Learning About the World from Language and Learning Language by Observing the World

Aida Nematzadeh

DeepMind
Learning About the World from Language

I am going to make pancakes for you. I have all the ingredients and now need to mix them together.

When does mixing begin and end?
Learning Language by Observing the World

I need to mix the eggs with the flour.

Je casse les oeufs.
Can we leverage this interaction between vision and language?
Obtaining real-world multimodal datasets of natural language and human activities is really hard.
Task: **Make Pancakes**

Narration: *Making pancakes is easy... first, you’ll need to mix sifted flour, sugar...*
Why Are Instructional Videos Interesting?

Some correspondence between language & video.

Naturalistic language and videos.

- Talking about things irrelevant to the task.
- Filmed in a variety of settings (not in a lab).
Many more domains (not just cooking).
Much larger datasets.
Multiple languages.
to Segment Actions
Learning About the World from Language
Why Segmenting Actions?

Crucial to understanding the world, remembering things, and planning.

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
The CrossTask Dataset [Zhukov et al., 2019]

Task: Make Pancakes

Steps: {add flour, add egg, whisk mixture, pour mixture ...}

Video:

| background | pour batter | background | remove pancake |

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

~70 percent of the videos is background regions.

Supervision: task ordering or set of steps
to Segment Actions
Learning About the World from Language

Do *task orderings* & *narrations* help unsupervised action segmentation?
Training Without Segment Labels

Video features:

 Actions $\alpha$:
 Video features $\chi$:

 Generative: $\max_{\theta} \sum_{\alpha} p_{\theta}(a, x)$  

 Discriminative: $\max_{\theta, \alpha} p_{\theta}(a|x)$

Weak-supervision for $a$:

- Likely ordering of the actions
- Time-aligned narration

[Richard et al. 2018, Sener and Yao 2018]

[Alayrac et al. 2016, Zhukov et al. 2019]
Structured:
Semi-Markov model

Unstructured:
Independent classifier at each time-step
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:
- \( s_1 \) → \( s_2 \)

Actions
- background
- pour batter
- remove pancake

Labels, l:
- \( l_1 \)
- \( l_2 \)
- \( l_3 \)
- \( l_4 \)
- \( l_5 \)
- \( l_t \)
- \( l_{t+1} \)

Features, x:
- \( x_1 \)
- \( x_2 \)
- \( x_3 \)
- \( x_4 \)
- \( x_5 \)
- \( x_t \)
- \( x_{t+1} \)

Video
Structured:
Semi-Markov model

Unstructured:
Independent classifier at each time-step

Generative: $p(a, x)$

Discriminative: $p(a|x)$
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]

Ground Truth: [Graphical representation with shaded and unshaded circles indicating accuracy and action recovery]
### Baselines

<table>
<thead>
<tr>
<th></th>
<th>Richard et al. 2018</th>
<th>Zhukov et al. 2019</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>generative model</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>step reordering</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>step repetitions</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>step duration</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Picked models with non-overlapping strengths.
### Supervised Results

<table>
<thead>
<tr>
<th></th>
<th>timestep accuracy</th>
<th>actions recovered</th>
<th>predicted bg %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordered uniform</td>
<td>8.1</td>
<td>12.2</td>
<td>73.0</td>
</tr>
<tr>
<td><strong>Unstructured</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discriminative linear</td>
<td>36.0</td>
<td>31.6</td>
<td>73.3</td>
</tr>
<tr>
<td>✓ Gaussian mixture</td>
<td>40.6</td>
<td>31.5</td>
<td>68.9</td>
</tr>
<tr>
<td><strong>Structured</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhukov et al. (2019)</td>
<td>18.1</td>
<td>45.4</td>
<td>90.7</td>
</tr>
<tr>
<td>SMM, discriminative</td>
<td>37.3</td>
<td>24.1</td>
<td>65.9</td>
</tr>
<tr>
<td>✓ SMM, generative</td>
<td>49.4</td>
<td>28.7</td>
<td>52.4</td>
</tr>
</tbody>
</table>
## Un- & Weakly-Supervised Results

<table>
<thead>
<tr>
<th></th>
<th>timestep accuracy</th>
<th>actions recovered</th>
<th>predicted bg %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>Ordered uniform</td>
<td>8.1</td>
<td>12.2</td>
</tr>
<tr>
<td><strong>Unsupervised</strong></td>
<td>HSMM</td>
<td>28.8</td>
<td>10.6</td>
</tr>
</tbody>
</table>

### Ordering Supervision
- Richard et al. (2018)
  - 14.0 12.1 49.8
- Zhukov et al. (2019) (w/o narr.)
  - 0.2 2.8 97.2
- HSMM + Ord
  - 8.3 7.3 70.6
- Ordering + Narration Supervision
  - Zhukov et al. (2019)
    - 1.8 24.5 97.2
- HSMM + Narr + Ord
  - 15.9 17.2 73.7

Task orderings & narrations prevent over/underpredicting background.
How Should We Model Action Segmentation?

Generative or discriminative models? Generative; discriminative ones overpredict bg.

Explicit structured modeling? Yes, especially for sequence-level metrics.

Do task ordering and narration help unsupervised models? Yes; prevent over/underpredicting bg.
segmentation examples

ground truth
uniform
hsmm
hsmm + narration + ordering
Zhukov et al. (2019)
Instructional Videos: Challenges

coloring by step

coloring by video
What Did We Learn? [Fried et. al, ACL 2020]

→ Narrations & task orderings are important inductive biases for action segmentation.

→ Need to report different metrics to evaluate action segmentation.

