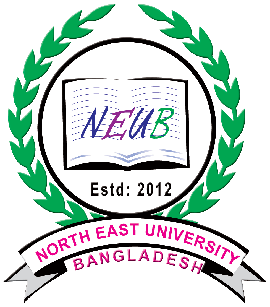
North East University Bangladesh

Department of Computer Science and Engineering



**Ingredients Generation from Food Image**

**By**

|  |  |  |
| --- | --- | --- |
| Jannatul Ferdous Jannah  Reg. No: 200103020003  BSc(Engg) in CSE  4th year 2nd semester | Mansura Mokbul  Reg. No: 200103020028  BSc(Engg) in CSE  4th year 2nd semester |  |

**Supervised By**

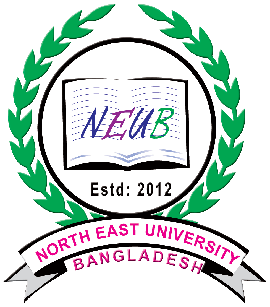
Razorshi Prozzol Talukder

Lecturer

Department of Computer Science and Engineering

29th November, 2023

**Ingredients Generation from Food Image**



A Project submitted to the Department of Computer Science and Engineering,

North East University Bangladesh, in partial fulfillment of the requirements  
for the degree of Bachelor of Science in Computer Science and Engineering

**By**

|  |  |  |
| --- | --- | --- |
| Jannatul Ferdous Jannah  Reg. No: 200103020003  BSc(Engg) in CSE  4th year 2nd semester | Mansura Mokbul  Reg. No: 200103020028  BSc(Engg) in CSE  4th year 2nd semester |  |

**Supervised By**

Razorshi Prozzol Talukder

Lecturer

Department of Computer Science and Engineering

29th November, 2023

**Recommendation Letter from Project Supervisor**

These Students *Jannatul Ferdous Jannah, Mansura Mokbul*, whose project entitled *“Ingredients Generation from Food Image.”,* is under my supervision and agrees to submit for examination.

Signature of the Supervisor ：

Razorshi Prozzol Talukder

Lecturer

Department of Computer Science and Engineering

North East University Bangladesh

**Qualification Form of BSc(Engg) Degree**

Student Name： Jannatul Ferdous Jannah, Mansura Mokbul

Thesis Title： Ingredients Generation from Food Image.

This is to certify that the thesis is submitted by the student named above in November, 2023. It is qualified and approved by the following persons and committee.

|  |  |
| --- | --- |
| **Head of the Dept.**  Rathindra Chandra Gope  Assistant Professor & Head  Department of CSE  North East University Bangladesh | **Supervisor**  Razorshi Prozzol Talukder  Lecturer  Department of CSE  North East University Bangladesh |

# Abstract

Food is a fundamental part of human experience. We like to collect our experiences through photographs, but these captured moments are more complex than they appear. We cannot demystify the creation of a dish by just looking at a food image. Therefore, in this paper, we introduce Recipes5k, a new large-scale, structured corpus of over one million cooking recipes. Using this data and a novel architecture, our system anticipates recipes of food images. The system uses language modeling to predict ingredients. It creates cooking instructions by conditioning on both inferred ingredients and its image. We postulate that this method will provide high-quality recipes by attending both images and ingredients.

