THE UNIVERSITY OF UTAH®

University of CINCINNATI

Commonsense Knowledge for Visual Activity Recognition

Tianyu Jiang

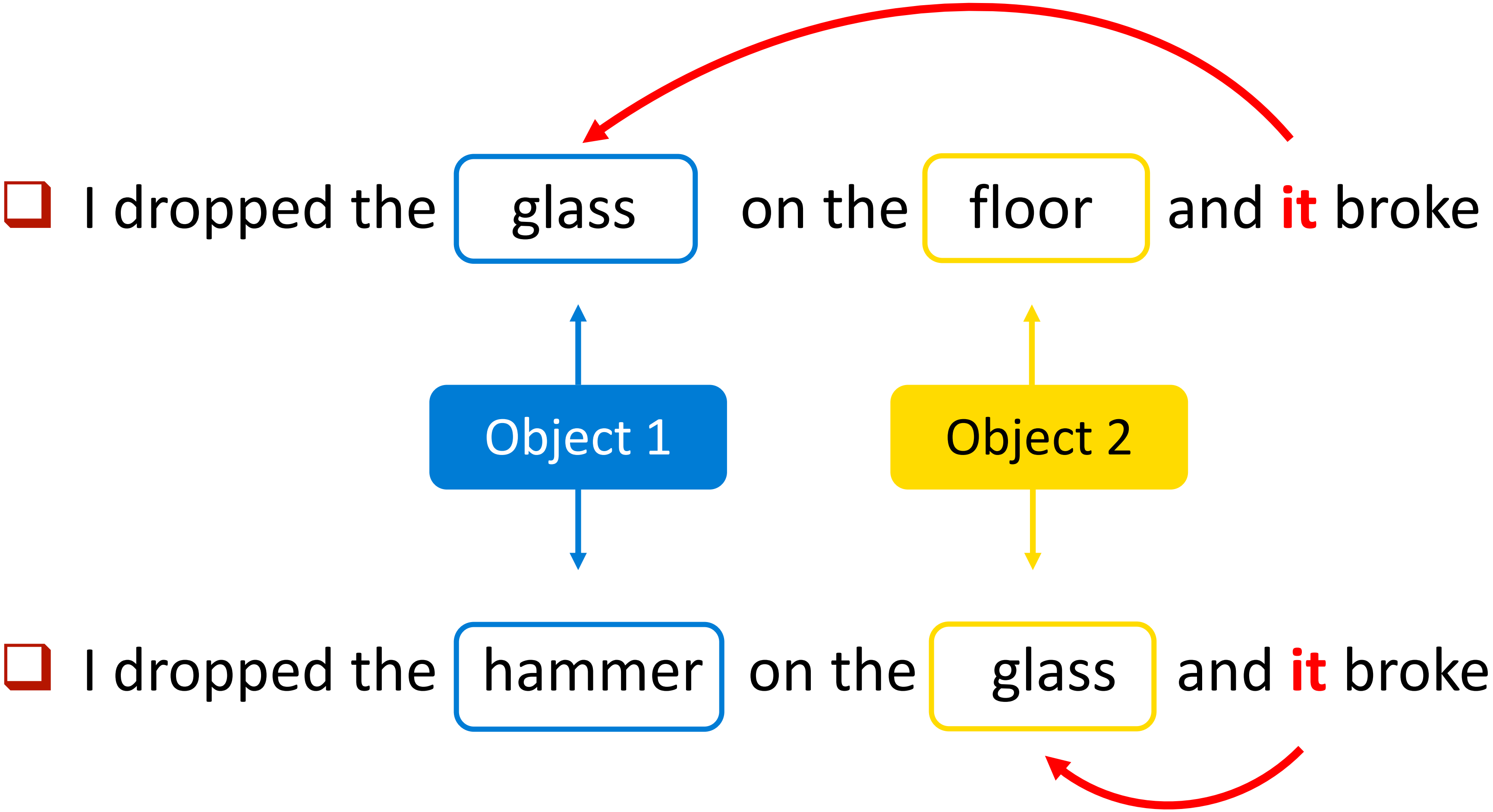
University of Cincinnati

Oct 17, 2023

Mary put the cake in the **oven**.

Why?

- “decorate the cake”
- “cut the cake”
- “eat the cake”
- “bake the cake”



- ❑ I dropped the glass on the floor and **it** broke
- ❑ I dropped the hammer on the glass and **it** broke

Sentence pair from http://demo.clab.cs.cmu.edu/11711fa20/slides/11711-01-introduction.pdf

