Learning from Narrated Videos

Jean-Baptiste Alayrac
jbalayrac.com

3rd Workshop on YouTube-8M
Large-Scale Video Understanding
28/10/2019
Success of Supervised Learning

Pose estimation
[Towards Accurate Multi-person Pose Estimation in the Wild, Papandreou, Zhu, Kanazawa, Toshev, Tompson, Bregler and Murphy, CVPR17]

Image Segmentation
[Mask R-CNN, He, Gkioxari, Dollár, and Girshisck, ICCV17]
Issues of Supervised Learning

Labels are expensive

Agreement: definition? granularity?
Issues of Supervised Learning

Labels are expensive

Even more problematic for videos!
Weakly supervised learning
Use weaker and readily available source of supervision

Training info: *image level label*

[Barnard et al’03], [Joulin et al’10], [Deselaers et al’12], [Song et al’14], [Wang et al’14], [Cinbis et al’15], [Oquab et al’15], [Kantorov et al’16], [Bilen and Vedaldi’16]...
Weakly supervised learning

*Use weaker and readily available source of supervision*

Training info: **video narration (ASR)**

[Alayrac et al’16/17], [Malmaud et al, 15], [Sener et al’15], [Huang et al’17], [Zhou et al’17], [Kuehne et al’17], ...
What are instructional videos?

- Depict complex, **goal-oriented** human activities (*e.g.* how to change a car tire)
- **Multimodal**: video and language
- Can be obtained at **scale** (*e.g.* on YouTube), without manual annotation
Glossary

Tasks: a complex human activity involving interacting with objects and/or performing multiple small actions.

Example: “make pancakes”, “change a car tire”, ...

Steps: an atomic action composing a task.

Example: “crack egg”, “remove tire”, ...
Overview of the talk

1) Leveraging the structure of narrated videos

Making Meringue
- Pour egg
- Add sugar
- Whisk mixture

Making Pancakes
- Pour mixture

Making Lemonade
- Pour water

Cross-task weakly supervised learning from instructional videos, Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic, CVPR2019

2) Leveraging the scale of narrated videos

HowTo100M: Learning a Text–Video Embedding by Watching Hundred Million Narrated Video Clips, Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic, ICCV2019
Cross-Task Weakly Supervised Learning from Instructional Videos, CVPR19

D. Zhukov* D. Fouhey G. Cinbis I. Laptev J. Sivic
How much can we leverage the structure in narrated videos and what can we get from that?
What do we mean by structure here?

**Task: Make Meringue**

1. **Pour egg**
2. **Add sugar**
3. **Whisk mixture**

**Structure within task**
What do we mean by structure here?

Task: Make Meringue
- pour egg
- add sugar
- whisk mixture

Structure within task

Task: Making Pancakes
- pour mixture

Structure across task

Task: Making Lemonade
- pour water
What do we mean by structure here?

Task: Make Meringue
- pour egg
- add sugar
- whisk mixture

Task: Making Pancakes
- pour water

Task: Making Lemonade
- pour mixture

Structure within task

Structure across task
Weakly supervised learning of step visual models

**Input**

→ A set of **tasks**
  
  ex: “Make Meringue”, “Make Pancakes”, “Change a car tire”, …

→ For each task, a **list of steps**:
  
  ```
  Make Pancake
  1) pour egg
  2) add milk
  3) whisk mixture
  ```

→ For each task, a **set of narrated videos**:
  
  “… now we pour the egg …”

**Output**

→ A visual **classifier** for each step

→ **Localize** each step in all videos
Our assumptions

- **Temporal ordering.** Steps always occur in the order given by the list of steps.

- **At least once.** We assume that for each video, each step occurs once.

- **Video and narration.** Correlation between video and language.
The approach

TL;DR: We jointly (i) learn step classifiers over pretrained visual features and (ii) localize where the steps happen in the video.