**Keywords:** Image modelling, Recipes5k, Language modelling, Feedforward Neural Network, ResNet50.

# Table of Contents

Abstract i

Table of Contents ii

List of Figures iii

Chapter 1 INTRODUCTION 1

1.1 Dataset Explain 1

1.1.1 Dataset Structure 1

1.2 Problem Formulation 2

Chapter 2 Related Work 2

Chapter 3 PROPOSED METHOD 3

3.1 Methodology 3

3.1.1 Dataset 4

3.1.2 Data Pre-process 5

3.1.3 Feature Extraction using ResNet-50 6

3.1.4 Feedforward Neural Network Model…………………………………………7

3.1.5 User Interface…………………………………………………………….……8

Chapter 4 RESULTS AND DISCUSSION 9

4.1 Result 9

4.2 Lacking and Limitation 10

4.3 Conclusion………………………………………………..…………………….…….10

References 10

**List of Figures**

Figure 1: Methodology 3

Figure 2: Our dataset (101 folders) [5] 4

Figure 3: Our image dataset (folder-1) [5] 5

Figure 4: Our text dataset [5] 5

Figure 5: AfterPre-proceed (200x200) [5] 6

Figure 6: AfterPre-process vocabulary calculation [5] 6

Figure 7: ResNet-50 architecture for feature extraction 6

Figure 8: Shape of the extracted feature [5] 7

Figure 9: Feedforward Neural Network model architecture 7

Figure 10: Streamlit app to take users input 8

Figure 11: Predicted Output 9

Figure 12: Testing Accuracy 9

**Chapter 1**

# INTRODUCTION

Food goes beyond mere sustenance; it acts as a universal language, uniting people worldwide. The internet and social media have transformed the way we share global food experiences, painting a rich tapestry of culinary stories through images. Each dish encapsulates a unique narrative, featuring diverse ingredients, preparation techniques, and cultural influences.

In the context of this study, we introduce a groundbreaking system capable of generating detailed food ingredients from images. Envision the ability to unravel the intricate details of a dish's ingredients simply by observing a picture. Our system accomplishes this feat through a sophisticated approach that takes into account both the image and the nuanced array of ingredients involved. This innovative methodology aims to reshape how we comprehend and share food experiences within the expansive and diverse realms of culinary culture.

## Dataset Explain

The Recipes5k is a dataset designed for ingredient recognition. It consists of 4,833 images, and 4,833 sentence, where each sentence is the combination of multiple ingredients. each composed of an image and the corresponding list of ingredients.

* + 1. **Dataset Structure**

The Recipes5k dataset appears to be designed for ingredient recognition, containing both images and corresponding lists of ingredients. Let's break down the key components of the dataset structure based on the information provided:

1. **Number of Samples:**The dataset consists of a total of 4,833 samples.
2. **Data Format for Each Sample:**Each sample in the dataset includes two main components: an image and a corresponding list of ingredients.
3. **Images:**There are 4,833 images in the dataset. These images are likely related to food items or dishes.
4. **Sentences:**Each sample is associated with a sentence that represents a combination of multiple ingredients. These sentences serve as textual annotations for the images.
5. **Ingredients:**The sentences contain lists of ingredients that are associated with the corresponding images. Each sentence is composed of multiple ingredients, and these ingredients presumably describe the contents of the corresponding image.
6. **Data Integration:** The combination of images and ingredient lists provides a multimodal dataset, where information from both visual (image) and textual (ingredient list) modalities is available for each sample.

In summary, the dataset is structured to facilitate training and evaluation of models for ingredient recognition tasks. The combination of images and corresponding ingredient lists allows for the development of models that can understand and predict the ingredients present in a given dish or food item based on visual information.

* 1. **Problem Formulation**

The dataset of Recipe5k is quite large; our GPU system could not process all the images together, and the server kept crashing. To solve that, we had to minimize our dataset.

And some foods look a lot alike, which makes it tricky for the system to tell them apart. To solve that, we made our system super accurate, using clever tricks to help it distinguish foods.

**Chapter 2**

# Related Work

The large scale food datasets such as Recipe1M [1], food 101[2], have developed significant advancement in visual food recognition using machine learning. Many studies are currently being conducted on food image classification, nutritional value extraction from the ingredients and image of the food, and recipe and ingredient extraction from the food image. Food- related tasks are also being considered in Natural language processing due to recipe generation tasks that come under language modeling. There have been many types of research done on extracting recipes from the food images using semantic segmentation, GAN method, and cross model embedding, which retrieves the similarities between two modalities. Learning common feature subspace is currently the mainstream of research. In Natural language processing, sequence to sequence modelling has played an important role in converting a sequence of one domain to another. The transformer model like ‘attention is all you need’ [4], has achieved great performance accuracy. In this project, this approach is utilized to extract ingredients and title from the image domain.

**Chapter 3**

# PROPOSED METHOD

Our methodology aims to predict a title and its ingredients from food images. This system predicts titles and generates ingredients by understanding both the images and ingredients.

