IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

Survey on Inverse Cooking: Recipe Creation from Food Images

Ashwini Hegde[1], Rabia Ishrath[2], Ranjitha Y[3], Yashaswini N[4], Sandesh R[5]
Students, Department of Computer Science and Engineering[1,2,3,4]
Faculty, Department of Computer Science and Engineering[5]
Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India

Abstract: *There haven't been many developments in the categorisation of specific ingredients for cooking. The issue is the dearth of accessible publicly edited records. In this study, the issue of automatically identifying a photographic meal for cooking and then outputting the proper recipe is addressed. There are significant overlaps between food items (also known as high intra-class similarity), which makes the chosen problem more challenging than previous supervised classification challenges because dishes from various categories may only superficially resemble one another in terms of visual data. Convolutional Neural Networks (short CNN) are used for object recognition or cookery court recognition, and the combination of these techniques and the search for nearest neighbours (Next-Neighbour Classification).*

Keywords: Inverse cooking, Image processing, Food recognition, Deep learning, Text generation

I. INTRODUCTION

A useful approach for food recording would be image recognition of food items. Then, taking a photo would serve as an adequate record. However, we are aware that food comes in a vast variety. There is a lot of variety, even within one food category. As a result, despite efforts, recognition performance for food items is still unsatisfactory. Once the food has been recognised, the appropriate recipe can be located. Food is essential for human survival. It not only gives us energy, but it also shapes our culture and identity. We are what we eat, as the proverb goes, thus activities associated to food, like preparing, eating, and talking about it, occupy a sizable portion of our daily lives. In the contemporary digital era, food culture has become more popular than ever thanks to the widespread use of social media for sharing food-related images. There are at least 300 million posts on Instagram when you search for #food, and at least 100 million posts are returned when you search for #foodie, demonstrating the undeniable importance of food in our culture. Also changing with time are eating habits and cooking customs. While most meals were once made at home, nowadays we regularly eat food that has been prepared by others (e.g., takeaways, catering and restaurants). As a result, it is difficult to determine exactly what we eat because access to comprehensive information about prepared food is restricted.

When compared to natural picture understanding, however, food recognition presents more difficulties because food and its components have a high intraclass variability and exhibit significant deformations due to cooking. In a cooked dish, ingredients frequently become obscured and arrive in a range of hues, shapes, and textures. Additionally, the recognition of visual ingredients involves sophisticated thinking and prior information (e.g., cake will likely contain sugar and not salt, while croissant will presumably include butter). Thus, the goal of food recognition is to push the boundaries of current computer vision systems and incorporate past knowledge to provide high-quality structured food preparation descriptions..

II. RELATED WORK

2.1 Food Understanding

By supplying reference benchmarks to train and compare machine learning methods, the development of large-scale food datasets, such as Food and Recipe, along with a recently organised food challenge2, has enabled considerable breakthroughs in visual food recognition. Due to this, there is now

There is a substantial body of literature in computer vision that deals with a range of food-related tasks, with an emphasis on image classification. Later efforts handle more difficult issues including calculating meal quantities, predicting the list of components, and determining the recipe for a given food image, among other things. Additionally,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

it offers a thorough cross-regional analysis of culinary recipes, taking into account pictures, recipe ingredients, and recipe features (such as style and course) and ingredients for recipes. Food-related activities have also been taken into account in the literature on natural language processing, where the synthesis of procedural text from flow diagrams or ingredient checklists has been explored in relation to recipe development.

2.2 Multi-Label Classification

The literature has invested a lot of time and energy in developing models and researching loss functions that are effective for this job in order to use deep neural networks for multi-label classification. Early approaches relied on single-label classification models and binary logistic loss, presuming that labels were independent of one another and omitting potentially important data. Using label power sets dependencies.

Power sets consider all possible label combinations, which makes them intractable for large scale problems. Another expensive alternative. To overcome this issue, probabilistic classifier chains and their recurrent neural network-based counterparts propose to decompose the joint distribution into conditionals, at the expense of introducing intrinsic ordering. Note that most of these models require to make a prediction for each of the potential labels. Moreover, joint input and label embedding have been introduced to preserve correlations and predict label sets. As an alternative, researchers have attempted to predict the cardinality of the set of labels; however, assuming the independence of labels. When it comes to multi-label classification objectives, binary logistic loss, target distribution cross entropy, target distribution mean squared error and ranking-based losses have been investigated and compared. Recent results on large scale datasets outline the potential of the target distribution loss.

2.3 Conditional Text Generation

Conditional text generation with auto-regressive models has been widely studied in the literature using both text-based, as well as image-based conditionings. In neural machine translation, where the goal is to predict the translation for a given source text into another language, different architecture designs have been studied, including recurrent neural networks, convolutional models and attention-based approaches. More recently, sequence -to-

sequence models have been applied to more open-ended generation tasks, such as poetryand story generation. Following neural machine translation trends, autoregressive models have exhibited promising performance in image captioning, where the goal is to provide a short description of the image contents, opening the doors to less constrained problems such as generating descriptive paragraphs or visual storytelling.

