Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition

Tianyu Jiang and **Ellen Riloff**
Kahlert School of Computing
University of Utah
Salt Lake City, UT 84112
{tianyu, riloff}@cs.utah.edu

Abstract

Situation recognition is the task of recognizing the activity depicted in an image, including the people and objects involved. Previous models for this task typically train a classifier to identify the activity using a backbone image feature extractor. We propose that commonsense knowledge about the objects depicted in an image can also be a valuable source of information for activity identification. Previous NLP research has argued that knowledge about the prototypical functions of physical objects is important for language understanding, and NLP techniques have been developed to acquire this knowledge. Our work investigates whether this prototypical function knowledge can also be beneficial for visual situation recognition. We build a framework that incorporates this type of commonsense knowledge in a transformer-based model that is trained to predict the action verb for situation recognition. Our experimental results show that adding prototypical function knowledge about physical objects does improve performance for the visual activity recognition task.

1 Introduction

Physical objects play an important role in our daily lives. People use different tools to achieve different goals in all kinds of situations. For example, we use a toothbrush to clean our teeth, a microwave oven to heat food, and a camera to take photos. The functions of physical objects is a type of commonsense knowledge that has been recognized to play an important role in natural language processing (Burstein, 1979; Jiang and Riloff, 2021).

Physical objects play an important role in computer vision as well. There are well-established computer vision tasks that aim to identify the objects in an image, such as object detection (Lin et al., 2014) and image classification (Deng et al., 2009; Krizhevsky, 2009). Recently, attention has been paid to more comprehensive image understanding, such as identifying the salient event depicted in an image as well as relevant people and objects. **Situation recognition** (Yatskar et al., 2016) is the task of producing a structured summary of an image that describes the main activity and the entities that fill semantic roles for that activity. The task was originally defined using frame structures from FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016) as the activity representation. For example, given the image shown in Figure 1, a system should identify a *baking* event (which is indexed in FrameNet as a type of *Cooking_creation* activity), and recognize the corresponding semantic role/value pairs associated with FrameNet's *Cooking_creation* frame. Models for this task usually follow a two-step pipeline: (1) predict a verb that describes the activity depicted in the image, and (2) identify the entities associated with each semantic role. Previous systems have relied solely on features extracted from the image and have not yet exploited any external commonsense knowledge.

(a) Input image.

BAKING	
ROLE	VALUE
AGENT	MAN
FOOD	COOKIE
FOODCONTAINER	COOKIE SHEET
HEATSOURCE	OVEN
PLACE	KITCHEN

(b) Annotated activity and semantic roles.

Figure 1: *Situation Recognition* involves predicting activities with semantic role/value pairs.

Our work focuses on the activity recognition (verb prediction) part of the situation recognition task. We hypothesize that (a) correctly identifying the activity in an image strongly depends on recognizing the objects that appear in the image, and (b) explicit commonsense knowledge about physical objects can also be beneficial. More specifically, our work is motivated by recent research emphasizing the importance of commonsense knowledge

about the prototypical functions of physical objects for language understanding (Jiang and Riloff, 2021, 2022). An intuitive extension to visual reasoning is that if an object appears in an image, especially when it is used by a person, the activity depicted in the image is likely to be the prototypical function associated with the object. For example, a woman holding a comb is probably brushing her hair, and a man holding a cookie sheet (as shown in Figure 1) is probably baking.

We explore these hypotheses by creating a transformer-based model that incorporates commonsense knowledge about the prototypical functions of physical objects for visual activity recognition. Our experimental results confirm that correctly identifying the objects in an image is very important for activity recognition, and we show that providing explicit knowledge about the prototypical functions of objects can improve performance for this task.

2 Related Work

Commonsense knowledge about physical objects has long been recognized to be important for natural language understanding (Burstein, 1979). Within the NLP community, a variety of recent projects have focused on acquiring and using different types of knowledge about physical objects, including relative physical knowledge (Forbes and Choi, 2017), relative spatial relations (Collell et al., 2018), semantic plausibility (Wang et al., 2018), object affordances (Persiani and Hellström, 2019), and object usage status (Jiang and Riloff, 2022). The work most relevant to our research is Jiang and Riloff (2021), which developed a NLP method to learn the most typical way that people use human-made physical artifacts. They used FrameNet frames as a representation for object functions and they created a dataset of physical objects paired with their prototypical function frames to evaluate their results. Our research incorporates their prototypical function data into a transformer-based model for visual activity recognition.

Visual reasoning tasks, such as visual question answering (Antol et al., 2015) and image captioning (Young et al., 2014), have been widely explored for understanding images and videos. Previous work has proposed to use external knowledge for visual tasks, such as image classification (Marino et al., 2017), object detection (Singh et al., 2018), and visual question answering (Wu et al., 2016).

