Identifying Visible Actions in Lifestyle Vlogs

Oana Ignat (oignat@umich.edu), Laura Burdick, Jia Deng, Rada Mihalcea

UNIVERSITY OF MICHIGAN

Introduction

Task

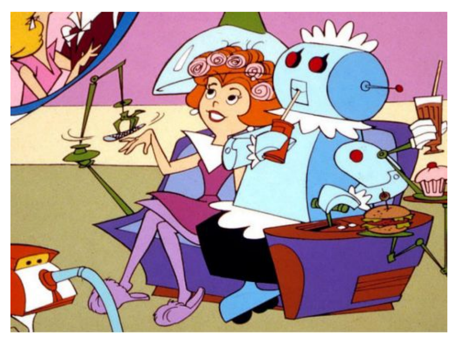
Given a video and its transcript, which human actions are visible in the video?

...
03:24 you're gonna actually cook it
03:27 and it you're gonna bake it for
03:30 about six hours it's definitely a
03:32 long time so keep in mind that it's
03:34 basically just dehydrating it
03:50 after what seems like an eternity in
03:53 the oven you're going to take it out
03:55 it's actually dehydrated at that point
03:57 which is fabulous because you can
03:59 pull it right off the baking sheet and
04:01 you're going to put it on to some
04:03 parchment paper and then you're
...

Action	Visible?
actually cook it	✓
bake it for	✓
take it out	✓
pull it right off the baking sheet	✓
put it on to some parchment paper	✓
so keep in mind that	x
seems like an eternity in the oven	x
dehydrated at that point which	x

Applications

- Video summarization
- Video-action mapping
- Action prediction



Proposed Solution

1. **Extract the actions** from the transcripts using a parser
2. Create a **dataset** with crowdsourced manual **annotations of visible actions** in videos
3. Evaluate a set of **single-modality baselines**:
   a. Text-based
   b. Video-based
4. Build a **multi-modal model** that combines visual and linguistic information

Mechanical Turk Task Description

- **Five miniclips** per task
- Up to **seven actions** per miniclip
- Each miniclip annotated by **3 workers**
- Last miniclip pre-labeled:
  - Two reliable annotators
  - Use it as **ground truth**

Is the action visible in the video?

seems like an eternity in the oven	○ Yes	● No	○ Not an action
take the tray out	● Yes	○ No	○ Not an action
dehydrated at that point which	○ Yes	● No	○ Not an action
pull it right off the baking sheet	● Yes	○ No	○ Not an action

Methods

Data Gathering Pipeline

1. Filter videos based on movement and text

Based on text

do not contain transcripts or # words / second < 0.5

Based on movement

2D correlation coefficient

Transcript

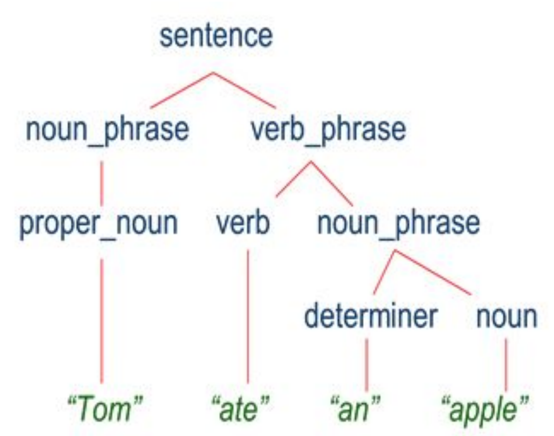
03:38 to try it out so you're adding all the
03:39 herbs in a mason jar and then you're
03:41 adding hot water and then I'm going to
03:43 put some cheesecloth over the top

English (auto-generated)

Try it out	3:38
Adding all the herbs in a mason jar	3:39
Adding hot water	3:41
Put some cheesecloth over the top	3:43

