How to Recycle: General Vision-Language Model without Task Tuning for Predicting Object Recyclability

Abstract

Waste segregation and recycling place a crucial role in fostering environmental sustainability. However, discerning the whether a material is recyclable or not poses a formidable challenge, primarily because of inadequate recycling guidelines to accommodate a diverse spectrum of objects and their varying conditions. We investigated the role of vision-language models in addressing this challenge. We curated a dataset consisting >1000 images across 11 disposal categories for optimal discarding and assessed the applicability of general vision-language models for recyclability classification. Our results show that Contrastive Language-Image Pre-training (CLIP) model, which is pretrained to understand the relationship between images and text, demonstrated remarkable performance in the zero-shot recyclability classification task, with an accuracy of 89%. Our results underscore the potential of general vision-language models in addressing real-world challenges, such as automated waste sorting, by harnessing the inherent associations between visual and textual information.

1 Introduction

Waste management, facilitated by proper recycling, is paramount in fostering sustainability, as it helps in the conservation of natural resources, mitigation of environmental harm, and re-utilization of recyclable materials instead of their disposal. The repercussions of inadequate waste disposal are severe, e.g., leaching toxic chemicals from batteries can pollute groundwater and decomposing organic waste in landfills can emit methane, a potent greenhouse gas. Since the 1960s, the per capita generation of garbage has doubled to $\sim$5 lbs/person/day [5]. A major obstacle to recycling is the widespread ambiguity surrounding the recyclability of certain items. The current classification bins labeled with generic categories such as paper, plastic, and metal often fall short of providing clear guidance. For instance, while cardboard falls under paper and is thus recyclable, does this include cardboard affixed with tape? Similarly, plastic bottles are generally recyclable, but what about those containing liquid?

Machine learning models offer a viable solution to these challenges. However, the absence of large-scale, labeled datasets annotated with recyclability is a fundamental obstacle. Additionally, the multifaceted nature of materials and their varying conditions, such as differing states of cardboard, necessitate models capable of discerning a potentially limitless array of categories. Thus, models that grasp detailed information about both the substance and its state are pivotal.

Our study explores the zero-shot classification capabilities of CLIP[4], a vision-language model trained on 400 million (image, text) pairs collected from the internet. Given its extensive and diverse training dataset, CLIP possesses significant potential to comprehend nuanced relationships between images and texts and exhibits proficiency in tasks for which it was not explicitly trained. Our findings reveal that CLIP can accurately classify materials according to their recyclability status, showcasing

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

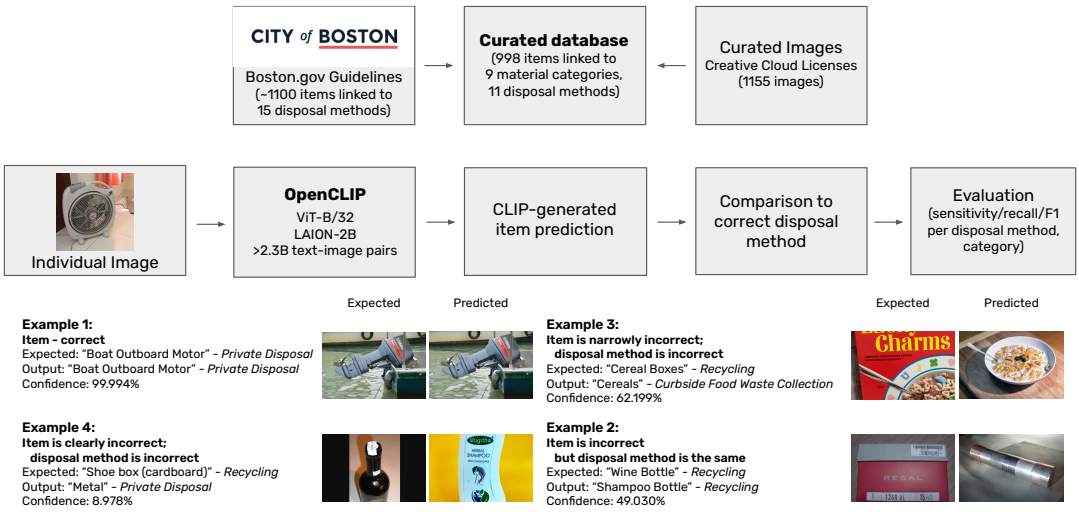

Figure 1: Overview: A database of 998 items and linked images was curated based on the City of Boston guidelines. CLIP-predicted item labels were converted to the disposal method based on the database and compared to the disposal method of the original image. Four examples are shown.

its potential as a valuable tool in offering informed guidance in waste management. Fig. 1 shows the overview of our approach.

2 Related Work

Prior work has explored the application of various machine learning methods for waste material classification. In 2016, Yang and G. [6] applied a convolutional neural network (CNN) on a custom dataset of ∼2500 images in 6 categories (paper, glass, plastic, metal, cardboard, trash). However, they found that the performance of their naive CNN performed poorly (27% using a 70/30 split), worse than SVM (63% accuracy). Several studies have subsequently employed variants of CNN to classify materials.

