Permutation Invariant Convolution
For Recognizing Long-range Activities

Noureldien Hussein  Efstratios Gavves  Arnold Smeulders

University of Amsterdam
Motivation
Motivation

High Jump
Motivation

High Jump

Temporal Pattern
Motivation

Temporal Pattern → CNN → Long Jump

Kernels
Motivation

Temporal Pattern

Repetition
Motivation

Temporal Pattern → CNN → Long Jump

Kernels
Motivation

Temporal Pattern → CNN → Long Jump

Permutation Invariant Kernels
Motivation

Temporal Pattern → CNN → Long Jump

Permutation Invariant Kernels
Permutation?
Motivation

--- Short-range Video (10 sec.) ---

run  jump  cross  fall
Motivation

Long-range Video (5 min.)

\( v_1 \)

- fridge
- meat
- cucumber
- rice
- present
- vegetables
- wash
Motivation
Motivation

$\nu_1$

$\nu_2$

- fridge
- meat
- cucumber
- rice
- present
- vegetables
- wash
Motivation

$v_1$

$v_2$

$v_3$
Motivation

$v_1$

$v_2$

$v_3$

- fridge
- meat
- cucumber
- rice
- present
- vegetables
- wash
Method
Solution

--- Making Coffee ---

\[ v_1 \]

\[ v_2 \]

\[ \vdots \]

\[ v_n \]

- take cup
- pour coffee
- pour sugar
- spoon sugar
- pour milk
- stir coffee
Solution

Making Coffee

$v_1$   $v_2$   $\vdots$   $v_n$

- take cup
- pour coffee
- pour sugar
- spoon sugar
- pour milk
- stir coffee

micro-level
Solution

--- Making Coffee ---

\[ \nu_1 \]

\[ \nu_2 \]

\[ \vdots \]

\[ \nu_n \]

- take cup
- pour coffee
- pour sugar
- spoon sugar
- pour milk
- stir coffee
Solution

Making Coffee

$v_1$

$v_2$

::

$v_n$

micro-level

macro-level

Preparing Coffee

take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee
\[ s = K \otimes X_w^T \] (1)
$$s = K \otimes X_w^T \quad (1)$$

$$s' = \text{max}_{\text{row}}(s) \quad (2)$$
\begin{align*}
\mathbf{s} &= K \otimes X_w^T \quad (1) \\
\mathbf{s}' &= \max_{\text{row}}(s) \quad (2) \\
\alpha &= \sigma \left[ f_\theta(s') \right] \quad (3)
\end{align*}
\[ s = K \otimes X_w^T \]  \hspace{1cm} (1)

\[ s' = \max_{\text{row}}(s) \]  \hspace{1cm} (2)

\[ \alpha = \sigma [ f_\theta (s') ] \]  \hspace{1cm} (3)

\[ y = \alpha \odot V \]  \hspace{1cm} (4)
\begin{align*}
    s &= K \otimes X_w^T \quad (1) \\
    s' &= \max_{\text{row}}(s) \quad (2) \\
    \alpha &= \sigma[f_\theta(s')] \quad (3) \\
    y &= \alpha \odot V \quad (4)
\end{align*}
\[ s = K \otimes X_w^T \] (1)

\[ s' = \max_{\text{row}}(s) \] (2)

\[ \alpha = \sigma[f_\theta(s')] \] (3)

\[ y = \alpha \odot V \] (4)
\[
\begin{align*}
    s &= K \otimes X^T_w \quad (1) \\
    s' &= \text{max}_\text{row}(s) \quad (2) \\
    \alpha &= \sigma [ f_\theta(s') ] \quad (3) \\
    y &= \alpha \odot V \quad (4)
\end{align*}
\]
\[ s = K \otimes X_w^T \quad (1) \]
\[ s' = \max_{\text{row}}(s) \quad (2) \]
\[ \alpha = \sigma[f_\theta(s')] \quad (3) \]
\[ y = \alpha \odot V \quad (4) \]
Solution

- Keys ($K$)
- Similarities ($s$)
- Max
- Values ($V$)
- $X_w$
- PIC Operation

$$y = \alpha \cdot f_{\theta} (s')$$
Solution

PIC Operation

PIC Layer
Solution

PIC Operation

PIC Layer
Predictions

Video Segment

\[ S_1 \rightarrow \ldots \rightarrow S_T \]

\[ \text{CNN} \]

\[ x_1 \rightarrow \ldots \rightarrow x_T \]

\[ \text{PIC} \]

\[ \ldots \]

\[ \text{PIC} \]

\[ Z \rightarrow \text{Dense} \]

\[ \text{Predictions} \]

\[ X'_w \]

\[ X'_w \rightarrow g_{\phi} \]

\[ g_{\phi} \rightarrow K \]

\[ K \rightarrow V \]

\[ V \rightarrow X_w \]

\[ X_w \rightarrow M \times C \]