Formally: This is done by an alternate optimization between the parameters of the step classifier ($F$) and the localization variable ($Y$) under specific constraints that reflects our assumptions.

\[
\min_{Y \in \mathcal{C}, F \in \mathcal{F}} \sum_{\tau} \sum_{v \in \mathcal{V}(\tau)} h(X^v, Y^v; F)
\]

Video $v \in \tau_1$

Alternate Optimization

Update $F$

Update $Y$

Constraints

Narration "[...] now I'm gonna pour some milk into the bowl and [...]"
Component based model for steps

Shared Components

Tasks $\tau$

Step Classifier

The CrossTask dataset

- Designed to assess the benefit of sharing knowledge across tasks:
  - 18 primary tasks, 2750 videos with full temporal annotation
  - 65 related tasks, 1950 videos without annotation
- Diverse set of tasks: Car maintenance, gardening, cooking, home repair
Results: gains from sharing

Localization metric (recall) on CrossTask

- Uniform
- Task specific step based
- Shared step based model
- Shared component based model
Results: novel task transfer

We train only on related task and transfer to the unseen primary task.
Qualitative results

Source Steps
From Related Tasks

Cut Steak
Cut Tomato
Add Tomato
Add Cherries to Cake

Unseen Task: Make
French Strawberry Cake

Cut Strawberry
Add Strawberry To Cake
HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, ICCV19
How much can we scale Instructional Video dataset and what can we get from that?
The HowTo100M dataset in numbers

- 23K human **tasks** scrapped from WikiHow
- 1.2M unique YouTube **videos** (duration 15 years)
- 136M **clips** with **narration** transcribed into text (mostly from ASR)
- Larger than any existing manually annotated captioning dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Clips</th>
<th>Captions</th>
<th>Videos</th>
<th>Duration</th>
<th>Source</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charades [48]</td>
<td>10k</td>
<td>16k</td>
<td>10,000</td>
<td>82h</td>
<td>Home</td>
<td>2016</td>
</tr>
<tr>
<td>MSR-VTT [58]</td>
<td>10k</td>
<td>200k</td>
<td>7,180</td>
<td>40h</td>
<td>Youtube</td>
<td>2016</td>
</tr>
<tr>
<td>YouCook2 [67]</td>
<td>14k</td>
<td>14k</td>
<td>2,000</td>
<td>176h</td>
<td>Youtube</td>
<td>2018</td>
</tr>
<tr>
<td>EPIC-KITCHENS [7]</td>
<td>40k</td>
<td>40k</td>
<td>432</td>
<td>55h</td>
<td>Home</td>
<td>2018</td>
</tr>
<tr>
<td>DiDeMo [15]</td>
<td>27k</td>
<td>41k</td>
<td>10,464</td>
<td>87h</td>
<td>Flickr</td>
<td>2017</td>
</tr>
<tr>
<td>M-VAD [52]</td>
<td>49k</td>
<td>56k</td>
<td>92</td>
<td>84h</td>
<td>Movies</td>
<td>2015</td>
</tr>
<tr>
<td>MPII-MD [43]</td>
<td>69k</td>
<td>68k</td>
<td>94</td>
<td>41h</td>
<td>Movies</td>
<td>2015</td>
</tr>
<tr>
<td>ANet Captions [26]</td>
<td>100k</td>
<td>100k</td>
<td>20,000</td>
<td>849h</td>
<td>Youtube</td>
<td>2017</td>
</tr>
<tr>
<td>TGIF [27]</td>
<td>102k</td>
<td>126k</td>
<td>102,068</td>
<td>103h</td>
<td>Tumblr</td>
<td>2016</td>
</tr>
<tr>
<td>LSMDC [44]</td>
<td>128k</td>
<td>128k</td>
<td>200</td>
<td>150h</td>
<td>Movies</td>
<td>2017</td>
</tr>
<tr>
<td>How2 [45]</td>
<td>185k</td>
<td>185k</td>
<td>13,168</td>
<td>298h</td>
<td>Youtube</td>
<td>2018</td>
</tr>
<tr>
<td><strong>HowTo100M</strong></td>
<td><strong>136M</strong></td>
<td><strong>136M</strong></td>
<td><strong>1.221M</strong></td>
<td><strong>134,472h</strong></td>
<td>Youtube</td>
<td>2019</td>
</tr>
</tbody>
</table>
How to collect HowTo100M?

