Recognition on videos
Video classification

• Significantly harder to collect data
• 1 video = hundreds of frames
• UCF 101: 13k videos from 101 actions
• HMDB: 7k videos from 51 actions
Is a video just a collection of images?
Is a video just a collection of images?
Is a video just a collection of images?
Is a video just a collection of images?
Is a video just a collection of images?

• Mostly a question of classes involved
• Actions involving object interactions can be identified from single image
  • e.g., riding horse, playing tennis, pitching ball
• Actions that occur in particular scenes can be identified from single images
  • e.g., diving, swimming, rock-climbing
Video recognition with convolutional networks
Extending convolutional networks to 3D
Extending convolutional networks to 3D

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best non-deep approach</td>
<td>85.9%</td>
</tr>
<tr>
<td>Temporal convolution</td>
<td>65.4%</td>
</tr>
</tbody>
</table>
The challenge of applying convnets to videos

• Each video = multiple images: memory constraints
• Cost of convolution increases with dimension: time constraints
Convnets and Videos

• Can’t use pretraining
• Can’t train with long videos
• 3D convolution too expensive
A step back: frame-based convnets

• Idea: train single-frame classifier
  • Can use pretraining
  • Can sample frames from videos
  • Cheap

• *Average* scores over frames

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (UCF 101)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best classical approach</td>
<td>85.9%</td>
</tr>
<tr>
<td>Temporal convolution</td>
<td>65.4%</td>
</tr>
<tr>
<td>Avg-of-frame-scores</td>
<td>72.8%</td>
</tr>
</tbody>
</table>
Bringing back motion

• How do we take motion into account?
• Key idea: use optical flow
• Add flow as additional channel?
  • Can’t use pretraining!
• Idea: Use two-stream architecture
Bringing back motion

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (UCF 101)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best dense trajectories + bag-of-words</td>
<td>85.9%</td>
</tr>
<tr>
<td>Temporal convolution</td>
<td>65.4%</td>
</tr>
<tr>
<td>Avg-of-frame-scores (color)</td>
<td>72.8%</td>
</tr>
<tr>
<td>Avg-of-frame-scores (flow)</td>
<td>81.2%</td>
</tr>
<tr>
<td>Avg-of-frame-scores (both)</td>
<td>86.2%</td>
</tr>
</tbody>
</table>

Beyond frame averaging

• Do better than averaging frames?
  • without increasing cost!

• Idea: video is *sequence* of frames

• Convnet converts each frame into feature vector

• Sequence of feature vectors $\rightarrow$ class scores?
Detour: Sequence modeling
Sequence modeling

• Internal state $h$
• Observations $x$
• See observations over time
• Update state

\[ h_t = f(x_t, h_{t-1}) \]
Problem: vanishing/exploding gradients

\[ h_t = f(x_t, h_{t-1}) \]

\[ \frac{\partial z}{\partial h_{t-1}} = \frac{\partial f(x_t, h_{t-1})}{\partial h_{t-1}} \frac{\partial z}{\partial h_t} \]
“Long Short-term Memory”

• Instead of repeated applications of arbitrary function \( f \)
• Have repeated application of something that does not decay or scale up
  • Addition!
(Not) LSTM - 1

\[ h_t = f(x_t, h_{t-1}) \]

\[ h_t = h_{t-1} + x_t \]

\[
\frac{\partial z}{\partial h_{t-1}} = \frac{\partial z}{\partial h_t}
\]
(Not) LSTM - 2

\[ h_t = f(c_t) \]

\[ c_t = c_{t-1} + x_t \]
(Not) LSTM - 3

\[
h_t = \underline{\times} f(x_t, h_{t-1})
\]

\[
h_t = f(c_t)
\]

\[
c_t = c_{t-1} + g_t
\]

\[
g_t = g(x_t)
\]
(Not) LSTM - 3

\[ h_t = f(x_t, h_{t-1}) \]

\[ h_t = f(c_t) \]

\[ c_t = f_t \odot c_{t-1} + i_t \odot g_t \]

\[ f_t = \sigma(W[h_{t-1}, x_t]) \quad i_t = \sigma(V[h_{t-1}, x_t]) \]

Long Short-Term Memory. Sepp Hochreiter and Jurgen Schmidhuber. In Neural Computation
Using LSTMs for frame accumulation

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (UCF 101)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple average</td>
<td>78.9%</td>
</tr>
<tr>
<td>LSTM</td>
<td>82.3%</td>
</tr>
</tbody>
</table>

