Food Image Recognition Using Very Deep Convolutional Networks

Hamid Hassannejad

2nd International Workshop on Multimedia Assisted Dietary Management
Oct. 2016

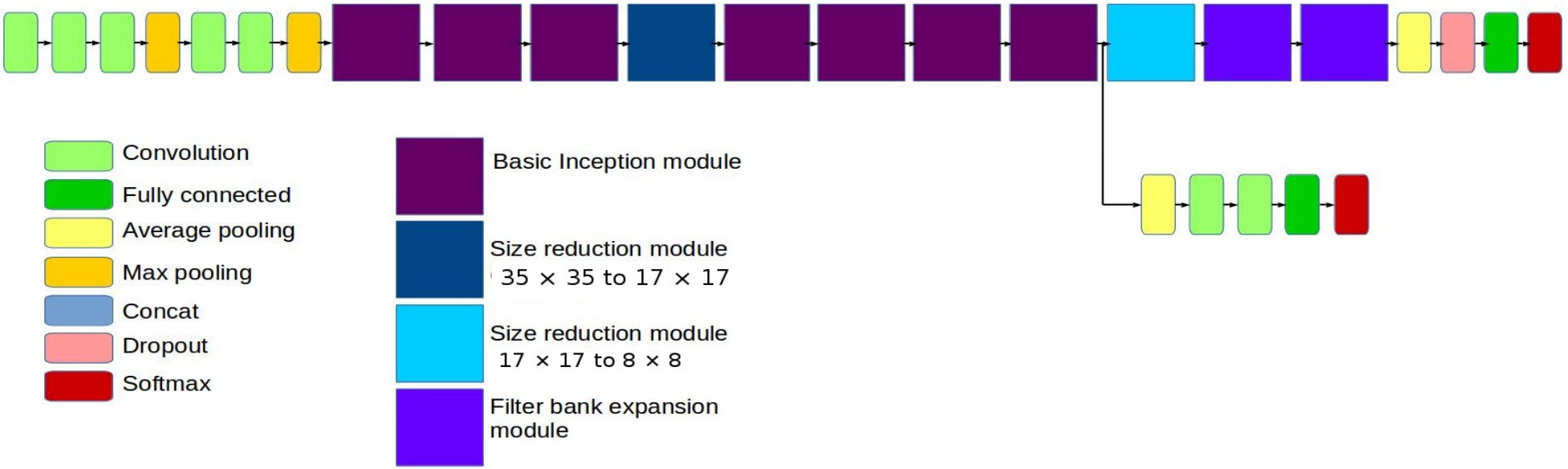
Outlines

- Deep learning and food recognition
- Inception model
- Experiments

Deep Learning on Image Recognition

	Year	Top-1 Err.	Top-5 Err.	N. layers	N. Params
AlexNet	2012	37.5%	17.0%	8	60 millions
GoogLeNet	2014	21.2%	5.6%	22	5 millions

Inception V3

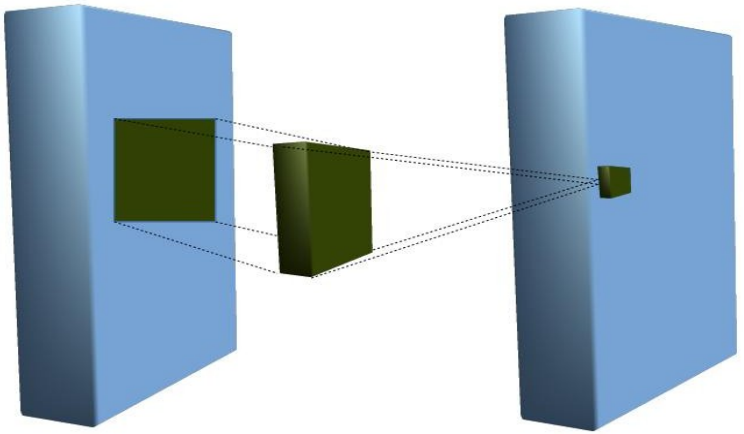


- 54 layers
- 25 million parameters
- On ILSVRC 2012: top-1 and top-5 error rates of 17.3% and 3.5%

