The beef with food recognition: a comparison of machine learning techniques

Nathan Spencer, Marco Piccirilli, Don Adjeroh, Gianfranco Doretto

Lane Department of Computer Science and Electrical Engineering, West Virginia University

Introduction

- **Food recognition:** the ability of a computer to identify different types of food in images
- Imagine a food processing plant that detects defective products
- What if you could track nutrition information by taking a picture of your meal?

Figure 1: Images of french onion soup (left) and macaroni and cheese (right) from the Food-101 dataset, which has 101 classes total

- Lukas Bossard et. al.[2] introduced Food-101 dataset (Fig. 1) for food recognition, achieved 56.4% acc. using convolutional neural network
- **Goal:** compare techniques for food recognition and improve on the accuracy reported by Bossard et. al.

Methodology

- Four classifiers were implemented and used to classify Food-101, and their resulting accuracies compared
  1. Bag of Words[1] **(BoW)**
  2. Improved Fisher Vector[1] **(IFV)**
  3. Convolutional Neural Network[3] **(CNN)**
  4. Fine-tuned Very Deep CNN[4] **(DCNN)**
- Food-101 set was used as is for consistency with Bossard et. al.; some incorrect labels exist (Fig. 2)

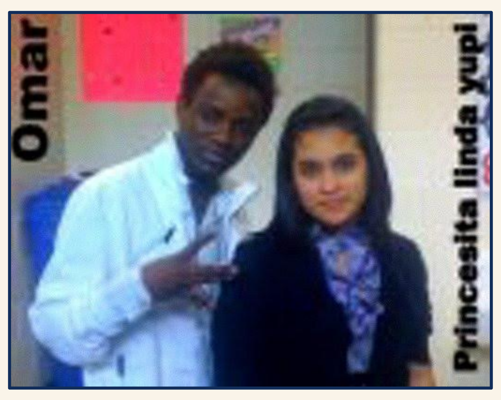

Figure 2: An applicable classifier must be able to deal with incorrect labels on training images such as this one, which is labeled as hummus

- Compared with two Bossard models: a CNN at 54.6% and a random forest (RF) model at 50.8%

Average Accuracies

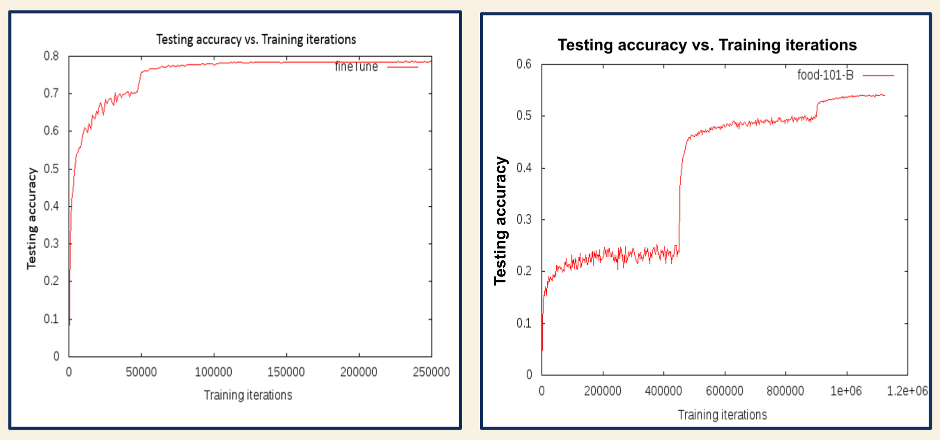

Figure 3: Accuracy for the DCNN over 250,000 training iterations (left), ultimately resulting in 78.6% accuracy. Accuracy for our CNN is shown for comparison (right), reaching 54.8% accuracy

- Accuracy for our CNN fell just short of the accuracy reported by Bossard et. al. (Table 1)
- Some variance expected due to random initialization of weights
- Fine-tuned DCNN outperforms Bossard et. al. CNN by a margin of **over 20%** (Fig. 3)

Table 1: Average accuracy for each of the four implemented classifiers as those reported by Bossard

	Model	Acc.
This work	BoW	27.6%
	IFV	42.2%
	CNN	54.8%
	DCNN	**78.6%**
Bossard et. al.	CNN	56.4%
	RF	50.8%

