iFood Challenge @Rice University

Fine-grained Classification of Food Images

Antonio Lara (jx24), Yue Zhuo (yz154)

comp540_yz154_jx24, Rice University

{antonio.lara,yuezhuo}@rice.edu

Problem

Predict the Fine-Grained Food-Category Label given an image

- 251 fine-grained (prepared) food categories from FGVC6
- 118,475 training images with human verified labels
- 11,994 validation images with human verified labels
- 28,377 testing images

Exploratory Data Analysis

- **Clusters Patterns**

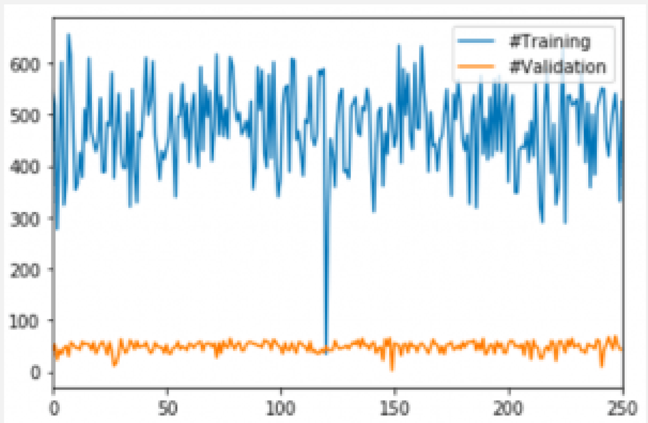

Figure 1:Number of Training and Validation Data by Class

- **Outliers**

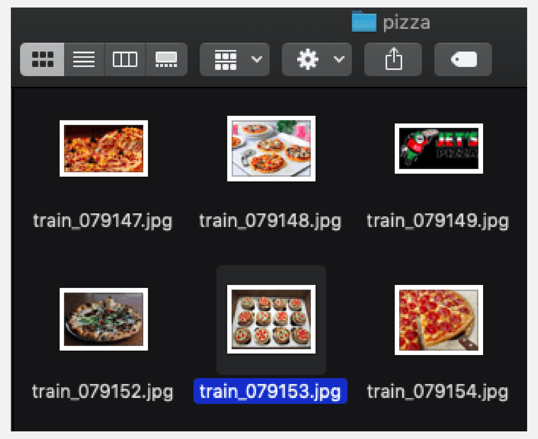

Figure 2:Wrong Classified Data

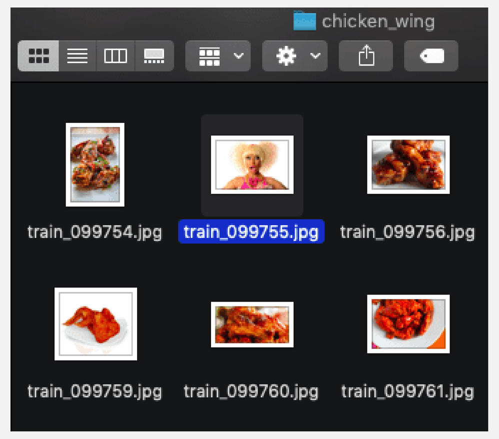

Figure 3:Data out of Sample Space

Data Pre-Processing

For better recognizing images and preventing overfitting:

- **Transformation** Randomly flipped, rotated, zoomed, etc.
- **Augmentation** Resized and normalized images into 224*224.

Figure 4:Image Data Examples after Pre-Processing

Challenges

- The classes are fine-grained and visually similar.
- The training images are crawled from the web, they often include images of raw ingredients or processed and packaged food items.

Model Architectures

- **ResNets** utilize skip connections to jump over some layers. Typical models are implemented with double- or triple- layer skips that contain nonlinearities (ReLU) and batch normalization in between.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 5:ResNet architecture.

- **Performances**

Res-	Net18	Net34	Net50	Net101	Net152	NeXt101
Top3 Acc	0.7123	0.7285	0.8302	0.8686	0.8680	0.8836

Ensemble Learning

Ensemble Learning combines the predictions from multiple models. It not only reduces the variance of predictions, but also can result in predictions that are better than any single model.

