A deep learning approach to classifying images of packaged versus unpackaged food

Shubham Bipin Kumar
Indiana University
Bloomington
sbipink@iu.edu

Himanshu Hansaria
Indiana University
Bloomington
hhansar@iu.edu

Abhinav Sinha
Indiana University
Bloomington
sinhabhi@iu.edu

Abstract

Taking images of food is quickly replacing written food diaries as the primary means of food logging for people who want to monitor their food choices. Automatically analyzing the content of these food images can provide users and health experts with invaluable information about a user's eating and possibly help plan diet charts, diagnose diet-related medical conditions, or plan treatments for existing conditions. There are several challenges with image-based food recognition – estimating the volume of food, recognizing occluded ingredients, etc. Food images from these logs often contain a mix of packaged and unpackaged food, which have fundamentally different structures. It adds further complexity to an already difficult recognition task. One way to simplify this problem is to classify images into two categories (packaged and unpackaged) during the preprocessing step and then feed these sets of images into separate models for further processing. In this paper, we tackle the problem of segregating packaged and unpackaged food images.

1. Introduction

The remarkable success of deep learning techniques in object detection combined with a surging need for food logging and dietary assessments presents Computer Vision researchers with the problem of identifying meaningful information from these food image logs. But even with advancements in image recognition technology and the abundance of food images on the internet, automatic food recognition remains a challenging problem. There are several reasons – 1. Contents of food dishes are objects with complex semantics with an undefined structure, therefore it's difficult for any model to learn features to recognize them. 2. Lack of universal datasets – foods popular in one region of the world may be unheard of in another region. This causes existing datasets to be very biased – leading to suboptimal model performance on real data. 3. Lack of labels is another problem. Even though there are millions of images of food on the internet, most of them lack labels, and even if they do – they are not distinctive. 4. Similar image structure but different nutritional content – Mayonnaise looks identical to Greek yogurt, but their nutritional content is very different. 5. Occlusion – it's very difficult to tell what a sandwich contains inside it just by looking at it from the outside. 6. Estimating portion size – estimating the volume of food is very difficult due to the lack of depth information in 2D images.

In this paper, we tackle a subproblem of the larger problem – classifying images of packaged vs unpackaged food. We start our project by exploring why this classification problem is important. In this process, we examine the existing work by Chung et al. that forms the foundation of our project.

We then investigate existing image recognition architectures that could be used for our use case. We then start with our implementation and experiment with various hyperparameter values. Next, we study our results by using saliency maps. We also visualize the misclassified and correctly classified images. We then discuss our results and conclude with our findings and future work.

2. Background and related work

A powerful idea highlighted by Chung et al. is that nutritional values (number of calories/estimates of nutrients) and sizes of food portions are not sufficient to predict the "healthiness" of a diet. For example, a portion of nuts could be more calorie-dense than a scoop of ice cream, but that does not mean it is unhealthy. Alternatively, consider a portion of plain white rice vs. a portion of mixed fruit -> both the portion sizes may be equal – but the health benefits of a plate of fresh fruit may outweigh that of a bowl of rice in most cases.

One potential candidate to predict the "healthiness" of food plates is the color variety [2]. The more colorful a plate of food is, the healthier it's likely to be. So, examining the variety of colors in the food images of a user could be used to predict whether their diet is healthy or not. However, the packaging in packaged foods adds to the color variations in the image even if the food item inside the package is not colorful. [3] Another challenge with

packaged food images is that it's very difficult to estimate the portion size by just examining its features. [3]

Here is where our classifier comes in. Suppose we split up packaged and unpackaged food. In that case, we can then run separate processing tasks on them to extract meaningful features for dietary assessments.

3. Methods

As we understand from the literature available on the image classification and identification as well our own experimentation with logistic regression and SVM models, we came to conclusion the deep learning model perform the best .

There are existing model like AlexNet and REsNet which works very well on image classification recognition task . AlexNet uses 5 convolutions layers, 3 max-pooling layers , 2 normalization layers , 2 fully connected layer and 1 softmax layer whereas ResNet uses double and triple layer skips that contain nonlinearities(Relu) and batch normalizations in between.

We decided to use transfer learning approach first and see how it performs on our dataset . We have used data augmentation and normalization on the dataset before training the model. AlexNet gave us a validation accuracy of 98% and test accuracy as 97.4%. Further ResNet as it is more complex model gave as even better validation accuracy of close to 99% which was expected in comparison to AlexNet.

Since these model have lot of parameters and have uses lot of resources we decided to further proceed with simple model and try various computer vision techniques to see how it performs in comparison with these models and if it give us an accuracy close to these models.

We started with a baseline model with3 blocks of sequential conv2d–maxpool layers with 64, 128 and 256 channels respectively. Conv2d kernel size: 3x3, and MaxPool: 2x2. No. of channels 64, 128, 256 respectively. This is followed by 2 fully connected layers with 128 and 256 neurons respectively . It gave us test accuracy of 80.22 %. This is a generic CNN model and widely used as baseline model.

