Recipe Prediction from Food Images using Transformer Model

Dipika Guhe[1], Samruddhi Patil[2], Samisha Ladkar[3], Simran Bhatia[4], Seema Shrawne[5]

[1,2,3,4] Student Researcher , Veermata Jijabai Technological Institute, Maharashtra, Mumbai
[5]Professor, Dept. of Computer Engineering, Veermata Jijabai Technological Institute, Maharashtra, Mumbai

---***---

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 Recipe1M, 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.*

Key Words: Image modelling, Resnet-50, Language modelling, Encoder-decoder transformer model, Attention mechanism.

1. INTRODUCTION

In every culture of human existence food plays a vital role. Food is what brings together people; overcoming all barriers of language, geographical differences, visual differences. We enjoy discussing all aspects of food as much as we enjoy consuming it. With the advent of the Internet and social media this discussion has gone global. Today food images from a large part of the shared repertoire on social media and search engines. The food items that these images represent have different make, different preparation strategy and even different geographical belonging. To be able to access the recipe of a dish from its image would revolutionise how the food industry functions. With new food trends emerging every year, for example veganism, keto diets, paleo diets it has become critical to understand how the dish was prepared.

In light of the foregoing, we propose in this study a system that can develop food recipes when given food images. It uses an architecture that builds upon a joint embedding of both images and recipes. The recipes consist of two parts- an unordered set of ingredients and a list of instructions.

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[3], nutritional value extraction from the ingredients and image of the food[5], and recipe and ingredient extraction from the food image[6,7]. 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[7], GAN method[6], and cross model embedding, which retrieves the similarities between two modalities. Learning common feature subspace is currently the mainstream of research[4].

In Natural language processing, sequence to sequence modelling has played an important role in converting a sequence of one domain to another[8]. The transformer model like 'attention is all you need'[9], has achieved great performance accuracy. In this paper, this approach is utilised to extract ingredients and recipe instructions from the image domain.

3. DATASET STRUCTURE

Recipe1M dataset is the biggest publicly available recipe dataset. The information each recipe contains is separated in two JavaScript Object Notation files.

The contents of the dataset are logically grouped into two layers. The first contains basic information including title, a list of ingredients, and a sequence of instructions for preparing the dish; all of this data is provided as free text. The second layer builds on the first by including any image related with the recipe, which are provided as RGB JPEG files. The id of each image defines the location of the image on disk.

Det_ingrs.json file contains the ingredients of all the recipes. Each of the documents in the json file has it's unique id which is used for mapping of ingredients and instructions. The 'valid' key indicates whether the corresponding ingredient text data is valid or not.

The layer1.json and det_ingrs.json files are used to create the vocabulary of words for instructions and ingredients respectively. Recipe1M also contains nutritional information on a subset of about 51k recipes. This dataset, combined with data scraped from various websites on vegetarian/vegan ingredients, was used to preprocess the recipes to create the nutritional labels. Nutritional information (i.e Total energy, protein, sugar, fat, saturates, and salt content) is only added for those recipes that contain both units and quantities. There are 50,637 recipes with nutritional information.

