Bird's Eye Review on Food Image Classification using Supervised Machine Learning

Shamay Jahan[1], Shashi Rekha. H[2], Shah Ayub Quadri[3]

[1]M. Tech., Semester 4, DoS in CSE, Visvesvaraya Technological University, Centre for PG Studies, Mysuru, India
[2]M.Tech., Assistant Professor, DoS in CSE, Visvesvaraya Technological University, Centre for PG Studies, Mysuru, India
[3]M.Tech., Data Scientist, Hyderabad, India

Abstract **– Food image recognition and classification is a challenging task in image recognition. Due to a variety of food dishes available, it becomes a very complicated task to correctly classify the food image as belonging to one of the predefined classes. Significant work has been carried out on food recognition and classification using Computer Vision. Food recognition is important to assess diet of people with diabetics and people suffering from various food allergies. Food recognition also helps in finding the calorie value of foods, its nutrition value, food preferences etc. This survey paper covers some of the work done in food image recognition and classification using Deep Convolutional Neural Networks (DCNN) using various parameters and models, and other machine learning techniques. The classification accuracy of various models is also mentioned. The on-going work on Indian food image classification is also mentioned, which aims to improve the classification accuracy by choosing the suitable hyperparameters.**

Keywords **– Food image recognition, image classification, Convolutional Neural Networks, Pattern Recognition, Supervised Machine Learning, computer vision**

I .INTRODUCTION

As a saying goes "***We are what we eat***", food plays a significant role in human life. People nowadays are more cautious about their eating habits. In order to avoid various diseases and to live a healthy lifestyle, eating right food is necessary. For this purpose identification of food is important. Food classification is more challenging than any other image classification due to huge diversity in food. Food has a distinct, unique colour, texture, shape, size etc., which may vary if the ingredients in food are changed. Machine learning techniques are extensively used in image classification of food images. A lot of work has been done on image recognition and classification using Bag of Features (BoF) model, Support Vector Machine (SVM), Neural Networks, Convolutional Neural Networks (CNN) etc. Among the various classification techniques, CNN has outperformed all the other conventional approaches.

This paper covers relevant work done on food image classification using various techniques. The main focus of this paper is to study various work done on deep learning model, CNN, by varying the hyperparameters, applying optimization, using different datasets, etc. A comparative study is also presented on various CNN models used for image recognition and classification.

CNN is used as a state of art for image classification because of its ability to automatically extract the features from the image that are necessary for image classification, unlike other approaches that rely on handcrafted features. It is concluded that CNN has outperformed all the other machine learning techniques in image classification task.

The main purpose of this paper is to deeply understand the various techniques used for image classification and their corresponding results. The paper highlights the power of deep learning algorithm, CNN for image recognition and classification tasks.

The following paper is organized as follows: Section II provides an insight on the various models used for food recognition and classification. In section III, the motivation to develop the CNN model is addressed. The proposed idea is discussed in section IV, which is aimed to improve the accuracy and be deployed for real-time use. The paper ends with the conclusion in section V.

II RELATED WORK

A. Using different Machine learning approaches

In [1], Abhishek Goswami and Haichen Lu have performed a comparative study on food image classification using various deep learning models and have provided the detailed description of the models and the obtained results.

In paper [1], 26,984 colour images of 20 different classes from across the globe have been chosen for classification. The images are divided into the following three categories: training (18,927 images), validation (5,375 images) and testing (2,682 images). The images were sourced from Google, Bing Image Search API. The images were preprocessed and resized to 32X32X3, which caused a loss of 10% of the images; because those images did not fit the dimensions specified.

The table provides the details on the category and the number of images in each category.

Table I Description of the food image dataset

Food item Class	Number of images per class
Biryani	865
Bratwurst	876
Burger	912
Burrito	888
Cordonbleu	1,018
Dal	1,031
Dumplings	1,091
Fried Rice	966
Fries	745
Ice-Cream	1,021
Lasagna	971
Naan	1,020
Padthai	919
Pasta	1,002
Pizza	885
Ramen	1,023
Roast Turkey	930
Samosa	897
Sandwich	847

The experiments were carried out on various machine learning models. The evaluation was the accuracy obtained from each model. A detailed description of each model is explained below and summarized in a tabular form along with the validation accuracy and test accuracy.