III. PRE-REQUISITES

3.1 Generating Recipes from Images

IVGenerating a recipe (title, ingredients and instructions) from an image is a challenging task, which requires a simultaneous understanding of the ingredients composing the dish as well as the transformations they went through, e.g., slicing, blending or mixing with other ingredients. Instead of obtaining the recipe from an image directly, we argue that a recipe generation pipeline would benefit from an intermediate step predicting the ingredients list. The sequence of instructions would then be generated conditioned on both the image and its corresponding list of ingredients, where the interplay between image and ingredients could provide additional insights on how the latter were processed to produce the resulting dish.

3.2 Cooking Instruction Transformation

Given an input image with associated ingredients, we aim to produce a sequence of instructions by means of an instruction transformer. Note that the title is predicted as the first instruction. This transformer is conditioned jointly on two inputs: the image representation and the ingredient embedding. We extract the image representation with a encoder and obtain the ingredient embedding by means of a decoder architecture to predict ingredients, followed by a single embedding layer mapping each ingredient into a fixed-size vector. The instruction decoder is composed of transformer blocks, each of them containing two attention layers followed by a linear layer. The first attention layer applies self-attention over previously generated outputs, whereas the second one attends to the model conditioning in order to refine

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

the self-attention output. The transformer model is composed of multiple transformer blocks followed by a linear layer and a SoftMax nonlinearity that provides a distribution over recipe words for each time step.

.

3.3 Ingredient Decoding

Which is the best structure to represent ingredients? On the one hand, it seems clear that ingredients are a set, since permuting them does not alter the outcome of the cooking recipe. On the other hand, we colloquially refer to ingredients as a list (e.g., list of ingredients), implying some order. Moreover, it would be reasonable to think that there is some information in the order in which humans write down the ingredients in a recipe. Therefore, in this subsection we consider both scenarios and introduce models that work either with a list of ingredients or with a set of ingredients. A list of ingredients is a variable sized, ordered collection of unique meal constituents.

3.4 Optimization

We train our recipe transformer in two stages. In the first stage, we pre-train the image encoder and ingredients decoder. Then, in the second stage, we train the ingredient encoder and instruction decoder. Note that, while training, the instruction decoder takes as input the ground truth ingredients. All transformer models are trained with teacher forcingexcept for the set transformer.

IV. EXPECTED OUTCOMES

From the uploaded food image name of the food and recipe should be displayed. It should also display calories. Origin of the food and YouTube video link will also be fetched along with the other information's.

V. CONCLUSION

In this paper, we introduced an image-to-recipe generation system, which takes a food image and produces a recipe consisting of a title, ingredients and sequence of cooking instructions. We first predicted sets of ingredients from food images, showing that modelling dependencies matters.

Then, we explored instruction generation conditioned on images and inferred ingredients, highlighting the importance of reasoning about both modalities at the same time. Finally, user study results confirm the difficulty of the task, and demonstrate the superiority of our system against state-of-the-art image-to-recipe retrieval approaches.

REFERENCES

[1]. Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In ECCV, 2014.

[2]. Micael Carvalho, R´emiCad`ene, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embedding. In SIGIR, 2018.

[3]. Jing-Jing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In ACM Multimedia. ACM, 2016.

[4]. Jing-Jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. Cross-modal recipe retrieval with rich food attributes. In ACM Multimedia. ACM, 2017.

[5]. Mei-Yun Chen, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung. Automatic chinese food identification and quantity estimation. In SIGGRAPH Asia 2012 Technical Briefs, 2012.

[6]. Xin Chen, Hua Zhou, and Liang Diao. Chinesefoodnet: A large-scale image dataset for chinese food recognition. CoRR, abs/1705.02743, 2017.

[7]. Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. Towards diverse and natural image descriptions via a conditional gan. ICCV, 2017.

[8]. Krzysztof Dembczy´nski, Weiwei Cheng, and EykeH¨ullermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In ICML, 2010.

[9]. Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In ACL, 2018.

[10]. Claude Fischler. Food, self and identity. Information (Inter- national Social Science Council), 1988.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, July 2022
Impact Factor: 6.252

[11]. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. CoRR, abs/1705.03122, 2017.

[12]. Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. CoRR, abs/1312.4894, 2013.

[13]. Kristian J. Hammond. CHEF: A model of case-based planning. In AAAI, 1986.

[14]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In CVPR, 2015.

[15]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[16]. Luis Herranz, Shuqiang Jiang, and Ruihan Xu. Modeling restaurant context for food recognition. IEEE Transactions on Multimedia, 2017.

[17]. Shota Horiguchi, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. Personalized classifier for food image recognition. IEEE Transactions on Multimedia, 2018.

[18]. Qiuyuan Huang, Zhe Gan, Asli C¸ elikyilmaz, Dapeng Oliver Wu, JianfengWang, and Xiaodong He. Hierarchically structured