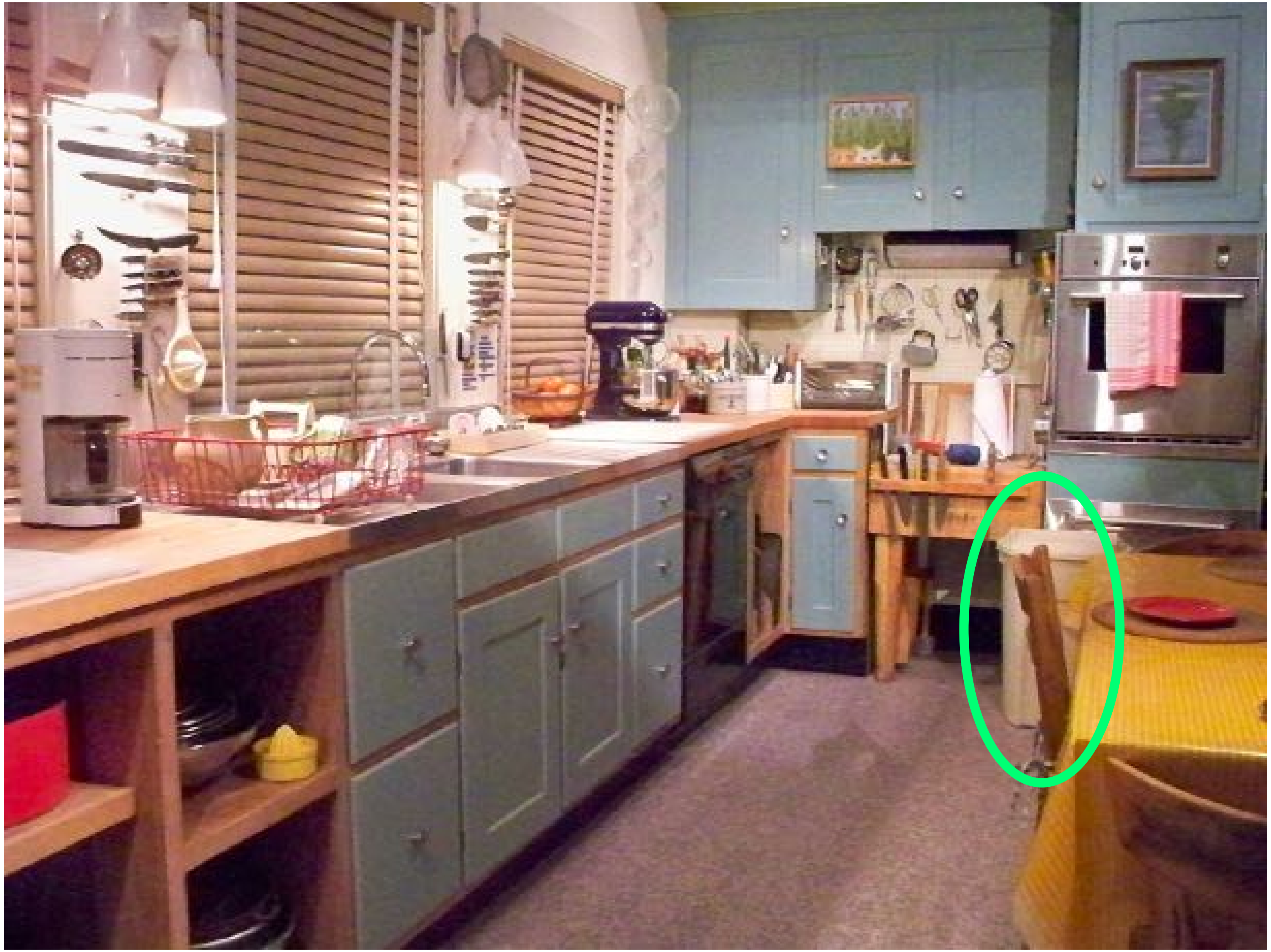


Photo from Davis & Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM. 2015 Aug 24;58(9):92-103.

Commonsense Knowledge

Explicitly Expressed

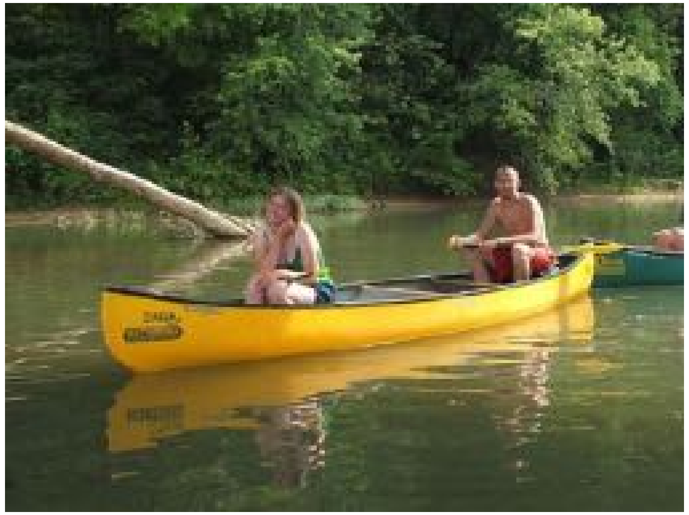
Implicitly Expressed

Challenge: How to enable an automatic system to understand the implicit language as humans?