Narayan [3] implemented multiple CNN models to categorize images into compostable, recyclable, and landfill divisions. For model training, the authors amassed 1218 images and initiated the models with weights pre-trained on ImageNet, achieving accuracies ranging from 0.82 to 0.88 across InceptionV3, Resnet50, Mobile Net, and PNAS Net models, with Resnet_50 performing the best.

In another study by Liu and Liu [2], a variety of CNN models were evaluated on a dataset comprising over 12,000 images (four classes: glass, metal, paper, and plastic) from five previous efforts. The results indicated that a 16-layer VGGNet variant, integrated with transfer learning, exhibited optimal performance (84.6% accuracy, 80/20 split testing). Furthermore, in 2022, Liu et al. [1] introduced a novel method named DSCAM (Depth-wise Separable Convolution Attention Module) specifically for waste material classification.

Each of these methodologies required the tuning of model parameters using a labeled dataset for training. In contrast, in our current study, we employ an innovative approach based on a model that has been pre-trained on text-image pairs, allowing for a complex understanding and classification of waste materials.

3 Methodology

Data Collection Recycling protocols vary significantly based on the geographical region and the capabilities of the respective processing facility. For instance, while broken glass shards are non-

recyclable in Boston, they might be accepted in other regions. In conducting the current study, we adhered to the local guidelines as stipulated by the Boston.gov Recycling & Trash Directory's waste management website. [1] This website offers recommended disposal methods for a wide range of items upon search. We curated all accessible items directly from the website, assembling a comprehensive list of 1100 items along with their corresponding disposal methods. These items were then categorized under 11 distinct disposal methods (4 methods with very few items were merged into "Special Instructions") as shown in Fig. 2.

To curate our database, we used Google Images and other websites to find eligible images (Creative Commons license) that correspond to each of the 1100 items. Our search was focused on acquiring the most precise and realistic images, prioritizing clarity and accurate depiction of the items. Despite our efforts, we encountered 112 instances where the objects were either indiscernible or lacked images available under a Creative Commons license. This was particularly true for specific chemical items with visual similarities, such as Acetone and Paint Stripper, distinguishable primarily through their labels. Consequently, these categories were excluded from our database. Additionally, we supplied multiple images for some objects that were underepresented in the boston.gov database (e.g. wood). Ultimately, our dataset comprised 998 items and 1155 images.

Contrastive Language-Image Pretraining (CLIP)

CLIP[4] is a vision-language model trained on 400 million (image, text) pairs collected from the internet. It minimizes the contrastive loss term between (image, text) pairs, mathematically represented by -

$$L = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N} \exp(s_{i,j}/\tau)}$$

where $s_{i,j}$ is the similarity score between the i^{th} image and the j^{th} text in a batch, N is the batch size, and τ is the temperature parameter controlling the concentration of the distribution. The model, therefore, learns to associate the correct image-text pairs by maximizing the similarity scores $s_{i,i}$ for correct pairs and minimizing the scores $s_{i,j}$ for incorrect pairs, thus effectively learning meaningful representations of images and text.

For the recyclability classification, we used the CLIP ViT-B/32 model trained with the LAION-2B English subset of LAION-5B [2]) using OpenCLIP[3].

4 Results and Analysis

Approach To evaluate the zero-shot accuracy of CLIP, the image to be analyzed and the list of 998 possible items from the Boston Recycling Database were provided as inputs. The similarity of the image to each of the items was computed, and the item with the highest similarity score was selected. The disposal method of the image and that of the predicted item were then compared to compute accuracy. This allowed for a more fine grained classification instead of simple recyclable vs. non-recyclable categorization. Although multi-class prediction poses increased difficulty, the capacity to discern, for instance, an item that falls under "electronics recycling" from "recycling", or an item categorized as "regular trash - Bulky items" requiring separate pickup from "regular trash", proves to be immensely beneficial for enhancing efficient sorting at the disposal stage.

Another layer of complexity in our classification endeavor arises from the extensive array of items within our dataset. For instance, 'gum' (regular trash) must be distinguished from 'pills' (hazardous waste); and area rug (regular trash) must be distinguished from 'carpet' (private disposal). The large number of items to curate may have also introduced errors. For instance, 'apple' belongs to 'curbside food waste collection'. But a CLIP prediction returned 'nectarine' which, according to the Boston Recycling site, belongs to 'regular trash.' Given that some cities categorize 'nectarine' under 'curbside compost', we speculate that there might have been a curation discrepancy. Such imperfections within the database are anticipated to adversely affect our results.

