RECIPE GENERATION FROM FOOD IMAGES USING DEEP LEARNING

Srinivasamoorthy.P[1], Dr. Preeti Savant[2]

[1]PG student, Department of Computer Application, JAIN University, Karnataka, India
[2]Assistant professor, Department of Computer Application, JAIN University, Karnataka, India

---***---

Abstract - *In the era of deep learning, image understanding is exploding not only in terms of semantics but also in terms of the generation of meaningful descriptions of images. This necessitates specific cross model training of deep neural networks that must be complex enough to encode the fine contextual information related to the image while also being simple enough to cover a wide range of inputs. The above-mentioned picture understanding challenge is exemplified by the conversion of a food image to its cooking description/instructions using CNN, LSTM, and Bi-Directional LSTM cross model training, this work provides a novel approach for extracting compressed embeddings of cooking instructions in a cookbook image. The varying length of instructions, the amount of instructions per dish, and the presence of many food items in a food image all provide significant challenges. Our model overcomes these obstacles by generating condensed embeddings of culinary instructions that are highly similar to the original instructions via transfer learning and multi-level error propagation across various neural networks. We experimented with Indian cuisine data (food image, ingredients, cooking instructions, and contextual information) collected from the web in this research. This The presented model has a lot of potential for application in information retrieval systems and can also be used to make automatic recipe recommendations.*

***Key Words*: LSTM, Bi-Directional, CNN, Concatenated, Independent, Sequential.**

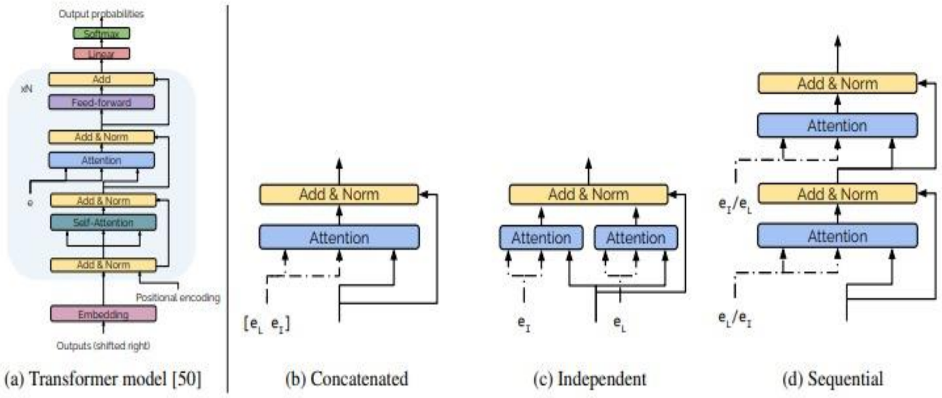
1.INTRODUCTION

Human survival depends on the availability of food. It provides us with energy as well as defining 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 up a significant portion of our daily life. In the digital age, food culture has spread further than ever before, with many individuals sharing photos of their meals on social media. On Instagram, a search for #food returns at least 300 million results, while a search for #foodie returns at least 100 million, indicating the obvious value of food in our society. Furthermore, over time, eating habits and cooking culture have evolved. Food was historically prepared at home, but we now consume food provided by others on a regular basis (e.g. takeaways, catering and restaurants). As a result, particular information about prepared foods is hard to come by, making it impossible to know exactly what we're eating. As a result, we believe inverse cooking systems are necessary, which can infer components and cooking directions from a prepared meal. In recent years, advances have been made in visual recognition tasks such as natural image categorization, object detection, and semantic segmentation. Food identification, on the other hand, has more challenges than natural picture interpretation since food and its components contain a lot of intraclass variability and severe deformations throughout the cooking process. Cooked food ingredients are often concealed and come in a variety of colours, sizes, and textures. Furthermore, detecting visual ingredients necessitates significant reasoning and prior knowledge (e.g.cake will likely contain sugar and not salt, while croissant will presumably include butter). As a result, 6 food recognition pushes current computer vision systems to think beyond the obvious in order to give high-quality structured meal preparation descriptions. Previous attempts to better understand food have mostly focused on food and ingredient categorization.

A system for comprehensive visual food recognition, on the other hand, should be able to recognize not only the type of meal or its ingredients, but also the method of preparation. A recipe is retrieved from a fixed dataset using an embedding space image similarity score in the picture-to-recipe issue, which has typically been regarded as a retrieval challenge. The size and variety of the dataset, as well as the quality of the learned embedding, have a big impact on how well these systems work. These techniques fail when the static dataset lacks a matching formula for the picture query, which is unsurprising. To get around the dataset restrictions of retrieval systems, the image-to-recipe problem might be restated as a conditional generation problem. As a result, in this paper, we offer a system for generating a cooking recipe from an image, replete with a title, ingredients, and cooking instructions. Figure 1 shows an example of a recipe created with our system, which predicts ingredients from a photo and then creates cooking instructions based on both the photo and the ingredients. To the best of our knowledge, our technology is the first to generate cooking recipes straight from food images. The task of creating instructions is modelled as a sequence generation problem with two modalities: an image and its predicted elements. Using their underlying structure, we define the ingredient prediction issue as a set prediction. Without penalizing for prediction order, we model ingredient dependencies, renewing the argument over whether or not order matters. We put our system to the test on the Recipe1M dataset, which includes photos, ingredients, and cooking instructions, and it performs admirably. In a human evaluation study, we show that our inverse cooking system outperforms previously introduced image-to-recipe retrieval methods by a significant margin. Furthermore, we show that food image-to-ingredient prediction is a challenging problem for humans to solve, and that our method can outperform them with a modest number of images. The following are the paper's main contributions: – We present an inverse cooking system that generates cooking instructions based on an image and its ingredients, as well as a study of different attention strategies for reasoning about both modalities simultaneously. – Ingredients are investigated as both a list and a set, and a new architecture for ingredient prediction is proposed that leverages component co-dependencies rather than enforcing order. – Using a user study, we show that ingredient prediction is a difficult problem, and that our suggested system beats image-to-recipe retrieval alternatives.