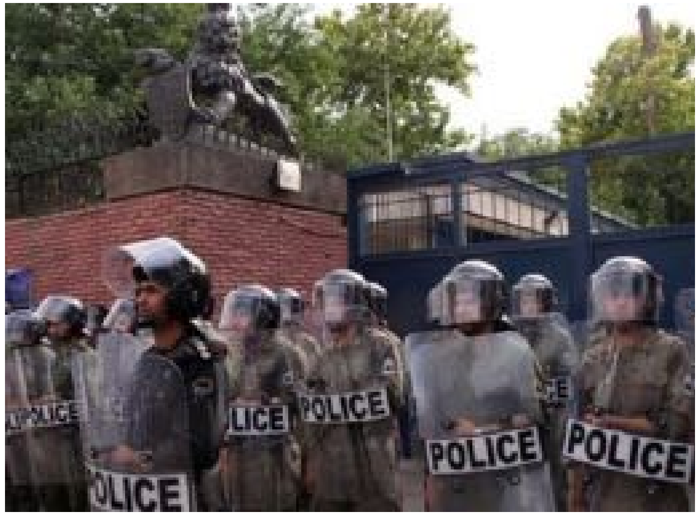
Approach: Build systems with commonsense knowledge.

Situation Recognition

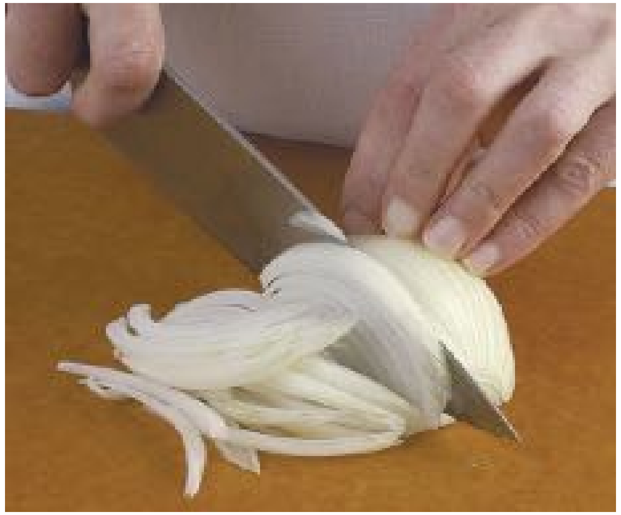
- is a task of recognizing the activity depicted in an image.[1]
- It identifies the people and objects involved in the activity and the roles these participants play.



