Computer Vision to reduce food wastage

Sharath Giri and Anitesh Reddy
April 28, 2023

Abstract

In the last decade, computer vision has taken massive strides towards helping humans to better understand the universe around us. This has aided the process in many fields and impacting individuals and society at various levels. Our focus is to utilize the advancements in science of computer vision to address a key issue of food wastage around the globe. We believe that there is a lot of food wastage in many industries due to poor feedback mechanism such as, food served in flights, buffets in restaurants, even at homes etc. Our project will explore the possibility of combining various established models to achieve this task for instance, GLIP, RESNET50 etc. Our intention is to create a model which can be used by airlines, restaurants and individuals to save food.

1. Introduction

Airlines spend around USD 20 Billion on food and beverages out of which approximately USD 4 Billion worth of food is wastage [2]. Absence of a structured feedback mechanism and government regulations are the major factors for this wastage. Roughly one-third of the edible parts of food produced for human consumption, gets lost or wasted globally, which is about 1.3 billion ton per year. In North-America and Europe around 10-15% of this wastage by consumers.[3] Our final goal is to create a model which can be run on an edge device and capture videos of the completed food trays and feeds it into the model which in turn gives feedback on food wasted. For instance, we would plant this device on top of service carts which will capture the video of trays being collected after food consumption. Similarly, these devices can be placed on top of collection trays at restaurants (Example of a food tray before and after consumption in a n aircraft: Figure1). Over a period of time, if we consistently get feedback that a dish is not completely consumed, we can conclude either the portion size can be reduced or the taste is to be question.

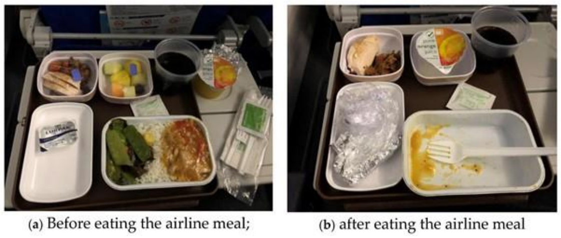

(a) Before eating the airline meal; (b) after eating the airline meal

Figure 1. Shows figure 1a and 1b, our model will read figure 1b and identify the different containers to predict how much of that dish is wasted

Photo credit: Fangzhou You ,Tracy Bhamra and Debra Lilley

Our model should be able to capture figure 1b and process it and give feedback that main dish is completely eaten and salad is half eaten. In some cases, we may face a situation where the container is not clearly visible and the model should be able to identify those and not give any feedback.

2. Related Work

There is quite some research done to quantify the amount of food that is being wasted along the supply chain from production to consumption. There are behaviour studies done as well to understand what is behaviour aspect is causing the wastage of food ref3. Airlines have also started to adopt AI to bring down food wastage.

Very recently, on 11th April 2023 Airbus announced usage that they are working on food scanner to help airlines reduce food wastage [3]. Similarly, Lufthansa has announced (on 23rd Mar 2023) that they will be using similar approach as our model to reduce food wastage but will only be able to roll it out in 2024 [4]. Limitics is a Singapore based start-up which is at forefront of the concept to use AI to reduce food wastage in many industries such as Airlines, Hospitality etc. They have partnered with Eithad Airways to come up with similar approach as ours and have been consuming data from aircrafts for over a year now. [1] Similarly, there are efforts made to reduce food wastage at restaurants using similar concepts. Overall industries are definitely moving in the direction of adopting AI to reduce food wastage. We haven't come across papers publishing usage of computer vision to reduce food wastage.