\[ M \times C \rightarrow h_{\psi} \]

\[ h_{\psi} \rightarrow y \]

\[ y \rightarrow 1 \times C \]

\[ 1 \times C \rightarrow y' \]

\[ y' \rightarrow 1 \times C' \]

\[ 1 \times C' \rightarrow \text{PIC} \]

\[ \text{PIC} \rightarrow \text{PIC Layer} \]

\[ \text{CNN Classifier} \]
Experiments
Permutation Invariance

Uniform $\nu$
Permutation Invariance

Uniform
Coarse Perm
Fine Perm
### Permutation Invariance

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Coarse</th>
<th>Uniform</th>
<th>Drop ↓</th>
<th>Fine</th>
<th>Uniform</th>
<th>Drop ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeception</td>
<td>84.6</td>
<td>82.2</td>
<td>2.4</td>
<td>84.6</td>
<td>81.9</td>
<td>2.7</td>
</tr>
<tr>
<td>PIC-Ordered</td>
<td>80.2</td>
<td>77.6</td>
<td>2.6</td>
<td>80.2</td>
<td>76.3</td>
<td>3.9</td>
</tr>
<tr>
<td>PIC</td>
<td>87.5</td>
<td>87.0</td>
<td>0.5</td>
<td>87.5</td>
<td>86.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Drop in Accuracy in Permuted Video Segments
## Permutation Invariance

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Coarse</th>
<th>Uniform</th>
<th>Drop ↓</th>
<th>Fine</th>
<th>Uniform</th>
<th>Drop ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeception</td>
<td>84.6</td>
<td>82.2</td>
<td>2.4</td>
<td>84.6</td>
<td>81.9</td>
<td>2.7</td>
</tr>
<tr>
<td>PIC-Ordered</td>
<td>80.2</td>
<td>77.6</td>
<td>2.6</td>
<td>80.2</td>
<td>76.3</td>
<td>3.9</td>
</tr>
<tr>
<td>PIC</td>
<td>87.5</td>
<td>87.0</td>
<td>0.5</td>
<td>87.5</td>
<td>86.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Drop in Accuracy in Permutated Video Segments
## Permutation Invariance

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Coarse</th>
<th>Uniform</th>
<th>Drop ↓</th>
<th>Fine</th>
<th>Uniform</th>
<th>Drop ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeception</td>
<td>84.6</td>
<td>82.2</td>
<td>2.4</td>
<td>84.6</td>
<td>81.9</td>
<td>2.7</td>
</tr>
<tr>
<td>PIC-Ordered</td>
<td>80.2</td>
<td>77.6</td>
<td>2.6</td>
<td>80.2</td>
<td>76.3</td>
<td>3.9</td>
</tr>
<tr>
<td>PIC</td>
<td>87.5</td>
<td>87.0</td>
<td>0.5</td>
<td>87.5</td>
<td>86.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Drop in Accuracy in Permuted Video Segments
Cascaded Layers

Making Coffee

$\nu_1$

micro-level

macro-level
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ActionVlad</td>
<td>83.07</td>
<td>----</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
<td>83.33</td>
</tr>
<tr>
<td>PIC-Global</td>
<td>86.76</td>
<td>85.68</td>
</tr>
<tr>
<td>Timeception</td>
<td>83.85</td>
<td>84.90</td>
</tr>
<tr>
<td>PIC</td>
<td>86.25</td>
<td>87.72</td>
</tr>
</tbody>
</table>

Increase in Accuracy with Cascaded Layers
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ActionVlad</td>
<td>83.07</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
</tr>
<tr>
<td>PIC-Global</td>
<td>86.76</td>
</tr>
<tr>
<td>Timeception</td>
<td>83.85</td>
</tr>
<tr>
<td>PIC</td>
<td>86.25</td>
</tr>
</tbody>
</table>

Increase in Accuracy with Cascaded Layers
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ActionVlad</td>
<td>83.07</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
<td>83.33</td>
<td>83.07</td>
<td>83.29</td>
</tr>
<tr>
<td>PIC-Global</td>
<td><strong>86.76</strong></td>
<td>85.68</td>
<td>85.68</td>
<td>85.42</td>
</tr>
<tr>
<td>Timeception</td>
<td>83.85</td>
<td>84.90</td>
<td>85.30</td>
<td>86.93</td>
</tr>
<tr>
<td>PIC</td>
<td><strong>86.25</strong></td>
<td><strong>87.72</strong></td>
<td><strong>88.02</strong></td>
<td><strong>89.84</strong></td>
</tr>
</tbody>
</table>

*Increase in Accuracy with Cascaded Layers*
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ActionVlad</td>
<td>83.07</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
</tr>
<tr>
<td>PIC-Global</td>
<td>86.76</td>
</tr>
<tr>
<td>Timeception</td>
<td><strong>83.85</strong></td>
</tr>
<tr>
<td>PIC</td>
<td><strong>86.25</strong></td>
</tr>
</tbody>
</table>

