Learning to Segment Actions from Observation and Narration

Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh
Task: make pancakes: \{add egg, add flour, ..., pour batter, remove pancake\}

Actions:

- background
- pour batter
- background
- remove pancake

Video:

Narration: hey folks here welcome to my kitchen ... pour a nice-sized amount ... change the angle to show ... and take it out

Challenges: visual diversity, noisy narration, varied task structure
How little supervision can we get away with?
Training Without Segment Labels

make pancakes: \{add egg, add flour, ..., pour batter, remove pancake\}

Actions: \(a\)

Video features: \(x\)

Hey folks here welcome to my kitchen... pour a nice-sized amount...

Generative: \[\max_{\theta} \sum_a p_\theta(a, x)\] [Richard et al. 2018, Sener and Yao 2018]

Discriminative: \[\max_{\theta, a} p_\theta(a|x)\] [Alayrac et al. 2016, Zhukov et al. 2019]

Weak-supervision for \(a\):
- Likely ordering of the actions
- Time-aligned narration
How little supervision can we get away with?

*Define a model that allows flexible training.*
Semi-Markov Model

Actions

Video
Semi-Markov Model

\[ \prod_{k} p(s_k | s_{k-1}) \]

Segments, s:

Actions

Video
Semi-Markov Model

\[ \prod_{k} p(s_k | s_{k-1}) p(\text{len}(s_k) | s_k) \]

Segments, \( s \):
- \( s_1 \)
- \( s_2 \)
- \( s_k \)...

Actions
- \( l_1 \)
- \( l_2 \)
- \( l_3 \)
- \( l_4 \)
- \( l_5 \)
- \( l_t \)
- \( l_{t+1} \)

Labels, \( l \):
- background
- pour batter
- remove pancake

Video
Semi-Markov Model

\[ p(s, l, x) = \prod_{k} p(s_k | s_{k-1}) p(\text{len}(s_k) | s_k) \prod_{t} p(x_t | l_t) \]

Segments, \( s \):

- \( s_1 \) background
- \( s_2 \) pour batter
- \( \ldots \)
- \( s_k \) remove pancake

Actions

Labels, \( l \):

- \( l_1 \)
- \( l_2 \)
- \( l_3 \)
- \( l_4 \)
- \( l_5 \)
- \( \ldots \)

Features, \( x \):

- \( x_1 \)
- \( x_2 \)
- \( x_3 \)
- \( x_4 \)
- \( x_5 \)
- \( \ldots \)

Video
CrossTask Dataset [Zhukov et al. 2019]

- 2,700 instructional YouTube videos, with transcribed narration
- 18 household tasks, e.g. cooking, changing a tire, assembling furniture
- Features from ConvNets trained on other related tasks
  - Action recognition [Carreira and Zisserman 2017; Kay et al. 2017]
  - Object classification [He et al. 2016; Russakovsky et al. 2015]
  - Audio classification [Simonyan and Zisserman 2015; Abu-El-Haija et al. 2016]
Evaluation

Two main metrics from past work:

- **Timestep accuracy (1-second intervals)** [Sener and Yao 2018, Richard et al. 2018, inter alia]
- **Action recovery (with one timestep per action)** [Alayrac et al. 2016, Zhukov et al. 2019]
How little supervision can we get away with?

First, compare models in a supervised setting.
Supervised Training

Generative: $p(a, x)$

Discriminative: $p(a|x)$

Structured: Semi-Markov model

Unstructured: Independent classifier at each time-step
## Supervised Results

<table>
<thead>
<tr>
<th>Non-Background Timestep Accuracy</th>
<th>Actions Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample labels from training</td>
<td>7%</td>
</tr>
<tr>
<td>Uniform length segments</td>
<td>8%</td>
</tr>
</tbody>
</table>
## Supervised Results

<table>
<thead>
<tr>
<th>Non-Background Timestep Accuracy</th>
<th>Actions Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample labels from training</strong></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Uniform length segments</strong></td>
<td></td>
</tr>
<tr>
<td>8%</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Unstructured discriminative</strong></td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td>28%</td>
</tr>
<tr>
<td><strong>Unstructured generative</strong></td>
<td></td>
</tr>
<tr>
<td>41%</td>
<td>32%</td>
</tr>
</tbody>
</table>

