epoch 100
   step 1386 training loss 0.001310884952545166
   Epoch (100 / 101)

   Train loss: 0.0013
   Test loss: 0.0013

   Start epoch 101
   step 1400 training loss 0.001502185594290495
   Epoch (101 / 101)

   Train loss: 0.0013
   Test loss: 0.0013

   Number of objects left in the bin: 2

   Even though my attempt did not improve the performance of the robot, from the video, I observed the grasping of smaller objects was from a more "reasonable" angle. Comparing the video recordings of the affordance and affordance_improved, I observed the following improvements that can't be reflected quantitatively:
   1. The robot can have better grasp angle for objects that are near each other. 2. The robot can have better grasp angle for objects at the bin corner or sides. For example, in affordance_improved, the drill and tennis ball were on the side of the bin, which would have been a difficult task (requires 1+ grasp attempts) for the robot, but here, the robot was able to pick up both objects with 1 attempt each. 3. The robot completes tasks with fewer attempts, especially with smaller and irregular objects. The improved robot move 13 objects with 19 attempts, but the default robot used 23 attempts. For scissors, clamp, and strawberry, the robot was able to grasp and successfully move them on its first attempt.

Analysis of why no improvement was seen in robot performance:
With finer angles to train, this might also have introduced more noise. In addition, the labelling size may needs to be larger to see improvements.

See video here on YouTube