GENERATION OF RECIPE USING FOOD IMAGES

Kotte Shivani[1]., B.Amulya[2].,Choudary Srivalli[3].,M.Pranavi[4].,T.Greeshma[5]

1 Assistant Professor, Department Of ECE., Malla Reddy College Of Engineering For Women.,Maisammaguda., Medchal.,Ts, India (✉shivani.kotte481@gmail.Com)

2, 3, 4, 5 B.TechECE, (19RG1A04C8, 19RG1A04D5, 19RG1A04F6, 19RG1A04H0), Malla Reddy College Of Engineering For Women., Maisammaguda., Medchal., Ts, India

Abstract

The appreciation for food is reflected in the popularity of food photography. Each dish has a backstory that's written out in a detailed recipe, and it's a shame that we can't peek inside the kitchen just by gazing at a photo of the food. Thus, in this study, we provide an inverse cooking system, which, given a picture of a dish, can reconstruct the recipe for making that dish. In order to provide cooking directions, our system first predicts ingredients as sets using a unique architecture, modelling their relationships without imposing any order, and then it simultaneously considers the picture and its inferred components. We conduct a thorough evaluation of the system as a whole on the massive Recipe1M dataset, demonstrating that (1) we outperform prior baselines for ingredient prediction, (2) we can obtain high-quality recipes by leveraging both image and ingredients, and (3) our system can generate more compelling recipes than retrieval-based approaches, as judged by humans. Code and models are made accessible to the general public.1 .

Introduction

We can't survive without food. It's not only the fuel that keeps us going; it's also what sets us apart as a people and a culture [10, 34]. As the adage goes, "you are what you eat," and thus it seems to reason that our lives revolve heavily on the preparation, consumption, and discussion of food. More individuals than ever before are promoting their culinary culture over the internet, with many posting photos of what they're eating on various social media platforms [31]. The unarguable worth that food has in our culture is shown by the fact that a query for #food on Instagram returns at least 300 million posts, while a comparable query for #foodie returns at least 100 million photos. In addition, both the way people cook and the foods they consume have changed throughout time. While most meals were cooked at home in the past, many of us today rely on other sources (such as delivery services, caterers, and restaurants) for our culinary needs. As a result, knowing exactly what went into a meal is easier than ever.

Figure 1: Example of a generated recipe, composed of a title, ingredients and cooking instructions. limited and, as a consequence, it is hard to know precisely what we eat. Therefore, we argue that there is a need for inverse cooking systems, which are able to infer ingredients and cooking instructions from a prepared meal.

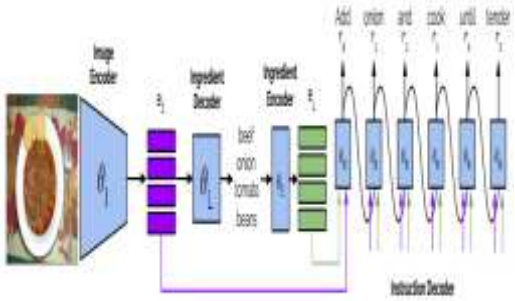
Outstanding progress has been made in natural picture classification [47, 14], object identification [42, 41], and semantic segmentation [27, 19] in recent years. However, food identification presents extra hurdles as compared to natural picture comprehension because to the significant intra class diversity of food and its components and

the substantial deformations that occur during cooking. The ingredients used in a meal might vary in size, shape, color, and texture. In addition, you need to use some very sophisticated logic and know what you're talking about in order to identify ingredients in a dish just by looking at it (a cake will probably have sugar but not salt, whereas a croissant would probably have butter). Because of this, recognizing food is a challenge for current computer vision systems, which need them to go beyond the visible and include past information to provide high-quality structured descriptions of food preparation.

In-Depth Knowledge of Food-Related Work.