Increase in Accuracy with Cascaded Layers
Key-Value Pair Kernels

\[ x_1 \quad x_T \quad x_N \]

\[ X_W \rightarrow \text{Max} \rightarrow S' \rightarrow f_\theta(\alpha) \rightarrow \mathbf{y} \]

Keys (K) Similarities (S)

Values (V)
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
</tr>
<tr>
<td>PIC-Inferred</td>
<td>82.55</td>
</tr>
<tr>
<td>PIC</td>
<td>86.27</td>
</tr>
</tbody>
</table>

Benefit of Shared v.s. Inferred Key-Value Kernels
Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
</tr>
<tr>
<td>PIC-Inferred</td>
<td>82.55</td>
</tr>
<tr>
<td>PIC</td>
<td><strong>86.27</strong></td>
</tr>
</tbody>
</table>

Benefit of Shared v.s. Inferred Key-Value Kernels
## Cascaded Layers

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Accuracy (%) @ Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Non-local</td>
<td>82.29</td>
</tr>
<tr>
<td>PIC-Inferred</td>
<td>82.55</td>
</tr>
<tr>
<td>PIC</td>
<td>86.27</td>
</tr>
</tbody>
</table>

**Benefit of Shared v.s. Inferred Key-Value Kernels**
Related Works

(a) Self-Attention
(b) Vector Aggregation
(c) Convolution
## Related Works

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SlowFast [47]</td>
<td>—</td>
<td>42.1</td>
</tr>
<tr>
<td>SlowFast-NL* [47]</td>
<td>—</td>
<td>42.5</td>
</tr>
<tr>
<td>3D CNN [6]</td>
<td>R101</td>
<td>35.5</td>
</tr>
<tr>
<td>3D CNN + TC [2]</td>
<td>R101</td>
<td>41.1</td>
</tr>
<tr>
<td>3D CNN + PIC</td>
<td>R101</td>
<td>42.7</td>
</tr>
<tr>
<td>3D CNN [7]</td>
<td>R101-NL</td>
<td>41.0</td>
</tr>
<tr>
<td>3D CNN + FB [7]</td>
<td>R101-NL</td>
<td>42.5</td>
</tr>
<tr>
<td>3D CNN + PIC</td>
<td>R101-NL</td>
<td><strong>43.8</strong></td>
</tr>
</tbody>
</table>

### Breakfast

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D CNN</td>
<td>I3D</td>
<td>80.64</td>
</tr>
<tr>
<td>3D CNN + Vlad [46]</td>
<td>I3D</td>
<td>82.67</td>
</tr>
<tr>
<td>3D CNN + Nonlocal [6]</td>
<td>I3D</td>
<td>83.79</td>
</tr>
<tr>
<td>3D CNN + Timeception [2]</td>
<td>I3D</td>
<td>86.93</td>
</tr>
<tr>
<td>3D CNN + PIC</td>
<td>I3D</td>
<td><strong>89.84</strong></td>
</tr>
</tbody>
</table>

### Charades

### Multi-Thumos

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D CNN</td>
<td>I3D</td>
<td>72.43</td>
</tr>
<tr>
<td>3D CNN + Timeception [2]</td>
<td>I3D</td>
<td>74.79</td>
</tr>
<tr>
<td>3D CNN + PIC</td>
<td>I3D</td>
<td><strong>78.31</strong></td>
</tr>
</tbody>
</table>
Performance Analysis

- Time (MS) vs. Layers
- Params (M) vs. Layers
- FLOPS (Mega) vs. Layers
- Accuracy (%) vs. Layers

- PIC
- Timeception
- Nonlocal
- Vlad
Performance Analysis
Performance Analysis

![Graphs showing performance metrics for different layers and models: Time, Params, FLOPS, and Accuracy.](image-url)
Performance Analysis

Graphs showing the performance analysis of different models:
- Time (MS) over Layers
- Params (M) over Layers
- FLOPS (Mega) over Layers
- Accuracy (%) over Layers

Models compared:
- PIC
- Timeception
- Nonlocal
- Vlad
Learned Concepts

- A01: Food Box (A05 Milk, A01 Cereals)
- A02: Pouring (A02 Coffee, A10 Tea)
- A07: Cutting (A07 Salad, A08 Sandwich)
Learned Concepts

A05
A01
Food Box (A05 Milk, A01 Cereals)

A02
A10
Pouring (A02 Coffee, A10 Tea)

A07
A08
Cutting (A07 Salad, A08 Sandwich)

Object-centric
Learned Concepts

Object-centric

- Food Box (A05 Milk, A01 Cereals)

Action-centric

- Pouring (A02 Coffee, A10 Tea)
- Cutting (A07 Salad, A08 Sandwich)
Learned Attention

Making Pancake (6 Minutes)
Learned Attention

Making Pancake (6 Minutes)
Thank You!