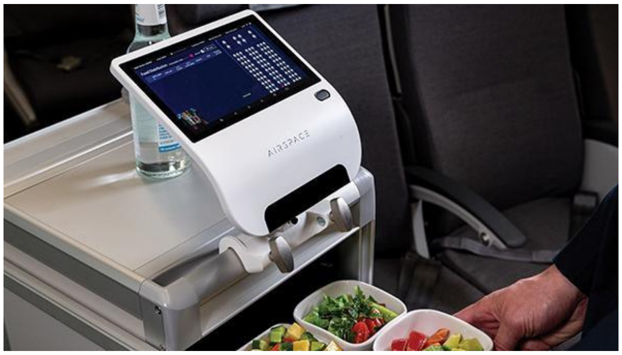


Figure 2. On 11th April 2023, Airbus Prototype of food scanner which will capture the images and process them to generate insights on food wastage

Photo credit: Airbus

3. Model

3.1. Data Collection

DALL-E for Image Generation: The first step in our proposed system is to use DALL-E, a generative deep learning model that produces high-quality images based on a given text prompt. In our case, we use DALL-E to generate images of meals with varying levels of food consumption. These images will then be processed by the GLIP model for cropping. Currently we are targeting on predicting the consumption of Rice only. Hence, we have generated an image dataset of around 900 images and manually classified them into 4 categories namely as eaten, over-served, tasted, untouched.

Apart from DALL-E we are collecting images manually which are corrupt. Our definition of corrupt image is an image where the food is not visible correctly. example: Figure3.

Test Data: We have manually clicked 256 images of food trays to test our model. In this tray, we included a main dish(Rice), salad, dessert (brownie) and bread. All permutations & combinations of 4 categories were applied to each of these dishes to come up with the test data set. We have randomly corrupted 20 images while taking these pictures.

3.2. Model Architecture

The proposed system consists of a combination of GLIP (Generative Latent Image Perception) followed by a model for identifying corrupted images and a fine-tuned ResNet50 architecture integrated with transformers for classifying the degree of food consumption. The system demonstrates promising performance in classifying unseen data, suggesting potential applications in the food industry for efficient food waste management and meal planning. High level model architecture is provided in figure4.

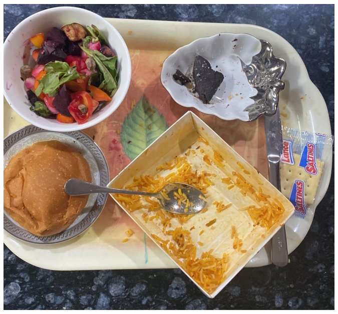


Figure 3. Sample image of a test image

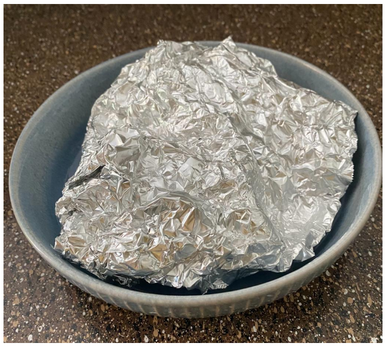


Figure 4. Image of a corrupt dish

GLIP Model for Image Cropping: We use the GLIP (Grounded Language-Image Pre-training) model [8] to automatically crop the generated meal images, focusing on the regions containing food items. This step ensures that our system processes only relevant portions of the images, which will improve the accuracy and efficiency of the subsequent classification model. We plan to test our model on unseen data as well. Thus, we plan on using few images manually obtained from the internet and try to crop out the relevant portions where the food is present.

Image Corruption Detection: Before classifying the degree of food consumption, it is crucial to ensure that the images used are of high quality and free from any corruption. Images are classified as corrupt if there are foils/tissue present in the tray or if the lid is closed. For this purpose, we employ a separate deep learning model specifically trained to identify corrupted images. This model acts as a preprocessing step, filtering out any unusable images before they are inputted into the main classification model. To ob-

