Space-Time Crop and Attend: Improving Cross-Modal Video Representation Learning

Mandela Patrick*, Yuki M. Asano*, Po-Yao Huang*, Ishan Misra, Florian Metze, Joao Henriques, Andrea Vedaldi

UNIVERSITY OF OXFORD FACEBOOK 2021 ICCV OCTOBER 11-17 VIRTUAL

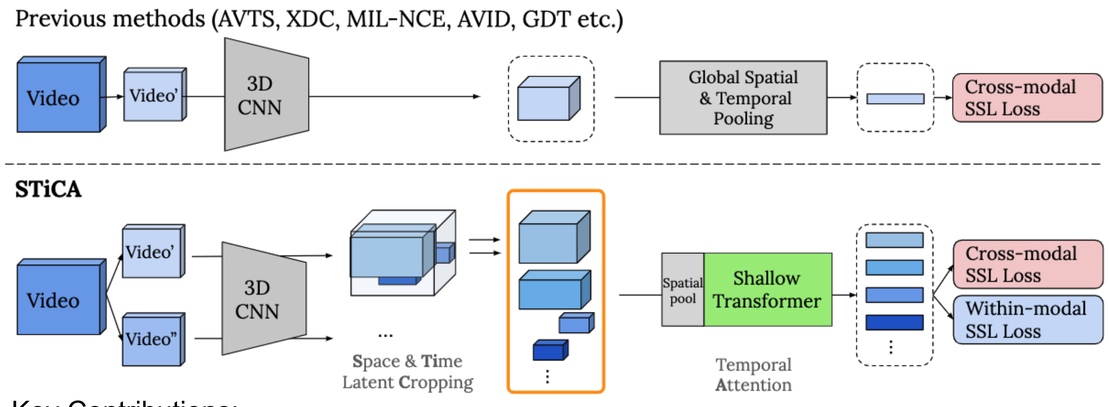
Motivation

We improve audio-visual self-supervised learning in two ways unique to the spatio-temporal aspect of videos. First, for space, we show that spatial augmentations such as cropping do work well for videos too, but that previous implementations, due to the high processing and memory cost, could not do this at a scale sufficient for it to work well. To address this issue, we first introduce Feature Crop, a method to simulate such augmentations much more efficiently directly in feature space. Second, we show that as opposed to naive average pooling, the use of transformer-based attention improves performance significantly, and is well suited for processing feature crops. Combining both of our discoveries into a new method, Space-Time Crop and Attend (STiCA) we achieve state-of-the-art performance across multiple video-representation learning benchmarks. In particular, we achieve new state-of-the-art accuracies of 67.0% on HMDB-51 and 93.1% on UCF-101 when pre-training on Kinetics-400.

Code and pre-trained models:

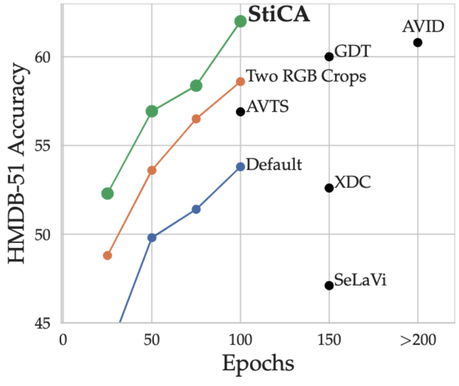
https://github.com/facebookresearch/GDT

STiCA: Space-Time Crop and Attend

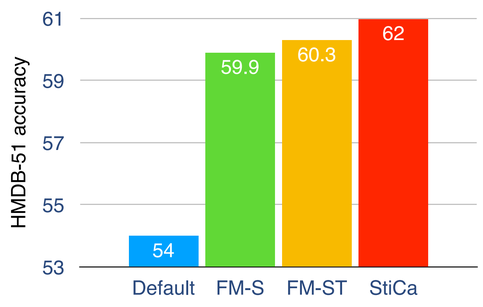


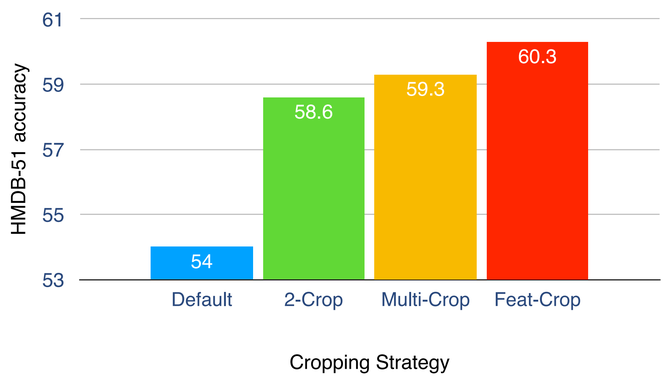
Key Contributions:
1. Feature-Crop Augmentation
2. Transformer for Late Temporal Attention Modelling