Motion



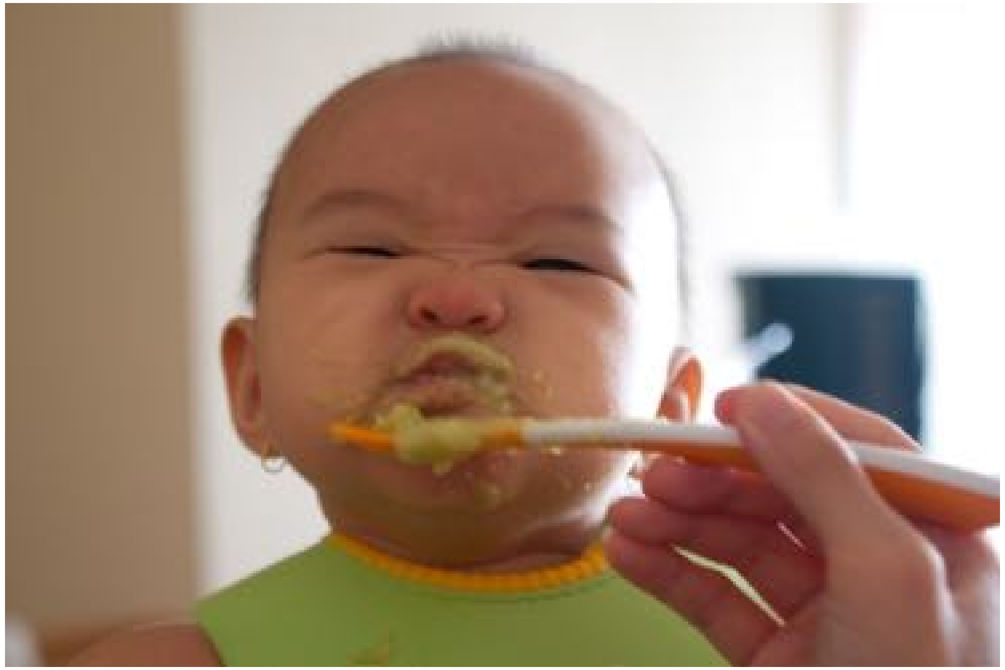
Protecting



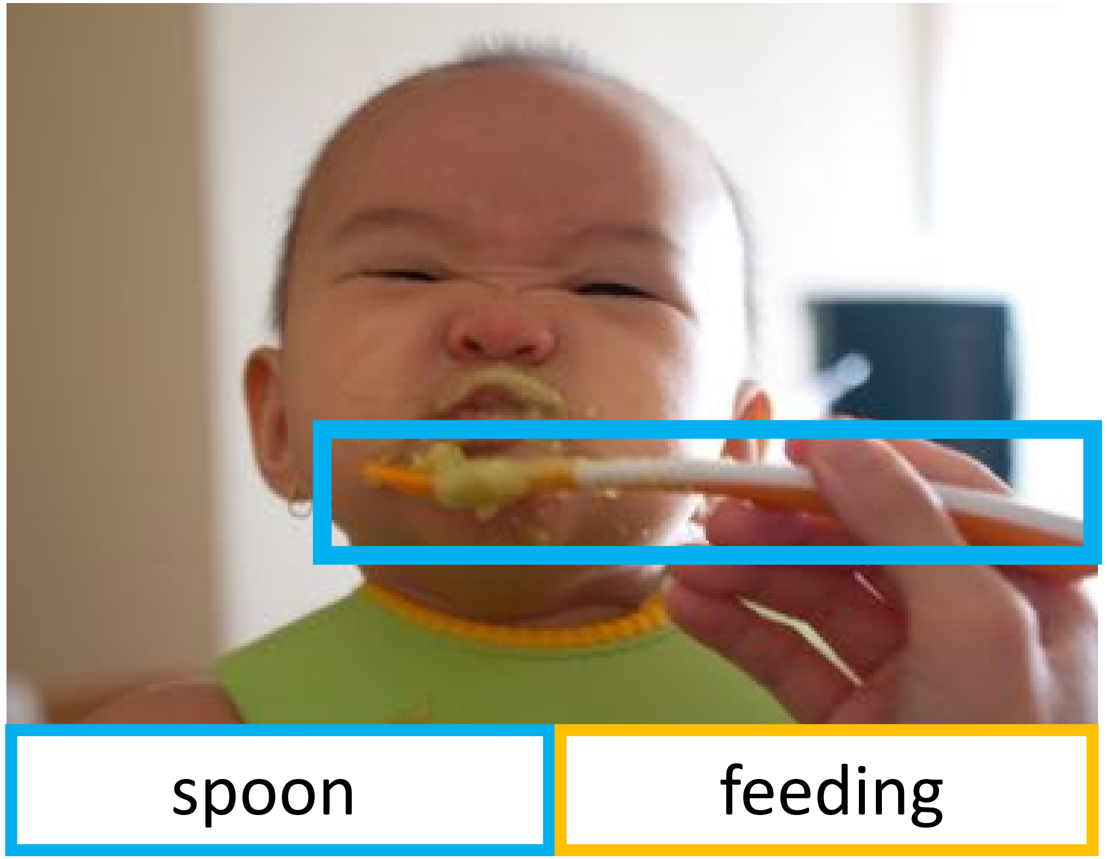
Cutting

[1] Yatskar et al. Situation recognition: Visual semantic role labeling for image understanding. CVPR 2016. Images are from its dataset.

What is happening in the image?



Knowing what objects exist can help identify the action!



spoon

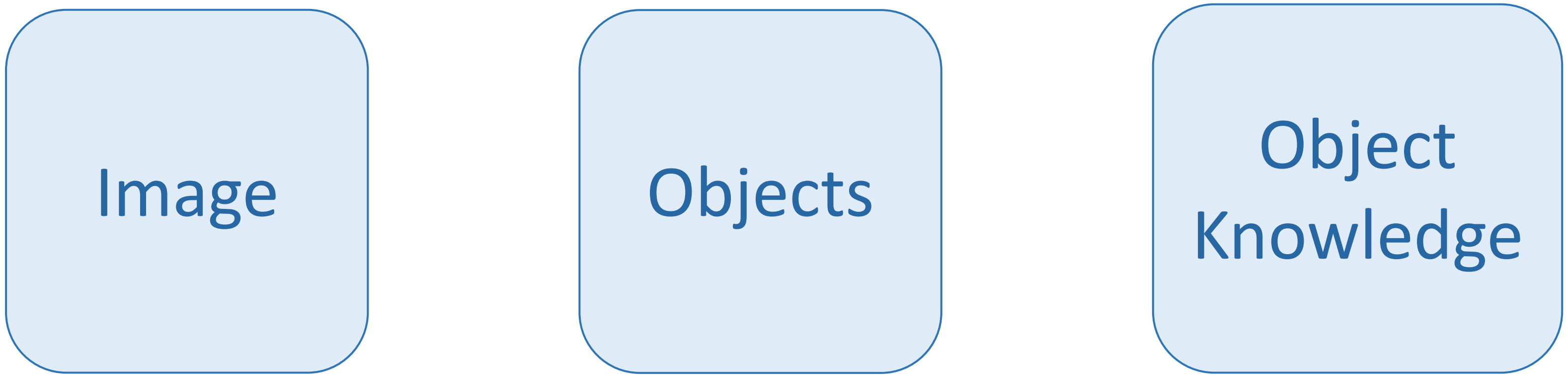
feeding

canoe → Motion

shield → Protecting

knife → Cutting

Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition [Jiang & Riloff, Findings of ACL 2023]



