SlowFast Networks for Video Recognition

Conference paper

C. Feichtenhofer [1] H. Fan [1] J. Malik [2] K. He [1]

[1]Facebook AI Research (FAIR)

[2]Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

IEEE Conference on Computer Vision and Pattern Recognition, June 2020

Introduction

- Paper published in 2020 by researchers at Facebook AI Research
- Proposes a novel architecture for video recognition
- Takes into account the fact that videos often contain both slow and fast components

Main Contributions

- Two-stream architecture with a slow pathway and a fast pathway
- Slow pathway captures high-level features that are invariant to slow changes
- Fast pathway captures motion information and other fast changes
- Outputs of the two pathways are fused at a later stage to make a final prediction

Motivation

- Videos often contain both slow and fast components
- These components are not equally important for recognition
- Slow components are often invariant across frames, while fast components are highly variable
- A network architecture that can capture slow and fast information separately could be more effective

Architecture

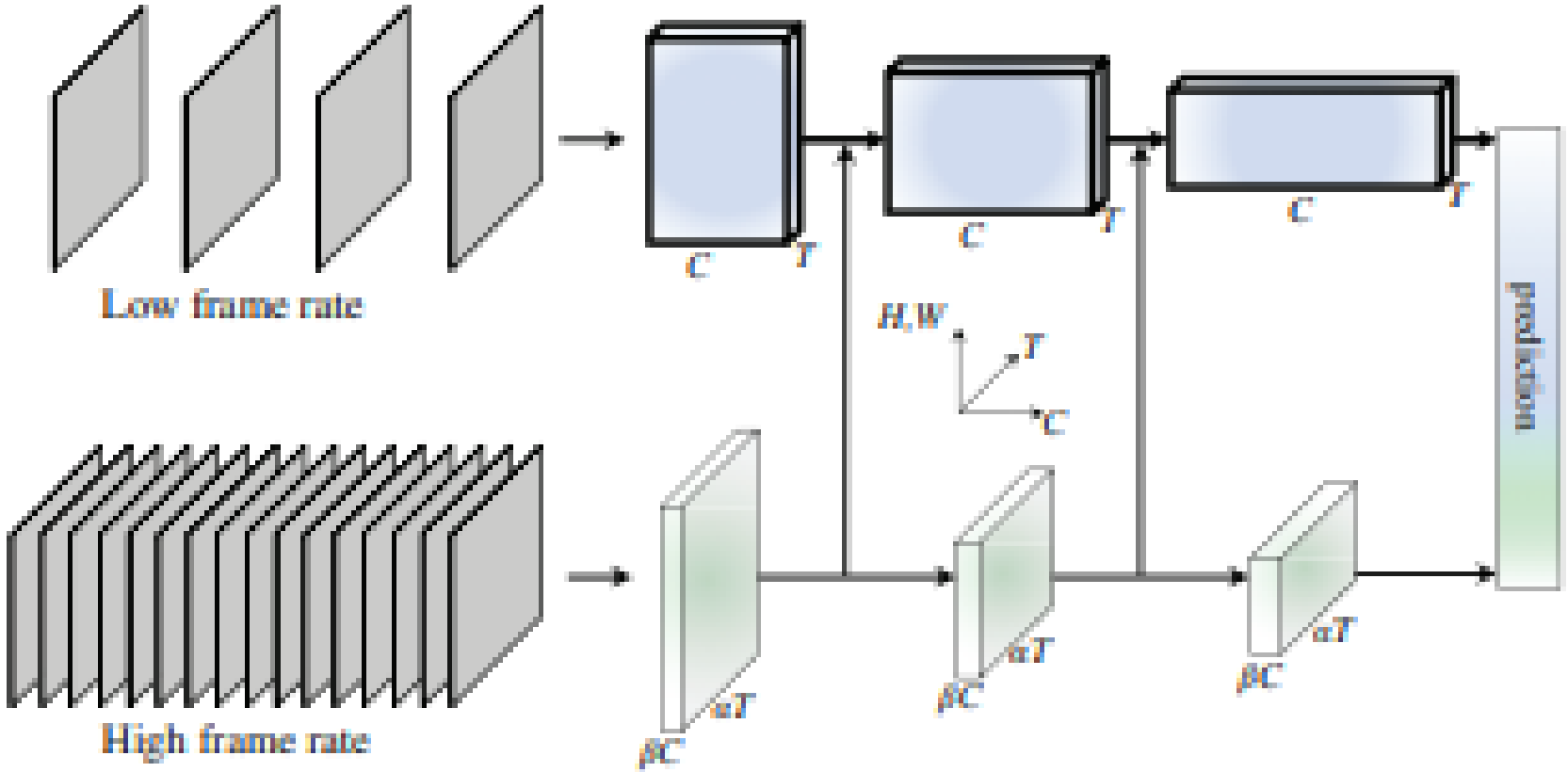

Figure 1. **A SlowFast network** has a low frame rate, low temporal resolution *Slow* pathway and a high frame rate, $\alpha\times$ higher temporal resolution *Fast* pathway. The Fast pathway is lightweight by using a fraction (β, e.g., 1/8) of channels. Lateral connections fuse them.

Technical Details

- ResNet-like backbone for both slow and fast pathways
- Slow pathway uses a low frame rate of 8 frames per second
- Fast pathway uses a high frame rate of 32 frames per second
- Pathways are fused at a later stage using a channel-wise concatenation operation

Training Strategy

- Pre-training the slow pathway
- Fine-tuning the entire network using a combination of slow and fast inputs

Results

model	flow	pretrain	top-1	top-5	GFLOPs × views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30

Implications

- Significant advance in the field of video recognition
- Has important applications in fields such as surveillance, autonomous driving, and human-computer interaction

Conclusion

- SlowFast network architecture is a significant step forward in the field of action recognition
- Takes a novel approach to capturing slow and fast information in videos
- Results demonstrate its effectiveness and efficiency