→ Deep visual features don’t capture the finer-grained differences between actions.
to Translate Words
Learning Language by Observing the World
Translating Words Without Supervision

I need to mix the eggs with the flour.

Je casse les oeufs.

different videos in each language (no paired data).
Two Ways to Translate Words Through Videos

Create a *paired* corpus using videos and then *discard* the videos,

or,

learn a *joint space* between languages and videos, where the visual encoder is shared.
Creating a Paired Corpus

2. Select the nearest neighbors for each video.
3. Create a paired corpus with the narrations associated to neighboring videos.
4. Calculate joint probability between all En-Fr word pairs.
Our Base Model Architecture

Mix the eggs with the flour

$\mathbf{WordEmbed} \; \mathcal{X}$

$\mathbf{Linear} + \mathbf{ReLU} + \mathbf{MaxPool}$

$[L, \text{d}_w]$

$\mathbf{AdaptLayer}$

Parameters: $W \in \mathbb{R}^{d_w \times d_w}$

$[L, \text{d}_w]$

$\mathbf{WordEmbed} \; \mathcal{Y}$

$y \in \mathcal{Y}$

$\mathbf{Linear}$

$d_i$

Contrastive Loss

Joint Embedding space

$I3D + \mathbf{Linear}$

$[L, \text{d}_w]$

Video

$z \in \mathcal{Z}$

Je casse les oeufs
The HowTo100M Dataset [Miech et al., 2019]

~100M video clips-narrations (ASR output).

~23k high-level tasks.
Does a Shared Visual Encoder Help?

<table>
<thead>
<tr>
<th></th>
<th>Dictionary (Conneau et al., 2017)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td><strong>English-French</strong></td>
<td></td>
</tr>
<tr>
<td>(reporting recall@1)</td>
<td></td>
</tr>
<tr>
<td><strong>Random Chance</strong></td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Paired Corpus</strong></td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Base Model</strong></td>
<td>9.1</td>
</tr>
</tbody>
</table>

Yes, especially for the visual words.
Failures When Pairing Two Languages

Videos are semantically similar but not the narrations.

Videos are semantically less similar.
Why Should We Use Videos at All?

Text-based methods align the space of the two languages.

Unsupervised approaches are possible:

- MUSE (Conneau et al., 2017)
- VecMap (Artetxe et al., 2018)
Why Should We Use Videos at All?

But these methods are sensitive to the similarity of languages and their training corpora.

Can grounding improve unsupervised word translation -- make it more robust?
Multilingual Visual Embeddings (MUVE)

$\mathbf{x} \in \mathcal{X}$

WordEmbed $\mathcal{X}$

$[L, d_w]$

$\mathbf{L} + \text{ReLU}$

MaxPool

$[L, d_w]$

AdaptLayer

Parameters: $W \in \mathbb{R}^{d_w \times d_w}$

$[L, d_w]$

WordEmbed $\mathcal{Y}$

$\mathbf{y} \in \mathcal{Y}$

Contrastive Loss

$\mathbf{d}$

I3D + Linear

Joint Embedding space

$\mathbf{z} \in \mathcal{Z}$

Video

$\mathbf{d}$

Linear

$\mathbf{d}$

$\mathbf{d}$

$+\ MUSE \rightarrow \text{MUVE}$
Extending the HowTo100M Dataset [Miech et al, 2019]
### Performance of Models Across Language Pairs

<table>
<thead>
<tr>
<th>reporting recall@1</th>
<th>En-Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MUSE</strong> (Conneau et al., 2017)</td>
<td>26.3</td>
</tr>
<tr>
<td><strong>VecMap</strong> (Artetxe et al., 2018)</td>
<td>28.4</td>
</tr>
<tr>
<td><strong>MUVE</strong> (ours)</td>
<td>28.9</td>
</tr>
<tr>
<td><strong>Supervised</strong></td>
<td>57.9</td>
</tr>
</tbody>
</table>
## Robustness to Dissimilarity of Corpora

<table>
<thead>
<tr>
<th>Reporting recall@10</th>
<th>MUSE (Conneau et al., 2017)</th>
<th>VecMap (Artetxe et al., 2018)</th>
<th>MUVE (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HowTo-En</td>
<td>45.8</td>
<td>45.4</td>
<td>47.3</td>
</tr>
</tbody>
</table>
Robustness to Amount of Training Data

- **Bar Chart:**
  - Title: "Wiki En-Fr Recall@10"
  - X-axis: "100%, 10%, 1%"
  - Y-axis: Values from 0 to 60
  - Categories: MUSE, VecMap, MUVE

- **Line Graph:**
  - Title: "Vocabulary Size vs. Recall"
  - X-axis: Vocabulary Size from 65k to 500
  - Y-axis: Recall
  - Data points for different vocabulary sizes
Robustness to Amount of Training Data
What Did We Learn? [Sigurdsson et. al, CVPR 2020]

Learning a joint multimodal space (sharing the visual encoder) improves word translations.

Grounding in videos improves text-based methods, especially in challenging situations.
Takeaways

Working with real-world data is an important first step for AI systems -- easy problems for humans are still challenging for our models.

Collecting less-noisy datasets is expensive. We need better ways of removing noise, e.g., through better grounding.
Towards Better Multimodal Features

Recent multimodal transformers show impressive performance on a range of benchmarks.

Do these models provide features that better ground language to vision? [Hendricks et. al, TACL 2021]
Acknowledgments

Daniel Fried, Gunnar Atli Sigurdsson, Jean-Baptiste Alayrac, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Chris Dyer, Stephen Clark, Phil Blunsom, and Andrew Zisserman.

Thanks!