## Methodology

## 

**Figure 1:** Methodology

### Methodology Details

**Dataset:** We Obtain our dataset of Recipes5k from kaggle.

**Pre-process**: Load images, labels, and ingredients. Pre-process images and tokenize titles and ingredients.

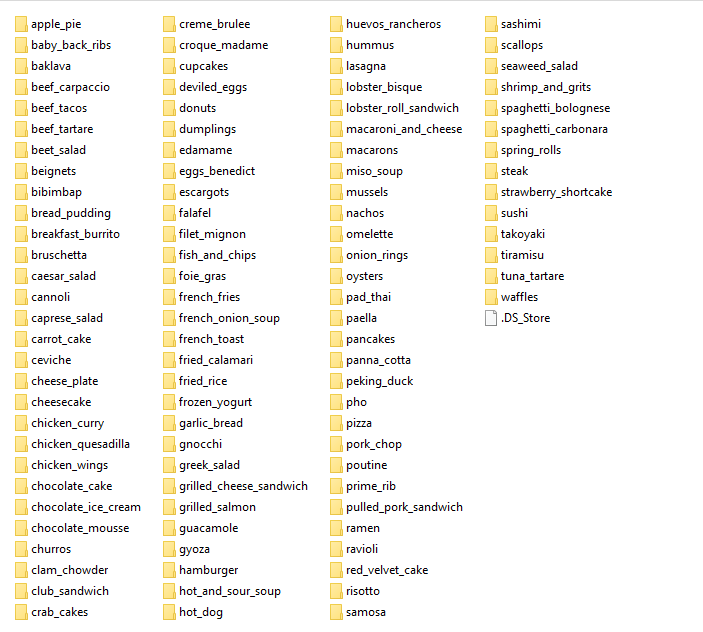
**Feature Extraction:** We used a pre-trained ResNet-50 model to extract features from images.

**Model:** Define the model for predicting titles and ingredients.

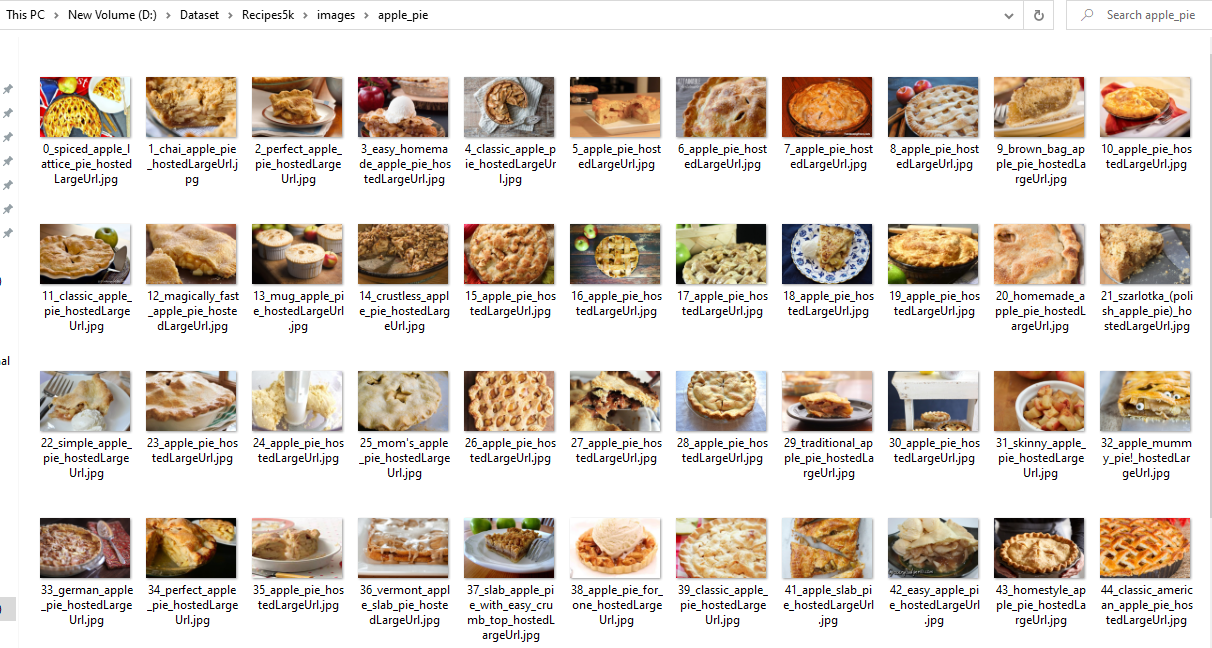
**User Interface:** Implement a Streamlit app for user interaction and prediction display.

