Label-Free Object Detection in Videos of Physical Interactions

Alex Barron, Todor Markov, and Zack Swafford

Introduction

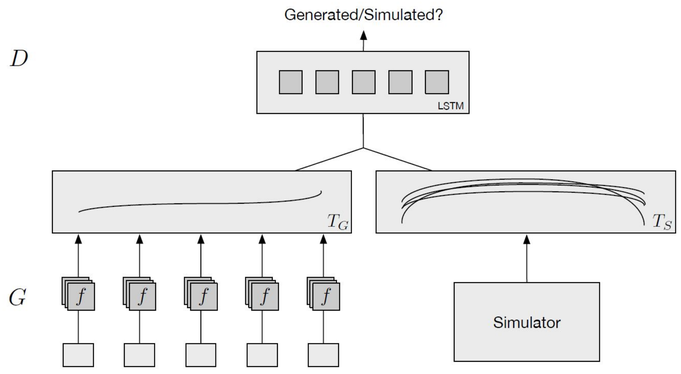
Well-labeled object tracking datasets, such as videos with object locations annotated, are hard to come by and expensive to create. It is therefore very helpful to develop algorithms that track objects in a unsupervised or semi-supervised way.

Implementation

Our method to detect and track objects in video utilizes a GAN in which the generator network extracts an object's trajectory from video data and the discriminator network determines if the trajectory seen was given by the generator network or drawn from a prespecified simulator. Once trained, the generator is able to determine object trajectory unsupervised from unlabeled videos, as desired.

Yet unpublished work from Stewart and Ermon demonstrates that this approach is legitimate in select cases; we have extended their examples here to other physical interactions such as real videos of thrown balls effected by gravity or simulated videos of balls bouncing in a Pong-like environment.

Model



We use a vanilla GAN with the architecture suggested by Stewart and Ermon. The generator is a simple CNN with 2 Conv-Relu-Pool-BatchNorm layers followed by a dense output layer which produces the desired 1 or 2 dimensional trajectory. The discriminator uses an LSTM and linear output layer to tell the simulated and generated trajectories apart.

To balance the power of the discriminator and generator, if the loss exceeds 1.35 or 0.75 respectively, we repeat the training step up to 10 times to equalize the losses. Label smoothing also aids the stability of training which we perform only on the positive discriminator labels by adding a normal distribution with standard deviation 0.2. We also intend to experiment with more stable GAN architectures such as LS-GAN.

Results

We see below a series of five stills from a representative video of a ball bouncing (in white) overlayed with our trained GAN's annotation (in yellow). Note that the labeling follows a reflection of the ball's true trajectory almost exactly--so it is a valid trajectory, although not precisely the desired one.

Such results are much as expected because we imposed no loss on following the true trajectory of the ball; the generator has learned to fool the discriminator by creating a real-looking trajectory inspired by the ball but not actually following the ball itself.

This aspect of the generator could easily be improved by incorporating into the final loss a term based on any of various segmentation algorithms, which would then give the net an understanding of the physical boundaries and existence of distinct objects whose trajectory is to be tracked. However, this approach becomes less practical when the generator itself would be relied upon more heavily to distinguish objects, e.g. in cases where a vanilla segmentation algorithm could not be applied.

Conclusion

Our results so far imply that carefully specifying the physics of the system is critical for achieving success using the label-free approach. This raises some questions regarding the scalability of this approach - more complicated environments will require more complicated simulators. Still, the label-free approach might be useful for particular cases where labeled data is extremely expensive or unavailable.