Significant progress in visual food identification has been made possible by the release of large-scale food datasets like Food-101 [1] and Recipe1M [45], as well as a recently conducted food challenge2. These datasets serve as reference standards for training and comparing machine learning algorithms. Therefore, there is a large body of work in computer vision that focuses on picture classification for a wide range of food-related applications [26, 39, 38, 33, 6, 24, 30, 60, 16, 17]. More advanced problems, such as determining how many calories are in a picture of food [32], determining how much food is in an image [5], predicting the list of ingredients [3, 4], and locating the recipe for a given image [54, 3, 4, 45, 2], are addressed in subsequent studies. In addition, [34] offers a comprehensive, cross-regional analysis of recipes that takes into account visuals, qualities (such as style and course), and components. Recipe creation has been examined in the context of creating procedural text from either flow graphs [13, 36, 35] or ingredients' checklists [21], both of which have been addressed in the natural language process ing literature. Multiple-label sorting. There has been much research on loss functions [12] and model designs [49, 8, 56, 37, 53] that make use of deep neural networks for multi-label classification. In the beginning, people used single-label classification models with binary logistic loss [3], which relied on the assumption of label independence and left out data that may be useful.

Label powersets are one approach to capture label dependencies [49]. Intractable for large-scale issues, powersets take into account every conceivable label combination. The cost of acquiring knowledge about the joint likelihood of the labels is another kind of costly alternative. Probabilistic classifier chains [8] and their recurrent neural network-based [53, 37] counterparts offer to breakdown the joint distribution into con dictional, at the cost of adding intrinsic ordering, in order to solve this problem. It's important to remember that most of these models need a prediction for each of the possible labels. To further retain correlations and forecast label sets, combined input and label embeddings [57, 25, 61] have been presented. Researchers have tried to make predictions about the cardinality of the collection of labels [43, 44], but they do so under the assumption that the labels are independent. Objectives for multi-label classification that have been studied and compared include the binary logistic loss [3], the cross entropy of the target distribution [12, 29], the mean squared error of the target distribution [56], and the ranking-based losses [12]. The potential of the target distribution loss has been recently shown by findings on large-scale datasets [29]. The ability to generate text based on certain conditions. There is a wealth of literature on conditional text generation using auto-regressive models, with research focusing on both text-based [48, 11, 50, 9] and image-based conditionings [52, 59, 28, 20, 23, 7, 46]. Predicting the target language version of a given source text is the main focus of neural machine translation.



Recipe generating model (Figure 2). The image encoder, which is parametrized by I, is used to extract the image characteristics ea. Predictions of ingredients, denoted by eL, are encoded with e to form ingredient embeddings. Parameterized using R, this culinary instruction decoder takes into account image embeddings eI, ingredient embeddings eL, and previously predicted words (r0,..., rt1) to produce a recipe title and a series of cooking stages.

Recurrent neural networks [48], convolutional models [11], and attention based techniques [50] are only few of the architectures that have been investigated. Recent work using sequence-to-sequence models has expanded into more creative areas, such as poetry [55] and narrative [23, 9] production. Auto regressive models, which are a recent development in the field of neural machine translation, have shown promise in the area of image captioning [52, 59, 28, 20, 7, 46], where the goal is to provide a short description of the image contents, and thus

have opened the door to less constrained problems like the generation of descriptive paragraphs [23] and the creation of visual stories [18].

Recipe creation using image processing

Generating a recipe (title, ingredients, and directions) from a picture is difficult because it calls for a multi-layered knowledge of not just the items that went into the meal but also the processes by which those ingredients were transformed, such as chopping, blending, or mixing. We propose an intermediate phase in a recipe generating pipeline that predicts the components list instead of immediately retrieving the recipe from a picture. Based on the picture and the list of ingredients, a set of instructions for making the meal would be created, with the potential for the interaction between the two to reveal new details on the preparation of the ingredients. Our method is shown in Figure 2. A food picture is sent into our recipe generation system, and the system spits out a set of cooking instructions using an instruction decoder that accepts two em beddings as input. The first one encodes the image's ingredients, while the second one represents the image's retrieved visual attributes. In Section 3.1, we present our transfomer-based instruction decoder. In Section 3.2, we will explicitly evaluate the transformer, analyze it, and tweak it such that it can forecast components without any particular sequence. In Section 3.3, we recap the optimization specifics.

