Video Swin Transformer

arXiv Jun, 2021
Outline

- Introduction
- Related Works
- Architecture
  - 3D Shifted Window-MSA Module
  - Pre-Training
  - Variants
- Experiments
- Conclusion
Introduction

- **Video Swin**
  - Pure transformer backbone architecture for video

- **Adapts Swin architecture to videos**
  - Takes advantage of spatio-temporal locality in videos
  - Swin, extended to the temporal domain
Related Works

- **CNNs**
  - C3D, I3D - 3D Convolutions
  - P3D - Disentangle spatial and temporal convolutions

- **Transformers to Compliment CNNs**
  - NLNet
  - GCNet
  - DNL

- **Vision Transformers**
  - Swin Transformer: Uses locality in images, general purpose backbone for image recognition.
  - TimeSformer: Examines variants of space-time attention; proposes factorized attention
  - ViViT: Examines four factorized spatial and temporal attentions
  - MViT: Reduces computation by pooling attention
Input includes a time dimension: number of Frames

Patches are 3D
  - Size $2 \times 4 \times 4 \times 3$ (Frames x Height x Width x Channels)

Clips are broken up into $(T/2)*(H/4)*(W/4)$ 3D tokens

Each patch token consists of a 96-dimensional features, up from 48

No downsampling along the temporal dimension
  - Spatial dimension still downsampled the same

Replaces MSA module with a 3D shifted window based MSA module
Swin

\[ H \times W \times 3 \]
Images

\[ \frac{H}{4} \times \frac{W}{4} \times 48 \]
Patch Partition

\[ \frac{H}{4} \times \frac{W}{4} \times C \]
Linear Embedding

\[ \times 2 \]
Swin Transformer Block

\[ \frac{H}{8} \times \frac{W}{8} \times 2C \]
Patch Merging

\[ \frac{H}{16} \times \frac{W}{16} \times 4C \]
Swin Transformer Block

\[ \times 6 \]
Patch Merging

\[ \frac{H}{32} \times \frac{W}{32} \times 8C \]
Swin Transformer Block

Stage 1

Stage 2

Stage 3

Stage 4

Video Swin

\[ T \times H \times W \times 3 \]
Videos

\[ \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times 96 \]
3D Patch Partition

\[ \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times C \]
Linear Embedding

\[ \times 2 \]
Video Swin Transformer Block

\[ \frac{T}{2} \times \frac{H}{8} \times \frac{W}{8} \times 2C \]
Patch Merging

\[ \frac{T}{2} \times \frac{H}{16} \times \frac{W}{16} \times 4C \]
Video Swin Transformer Block

\[ \times 6 \]
Patch Merging

\[ \frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8C \]
Video Swin Transformer Block

Stage 1

Stage 2

Stage 3

Stage 4
Swin vs Video Swin Transformer Block Architecture

**SWIN**

- MLP
- LN
- W-MSA
- LN

**Video SWIN**

- MLP
- LN
- 3D W-MSA
- LN
Window Construction

Token

Window

Configuration

(2 x 4 x 4 x 3) (T x H x W x RGB)

(4 x 4 x 4)

(2 x 2 x 2)
Layer 1

Layer l+1

A local window to perform self-attention

A patch

Layer 1

Layer l+1

3D tokens: $T' \times H' \times W' = 8 \times 8 \times 8$
Window size: $P \times M \times M = 4 \times 4 \times 4$

Layer 1

# window: $2 \times 2 \times 2 = 8$

Layer l+1

# window: $3 \times 3 \times 3 = 27$

3D local window to perform self-attention

A token
3D Shifted Window MSA Module

- Global Self-Attention would be too costly for videos

- Solution, as in Swin Transformer:
  - Video Swin advocates for using the spatiotemporal locality of videos
    - Pixels closer in spatiotemporal distance are more likely correlated
Multi-head self-attention
(Non-overlapping 3D Windows)

$T' \times H' \times W'$ 3D tokens and a 3D window size of $P \times M \times M$

- $T'$ is # of 3D tokens in the temporal dimension
- $H'$ is # of 3D tokens high
- $W'$ is # of 3D tokens wide

$\left\lfloor \frac{T'}{P} \right\rfloor \times \left\lfloor \frac{H'}{M} \right\rfloor \times \left\lfloor \frac{W'}{M} \right\rfloor$ non-overlapping 3D windows
3D Shifted Windows

- MSA is applied within each non-overlapping 3D window
  - There are no connection between windows

- Shifted windows allow for cross-window connections
  - Maintains computational efficiency
3D Shifted Windows (Cont.)

We shift by \((P/2, M/2, M/2)\) tokens from that of the preceding layer.

