Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning

AJ Piergiovanni · Weicheng Kuo · Anelia Angelova · CVPR (2023)

Download Google Scholar Copy Bibtex

Abstract

We present a simple approach which can turn a ViT encoder into an efficient video model, which can seamlessly work with both image and video inputs. By sparsely sampling the inputs, the model is able to do training and inference from both inputs. The model is easily scalable and can be adapted to large-scale pre-trained ViTs without requiring full finetuning. The model achieves SOTA results.

Research Areas

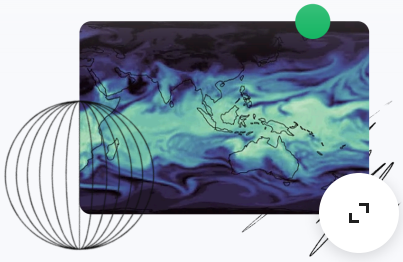
Machine perception

Research

Meet the teams driving innovation

Our teams advance the state of the art through research, systems engineering, and collaboration across Google.

See our teams



Follow us

Explore our other initiatives

Google AI

Discover how Google AI is committed to enriching knowledge and solving complex challenges

- Products
- Build
- Research
- Responsibility
- Societal Impact
- About

Google Cloud

High-performance infrastructure for cloud computing, data analytics & machine learning

- Overview
- Solutions
- Products
- Pricing
- Resources

Google DeepMind

Our mission is to build AI responsibly to benefit humanity

Google Labs

Explore the future of AI responsibly with Google Labs