Image-to-text based Recipe Recommendation System

Devender
Indraprastha Institute of Information Technology Delhi
devender18334@iiitd.ac.in

Mohammad Sajeel Khan
Indraprastha Institute of Information Technology Delhi
sajeel18342@iiitd.ac.in

Nischal Garg
Indraprastha Institute of Information Technology Delhi
nischal18350@iiitd.ac.in

1. MOTIVATION AND PROBLEM STATEMENT

1.1 Motivation

The world is a collection of dialects and languages, making it challenging for one language's trained technologies to properly assess other languages' queries and make predictions based on them. The standard query input, "text", also requires a certain level of literacy on the user's part. So, we propose a visual input query for our IR system that aims to eliminate these problems.

1.2 Problem Statement

As the world, with each passing day, is coming closer than ever, we think that a visual approach to problems around us would tremendously help as it is least prone to error and requires the least effort on the user's part. We would use a visual-based approach to solve ingredient identification to recommend recipes made from these identified ingredients.

1.3 Task

We plan to make an image-to-text IR system. The job would be to correctly identify the ingredients from various transfer learning methods such as pre-trained CNNs(like ResNet, GoogleNet etc.) in forms of a string. After the procurement of these strings, we would form a suitable query for the IR model to obtain the summarised form of the recipe that these identified ingredients could make. We also plan to incorporate the calorie calculation of the recipe along with its summarised form.

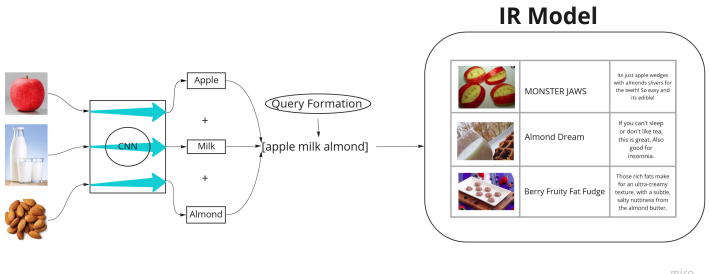

Figure 1: Working flow of the task

1.4 Dataset

The recipe recommendation will be based on several public datasets available such as Recipe1M, Recipe Box, Food101 or will be live-crawled on websites such as allrecipes.com For the image classification of the ingredients, we will have to create our food ingredients dataset by combining various Kaggle datasets of food images(fruits, dairy, vegetables, nuts etc.)

2. LITERATURE REVIEW

1. Cross-Modal Recipe Retrieval: How to Cook this Dish? [1]
   This paper focuses on recipe retrieval based on the image of a dish using cross-modality analysis. The proposed model of Stacked Attention Network (SAN) is trained on a joint space of image and recipe pairs. The learning occurs at the region level for an image and ingredient level for a recipe.
   - Dataset - 61,139 image-recipe pairs crawled from several websites and covered different kinds of food, mainly Chinese Cuisine.
   - Evaluation Criteria- The authors used two metrics for performance evaluation - Mean Reciprocal rank(MRR) and Recall at Top-K(R@K).
   - Result Analysis - Compared to DeViSE, the SAN model showed a relative improvement of 130% in MRR and higher R@K.
   - Limitations - In the cases where some ingredients, especially key ingredients, are unknown, the model fails to retrieve relevant recipes.
2. Recipe Recognition with large Multimodal food dataset [2]
   This paper explores several vision-based and text-based technologies for image recipe recognition. Several classification algorithms are run using visual information, textual information, and fusion to classify the dish. For any query image, the result is a ranking of the three best categories (along with links to recipe webpages) with a positive score for correct match and a negative score for the incorrect match.
   - Dataset - UPMC Food-101, containing about 100,000 recipes for a total of 101 food categories collected from the web

- Result Analysis - The best validation accuracy of 85.10% is achieved by using fusion features(TF-IDF and OverFeat CNN deep features)

3. Intelligent Menu Planning: Recommending Set of Recipes by Ingredients [3]
   This paper recommends menus to the user by ingredients given by the user as input. The recipe menu is collected from food sites and then a recipe relevance graph is formed with the graph node as the recipe name and the ingredients in it.

   - Dataset: The recipes were collected from food.com, and there were 226,025 total recipes and a total of 5,073 kinds of ingredients in all the recipes.
   - Evaluation Criteria : The authors used two evaluation metrics Tag Entropy and Tag Co-Occurance Density.
   - Recipe similarity is calculated using Jaccard Similarity. If two recipes have the same ingredients, then they are termed equivalence with one another.
   - Once the recipe relevance graph is formed, the menus are generated using Connector-Steiner Tree Algorithm and the set of recipes are given to the user.

4. Recipe Retrieval Using Ingredient Recognition [4]
   The paper proposes a recipe retrieval system which involves object recognition through images as input query as opposed to the classic text-based input system. The user simply clicks pictures of the ingredients and the recipes are retrieved based on a certain ranking system. The proposed system solves the problem of language barrier. Besides, the unstructured food images on the web can make recipe retrieval harder.

   - The process begins with the user feeding ingredient images as input. The images will then be matched with the dataset that was created manually using feature extraction.
   - The system utilizes data classification algorithm KNN – to retrieve recipes that include identified ingredients.
   - The list of retrieved recipes is sorted based on the percentage of ingredients that match with the input ingredient set.