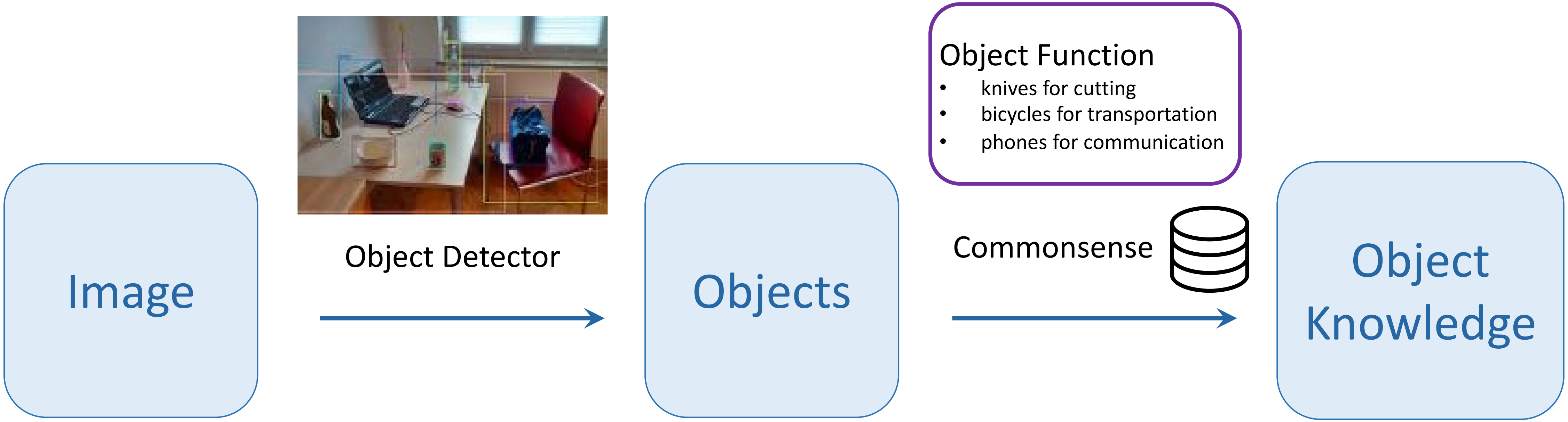
Image

Objects

Object Knowledge

Jiang, Tianyu, and Ellen Riloff. "Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition." *Findings of the Association for Computational Linguistics: ACL 2023*.

Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition [Jiang & Riloff, Findings of ACL 2023]



Image

Object Detector

Objects

Object Function

- knives for cutting
- bicycles for transportation
- phones for communication

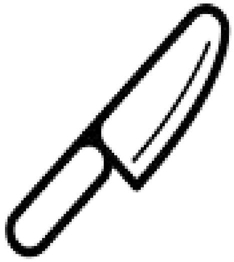
Commonsense

Object Knowledge

Photo from: https://en.wikipedia.org/wiki/Object_detection#/media/File:Detected-with-YOLO--Schreibtisch-mit-Objekten.jpg

Learning Prototypical Functions for Physical Artifacts [Jiang & Riloff, ACL 2021]

spears for hunting



knives for cutting

pots for cooking

Jiang & Riloff. Learning prototypical functions for physical artifacts. ACL 2021.

Learning Prototypical Functions for Physical Artifacts [Jiang & Riloff, ACL 2021]

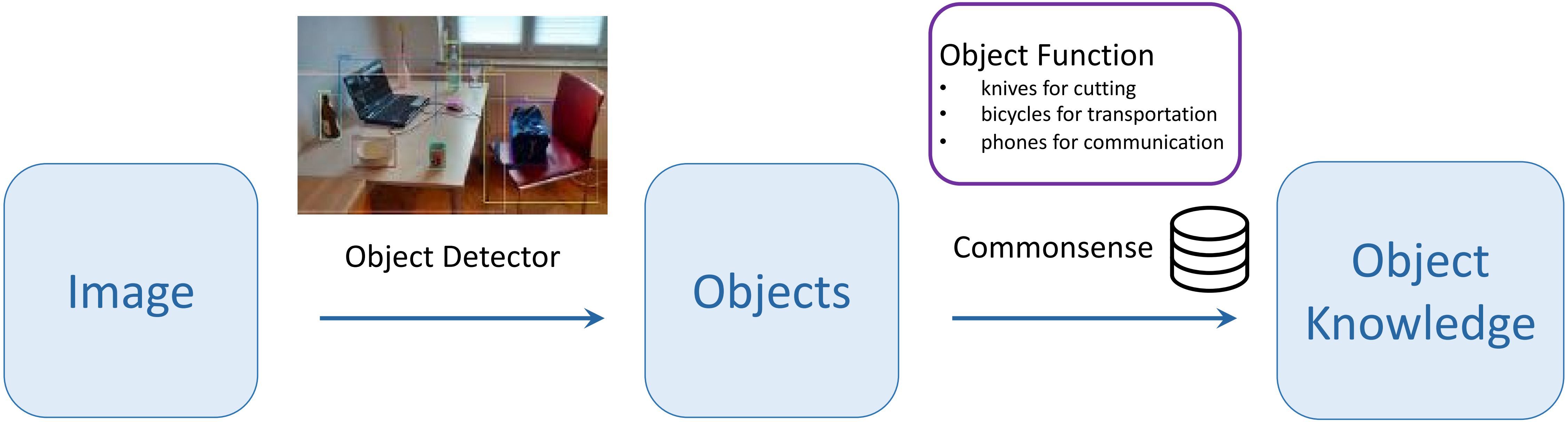
We manually selected frames from FrameNet to represent the functions of physical objects.