1) Linear Classifier on raw image pixels: Linear classifier works as a *template match*, where each row of learned weights corresponds to a template for a particular class. Raw image pixels were used as a feature and both SVM (Support Vector Machine) classifier and Softmax classifier were used.

The learning rate was set to 1e-07 with the regularizing strength of 2.5e+04. The best validation accuracy was provided by the SVM classifier.

2) Six fully connected Neural Networks (NN) on raw image pixels: The network architecture had six fully connected layers. ReLU non-linear operation was performed on all layers and softmax loss function was applied to the last layer. Raw image pixels were again used as a feature.

The learning rate was set to 1e-03. The model was trained and validated on 20 epochs and best classification accuracy obtained on validation set was 0.19. **Adam** optimizer was used for parameter update rule. The test accuracy was 0.18.

One important observation was **batch normalization** was useful for improving the performance of training the model.

3) Image features with SVM and two layered fully connected NN: Image features were extracted and a feature vector was formed by concatenating the HOG (Histogram Oriented Gradients) and colour histogram. Each image had 155 features.

Linear SVM Classifier: Validation accuracy obtained was 0.21 with the learning rate of 1e-03 and the regularization strength of 1e+00. SGD (Stochastic Gradient Descent) was used as the update rule.

NN classifier: Validation accuracy obtained was 0.26 and the test accuracy was 0.27. The learning rate was set to 0.9 with the regularization strength of 0 with SGD update rule.

4) Convolutional Neural Networks (CNN): The CNN architecture had five convolutional and max-pooling layers with two fully connected layers. The kernel size was fixed to 32x32 with 32 filters in each convolutional layer. A dropout of 0.75 was applied to each layer. In the last layer, softmax classifier with cross-entropy loss was used. The best validation and test accuracy of 0.4 were achieved using Adam update rule over 25 epochs. The learning rate was set to 1e-04.

*5) Transferred learning using VGG-16 pre-trained model:*To further improve the accuracy obtained by the CNN model, a VGG-16 model[2] pre-trained on ImageNet was used. The VGG-16 model was modified by dropping the last fully connected layer and was replaced with 20 outputs. The last layer was trained with 10 epochs then the whole network was trained for 10 more epochs.

The dataset was pre-processed and cropped to 255X255. The training set was horizontally flipped to one half probability, and the entire dataset was subtracted with VGG colour mean.

B. CNN with hand-crafted features, Fisher Vector with HOG and Colour Patch

Yoshiyuki KAWANO, Keiji YANAI, in their work [2] have tried to improve the classification accuracy by using Deep Convolutional Neural Networks with the traditional hand-crafted features, Fisher Vectors with HoG and colour patches which is the work done by Chatfield et al. on generic dataset such as PASCAL VOC 2007 and Caltech-101/256.

UEC-FOOD100 dataset containing 100 classes, with each class containing more than 100 images of food was used in their experiment. Food photo also has a boundary box indicating the location of food. Food dataset with 70 different food items was used for the experiment. The earlier recorded accuracy on this dataset was 59.6%.

Fig. 1 shows the 70 classes of colour food images used to train the model.

