More Than Eleven Thousand Words: Towards Using Language Models for Robotic Sorting of Unseen Objects into Arbitrary Categories

Anonymous Author(s)
Affiliation
Address
email

Abstract:
We consider the task of automatically sorting previously unseen objects into arbitrary categories. We aim to sort into general, high-level categories in contrast to traditional methods that sort on visually discernible features or by other sensor measurements. This paper explores a method where we divide the categorization into two sub-tasks: object detection and categorization. In a set of experiments, it is shown that splitting the categorization task into a two-stage process removes highly important information for robust categorization and performs far worse than using an open vocabulary object detector. We hope these results are helpful for exploring the limits of Language Models for robotic tasks.

Keywords: Language Models, Robots, Learning, Sorting

1 Introduction

We recognize the ubiquity of sorting tasks. From industrial settings, second-hand stores, and household services, the act of sorting objects provides value. So far, however, many sorting tasks have been restricted to human execution. Considering this, we consider the question: *Can we devise a robotic system that can sort objects efficiently, in a highly general manner?*

A usual approach to sorting systems is training on a fixed set of categories, such as color, material, or other visually discernible or measurable features with sensors or cameras. Compared to humans, these methods suffer from two significant distinctions that hinder their flexibility: (i) they need to be retrained when presented with new categories, and (ii) the categories must be discernible from sensor input. These characteristics limit the use of robotics in cases where categories might change often and are provided as high-level expressions. Consider a service robot in a specific household where the pots and pans are stored in one drawer and children's toys in another. Reprogramming the sorting system for each house is limiting. Another application is second-hand stores, such as Goodwill, which might have predefined categories they sort after, determined by factors such as target groups, seasons, or campaigns. These categories might change rapidly, and to compete with humans, the sorting systems should allow rapid switching between categories. From an environmental perspective, improving the throughput of the reuse industry will positively impact the circular economy and lower the demand for production.

An apparent challenge is that the categories are previously unseen, separating this from traditional classification in computer vision. We must capture the relation between an arbitrary object and an arbitrary category. A challenging aspect of this task is the wide range of objects and categories we can encounter. For many categories, modeling the relationship between the sensor observation and category membership relies on a complex semantic understanding. Consider the class of children's toys, a category containing objects of all shapes, weights, colors, and sizes.

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

Several works have recently shown that combining Language Models (LMs) with robotics systems allows for understanding natural language and complex reasoning for long-horizon planning. [1, 2] An emerging challenge is ensuring that the proposed actions are executable in practice. Given the problem of flexible and high-level sorting and the advancements of LMs for robotics, we want to answer the pertinent question: *Can we incorporate the semantic understanding of LMs into a robotic system to achieve efficient yet flexible sorting behavior?*

The sorting task can be separated into the following tasks: (i) Detect a single object in the scene and determine its category and subsequently (ii) determine pick-and-place positions, and lastly (iii) execute the pick-and-place action with the robot. We recognize that detection and categorization is the primary limiting factor for allowing a larger set of classes and objects. Responding to this challenge, we explore a novel approach for categorization by separating the task into two steps. First, perform classification, namely assigning the membership to a specific class, and categorization, assigning this class to a higher-level category. Figure 2 shows a schematic view of this approach.

We conduct experiments on images to measure the performance of the methods. The results show that it is beneficial to predict the category directly from the image, pointing to the fact that visual appearance provides essential cues for determining the category of the object.

Section 2 will provide an overview of the related work, Section 3 will describe the approach of our method, Section 4 will present experimental results, and Section 5 will conclude and point to further research.

2 Related work

Robotic sorting Several systems perform sorting of objects into predefined categories; however, to the best of our knowledge, no previous work has considered the problem of open vocabulary robotic object sorting. Several works have investigated the sorting of objects based on specific properties such as color, shape, and material. [3, 4, 5] An application for these methods is waste sorting, where Lukka et al. [6] performs sorting by material properties. Similarly, Kujala et al. [7] sorts objects by color off a conveyor belt.

Automatic sorting Guérin et al. [8] explore the problem of sorting objects without specified categories. They use a Convolutional Neural Network in combination with a clustering algorithm to group objects into a given set of bins. Their method is, however, restricted to sorting objects of similar appearance. The sorting system is also unaware of the semantic meaning of the different classes, and the system does not allow us to provide categories.

Language models for robotics Zeng et al. [2] show that combining LMs with Visual Language Models (VLMs), allowing them to communicate through language, results in an overall system with a high understanding of the scene. Specifically, they use VLMs to inform the system what objects are in the scene, and a prompt format guides the LM to output code-like responses, e.g., pick-and-place actions. Ahn et al. [1] compares the predictions from the LM to affordance functions, predicting the most relevant action given the robot's surroundings and the given task description.

Object retrieval Nguyen et al. [9] focus on the problem of retrieving the most relevant object given a command containing a verb, imbuing the robot with semantic understanding. Their approach is limited to retrieval given an action that can be done with the object, whereas our approach considers a free space of categories.

3 Proposed method

We propose SortingBot, which consists of several parts. We divide the challenge of sorting the objects into (i) object detection and classification and (ii) grasp synthesis and pick-and-place operations execution. This paper focuses on the former task and investigates how LMs could be used for this purpose.

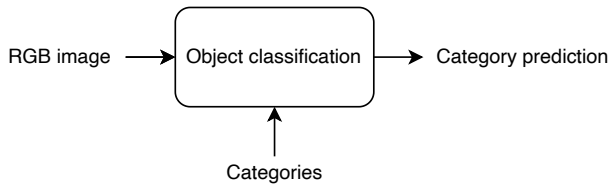

Figure 1: Schematic for the approach of direct categorization.

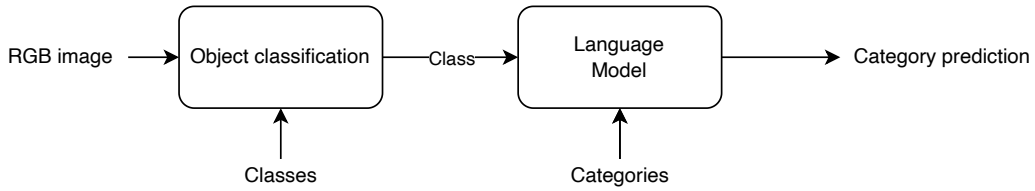

Figure 2: Schematic for the approach of separating classification and categorization.

We will discuss two approaches to solving object categorization:

1. Categorizing directly using an open vocabulary object detector.
2. First, detect objects using a large vocabulary and then categorize them using a language model.

These will be described in detail in Section 3.2 and Section 3.1.

3.1 Categorization directly

Here, we aim to categorize the objects directly, meaning we find the associated category without classifying the objects first. This aligns with the task of an open-language object detector, which can detect objects using an arbitrary vocabulary.

We hypothesize that this method will work well for categories within its training distribution. This is not restricted to purely classes that the detector is trained on, but also includes the level of abstraction. Perhaps classes such as "cat," "poodle," and "bicycle" are more straightforward to recognize than "travel-related," "furniture," and "kids toys"?

3.2 Separating detection and categorization

An overview of this approach is shown in Figure 2. In the spirit of Ahn et al. [1], we incorporate an LM in the task of assigning the object to a category. We achieve this by splitting the task into classification and categorization. In short, we use an object classification model to obtain $P(\text{class}|\text{image})$, where "class" is an class in the vocabulary of the classification model (e.g. "apple"). Subsequently, we query an LM to obtain $P(\text{category}|\text{class})$, where "category" is one of the given categories (e.g. "food").

We can express the original task of estimating the category C from an image I as finding:

$$C^* = \arg\max_{C} P(C|I).$$

Incorporating an LM, we reformulate the categorization problem to depend on the object's class. Then, we want to assign two attributes to the object in the image I, namely the class of the object O and its corresponding category C. We are then interested in maximizing:

$$(C^*, O^*) = \arg\max_{C,O} P(C \cap O|I).$$

We write the above expression as the two conditional probabilities:

$$(C^*, O^*) = \arg\max_{C,O} P(C|O \cap I)P(O|I).$$

We now make two assumptions to allow the separation of concerns. First, we assume that $P(O^*|I) = 1$, meaning that we have a perfect object classifier predicting only one relevant class. The maximization is then only over the categories. Secondly, we assume that information about the object's class is sufficient to estimate the category, meaning that the category C is conditionally independent of the image I. We simplify the expression to:

$$\hat{C}^* = \arg\max_{C} P(C|O). \tag{1}$$

With classification, we mean detecting the objects in the scene and classifying them into a set of classes. The classes can be fixed, and the classifier can be trained specifically for them. Therefore, this task can be performed by an object detector providing instance masks. For categorization, we can use an LM to get the probability that an object class is assigned to a given class by careful prompt engineering. The details are provided in Section 4.

We hypothesize that this two-stage approach is beneficial when it is challenging to determine the category given the detected object class. This might be the case when reasoning is required to determine the category. For example, it might be easier for an LM to capture that and "apple" should be in the "food" category than it is for an object detector to detect something as "food" in the image directly.

4 Experiments

Here, we present experimental results from each of the presented methods in Section 3 for categorizing objects. We benchmark the methods on a task of classifying images of singular objects into a set of given categories.

For the direct categorization approach, we use an open-language object detector. Specifically, we choose ViLD [10], representing state of the art on several datasets. Pre-trained models are publicly available.

We choose ViLD as the object detector with a fixed vocabulary and GPT-3 as the LM [11] for separating the classification and categorization. As we want the object detector to be as specific as possible, we use the class labels from the Tencent ML Image dataset [12] as this vocabulary. It combines the categories from both ImageNet [13] and Open Images [14], resulting in 11 166 categories. We do not do any fine-tuning of the object detector on the Tencent ML dataset and note that this might help improve this approach.

We use GPT-3 with the following prompt:

```
prompt = "Classify each of the following objects as either "
prompt += ", ".join(category_list[:-1])
prompt += " or " + category_list[-1] + ". "
prompt += "Object: " + detected_object + " Label:"
```

where `detected_object` is the detected object class, and `category_list` is the each of the provided categories. We make a prompt for each category, where the relevant category is appended to the prompt. Using the language model, we choose the category corresponding to the most probable prompt.

4.1 Open source images in the wild

To indicate the robustness of the two methods, we sample a small set of 23 images from the web and measure the classification accuracy for each method.

Table 1: The different sets of categories used in the experiments.

Narrow categories	Broad categories	YBC categories
newer books	eating related	food items
LP records	entertainment	kitchen items
CDs and DVDs	decorative	tools
small interior items		
crockery		
mugs and glasses		
serving bowls		
cutlery		
vases		
ornaments		
candlesticks		
tablecloths		
decorative pillows		
wool blankets		
pictures and paintings		
kitchen utensils		
small furniture		
toys and games		
working electrical items		
working kitchen equipment		

Figure 3: ViLD categorization.

Motivated by the second-hand industry, we sort the images into a set of categories that might be used when receiving second-hand goods. These are listed in the first column of Table 1.

The accuracy for each method is listed in the first column of Table 2. Here, the direct approach is an order of magnitude better than the two-stage approach, at 45% in comparison to 4.5%. Figure 3 and Figure 4 show the result on a sample of images using the open vocabulary object detector and the two-stage approach. Here too, we see that the direct approach yields higher accuracy.

Figure 4: Categorization results using two-stage approach (ViLD+GPT-3).

Figure 5: ViLD categorization on the broad categories.

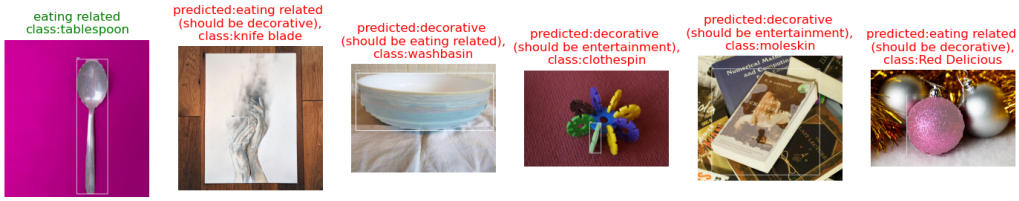

Figure 6: Categorization results using two-stage approach (ViLD+GPT-3) on the broader categories.

