SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He

In standard image recognition tasks, the spatial dimensions are treaded symmetrically. However, for video understanding tasks, all spatiotemporal orientations are not equally likely, so there is no reason to treat space and time the same way. This can be understood by seeing how spatial and temporal content varies differently. The spatial semantics evolve slowly (the identity of hands does not change in the span of an action); while the motion being performed evolves much faster than the object's semantics. Based on this intuition, the paper proposes a two-pathway **SlowFast** model, where the slow path operates over low frame rates with a slowly refreshing speed, while the fast pathway operates over high frame rates while having fewer channels and weaker ability to model spatial information. The two pathways are fused by lateral connection along the paths at the end of each block.

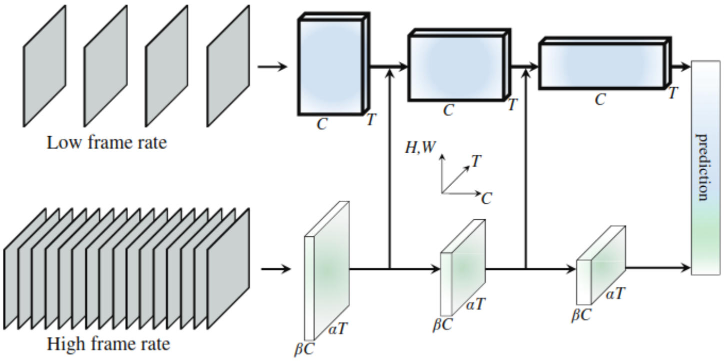
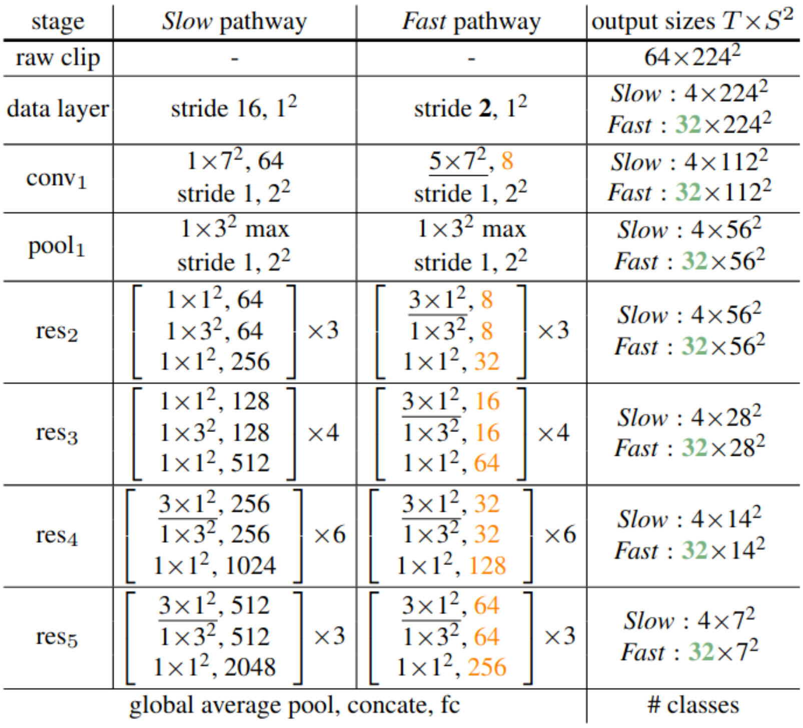


Figure 1. **A SlowFast network** has a low frame rate, low temporal resolution *Slow* pathway and a high frame rate, $\alpha\times$ higher temporal resolution *Fast* pathway. The Fast pathway is lightweight by using a fraction (β, *e.g.*, 1/8) of channels. Lateral connections fuse them.

The slow pathway takes as input T frames from a clip of T x N frames, so the input to the slow pathway has a temporal sampling stride of N. For example, for a video with 30 fps, N equals 16 in order to roughly sample two frames per second. On the other hand, the fast pathway extracts M x T frames from all of N x T frames, if M = N, then all of the input frames are considered. In the paper they consider M = 8, so for a 30fps video, the input is roughly 16 frames, sampling one frame from two input ones. However, in order to reduce the computation, the channel dimension is reduced by a factor of B, which generally equals 1 / M. This way, when we want to fuse the fast pathway into the slow pathway, we can reshape the inputs from (C/B) x (TxM) x H x W into (C/B x M) x T x H x W; followed by a sum (Time-to-channel). Other possible lateral connections are also considered: Time-strided sampling, where we sample T frames from M x T ones followed by concatenation, or Time-strided convolution with a stride of M and output channels of 2BC. They all perform roughly the same, with Time-strided convolution performing slightly better.



stage	*Slow* pathway	*Fast* pathway	output sizes $T\times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv$_1$	1×7^2, 64 stride 1, 2^2	$\underline{5\times7^2}$, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool$_1$	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res$_2$	$\begin{bmatrix}1\times1^2, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	$\begin{bmatrix}\underline{3\times1^2}, 8\\1\times3^2, 8\\1\times1^2, 32\end{bmatrix}\times3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res$_3$	$\begin{bmatrix}1\times1^2, 128\\1\times3^2, 128\\1\times1^2, 512\end{bmatrix}\times4$	$\begin{bmatrix}\underline{3\times1^2}, 16\\1\times3^2, 16\\1\times1^2, 64\end{bmatrix}\times4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res$_4$	$\begin{bmatrix}\underline{3\times1^2}, 256\\1\times3^2, 256\\1\times1^2, 1024\end{bmatrix}\times6$	$\begin{bmatrix}\underline{3\times1^2}, 32\\1\times3^2, 32\\1\times1^2, 128\end{bmatrix}\times6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res$_5$	$\begin{bmatrix}\underline{3\times1^2}, 512\\1\times3^2, 512\\1\times1^2, 2048\end{bmatrix}\times3$	$\begin{bmatrix}\underline{3\times1^2}, 64\\1\times3^2, 64\\1\times1^2, 256\end{bmatrix}\times3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

The final results show that by adding the fast pathway, the model is able to better recognize actions with high-speed motions like hand clapping.