Calorie estimation from food images

Kaixi Ruan and Lin Shao
kaixi@stanford.edu, lins2@stanford.edu

Abstract

The project consists of two steps: identifying food from an image, and convert the food identification into a calorie estimation. We perform food image classification using SVM and deep learning algorithms. Different features such as LBP, HOG, and CNN) are explored and compared. For the calorie estimation step, we create a "calorie map" for the image classification labels.

Introduction

This project is motivated by calorie-tracker applications. The idea is, instead of entering meal details to track the daily consumption, it would be easier to ask users to take a photo of their meal, and from that photo to give a calorie estimation.
For the scope of this project, we ignore the scale and quantity of food in an image and treat it as an image classification problem and we then convert the predicted labels to calorie estimations using online food database.
We mainly use SVM and CNN models in our implementation. Different features used in image processing are used and compared to the CNN-SVM model.

Dataset

We mainly used the Pittsburgh Fast-food Image Dataset (PFID): we choose 1359 fast food images taken in laboratory or restaurants. We labeled the data set with 10 general categories: burrito, salad, donuts, bread sandwich, pie, burger, toast sandwich, chicken, pizza and bread.

Figure 1: Ten categories we labeled for PFID data set

Feature extraction

In order to run the SVM algorithm, we need to extract features from the data set. Here, we explore some of the classical features used in image recognition.

- Color Histogram: represents the distribution of colors in an image.
- Histogram of Oriented Gradients (HOG): the technique counts occurrences of gradient orientation in localized portions of an image, which captures shape information. (Figure 1)

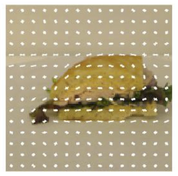

Figure 2: HOG feature

- Local Binary Pattern (LBP): LBP gives a histogram of the frequency of each higher illuminance for each patch (Figure 1). It emphasizes texture information within each patch.

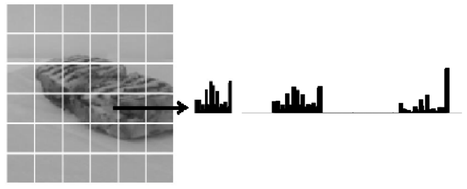

Figure 3: LBP feature

SVM results on HOG and LBP

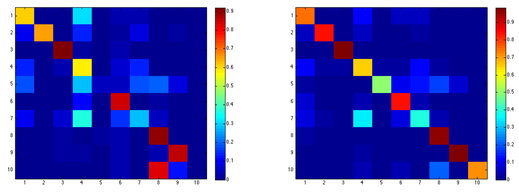

Figure 4: Confusion Matrix for HOG (left) and LBP(right)

CNN-SVM

Given the small dataset, we choose to fine-tune the CNN model of BVLC Reference CaffeNet. The architecture of CNN model is shown below.

Layers	Type
fc7 relu7 drop7	InnerProduct ReLU Dropout
fc6 relu6 drop6	InnerProduct ReLU Dropout
conv5 relu5 pool5	Convolution ReLU Pooling
conv4 relu4	Convolution ReLU
conv3 relu3	Convolution ReLU
conv2 relu2 pool2 norm2	Convolution ReLU Pooling LRN
conv1 relu1 pool1 norm1	convolution ReLU Pooling LRN

Table 1: Architecture of CNN model

We fixed the first fifth layers and fine tune on the last two layers. We choose the model after 3000 iterations. The learning accuracy curve on Validate Set is shown below.

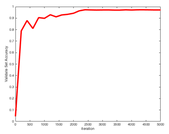

Figure 5: Accuracy Curve

To combine CNN and SVM, we extract the sixth layer from the model as the input features in SVM. The result (confusion matrix on the test set) is shown below.

Confusion matrix for CNN-SVM

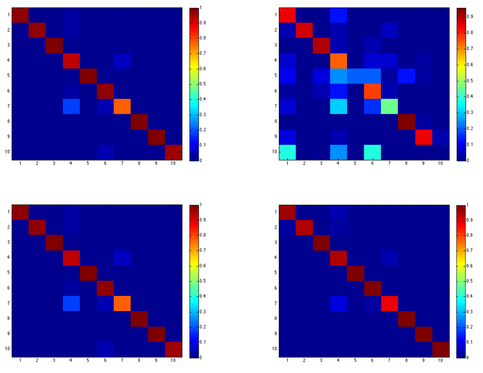

Figure 6: Confusion Matrix using CNN features: fc6 no fine-tuning(top left) fc7 no fine-tuning(top right) fc6 fine-tuning (down left) fc7 fine-tuning (down right)

Conclusion

Comparing the confusion matrices generated using different features, we observe that

- HOG feature can not distinguish between different food with similar shapes, such as chicken and bread, while these categories can be well separated base on their different textures.
- Compared to the standard image classification features, the CNN features result in a net improvement in prediction. Although features like HOG and LBP provide local information about the image, they assume fixed patch size thus might not capture all the information available; CNN deals with this problem by providing a more flexible framework.

References

[1] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 289–292, Nov 2009.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

Acknowledgements

We'd like to thank Intel Labs Pittsburgh for the publication of the food image data set. We also want to thank all CS229 staff for their support during the quarter.