Activity Recognition

Computer Vision
Fall 2018
Columbia University

Many slides from Bolei Zhou
• How are they going? About 30 teams have requested GPU credit so far

• Final presentations on December 5th and 10th

  • We will assign you to dates soon

• Final report due December 10 at midnight

• Details here: http://w4731.cs.columbia.edu/project
Challenge for Image Recognition

• Variation in appearance.
Challenge for Activity Recognition

- Describing activity at the proper level

Image recognition? Skeleton recognition? No motion needed? Which activities?
Challenge for Activity Recognition

• Describing activity at the proper level

A chain of events
Making chocolate cookies
What are they doing?
What are they doing?

Barker and Wright, 1954
Vision or Cognition?
Video Recognition Datasets

• KTH Dataset: recognition of human actions
• 6 classes, 2391 videos

https://www.youtube.com/watch?v=Jm69kbCC17s
Recognizing Human Actions: A Local SVM Approach. ICPR 2004
Video Recognition Datasets

• UCF101 from University of Central Florida
• 101 classes, 9,511 videos in training

https://www.youtube.com/watch?v=hGhuUaxocIE
Video Recognition Datasets

• Kinetics from Google DeepMind
• 400 classes, 239,956 videos in training

https://deepmind.com/research/open-source/open-source-datasets/kinetics/
Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowdsourced video dataset

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV’16

http://allennlp.org/plato/charades/
Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowdsourced video dataset

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV’16
Example annotated videos from the Charades dataset

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding
Video Recognition Datasets

• Something-Something dataset: human object interaction

• 174 categories: 100,000 videos

- Holding something
- Turning something upside down
- Turning the camera left while filming something
- Opening something

Poking a stack of something so the stack collapses
Plugging something into something

https://www.twentybn.com/datasets/something-something
<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Activity Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Activity</td>
</tr>
<tr>
<td>234</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>7</td>
<td>123</td>
</tr>
<tr>
<td>67</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>97</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>78</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>81</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>64</td>
<td>35</td>
<td>46</td>
</tr>
<tr>
<td>75</td>
<td>464</td>
<td>576</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>64</td>
<td>3</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>234</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>7</td>
<td>123</td>
</tr>
<tr>
<td>67</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>97</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>78</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>81</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>64</td>
<td>35</td>
<td>46</td>
</tr>
<tr>
<td>75</td>
<td>464</td>
<td>576</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>64</td>
<td>3</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>234</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>7</td>
<td>123</td>
</tr>
<tr>
<td>67</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>97</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>78</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>81</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>64</td>
<td>35</td>
<td>46</td>
</tr>
<tr>
<td>75</td>
<td>464</td>
<td>576</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>64</td>
<td>3</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

The table shows the relationship between Width, Height, and Activity Labels over Time.
Single-frame image model

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

41.1%  40.7%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

Single Frame

Late Fusion

Early Fusion

41.1%

40.7%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

Single Frame: 41.1%
Late Fusion: 40.7%
Early Fusion: 38.9%

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Multi-frame fusion model

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
mountain unicycling: 0.280

canyoning: 0.164

base jumping: 0.124
Sequence of frames?
Recurrent Neural Networks (RNNs)

In the above diagram, a chunk of neural network, $A$, looks at some input $x_t$ and outputs a value $h_t$. A loop allows information to be passed from one step of the network to the next.
Recurrent Neural Networks (RNNs)

A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.
When the gap between the relevant information and the place that it’s needed is small, RNNs can learn to use the past information.
Long-term dependencies - hard to model!

But there are also cases where we need more context.

Credit: Christopher Olah
From plain RNNs to LSTMs

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Credit: Christopher Olah
From plain RNNs to LSTMs

(LSTM: Long Short Term Memory Networks)

Credit: Christopher Olah
The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.
LSTMs Step by Step: Forget Gate

Should we continue to remember this “bit” of information or not?

\[ f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f) \]

The first step in our LSTM is to decide what information we’re going to throw away from the cell state. This decision is made by a sigmoid layer called the “forget gate layer.”

Credit: Christopher Olah
LSTMs Step by Step: Input Gate

Should we update this “bit” of information or not? If so, with what?

The next step is to decide what new information we’re going to store in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Next, a tanh layer creates a vector of new candidate values, $\tilde{C}_t$, that could be added to the state.

\[
i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)
\]
\[
\tilde{C}_t = \text{tanh}(W_C \cdot [h_{t-1}, x_t] + b_C)
\]
LSTMs Step by Step: Memory Update

Decide what will be kept in the cell state/memory

\[ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \]

Forget that

Memorize this

Credit: Christopher Olah
LSTMs Step by Step: Output Gate

Should we output this “bit” of information?

