Food Image Classification with Convolutional Neural Networks

Malina Jiang (malinaj@stanford.edu) -- https://youtu.be/iJZPs5GE2Hs

Introduction

- Food images dominate across social media platforms, driving the restaurant and travel industries, but are still relatively unorganized.
- The ability to properly label / classify food images could lead to better recommendation systems (matching food based on an individual's tastes and preferences, or diet).
- Input is a food image, output is label prediction by CNN (whether trained for scratch or pretrained).

Data and Features

- Food-101[1]: Total of 101,000 images from 101 distinct classes of food, with 1000 images per class. Of these 750 are training images that may be noisy or even mislabeled; 250 are correctly labeled validation images.
- ImageNet[2]: Used only during transfer learning as pre-trained weights, not directly. Of 1000 classes, 10 are food-related.
- Images are color-normalized and augmented through scaling, rotation, flipping, etc.

Figure 1: Food-101 examples, from left to right: 'ramen', 'pizza', 'apple pie'

1-cheesecake
2-tiramisu
3-chocolate cake
4-chocolate mousse

Model

- Baseline model was a shallow CNN with filters of the same size, then a fully-connected layer.
- Transfer learning was performed with VGG16, ResNet50, and InceptionV3, with top layer removed and retrained on the 101 food classes. More layers were incrementally unfrozen to improve performance.

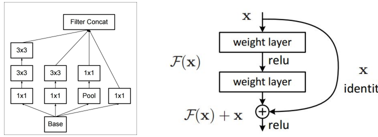
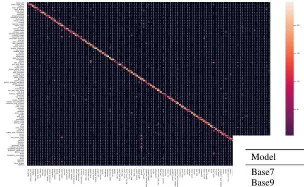


Figure 3: Inception module[4] and residual block[3]

- Loss function was categorical cross-entropy loss:

$$L(y, \hat{y}) = -\sum_{c=1}^{M} y * log(\hat{y})$$

Results



Left: Confusion matrix for InceptionV3 model

Right: Model results and details

Model	Res	Epochs	Layers	Params	Model Details	Top-1	Top-5
Base7	64	8	7	-	-	28.3	-
Base9	128	5	9	-	-	36.3	-
AlexNet	227	50	9	29M	-	25.7	-
AlexNet	227	50	9	25M	data augmentation, removed dropout	32.8	61.9
AlexNet	227	50	9	59M	same padding	32.5	61.4
VGG16	224	50	19	15M	terminated early, too slow	18.5	42.1
ResNet50	224	23	52	24M	-	39.0	67.1
ResNet50	224	14	52	24M	modified optimizer	42.8	71.4
InceptionV3	299	8	50	24M	top-layer training	43.1	-
InceptionV3	299	50	50	24M	top-*N*-layer training, custom preprocessing	**61.4**	**85.2**

Table 1: Model accuracy results

Discussion

- On smaller models, issue was underfitting. With larger models, issue was generally overfitting, though data augmentation and model structure (e.g. residual blocks in ResNet50) mitigated this.
- Highest accuracy model was InceptionV3 pre-trained on ImageNet and with top few layers made trainable, with **61.4%** top-1 and **85.2%** top-5 accuracy.
- Higher top-5 accuracy shows that model is confused by visually-similar food types.

Figure 2: Mixed images of tiramisu, chocolate cake, chocolate mousse, and cheesecake (actual order in bottom left)

Future

- Hyperparameter search and optimization.
- Bounding box image preprocessing model.
- Train separate models for food sub-categories.

References

[1] Bossard et al., "Food-101–mining discriminative components with random forests," in European Conference on Computer Vision, pp. 446–461, Springer, 2014.
[2] Deng et al., "Imagenet: A large-scale hierarchical image database," in2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.
[3] He et al., "Deep residual learning for image recognition," inProceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
[4] Szegedy et al.,"Going deeper with convolutions," inProceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.