WhyAct: Identifying Action Reasons in Lifestyle Vlogs

Oana Ignat, Santiago Castro, Hanwen Miao, Weiji Li, Rada Mihalcea
oignat@umich.edu

Motivation

Action Recognition systems with known intent perform better. Knowing the reason for performing an action is an important step for **understanding that action.**

Causal reasoning has **direct applications** on many real-life settings, for instance to **understand the consequences of events** (e.g., *if there is clutter, cleaning is required*), or to **enable social reasoning** (e.g., *when guests are expected, cleaning may be needed*) -- see Fig 1.

Data Collection

YouTube — **Lifestyle Vlogs:**

*my morning **routine***
*my everyday **routine***
...

ConceptNet — *Clean* **is motivated by ...**

company was coming
remove dirt
...

Data Pre-processing

Reason Clustering
Transcript Filtering
Video Filtering

Initial	9,759
Actions with reasons in ConceptNet	139
Actions with at least 3 reasons in CN	102
Actions with at least 25 video-clips	25

Table 1: Statistics for number of collected actions at each stage of data filtering.

Data Annotations

amazon mechanical turk

3 annotators per video

Moderate Agreement (0.6 kappa)

They are asked to identify:

(1) the reasons **shown** or **mentioned in the video for performing a given action**

(2) are the reasons **mentioned verbally, shown visually**, or **both**

(3) other reasons

(4) confidence

Video-clips	1,077
Video hours	107.3
Transcript words	109,711
Actions	24
Reasons	166