For any given physical object, our model will select a frame that best represents the object’s function.

Frames	Objects
Wearing	hat, shirt
Containing	basket, luggage
Self_motion	bicycle, yacht
Protecting	armor, helmet
Cutting	knife, scissors

Jiang & Riloff. Learning prototypical functions for physical artifacts. ACL 2021.

Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition [Jiang & Riloff, Findings of ACL 2023]



Image

Object Detector

Objects

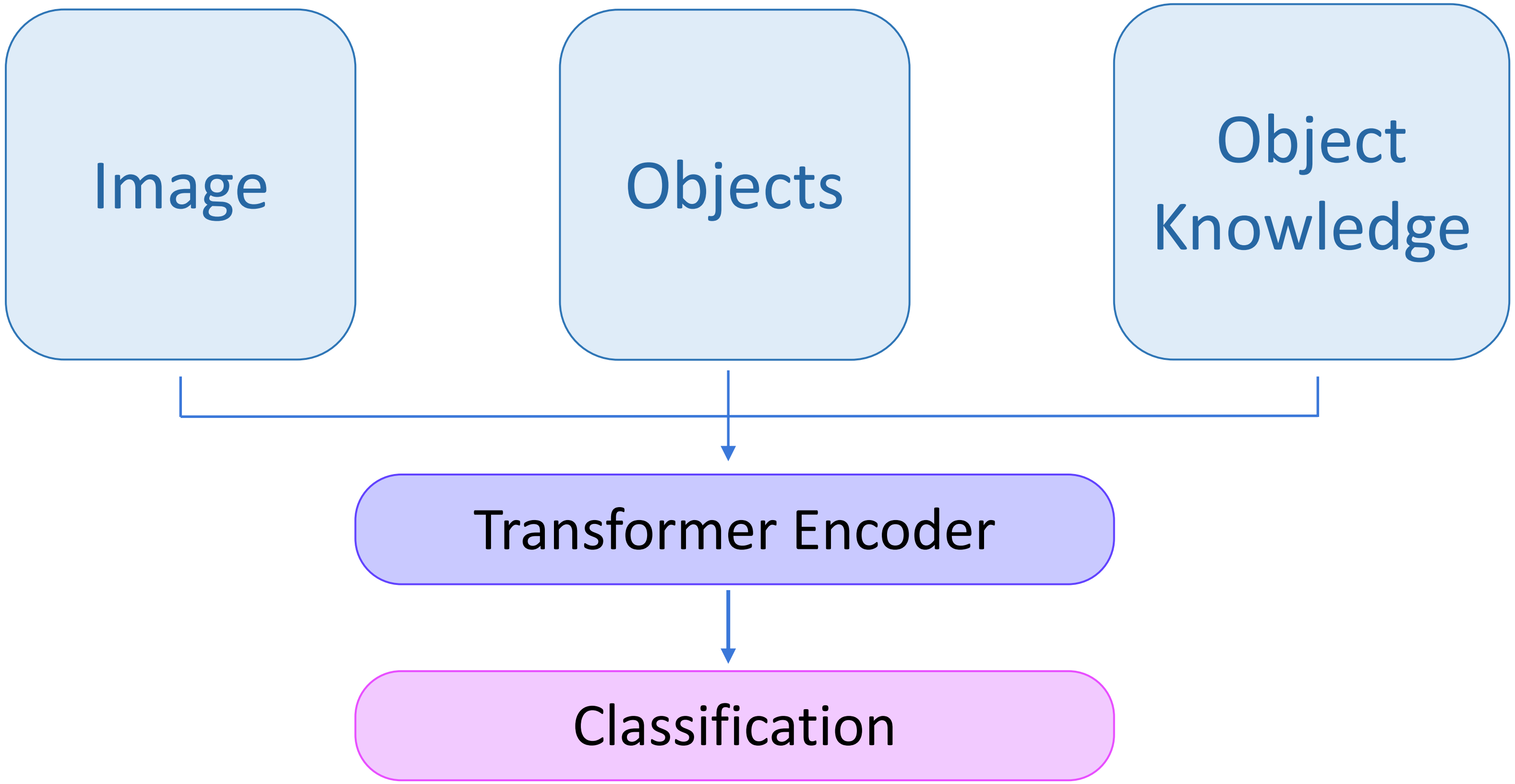
Object Function
- knives for cutting
- bicycles for transportation
- phones for communication

