CVPR 2018
Salt Lake City

Temporal Reasoning in Videos using Convolutional Gated Recurrent Units

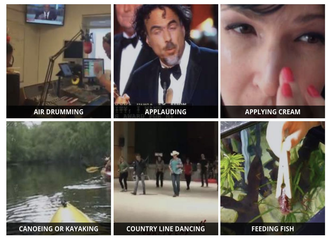
Debidatta Dwibedi, Pierre Sermanet, Jonathan Tompson

{debidatta, sermanet, tompson} @ google.com

Google

Problem Setting

Action Recognition in Videos



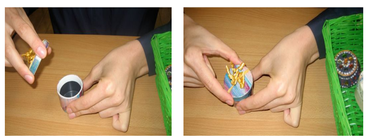
Kinetics

UCF101

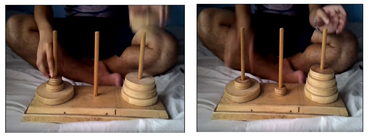
It is possible to predict action in video from single frame

Human-object Interaction videos

Opening



Closing



Transfer from left to middle

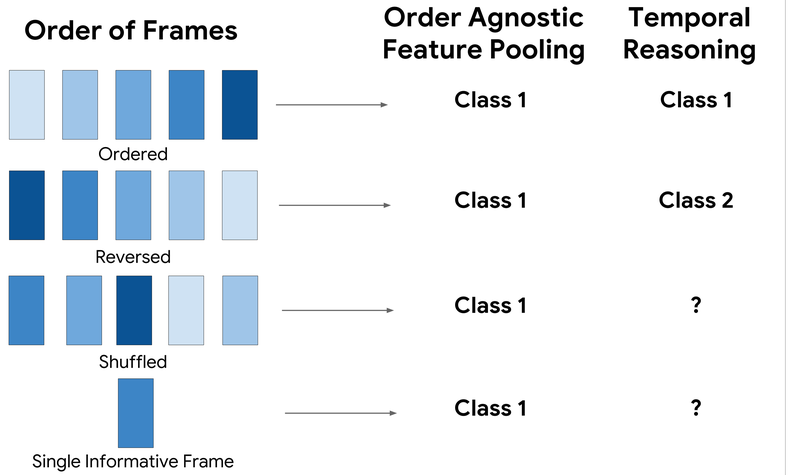
Transfer from middle to left

Action in video is ambiguous if only one frame is considered

Research Questions

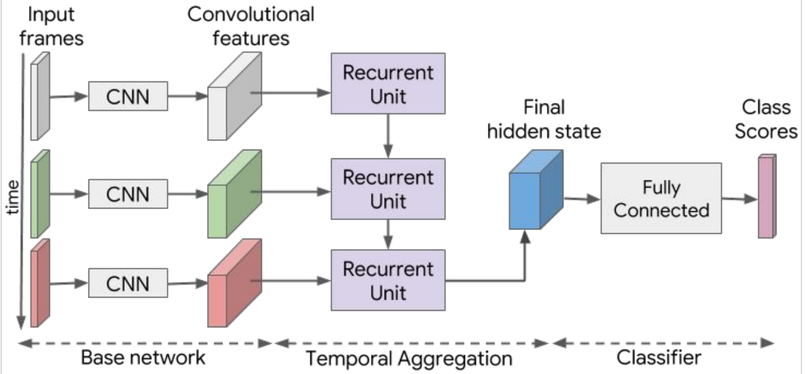
1) **Are all action recognition problems equivalent?**
2) **Will there be one architecture for all action recognition tasks?**
3) **What is encoded in the hidden state of the recurrent units?**

Action Recognition Problems



Our Solution: RNNs for Action Recognition

Recurrent Units = LSTM, GRU, ConvLSTM, ConvGRU



Quantitative Results

Kinetics

Method	Accuracy @ 1	Accuracy @ 5
Spatio-temporal Averaging	71.5	89.5
GRU	70.6	88.4
ConvGRU	70.0	88.1
I3D	**71.6**	**90.2**

Recurrent Units do not provide performance boosts when problem is not sequential

Human-object Interaction videos

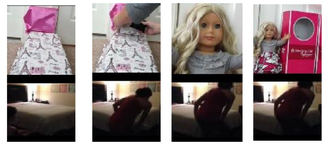
Method	Accuracy @ 1	Accuracy @ 5
Spatio-temporal Averaging	20.5	48.2
GRU	35.4	63.3
I3D (Kinetics pre-trained)	39.9	67.8
ConvGRU	43.7	71.4
ConvGRU (Large)	**45.9**	**74.5**

Recurrent Units are effective for temporal reasoning tasks

Qualitative Results

Action-conditioned nearest neighbours give us interesting insights into what is encoded in the hidden states of the recurrent units.

Folding a napkin



Making a bed

Left video: State changes as video proceeds aligning similar frames
Right video: State is fixed even as the video proceeds, ignoring unnecessary frames

Nearest neighbours in hidden states encode meaningful state transitions