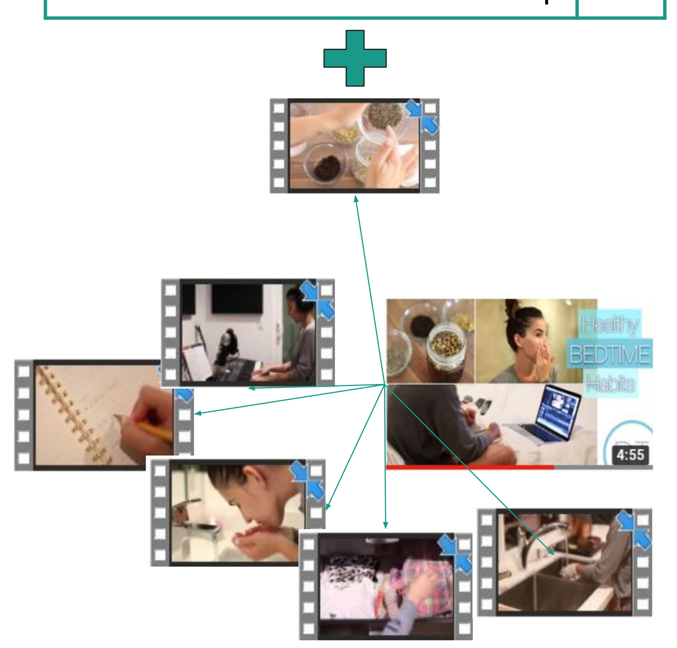
2. Extract Actions

Stanford constituency parser to extract verb phrases



3. Generate Miniclips

- Map actions to miniclips according to the **time** they appear in the transcript
- Misalignment, use a **time window (± 15 seconds)**

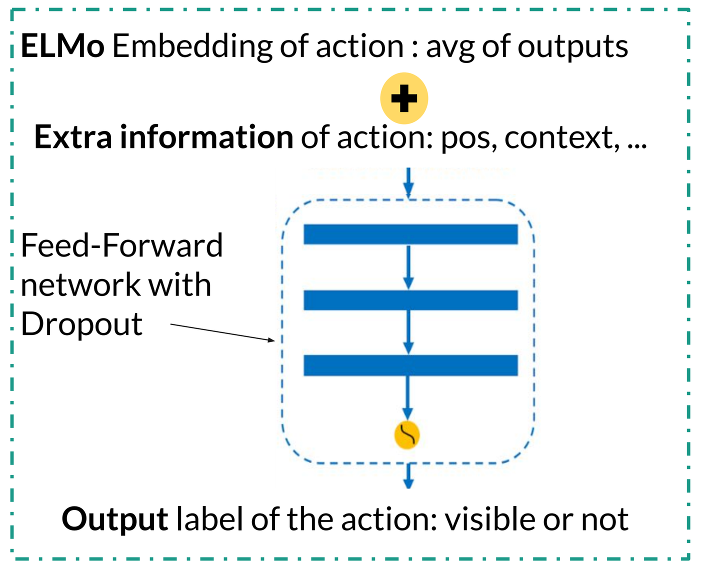


Data Representation

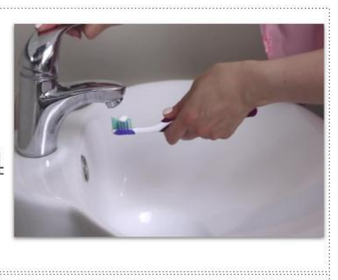
- Action Embeddings
- Part of Speech (POS) Embeddings
- Context Embeddings
- Concreteness Score
- Frame-level: Inception V3
- Sequence-level: C3D

Textual (Action Embeddings, POS Embeddings, Context Embeddings, Concreteness Score)

Visual (Frame-level: Inception V3, Sequence-level: C3D)