### Dataset

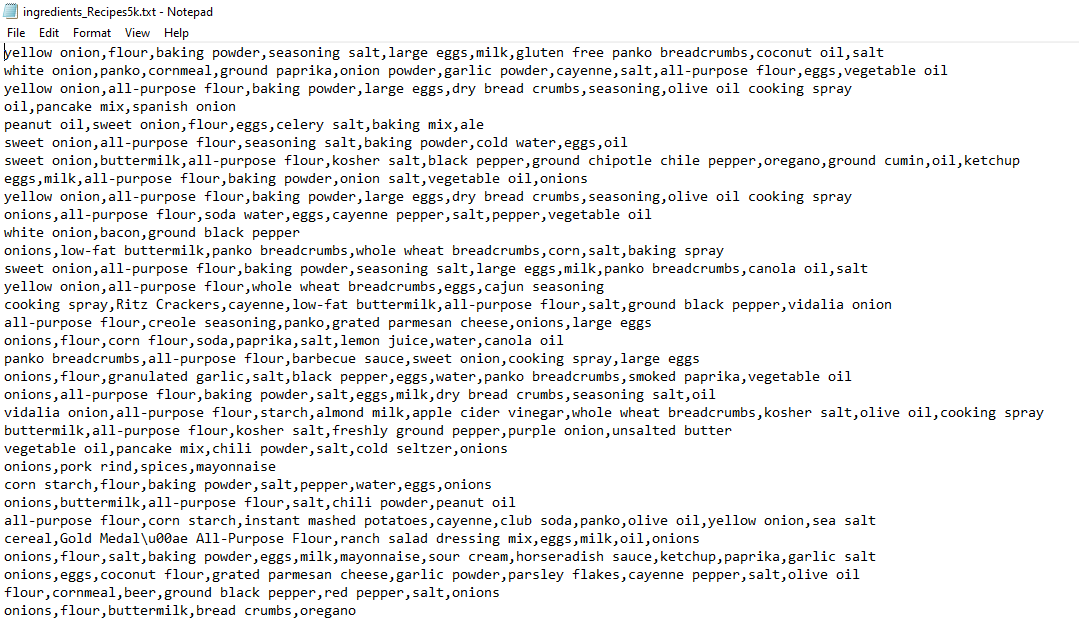
Our dataset is combination of images and ingredients. There are 101 subfolders under the Recepie5k folder where each subfolder represents different categories food each with multiple images and a text file for the ingredients.



**Figure 2:** Our dataset (101 folders) [5]



**Figure 3:** Our image dataset (folder-1) [5]



**Figure 4:** Our text dataset [5]

### Data Pre-process

Load the images along with their labels, and ingredients after resizing them to 224 x 224 pixels. Where ingredients are split and stored in a list, and titles are derived from the directory structure. Then tokenization is performed on both titles and ingredients. Later vocabulary sizes are calculated for titles and ingredients separately based on the unique tokens present in the tokenized lists.



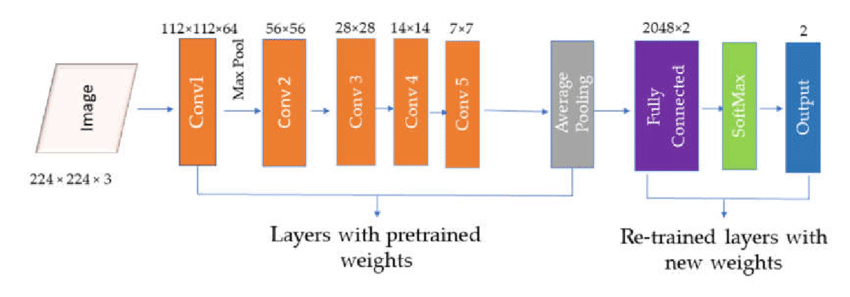
**Figure 5:** AfterPre-proceed (200x200) [5]



**Figure 6:** AfterPre-process vocabulary calculation [5]

* + 1. **Feature Extraction using ResNet-50**

ResNet-50 is a deep convolutional neural network architecture with 50 layers. By featuring residual blocks that employ skip connections to enable the training of deep networks. and it is commonly used for image classification tasks.