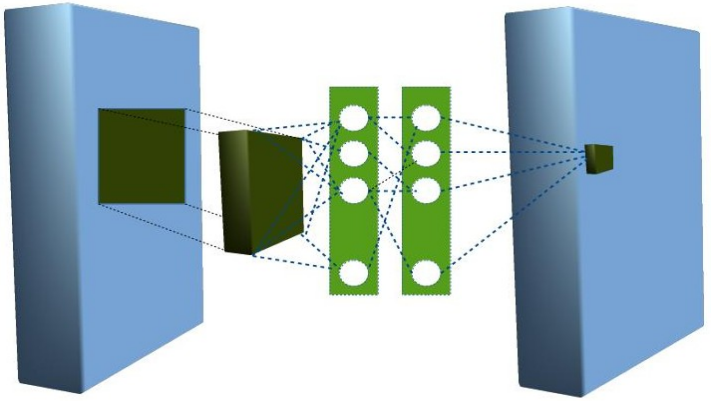
Inception

- Increasing the size of network can improve it, but the size of parameters and computational cost would increase dramatically.
- Google approached these issues by proposing a deep network whose architecture is based on ”Inception modules” and is inspired by two main concepts:
  - Network-in-Network
  - sparse networks

•Network-in-Network



Linear convolution layer



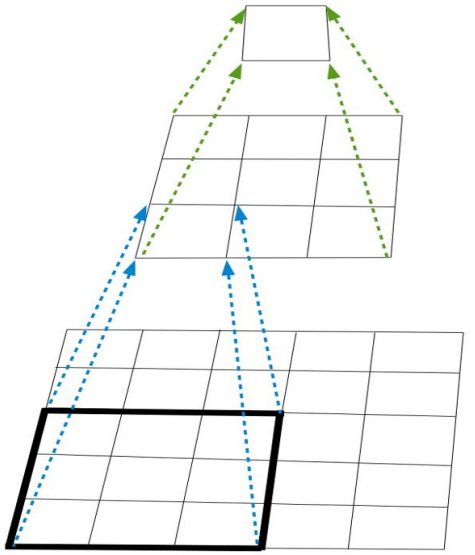
Multilayer perceptron layer

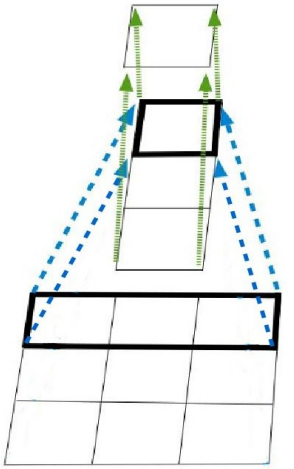
Design Principles

- Optimal performance of the network can be reached by properly balancing the number of filters (convolutions) per stage and the depth of the network.
- Avoid representational bottlenecks, especially early in the network. In other words, the representation size should gently decrease from the inputs to the outputs before reaching the final representation used for the task at hand.
- Spatial aggregation can be operated on lower-dimensional embeddings without much or any loss in representational power.
- Higher-dimensional representations are easier to process locally within a network. Increasing the activations per tile in a convolutional network allows for more disentangled features.