Layer 1
# window: 2x2x2=8

Layer l+1
# window: 3x3x3=27
Shift by \((P/2, M/2, M/2)\)
Shift by \(( P / 2 , M / 2 , M / 2 )\)
Shift by \( \left( \frac{P}{2}, \frac{M}{2}, \frac{M}{2} \right) \)
Shift by \((P/2, M/2, M/2)\)

(You will not see the time shift on a 2D space!)
Efficient Batch Computation
(Maintained From Swin Transformer)
Two consecutive Video Swin Transformer Blocks are computed as:

\[
\hat{z}^l = \text{3DW-MSA} \left( \text{LN} \left( z^{l-1} \right) \right) + z^{l-1}, \\
z^l = \text{FFN} \left( \text{LN} \left( \hat{z}^l \right) \right) + \hat{z}^l, \\
\hat{z}^{l+1} = \text{3DSW-MSA} \left( \text{LN} \left( z^l \right) \right) + z^l, \\
\tilde{z}^{l+1} = \text{FFN} \left( \text{LN} \left( \tilde{z}^{l+1} \right) \right) + \tilde{z}^{l+1},
\]

This design introduces connections between neighboring non-overlapping windows.
\[ \hat{\mathbf{z}}^l = 3\text{DW-MSA} \left( \text{LN} \left( \mathbf{z}^{l-1} \right) \right) + \mathbf{z}^{l-1}, \]
\[ \mathbf{z}^l = \text{FFN} \left( \text{LN} \left( \hat{\mathbf{z}}^l \right) \right) + \hat{\mathbf{z}}^l, \]
\[ \hat{\mathbf{z}}^{l+1} = 3\text{DSW-MSA} \left( \text{LN} \left( \mathbf{z}^l \right) \right) + \mathbf{z}^l, \]
\[ \mathbf{z}^{l+1} = \text{FFN} \left( \text{LN} \left( \hat{\mathbf{z}}^{l+1} \right) \right) + \hat{\mathbf{z}}^{l+1}, \]
3D Relative Position Bias

\[ \text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d + B})V, \]

Matrix B: Size \((PM^2)^2\)

Matrix B\(^{\wedge}\): Size \((2P-1) \times (2M-1)^2\)

- \(PM^2 = \) tokens in a window
### Initialization from Pre-trained Model

<table>
<thead>
<tr>
<th>Mismatched Component</th>
<th>Swin 2D</th>
<th>Swin 3D</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Embedding Layer</td>
<td>48xC</td>
<td>96xC</td>
<td>Duplicate original matrix values, then rescale by 0.5 to fix mean and variance</td>
</tr>
<tr>
<td>Relative Position Bias Matrix</td>
<td>(2M-1, 2M-1)</td>
<td>(2P-1, 2M-1, 2M-1)</td>
<td>Duplicate original spatial matrix values along temporal dimension.</td>
</tr>
</tbody>
</table>
Ablation on Pretrained Initialization Methods

- **Linear Embedding Layer**
  - Inflated = average values across time
  - Center = copy values to middle frame
  - No significant change

- **Relative Position Bias**
  - Duplicate = copy values across time
  - Center* = copy values to middle frame, mask rest
  - No change at all
# Architecture Variants

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Scale</th>
<th>Linear Embedding Dimension</th>
<th>Transformers per stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swin-T</td>
<td>0.25x</td>
<td>96</td>
<td>2, 2, 6, 2</td>
</tr>
<tr>
<td>Swin-S</td>
<td>0.5x</td>
<td>96</td>
<td>2, 2, 18, 2</td>
</tr>
<tr>
<td><strong>Swin-B</strong></td>
<td>1.0x</td>
<td>128</td>
<td>2, 2, 18, 2</td>
</tr>
<tr>
<td>Swin-L</td>
<td>1.5x</td>
<td>192</td>
<td>2, 2, 18, 2</td>
</tr>
</tbody>
</table>
Experiments