Table 2: Data Statistics

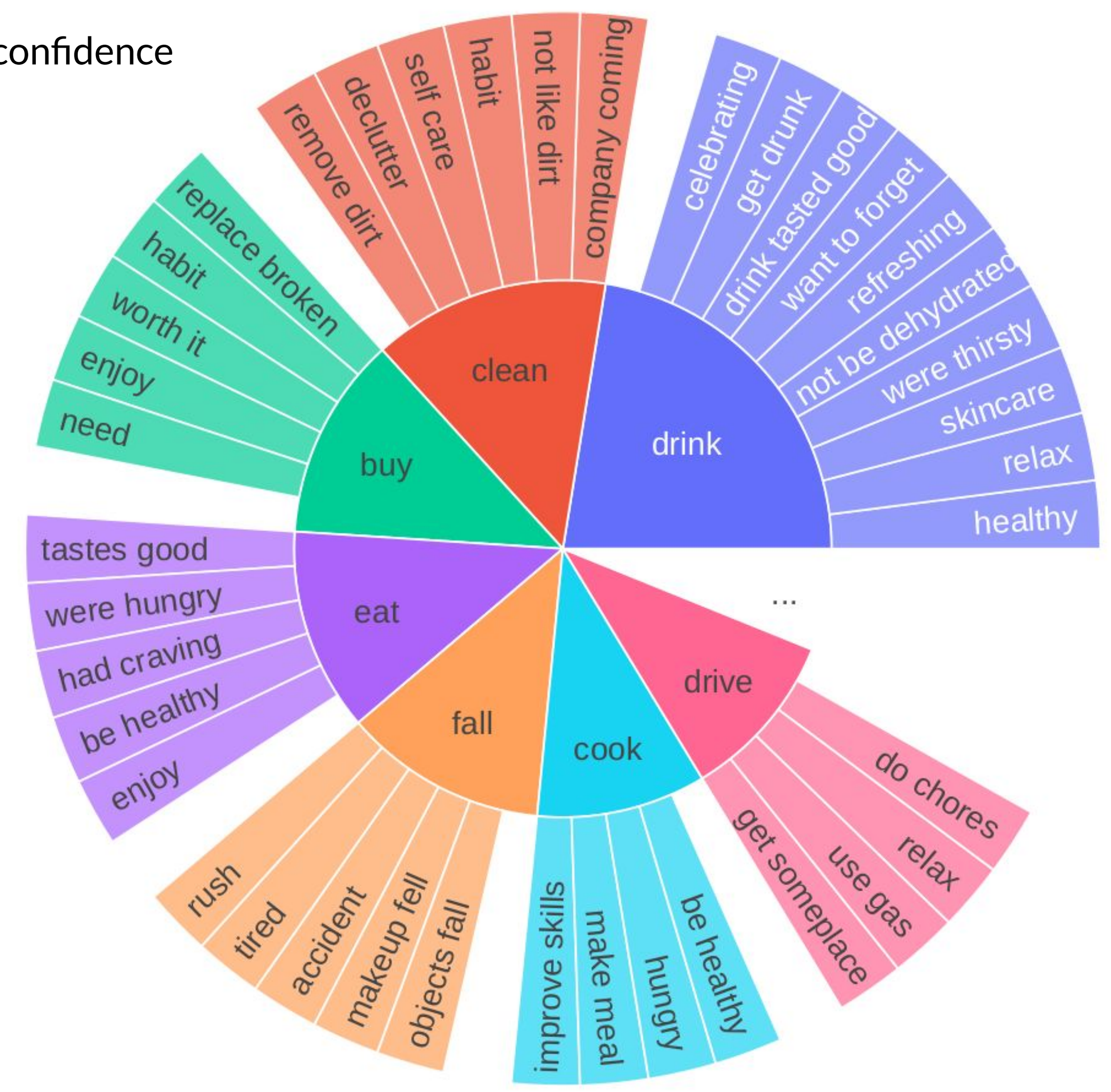
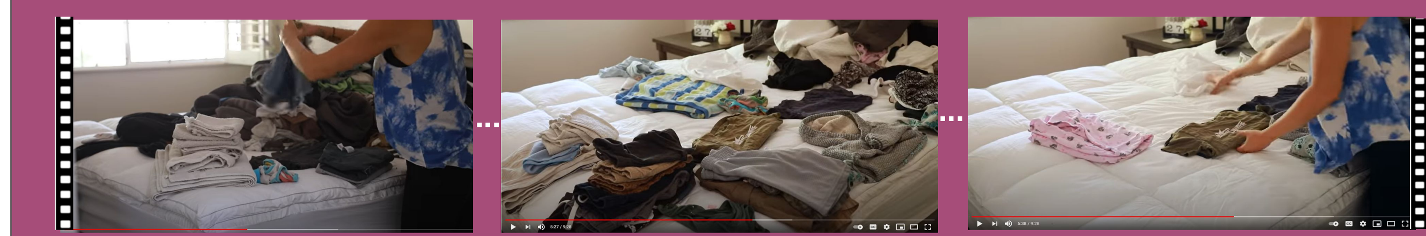


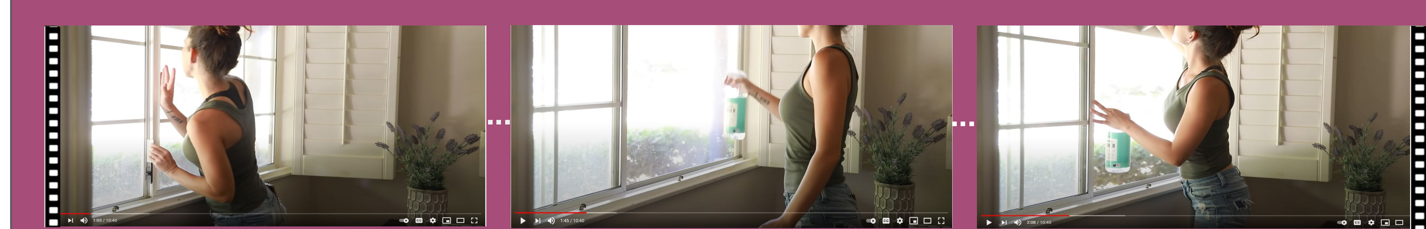
Fig 2: Distribution of the first seven actions, in alphabetical order, and their reasons

Why is the person cleaning?