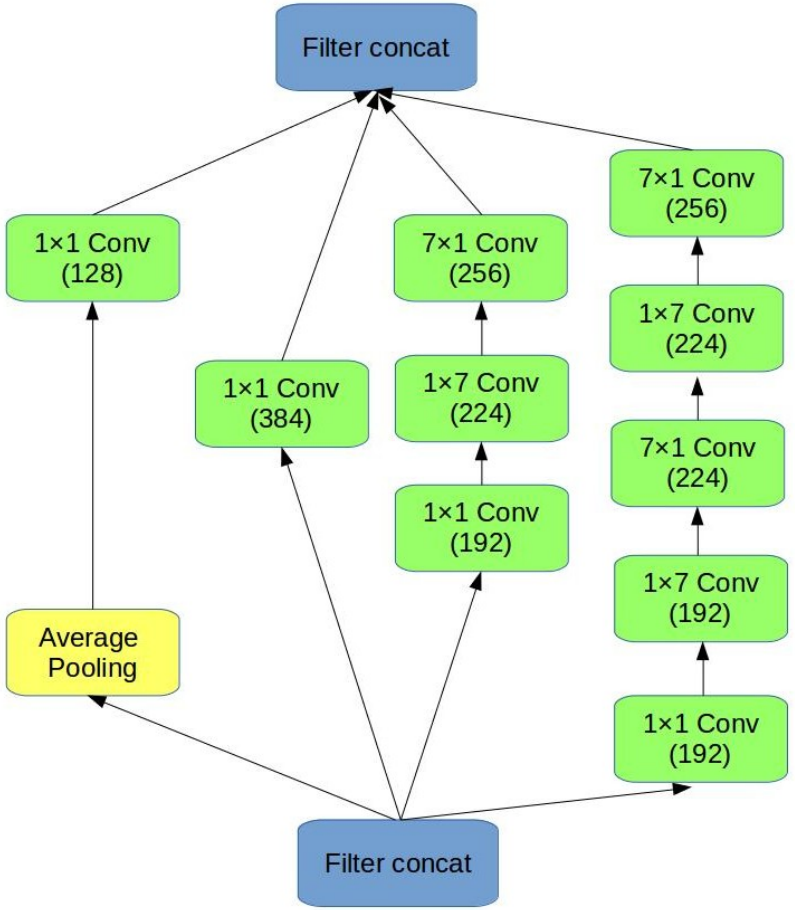
Factorization

- Two main techniques are applied in order to increase computational efficiency:
  - factorization into smaller convolutions.
  - spatial factorization into asymmetric convolutions.