Datasets

- Kinetics-400
- Kinetics-600
- Something-Something v2
<table>
<thead>
<tr>
<th>Method</th>
<th>Pretrain</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Views</th>
<th>FLOPs</th>
<th>Param</th>
</tr>
</thead>
<tbody>
<tr>
<td>MViT-B, 32x3 [10]</td>
<td>-</td>
<td>80.2</td>
<td>94.4</td>
<td>1 x 5</td>
<td>170</td>
<td>36.6</td>
</tr>
<tr>
<td>MViT-B, 64x3 [10]</td>
<td>-</td>
<td>81.2</td>
<td>95.1</td>
<td>3 x 3</td>
<td>455</td>
<td>36.6</td>
</tr>
<tr>
<td>TimeSformer-L [3]</td>
<td>ImageNet-21K</td>
<td>80.7</td>
<td>94.7</td>
<td>1 x 3</td>
<td>2380</td>
<td>121.4</td>
</tr>
<tr>
<td>ViT-B-VTN [29]</td>
<td>ImageNet-21K</td>
<td>78.6</td>
<td>93.7</td>
<td>1 x 1</td>
<td>4218</td>
<td>11.04</td>
</tr>
<tr>
<td>ViViT-L/16x2 [1]</td>
<td>ImageNet-21K</td>
<td>80.6</td>
<td>94.7</td>
<td>4 x 3</td>
<td>1446</td>
<td>310.8</td>
</tr>
<tr>
<td>ViViT-L/16x2 320 [1]</td>
<td>ImageNet-21K</td>
<td>81.3</td>
<td>94.7</td>
<td>4 x 3</td>
<td>3992</td>
<td>310.8</td>
</tr>
<tr>
<td>ip-CSN-152 [56]</td>
<td>IG-65M</td>
<td>82.5</td>
<td>95.3</td>
<td>10 x 3</td>
<td>109</td>
<td>32.8</td>
</tr>
<tr>
<td>ViViT-L/16x2 [1]</td>
<td>JFT-300M</td>
<td>82.8</td>
<td>95.5</td>
<td>4 x 3</td>
<td>1446</td>
<td>310.8</td>
</tr>
<tr>
<td>ViViT-L/16x2 320 [1]</td>
<td>JFT-300M</td>
<td>83.5</td>
<td>95.5</td>
<td>4 x 3</td>
<td>3992</td>
<td>310.8</td>
</tr>
<tr>
<td>ViViT-H/16x2 [1]</td>
<td>JFT-300M</td>
<td>84.8</td>
<td>95.8</td>
<td>4 x 3</td>
<td>8316</td>
<td>647.5</td>
</tr>
<tr>
<td>Swin-T</td>
<td>ImageNet-1K</td>
<td>78.8</td>
<td>93.6</td>
<td>4 x 3</td>
<td>88</td>
<td>28.2</td>
</tr>
<tr>
<td>Swin-S</td>
<td>ImageNet-1K</td>
<td>80.6</td>
<td>94.5</td>
<td>4 x 3</td>
<td>166</td>
<td>49.8</td>
</tr>
<tr>
<td>Swin-B</td>
<td>ImageNet-1K</td>
<td>80.6</td>
<td>94.6</td>
<td>4 x 3</td>
<td>282</td>
<td>88.1</td>
</tr>
<tr>
<td>Swin-B</td>
<td>ImageNet-21K</td>
<td>82.7</td>
<td>95.5</td>
<td>4 x 3</td>
<td>282</td>
<td>88.1</td>
</tr>
<tr>
<td>Swin-L</td>
<td>ImageNet-21K</td>
<td>83.1</td>
<td>95.9</td>
<td>4 x 3</td>
<td>604</td>
<td>197.0</td>
</tr>
<tr>
<td>Swin-L (384↑)</td>
<td>ImageNet-21K</td>
<td>84.6</td>
<td>96.5</td>
<td>4 x 3</td>
<td>2107</td>
<td>200.0</td>
</tr>
<tr>
<td>Swin-L (384↑)</td>
<td>ImageNet-21K</td>
<td><strong>84.9</strong></td>
<td><strong>96.7</strong></td>
<td>10 x 5</td>
<td>2107</td>
<td>200.0</td>
</tr>
</tbody>
</table>

*Comparison on K400; Some entries (mostly CNN-based) have been removed from table*
<table>
<thead>
<tr>
<th>Method</th>
<th>Pretrain</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Views</th>
<th>FLOPs</th>
<th>Param</th>
</tr>
</thead>
<tbody>
<tr>
<td>SlowFast R101+NL [13]</td>
<td>-</td>
<td>81.8</td>
<td>95.1</td>
<td>$10 \times 3$</td>
<td>234</td>
<td>59.9</td>
</tr>
<tr>
<td>X3D-XL [12]</td>
<td>-</td>
<td>81.9</td>
<td>95.5</td>
<td>$10 \times 3$</td>
<td>48</td>
<td>11.0</td>
</tr>
<tr>
<td>MVIT-B-24, 32x3 [9]</td>
<td>-</td>
<td>83.8</td>
<td>96.3</td>
<td>$5 \times 1$</td>
<td>236</td>
<td>52.9</td>
</tr>
<tr>
<td>TimeSformer-HR [3]</td>
<td>ImageNet-21K</td>
<td>82.4</td>
<td>96</td>
<td>$1 \times 3$</td>
<td>1703</td>
<td>121.4</td>
</tr>
<tr>
<td>ViViT-L/16x2 320 [1]</td>
<td>ImageNet-21K</td>
<td>83.0</td>
<td>95.7</td>
<td>$4 \times 3$</td>
<td>3992</td>
<td>310.8</td>
</tr>
<tr>
<td>ViViT-H/16x2 [9]</td>
<td>JFT-300M</td>
<td>85.8</td>
<td>96.5</td>
<td>$4 \times 3$</td>
<td>8316</td>
<td>647.5</td>
</tr>
<tr>
<td>Swin-B</td>
<td>ImageNet-21K</td>
<td>84.0</td>
<td>96.5</td>
<td>$4 \times 3$</td>
<td>282</td>
<td>88.1</td>
</tr>
<tr>
<td>Swin-L (384↑)</td>
<td>ImageNet-21K</td>
<td>85.9</td>
<td>97.1</td>
<td>$4 \times 3$</td>
<td>2107</td>
<td>200.0</td>
</tr>
<tr>
<td>Swin-L (384↑)</td>
<td>ImageNet-21K</td>
<td><strong>86.1</strong></td>
<td><strong>97.3</strong></td>
<td>$10 \times 5$</td>
<td>2107</td>
<td>200.0</td>
</tr>
</tbody>
</table>