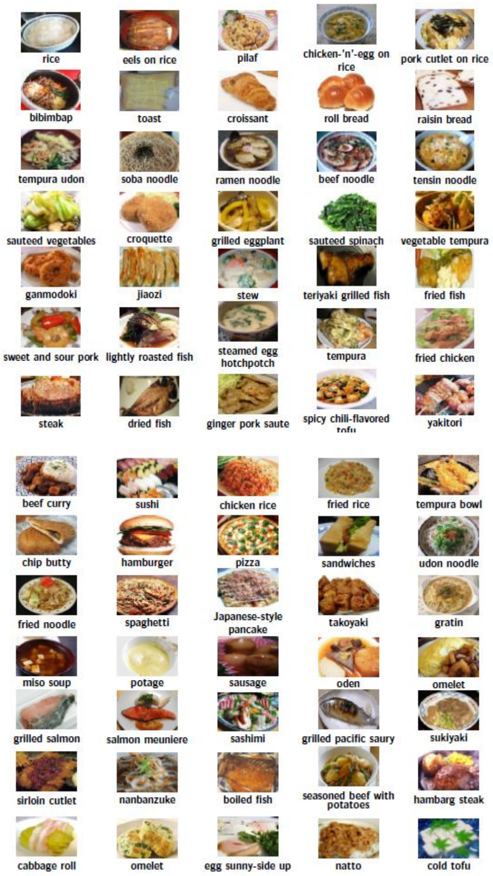


Fig. 1 Dataset with 70 classes of food images

The pre-trained DCNN with the ILSVRC 1000-class dataset was used as a feature extractor. The conventional features, such as ROOTHoG and colour patches were integrated with the DCNN. One-vs-rest linear classifier with 5-fold validation was used for food classification.

The classification accuracy was evaluated based on various features and their corresponding accuracy is mentioned in the following table.

Table II Feature selection and the Classification accuracy

Feature	Accuracy (%)
RootHoG-FV	50.14
Colour-FV	53.04
RootHoG-FV + Colour-FV	65.32
DCNN	57.87
RootHoG-FV + Colour-FV+DCNN	72.26 (top-1) 92.00 (top-5)

C. CNN with pre-training and fine-tuning

Keiji Yanai, Yoshiyuki Kawano further improved their obtained accuracy by using 2000 categories of food images on the pre-trained DCNN model [5].

The pre-trained model is the modified network structure of the AlexNet, which acts as a feature extractor. In relation to the work done by Oquab et al. [6], the features of DCNN were improved by extracting 1000 food-related classes from 21,000 categories ImageNet and adding them to the ILSVRC 1000 ImageNet, so as to pre-train the DCNN.

It took about one week to pre-train the DCNN on the NVidia GeForce TITAN BLACK GPU and 6GB RAM. Caffe was used to pre-train the model.

The model was trained on Japanese food item dataset, UEC-FOOD100 dataset and theUEC-FOOD256 dataset that is openly available for public use. Each class consisted of more than 100 images in each class. The best classification accuracy achieved on the fine-tuned UEC-FOOD100 and UEC-FOOD256 is 78.77% and 67.57% respectively.

Fig. 2: Example images in the dataset. (From the top) ramen noodles, dipping noodles (tsukemen), sushi and omelette.

D. CNN with parameter optimization and colour importance for feature extraction.

A very interesting work has been done by Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa who have developed CNN model by parameter optimization [3]. Colour is a very important factor in the food image recognition task. This important observation is made by them, i.e., feature kernels heavily rely on colour. This means that the colour images provide better understanding and better learning to the CNN model. This will be a useful factor for the model to learn more detailed features, which will obviously improve the classification accuracy. The dataset was created from the Food Logging (FL) system available for public use. The dataset consisted of 1, 70,000 images of everyday food meals belonging to 10 commonly logged meals category. The 80X80 images were cropped to 64x64 using cuda-convent python module.

A comparative study was conducted to measure the performance of CNN model vs the traditional hand-crafted SVM. It was observed that the SVM obtained 50-60% accuracy in comparison to two-layered CNN which achieved more than 70% accuracy.

CNN model was 5X5 filter size, one time normalization with 6-fold cross validation provided a good accuracy of 73.70%.

Table III provides the information on the images chosen from two months FL of 2013.

Table III Common meals from Two months FL

Food Item	Number of images per class
Cold Tofu	1,434
Curry and Rice	979
Deep-fried Chicken	910
Green salad	3,603
Grilled Salmon	951
Miso Soup	5,786
Natto	2,338
Ramen	1,070
Rice	11,560
Yoghurt	1,724

A comparative study was done using CNN and SVM hand-crafted features. It was observed that the accuracy of SVM was 50-60%, whereas CNN achieved more than 70% accuracy. Also, the colour features further enhanced the accuracy of CNN. Thus, it was concluded that colour features dominate food recognition process which is in-line with the work done by Bosh [4], where hand-crafted colour features were regarded as best in hand-feature extraction.

