CHEFNET: IMAGE CAPTIONING AND RECIPE MATCHING ON FOOD IMAGE DATASET WITH DEEP LEARNING

Kaylie Zhu, Harry Sha & Chenlin Meng

{kayliez, harry2, chenlin}@cs.stanford.edu

Introduction

Food is central to human life, but finding nutritious and satisfying food is not easy. In this context, we introduce Chefnet, an extension to the im2recipe model, which matches food images to their recipes. Modifications include incorporating newer and deeper CNN models - Densenet121, and Resnet101.

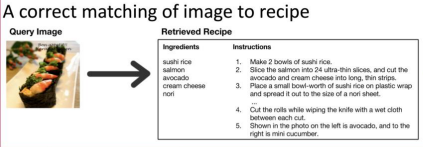
Data

We use the Recipe1M dataset for training and validation. Exact duplicates and recipes that share the same images were manually removed, as were instances with unwanted characters or without discriminative food properties. We investigate model performance with the reduced training and validation sizes of 20,000 and 2,000 in alignment with our computational resources (Figure 4).

Illustration & Demonstration

Example 1:

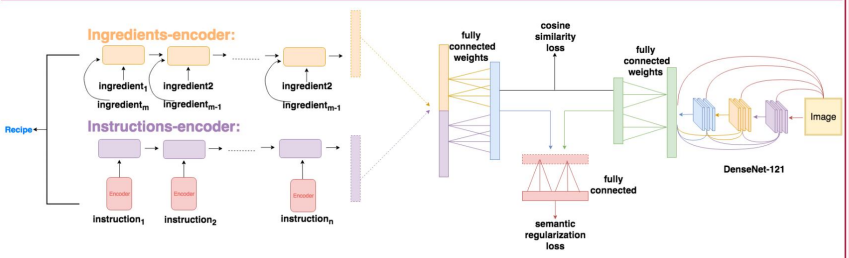
A correct matching of image to recipe



Example 2:

An incorrect matching of image to recipe

Model



We obtained recipe embeddings by concatenating the cooking instruction encodings and the recipe list encodings. The former was determined by an LSTM, and the latter by a bidirectional LSTM due to its inherent unordered nature. We used a CNN for the image embeddings. In particular, we experimented with Densenet121, Resnet50 and Resnet101 models. The image and recipe embeddings were then projected onto the same embedding space using fully connected layers. Our goal is to maximize the cosine similarity between the embeddings of matching food-recipe pairs, and to minimize that between the encodings of non-matching ones.

Procedure

Given the embeddings, we obtain class probabilities followed by a softmax activation. While learning the model, we first fix the weights of the image network and learn the recipe encodings. Then we freeze the recipe encodings, and learn the image network.

Discussion and Results

We compare the performance (Figure 1) and computation costs (Figure 2) of using Densenet121, Resnet101, and Resnet50 models, each with their respective optimal learning rates, which we attain through experiments (Figure 3). We find that DenseNet achieves the lowest cosine loss at a negligibly higher computation cost compared to Resnet50. Thus, when training for longer on the full data set, we expect our Densenet model to perform better relative to the ResNet-50, which is used in the original paper.

Figure 1

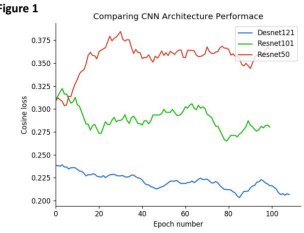


Figure 3

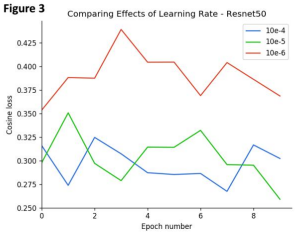


Figure 2

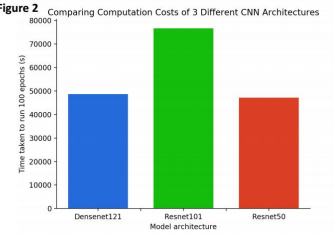
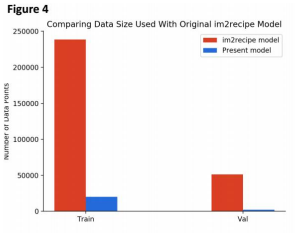


Figure 4



Reference

[1]Salvador, Amaia, Hynes, Nicholas, Aytar, Yusuf, Marin, Javier, Ofli, Ferda, Weber, Ingmar, and Torralba, Antonio. Learning cross-modal embeddings for cooking recipes and food images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017

[2]Kaiming He, Xiangyu Zhang, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. 2015.

[3]Chang Liu, Yu Cao, Yan Luo, Guanling Chen, and Vinod Vokkarane, Yunsheng. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. 2016. doi: arXiv:1606. 05675.

[4]K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015

[5]K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 1, 3