We felt that we can further increase the number of neurons in both conv and fully connected layer as well as add dropouts to prevent the model from overfitting by reducing the capacity or thinning the network during training.

we used 3 blocks of sequential conv2d–maxpool layers with 64, 128 and 256 channels respectively. Conv2d kernel size: 3x3, MaxPool: 2x2. No. of channels 64, 128, 256 .This is followed by 2 fully connected layers with 512 and 1024 neurons respectively . We further added dropout of 0.5 for all the 3 convolutional layer (having experimented with 0.3 ,0.4) as it gave the best result.We ran this for 5 epochs and got and accuracy of 91.5%.

We used 'relu' as the activation because it's a non saturating activation function and we we resized all images to 128x128 and rescaled to pixel values between 0 and 1 since it's leads to better numerical stability and better convergence of the neural network.

We further number of conv layers and number of epochs . In addition to existing 3 conv layers we added one more conv layer after 3 existing conv layer with 512 neurons and increased the epochs to 10 as the later conv layer learns higher level features it gave us an accuracy of 96.5%. We further increased the accuracy and used early stopping as well to get better generalization and avoid overfitting . We Realized that around 20 epochs is best and we achieved an final test accuracy of 96.7 % which is comparable to the test accuracy of the AlexNet with transfer learning but with much simpler architecture. Results included in Table 1. Model architecture in Figure 3.

4. Results

From the training, validation and test accuracies using the method of early stopping, we found that the model gave best performance for training after 16 epochs. We confirmed that the model is not overfitting since the test accuracies were very similar i.e. 96.70% in comparison to the validation accuracy of 97.50%.

Figure 1: Line1 – correctly classified, line 2: incorrectly classified

Based on the correct and misclassified images as in the above figure, it seems that blurry and occluded images results in misclassification and different variations of correctly identified images are visible even in transparent packages even though It may sometimes be a challenge in the coming analysis

Figure 1: Image ‘a’ on the left, Image ‘b’ on the right

Figure 1: Feature maps image ‘a’

Figure 1: Feature maps image ‘b’

As, we can observe from the feature maps in the figure, that the shape and the text has the neurons firing while correctly classifying the image, shows us that the textual and shape features contributes to the image being classified as packaged food.

Similarly, we can observe that when the feature maps, don’t give significant firing in the regions concerning the text, and identify the texture and shape as a whole, it may lead to incorrect classification. Also, there is an ambiguity about the definition of the image as a packaged/ unpackaged food as the food is in a transparent packaging, which might also be a challenge for it’s classification as packaged or unpackaged.

5. Discussion

Our research question was how the neural networks is able to classify packaged versus unpackaged food, and based on the analysis of the results, we were able to significantly answer the research question based on the activation maps of the convolutional layer of the neural networks.

6. Conclusion

The study aimed at identifying separating packaged food from unpackaged food in the images, and we were able to bring up a model to correctly classify images as packaged/unpackaged food, thereby helping real-world images in the area of food logging and classification pipeline to further analyze the packaged food separately.

We plan to further integrate this work with the bigger part of the food identification process, and further identify the contents of the packaged food as well.

We learned a lot during the preparation of the project and would like to thank our professor David Crandall and Weslie Khoo for the constant support and guidance during the entire process.

Model Params	Training accuracy	Validation accuracy	Test accuracy
3 conv, 2 fully, 5 epochs	85.5%	82.3%	80.22%
3 conv, 2 fully, 5 epochs with dropout and increased neurons	88.87%	89.2%	87.2%
4 conv, 2 fully, 10 epochs	90.73%	89.9%	91.1%
4 conv, 2 fully, 50 epochs, early stopping	98.43%	96.30%	96.70%
Transfer learning with AlexNet	99.13%	98.4%	97.40%

Table 1

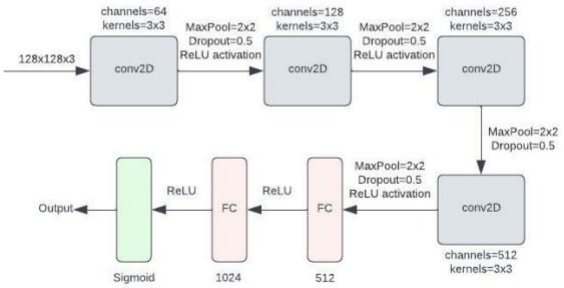

Figure 3

References

[1] Chung, C., Ramos, A., Chiang, P., Wu, C., Tan, C.A., Khoo, W., & Crandall, D.J. (2021). Computer Vision for Dietary Assessment.

[2] König, L.M., Renner, B. Boosting healthy food choices by meal colour variety: results from two experiments and a just-in-time Ecological Momentary Intervention. BMC Public Health 19, 975 (2019). https://doi.org/10.1186/s12889-019-7306-z