E. CNN with Affine Data Transformation Technique

Yuzhen Lu [7] has used a small dataset consisting of 5,822 colour images belonging to 10 different classes. The classification accuracy is evaluated over Bag-of-features (BoF) model integrated with SVM model and the five layered CNN.

The images were sourced from ImageNet. To increase the images in the dataset, affine transformations were applied to the dataset.

Table IV provides the images per category that was used in the experiment.

Food Item	Number of images
Apple	1,050
Banana	310
Broccoli	327
Burger	519
Egg	626
French fry	296
Hotdog	639
Pizza	1,248
Rice	352
Strawberry	455

Prior to training the model, the images were down-sampled to 128X128. The entire dataset was divided into the training set with 4,654 images and the testing set with 1,168 images.

In the experiment, BoF model with SIFT (Scale Invariant Feature Transform) descriptors was used to extract features that were then fed to linear SVM for image classification. This method was implemented by means of the VLFeat library [8]. The feature descriptors are unaffected by position, occlusion, illumination, the perspective of view, scaling etc.

The conventional powerful BoF model with the most robust and popular feature descriptor obtained the classification accuracy of 68% for training and 56% on test images.

The same dataset was used to train the four layered CNN, with three convolutional- pooling layers and one fully connected layer. ReLU activation function was applied to the CNN. The SGD (Stochastic Gradient Descent) with cross-entropy loss was used to minimize the loss and improve the accuracy of the CNN model. To prevent model overfitting, a dropout of 0.2 was applied to the third convolution-pooling layer and 0.5 was applied to the fully connected layer.

A dynamically updated learning rate was used, given by an exponential function of cost $\eta = \eta 0 \times \exp(C)$, where $\eta 0$ is set to 0.00l through trials and errors and C is the training loss.

The model was trained on GPU with keras library in the Spyder environment.

The CNN was initially trained without data expansion techniques with 100 epochs. The training accuracy obtained was 95% and the testing accuracy was 74%. The model encountered overfitting only after 10 epochs. This was due to limited images available for training.

To avoid issues of overfitting and to improve the accuracy, the model was trained by expanding the available dataset by applying transformations, such as rotation, scaling (horizontal and vertical), etc. The CNN model showed better performance with the expanded dataset and the model became more generalized i.e., issue of overfitting was completely eliminated. The obtained test accuracy after 100 epochs was 87%.

The training cycles were increased to 400 epochs and the obtained test accuracy was more than 90%.

III MOTIVATION

From the above study, it is clear that the CNN is the state of art for food image recognition and classification task. In comparison to the various supervised machine learning techniques, it is concluded that CNN performs better in terms of image classification tasks. The CNN automatically learns the features which are very helpful for food image classification.

Various datasets have been used for food image classification. In [9], Pittsburgh Fast-Food Image Dataset consisting of American fast food is used for food classification using the mobile phone for dietary assessment. In another experiment, Hoashi et al. [10] have used 85 Japanese food items for classification and achieved 62.5% accuracy. Some other work on food image classification is mentioned in section II.

Some observations made are as follows:

- Chinese, Japanese, American fast-food, etc., have been chosen for classification. But, the essence of Indian food is missing in these datasets.

- Various models have been deployed and tested to improve the accuracy of food image classification. All conclude that CNN is the best model for image classification task, particularly food image classification. Food image recognition and classification is more fine-grained than any other image recognition because of its unique textures, shape, size colours etc. Even same food may look different, because of its way of presentation.