**Figure 7:** ResNet-50 architecture for feature extraction

**ResNet-50 Architecture:**

**Input:** ResNet-50 takes an input image of size 224x224 pixels.

**Convolutional Layers**: The network starts with a series of convolutional layers, including batch normalization and ReLU activation functions.

**Residual Blocks**: The unique feature of ResNet is the residual blocks. These blocks contain shortcut connections that skip one or more layers, helping with the training of very deep networks.

**Pooling Layers**: Spatial pooling (average pooling) reduces the spatial dimensions of the feature maps.

**Fully Connected Layer**: The network concludes with a fully connected layer for classification. But since we are interested in the features, not the final classification we removed this layer.

**Output:** The final output is a feature vector that represents the image content. Where the extracted features are then converted to a NumPy array and labels (titles and ingredients) are mapped to numerical indices for model training using dictionaries.

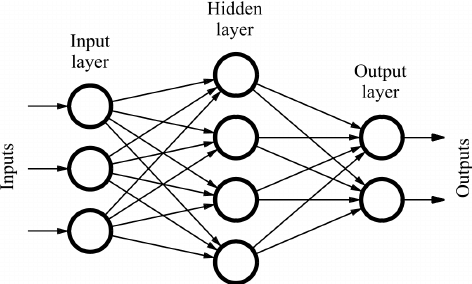
After loading the pre-trained ResNet-50 model extract the features:



**Figure 8:** Shape of the extracted feature [5]

**3.1.4 Feedforward Neural Network Model:**

A Feedforward Neural Network model consists of three linear (fully connected) layers. It takes an input of size, applies a linear transformation with a ReLU activation function, and then outputs two sets of predictions: one for titles and one for ingredients. Each set of predictions is obtained from a separate linear layer.



**Figure 9**: Feedforward Neural Network model architecture

**Feedforward Neural Network Architecture:**

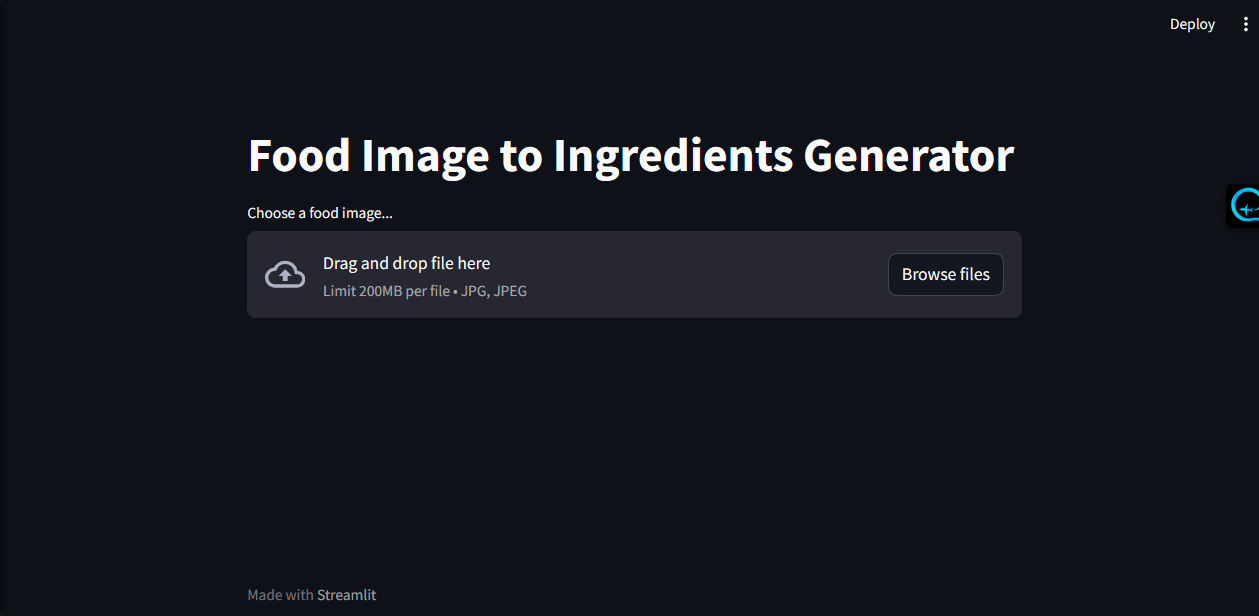
**Input Layer:** The input features are the extracted image features from the ResNet-50 model. These features serve as the input to our feedforward neural network.

