Deep Learning Based Food Recognition

Dongyuan Mao, Qian Yu, Jingfan Wang

Motivation

Food is the cornerstone of people's life. Nowadays more and more people cares about the dietary intake since unhealthy diet leads to numerous diseases, like obesity and diabetes. Accurately labelling food items is significantly essential to keep fit and live a healthy life. However, currently referring to nutrition experts or Amazon Mechanical Turk is the only way to recognize the food items.

In this project, we propose a deep learning based food image recognition algorithms to improve the accuracy of dietary assessment. We applied convolutional neural network to solve this problem. Especially GoogLeNet Inception V3 and GoogLeNet Inception-ResNet were performed for classifying the food images on the slim version of Tensorflow. Optimizer like RMSprop or Adamare also tried to optimize the model.

In this project, we will work to solve two key issues:

1. What are the methods to improve the accuracy of GoogLeNet Inception V3 and GoogLeNet Inception-ResNet? How accurate they can finally achieve?
2. How do these two GoogLeNet based methods compared with other methods like VGG and SVM by means of Top 1 accuracy and Top 5 accuracy?

Dataset and Methods

Dataset

Deep learning-based algorithms requires large dataset. We decided to use the novel and challenging dataset called ETHZ-FOOD-101, which consists of 101 food categories with 101,000 images.

Image Processing

- Reason: The environmental background varies a lot in different food pictures. Those environmental factors are the color temperatures, luminance and so on.
- Method: Grey World method, and Histogram equalization.

Figure 1. Image processing: (left) raw image, (middle) perform Grey World method, (right) perform Histogram Equalization on the middle image.

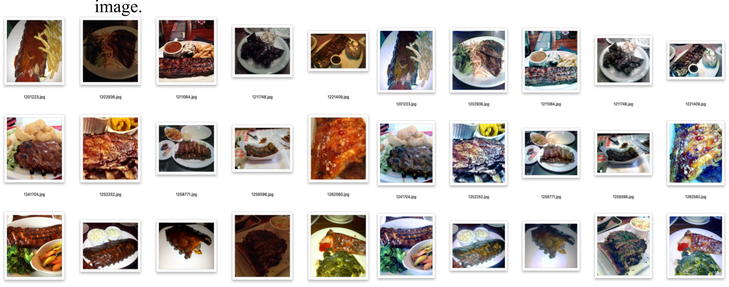

Figure 2. Image processing: (left) raw images, (right) processed images

Training processes

Transfer learning: we can make full use of the pre-trained model and get the feature based on the ImageNet dataset.
Algorithms: GoogLeNet Inception V3 and GoogLeNet Inception-ResNet

Last layer training → (1) Use the pre-trained model. (2) Change the last layer of the network and recognize classes of food images. The dimension of the last layer is 2048*101 instead of 2048*1001.

Full layer training → (3) Train with raw images and processed images. (4) Evaluate the results. 80% of the data is training data; while 20% of the data is testing data. (5) Fine tuning the network, including tuning learning rate, learning rate decay, weight decay, performing different Optimizer: Adam and RMSprop

Methods and Results

Calculation setup

- GPU: Amazon AWS GPU. The AWS g2 instance: NVIDIA GRID K340 with 1536 CUDA cores and 4GB memory size.
- The framework for deep learning: the latest slim version of Tensorflow.

Algorithm

- **GoogLeNet or Inception V1**
  - ✓ The Inception deep convolutional architecture was introduced, with the advantages of less parameters (4M, compared to AlexNet with 60M).
  - ✓ Average Pooling instead of Fully Connected layers at the top of the ConvNet was applied to eliminate unnecessary parameters。
- **GoogLeNet Inception V3**:
  - ✓ Improved inception V1 by additional factorization ideas in the third iteration
- **GoogLeNet Inception-ResNet**:
  - ✓ Designed Inception-ResNet to make full use of residual connections.
  - ✓ Training with residual connections accelerates the training of Inception networks significantly, by utilizing additive merging of signals.