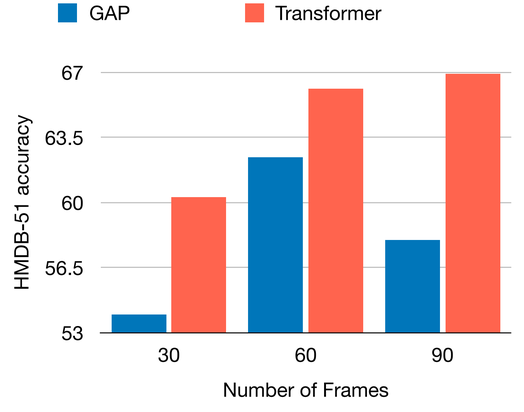
Performance Highlight: HMDB-51 accuracy vs epoch



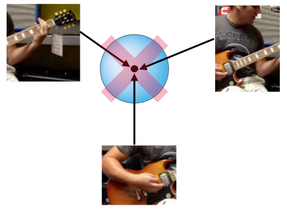
Analysis



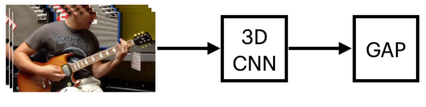




Problems in Multi-Modal Video Contrastive Learning

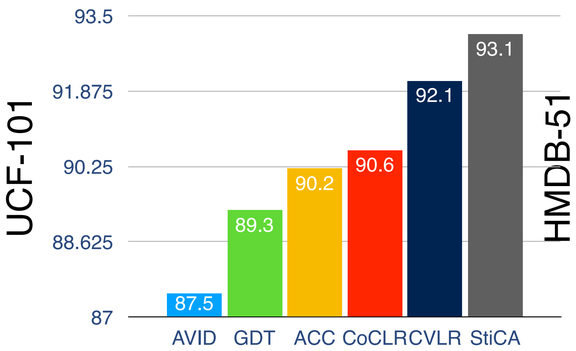


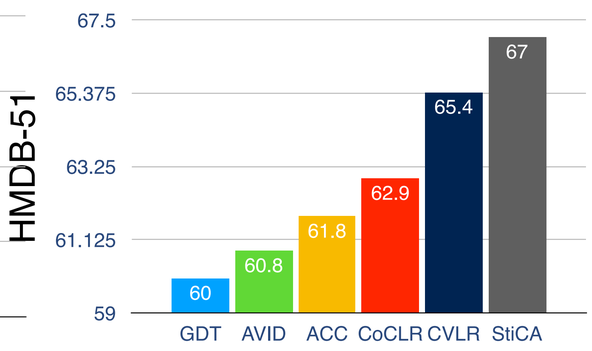
Within modal spatial invariance are not learned.



High-level temporal information is discarded.

Comparison to state-of-the-art





Reference:

Hu et al, “Deep multimodal clustering for unsupervised audiovisual learning”
Ma et al, “Learning audio-visual representations with active contrastive coding”
Patrick et al, “Multi-modal self-supervision from generalized data transformations”
Qian et al, “Spatiotemporal contrastive video representation learning”
Han et al, “Self-supervised co-training for video representation learning”
Morgado et al, “Audio-visual instance discrimination with cross-modal agreement”

Acknowledgements:

We are grateful for support from the Rhodes Trust (M.P.), Facebook (M.P.), EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines \& Systems [EP/L015897/1] (M.P. and Y.A.), the Qualcomm Fellowship (Y.A.), and the Royal Academy of Engineering under the Research Fellowship scheme (J.F.H.).
This work is also supported by the DARPA grants funded under the AIDA program (FA8750-18-2-0018) and the GAILA program (award HR00111990063) (P.H.).
We also thank Tengda Han for helpful discussions and feedback.