Temporally Coherent Embeddings for Self-Supervised Video Representation Learning

Joshua Knights, Ben Harwood, Daniel Ward, Anthony Vanderkop, Olivia Mackenzie-Ross, Peyman Moghadam
Robotics and Autonomous Systems, Data61 CSIRO, Brisbane, QLD 4069, Australia
{firstname.lastname}@csiro.au

In this paper we present TCE: A method for self-supervised learning from unlabelled video data. Mimicking the smoothness of the real world, we enforce similarity between nearby frames and dissimilarity between videos to create a temporally coherent embedding space with a 2D-CNN backbone. We demonstrate the downstream benefits of our approach by achieving state-of-the-art results across multiple action recognition datasets, with a top-1 accuracy of 71.2% on UCF101 and 36.6% on HMDB51.

Motivation

- A major bottleneck for the performance of ML models is a lack of labelled data for training.
- We believe that in the same way that the real world is temporally smooth, a strong pre-trained embedding should also demonstrate smooth behaviour over time.

Proposed Model

We use a contrastive loss function to enforce similarity between nearby frames, while encouraging dissimilarity between frames from other videos.

In addition, we implement a hard negative mining approach to find increasingly difficult negative examples as training progresses.

As training progresses, the network will select potential negatives that exist closer and closer to the positive examples in the embedding space.

Results

Ablation: Hard Negative Mining

We propose TCE, a self-supervised approach to learning from unlabelled video data, exploiting the inherent smoothness of the real world:

- We achieve state-of-the-art results for HMDB51, and for UCF101 approaches pre-trained on UCF101.
- TCE achieves results on-par or superior to current action recognition state-of-the-art.
- We demonstrate strong spatio-temporal features can be learnt by 2D CNNs, given appropriate formulation of the pre-training loss.
- Hard negative mining approach ameliorates vanishing gradient issue selecting negatives on large unlabelled datasets.

Conclusion