Step 1: WikiHow

Result: list of 130k tasks

... How to be healthy
How to cook quinoa in a Rice Cooker
How to Sew an Apron
How to Break a Chain
How to April Fool your Girlfriend
...

Annotation cost: 0
How to collect HowTo100M?
Step 2: Filter task by verb to keep visual tasks

Result: list of 23k tasks

...  
How to Be healthy  
✓ How to Cook quinoa in a Rice Cooker  
✓ How to Sew an Apron  
✓ How to Break a Chain  
How to April Fool your Girlfriend  
...

Annotation cost: 8 hours for Antoine
How to collect HowTo100M?

Step 3: YouTube queries for videos with captions

Result: 1.2 M unique videos

Annotation cost: 0
How to collect HowTo100M?

Step 4: Create clips

Result: 136M narrated clips

Annotation cost: 0
Learning a visual-text embedding on HowTo100M

Pre-trained word2vec word embeddings (dim=300) (No stop words)


Learning a visual-text embedding on HowTo100M

\[ S_{i,j} = S(X_i, Y_j) \text{ (dot product)} \]

\[ \forall (i, j), \ j \neq i, S_{i,i} > S_{i,j}, S_{i,i} > S_{j,i} \]

\[ L = \frac{1}{B} \sum_{i=1}^{B} \sum_{j \neq i} \left[ \max(0, m + S_{i,j} - S_{i,i}) + \max(0, m + S_{j,i} - S_{i,i}) \right] \]
Evaluation procedure

Text to video retrieval: **YouCook2, MSRVTT, LSMDC**

**Answering the phone**

Action localization: **CrossTask**

- Loose bolt
- Jack car
- Remove wheel
Beauty of having a joint text and video embedding: 
*In both cases, we can evaluate without finetuning!*
Within domain: YouCook2 retrieval (YouTube cooking videos)

YouCook2 (R@10)

- Random
- Trained on HowTo100M
- Trained on YouCook2
- PT HowTo100M + Finetune YouCook2
Within domain: CrossTask action localization

Localization metric (recall) on CrossTask

- Uniform
- Task specific step based
- Shared step based model
- Shared component based model
- Trained on HowTo100M
- Supervised CrossTask

Recall
Out of domain: MSRVTT (popular & generic YouTube videos)

R@10

- Random
- Trained on HowTo100M
- Trained on MSR-VTT
- PT HowTo100M + Fine-Tune MSR-VTT
- JSFusion (Previous SoTA)

JSFusion: A joint sequence fusion model for video question answering and retrieval, Yu et al. ECCV2018
Out of domain ++: LSMDC (movies)

R@10

- Random
- Trained on HowTo100M2
- PT HowTo100M + Fine-tune LSMDC
- JSFusion

JSFusion: A joint sequence fusion model for video question answering and retrieval, Yu et al. ECCV2018
Coming back to the original question: scale matters!
Online web demo

http://howto100m.inria.fr/
3 Discussion
Summary

1) Leveraging the structure of narrated videos

Cross-task weakly supervised learning from instructional videos, Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic, CVPR2019

2) Leveraging the scale of narrated videos

HowTo100M: Learning a Text–Video Embedding by Watching Hundred Million Narrated Video Clips, Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic, ICCV2019
Future directions

Dealing with the noise. In 50% of the cases, video and narration are not matching. Something should be done!

Still relying on pretrained features (obtained from Kinetics or ImageNet) the story is not complete.

The dream: end to end learning directly from HowTo100M.