Situation recognition is a task of recognizing the activity depicted in an image, including the people and objects involved in the activity and the roles these participants play. Yatskar et al. (2016) introduced the **imSitu** dataset, which associates images with a verb that describes the main action, and a set of semantic roles derived from FrameNet (Ruppenhofer et al., 2016). They tackled this problem by first applying the VGG network (Simonyan and Zisserman, 2014) to extract features from the image and then building a CRF model to jointly predict the verb and semantic roles. Several research efforts have further explored this task. Suhail and Sigal (2019) used a graph neural network to capture the relations between semantic roles. Pratt et al. (2020) used a LSTM to jointly classify verbs and semantic roles. Cooray et al. (2020) cast situation recognition as a query-based visual reasoning problem and further handled inter-dependencies between queries to overcome the sparsity issues of semantic roles. Recently, Cho et al. (2022) proposed a collaborative framework using two transformer modules, and Li et al. (2022) used contrastive learning to distinguish the correct activities from negative examples. All of these prior efforts have relied solely on features extracted directly from the image. Our work aims to show that explicitly providing commonsense knowledge about objects can also be beneficial for visual activity recognition.

3 Methods

Given an image, the visual activity recognition task predicts a verb that describes the main activity in the image. Figure 2 shows the framework of our model called **ARF** (**A**ctivity **R**ecognition with **F**unctions), which takes 3 sources of input: 1) the image, 2) nouns corresponding to the objects in the image, and 3) the names of FrameNet frames that describe the prototypical functions of the objects. We use the CLIP (Radford et al., 2021) model, which has been pre-trained on both images and text, to generate an encoding for each of the 3 types of input. Finally, we give the concatenated representation vectors as input to a transformer model that is trained to predict a verb for activity recognition.

3.1 Notation

The task can be denoted as given the ith image I_i ($i = 1..n$), the system should predict the correct activity verb v_i^*. The score for the jth candidate

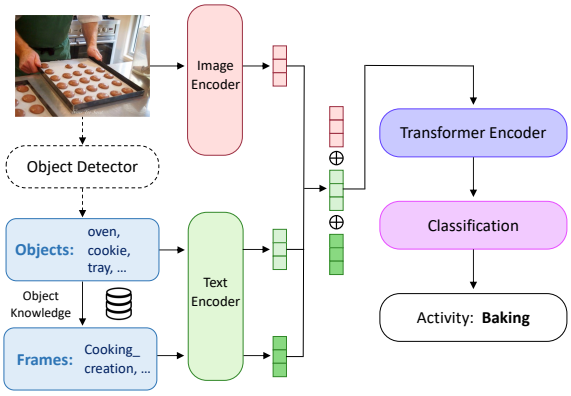

Figure 2: Overview of the ARF architecture.

verb being the activity for image I_i is defined as:

$$\Pr(v_i^j|I_i) = \frac{\exp(g(I_i, v_i^j))}{\sum_{k=1}^{m} \exp(g(I_i, v_i^k))} \quad (1)$$

where $g(\cdot)$ is a function produced by our model for scoring the assignment of a verb to the image, and m is the total number of candidate verbs. We use negative log likelihood as our loss function:

$$\mathcal{L} = -\sum_{i=1}^{n} \log \Pr(v_i^*|I_i) \quad (2)$$

3.2 Object Recognition

Ideally, we would use an Object Detector to identify the objects in an image for our experiments. However, the object detectors that are most readily available use categories that do not cover the range of object types that we need. For example, object detection datasets often contain a number of animate objects such as people and animals. As an alternative, we turned to image captioning systems. For our first set of experiments, we used a state-of-the-art image captioning model called OFA (Wang et al., 2022) to generate 10 different sentences that describe the image. We set beam size 10 and diversity 10. We then extracted the nouns from these sentences to create a set of words that (hopefully) include the objects.

However, even though the image captioning system often generated reasonable captions, the most relevant objects were frequently omitted from the caption, or misidentified.[1] Since the goal of our research is to determine whether *adding* explicit knowledge about an object improves performance, we cannot truly assess the value of such knowledge when we do not know what objects appear in the image. Developing better methods to identify specific objects in an image is an important direction for future research in computer vision. For now, we continued our investigation by performing additional experiments with the gold nouns in the imSitu dataset. These experiments essentially evaluate the impact of adding object knowledge when the objects have been perfectly identified by an oracle.

3.3 Prototypical Function Knowledge

We obtained the knowledge of what an object is typically used for from a dataset[2] created by (Jiang and Riloff, 2021). Their data contains a list of physical objects represented as WordNet synsets (Miller, 1995), and each object is paired with a human-annotated frame from FrameNet that represents its prototypical function. For example, *knife* is paired with the *Cutting* frame.