Class-by-Class Accuracies

Table 2: Accuracy by class for BoW and IFV

Class	BoW	IFV	Class	BoW	IFV	Class	BoW	IFV	Class	BoW	IFV	Class	BoW	IFV
apple_pie	10.8%	14.8%	chicken_wings	24.0%	46.8%	french_fries	34.0%	60.4%	lobster_bisque	50.4%	59.6%	pulled_pork_sandwich	15.2%	28.8%
baby_back_ribs	22.8%	40.0%	chocolate_cake	22.4%	30.4%	french_onion_soup	36.4%	58.0%	lobster_roll_sandwich	12.8%	27.6%	ramen	40.0%	52.8%
baklava	23.2%	43.2%	chocolate_mousse	13.2%	18.8%	french_toast	12.4%	32.8%	macaroni_and_cheese	28.8%	38.0%	ravioli	17.6%	20.8%
beef_carpaccio	36.0%	48.0%	churros	25.6%	54.8%	fried_calamari	24.4%	47.2%	macarons	48.8%	73.6%	red_velvet_cake	36.8%	50.8%
beef_tartare	14.0%	20.8%	clam_chowder	44.4%	62.8%	fried_rice	43.2%	58.0%	miso_soup	70.4%	77.6%	risotto	33.2%	46.8%
beet_salad	20.0%	34.0%	club_sandwich	24.0%	49.6%	frozen_yogurt	49.2%	69.6%	mussels	52.4%	67.2%	samosa	16.4%	32.0%
beignets	49.2%	64.8%	crab_cakes	7.6%	20.4%	garlic_bread	27.6%	36.8%	nachos	22.8%	32.4%	sashimi	30.8%	54.0%
bibimbap	53.6%	60.4%	creme_brulee	40.0%	58.8%	gnocchi	25.2%	31.6%	omelette	10.0%	22.4%	scallops	14.4%	21.2%
bread_pudding	11.6%	13.2%	croque_madame	30.8%	44.8%	greek_salad	22.8%	43.2%	onion_rings	45.6%	60.8%	seaweed_salad	54.8%	73.6%
breakfast_burrito	5.6%	17.2%	cup_cakes	43.2%	65.2%	grilled_cheese_sandwich	10.0%	28.8%	oysters	55.6%	76.4%	shrimp_and_grits	16.4%	38.0%
bruschetta	10.0%	19.2%	deviled_eggs	44.4%	66.8%	grilled_salmon	6.0%	14.4%	pad_thai	39.2%	54.8%	spaghetti_bolognese	50.8%	63.2%
caesar_salad	28.8%	43.2%	donuts	14.0%	42.4%	guacamole	19.2%	30.0%	paella	31.2%	43.2%	spaghetti_carbonara	57.6%	72.0%
cannoli	30.4%	41.6%	dumplings	51.6%	70.8%	gyoza	15.2%	42.4%	pancakes	29.2%	45.2%	spring_rolls	24.4%	44.8%
caprese_salad	17.2%	34.8%	edamame	64.4%	83.2%	hamburger	20.8%	28.8%	panna_cotta	26.8%	28.0%	steak	8.4%	14.0%
carrot_cake	22.0%	33.2%	eggs_benedict	24.8%	53.2%	hot_and_sour_soup	70.4%	78.0%	peking_duck	21.2%	44.8%	strawberry_shortcake	13.6%	31.6%
ceviche	10.0%	14.8%	escargots	30.4%	42.4%	hot_dog	26.0%	40.0%	pho	63.6%	77.2%	sushi	15.6%	40.8%
cheese_plate	16.0%	40.0%	falafel	21.2%	29.2%	huevos_rancheros	9.6%	20.4%	pizza	46.4%	60.8%	tacos	10.8%	27.2%
cheesecake	20.4%	34.0%	filet_mignon	12.0%	19.2%	hummus	18.8%	28.0%	pork_chop	8.4%	15.2%	takoyaki	22.0%	49.6%
chicken_curry	14.8%	20.8%	fish_and_chips	18.0%	38.8%	ice_cream	22.8%	28.0%	poutine	26.8%	44.0%	tiramisu	33.2%	41.2%
chicken_quesadilla	15.6%	40.4%	foie_gras	8.8%	16.8%	lasagna	19.2%	20.8%	prime_rib	37.2%	50.4%	tuna_tartare	12.4%	16.0%
												waffles	29.2%	50.4%

- IFV outperforms BoW for all 101 classes of Food-101 (Table 2)
- Even so, accuracies vary from 13.2% to 83.2%
- Further development needed to calculate for CNN, DCNN

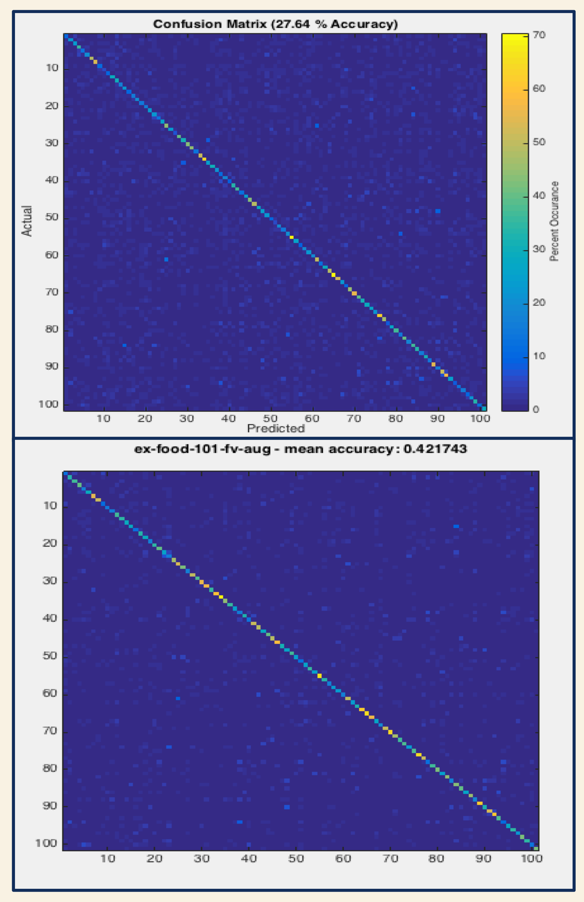

Figure 4: Confusion matrices for BoW (top) and IFV (bottom)

So what's the beef?

