Unsupervised Learning from Narrated Instruction Videos

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, Simon Lacoste-Julien

INRIA/ENS, Paris, France

The problem:
Automatically learn the main steps to complete a given task from narrated instruction videos.

Input: A set of narrated instruction videos.

Approach
Video 2
Experimentally demonstrated
Report a plant
Video 4
Each task is performed by
et al.
Visual and linguistic representations of the steps
transcripts.
Collected and annotated a
Discriminative video
dataset

Experiments

Contributions:
i. Collected and annotated a new dataset of narrated instruction videos,
ii. Developed an unsupervised learning method that takes advantage of the complementary nature of the text and video,
iii. Experimentally demonstrated recovery of main steps and their locations in video.

Outputs:
• List of k (input) main steps
• Visual and linguistic representations of the steps
• Temporal localization of each step in the videos

A new dataset

• 5 tasks:
  - Changing a car tire
  - Report a plant
  - Make coffee
  - Perform CPR
  - Jump a car battery

• 30 videos per task (total of 800,000 frames).
• Manual correction of the ASR transcripts.
• Manual annotation of 7-10 main steps and time localization in each video (only used for evaluation).

Approach

Assumptions
Each task is performed by an ordered sequence of steps.
People do what they say (roughly) when they say it.

Two linked clustering steps
1. Text clustering: multiple sequence alignment
2. Discriminative video clustering under text constraints

Multiple sequence alignment (MSA)
• Text signals are first processed into sequences of direct object relations: Let’s now jack the car.
  
  DOBJ = (jack,car)

INDIVIDUAL INPUT SEQUENCES:

OUTPUT OF THE MSA:
Novel Formulate MSA as a QP, and approximately solve it with Frank-Wolfe.

FINAL OUTPUT:

Discriminative clustering
under text constraints

CONSTRAINED OPTIMIZATION PROBLEM:

\[
\min_{Z} \quad h(Z) \\
\text{s.t.} \\
Z \in \mathbb{Z}^{S \times T} \\
|Z| \geq 1 \\
\text{where} \quad h(Z) \quad \text{is a discriminative clustering cost} \quad [1]:
\]

\[
\begin{align*}
& h(Z) = \min_{Z} \quad \frac{1}{2} \left\| W \right\|_2^2+ \frac{1}{2} \left\| W \right\|_1 + \lambda \sum_{i=1}^{K} \left\| \sum_{j \in S} d_{ij} \right\|_1 \\
& \text{where} \quad W = [W_{ij}] \quad \text{is the} \quad K \times S \text{weight matrix,} \\
& d_{ij} = \text{direct object distance of} \quad (i,j) \quad \text{matrix}.
\end{align*}
\]

OPTIMIZATION METHOD [2,3]:
• Optimize convex relaxation using Frank-Wolfe
• Use DP as the linear oracle
• Cost classifier based rounding

References

Check out our project webpage for code/data!