11-777 Spring 2021 Class Project

Yun Cheng* **Yuxuan Liu*** **Tiffany Ma*** **Erin Zhang***
{yuncheng, yuxuanli, tma1, xiaoyuz1}@andrew.cmu.edu

Abstract

Template for 11-777 Reports using the ACL 2021 Style File

1 Introduction and Problem Definition

*Everyone Contributed Equally – Alphabetical order

2 Related Work and Background (1-1.5 pages)

Literature 1

Literature 2

Literature 3

Literature 4

3 Task Setup and Data (1 page)

The main task is to segment egocentric (first-person) cooking videos from EPIC-KITCHENS dataset into action-object pairs. Given a video clip in the form of a sequence of frames, we want to identify the type of actions as well as their start and end time in the given video.

3.1 Dataset

We use the largest egocentric (first-person) dataset EPIC-KITCHENS-100, which features 100 hours, 700 variable-length videos with 90K actions of 37 participants (Damen et al., 2020). Compared to YouTube-based datasets such as HowTo100M (Miech et al., 2019), EPIC-KITCHENS contains activities that are non-scripted and thus capture more natural settings such as parallel tasking . The egocentric view provides a unique perspective on people-object interactions, attention, and intention. Meanwhile, it also imposes extra challenges compared to third-person datasets like YouCook2 (Zhou et al., 2018). One of the challenges is that certain actions, such as eating and drinking, cannot be directly observed due to the limited field of view. Other challenges include unseen participants, unseen cooking actions, frame noises from different sources (i.e. background and lighting), long videos with many action instances, fragmentation of segments resulted from interleaving actions in multi-tasking, and weaker temporal correlations in objects interfering the correlations in actions.

3.2 Task formulation

There are two input modalities: video frames of egocentric cooking scenes and narrations describing the action in the scenes. The narrations are transcribed from the audio in the form of imperative phrases: verb-noun with optional propositional phrase. The goal is to predict a verb class as well as a noun class for each frame to identify the action in the segments. Afterwards, we combine the two classes into a tuple as the final output class label.

Formally, the visual input consists of a sequence of M RGB frames in temporal order, denoted as $F = (\mathbf{f}_i)_{i=1}^{M}$. The RGB frames are sampled from untrimmed videos at a rate of 50 frames per second. The textual input is a sequence of N audio-transcribed narrations in temporal order, denoted as $C = (\mathbf{c}_i)_{i=1}^{N}$. Our goal is to infer the action-object class label for each frame. The ground truth is given by $Y = (\mathbf{y}_i)_{i=1}^{M}$. Each

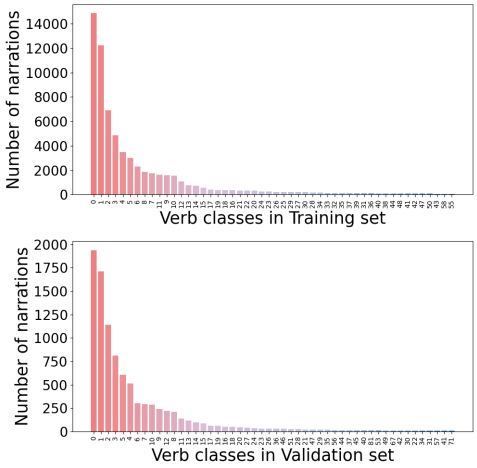

Figure 1: Frequency distribution of 50 most frequent verb class in training and validation set

$\mathbf{y}_i \in \{0,1\}^K \times \{0,1\}^L$ is a tuple of one-hot vectors encoding the true verb and noun class, where K is the number of verb classes and L is the number of noun classes.

3.3 Dataset Statistics

3.3.1 Text Analysis

Narrations in EPIC-KITCHENS-100 are mainly imperative phrases in the form of verb-noun with optional propositional phrase (e.g. *put down plate*, *put container on top of counter*). Each annotation includes start/stop timestamps and frames, action verbs and object nouns, which are extracted from the corresponding narration. Verbs and nouns are further classified into classes based on their semantic meaning. For example, *grab* and *get* belong to the same verb class. There are a total of 97 verb classes and 300 noun classes in the training and validation set.