For each object in an image, we aim to use its function frame to help with activity identification. However, Jiang and Riloff (2021) and imSitu (Yatskar et al., 2016) used different subsets of frames from FrameNet. We felt that it made sense to align them, so we used the inter-frame relations provided by FrameNet to map the prototypical function frames to imSitu's frames. For each function frame, we create a mapping to all of the imSitu frames that are within one hop via any frame relation. Finally, we associate each object with its corresponding imSitu frames.

3.4 Activity Recognition Model

We use CLIP ViT-B/32 (Radford et al., 2021) as the backbone model to encode the image and text. For each example, we first apply CLIP's image encoder to produce an image feature vector. Then we use CLIP's text encoder to generate an embedding for each object (noun) and average the object vectors. For each object, we also collect its prototypical function frames and use CLIP's text encoder again to generate embeddings for each frame's name, then average those vectors. If there is no object, or no associated frame, then we encode an empty string.

Next, we build a transformer model consisting of 6 encoding layers and a classification layer on

[1] One likely reason is that the images are in low resolution and many objects are small, such as a pencil.

[2] https://github.com/tyjiangU/physical_artifacts_function

Model	Dev Acc	Test Acc
Yatskar et al. (2016)	32.3	32.3
Cooray et al. (2020)	38.0	38.2
Pratt et al. (2020)	39.6	39.9
Suhail and Sigal (2019)	43.2	43.3
Cho et al. (2022)	44.4	44.7
Li et al. (2022)	-	45.6
ARF	46.2	46.4
ARF+nouns$_C$	46.6	46.5
ARF+nouns$_C$+func	46.9	47.2
ARF+nouns$_G$	69.2	69.5
ARF+nouns$_G$+func	72.0	71.9

Table 1: Experimental results.

top. As input, the model takes the concatenation of all 3 vectors (corresponding to image, objects and functions). The classifier then selects the most probable action verb from all 504 candidate verbs used in the imSitu dataset.

4 Evaluation

The imSitu data contains 126,102 images, with manually annotated activity verbs and frame structures. We follow the same data split (train 75,702, development 25,200, test 25,200) as Yatskar et al. (2016). We report verb prediction accuracy on both the development and test sets. When fine-tuning the transformer, we use batch size 32, hidden vector dimension 512, AdamW optimizer with learning rate 1e-4 and train for 10 epochs.

4.1 Experimental Results

Table 1 compares our model with six previous methods described in Section 2. The ARF row shows the performance of our basic model using only image input. Our model performs a little better than previous systems, probably due to the CLIP model which is quite good. Also, the other models are trained for the full situation recognition task, whereas our model is trained solely for the verb prediction task.

The next two rows show results when adding embeddings for the nouns extracted from the captioning system (nouns$_C$) and when using the nouns as well as their function frames (nouns$_C$+func). The nouns alone produce just a tiny improvement, but adding the function frames improves a bit more. We believe that these results are primarily due to the limitations of the captioning system.

(a) spoon → Ingestion (b) shield → Protecting

(c) canoe → Motion (d) knife → Cutting

Figure 3: Object triggering function frames. The gold verbs are (a) feeding (b) guarding (c) floating, and (d) slicing.

The last two rows in Table 1 show the performance when using the gold nouns (nouns$_G$) and when using the gold nouns plus their associated function frames (nouns$_G$+func). These results show a huge performance boost simply from correctly identifying all the objects in the image. And providing the external knowledge about their prototypical functions further improves performance. In the next section, we try to better understand the role that objects play.

4.2 Analysis

Figure 3 shows some examples of how the functions of objects in the image can help identify the main activity. Consider subfigure (a), we see a hand-held **spoon** in front of the baby's mouth; the baby is expressing their like or dislike by making a grimace; there is some green substance (presumably food) both on the face and spoon. We don't see a series of continuous actions, yet we know it is a feeding event because of our commonsense knowledge. Similarly for the other images in Figure 3, from the shields, we can infer *Protecting*; looking at the canoe, we know it is *Motion*; and the knife is a good indicator for *Cutting*.