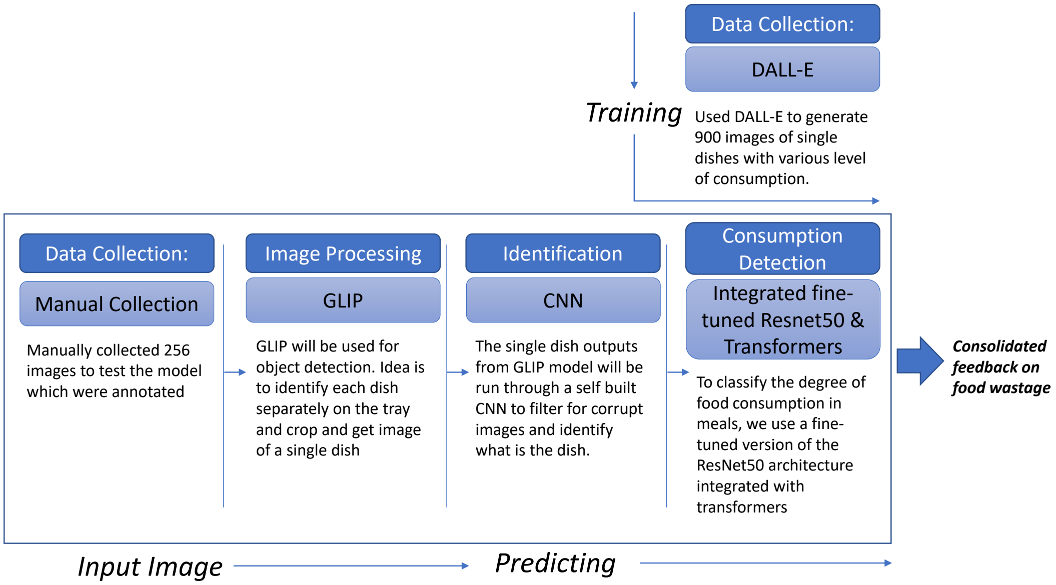

Data Collection:
DALL-E
Training
Used DALL-E to generate 900 images of single dishes with various level of consumption.
Data Collection:
Manual Collection
Manually collected 256 images to test the model which were annotated
Image Processing
GLIP
GLIP will be used for object detection. Idea is to identify each dish separately on the tray and crop and get image of a single dish
Identification
CNN
The single dish outputs from GLIP model will be run through a self built CNN to filter for corrupt images and identify what is the dish.
Consumption Detection
Integrated fine-tuned Resnet50 & Transformers
To classify the degree of food consumption in meals, we use a fine-tuned version of the ResNet50 architecture integrated with transformers
Consolidated feedback on food wastage
Input Image
Predicting

Figure 5. Here we show a figure detailing our model components. The input for prediction will be an image of tray after consumption of food and output will be food wastage by dish.

tain the training dataset for this model, we are planning on creating the dataset at home and capture the images through our mobile.

Fine-tuned ResNet50 Model with Transformers for Consumption Detection: To classify the degree of food consumption in meals, we use a fine-tuned version of the ResNet50 architecture integrated with transformers [7]. This hybrid model combines the strengths of both CNNs and transformers, resulting in a powerful and accurate system for classifying food consumption levels. We train the model on a dataset of labelled meal images, covering a wide range of consumption levels. The fine-tuned ResNet50 model with transformers is then able to classify the degree of food consumption in unseen meal images. At the moment we have classified single dish uncorrupted images using ResNet50 and gained an accuracy of 77.8% on 15 images. We have observed that the model is performing well even on unseen data such as Noodles or different types of plates.

Parameters used: Restnet50: batch size = 64, nepochs = 10, learningRate = 1e-4 and transformers: batch size = 32, n epochs = 10, learningRate = 1e-5

4. Experiments & Results:

When we started experimenting the model, we have come across quite a few huddles:

- Incorrect bounding boxes are generated by GLIP when the dish is partially eaten. For example: As shown in figure:6, intention was to detect the whole serving dish but GLIP is capturing a cropped image and when we use this image to detect the food wastage, model is obviously generating incorrect category.
  Temporary fix: instead of giving prompt as Rice/Noodles and bread , we are giving prompt as box and flat plate. This is a temporary fix as the shape of the dish changes or having multiple dishes of same shape. Long term solution is to try another model to object detection.