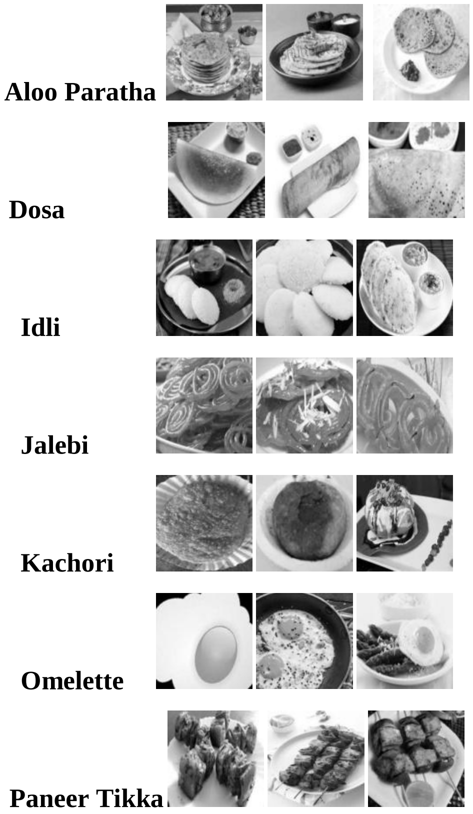
- All the previous work focuses on improving classification accuracy so that the models can be used for dietary assessment, identify calorie count in food, identify nutritional value of food etc.

IV. PROPOSED WORK

The inspiration to propose the CNN model for Indian food image classification arouse after studying the previous work done on food image classification with CNN and other techniques. The proposed work is closely related to the work done by Yuzhen Lu in [7]. The only dataset available on Indian food snacks is available on [12]. The dataset consists of Indian snack images belonging to 10 different categories. Due to the fewer images available, affine transformations are applied on the dataset to increase the number of images. The reason is that CNN requires more images to improve its accuracy. Another way is to use the pre-trained CNN such as AlexNet, GoogleNet, etc., to improve the classification accuracy.

The Indian snack dataset, after applying affine transformations consists of 60,000 grayscale images. The entire dataset is divided into the training set (50,000 images) and the testing set (10,000 images). Each image is of size 280X280 pixels. Since it is a grayscale image, 0 represents black and 255 represents white.

The ten classes of food include the following;



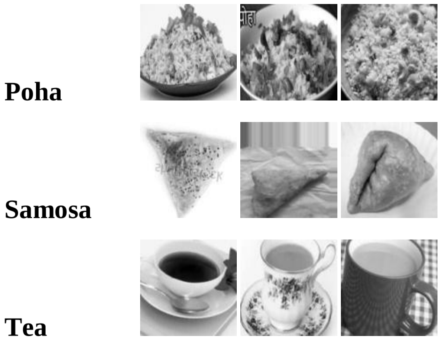


Fig. 3: Indian Food image dataset

The proposed CNN architecture consists of five convolution-pooling layers with one fully connected (dense) layer and an output layer with 10 neurons, representing 10 output channels. The non-linear ReLU operation is performed on each convolutional, pooling and the fully connected layer. The kernel size is set to 5X5 which remains unchanged for all the five convolutional layers. The window size for pooling layer is also fixed to 5X5. The input will be fed into the CNN through the input layer in a mini-batch of 400 images per batch. The model is trained on ***'Adam'*** optimizer with softmax classifier used in the fully connected layer, with the categorical cross-entropy cost function. The learning rate is set to 1e-03. A dropout of 0.8 is applied to the fully connected layer. The model is tested for its correct classification on the given input image.

The test accuracy of at least 95% is expected in just one epoch (Backpropagation algorithm).

The proposed work finds its application in Computer Vision, where robots can be used to serve as butlers in hotels. Many developed countries already have robots in service for hotels. Also, the proposed model can be optimized for use on mobile devices. The people just need to scan the food image using a smartphone and the model must correctly label the image. This is very useful for people trying food in another region or trying a new dish so that they know exactly what the particular dish is. People allergic to certain kind of food, vegetarians may find this mobile application a magical stick in their hands. The model will be trained on GPU, Ge-Force 1050Ti, 4GB card.

The model is coded in python and developed, trained and tested in Spyder environment using Keras deep learning library.