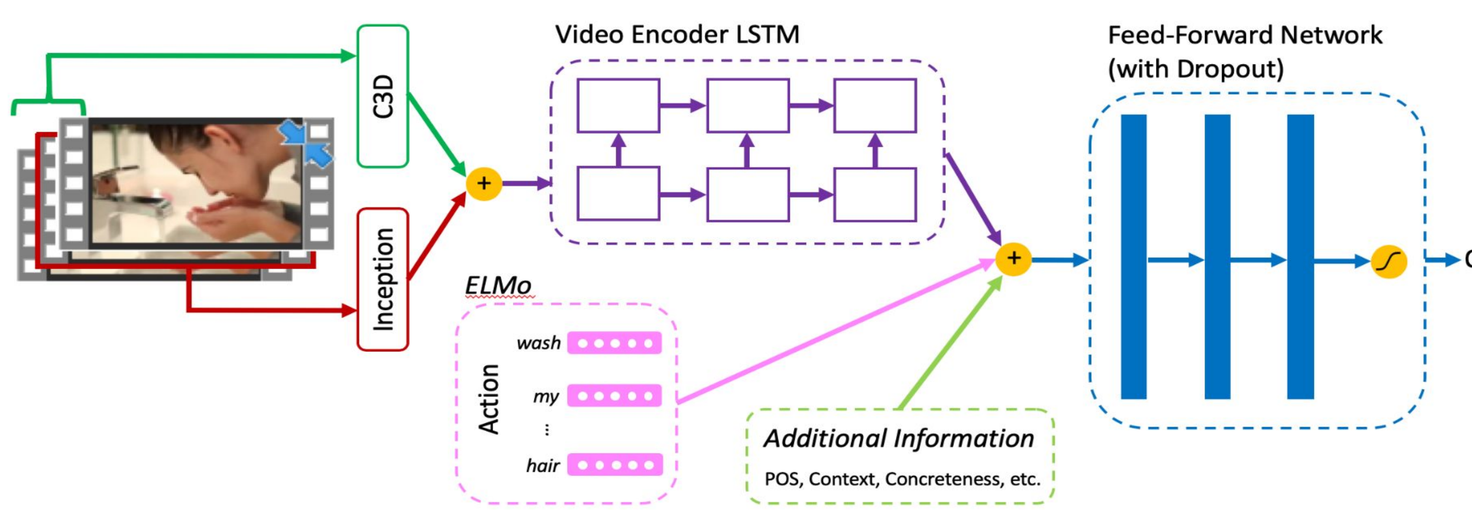


Action: brush my teeth
Object detected: toothbrush
similarity(brush, toothbrush) = 0.94



Action: chop my vegetables
Object detected: carrot
similarity(vegetables, carrot) = 0.9

Overview of Multi-modal Architecture



Results

Related Datasets

Dataset	#Actions	#Verbs	Implicit	Labels
Ours	4340	580	✓	✓
VLOG (Fouhey et al., 2018)	-	-	✓	✓
Kinetics (Kay et al., 2017)	600	270	x	x
ActivityNet (Fabian Caba Heilbron and Niebles, 2015)	203	-	x	x
AVA (Gu et al., 2017)	80	80	✓	x
Charades (Sigurdsson et al., 2016)	157	30	x	x

*# Actions: # action classes (other datasets) or # unique visible actions (ours); #Verbs: # unique verbs in the actions; **Implicit** vs. **Explicit** data gathering; **Labels** refers to label type: post-defined: ✓, pre-defined: x*

Data Statistics

Videos	177
Video hours	21
Transcript words	302,316
Miniclips	1,268
Actions	14,769
Visible actions	4,340
Non-visible actions	10,429

Data Split

- One Youtube channel for **Test**
- One Youtube channel for **Validation**
- The rest (8 channels) for **Training**

	Train	Test	Validation
# Actions	11,403	1,999	1,367
# Miniclips	997	158	113

Evaluation

Method	Input	Accuracy	Precision	Recall	F1
Baselines					
Majority	Action	0.692	0.692	1.0	0.81
Threshold	Concreteness	0.685	0.7	0.954	0.807
Feature-based Classifier	$Action_G$	0.715	0.722	0.956	0.823
	$Action_G$, POS	0.701	0.702	**0.986**	0.820
	$Action_G$, $Context_S$	0.725	0.736	0.938	0.825
	$Action_G$, $Context_A$	0.712	0.722	0.949	0.820
	$Action_G$, Concreteness	0.718	0.729	0.942	0.822
	$Action_G$, $Context_S$, Concreteness	**0.728**	**0.742**	0.932	**0.826**
LSTM	$Action_G$	0.706	0.753	0.857	0.802
ELMo	$Action_G$	**0.726**	**0.771**	**0.859**	**0.813**
YOLO	Miniclip	0.625	0.619	0.448	0.520
Multimodal Neural Architecture (Figure 5)					
Multi-modal Model	$Action_E$, Inception	0.722	0.765	0.863	0.811
	$Action_E$, Inception, C3D	0.725	0.769	0.869	0.814
	$Action_E$, POS, Inception, C3D	0.731	0.763	0.885	0.820
	$Action_E$, $Context_S$, Inception, C3D	0.725	**0.770**	0.859	0.812
	$Action_E$, $Context_A$, Inception, C3D	0.729	0.757	0.895	0.820
	$Action_E$, Concreteness, Inception, C3D	0.723	0.768	0.860	0.811
	$Action_E$, POS, $Context_S$, Concreteness, Inception, C3D	**0.737**	0.758	**0.911**	**0.827**

$Action_G$ *indicates action representation using GloVe embedding, and* ***$Action_E$*** *indicates action representation using ELMo embedding.* ***$Context_S$*** *indicates sentence-level context, and* ***$Context_A$*** *indicates action-level context.*

Download

*The **dataset** and the **code** introduced in this paper are publicly available at lit.eecs.umich.edu/downloads.html*

Acknowledgements

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), by the John Templeton Foundation (grant #61156), and by DARPA (grant #HR001117S0026-AIDA-FP-045).