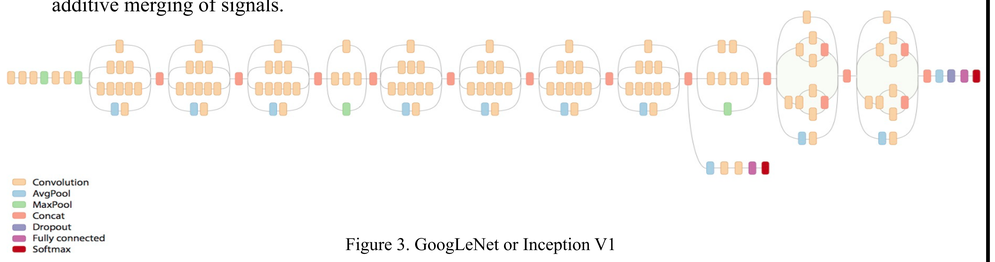

Figure 3. GoogLeNet or Inception V1

Results

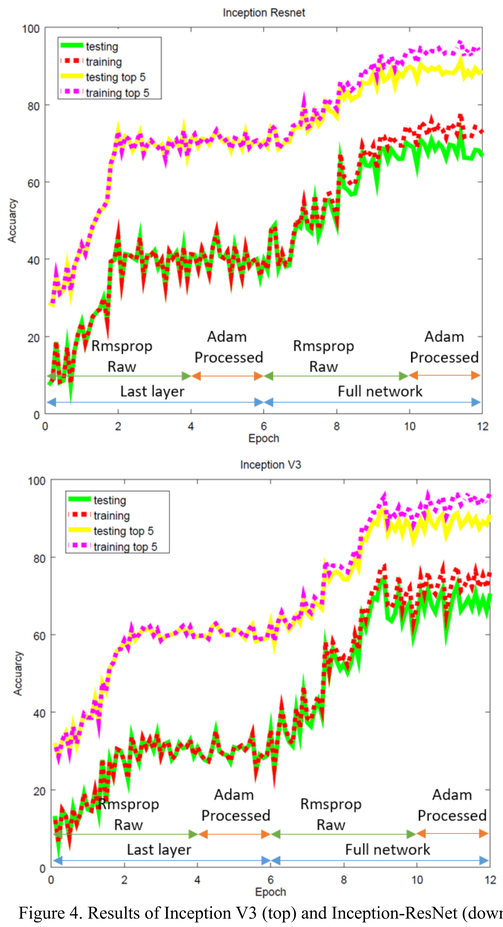

Figure 4. Results of Inception V3 (top) and Inception-ResNet (down)

The following questions can be answered by Figure 4.

- How is training on the last layer different from training on the full layers?
- How do the two optimizers affect differently?
- Which deep learning method has better performance?
- Is there any over-fitting problem?
- How is the effect of image processing?
- How the loss change over epochs?

Table 1. Accuracy comparison

Method	Top 1 Accuracy	Top 5 Accuracy
SVM from Bossard, L et.al	50.76%	NA
CNN-based Approach from Lukas et. al	56.40%	NA
RFDC-based Approach from Lukas et. al	50.76%	NA
Inception V3 (last layer training)	35.32%	62.97%
Inception-ResNet (last layer training)	42.69%	72.78%
Inception V3 (full layer training)	70.60%	90.91%
Inception-ResNet (full layer training)	**72.55%**	**91.31%**

Discussion & Conclusion

- ✓ Inception-ResNet (full layer training) with Top 1 accuracy of 72.55% and Top 5 accuracy of 91.31% achieves the best accuracy compared with the methods in this project or papers.
- ✓ Training on the full layers can boost the accuracy compared with training only on the last layer.
- ✓ When only retraining the softmax layer of the network, Inception-ResNet has a better performance than Inception V3. When retraining all network layers, Inception V3 and Inception ResNet has similar performance.
- ✓ Processing image improves the accuracy by about 3%.
- ✓ Overall, RMSprop and Adam perform similarly. For each process, we first use RMSprop and then Adam, because loss under RMSprop drops much quicker at the beginning and Adam is more likely to converge on the global minimum at the end with less jitter.

Future Work

- ✓ Explore which features affect the classification accuracy the most and present the most of the food.
- ✓ Design new architecture based on the two architecture we are using, including adding dropout layer, performing new max/pool and convolution.
- ✓ Instead of applying pre-trained model, try to train the whole layers of the new architecture by ourselves.
- ✓ Try other architecture in order to get full understanding of architecture comparison and why architectures perform differently.
- ✓ Try adding bounding box on the processed images.

References

[1] Martin, C., Correa, J., Han, H., Allen, H., Rood, J., Champagne, C., Gunturk, B., Bray, G.: Validity of the remote food photography method (RFPM) for estimating energy and nutrient intake in near real-time. Obesity (2011)
[2] Noronha, J., Hysen, E., Zhang, H., Gajos, K.Z.: Platemate: crowdsourcing nutritional analysis from food photographs. In: ACM Symposium on UI Software and Technology (2011)
[3] ETHZ-FOOD-101, https://www.vision.ee.ethz.ch/datasets_extra/food-101/
[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9).
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.
[6] Szegedy, C., Ioffe, S., & Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261.
[7] Bossard, L., Guillaumin, M., & Van Gool, L. (2014, September). Food-101–mining discriminative components with random forests. In European Conference on Computer Vision (pp. 446-461). Springer International Publishing.
[8] Cadène, R., Thome, N., & Cord, M. (2016). Master's Thesis: Deep Learning for Visual Recognition. arXiv preprint arXiv:1610.05567.
[9] Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., ... & Murphy, K. P. (2015). Im2Calories: towards an automated mobile vision food diary. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1233-1241).

Acknowledgements

In particular, we would like to thank Professor Andrew Ng, Professor John Duchi and Sheng Hao for their help and support throughout the project.