[1] https://www.boston.gov/departments/public-works/recycling-boston

[2] https://laion.ai/blog/laion-5b/

[3] https://github.com/mlfoundations/open_clip

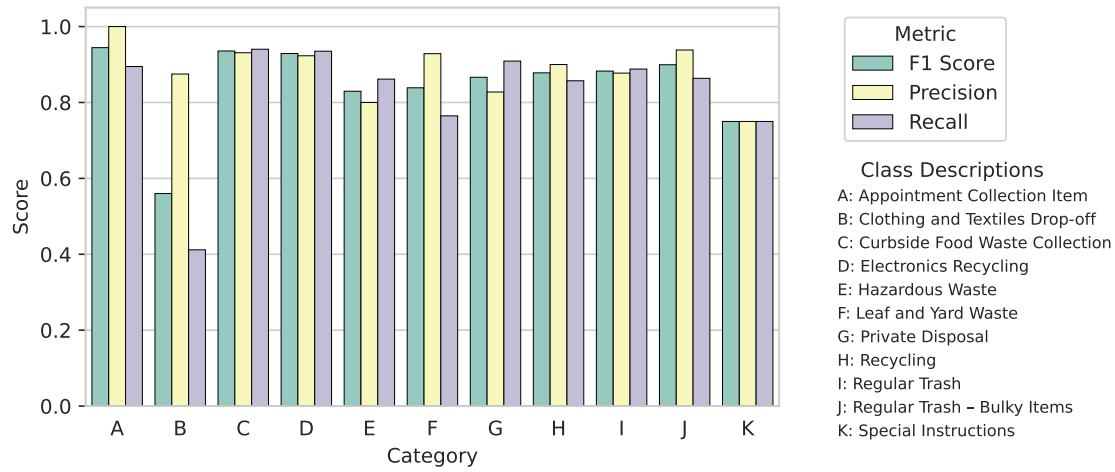

Figure 2: CLIP performance measures for multi-class prediction.

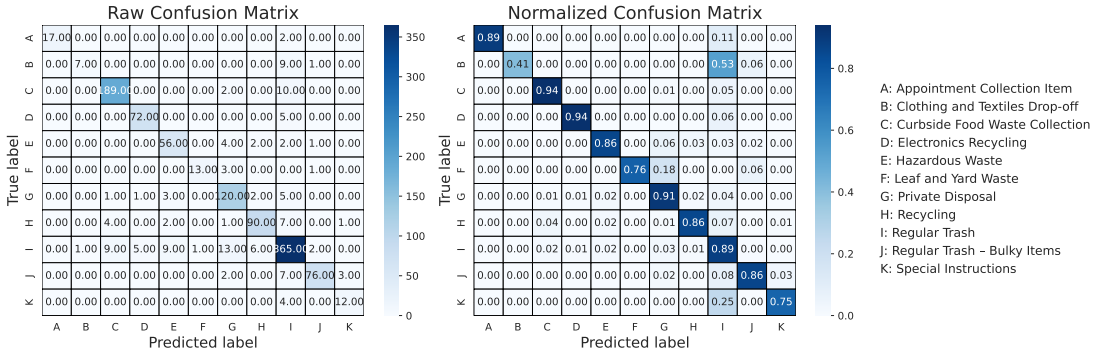

Figure 3: The CLIP predictions in raw (left) and normalized (right) confusion matrices

Performance of CLIP For image identification, CLIP correctly identified 70.8% of the images. Given the difficulty of item-specific prediction, this attests to the excellent performance of CLIP. For disposal method prediction, the overall accuracy was 89%. The mean precision, recall, and F1 for each category (1-vs-many) were 0.890, 0.831, and 0.853, respectively (Fig. 2). The confusion matrices (raw and normalized) are shown in Fig. 3. These matrices highlight difficult cases, e.g., the lowest accuracy category is "Clothing and textiles - Drop-off", for which 8 are predicted correctly, 8 are predicted to be "Regular Trash", and the last is predicted to be "Regular Trash - Bulky items".

5 Discussion

Our results demonstrate the zero-shot capabilities of CLIP for recyclability classification without task-specific tuning. In comparison to the simpler tasks in the past (e.g., recycle vs trash vs compost), our problem encompassing a comprehensive city guideline was more challenging. In addition to the 11 disposal methods, our database had a much larger variety of items compared to existing datasets. CLIP showed high accuracy in correctly classifying the disposal method.

To enhance accuracy, it is imperative to augment both the quantity of items and the number of images per item within the dataset. A major constraint for our dataset was the limited number of publicly available images with a Creative Commons license. Our examination of incorrect predictions shows the promise of additional curation efforts. Furthermore, leveraging the confidence level in our predictions can be helpful for better classification. Those with low confidence could potentially be flagged for manual inspection rather than returning a disposal method.

References

[1] F. Liu, H. Xu, M. Qi, D. Liu, J. Wang, and J. Kong. Depth-wise separable convolution attention module for garbage image classification. *Sustainability*, 2022.

[2] K. Liu and X. Liu. Recycling material classification using convolutional neural networks. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 83–88, 2022.

[3] Y. Narayan. Deepwaste: Applying deep learning to waste classification for a sustainable planet. 2021. URL `https://arxiv.org/abs/2101.05960`.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

[5] United States Environmental Protection Agency. National overview: Facts and figures on materials, wastes and recycling, 2022. URL `https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/national-overview-facts-and-figures-materials`.

[6] M. Yang and T. G. Classification of Trash for Recyclability Status. *Stanford University, CS229 Project Report*, 2016.