5. Images & Recipes: Retrieval In The Cooking Context [5]
   The paper deals with the problem of alignment of images and recipe. It focuses on efficient techniques of retrieval between the components i.e. recipe texts and cooked dish images. The paper carries out two independent deep neural networks which is aimed to learn representation of the recipe texts and images of the cooked dishes.

   - The model is trained by "cross-modal alignment" which ensures that the representation of a recipe and its corresponding dish image is similar while non-corresponding pairs are dissimilar.
   - The paper discusses the model and its capabilities when dealing with traditional problems in the field of computational cooking. It analyzes several downstream tasks and evaluates its effectiveness through quantitative analysis.
   - The dataset that was used to train the models was Recipe1M which comprises over 1 million pairs that have aligned text documents corresponding to relevant dishes with matching images.

6. Recipe Retrieval with visual query of ingredients [6]
   The paper proposes a method wherein the input queries are not based on text but on images. Textual input often results in loss of information due to inaccurate description or the lack of knowledge on part of the user. The proposed system will use the images of ingredients as input queries which will be passed through a convolutional neural network to retrieve cuisine images.

   - The dataset used was Recipe1M which has over 10 lakh structured recipes and over 8 lakh associated images collected from multiple cooking websites. Images were used from Google in case they were missing in the original dataset.
   - The user inputs ingredient images to the system. Convolutional Neural Networks are then used to compute the matching score between ingredients and the learned representations.
   - For the purpose of evaluation, median retrieval rank and recall at top K are used.

   Thus the paper's contributions are significant and multifold. It proposes an image-to-image recipe retrieval system. It adds data to expand an already existing dataset. Most importantly, it proposes a neural network for the proposed task.

3. PROPOSED METHOD

We propose a method of identifying the food items through images by training a Convolutional Neural Network(CNN) and forming a query from the textual outputs of the CNN. And then feed the formed query to the Information Retrieval system that would have the recipes from the popular food-recipes dataset such as Recipe1M and Food101. In short, we are creating an image-to-text retrieval system, where users would get the textual information(recipe documents) from the images of the food items from which they would like to create a recipe.

4. BASELINE RESULTS

For the purpose of achieving baseline results and applying our methodology, we have worked with a reduced dataset of about 25,000 recipes. Currently, we have prepared our model to handle only the ingredients that are either fruits or vegetables. The current set of identifiable ingredients includes 13 different types of fruits and 12 different types of vegetables. A total of 15 items have been used for each entity, of which 12 have been used for training and 3 have been used for testing. We fine-tuned the pre-trained Googlenet model on 300 images and in over 100 epochs achieved the best validation accuracy of 78.66% on 75 images. The progress of the project can be looked at in our **Github** repository.

Using the convolutional neural network, we have now obtained the name of the ingredient in text format and have hence achieved image-to-text conversion of ingredients. We will now be using the text format of ingredients as keywords to perform a boolean search on the documented set of about 25,000 recipes. The recipes dataset is taken from Recipe Box which contains structured recipes scrapped from food websites. To achieve baseline results, we will use the 'AND' operator between all the ingredients to maximize ingredient utilization.

5. FUTURE WORK

We aim to build on our self-compiled dataset to incorporate more fruits and vegetables. Apart from that, we aim to move beyond the scope of fruits and vegetables, to add variety to our identifiable ingredients set. We can add more dairy products such as milk, curd and paneer. We also wish to add different types of nuts, pulses, rice, meat etc. The images being used currently to train our model are lab images even though in practice, the system will have to evaluate user-clicked images. Hence, we not only aim to expand our dataset with more ingredients and images but also wish to do that using our own photos to achieve higher accuracy. We also aim to explore beyond boolean retrieval methods to make our information retrieval system more robust and industry-grade that would provide the most significant recipes.

APPENDIX

A. ADDITIONAL AUTHORS

Additional authors: Sandeep Singh (Indraprastha Institute of Information Technology Delhi, email: `sandeep18363@iiitd.ac.in`) and Shekhar Shukla (Indraprastha Institute of Information Technology Delhi, email: `shekhar18414@iiitd.ac.in`).

B. REFERENCES

[1] Chen, J., Pang, L. and Ngo, C., 2018. Cross-modal recipe retrieval with stacked attention model. Multimedia Tools and Applications, 77(22), pp.29457-29473.

[2] Wang, X., Kumar, D., Thome, N., Cord, M. and Precioso, F., 2015, June. Recipe recognition ith large multimodal food dataset. In 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 1-6). IEEE.

[3] Kuo, F., Li, C., Shan, M., & Lee, S. (2012). Intelligent menu planning. Proceedings of the ACM Multimedia 2012 Workshop on Multimedia for Cooking and Eating Activities - CEA '12. doi:10.1145/2390776.2390778

[4] Nerkar, M., Vajpayee, P., Ganjoo, V., Suryawanshi,, V. and Sonawane, P., 2017. RECIPE RETRIEVAL USING INGREDIENT RECOGNITION. International Journal of Advance Engineering and Research Development, 04(5).

[5] Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N. and Cord, M., 2018, June. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 35-44).

[6] Lien, Y. C., Zamani, H., Croft, W. B. (2020, July). Recipe Retrieval with Visual Query of Ingredients. In Proceedings of the 43rd International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 1565-1568).

[7] Food images are taken from **Fruit 360** on Kaggle

[8] Recipes taken from **Recipe Box**