4. PRE-PROCESSING

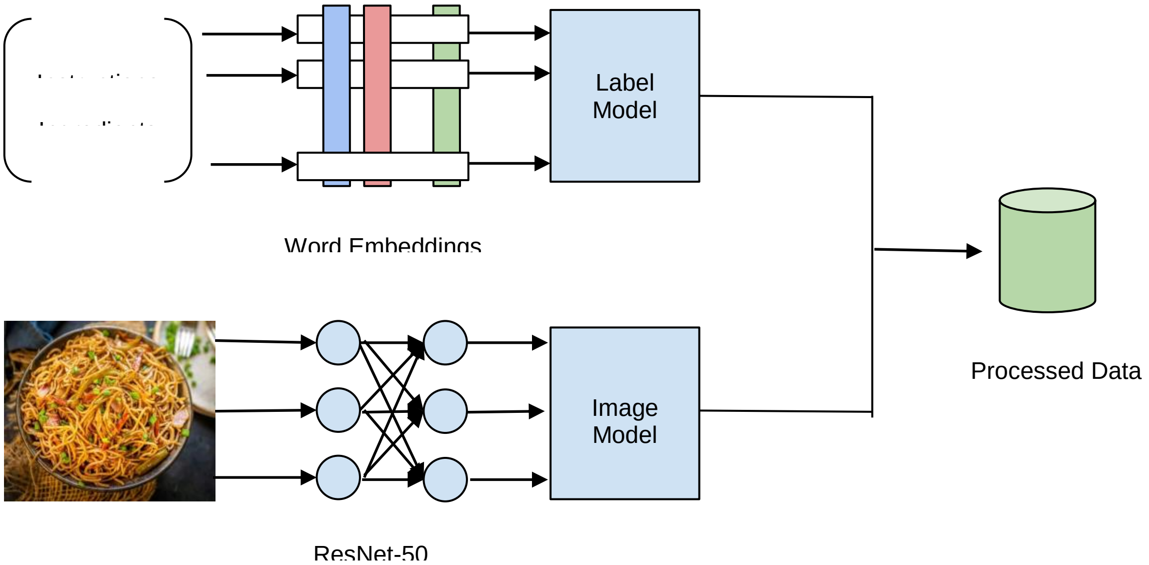


Figure 1: Preprocessing of instructions,ingredients and images to build processed data from Recipe1M Dataset

4.1 Label Modelling

The raw data obtained from Recipe1M[1] dataset includes two types of data- recipes that are in the form of text data and food images in image format. The Raw data is first divided into train, test and validation subsets. Since the recipes are obtained from different online websites they lack standardisation hence certain transformations are applied to them to bring them to a unique format.

The transformations applied are :-

1. Replacement of special characters such as punctuation marks - { } ; + , [ ] %
2. Clustering of ingredients that belong to similar categories such as different types of breads - brown, white, whole wheat, pita, focaccia are all grouped together and named as just bread.
3. Adding special tokens <start> and <end> indicating the start and end of the recipe as well as each individual instruction.

After the above basic transformations have been applied the next step is to process the two parts of the recipe- ingredients and instructions separately. For the ingredients plural words are converted to singular such as onions is written as onion. Any ingredient that occurs less than 10 times is discarded. For cooking instructions words with less frequency are replaced with the unknown word token and the sentences are tokenized. In the end we obtain a set of processed Vocabulary containing 1488 ingredients and 23231 words in instructions.

Both the vocabularies are combined using one hot encoding to generate a lookup table in the form of a 2D tensor which is permuted, reshaped and linearly transformed to produce a 1D tensor which is then sent as an input to a label encoder thus establishing the basis for a label model.

4.2 Image Modelling

In this subsection, we try to get features from a given input image and train it to our requirements. Layer 2 photos are pre-processed by applying several modifications to make them 256 X 256 pixels in size. Features are extracted from the processed images. These features are fed to a pretrained model Resnet50 [11] whose last layer has been removed and a new transfer model is created. This model is optimized using adam optimizer[10] with a learning rate of 0.01. These two models generated in section 4.1 and 4.2 are combined together to predict ingredients and instructions discussed in section 5.1 and 5.2.

5. RECIPE GENERATION

5.1 Recipe Generation Model

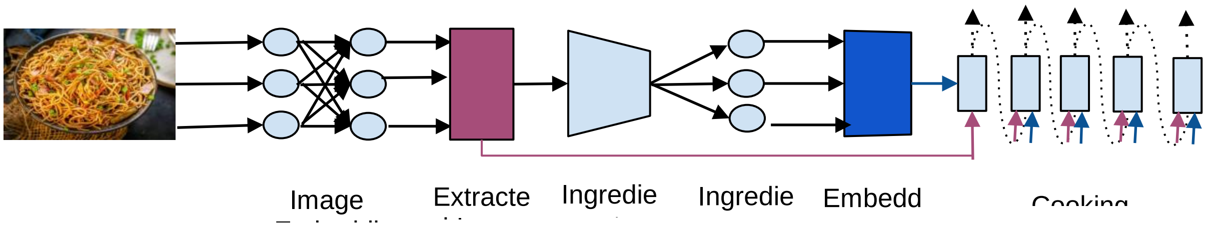


Figure 2: Embedding the food image to extract features which are utilised by ingredient Prediction transformer model. The inferred ingredients are again encoded to be processed by Instruction prediction transformer model along with previously embedded image features.