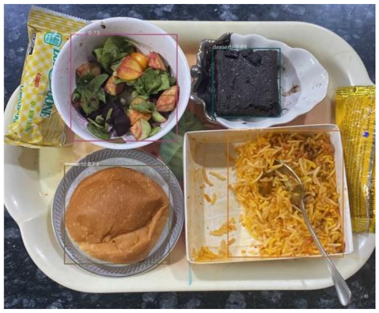


Figure 6. Images shows incorrect bounding box generated for Rice by GLIP

- GLIP not capturing empty and corrupt bowls. This basically means if the food is completely eaten we will never get a feedback on it which is a major drawback.

Solution, is to try another model which is able to capture even if the dish is empty.
Solution: We tried to shift to YOLOv7 [9] for object detection. Without fine tuning the model, we were ending up with objects being detected as bowls (example: Figure 7). So, we had to fine tune YOLOv5, for which we annotated images using VGG [5][6]
We created 30 test images and annotated them using VGG and fine tuned YOLOv7. Unfortunately, precision and recall were very poor and we believe it is because data set for fine tuning is very small for all four categories. Please check figure 8 for the new architecture is the model.

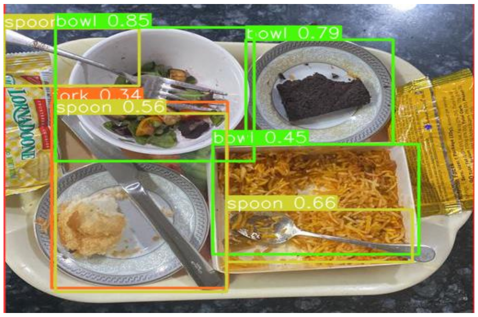
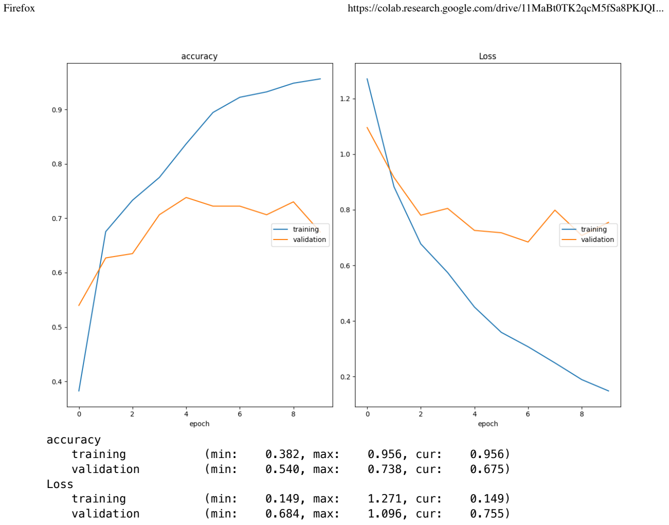


Figure 7. sample of image being annotated using YOLOv5

- If the containers are placed over each other the bottom container is not captured and will be disregarded. We are fine to take this approach for now as detection of food wastage is notible without having a view of the dish.
- Since GLIP was not detecting corrupt images, we did not need our own CNN model to detect them. So we removed it from the architecture.

In-conclusion, we could'nt stick to our original plan of using GLIP for object detection. When we tried to use YOLO and fine tune it to detect salad / Rice/ bread and dessert the results were not impressive. So, we ended up using Faster-RCNN which is very good at predicting bowls. Drawback for this model is that it wont be able to differentiate between the dishes which is not very helpful measure accuracy of food wastage. But we had to choose best option we had, so we went ahead with Faster-RCNN. Please check figure 9 for the final model architecture.

