Learning Correspondence from the Cycle-consistency of Time

Purvanshi Mehta

November 15, 2019

1 Summary

1.1 Problem

Learning representations for visual correspondence from pixel-wise to object-level have been widely explored within supervised learning settings which has relied highly on human annotations. The paper aims to learn representations that support reasoning at various levels from scratch without human supervision.

1.2 Innovation

The key idea for self supervision is that unlimited supervision can be obtained by tracking backward and the forward and using the inconsistency between the start and end points as the loss function.

1.3 Contributions

The overall goal of the problem can be formulated as learning a feature space by tracking a patch, extracted from an image I_t backwards and then forward in time while minimizing the cycle-consistency loss. Learning the feature space is determined by tan operation which takes as inputs the features of a current patch and a target image and returns the image feature region with maximum similarity.

1.4 Evaluation

Experimentation has been shown on various tasks like instance propagation(DAVIS - 2017), pose keypoint propagation(JHMDB), semantic and instance propagation(VIP), texture propagation and video frame reconstructions. Their model seems to improve performance as training continues and more data is provided but the model converges asymptotically in terms of training. The evaluation has been done using various techniques- optical flow, DeepCluster etc. Their method clearly outperforms the rest of the methods on all experiments.

1.5 Substantiation

The authors claim to learn representations that support various levels of visual correspondence in a unsupervised manner. The experimentation is extensive in terms of types of methods used to compare their results with and the number of domains covered. The authors also emphasize the importance of deciding what to track during training time. Tracking a patch containing two objects in it which actually diverge in the next frame is non trivial.