International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 1324-1326 www.ijaem.net ISSN: 2395-5252

Survey on Reverse Cooking: Recipe Creation from Food Images

Ashwini Hegde, Rabia Ishrat, Ranjitha Y, Yaswaswini N

Submitted: 10-07-2022 Revised: 18-07-2022 Accepted: 23-07-2022

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 fornearest neighbours (Next-Neighbour Classification).

KEYWORDS: Inversecooking, Image processing, Food recognition, Deeplearning, Text generation

I. INTRODUCTION

The Image recognition of food items would be a good solution to food recording. Taking a picture would then be a sufficient record. However, we know that there is a wide diversity of types of food. Even within the same food category, there is considerable diversity. Therefore, despite the attempts at food item recognition, recognition performance is not yet satisfactory. Once the food is identified, proper recipe can be found accordingly. Food is fundamental to human existence. Not only does it provide us with energy—it also defines our identity and culture. As the old saying goes, we are what we eat, and food related activities such as cooking, eating and talking about it take a significant portion of our daily life. Food culture has been spreading more than ever in the current digital era, with many people sharing pictures of food they are eating across social media. Querying Instagram for #food leads to at least 300M posts; similarly, searching for #foodie results in at least 100M posts, highlighting the unquestionable value that food has in our society. Moreover, eating patterns and cooking culture have been evolving over time. In the past, food was mostly prepared at home, but nowadays we frequently consume food prepared by third parties (e.g., takeaways, catering and restaurants). Thus, the access to detailed information about prepared food is limited and, as a consequence, it is hard to know precisely what we eat.

However, when comparing to natural image understanding, food recognition poses additional challenges, since food and its components have high intra class variability and present heavy deformations that occur during the cooking process. Ingredients are frequently occluded in a cooked dish and come in a variety of colours, forms and textures. Further, visual ingredient detection requires high level reasoning and prior knowledge (e.g., cake will likely contain sugar and not salt, while croissant will presumably include butter). Hence, food recognition challenges current computer vision systems to go beyond the merely visible, and to incorporate prior knowledge to enable high-quality structured food preparation descriptions.

II. RELATED WORK

Food Understanding. The introduction of large-scale food datasets, such as Foodand Recipe, together with a recently held food challenge2 has enabled significant advancements in visual food recognition, by providing reference benchmarks to train and compare machine learning approaches. As a result, there is currently
a vast literature in computer vision dealing with a variety of food related tasks, with special focus in image classification. Subsequent works tackle more challenging tasks such as estimating the number of calories given a food image, estimating food quantities, predicting the list of present ingredients and finding the recipe for a given image. Additionally, it provides a detailed cross-region analysis of food recipes, considering images, attributes (e.g., style and course) and recipe ingredients. Food related tasks have also been considered in the natural language processing

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 1324-1326 www.ijaem.net ISSN: 2395-5252

literature, where recipe generation has been studied in the context of generating procedural text from either flow graphs or ingredients' checklists.

Multi-label classification. Significant effort has been devoted in the literature to leverage deep neural networks for multi-label classification, by designing models and studying loss functions well suited for this task. Early attempts exploit single-label classification models coupled with binary logistic loss, assuming the independence among labels and dropping potentially relevant information. One way of capturing label dependencies is by relying on label power sets. 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.

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

Generating recipes from images. Generating 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.

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 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.

.

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.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 7 July 2022, pp: 1324-1326 www.ijaem.net ISSN: 2395-5252

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.

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.

[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