**Motivation:**

- Video action recognition is a fundamental yet challenging task in the field of computer vision.
- Short-range motion features and long-range dependencies are two complementary and vital cues for action recognition in videos.
- It is still unclear how to capture temporal information with complex evolution on multiple ranges using an efficient and effective way.

**Feature interchange:** the features from the colored regions bi-directionally shift in the feature map of video models.

**Contribution:**

- Perform channel-wise temporal interchange (CTI) along the temporal dimension to effectively encode short-range motion features.
- Construct graph-based regional interchange (GRI) module to learn efficiently long-range dependencies using graph convolution.
- Propose a novel multi-range feature interchange (MFI) network to integrate the proposed two modules. Achieves competitive results by using very limited computing cost.

**Network Architecture**

- **Channel-wise Temporal Interchange (CTI) Module**
  - The temporal difference can be obtained by calculating the difference between the features of two consecutive frames.
  - Temporal interchange operation.
    
```
H′_t[k, w, c] = H_t[k, w, c] − H_{t−1}[k, w, c], \quad t \in [1, T−1].
```

**Channel-wise Temporal Interchange (CTI) Module**

- Transform from the features in a regular feature map to the state nodes in a non-grid graph.

```
W_c = [\text{Conv}_{\text{in}} \otimes \Phi_c(X)]^c, \quad V_c = R^c.
```

- Graph Convolutional Operation. The nodes propagate their state with each other.

```
V_c = ReLU(\sum_{c=1}^{C} W_c \otimes V_c).
```

- Reverse the output into the regular feature maps to be compatible with CNN models.

```
Y_{\text{out}} = \phi\left(W_{\text{out}} \otimes V_{\text{out}}\right)
```

**Architecture Details**

**Experimental Results**

- **Table 1:** The comparison of performance on Something-Something V1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>#Frames</th>
<th>FLOPs (Gops)</th>
<th>Val Top1 (%)</th>
<th>Val Top5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN</td>
<td>BNInception</td>
<td>8</td>
<td>33G</td>
<td>39.3</td>
<td>66.1</td>
</tr>
<tr>
<td>TSM</td>
<td>BNInception</td>
<td>8</td>
<td>33G</td>
<td>39.7</td>
<td>66.6</td>
</tr>
<tr>
<td>MultiScale TSN</td>
<td>BNInception</td>
<td>8</td>
<td>33G</td>
<td>34.4</td>
<td>73.2</td>
</tr>
<tr>
<td>TSM</td>
<td>ResNet-50</td>
<td>16</td>
<td>33G</td>
<td>43.3</td>
<td>78.5</td>
</tr>
<tr>
<td>ECOA1</td>
<td>BNInception</td>
<td>16</td>
<td>33G</td>
<td>37.6</td>
<td>77.8</td>
</tr>
<tr>
<td>ECOA2</td>
<td>BNInception</td>
<td>32</td>
<td>33G</td>
<td>37.8</td>
<td>75.7</td>
</tr>
</tbody>
</table>

- **Table 2:** The comparison on UCF101 and HMDB51.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Frames</th>
<th>UCF101</th>
<th>HMDB51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stream TSN</td>
<td>8</td>
<td>89.8</td>
<td>89.5</td>
</tr>
<tr>
<td>ECOA1</td>
<td>8</td>
<td>89.4</td>
<td>89.1</td>
</tr>
<tr>
<td>ECOA2</td>
<td>32</td>
<td>89.0</td>
<td>88.7</td>
</tr>
</tbody>
</table>

**References**


**Contact Info**

- WeChat: whitesk1973@gmail.com

- Github: Whitesk1973

- Email: whitesk1973@gmail.com

Fig. 4: Some prediction examples on Something-Something V1. The top 2 predictions with green text indicating a correct prediction.