Transforming Recipe Books The goal is to use an instruction transformer [50] to convert an input picture and its related components into a set of instructions, R = (r1,..., rT), where each rt represents a word in the set of instructions. The title serves as the initial directive, so keep that in mind. The image representation eI and the ingredient embedding eL are both used as inputs to this transformer, which is conditioned on both. We use a ResNet-50 [15] encoder to extract the picture representation and a decoder architecture to forecast the ingredients, followed by a single embedding layer mapping each ingredient into a fixed-size vector to generate the ingredient embedding eL. Each transformer block in the instruction decoder consists of two attention layers and a linear layer [50]. The first layer of attention uses self-attention to enhance the outputs it has already created, while the second layer uses model conditioning to further improve the output it has already generated. For each time step t, the transformer model generates a distribution across recipe words by first applying a linear layer, and then a softmax nonlinearity. The single-modality focus of the transformer model is seen in Figure 3a. In contrast, our recipe generator relies on input from two different places: the image features eI R P de and the ingredient embeddings eL R K de (where P and K are the number of image and ingredient features, respectively, and de is the embedding dimensionality). Therefore, we need the focus of our cognition to be able to reason about both modalities at once, therefore directing the process of instruction generation. We do this by looking at three distinct fusion approaches (shown in Figure 3). The stringing together of focus. The eI and eL image and ingredient embeddings are first concatenated across the first dimension econcat R (K+P)de. After then, we use attention to focus on the new, merged embeddings. - Unbiased focus. This approach takes into account the bi-modal condition with two levels of attention. In this situation, one layer is concerned with the eI embedding of the picture, while the other is concerned with the in

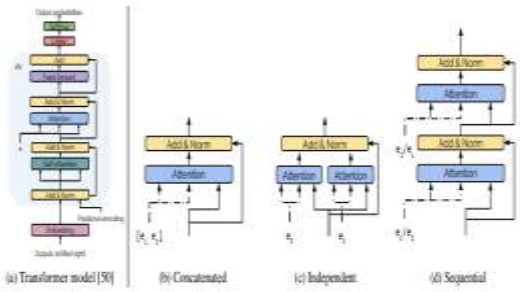


Figure 3: Attention strategies for the instruction decoder. In our experiments, we replace the attention module in the transformer (a), with three different attention modules (b-d) for cooking instruction generation using multiple conditions.

gredient embeddings eL. The output of both attention layers is combined via summation operation. – Sequential attention. This strategy sequentially attends over the two conditioning modalities. In our design, we consider two orderings: (1) image first where the attention is first computed over image embeddings eI and then over ingredient embeddings eL; and (2) ingredients first where the order is flipped and we first attend over ingredient embeddings eL followed by image embeddings eI .

Instruction decoder attention techniques are shown in Figure 3. Our studies include swapping out the transformer's (a) attention module with one of three alternative attention modules (b–d) designed to convey culinary instruction. generation based on a combination of criteria. embedded gredients eL. Sequential attention

uses a summing procedure to aggregate the results of the two attention layers. This method employs a sequential approach to both conditioning methods. For our design, we take into account two possible orders: (1) image first, in which the attention is calculated over image embeddings eI before moving on to ingredient embeddings eL; and (2) ingredients first, in which the order is reversed and we first attend over ingredient embeddings eL before moving on to image embeddings eI.

Experiments

The dataset and implementation details are described below, and then the attention techniques provided for the cooking instruction transformer are thoroughly analysed. In addition, we provide a quantitative comparison of the suggested ingredient prediction models to established norms. Finally, a detailed user research and a comparison with retrieval-based models are shown to demonstrate the efficacy of our inverse cooking system.

