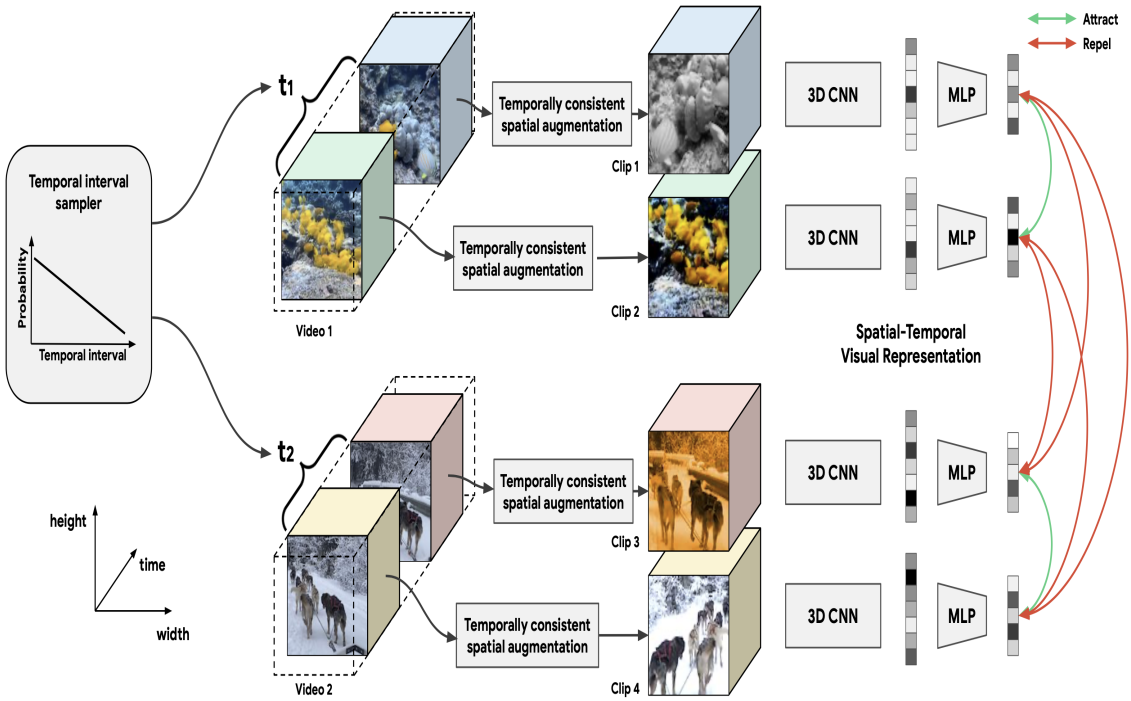
Spatiotemporal Contrastive Video Representation Learning

arXiv Paper arXiv.2008.03800

This repository is the official TF2 implementation of Spatiotemporal Contrastive Video Representation Learning.



Description

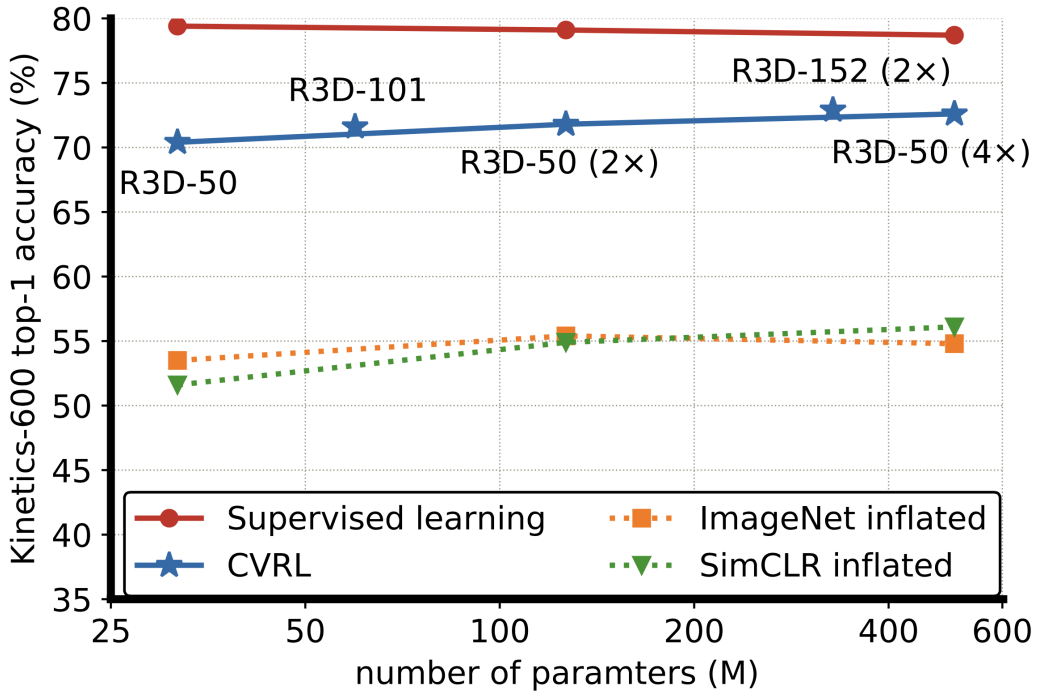
We present a self-supervised Contrastive Video Representation Learning (CVRL) method to learn spatiotemporal visual representations from unlabeled videos. Our representations are learned using a contrastive loss, where two augmented clips from the same short video are pulled together in the embedding space, while clips from different videos are pushed away. CVRL significantly closes the gap between unsupervised and supervised video representation learning.

We release the code and pre-trained models.

More pre-trained model checkpoints and a detailed instruction about the code will be updated.

Experimental Results

Kinetics-600 top-1 linear classification accuracy



Pre-trained Model Checkpoints

We provide model checkpoints pre-trained on unlabeled RGB videos from Kinetics-400 and Kinetics-600. All models are trained scratch with random initialization.

We also provide a baseline model checkpoint of "ImageNet inflated" we used in the paper. The model has the same architecture as 3D-ResNet-50 (R3D-50), with model weights inflated from a 2D ResNet-50 pre-trained on ImageNet.

Model	Parameters	Dataset	Epochs	K400 Linear Eval.	K600 Linear Eval.	Checkpoint
R3D-50 (1x)	31.7M	ImageNet	-	53.5%	54.7%	ckpt (127 MB)
R3D-50 (1x)	31.7M	Kinetics-400	200	63.8%	-	ckpt (127 MB)
R3D-50 (1x)	31.7M	Kinetics-400	800	66.1%	-	ckpt (127 MB)
R3D-50 (1x)	31.7M	Kinetics-600	800	68.5%	70.4%	ckpt (127 MB)

Citation

```
@inproceedings{qian2021spatiotemporal,
  title={Spatiotemporal contrastive video representation learning},
  author={Qian, Rui and Meng, Tianjian and Gong, Boqing and Yang, Ming-Hsuan and Wang, Huisheng and Belongie, Serge and Cui, Yin},
```

```
  booktitle={CVPR},
  year={2021}
}
```