- [ ] company was coming
- [ ] do not like dirtiness
- [x] declutter
- [ ] remove dirt

*"..**just put everything in everybody's drawers** and slowly but surely we are down to the bedding I always do it on a strip bed because I usually clean the bedding .."*



- [x] company was coming
- [ ] do not like dirtiness
- [ ] declutter
- [x] remove dirt

*"".. blow up mattress so **it feels really nice to be able to have this space for them** in the summer months **I focus on [cleaning] the windows** more because I noticed .."*

Fig 1: Overview of our task: automatic identification of action reasons in online videos. The reasons for *cleaning* change based on the visual and textual (video transcript) context. The figure shows two examples from our WhyAct dataset.

Task: Human Action Reason Identification in online vlogs

Dataset: WhyAct of **1,077 (action, context, reason)** tuples

Models: single and multi-modality

Data Analysis: what kind of actions are depicted

Future work: incorporate reasons in action recognition models

Data and Code:
https://github.com/MichiganNLP/vlog_action_reason

Experiments

The methods we run are **unsupervised with fine-tuning on development set.**

	Test	Development
Actions	24	24
Reasons	166	166
Video-clips	853	224

Table 3: Statistics for the experimental data split.

Features

Text: Transcripts and **reasons** are represented using **Sentence-BERT**

Video: I3D and **Bag of objects** and collection of **Automatic Captions**

Baselines

Textual Similarity: cosine (transcript, reason)

Natural Language Inference (NLI):

Premise — *Entails?* → *Hypothesis*

- Transcript
- Bag of Objects (*Detectron 2*)
- Dense Video Captions

reason

Multimodal

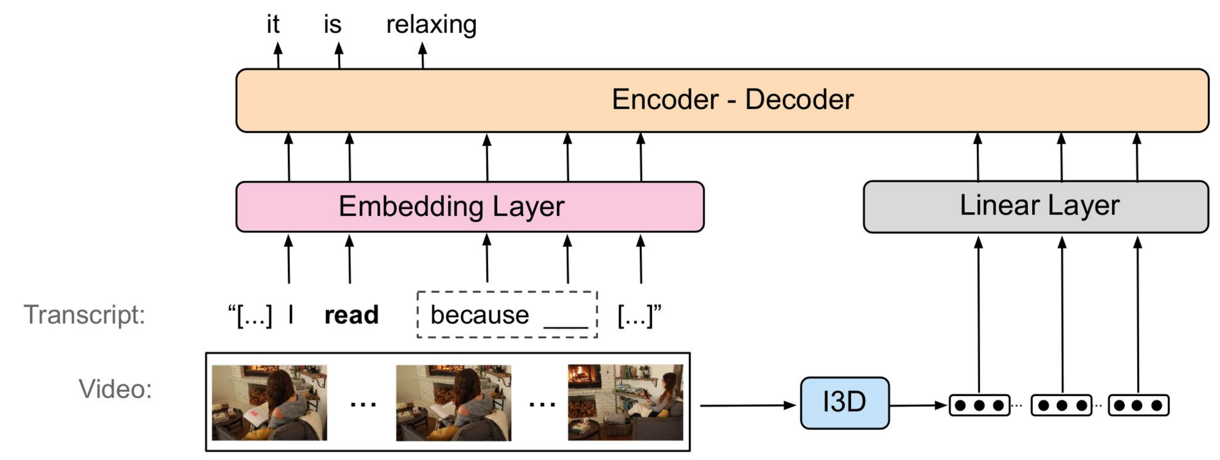


Fig 3: Overview architecture of our **Multimodal Fill-in-the-blanks model**. The span of text *"because ___"* is introduced in the video transcript, after the appearance of the action. This forces the **T5** model to generate the words missing in the blanks.