We define the frequency of a verb/noun class as the number of narrations that contain a verb/noun from that class. Both verb and noun classes have a heavy tailed distribution with tail classes ($\leq$ 1/15 of the maximum frequency) accounting for 13.02% and 11.67% total verbs and 5.38% and 1.85% total nouns in the training and validation set respectively (Figure 1). Such a distribution indicates the intrinsic complexity and entropy of the text data. The training and validation set have similar composition: in the validation set, there are no unseen verb class and only four unseen noun classes, accounting for 0.03% of all narrations. Narration timestamps are relatively complete: only 0.17% in training and 0.72% in validation are missing. On the other

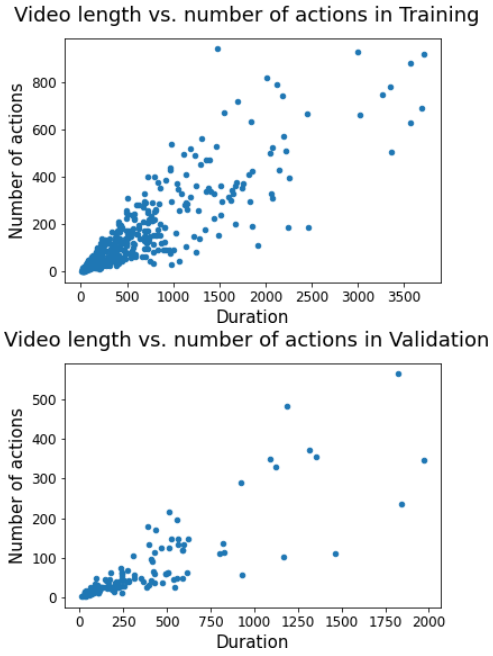

Figure 2: Number of action in each video against video length (in seconds)

hand, since there were no constraints on the recording duration, we observe a great variability across videos. Average sentence length of training and validation set is 15.1 and 14.8 with standard deviation of 6.3 and 6.0 words, respectively. Average number of actions per video is 135.8 (training) and 70.1 (validation) with standard deviation of 167.7 and 93.2. More distribution statistics can be found in Appendix A Table 1.

A natural assumption of our task is that there is none or minimal overlapping between action segments, i.e. only one action in almost all time frames. We check that there are at most 4 narrations in parallel in training and 3 in validation; only 3832 (5.70%) and 617 (0.92%) pairs of consecutive actions overlap for more than 1 second. We also inspect the feature embeddings of the verb and noun classes. Using GloVe word vectors pre-trained on Twitter (200d vectors) (Pennington et al., 2014), we do not notice significant interclass or intraclass clustering effect (Appendix A Figure 4).

3.3.2 Video Frame Analysis

We extract 1920 $\times$ 1080 RGB frames from the videos at a sampling rate of 50 FPS. Each frame is identified by participant id, video id, and a start/end frame number. More than half of the total 700 videos in the dataset have less than 25,000 frames.

Videos in EPIC-KITCHENS-100 have varied length with the longest video of 3708 seconds and shortest video of 10 seconds. 85.7 % of the videos are shorter than 1000 seconds and 66.0 % are less than 500 seconds (Appendix A Table 1, Figure 5). We also see that the number of narrations grows roughly linearly with video length (Figure 2).

We compile all training and validation samples of a given verb class and compute the average number of frames for this class (Appendix Figure 8). The top-10 verb classes with the most number of frames include actions like *grate, wait, prepare, knead, stir*, and *cut*; while those with the least number of frames contain actions like *bend, turn-off, turn-on, take, close*. It seems that actions involved during cooking take longer than those related to intermediate preparatory steps, and the average length of the action aligns with how people would respond if asked about which action would take longer. For most verb classes, the average number of frames in each class are roughly the same in both training and validation set, except a few where the validation sets have more frames. We also count the total number of frames for each verb class, summed over all training and validation samples in the class. We notice that such frequency corresponds to the trend of verb-class frequency in the annotations (Appendix A Figure 9). This indicates that within the dataset, the frequency of the verb class correlates to the amount of visual information in the dataset.