Images with and without Objects However, not all images contain "salient" physical objects. For example, imagine a picture showing a man running on a trail. The man is wearing clothes, which usually does not help with identifying the running activity (people generally wear clothes). In order to tease apart the images with and without salient ob-

Model	w/ Func	w/o Func
ARF	46.0	46.4
ARF+nouns$_G$	70.4	68.5
ARF+nouns$_G$+func	75.1	70.4

Table 2: Comparing performance on images with and without physical objects that have function frames.

jects, we divided the dev set into two subsets: one set (*w/ Func*) contains 8,957 images where at least one gold noun is associated with a function frame, and the other set (*w/o Func*) contains 16,243 images for which no nouns map to any frames. Since the gold annotations only provide semantic role values that are associated with the main activity, it is safe to assume that the *w/ Func* set of images would contain salient objects. Table 2 compares the performance of our systems on each subset of data. We see that performance is nearly identical when only using image features. Adding the gold nouns produces a big performance gain for both groups, although it benefits the *w/ Func* subset a little more. When the function frame knowledge is introduced, we see more separation: the images that depict physical objects associated with functions benefit more from having external knowledge about functions. This result confirms that the knowledge is beneficial in the expected way.

Which Semantic Categories Matter? The performance gap between ARF+nouns$_G$ and ARF is substantial, and we were curious to understand what types of nouns contributed the most. So we conducted another set of experiments on the dev set to identify certain types of semantic roles.

There are 190 different semantic roles in the data, but we are primarily interested in understanding the importance of physical objects. So we coarsely grouped the semantic roles into 3 categories roughly corresponding to *People*, *Locations* and *Objects*. To keep things manageable, we identified the 16 most frequent semantic roles that appear in at least 2,000 images and manually assigned them to the 3 categories. The *People* category includes *agent*, *agentpart*, *victim*, and *coagent*. The *Locations* category contains *place* and *destination*. The *Objects* category contains *tool*, *item*, *substance*, *object*, *container*, and *vehicle*. We disregarded a few semantic roles that are highly ambiguous (e.g., *source* can be both a location and object).

Table 3 shows our experimental results. Each experiment collected all images containing at least one instance of a relevant semantic role and then evaluated performance on those images both with and without the gold annotated nouns. For example, the *Objects* column shows that our model achieved 72.2% accuracy on the images that contain at least one object when it was given the nouns. But performance dropped to 37.2% accuracy on those same images without the nouns. In contrast, providing the gold nouns had much less impact on the other sets of images, which contain *People* or *Locations* but not necessarily *Objects*.

	People	Locations	Objects
with Nouns	69.3	69.2	72.2
without Nouns	61.4	64.4	37.2

Table 3: Performance with and without the nouns for specific semantic roles.

Salient Objects Another challenge is how to find the "salient" objects that play important roles in the image, and from which we have a better chance of identifying the main activity. We count the number of physical objects (not in the *People* or *Locations* semantic category) for all images. We find that nearly 40% of images are annotated with two or more objects. In our ARF model, when there are multiple objects in the image, we simply use the average of each object's embedding, which could potentially be improved by giving more weight to the most salient object. This issue may be even more important when using object detection systems because they may identify more objects (the gold annotation only contains objects that belong to a pre-defined semantic role)! This is an important issue to study in future work.

5 Conclusion

The prototypical functions of physical objects is a type of commonsense knowledge that is important for NLP. In this work, we showed that it can be a useful source of information for image understanding as well. Specifically, we tackled the situation recognition task by building a transformer model that incorporates the functions of objects to predict the activity in an image. The experiments show that knowledge of the objects and their prototypical functions can improve performance on this task. However, automatically recognizing the objects in an image remains a challenge, and exploiting better object detection methods is an important direction for future work.

6 Limitations

For image captioning, we used the pre-trained OFA model for zero-shot inference. We did not explore every state-of-the-art model or fine-tune OFA specifically on the imSitu dataset. Other image captioning systems could yield better results. The gap between automatic object recognition and using gold nouns confirms that correctly identifying the objects in an image is very important for activity recognition. Also, we are not certain that mapping the Jiang and Riloff (2021) function frames to the imSitu frames is strictly necessary.

Acknowledgements

We thank the Utah NLP group for their constructive comments. We also thank the anonymous ACL reviewers for their valuable suggestions and feedback.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision (ICCV 2015)*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*.

Mark H. Burstein. 1979. The use of object-specific knowledge in natural language processing. In *Proceeding of the 17th annual meeting on Association for Computational Linguistics (ACL 1979)*.

Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Thilini Cooray, Ngai-Man Cheung, and Wei Lu. 2020. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Tianyu Jiang and Ellen Riloff. 2021. Learning prototypical functions for physical artifacts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL 2021)*.

Tianyu Jiang and Ellen Riloff. 2022. Identifying physical object use in sentences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV 2014)*.

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Michele Persiani and Thomas Hellström. 2019. Unsupervised inference of object affordance from text corpora. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision (ECCV 2020)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML 2021)*.

Josef Ruppenhofer, Michael Ellsworth, Myriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. 2018. Dock: Detecting objects by transferring common-sense knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*.

Mohammed Suhail and Leonid Sigal. 2019. Mixture-kernel graph attention network for situation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning (ICML 2022)*.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.