Spatial Attention Selection for Fine-grained Action Recognition

Dichao Liu, Yu Wang, Jien Kato (Nagoya University)

1. Introduction
Among all computer vision tasks, fine-grained action recognition is an extremely difficult problem because very detailed visual information needs to be excavated as thoroughly as possible. Traditional researches select attention region by hand-crafted methods. In the paper, we try to use Spatial Transformer Network (ST-network)[1] learn this information in end-to-end approach.

As shown in Fig.1, in ST-network, firstly, a localization net is used to learn a set of parameters \( \theta \), which is actually a 6-dimensional matrix \( \left[ \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \theta_{31}, \theta_{32} \right] \) applying affine transformation. Then \( \theta \) is applied on input to generate new inputs, which will be fed into recognition network.

Fig.1. Process of ST-network

2. Approach

\(<2 \cdot 1>\) Attention region selection: we first select attention region by the hand-crafted method in (2). Then we also try to use end-to-end approach to get the attention region.

\(<2 \cdot 2>\) Regressive initialization of ST-network: we only update last 2 sampler parameters of \( \theta \left( [\theta_{11}, \theta_{21}] \right) \), which are used for translation. At first, we initialize the ST-network by regression with the regression target of \( [\theta_{11}, \theta_{21}] \) so that initially, the ST-network will try to select the center region of each frame.

\(<2 \cdot 3>\) CNN training of different inputs: to evaluate the effectiveness, we train CNNs of inputs obtained by 4 different ways. Consider the original size of a frame is \( w \times h \). The size of attention region and recognition network input is \( v \times v \) (\( h>v>\)).

1. Random cropping: randomly crop \( v \times v \) region from \( w \times h \) frames.

2. Attention region-based random cropping: Randomly crop \( w \times v \) region from \( w \times h \) frames with making sure that the \( v \times v \) attention region is included in the \( w \times w \) region. Then randomly crop \( v \times v \) region from the \( w \times w \) region.

3. Cropping by attention region: Crop \( v \times v \) region at the location where exactly the attention region is.

4. Cropping by ST-network: Randomly crop \( w \times w \) region from \( w \times h \) frames with making sure that the \( v \times v \) attention region is included in the \( w \times w \) region. Then use ST-network to learn the \( v \times v \) region from the \( w \times w \) region.

3. Experiments

\(<3 \cdot 1>\) Dataset: MPII Cooking Activities Dataset[3] is a dataset of cooking activities. The dataset contains 5609 clips, 3748 of which are labeled as one of 64 distinct cooking activities and the rest 1861 are labeled as background activity.

\(<3 \cdot 2>\) Preprocessing: We first resized all frames to be 448 × 594. Then we get attention region with size of 224 × 224 as [2]. We traversed the frames by the step of 4 pixels and we updated the attention region position every 10 frames.

\(<3 \cdot 3>\) Training details: Both ST-network and recognition networks in this paper are based on the very deep VGG-16 model[4]. We set the batch size as 128. We start the learning rate as 10⁻⁴. When ST-network and recognition networks are trained together, the learning rate of the latter is 10 times as the former.

\(<3 \cdot 4>\) Results and the conclusion: the results are shown as Table 1. Especially, ① is actually trained by adding ST-network to ②, achieving dramatically improvement over ②. Also, ④ performs better than ③, which shows that ST-network can effectively learn spatial attention information from frames.

Table 1 Results of evaluation

<table>
<thead>
<tr>
<th>Input-obtaining methods</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>①</td>
<td>19.04%</td>
</tr>
<tr>
<td>②</td>
<td>20.36%</td>
</tr>
<tr>
<td>③</td>
<td>27.73%</td>
</tr>
<tr>
<td>④</td>
<td>28.38%</td>
</tr>
</tbody>
</table>

Acknowledgement: This research is supported by the JSPS Grant-in-Aid for Scientific Research B (No. 26280057), the JSPS Grant-in-Aid for Challenging Exploratory Research (No.16K12460), and the JST Center of Innovation Program.

References