The dataset also provides bounding-box annotations for each frame, where it only distinguishes between two categories: hands and objects. Only active objects are annotated, so the number of object bounding-boxes in a frame approximates the number of objects that the person interacts with. We compute the average number of hand bounding-boxes appearing in a frame of each verb class. Class with less than 1.5 hand bounding-boxes include actions like *take, put-on, open, pull-down, walk*, and these correspond roughly to human impression on how many hands are needed for performing the action. We also compute the average number of objects bounding boxes in a frame of a given verb class. Verb classes with less than 1.8 object bounding boxes include actions like *open, close, shake, check, fold*, and *drink* (Appendix A Figure 6, 7). The average numbers of hand and object bounding boxes for the training and validation sets are mostly equal, despite the validation set misses a few verb classes. Full details can be found in Appendix A.

3.4 Metrics

4 Models (2 pages)

4.1 Baselines

Both existing baselines explained with citations and novel ones missing from the current literature

4.2 Proposed Approach

5 Results (1 page)

The columns above are just examples that should be expanded to include all metrics and baselines.

Methods	Dev		Test	
	Accuracy ↑	L_2 Error ↓	Accuracy ↑	L_2 Error ↓
Previous Approach 1 ()				
Previous Approach 2 ()				
Previous Approach 3 ()				
Proposed Method				

6 Analysis (2 pages)

This section should include at least two to three plots

6.1 Ablations and Their Implications

6.2 Qualitative Analysis and Examples

This section should likely contain a table of examples demonstrating how the current approach succeeds/fails.

References

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. Rescaling egocentric vision. *CoRR*, abs/2006.13256.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.

Appendix A Data Analysis

In this section, we present the full details of our data analysis.

	Training				Validation			
	Max.	Min.	Avg.	Std.	Max.	Min.	Avg.	Std.
Verb class frequency	14848	73	1314	2829	1937	71	191	398
Noun class frequency	3617	178	724	655	430	25	108	92
Sentence length	77	3	15.1	6.3	71	3	14.8	6.0
Actions per video	940	1	136	168	564	3	70	93
Frames per verb class	2129212	20165	225170	408408	407425	2702	42016	76950
Video length	3708	10	543	645	1969	11	344	377

Table 1: Statistics of EPIC-KITCHENS-100 training and validation set

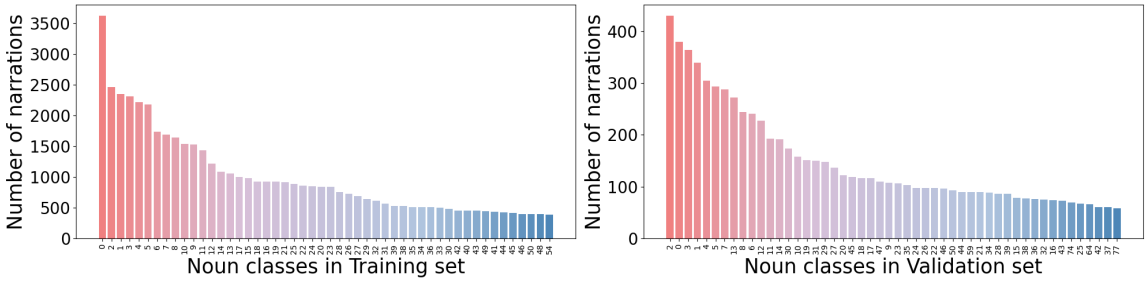

Figure 3: Frequency distribution of 50 most frequent noun classes in training and validation set

Figure 4: Example of visualizing feature embeddings of verb and noun classes in 2D and 3D space

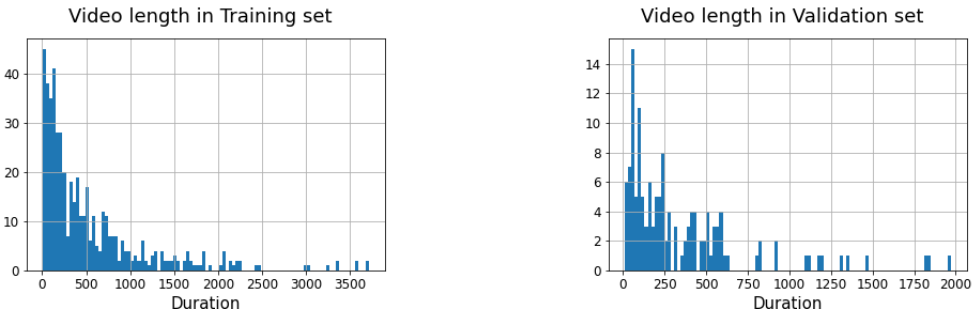

Figure 5: Distribution of video length (in seconds)

