A Computer Vision Solution to Cross-cultural Food Image Classification and Nutrition Logging

Rohan Sethi
rsethi1@luc.edu

George K. Thiruvathukal
Loyola University Chicago, gkt@cs.luc.edu

Follow this and additional works at: https://ecommons.luc.edu/cs_facpubs

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution 4.0 International License.
© The Authors, 2024.
Abstract
The US is a culturally and ethnically diverse country, and with this diversity comes a myriad of cuisines and eating habits that expand well beyond that of western culture. Each of these meals have their own good and bad effects when it comes to the nutritional value and its potential impact on human health [1]. Thus, there is a greater need for people to be able to access the nutritional profile of their diverse daily meals and better manage their health [1]. A revolutionary solution to democratize food image classification and nutritional logging is using deep learning to extract that information from analyzing images a user inputs. However, current computer vision (a subspecialty of deep learning) applications that are used to classify foods are limited by the western-biased datasets they are trained on [2]. Additionally, a diverse image dataset for training computational models for classification is not plausible as there are just too many cuisines for any model to [1]. Clearly there is a need for a pipeline that can learn to predict new categories of foods continuously. In this project, we propose to design an adaptable prototype pipeline using hierarchical neural networks which can classify international food images the model has rarely or never seen.

Research Questions
- Would image, text, or a combination of both data modalities aide in prediction of new cuisines?
- Would a hierarchical classification network or vanilla feedforward network prove to be better at predicting foods?
  - Should categories in the hierarchy be fixed or determined in an unsupervised fashion or a combination of both?

Methods and Design
The dataset contained 4000+ images and further information about the recipes documented. 80-10-10 data split applied. We extracted the images, cuisine, diet, course, and instructions features. Below are the steps we have taken so far:
- To address our first question, we first needed to convert the food images and the instructions into vectors. For this we utilized pretrained food image feature extractor from TensorFlow Hub and word2vec to convert images/text into numbers [6][7].
- We then trained 3 inputs: only instructions, only images, sum of both
- We then reduced the dimensionality of the features. This was to ensure that the sheer size of features does not prevent recipe correlations. We visually analyzed 3 principal components
- Finally, we clustered the data using KMeans clustering (n=56). The number we chose the max number of categories that the recipes belong, however this is arbitrary
- We also applied the pre-labeling of diet, cuisine, and course to see how the feature vector predictions compare to real predictions of recipe clustering

Preliminary Results

Discussion/Next Steps
It is clear to see from the results that the image data demonstrates no clear separation of recipes, while text data demonstrates a few clusters of categories. We believe that combining both data modalities provides a balance of having some categorical separation and room for hierarchical separation beyond the first.

Additionally, we don’t think all hierarchical-steps in the prediction framework should be fixed since the results show how difficult it is to see the extracted features matching the dataset applied labels. Rather having one layer fixed and the rest automatically generated allows for both flexibility and adaptability for new entries.

Along these lines, we plan to retrain image feature vector to predict cuisines. We believe the prediction framework would benefit from dividing the recipes into multiple categories (56 in this case) and within each of those categories have another set of networks to separate into cluster-based categories. This would allow for ease of adding in new recipes as only a subset of networks would need to be retrained to accommodate new food items and require less image/text data to train on since there would be less parameters to fit.

References
6. https://tful.google/imagenet/mobilenet_v1_100_224/feature_vector5
7. https://tful.google/seefood/segmenter/mobile_food_segmenter_V1/1

For these experiments, we originally hoped to run a crowd sourcing study to collect real time different food cuisines that residents of the US eat, which would require some form of approval/review. However, due to time constraints we found a precompiled Indian food dataset with over 50+ different cuisines and utilized this dataset for initial results [5].