Commonsense

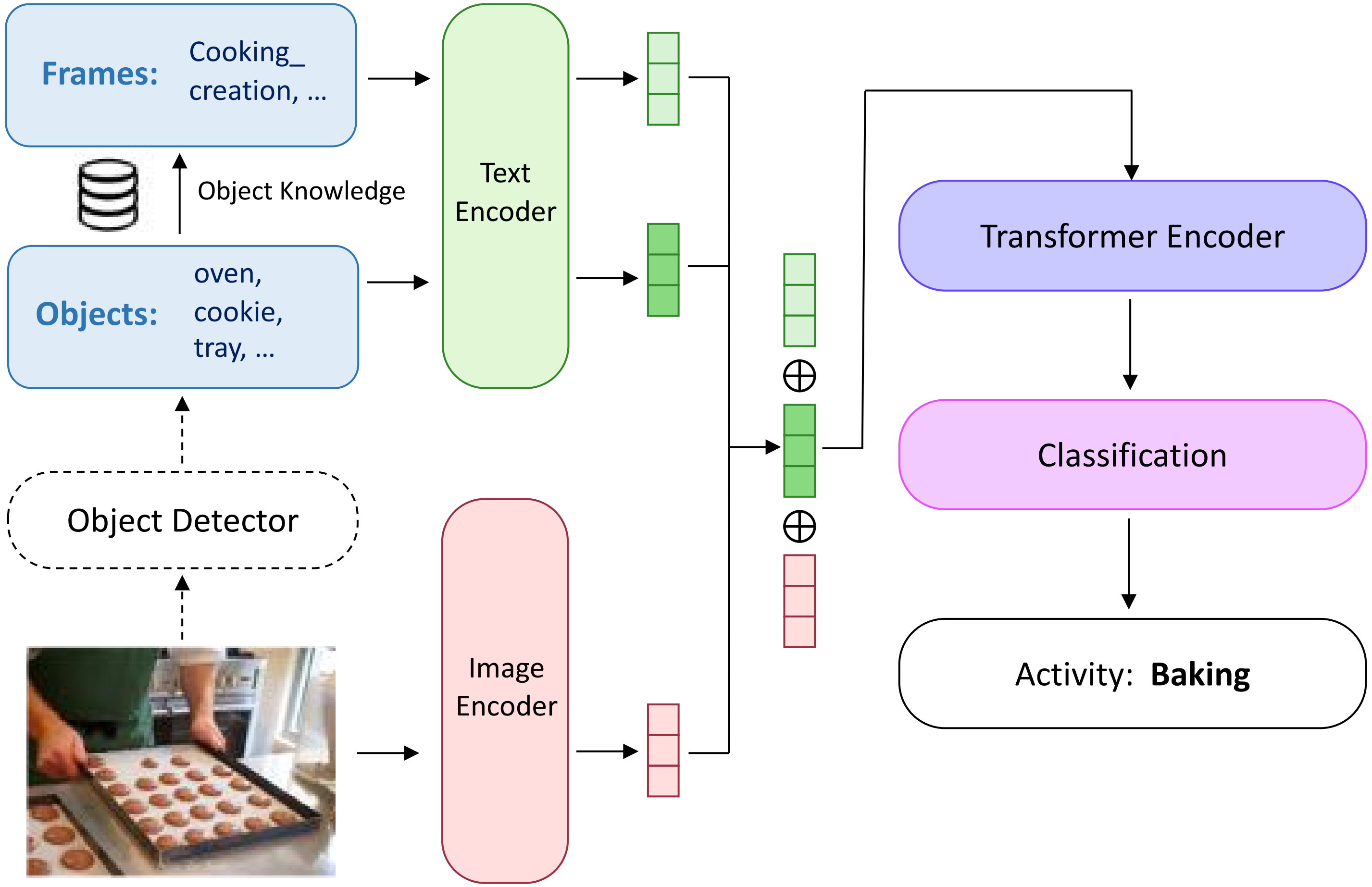
Object Knowledge

Photo from: https://en.wikipedia.org/wiki/Object_detection#/media/File:Detected-with-YOLO--Schreibtisch-mit-Objekten.jpg

Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition [Jiang & Riloff, Findings of ACL 2023]