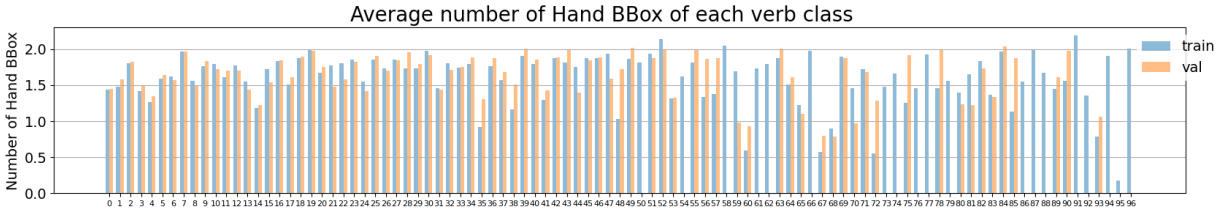

Figure 6: Average number of hand bounding-boxes in each frame of given verb class

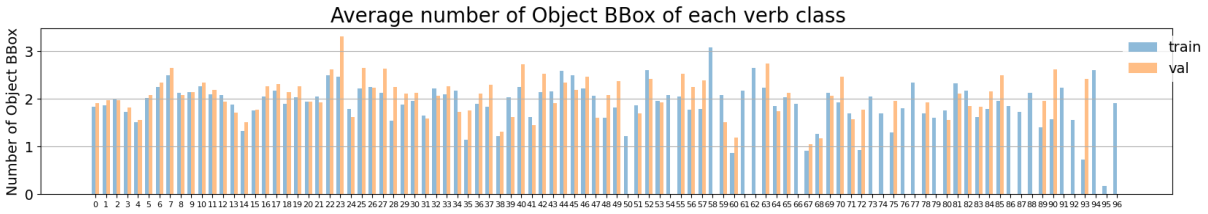

Figure 7: Average number of object bounding-boxes in each frame of given verb class

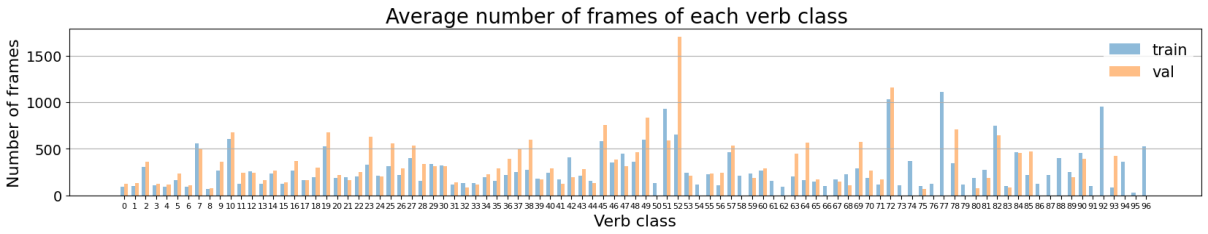

Figure 8: Average number of frames in a narration of a given verb class in training and validation set

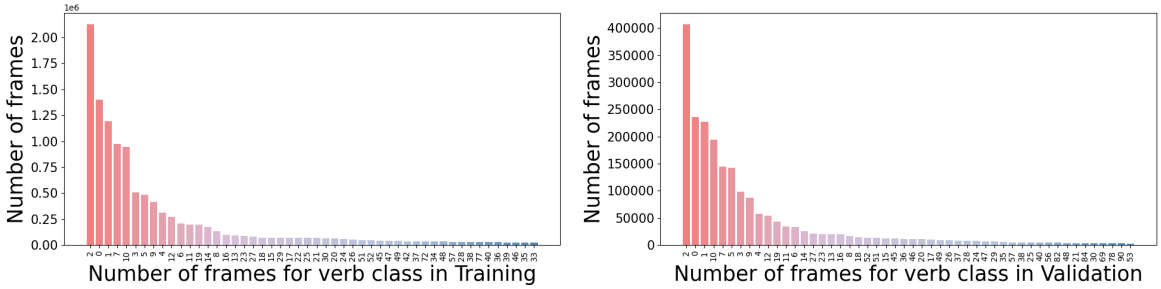

Figure 9: Distribution of number of frames in each video