Learning from Unlabeled Video

Carl Vondrick
Columbia University
Survivor Bias of Video Data
Survivor Bias of Video Data

mountain unicycling: 0.280
canyoning: 0.164
base jumping: 0.124
Survivor Bias of Video Data

- whitewater kayaking: 0.272
- base jumping: 0.247
- kayaking: 0.241
The **Oops!** dataset
Oops! Predicting Unintentional Action
CVPR 2020
oops.cs.columbia.edu

Learning from unlabeled video
Perceptual Clues

1) Predictability
   Ranzato 2014, Han 2019, …

2) Temporal Order
   Misra 2016, Wei 2018, …
3) Video speed as self-supervised clue

The key piece of evidence was a surveillance video of the shooting, which the jury saw both in real time and in slow motion. The jury found that Mr. Lewis had acted with premeditation, and he was sentenced to death.

Mr. Lewis appealed the decision, arguing that the slow-motion video was prejudicial. Specifically, he claimed that watching the video in slow motion artificially stretched the relevant time period and created a “false impression of premeditation.” Did it?
3) Video speed as self-supervised clue

Visualizing Features

Query  Nearest Neighbors

animals

sports

leaning over

crying/kid

Fit linear model to classify intentionality
What’s missing?

![Graph showing error comparison across different categories]

- Environmental
- Unexpected
- Multi-agent
- Limited Skill
- Planning Error
- Single-agent
- Execution Error
- Limited Visibility
- Limited Knowledge

Legend:
- Human
- Ours (self-supervised)
- Kinetics (supervised)

Error (lower is better)
Natural Synchronization

Vision

Speech
Ackee seems to be:
• edible
• white/yellow
• washable
• sticky
• larger than cherry tomato

“I’m going to go in with the actual ackee I rinsed off earlier”
Word Learning from Vision

VisualBERT, VILBERT, VideoBERT, LXMERT, …

Learn what “stir” means → Learn how to learn what “stir” means

“I turn on the fire and then I [???] the pasta”
Learning to Learn Words

An overturned 🎌 on the beach...

A cat watches a toy airplane...

A woman on a chair with...

Transformers as Meta-Learners

Transformers as Meta-Learners

Implement with cross entropy loss

Meta-Learning Episodes

New Words Episode

- **Test example**: An overturned ? on the beach...
- **Training example**: A cat watches a toy airplane...
- **Training example**: A woman on a chair with...

Composition Episode

- **Test example**: ? ?
- **Training example**: Next, we stir the pasta
- **Training example**: I'm going to cut the paneer into cubes

Mode 1: Language Modeling

Mode 2: Word Acquisition

Language Modeling

Language Modeling

- BERT pretrained: 19% drop
- BERT + vision: 18% drop
- Meta-Learned: 11% drop

Word Acquisition

Training Set

get **avocado**

still taking **skin** off fish with a knife

stir rice into pan

Test Example

**avocado**

Word Acquisition

Training Set
- Open the cupboard
- Wash plates with rag
- Close oven

Test Example
- Switch off oven on the bottom right

Novel word acquisition

Visualizing Learned Process

Reference Example 1
- Text
- Image

Reference Example 2
- Text
- Image

Target Example
- Text
- Image

Visualizing Attention
Green boxes impact green prediction the most

Training Set
- close food container

Test
- rinse container
- put spoon
- cut cherry tomatoes

expert.cs.columbia.edu
Learning from Unlabeled Video

Carl Vondrick
Columbia University