Results

Method	Input	Accuracy	Precision	Recall	F1
BASELINES					
Cosine similarity	Transcript	57.70	31.39	55.94	37.64
	Causal relations from transcript	50.85	30.40	68.91	39.73
SINGLE MODALITY MODELS					
Natural Language Inference	Transcript	**68.41**	**41.90**	48.01	40.78
	Video object labels	54.49	31.70	59.93	36.79
	Video dense captions	49.18	29.54	68.47	37.40
	Video object labels & dense captions	36.93	27.34	87.97	39.11
Fill-in-the-blanks	Transcript	44.04	30.70	87.10	**43.59**
MULTIMODAL NEURAL MODELS					
Fill-in-the-blanks	Video & Transcript	32.6	27.56	**94.76**	41.11

Table 4: Results on test data

Analysis

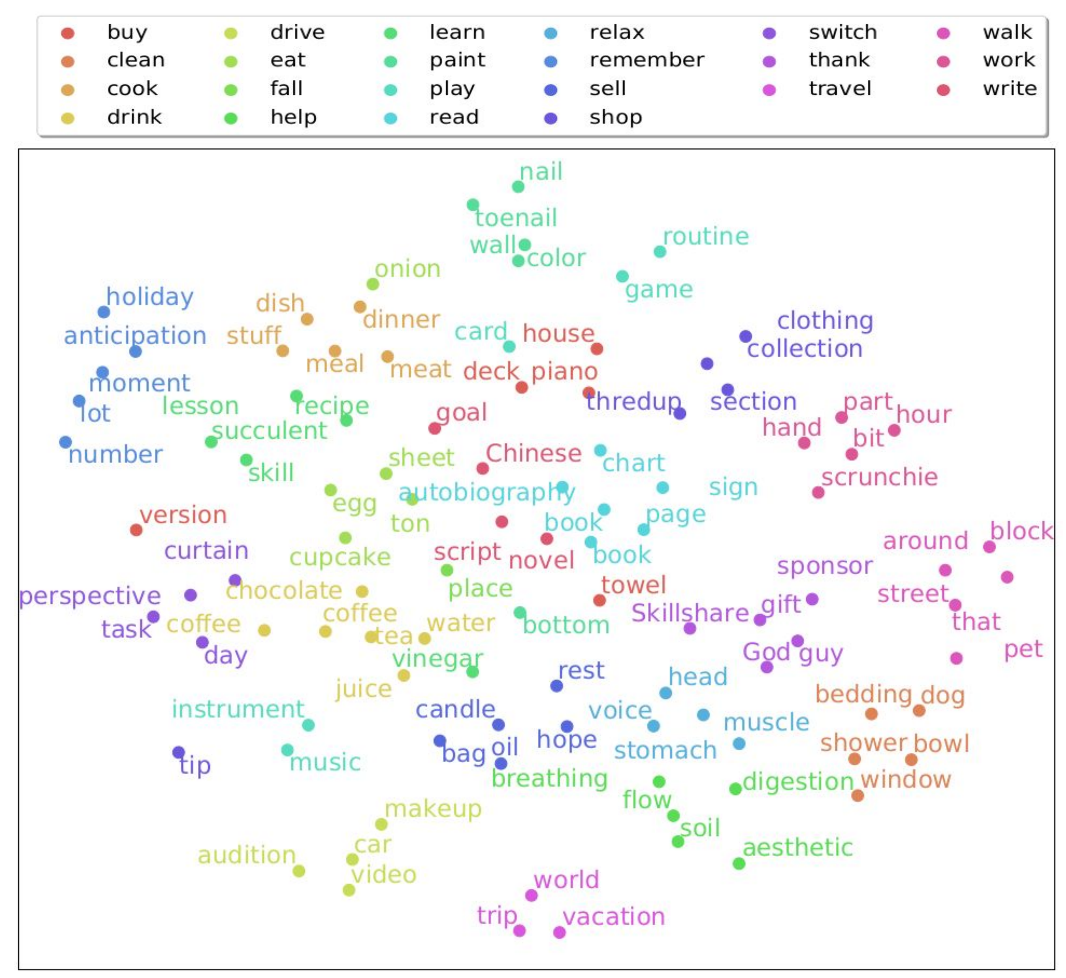


Fig 4: **t-SNE** representation of the five most frequent direct objects for each action/verb in our dataset. Each **color** represents a different **action**.