V. CONCLUSION

In this paper, the various food recognition and classification methods have been discussed in brief. The problem statement, various parameters are chosen, various dataset etc., for the various models have also been defined. The motivation for the proposed work is to classify the Indian food snacks and achieve better classification accuracy than the previous work done is also explained. The applications of the proposed model is also mentioned which can be the future work of the proposed system. The model accuracy is expected about 95% which shall be the best classification accuracy achieved so far using the DCNN architecture by applying affine transformations to the available limited dataset, containing grayscale images.

In the future, more datasets on Indian regional food images can be constructed consisting of colour images, as it is already studied in [3] the importance of colour in food image recognition and classification task. The more images available for training, the better the model will learn the features and better the accuracy will be. More convolutional-pooling layers can be added, altering the hyperparameters can also contribute to better classification accuracy.

ACKNOWLEDGEMENTS

I express my gratitude to our professor Mr K. Thippeswamy for motivation to write a technical paper. I am very thankful for my guide Mrs Shashi Rekha. H, VTU PG Centre, Mysuru for all the guidance and support in making this project successful and guiding me throughout the write up of paper as her research area is Deep CNN.

I equally thank Mr Shah Ayub Quadri, Data Scientist and Deep learning expert, Hyderabad, for helping in project development, imparting necessary knowledge related to machine learning and for making corrections in the paper.

REFERENCES

[1]. Abhishek Goswami, Haichen Liu, "Deep Dish : Deep Learning for Classifying Food Dishes". http://cs231n.stanford.edu/reports/2017/pdfs/6.pdf)

[2]. Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proc. of ACM UbiComp Workshop , 2014.* (http://ubicomp.org/ubicomp2014/proceedings/ubicomp_adjunct/workshops/CEA/p589-kawano.pdf*)*

[3]. Makoto Ogawa, Kiyoharu Aizawa, Hokuto Kagaya, "Food Detection and Recognition Using Convolutional Neural Network". (https://www.researchgate.net/publication/266357771)

[4]. M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp. Combining global and local features for food identi_cation in dietary assessment. In *IEEE ICIP*, pages 1789{1792, 2011.

[5]. Y. Kawano and K. Yanai, "Food Image Recognition Using Deep Convolutional Network with Pre-Training and Fine-Tuning". (http://ieeexplore.ieee.org/document/7169816/)

[6]. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[7]. Yuzhen Lu, "Food Image Recognition by Using Convolutional Neural Networks (CNNs)". (https://arxiv.org/abs/1612.00983)

[8]. A. Vedaldi and B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, in Proceedings of the 18th ACM international conference on Multimedia, 2008, pp.1469-1472.

[9]. M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. Pfid: Pittsburgh fast-food image dataset. In *IEEE ICIP*, 2009.

(http://ieeexplore.ieee.org/document/5413511/)

[10]. H. Hoashi, T. Joutou, and K. Yanai. Image recognition of 85 food categories by feature fusion. In *IEEE ISM*, pages 296{301, 2010. (https://www.researchgate.net/publication/321070597_Food_photo_recognition_for_dietary_tracking_system_and_experiment)

[11]. F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp. The use of mobile devices in aiding dietary assessment and evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 4(4):756{766, 2010. (https://dl.acm.org/citation.cfm?id=2986042)

[12]. http://github.com/NavinManaswi/IndianSnacks/tree/master/Indian Snacksdatasetand%20 code

BIOGRAPHIES

Shamay Jahan is pursuing her M. Tech., in Computer Science and Engineering, at VTU PG Centre, Mysuru. Her field of interest includes image processing, machine learning, etc. She is now developing the CNN model for Indian food Image Classification

Shashi Rekha. H, is working as an Assistant Professor at VTU PG Centre, Mysuru. Her research interests are Image Classification, Data Mining in E-Health, Pattern Recognition, etc. She is pursuing her research in Big Data Analytics.

Shah Ayub Quadri has completed his M. Tech., in Software Engineering. He is working as a Data Scientist in Hyderabad. His area of interests are Programming in Python, R, C# etc., Image Processing, Data Science, Machine Learning, Artificial Intelligence etc.