Recreating a recipe means to unravel the preparation process from just raw ingredients to a meal. The process transforms ingredients by various activities like cutting, blending, and then cooking. Therefore, we demonstrate our system to generate sequenced cooking instructions not only on food images but also the already extracted ingredients simultaneously.

The system is predicated on the Transformer model which leverages encoder-decoder architecture with attention mechanisms. The Recipe Generation takes place in two parts :

1) Ingredient prediction 2) Cooking Instruction Prediction. In subsection 5.2, we demonstrate how we extract features from food images and further encode them to predict ingredients. In subsection 5.3, we look at multimodal embedding of images and deduced ingredients in order to decode instructions.

5.2 Ingredient Prediction

In this subsection, we concentrate at the intermediate step of predicting the ingredients list in a recipe generation process. We extract the image representation with a ResNet-50 encoder and obtain the ingredient embedding by means of a Transformer architecture to predict ingredients.

The ingredients are represented in a set structure. A set of ingredients is a variable sized, unordered collection of unique meal constituents. The ingredients in a set are not always independent, e.g. salt and pepper frequently appear together. Moreover, permuting them does not affect the ultimate output. Therefore, we model the dependency between ingredients without imposing any order and predict as a set of ingredients rather than a list (which enforces order).

We formulate the ingredient prediction problem to a set prediction problem. We train a transformer network to predict ingredients given an image by minimizing the negative log-likelihood loss. We train the same transformer by randomly shuffling the ingredients (thus, removing order from the data).

For ingredient prediction, we use a transformer with 4 blocks and 2 multi-head attentions, each one with dimensionality of 256. To obtain image embeddings we use the last convolutional layer of the ResNet-50 model (Image model). Both image and ingredient embeddings are of dimension 512. We keep a maximum of 20 ingredients per recipe.

5.3 Instruction Prediction

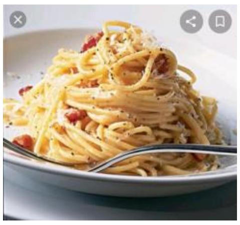
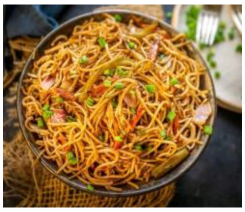
In this subsection, we concentrate on predicting recipe instructions from the image embedding obtained from the Resnet-50 model and ingredient embedding obtained from label modeling.

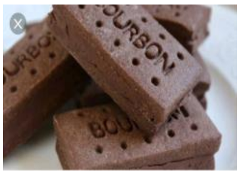
We used a concatenated attention mechanism to combine image and ingredient embeddings. This strategy first concatenates both image and ingredient embeddings with a dimension of vector as 512. Then the attention is applied to the combined embedding. This combined embedding and the instruction embedding are provided as input to the same transformer model used in ingredient prediction. The instructions are predicted based on the concatenated vector and previously predicted instructions.

The model is trained on the basis of the likelihood of maximization on the log of probability of similarity between image ingredient pair and instruction. The first instruction retrieved is the title of the recipe followed by other cooking instructions. For instruction prediction, we used a transformer with 16 blocks and 8 multi-head attentions, each one with dimensionality 64. The maximum number of words in the instruction are restricted to 150. All other dimensions of image and ingredient embeddings are the same as the ingredient prediction i.e. 512.

6. RESULTS

The output of the model is shown in the table below. The testing of the model is executed using human based outputs. We compare the original human written recipes and the predicted recipe and find the similarities between the two. As using sequence to sequence modelling it is hard to achieve perfect accuracy, our model is able to give convincing results.

