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

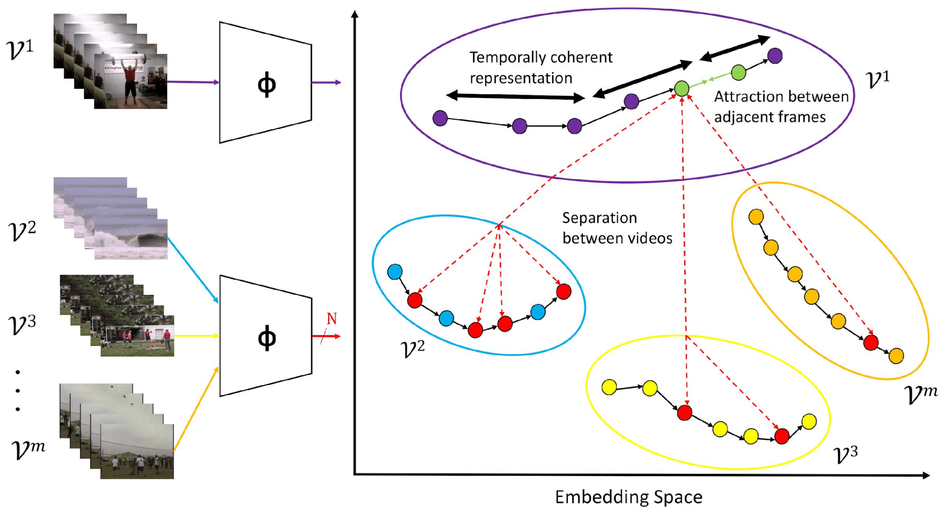

Figure 1: *TCE* uses a contrastive loss to enforce similarity between nearby frames, while encouraging dissimilarity to frames from other videos

- We use a contrastive loss function to enforce similarity between nearby frames, and dissimilarity between frames from different video sequences
- 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

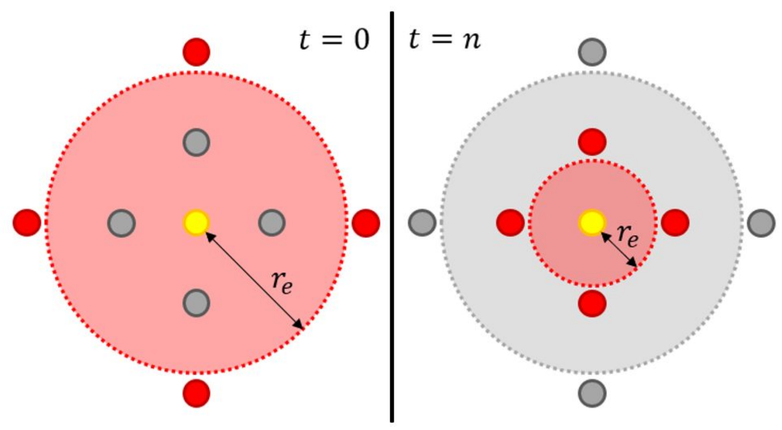

Figure 2: As our training progresses, we allow the network to train using hard negatives closer and closer to the contrastive anchor

Results

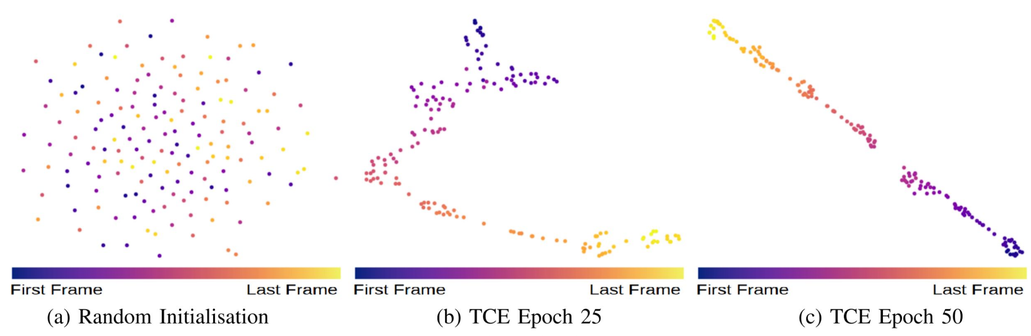

Figure 3: *TCE* produces an increasingly smooth embedding space as training progresses

Method	Backbone	Params	2D-CNN	Pre-Training	UCF101(%)	HMDB51(%)
3DRotNet [34]	3D ResNet-18	34×10^6	✗	Kinetics400	62.9	33.7
3DCubicPuzzles [10]	3D ResNet-18	34×10^6	✗	Kinetics400	65.8	33.7
DPC [5]	3D ResNet-18†	14×10^6	✗	Kinetics400	68.2	34.5
TCE (Ours)	2D ResNet-18	11×10^6	✓	Kinetics400	68.8+	34.2
TCE (Ours)	2D ResNet-50	23×10^6	✓	Kinetics400	71.2	**36.6**
DPC [5]	3D ResNet-34†	33×10^6	✗	Kinetics400	**75.7**	35.7
Motion & Appearance [35]	C3D	11×10^6	✗	UCF101	48.6	20.3
Shuffle and Learn [12]	AlexNet	61×10^6	✓	UCF101	50.9+	19.8
VideoGAN [4]	C3D	11×10^6	✗	UCF101	52.1	-
Arrow of time [36]	AlexNet	61×10^6	✓	UCF101	55.3	-
CMC [17]	CaffeNet ×2*	$58 \times 10^6 \times 2$	✓	UCF101	55.3	-
OPN [11]	VGG-M-2048	8.6×10^6	✓	UCF101	59.8	23.8
DPC [5]	3D ResNet-18†	14×10^6	✗	UCF101	60.6+	-
Skip-Clip [8]	3D ResNet-18	34×10^6	✗	UCF101	64.4+	-
Video Clip Ordering [13]	R3D	14×10^6	✗	UCF101	64.9+	29.5
***TCE* (Ours)**	2D ResNet-18	11×10^6	✓	UCF101	**68.2+**	**31.7**

Table 1: Comparison between *TCE* and other state-of-the-art approaches

Ablation: Hard Negative Mining

Figure 5: UCF101 Top-1 Accuracy for *TCE*. Results reported on the first split of UCF101 using a ResNet18 backbone.

Conclusion

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

FOR FURTHER INFORMATION
Joshua Knights
joshua.knights@csiro.au

REFERENCES
See original paper for references corresponding to Table 1