**Hidden Layer:** There is one hidden layer with a ReLU activation function. Where the hidden layer applies a linear transformation to the input features. And then The ReLU activation function is applied element-wise to introduce non-linearity to the model.

**Output Layer:** There are two output layers, one for predicting titles and one for predicting ingredients. Each output layer is connected to the hidden layer.

**3.1.5 User Interface**

We implement the Streamlit app so that users can upload a food image and then the app can display the uploaded image along with the generated predictions for the title and ingredients.



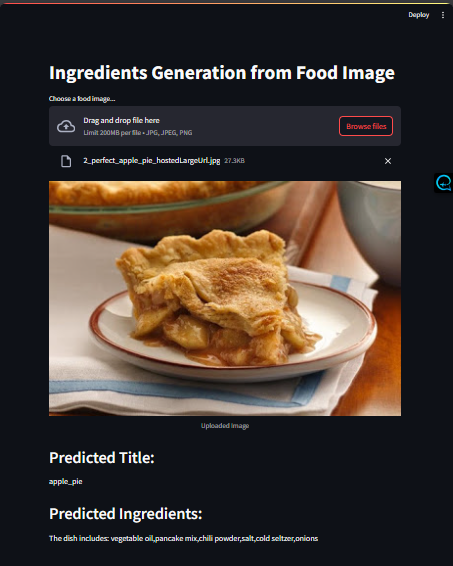
**Figure 10:** Streamlit app to take users input

**Chapter 4**

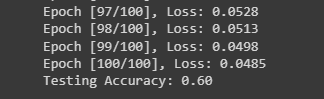
# RESULTS AND DISCUSSION

**4.1 Result**

When a user uploads a food image, this system will predict the closer title, and ingredients for the input image.



**Figure 11:** PredictedOutput



**Figure 12:** Testing Accuracy

**4. 2 Lacking & Limitation**

1. **Similar Color, Size, and Shape:**The system may struggle to distinguish between ingredients that share similar visual attributes, such as color, size, and shape. For instance, if two ingredients have similar shades of color, it might be challenging for the system to accurately differentiate between them.
2. **Limited Training Data:**

Training a system to accurately identify ingredients from images requires a vast and diverse dataset. Limited access to high-quality, diverse food images could hinder the system's ability to generalize across different cuisines and dishes.

1. **Ingredient Combinations:**Dishes often contain complex combinations of ingredients, and the interactions between them can influence the final flavor. Capturing these interactions accurately through image analysis alone can be challenging.
2. **Accuracy:**Test accuracy for the title is average but accuracy for ingredients is quite low. Better model such as Transformer model might be a good solution for better accuracy.

# 4.3 Conclusion

In this project we proposed a system that can generate ingredients of food items from their images which is given as an input. First ingredients are extracted from the images and then these predicted ingredients and title combine displayed. Our system highlights the importance of reasoning about both modalities simultaneously. It is hard to achieve good accuracy as compared to human based ingredients from the machine learning trained model. But our model is giving a convincing result.

# References

1. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and FoodImages.
2. Kaggle Food 101, pictures of 101 types of foods.
3. Guhe, D., Patil, S., Ladkar, S., Bhatia, S., & Shrawne, S. (2021). "Recipe Prediction from Food Images using Transformer Model". International Research Journal of Engineering and Technology (IRJET).
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). “Attention Is All You Need”. [arXiv preprint arXiv:1706.037621](https://arxiv.org/abs/1706.03762)
5. Our dataset available:

<https://drive.google.com/drive/folders/1Q_QBqev3MFba8nL8f75CwpxowIP5Y4__?usp=drive_link>