- While the DCNN classifier improves on current accuracies, it has tremendously many parameters: ~144 million (Fig. 5)
  - Therefore must be pre-trained on a larger dataset (Imagenet 2012)
- Larger dataset of food may produce features more useful for fine-tuning on Food-101
- Large memory needs of DCNN training also mandates small batch size on machines without sufficient GPU memory
  - This can create problems with instability in optimization of loss function (Fig. 6)
- Food-specific pre-training data and access to more memory could result in application-quality accuracy for food recognition

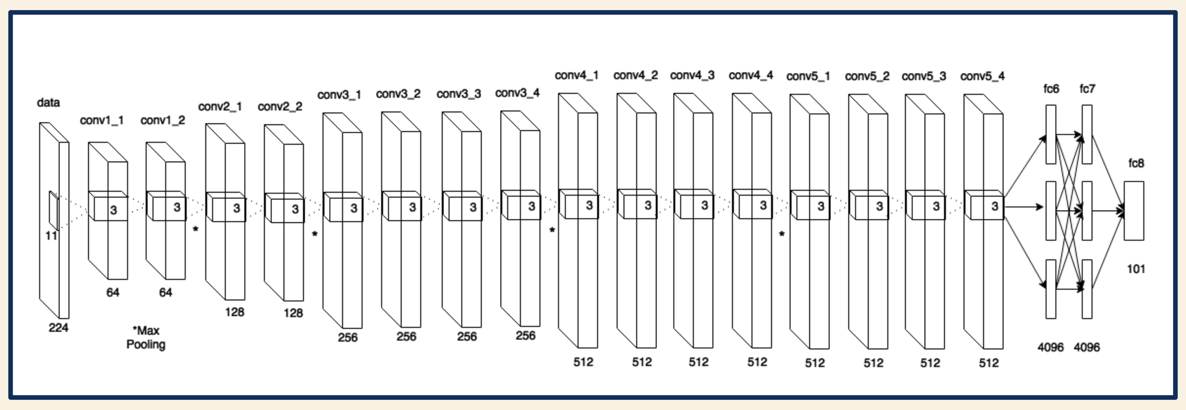

Figure 5: The DCNN uses 19 weight layers, resulting in a very memory-needy network relative to a more conventional 10-layer CNN

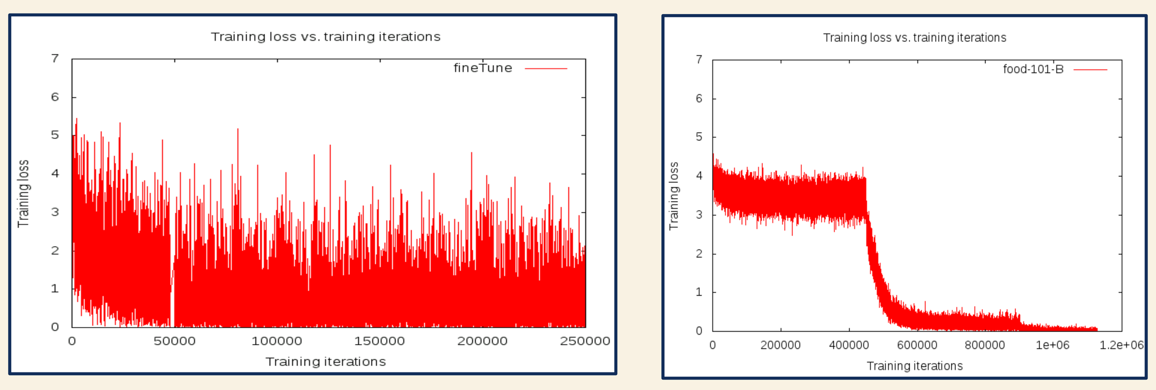

Figure 6: The training loss should be minimized as the training progresses, but small batch sizes make the descent quite noisy (4, left) compared to larger batches (64, right)

Sponsored by the WVU Office of the Provost with partial funding from the WVU Eberly College of Arts and Sciences, Statler College of Engineering and Mineral Resources, and the Davis College of Agriculture, Natural Resources, and Design.

[1] A. Vedaldi and B. Fulkerson. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/. 2008.
[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. "Food-101 – Mining Discriminative Components with Random Forests". In: European Conference on Computer Vision. 2014.
[3] Jia, Y. Caffe: An open source convolutional architecture for fast feature embedding.http://caffe.berkeleyvision.org/, 2013.
[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

West Virginia University