	Title: Spaghetti carbonara Ingredients: cheese, pepper, egg, oil, spaghetti, clove, bacon, salt, butter, parsley Instructions: 1. Cook the spaghetti according to package directions. 2. Meanwhile, in a large skillet, cook the bacon over medium heat until crisp. 3. Remove the bacon from the skillet and drain on paper towels. 4. Reserve 1 tablespoon of the bacon drippings in the skillet. 5. Add the butter and garlic to the skillet and cook over medium heat until the garlic is golden brown. 6. Add the cooked spaghetti and toss to coat. 7. Add the reserved bacon drippings and toss to coat. 8. Add the eggs, parmesan cheese, salt and pepper and toss to combine. 9. Serve immediately.
	Title: Ramen noodles Ingredients: oil, soy_sauce, pepper, onion, clove, noodles, sugar Instructions: 1. Cook noodles according to package directions. 2. Drain and set aside. 3. In a large skillet, heat oil over medium heat. 4. Add garlic and red pepper flakes. 5. Cook for 1 minute. 6. Add green onions and cook for 1 minute. 7. Add noodles and soy sauce. 8. Cook for 2 minutes.

	Title: Chocolate cake Ingredients: sugar, flour, cocoa, salt, egg, butter, baking_powder, milk, extract Instructions: 1. Preheat the oven to 350 degrees f (175 degrees c). 2. Grease and flour a 9 inch square pan. 3. In a large bowl, cream together the butter and sugar until light and fluffy. 4. Beat in the eggs one at a time, then stir in the vanilla. 5. Combine the flour, cocoa, baking powder and salt; stir into the creamed mixture alternately with the milk. 6. Pour batter into the prepared pan. 7. Bake in a preheated oven for 30 to 35 minutes, or until a toothpick inserted into the center of the cake comes out clean.

7. CONCLUSION

In this paper we proposed a system that can generate recipes 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 image features are jointly embedded to get the recipes. Our system highlights the importance of reasoning about both modalities simultaneously. It is hard to achieve good accuracy as compared to human based recipes from the machine learning trained model. But our model is giving a convincing result as compared to other image to recipe prediction models.

ACKNOWLEDGEMENT

We thank Ms. Seema Shrawne for her invaluable assistance and direction during the project's development. We'd like to express our gratitude to Veermata Jijabai Technological Institute for providing access to necessary equipment and setup for training and testing of the model. Thank you to the VJTI Centre of Excellence and our peers for sharing their perspectives on difficult decisions.

REFERENCES:

[1] Javier Mar´ın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, Antonio Torralba "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and FoodImages Massachusetts Institute of Technology Qatar Computing Research Institute, HBKU Universitat Polit `ecnica de Catalunya.

[2] Kaggle Food 101, pictures of 101 types of foods.

[3] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, Xiaoou Tang "Residual Attention Network for Image Classification"

[4] Alexander Mordvintsev, Software Engineer, Christopher Olah, Software Engineering Intern and Mike Tyka, "Inceptionism: Going Deeper into Neural Networks"

[5] Parisa Pouladzadeh, Shervin Shirmohammadi, Rana Almaghrabi "Measuring Calorie and Nutrition from Food Image" Distributed and Collaborative Virtual Environment Research Laboratory University of Ottawa, Ottawa, Canada.

[6] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, Yanbin Hao "GAN: Cross-modal Recipe Retrieval with Generative Adversarial Network" City University of Hong Kong

[7] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, Antonio Torralba "Learning Cross-modal Embeddings for Cooking Recipes and Food Images" Universitat Polit`ecnica de Catalunya Massachusetts Institute of Technology.

[8] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, Lukasz Kaiser "Multitask sequence to sequence learning " Google brain.Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun. "Deep Residual Learning for Image Recognition" arXiv:1512.03385, 2015

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Llion Jones, Łukasz Kaiser, Illia Polosukhin "Attention is all you need " Google Brain.

[10] Diederik P. Kingma, Jimmy Ba "Adam: A Method for Stochastic Optimization". arXiv preprint arXiv: 1412.6980, 2014.

[11] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun. "Deep Residual Learning for Image Recognition" arXiv:1512.03385, 2015