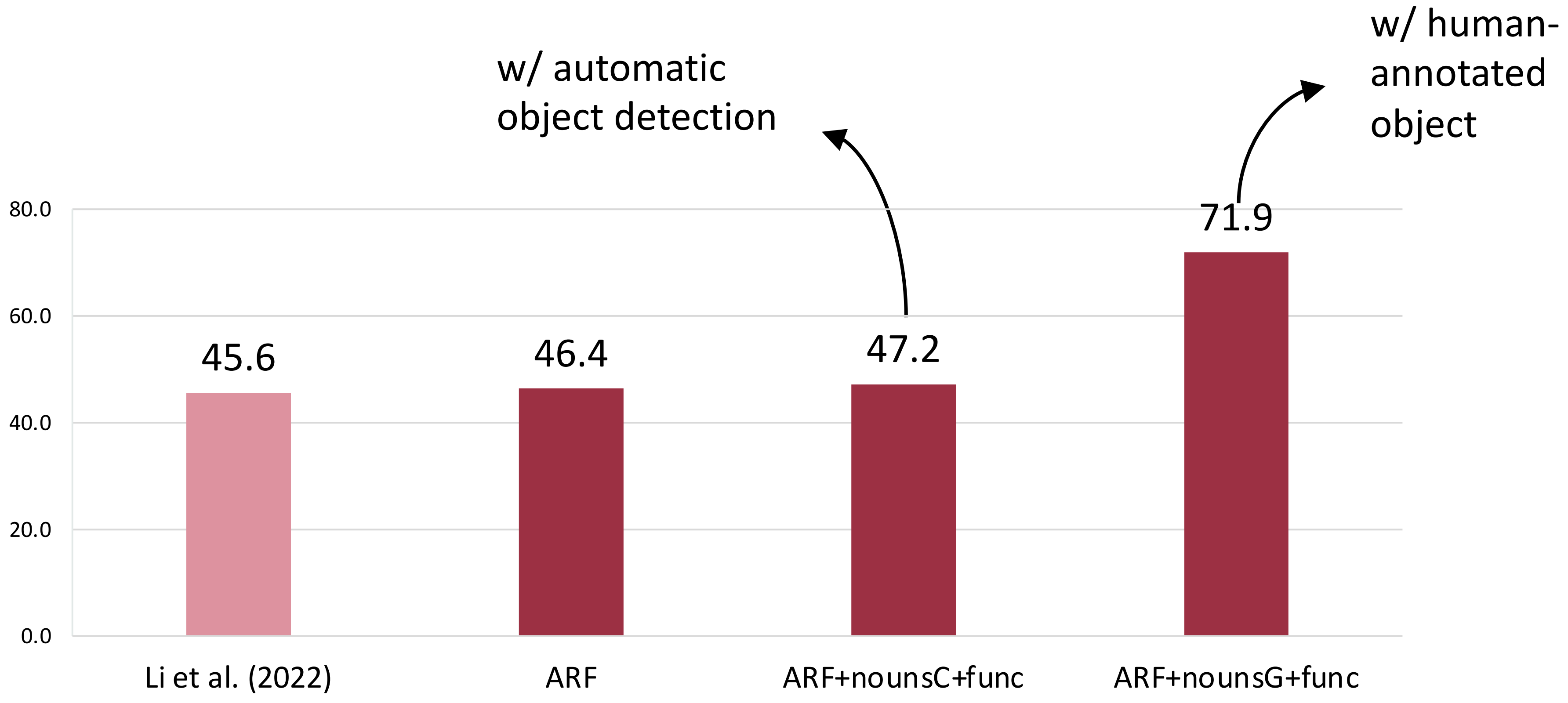


Model ARF (**A**ctivity **R**ecognition with **F**unctions)



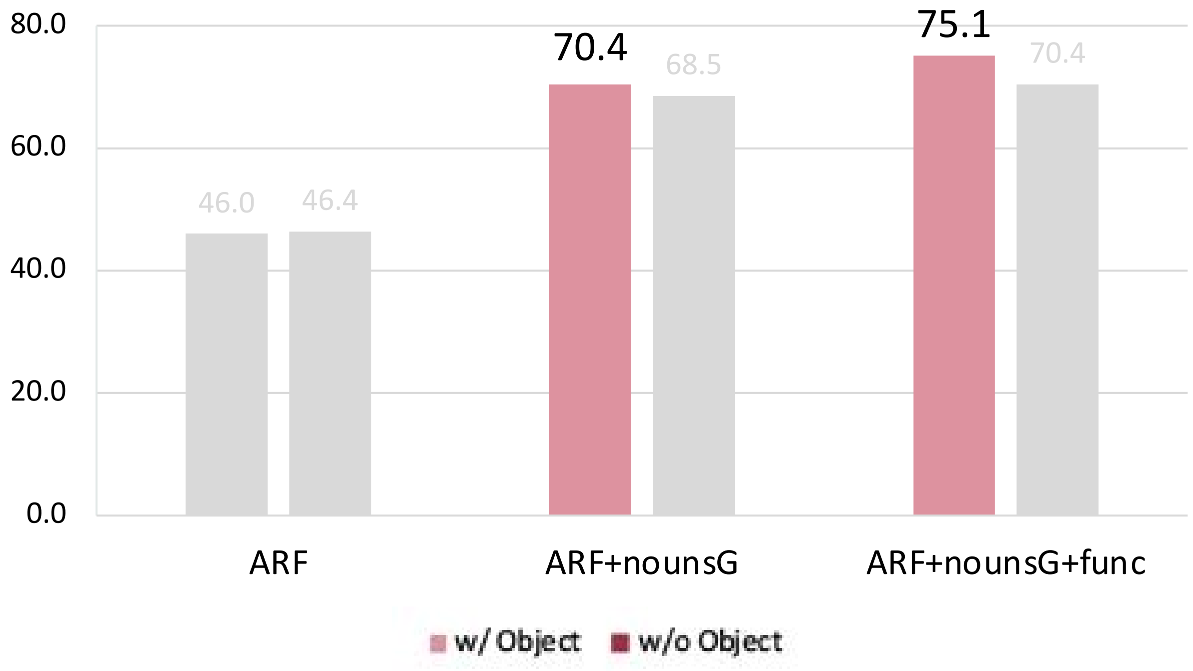
[Jiang & Riloff, Findings of ACL 2023]

Experimental Results



[Jiang & Riloff, Findings of ACL 2023]

Experimental Results



For images with objects, adding the functional knowledge get better performance

[Jiang & Riloff, Findings of ACL 2023]

Summary & Future Directions

- Commonsense knowledge of object functions can benefit visual activity recognition.
- Our system can benefit from a better object detector.
- Future: Incorporating other commonsense knowledge into vision-language models. For example, reasoning ability of physical events.

THANKS!

Questions?