*FLOPs are in gigaFLOPs*
<table>
<thead>
<tr>
<th>Method</th>
<th>Pretrain</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Views</th>
<th>FLOPs</th>
<th>Param</th>
</tr>
</thead>
<tbody>
<tr>
<td>TimeSformer-HR [3]</td>
<td>ImageNet-21K</td>
<td>62.5</td>
<td>-</td>
<td>1 x 3</td>
<td>1703</td>
<td>121.4</td>
</tr>
<tr>
<td>SlowFast R101, 8x8 [13]</td>
<td>Kinetics-400</td>
<td>63.1</td>
<td>87.6</td>
<td>1 x 3</td>
<td>106</td>
<td>53.3</td>
</tr>
<tr>
<td>TSM-RGB [27]</td>
<td>Kinetics-400</td>
<td>63.3</td>
<td>88.2</td>
<td>2 x 3</td>
<td>62</td>
<td>42.9</td>
</tr>
<tr>
<td>MSNet [23]</td>
<td>ImageNet-21K</td>
<td>64.7</td>
<td>89.4</td>
<td>1 x 1</td>
<td>67</td>
<td>24.6</td>
</tr>
<tr>
<td>TEA [26]</td>
<td>ImageNet-21K</td>
<td>65.1</td>
<td>89.9</td>
<td>10 x 3</td>
<td>70</td>
<td>-</td>
</tr>
<tr>
<td>bLVNet [11]</td>
<td>SSV2</td>
<td>65.2</td>
<td>90.3</td>
<td>1 x 1</td>
<td>129</td>
<td>40.2</td>
</tr>
<tr>
<td>ViViT-L/16x2 [1]</td>
<td>-</td>
<td>65.4</td>
<td>89.8</td>
<td>-</td>
<td>903</td>
<td>352.1</td>
</tr>
<tr>
<td>MViT-B, 64x3 [10]</td>
<td>Kinetics-400</td>
<td>67.7</td>
<td>90.9</td>
<td>1 x 3</td>
<td>455</td>
<td>36.6</td>
</tr>
<tr>
<td>MViT-B-24, 32x3 [10]</td>
<td>Kinetics-600</td>
<td>68.7</td>
<td>91.5</td>
<td>1 x 3</td>
<td>236</td>
<td>53.2</td>
</tr>
<tr>
<td>Swin-B</td>
<td>Kinetics-400</td>
<td>69.6</td>
<td>92.7</td>
<td>1 x 3</td>
<td>321</td>
<td>88.8</td>
</tr>
</tbody>
</table>
Ablation Study: Attention with Swin-T on K400

- Not running into memory constraints due to shifted windows
- Split model has less efficient temporal modeling
- Factorized approach has to compute temporal and spatial information separately
Ablation Study
Pre-Training and Learning Rate

By lowering the learning rate of the backbone relative to the head, the model maintains pre-trained parameters longer.

Allows for better generalization.

Future Study: How to better use pre-trained weights?

<table>
<thead>
<tr>
<th>ratio</th>
<th>Pretrain</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1×</td>
<td>ImageNet-1K</td>
<td>80.6</td>
<td>94.6</td>
</tr>
<tr>
<td>1.0×</td>
<td>ImageNet-1K</td>
<td>80.2</td>
<td>94.2</td>
</tr>
<tr>
<td>0.1×</td>
<td>ImageNet-21K</td>
<td>82.6</td>
<td>95.7</td>
</tr>
<tr>
<td>1.0×</td>
<td>ImageNet-21K</td>
<td>82.0</td>
<td>95.3</td>
</tr>
</tbody>
</table>
Conclusion: Advantages

● Smaller model size
● Requires less pre-training data
● Optimization allows for:
  ○ Joint spatiotemporal self-attention without running out of memory
  ○ Lower computation costs
Conclusion: Disadvantages

- Requires pre-training
- All ablation studies were on K400
  - Missing potentially interesting evaluation on temporally-intensive dataset
- Not much novelty
  - Swin applied to videos


Both from Microsoft Research Asia