4.1.1 Categorization into broader categories.

Motivated by the hypothesis in Section 3.2, we test the methods on fewer and broader categories. These are listed in the second column of Table 1.

The second column of Table 2 lists the respective accuracies on this task. As previously, the direct method outperforms the two-stage approach. However, both are close to the performance of a random classifier, which would yield an accuracy of 33%. This is also visible in the visualization on the subset of images shown in Figure 5 for the direct approach and Figure 6 for the two-stage method.

4.2 Categorization of YCB-objects

We apply both approaches to images of objects in the YCB-dataset [15], which amounts to 48 images. We use the categories in the third column of Table 1.

The results are summarized by their accuracy in Table 2. The direct approach shows superior performance in this task too. The direct approach results for a set of images are shown in Figure 7, and comparing these to Figure 8 showing the results of VILD+GPT-3, we can see that the direct method performs more robustly.

4.3 Discussion

For both methods, a significant failure mode is inappropriate bounding boxes. In the case where the bounding box is too small, the categorization is challenging due to the lack of knowledge of

Table 2: The accuracy of the two methods for the narrow and broad set of categories

Method	Narrow categories	Broad categories	YCB images
ViLD	0.455	0.409	0.638
ViLD+GPT-3	0.045	0.364	0.553

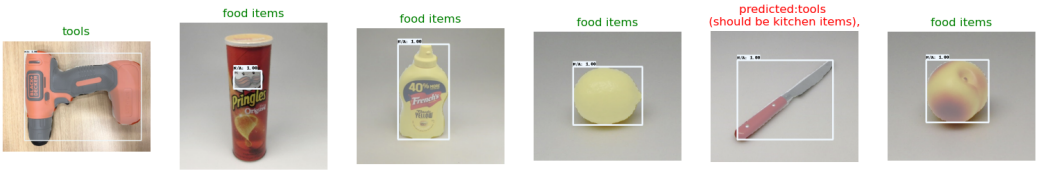

Figure 7: ViLD categorization of the YCB objects.

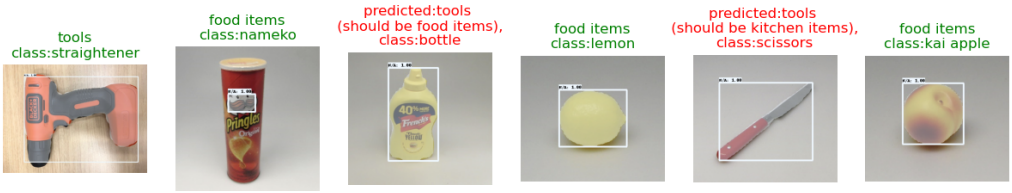

Figure 8: Categorization results using two-stage approach (ViLD+GPT-3) on the YCB objects.

the context. For example, consider the children's toy in Figure 3. In this case, both methods fail to categorize the whole object correctly.

Looking at Table 2, we see that ViLD directly performs better than ViLD+GPT-3 overall. This points to the fact that important information is lost in the classification process. Take the drawing of the hand in Figure 4 as an example. Here, ViLD detects a bounding box around the finger of the drawn hand. How this is handled further is different for the two methods. For ViLD+GPT-3, this is classified as a "knife blade", which arguably has a similar appearance as the depicted finger. This is then categorized as "cutlery" by GPT-3. Applying ViLD directly to this bounding box, however, results in the correct prediction of "decorative". We theorize that the texture and more minor details are highly relevant to the correct classification, and these crucial cues are lost when first conducting the classification task.