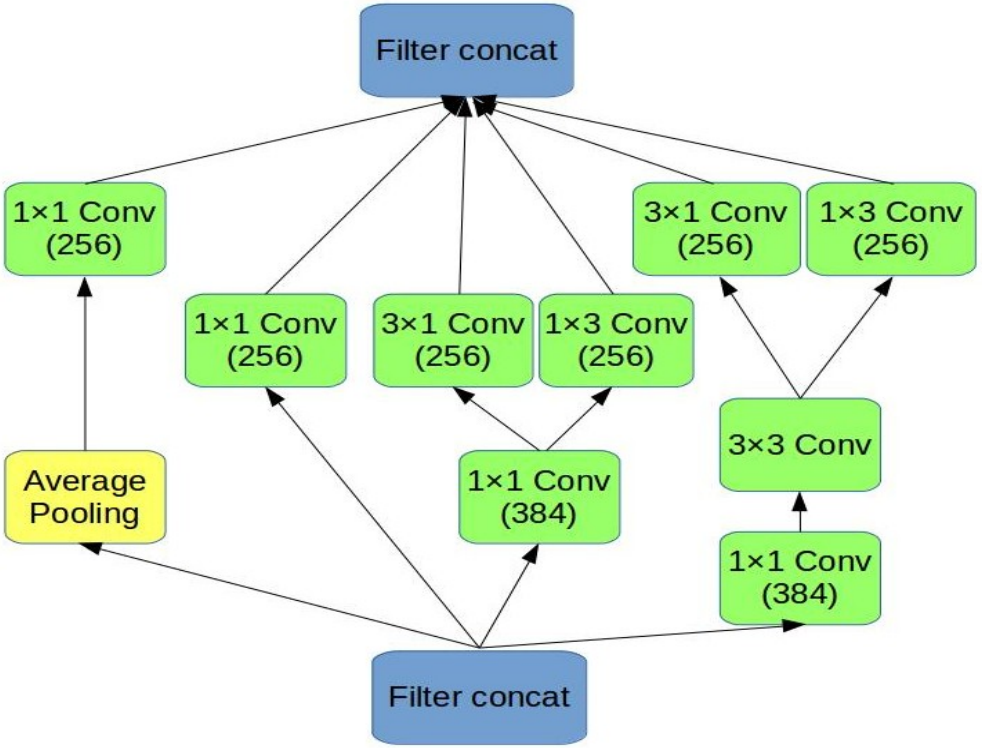




Inception Modules

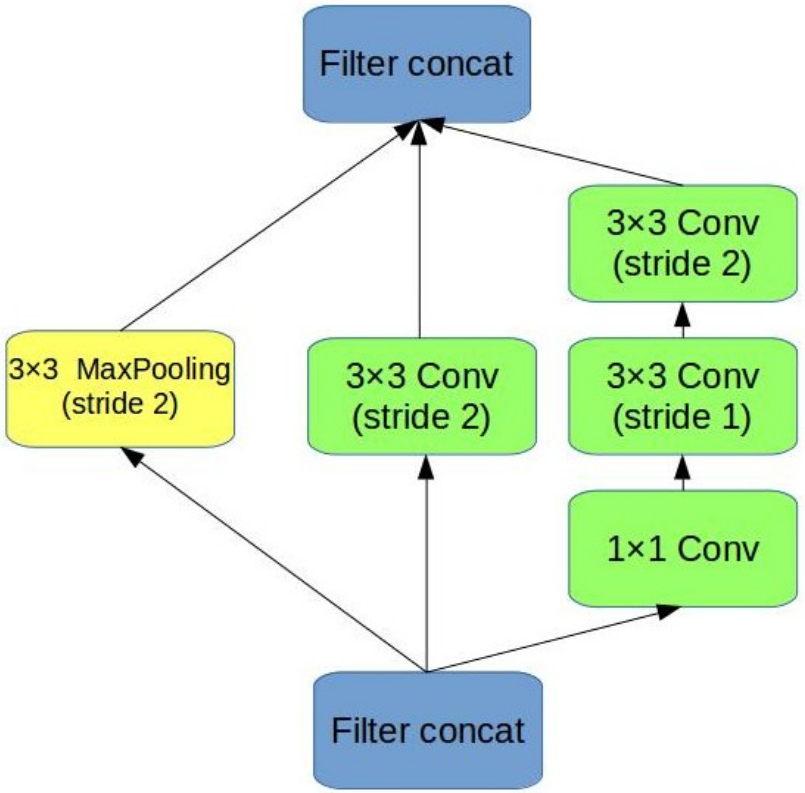


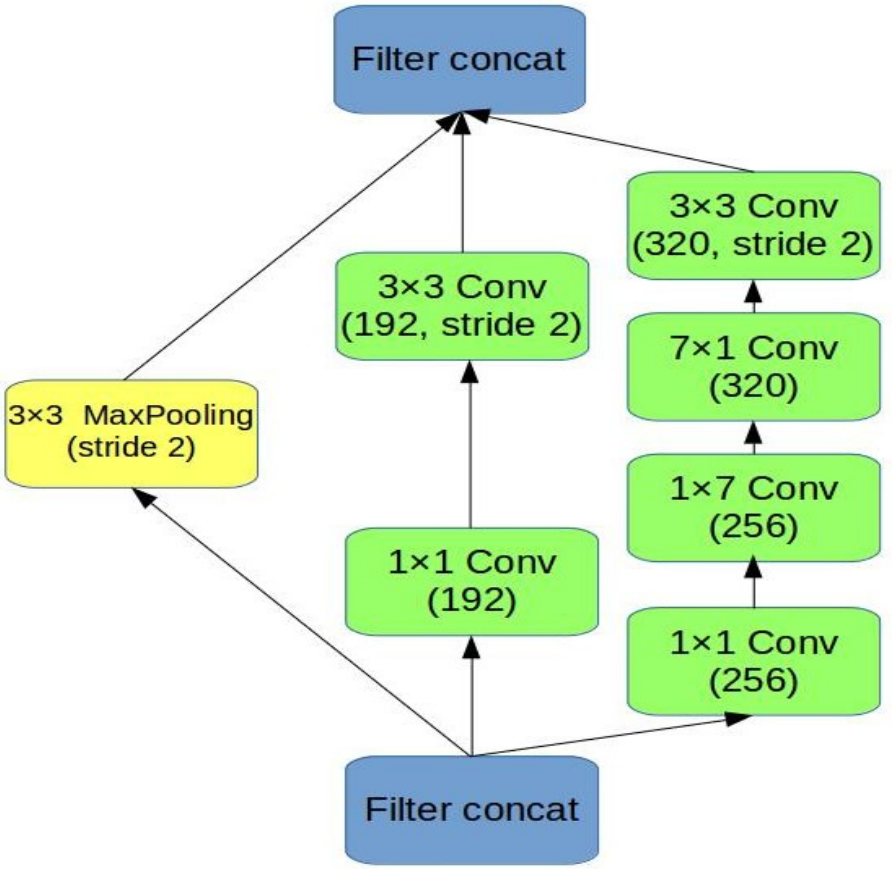
Basic modules: There are seven basic modules in the model, which are designed to approximate the optimal local sparse structure. They factorize a bigger 17×17 grid as two consecutive 7×7 convolutions.



Filter bank expansion module: Two filter bank expansion modules are used on the coarsest (8 × 8) grids to promote high-dimensional representations by expanding the filter bank outputs.

Inception Modules





Size reduction modules: Two Inception modules with different depths, that reduce the grid size while expanding the number of filter banks. They are used to reduce model dimension wherever the computational requirements would be too heavy otherwise. The left one is a reduction module from 35×35 to 17 × 17 and the right one from 17 × 17 to 8 × 8.

Test Datasets

- **ETH Food-101** : 101 food and dessert categories. 101, 000 images. No bounding boxes. Divided to training and test sets.
- **UEC FOOD 100** : 100 food categories (popular in Japan). more than 14, 000 images. There is bounding box for each image. No training and test sets definition.
- **UEC FOOD 256** : The same as UEC FOOD 100, but considering 256 international food categories and including about 32, 000 images.

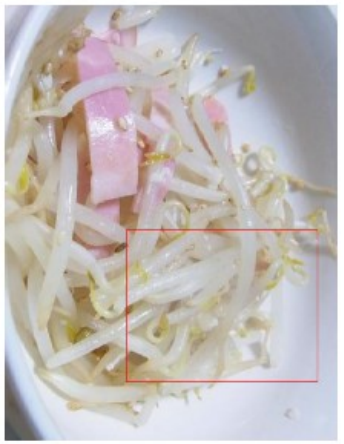
Experiment

- To train such a deep model successfully from scratch, we would have needed millions of images. However, the available food image datasets provide a small fraction of such requirements.

- To tackle the problem, a series of random distortions were applied to the training images to artificially expand the datasets:
  - Randomly cropping the images.
  - Resizing the cropped piece to 299 × 299.
  - Distorting the image brightness.
  - Distorting the image contrast.
  - Distorting the image saturation.
  - Distorting the image hue.

Artificial Samples

Original Images

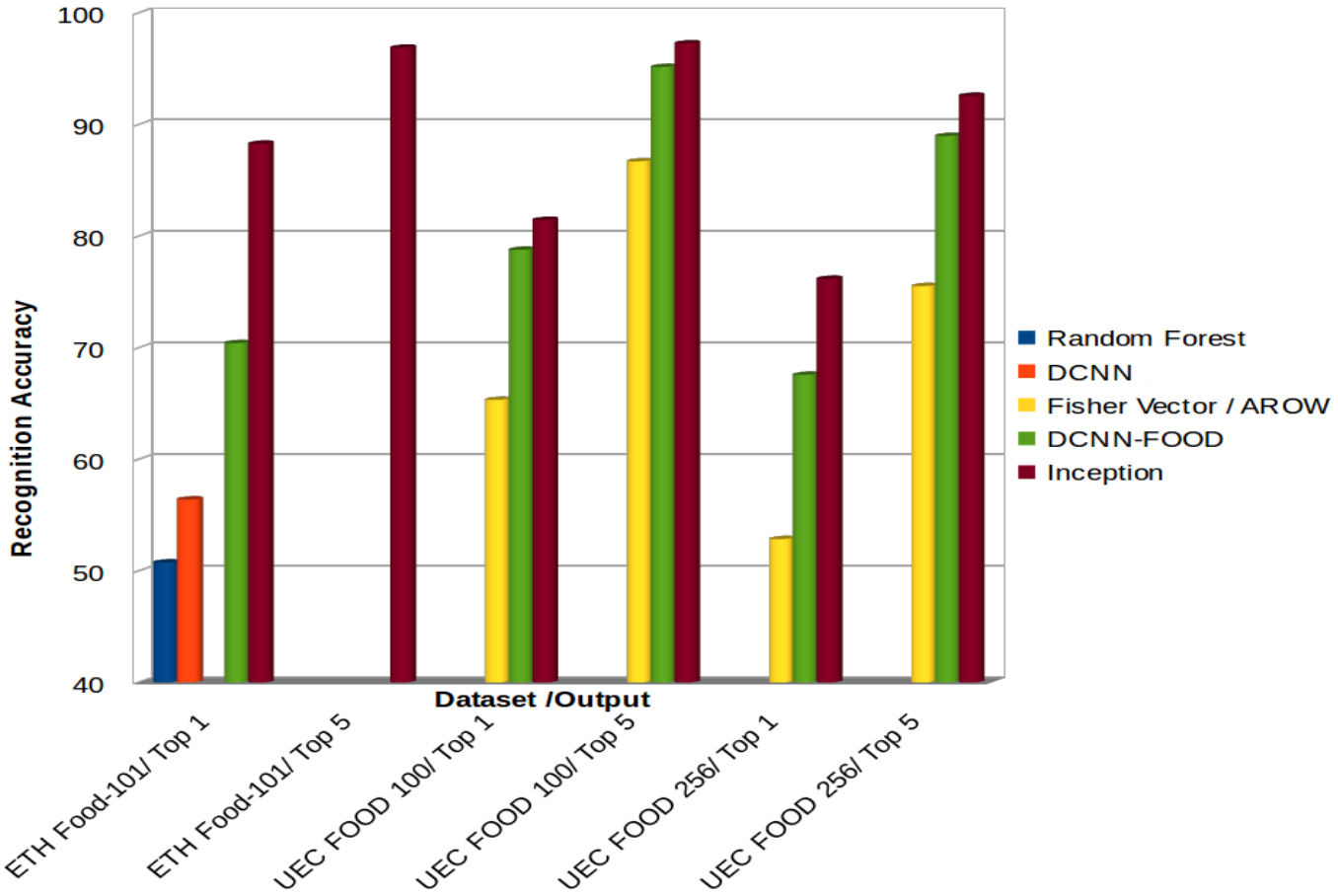


Distorted Images



Experimental Results

Method	ETH Food-101 (%)		UEC FOOD 100 (%)		UEC FOOD 256 (%)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CNN	56.40					
DCNN-FOOD	70.41		78.77	95.15	67.57	88.97
Inception V3	**88.28**	**96.88**	**81.45**	**97.27**	**76.17**	**92.58**



Next?

- Next model? Inception V4, ...
- Larger training dataset.
- Mobile implementation.

Food Image Recognition Using Very Deep Convolutional Networks

Hamid Hassannejad

2nd International Workshop on Multimedia Assisted Dietary Management
Oct. 2016