2D Deep Video Capsule Network with Temporal Shift for Action Recognition

T. Voillemin - H. Wannous - J-P. Vandeborre
Introduction

Motivations & challenges

• Action Recognition for Virtual/Augmented reality devices

• Light but efficient algorithm for real-time applications

• Working with easily accessible data as RGB and depth video stream

Garcia-Hernando et al., CVPR, 2018

De Smedt et al., SHREC, 2017
Introduction

Existing approaches

- **For lightness:**
  - Capsule Network (Sabour et al., ANIPS, 2017)

- **For efficiency:**
  - DeepCaps (Rajasegaran et al., CVPR, 2019)
  - Temporal Shift Module (Lin et al. ICCV, 2019)
Proposed Method

Temporal Shift on Capsule Layer

Original Capsule Layer

Shift of first kernels on first capsule

Shift of first kernels on every capsules

Shift of every kernels on first capsule
Proposed Method

Network Architecture

ShiftConvCaps

Capsule Layer Group

Full Network
Experiments

Accuracy comparaison over DHG dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSM [2]</td>
<td>24 M</td>
<td>58.66</td>
</tr>
<tr>
<td>2D 3DCNN Fusion [3]</td>
<td>140 M</td>
<td>74.41</td>
</tr>
<tr>
<td>Ours</td>
<td>7 M</td>
<td>68.98</td>
</tr>
</tbody>
</table>

Comparaison on DHG28 [5]

[2] Lin et al., ICCV, 2019
[3] Zhang et al., Electronics, 2019
[4] Garcia-Hernando et al., CVPR, 2018
[5] De Smedt et al., SHREC, 2017
Experiments

Accuracy comparison over FPHA dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two stream-color [1]</td>
<td>46 M</td>
<td>61.56</td>
</tr>
<tr>
<td>Two stream-flow [1]</td>
<td>46 M</td>
<td>69.91</td>
</tr>
<tr>
<td>Two stream-all [1]</td>
<td>181 M</td>
<td>75.30</td>
</tr>
<tr>
<td>TSM [2]</td>
<td>24 M</td>
<td>71.57</td>
</tr>
<tr>
<td>Ours</td>
<td>4 M</td>
<td>76.72</td>
</tr>
</tbody>
</table>

Comparaison on FPHA [4]

Garcia-Hernando et al., CVPR, 2018

[2] Lin et al., ICCV, 2019
[3] Zhang et al., Electronics, 2019
[4] Garcia-Hernando et al., CVPR, 2018
[5] De Smedt et al., SHREC, 2017
## Experiments

### Accuracy comparison over FPHA dataset

<table>
<thead>
<tr>
<th>Temporal Shift</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No shift</td>
<td>70.01</td>
</tr>
<tr>
<td>Shift first kernels on first capsules</td>
<td>64.14</td>
</tr>
<tr>
<td>Shift first kernels on every capsules</td>
<td>74.33</td>
</tr>
<tr>
<td>Shift every kernels of first capsules</td>
<td>76.72</td>
</tr>
</tbody>
</table>

**Final testing accuracy**

**Training accuracy evolution over epochs**
Conclusion

Contributions & perspectives

- First 2D Capsule Network for video understanding
- Implementation of temporal shift over capsule layer
- Outperform or near state-of-the-art with 10 to 40 times less parameters

- Adapt capsules for hand skeleton data
- Use it in real-time directly on augmented/virtual reality devices
Thank you for your attention

T. Voillemin - H. Wannous - J-P. Vandeborre