This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we’re going to output. Then, we put the cell state through tanh (to push the values to be between $-1$ and $1$) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.
Complete LSTM - A pretty sophisticated cell

Credit: Christopher Olah
Show and Tell: A Neural Image Caption Generator

A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.
Multi-frame LSTM fusion model

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015
Motivation: Separate visual pathways in nature

Dorsal stream (‘where/how’) recognizes motion and locates objects

Ventral (‘what’) stream performs object recognition

https://en.wikipedia.org/wiki/Two-streams_hypothesis
2-Stream Network

Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014
Temporal segment network

3D convolutional Networks

2D convolutions

3D convolutions

Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015
3D convolutional Networks

• 3D filters at the first layer.
Temporal Relational Reasoning

- Infer the temporal relation between frames.

Poking a stack of something so it collapses
Temporal Relational Reasoning

- It is the temporal transformation/relation that defines the activity, rather than the appearance of objects.

Poking a stack of something so it collapses
Temporal Relations in Videos

Pretending to put something next to something

2-frame relations

3-frame relations

4-frame relations
Framework of Temporal Relation Networks

Pretending to put something next to something
Something-Something Dataset

- 100 K videos from 174 human-object interaction classes.

Moving something away from something

Plugging something into something

Pulling two ends of something so that it gets stretched
Jester Dataset

• 140 K videos from 27 gesture classes.

Zooming in with two fingers

Thumb down

Drumming fingers
# Experimental Results

- On Something-Something dataset

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc.(%)</th>
<th>Top5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>single frame</td>
<td>11.41</td>
<td>33.39</td>
</tr>
<tr>
<td>2-frame TRN</td>
<td>22.23</td>
<td>48.80</td>
</tr>
<tr>
<td>3-frame TRN</td>
<td>26.22</td>
<td>54.15</td>
</tr>
<tr>
<td>4-frame TRN</td>
<td>29.83</td>
<td>58.21</td>
</tr>
<tr>
<td>5-frame TRN</td>
<td>30.39</td>
<td>58.29</td>
</tr>
<tr>
<td>7-frame TRN</td>
<td>31.01</td>
<td>59.24</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td>33.01</td>
<td>61.27</td>
</tr>
<tr>
<td>MultiScale TRN (10-crop)</td>
<td><strong>34.44</strong></td>
<td><strong>63.20</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yana Hasson</td>
<td>25.55</td>
</tr>
<tr>
<td>Harrison.AI</td>
<td>26.38</td>
</tr>
<tr>
<td>I3D by [8]</td>
<td>27.23</td>
</tr>
<tr>
<td>Guillaume Berger</td>
<td>30.48</td>
</tr>
<tr>
<td>Besnet (Top1 on leaderboard)</td>
<td>31.66</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td><strong>33.60</strong></td>
</tr>
</tbody>
</table>
Experimental Results

- On Jester dataset

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc. (%)</th>
<th>Top5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>single frame</td>
<td>63.60</td>
<td>92.44</td>
</tr>
<tr>
<td>2-frame TRN</td>
<td>75.65</td>
<td>94.40</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td>93.70</td>
<td>99.59</td>
</tr>
<tr>
<td>MultiScale TRN (10-crop)</td>
<td><strong>95.31</strong></td>
<td><strong>99.86</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20BN’s Jester System</td>
<td>82.34</td>
</tr>
<tr>
<td>VideoLSTM</td>
<td>85.86</td>
</tr>
<tr>
<td>Guillaume Berger</td>
<td>93.87</td>
</tr>
<tr>
<td>Ford’s Gesture Recognition System</td>
<td>94.11</td>
</tr>
<tr>
<td>Besnet (Top1 on leaderboard)</td>
<td>94.23</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td><strong>94.78</strong></td>
</tr>
</tbody>
</table>
Importance of temporal orders
How well are they diving?

Pirsiavash, Vondrick, Torralba. Assessing Quality of Actions, ECCV 2014
How well are they diving?

1. Track and compute human pose
How well are they diving?

1. Track and compute human pose

2. Extract temporal features
   - take FT and histogram?
   - use deep network?
**How well are they diving?**

1. Track and compute human pose

2. Extract temporal features
   - take FT and histogram?
   - use deep network?

3. Train regression model to predict expert quality score
Assessing diving

Estimated pose

Groundtruth score: 72.0
Predicted score: 73.8
Summarizing
Assessing figure skating