Results: To understand how the model is performing , we ran accuracy score only for Resnet50 in isolation by using approximately 20% of the images generate from DALLE. We achieve accuracy of around 75% for predicting the correct class out of the four. Please check figure 8: for training /validation accuracy.



```
accuracy
    training           (min:    0.382, max:    0.956, cur:    0.956)
    validation         (min:    0.540, max:    0.738, cur:    0.675)
Loss
    training           (min:    0.149, max:    1.271, cur:    0.149)
    validation         (min:    0.684, max:    1.096, cur:    0.755)
```

Figure 8. training and validation accuracy for integrated Resnet50 and transformer model

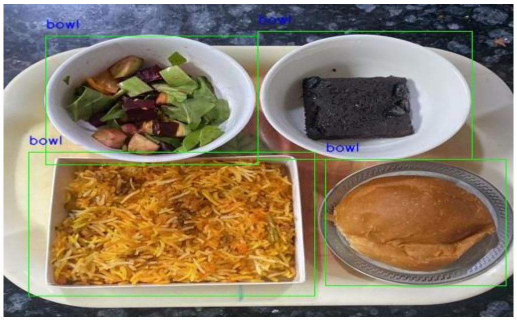


Figure 9. sample of image being annotated using Faster-RCNN

5. Future Goal:

We would like to generalize the model to capture various shapes and types of dishes. Model should be able to differentiate between corrupt and empty dishes and on top of it we would want to make this model run on an edge device and also increase the scope of detecting more dishes. Aim is to improve the accuracy of the model by collecting more the data and test out other modelling options. Ideally would like to partner with a restaurant or a small airline and start the process of data collection once we manage to make the model run on edge device.

6. Learnings & Conclusions:

- Data collection has been the key for the whole project. Apart from just collecting the data, pre-processing them was very crucial.
- Small sample of images (600) were sufficient to fine tune resnet50 model to a reasonably good extent
- We have faced unexpected hurdles which we fixed with

temporary solutions but to solve them holistically, we would need bigger data set

- In conclusion, the results we achieved are not overwhelmingly great as the model is not generalized enough. We believe with more data set (i.e annotated images) we should be able to build a more generalized accurate model.

References

[1] `https://lumitics.com/`.

[2] `https://simpleflying.com/airlines-4-billion-untouched-food-drink/#:~:text=The%20International%20Air%20Transport%20Association,a%20value%20of%20%244%20billion`, 2022.

[3] `https://aviationweek.com/air-transport/interiors-connectivity/tracking-inflight-catering-can-reduce-foodwaste#:~:text=Airbus'%20Food%20Scanner%20tracks%20onboard%20food%20and%20beverage%20consumption.&text=To%20help%20support%20airlines'%20goals,tracks%20and%20controls%20inflight%20catering`, 2023.

[4] `https://simpleflying.com/lufthansa-using-ai-to-reduce-food-waste/`, 2023.

[5] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: X.Y.Z, Accessed: INSERT$_{D}ATE_{H}ERE$.

[6] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. `doi:10.1145/3343031.3350535`.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[8] Harold Liunian Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2020.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2016.