Revisiting video classification architectures

Revisiting video classification architectures

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th># Input Frames</th>
<th>Temporal Footprint</th>
<th># Input Frames</th>
<th>Temporal Footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvNet+LSTM</td>
<td>9M</td>
<td>25 rgb</td>
<td>5s</td>
<td>50 rgb</td>
<td>10s</td>
</tr>
<tr>
<td>3D-ConvNet</td>
<td>79M</td>
<td>16 rgb</td>
<td>0.64s</td>
<td>240 rgb</td>
<td>9.6s</td>
</tr>
<tr>
<td>Two-Stream</td>
<td>12M</td>
<td>1 rgb, 10 flow</td>
<td>0.4s</td>
<td>25 rgb, 250 flow</td>
<td>10s</td>
</tr>
<tr>
<td>3D-Fused</td>
<td>39M</td>
<td>5 rgb, 50 flow</td>
<td>2s</td>
<td>25 rgb, 250 flow</td>
<td>10s</td>
</tr>
<tr>
<td>Two-Stream I3D</td>
<td>25M</td>
<td>64 rgb, 64 flow</td>
<td>2.56s</td>
<td>250 rgb, 250 flow</td>
<td>10s</td>
</tr>
</tbody>
</table>
Pre-training 3D convolutional networks
Pre-training 3D convolutional networks
Revisiting video classification architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>UCF-101</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
</tr>
<tr>
<td>(a) LSTM</td>
<td>81.0</td>
</tr>
<tr>
<td>(b) 3D-ConvNet</td>
<td>51.6</td>
</tr>
<tr>
<td>(c) Two-Stream</td>
<td>83.6</td>
</tr>
<tr>
<td>(d) 3D-Fused</td>
<td>83.2</td>
</tr>
<tr>
<td>(e) Two-Stream I3D</td>
<td>84.5</td>
</tr>
</tbody>
</table>
Towards better video datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Actions</th>
<th>Clips</th>
<th>Total</th>
<th>Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMDB-51 [15]</td>
<td>2011</td>
<td>51</td>
<td>min 102</td>
<td>6,766</td>
<td>3,312</td>
</tr>
<tr>
<td>UCF-101 [20]</td>
<td>2012</td>
<td>101</td>
<td>min 101</td>
<td>13,320</td>
<td>2,500</td>
</tr>
<tr>
<td>Kinetics</td>
<td>2017</td>
<td>400</td>
<td>min 400</td>
<td>306,245</td>
<td>306,245</td>
</tr>
</tbody>
</table>

Other video problems

- Action “detection / localization” in time
- Temporally consistent object detection / semantic segmentation
- Tracking
Vision and Language
Image captioning - The task

A group of young men playing soccer.
Image captioning - why?

• Alt-text for visually impaired
• Test for true understanding?
Image captioning - evaluation

- Given computer-generated caption and human caption, compute match
- BLEU from machine translation community
- Computes (modified) n-gram precision

Reference: A group of people playing soccer
Candidate: People playing baseball.
BLEU: 1/3
Image captioning


Show and tell: A neural image caption generator
Generating sequences with LSTMs

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \ldots \]
Generating sequences with LSTMs
Generating sequences with LSTMs
Evaluation Metrics

![Evaluation Metrics Graph](image)

Slide credit: Larry Zitnick
Evaluation Metrics

Human captions

Slide credit: Larry Zitnick
A man riding a wave on a surfboard in the water.
A man riding a wave on a surfboard in the water.

“surfboard”
The post-captioning world

• Captioning is hard to evaluate!
  • Frame task so that it is easy to evaluate objectively

• Datasets are biased!
  • Control dataset bias
I'm going to crush the rebellion... but first, let me take a selfie. #captionbot

I am not really confident, but I think it's a man taking a selfie in front of a building.
Reasoning

• Want vision systems to reason about what is going on
  • Identify objects and scenes
  • Identify relationships between objects
  • Understand physics of the world
  • Understand social interactions, intent etc.
  • Incorporate knowledge: common sense, pop culture, ...
Visual Question Answering

• Direct motivation: assistive technology
• Indirect motivation: sandbox for reasoning

“We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people’s expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot! Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image.”
Methods for VQA

How many bikes are there?
Methods for VQA

The Unreasonable Effectiveness of Baselines

Compositional reasoning

What is the color of the kitten to the left of the blue kitten?
What is the color of the kitten to the left of the blue kitten?
Compositional reasoning

- Look left
- Triangle detector
- Blue detector
- And
- Get color

Red
Compositional reasoning

What is the color of the kitten to the left of the blue kitten?
Compositional reasoning

What is the color of the kitten to the left of the blue kitten?

- Look left
- Kitten detector
- Blue detector
- And
- Get color
- Kitten detector
- Blue detector
- And
- Look left
- Get color
Compositional reasoning

- How do we learn a mapping from language to trees?
  - Problem: semantic parsing
  - Option 1: Syntactic parsing
  - Option 2: Use supervision

*Neural module networks.* Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. CVPR 2016


*Learning to reason: End-to-end module networks for visual question answering.* Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko. ICCV 2017

*Inferring and Executing Programs for Visual Reasoning.* Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. ICCV, 2017
Compositional reasoning

**Question:** Are there more cubes than yellow things?

**Answer:** Yes

Program Generator

- Are
- there
- more
- cubes
- than
- yellow
- things?

Execution Engine

- greater than
- count
- filter color [yellow]
- <SCENE>
- count
- filter color [yellow]
- filter shape [cube]

CNN

Predicted Program
The problem with VQA

• Dataset biases allow cheating
  • Only-question Bag-of-Words: 53.7% (vs ~65% for state-of-the-art)

• Require common sense to answer
  • “What is the moustache made of?”

• Hard to diagnose error
  • Is the problem understanding the question?
  • Or understanding the image?
Clever Hans
Current state of vision and language

• Still an active area of research
• Much better datasets but...
• Non-compositional models still win out
• The search for a better task continues