Dataset

The Recipe1M dataset [45], comprised of 1 029 720 recipes gathered from cookery websites, is used for both model training and testing. With each recipe including a title, list of ingredients, list of cooking directions, and (optionally) a picture, the dataset totals 720 639 training, 155 036 validation, and 154 045 test examples. After filtering out recipes with less than two ingredients or two steps, we were left with 252 547 training, 54 255 validation, and 54 506 test samples, all of which included photos. Recipes are very unstructured and include often redundant or very tightly specified culinary components (for example, olive oil, virgin olive oil, and Spanish olive oil are all different items) since the dataset was gathered via scraping cookery web pages. More than 400 varieties of cheese and more than 300 varieties of pepper are included in the ingredient of pabulary as well. As a consequence, there are 16,823 separate components in the raw dataset, which we pre-process in order to make more manageable. First, we remove plurals and ingredients that appear less than 10 times in the dataset, then we merge ingredients that share the first or last two words (for example, be con cheddar cheese is merged into cheddar cheese), and finally, we cluster the ingredients that have same word in the first or last position (for example, gorgonzola cheese or cheese blend are clustered together into the cheese category). In all, we cut the number of possible constituents from almost 16,000 down to only 1,488. Words that occur fewer than ten times in the dataset are removed using tokenization from the raw text of the cooking instructions and replaced with an unknown word token. In addition, we provide recipe-specific beginning and ending tokens.

Table 1: Model selection (val). Left: Recipe perplexity (ppl).

Model	ppl
Independent	8.59
Seq. img. first	8.53
Seq. ing. first	8.61
Concatenated	**8.50**

Model	IoU	F1
FF_{BCE}	17.85	30.30
FF_{IOU}	26.25	41.58
FF_{DC}	27.22	42.80
FF_{TD}	**28.84**	**44.11**
TF_{list}	29.48	45.55
TF_{list} + shuf.	27.86	43.58
TF_{set}	**31.80**	**48.26**

Correct: IoU& F1 standards for global ingredients. The total number of terms in the resulting "recipe vocabulary" is 23,231.

Modalities of Application

We randomly choose middle 224 224 pixels for assessment, and we scale pictures to 256 pixels on the shortest side for training. We employ a 16-block, 8-multi-head attention transformer with 64-dimensionality nodes to decode instructions. We use a 256-dimensional transformer with 4 blocks and 2 multi-head attentions for the decoder's ingredient. We utilize ResNet-50's last convolutional layer to generate embeddings for images. The dimensions of the picture and ingredient embeddings are both 512. Only recipes with no more than 20 ingredients and 150 words of instructions are kept. using a pa tierce of 50 and close attention paid to validation loss, the models are trained using Adam optimizer [22] until the early-stopping criterion is fulfilled. We use PyTorch4 [40] to implement all of our models. The appendices provide further information on the implementation process.

Conclusion

In this article, we presented a method for automatically creating recipes from food photos, complete with titles, lists of ingredients, and step-by-step directions. First, we used food image prediction to demonstrate the importance of modelling dependencies by predicting sets of components. The need of concurrently thinking about both modalities was then highlighted when we studied instruction creation based on visuals and inferred components. Finally, findings from a user study corroborate the challenge of the job and show that our system outperforms the current gold standard for image-to-recipe retrieval.

References

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

[2] Micael Carvalho, Remi Cad ´ ene, David Picard, Laure Soulier, ` Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. 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 Dembczynski, Weiwei Cheng, and Eyke ´ Hullermeier. 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 (International 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, Jianfeng Wang, and Xiaodong He. Hierarchically structured reinforcement learning for topically coherent visual story generation. CoRR, abs/1805.08191, 2018.

[19] Simon Jegou, Michal Drozdzal, David Vazquez, Adriana ´ Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In CVPR-W, 2017.

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015