Here we ensemble 4 models with Top-3 accuracy higher than 0.8 (ResNet-50, ResNet-101, ResNet-152 and ResNeXt-101)

- **Simple Arithmetic Average**

$$finalpred = \frac{pred_1 + pred_2 + pred_3 + pred_4}{4}$$

Top-3 Acc: 0.8876

- **Weighted Arithmetic Average**
  - Take the inverse of RMSEs as weights

$$finalpred = \frac{\frac{1}{RMSE_1}pred_1 + \frac{1}{RMSE_2}pred_2 + \frac{1}{RMSE_3}pred_3 + \frac{1}{RMSE_4}pred_4}{\frac{1}{RMSE_1} + \frac{1}{RMSE_2} + \frac{1}{RMSE_3} + \frac{1}{RMSE_4}}$$

  Top-3 Acc: Top-3 accuracy 0.8923
  - *Manually tune weights based on validation*

  Top-3 Acc: Top-3 accuracy 0.9013

Submission Timeline

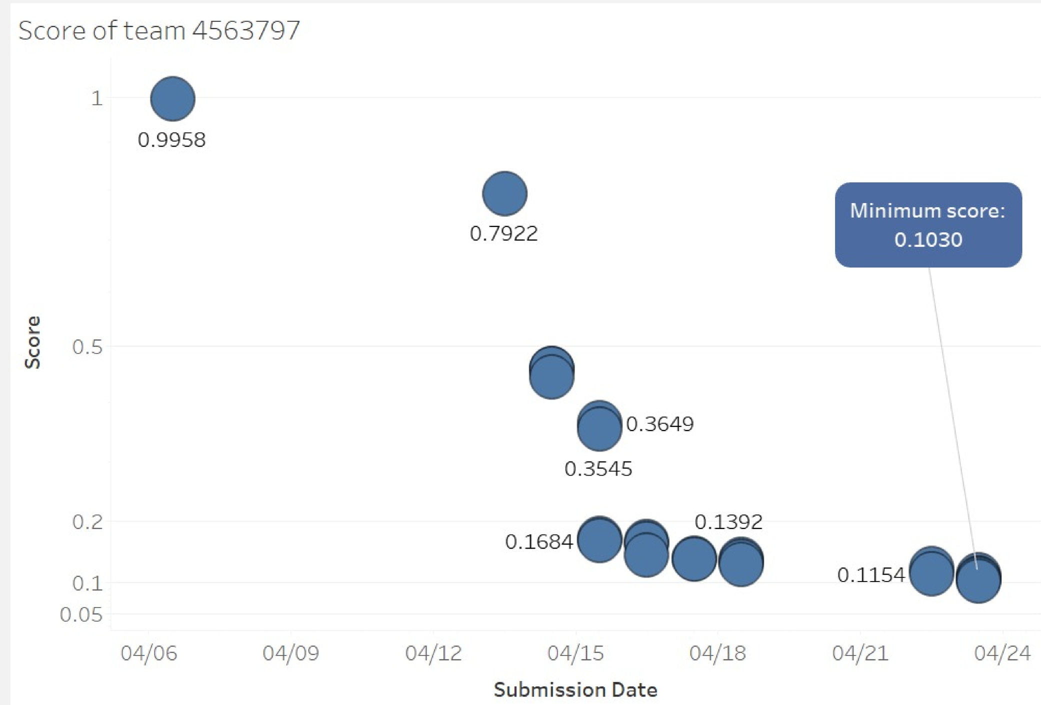

Figure 6:Kaggle Submission Timeline

Conclusions

- **Best Model** Ensemble of ResNet-50, ResNet-101, ResNet-152 and ResNeXt-101
- **Result** Score: 0.10301
- **Improvement**
  - Filter outliers in pre-processing to eliminate noise (K-means or downloading other food image datasets).
  - Better ensemble methods (bagging, boosting, etc.)

References

[1] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[2] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[4] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[9] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4696–4705, 2019.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference*