CS435: Introduction to Deep Learning

Deep Learning Proposal
Food Image Recognition

Team Members

Hanan Elkhateeb (20)

Youmna Dwidar (73)

Table of Contents

1 Problem statement

Our main problem here is automatic food recognition. Food is an important part of our everyday life. It got into digital life by the richness of photography in social media and dedicated photo sharing sites. This implies that automatic recognition of Food would not only help users effortlessly organize their extensive photo collections but would also help online photo repositories make their content more accessible. Additionally, food journal apps are now used to help patients estimate and track their daily caloric intake to modify their food habits and maintain a healthy diet. However, Current food journaling applications such as MyFitnessPal App require users to enter their meal information manually. Some people found that to be time consuming.

2 Current state of the art accuracy

We found that we have mainly two famous datasets. So here , we are going to introduce the state of the art accuracy for both of them.

First, we have the dataset Food101 with top-1 accuracy of 92.47%.

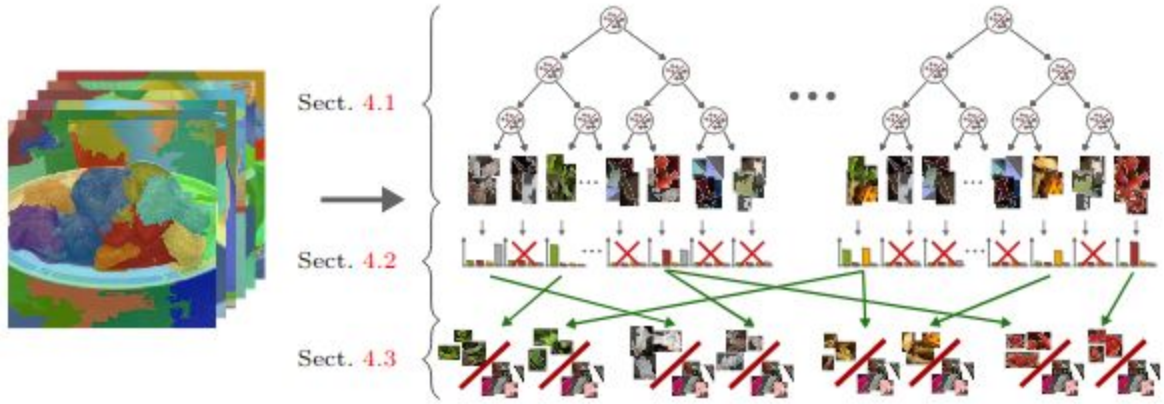
Secondly, we have the newest dataset foodX-251 with state-of-the-art 83% top-3 accuracy.

3 Short survey of models and solutions

First, we will go through the different solutions and models for the Food-101 dataset as it was addressed by several experiments and models. Secondly, we will address the newest dataset which is foodX-251. As far as we know, It was only mentioned in its release paper and hasn't been addressed yet by research due to its recent release.

Food-101

It started with accuracy 50.76% with the paper release model. They constructed the food images by collecting a novel real-world food dataset by downloading images from foodspotting.com. Then, all images were rescaled to have a maximum side length of 512 pixels and smaller ones were excluded from the whole process.



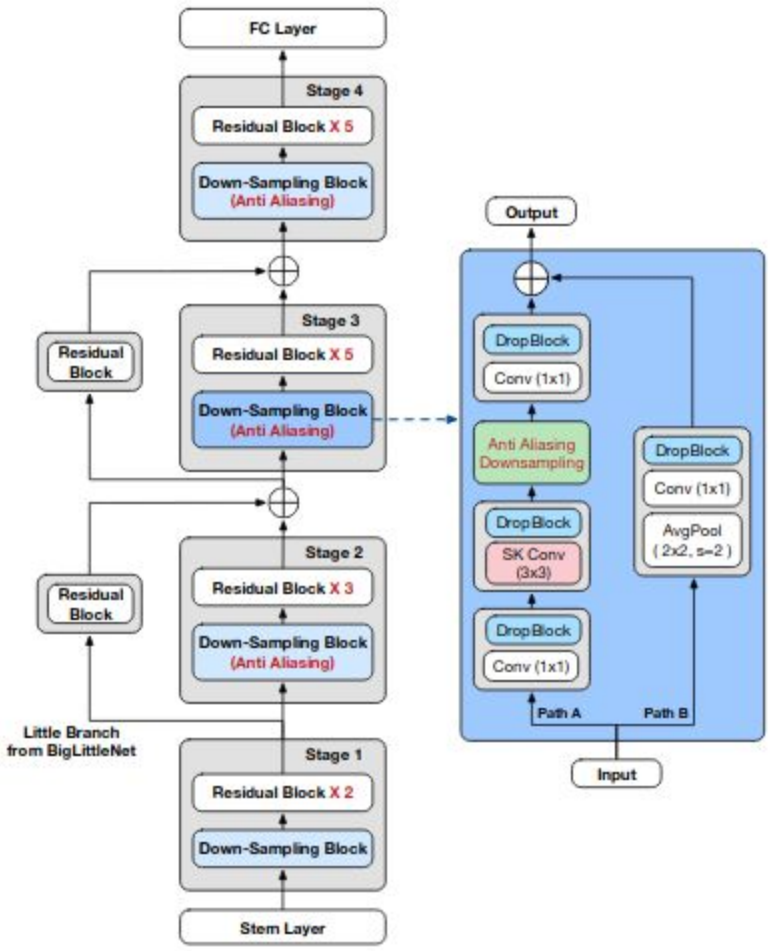
A Random Forest is used to hierarchically cluster superpixels of the training set. Then, discriminative clusters of super pixels in the leaves are selected and used to train the component models.After mining, the RF is not used anymore.For each class, they select the top N leaves and train for each one a linear binary svm to act as a component model.

Then we found some work on the food-101 dataset that includes training a well-known deep learning architectures.

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	18.8	43.5
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

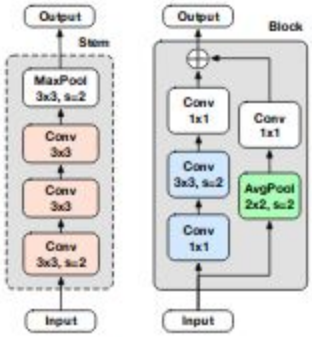
Lastly, we are introducing the state-of-the-art model for this dataset.



Assembling techniques into ResNet-50.

They applied network tweaks such as ResNet-D, SK, Anti-alias, DropBlock, andBigLittleNet to ResNet. In more detail, ResNet-D and SK apply to all blocks in all stages.

Downsampling with anti-aliasing only applies to the downsampling block from Stage 2 to Stage 4. Drop-Block only applies to all blocks in Stage 3 and Stage 4. Little-Branchfrom BigLittleNet uses one residual block with smaller width.



(b) ResNet-D

ResNet-D is a minor adjustment to the vanilla ResNet network architecture mode.

FoodX-251

The model used is the ResNet-101 architecture.

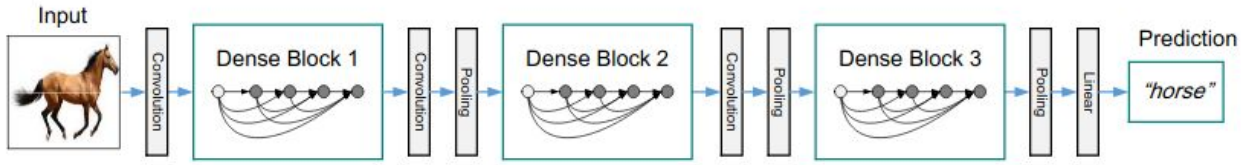
Method	Top-3 Error %		
	Val.	Test	
		Public	Private
ResNet-101 (*finetune* last-layer)	0.36	0.37	0.37
ResNet-101 (finetune all-layers)	0.16	0.17	0.17

4 Detailed description of the used model and why is it used

4.1 Detailed description

We decided to go with the DenseNet architecture. The detailed architecture is discussed here.

The high level architecture of the denseNet is described in the figure below.

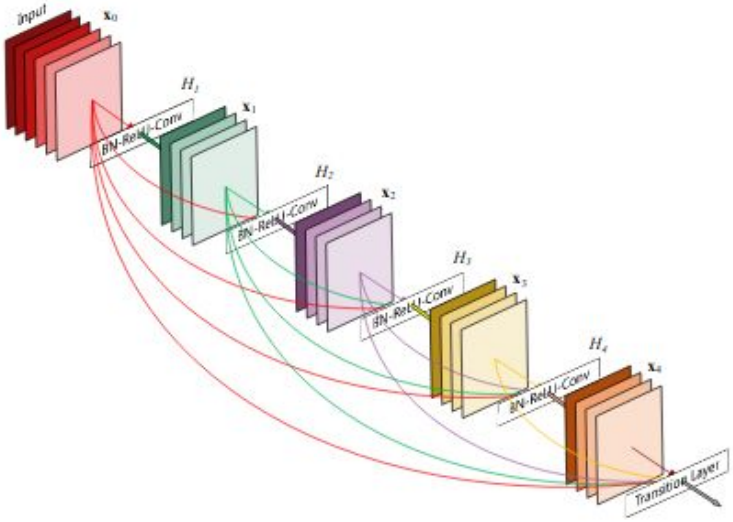


The Dense Connectivity is inspired by the ResNets. To further improve the information flow between layers, they proposed a different connectivity pattern.Direct connections from any layer to all subsequent layers as shown in the figure.

The l'th layer receives the feature-maps of all preceding layers,x0,...,x`−1, as input:

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]),$$

Where[x0,x1,...,x`−1] refers to the concatenation of the feature-maps produced in layers 0 , ... , l−1



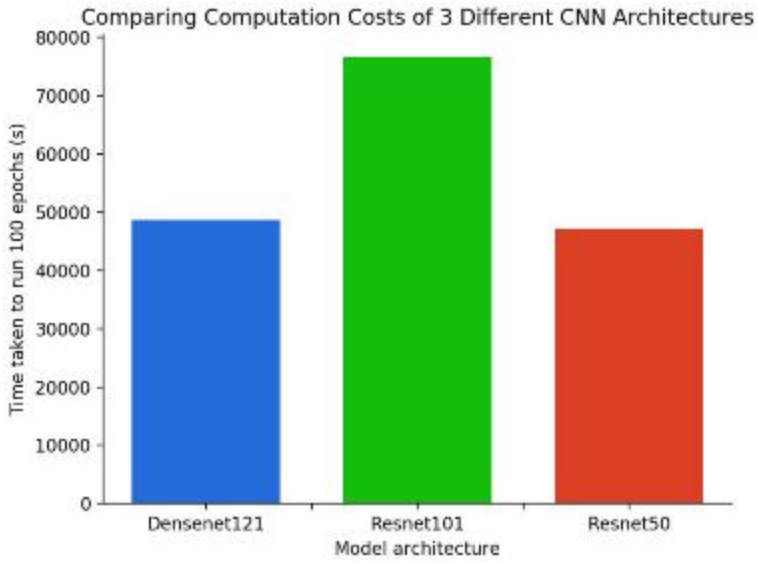
A 5 layer dense block with growth rate of k = 4

The growth rate If each function Hl produces k feature-maps, it follows that the l'th layer has k0 + k x (l−1) input feature-maps, where k0 is the number of channels in the input layer. An important difference between DenseNet and existing network architectures is that DenseNet can have very narrow layers,e.g.,k= 12

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

4.2 Why are we using this architecture

While searching for this problem, we came across a similar problem which is translating the food recipe from the food picture [ChefNet]. The paper used this architecture for training the photo part and They were mentioning that the computational cost to train their dataset by DenseNet-121 was a little higher than ResNet-50, but was nearly half the computational cost for training by ResNet-101.



Then, we read the architecture of the DenseNet which was inspired by the ResNet and gave a better performance. So , we decided to try this architecture for its lower computational cost and better performance in several problems.

5 Food Recognition Datasets Study

We provide a detailed study and analysis of the food recognition datasets exploited in the literature. We then discuss the motivations that encouraged us to settle on our dataset of interest.

5.1 Datasets Description and Comparison

As for food classification problem, there are limited datasets. Earlier works tried to tackle this issue by collecting training data using human annotators or crowd-sourcing platforms.

In this table, we give a detailed comparison between the food recognition datasets used in literature, considering their convenience, suitability, size and source.

Dataset Name	**Number of Categories**	**Total images**	**Description**
ETHZ Food-101	101	101,000	- The images are downloaded from a photo sharing website for food items (foodspotting.com). - The test data was manually cleaned by the authors whereas the training data consists of cross-category noise, i.e., images with multiple food items labeled with a single class.
UPMC Food-101	101	90,840	- This dataset has the same 101 categories as ETHZ Food-101 but the images are downloaded using Web search engine.
Food50	50	5000	- Those food recognition datasets have fewer food categories. - The images are downloaded using Web search engine.
Food85	85	8500	
CHO-Diabetes	6	5000	
Bettadapura et al	75	4350	
UEC256	256	At least 100 per class	- consists of 256 categories with a bounding box indicating the location of its category label. However, it mostly contains Japanese food items. - The images are downloaded using Web search engine.

ChineseFoodNet	208	185,628	- This dataset is restricted to chinese food items only - The images are downloaded using Web search engine.
NutriNet dataset	520	225,953	- Images are from 520 food and drink categories but are limited to Central European food items. - The images are downloaded using Web search engine.
FoodX-251	251	158,846	- Dataset consists of miscellaneous food items from various cuisines - The images are downloaded using Web search engine.

5.2 Dataset Election

In our work, we will use the FoodX-251 dataset due to the following reasons:

1. The availability and adequacy of the dataset; it is the newest work in food recognition problem and its size is adequate for training and evaluation purposes.
2. It provides more classes and images than existing datasets.
3. Features miscellaneous classes as opposed to a specific cuisine/food type.

6 Evaluation metrics

We will follow the same metric used at (P Kaur et al). For each image i, an algorithm will produce 3 labels l_{ij} , j = 1, 2, 3, and has one ground truth label g_i . The error for that image is:

$$e_i = \min_j d(l_{i,j}, g_i),$$

Where,

$$d(x, y) = \begin{cases} 0, & \text{if } x = y. \\ 1, & \text{otherwise.} \end{cases}$$

The overall error score for an algorithm is the average error over all N test images:

$$score = \frac{1}{N} \sum_i e_i.$$

7 Graduation Project Problem Statement

Most of the IT applications require storing and retrieving information from databases. We can store a big amount of data in databases but The data retrieval requires knowledge of domain- specific language like SQL (Structured Query Language) so casual users who don't have any technical background won't be able to access data. Our project aims to develop a system which will accept English natural-language statements from users and convert it into SQL query to be executed on a database.

8 References

Bossard et al., "Food-101–mining discriminative components with random forests," in European Conference on Computer Vision, pp. 446–461, Springer, 2014. link

Gao Huang et al., 2018."Densely Connected Convolutional Networks". link

Jungkyu Lee, Taeryun Won and Kiho Hong.2020."Compounding the Performance Improvements of Assembled Techniques in a Convolutional Neural Network ''. link

Kaylie Zhu, Harry Sha and Chenlin Meng.2018." ChefNet: Image Captioning and Recipe Matching on Food Image Dataset with Deep Learning". link

Malina Jiang.2019."Food Image Classification with Convolutional Neural Networks". link

P Kaur et al.,2019. "FoodX-251: A Dataset for Fine-grained Food Classification". arXiv preprint arXiv:1907.06167. link