Referring to Equation 1, we see that assuming the conditional independence $P(\text{category}|\text{class} \cup \text{image}) \approx P(\text{category}|\text{class})$ is a faulty assumption, as in many cases information such as texture is highly relevant to predicting the category. In addition, the assumption of a perfect classifier is also not valid, as the categorization fails frequently due to the wrong class being detected.

5 Conclusion

In this paper, we compare two approaches for sorting objects in RGB images motivated by improving robotic sorting. We implement an approach where the task of determining the class of an object (e.g., "apple") is separated from categorizing the object into categories (e.g., "food"). In a set of experiments, we measure the accuracy of previously unseen images and categories. After that, we present insights into where the different methods fail and succeed. The results indicate that information about appearance, such as detailed texture, is highly relevant to determining the category. Therefore, a direct approach behaves more robustly in our experiments.

As directions for further work, improving direct, open-vocabulary categorization methods emerges as a promising direction. A study of the effect of prompt engineering could provide some interesting hints on how to improve VLMs in general. Additionally, in the case of second-hand stores, it is not only crucial to group objects into categories, but also evaluate whether the object is in good condition and applicable for reselling. This new dimension is helpful to consider in further work to progress towards more effective second-hand sorting systems. Another point of further work is investigating

the gain of using RGB-D images for classification, as such observations are typically available on modern robotic platforms. This might help increase the detection, classification, and categorization robustness.

References

[1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, and M. Yan. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Apr. 2022. URL http://arxiv.org/abs/2204.01691. arXiv:2204.01691 [cs].

[2] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, May 2022. URL http://arxiv.org/abs/2204.00598. arXiv:2204.00598 [cs].

[3] T. Dewi, P. Risma, and Y. Oktarina. Fruit sorting robot based on color and size for an agricultural product packaging system. *Bulletin of Electrical Engineering and Informatics*, 9(4):1438–1445, Aug. 2020. ISSN 2302-9285. doi:10.11591/eei.v9i4.2353. URL https://beei.org/index.php/EEI/article/view/2353. Number: 4.

[4] J. Guérin, O. Gibaru, S. Thiery, and E. Nyiri. Clustering for Different Scales of Measurement - the Gap-Ratio Weighted K-means Algorithm, Mar. 2017. URL http://arxiv.org/abs/1703.07625. arXiv:1703.07625 [cs, stat].

[5] M. Gupta and G. S. Sukhatme. Using manipulation primitives for brick sorting in clutter. In *2012 IEEE International Conference on Robotics and Automation*, pages 3883–3889, May 2012. doi:10.1109/ICRA.2012.6224787. ISSN: 1050-4729.

[6] T. J. Lukka, Timo Tossavainen, T. Tossavainen, J. V. Kujala, and T. Raiko. ZenRobotics Recycler - Robotic Sorting using Machine Learning. Jan. 2014. MAG ID: 2560865072.

[7] J. V. Kujala, T. J. Lukka, and H. Holopainen. Classifying and sorting cluttered piles of unknown objects with robots: A learning approach. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 971–978, Oct. 2016. doi:10.1109/IROS.2016.7759167. ISSN: 2153-0866.

[8] J. Guérin, S. Thiery, E. Nyiri, and O. Gibaru. Unsupervised robotic sorting: Towards autonomous decision making robots, Apr. 2018. URL http://arxiv.org/abs/1804.04572. arXiv:1804.04572 [cs].

[9] T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex. Robot Object Retrieval with Contextual Natural Language Queries, June 2020. URL http://arxiv.org/abs/2006.13253. arXiv:2006.13253 [cs].

[10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, May 2022. URL http://arxiv.org/abs/2104.13921. arXiv:2104.13921 [cs].

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

[12] B. Wu, W. Chen, Y. Fan, Y. Zhang, J. Hou, J. Liu, and T. Zhang. Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning, Feb. 2020. URL http://arxiv.org/abs/1901.01703. arXiv:1901.01703 [cs].

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi:10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

[14] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[15] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, Sept. 2015. ISSN 1558-223X. doi:10.1109/MRA.2015.2448951. Conference Name: IEEE Robotics & Automation Magazine.