Uniform length segments
Supervised Results

### Non-Background Timestep Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Sample labels from training</th>
<th>Uniform length segments</th>
<th>Unstructured discriminative</th>
<th>Unstructured generative</th>
<th>Structured discriminative</th>
<th>Structured generative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>7%</td>
<td>8%</td>
<td>30%</td>
<td>41%</td>
<td>37%</td>
<td>49%</td>
</tr>
</tbody>
</table>

### Actions Recovered

<table>
<thead>
<tr>
<th>Category</th>
<th>Sample labels from training</th>
<th>Uniform length segments</th>
<th>Unstructured discriminative</th>
<th>Unstructured generative</th>
<th>Structured discriminative</th>
<th>Structured generative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>8%</td>
<td>12%</td>
<td>28%</td>
<td>32%</td>
<td>24%</td>
<td>29%</td>
</tr>
</tbody>
</table>
How little supervision can we get away with?

Train the structured, generative model without labels.
Training without Segment Labels

\[ p(x) = \sum_{a_{1:T}} \prod_{t} p(a_t|a_{t-1}) p(x_t|a_t) \]

Maximize \( \log p(x) \) (gradient-based soft EM [Eisner 2016])

- **Ordering Supervision**
  Use a typical ordering of steps for each task, *e.g.* add flour, add sugar, ... [Zhukov et al.]
  Constrain \( p(a_t|a_{t-1}) \) to enforce this ordering over segments in all videos for the task

- **Narration Supervision**
  Use label—narration similarity and time alignment to constrain labels [Zhukov et al.]
  In training, constrain the sum over label assignments \( \sum_{a_{1:T}} \)
Training without Segment Labels

Task: *make pancakes*

- BKG
- add flour
- add sugar
- whisk mixture
- pour milk
- pour egg
- pour mixture into pan
- flip pancake
- take pancake from pan

Ground truth

Zhukov et al. (w/ Order & Narr)
Task: *make pancakes*

- **Ground truth**
- **Zhukov et al. (w/ Order & Narr.)**
- **Semi-Markov, no Order or Narr.**

The diagram shows the time (seconds) for each step in the process, from 0 to 350 seconds.
Training without Segment Labels

Task: *make pancakes*

- **Ground truth**
- **Zhukov et al. (w/ Order & Narr)**
- **Semi-Markov, no Order or Narr.**
- **Semi-Markov w/ Order & Narr.**

- BKG
- add flour
- add sugar
- whisk mixture
- pour milk
- pour egg
- pour mixture into pan
- flip pancake
- take pancake from pan

<table>
<thead>
<tr>
<th>time (seconds)</th>
<th>0</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
</tr>
</thead>
<tbody>
<tr>
<td>BKG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add flour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add sugar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>whisk mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pour milk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pour egg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pour mixture into pan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flip pancake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>take pancake from pan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Training without Segment Labels

<table>
<thead>
<tr>
<th>Non-Background Timestep Accuracy</th>
<th>Actions Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ordered uniform</strong></td>
<td>8%</td>
</tr>
<tr>
<td><strong>Zhukov et al.</strong></td>
<td>2%</td>
</tr>
<tr>
<td><strong>Semi-Markov +Order</strong></td>
<td>8%</td>
</tr>
<tr>
<td><strong>Semi-Markov +Order +Narr.</strong></td>
<td>16%</td>
</tr>
</tbody>
</table>

Zhukov et al.
Effects of Supervision

- Zhukov et al. (discriminative)
- Supervised: Unstructured
- Supervised: Structured
- Ordering Constraints
- Ordering+Narration
- Semi-Markov (generative)
- Baselines
How little supervision can we get away with?

Weak supervision from narration helps substantially!
Thank you!

QA Sessions:
Monday, July 6. 4B: Language Grounding-1. 18:00-19:00 UTC+0
Monday, July 6. 5B: Language Grounding-2. 21:00-22:00 UTC+0

github.com/dpfried/action-segmentation