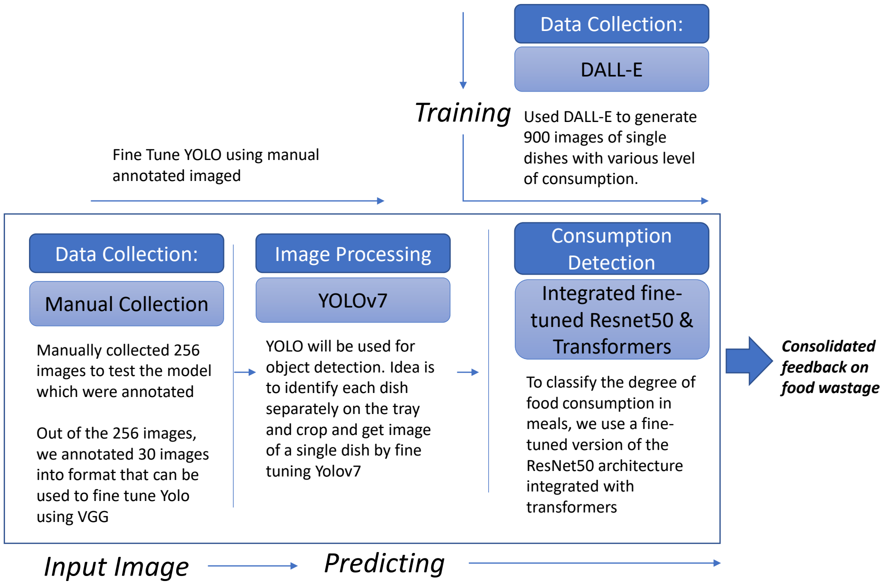

Data Collection:
DALL-E
Training
Used DALL-E to generate 900 images of single dishes with various level of consumption.
Fine Tune YOLO using manual annotated imaged
Data Collection:
Manual Collection
Manually collected 256 images to test the model which were annotated
Out of the 256 images, we annotated 30 images into format that can be used to fine tune Yolo using VGG
Image Processing
YOLOv7
YOLO will be used for object detection. Idea is to identify each dish separately on the tray and crop and get image of a single dish by fine tuning Yolov7
Consumption Detection
Integrated fine-tuned Resnet50 & Transformers
To classify the degree of food consumption in meals, we use a fine-tuned version of the ResNet50 architecture integrated with transformers
Consolidated feedback on food wastage
Input Image
Predicting

Figure 10. Here we show a figure detailing how we integrated YOLO fine tuned model into the over all architecture

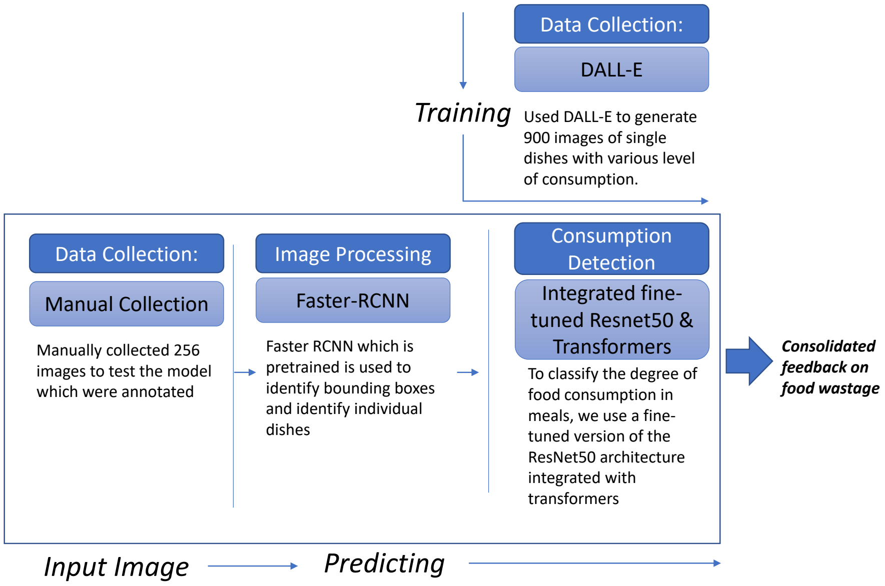

Data Collection:
DALL-E
Training
Used DALL-E to generate 900 images of single dishes with various level of consumption.
Data Collection:
Manual Collection
Manually collected 256 images to test the model which were annotated
Image Processing
Faster-RCNN
Faster RCNN which is pretrained is used to identify bounding boxes and identify individual dishes
Consumption Detection
Integrated fine-tuned Resnet50 & Transformers
To classify the degree of food consumption in meals, we use a fine-tuned version of the ResNet50 architecture integrated with transformers
Consolidated feedback on food wastage
Input Image
Predicting

Figure 11. Here we show a figure detailing final architecture of the model