2. LITERATURE REVIEW

Comprehension of food. By providing reference benchmarks for training and comparing machine learning algorithms, large-scale food datasets like Food-101 and Recipe1M, as well as the new Food challenge2, have supported substantial improvements in visual food recognition. As a result, there is already a substantial body of work in computer vision dealing with a variety of food-related activities, with a focus on image classification. More challenging tasks, such as estimating the number of calories in a given food image, estimating food quantities, predicting the list of ingredients present, and establishing the recipe for a given image, are addressed in subsequent studies. Visuals, features (such as style and course), and recipe ingredients are all considered in this broad cross-region analysis of culinary recipes. In the natural language processing literature, recipe creation has been studied in the context of constructing procedural text from either flow graphs or ingredient checklists, with food-related issues taken into account.



(a) Transformer model [50] (b) Concatenated (c) Independent (d) Sequential

Fig-1.Process Image

Classification multiple labelled classification A lot of work has gone into developing models and researching loss functions that are best suited for multi-label classification with deep neural networks in the literature. Early methods relied on single-label classification models with binary logistic loss, which assumed label independence while rejecting potentially relevant data. One way to capture label dependencies is to use label powersets. Powersets assess all possible label combinations for large-scale challenges, rendering them intractable. Calculating the labels' aggregate probability is another costly option. Probabilistic classifier chains and their recurrent neural network-based counterparts propose dissecting the joint distribution into conditionals at the expense of intrinsic ordering to solve this difficulty. It's worth mentioning that the bulk of these models require you to predict each of the possible labels. In addition, combined input and label embeddings have been constructed to preserve correlations and forecast label sets. As an alternative, researchers have attempted to predict the cardinality of a set of labels based on label independence.

When it comes to multi-label classification targets, binary logistic losses [target distribution cross entropy, target distribution mean squared error, and ranking-based losses have all been investigated and compared Recent results on large datasets have shown the potential of the target distribution loss.

Condition-based text generation. In the literature, the use of both text-based and image-based conditionings in conditional text creation using auto-regressive models has received a lot of attention. The goal of neural machine translation is to predict how a given source text will be translated into a different language.

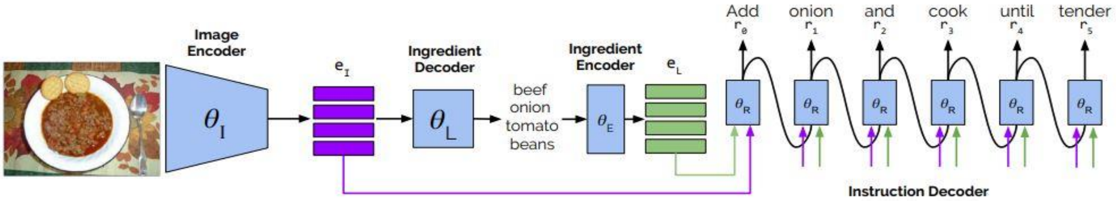


Fig -2: Working

Various architecture designs have been examined, including recurrent neural networks, convolutional models, and attention-based approaches. More open-ended generation tasks, such as poetry and storey generation, have recently been applied with sequenceto-sequence models. Following the trend of neural machine translation, autoregressive models have shown promise in picture captioning, where the goal is to produce a brief description of the image contents, potentially opening the door to more limited tasks such as creating descriptive paragraphs or visual storytelling.

3. CONCLUSION

In this paper, we offer a structure-aware generation network (SGN) for recipe generation, and we are the first to employ the concept of inferring the target language structure to direct the text generation process. We offer successful solutions to a variety of complex problems, like extracting paragraph structures without supervision, building tree structures from photographs, and using the generated trees for recipe creation. To label recipe tree structures, we employ an unsupervised technique to expand ON-LSTM. We suggest using RNN to infer tree structures from food pictures, and then using the inferred trees to improve recipe formulation. We ran thorough testing on the Recipe1M dataset for recipe generation and came up with cutting-edge findings. We also show that unsupervised tree topologies can improve food cross-modal retrieval abilities by adding structural information to cooking instruction representations. We used quantitative and qualitative analyses for the food retrieval task, and the results show that the model with tree structure representations outperformed the baseline model by a large margin. We may extend the framework in future work to generate tree topologies for various types of lengthy paragraphs because our recommended method provides sentence-level structural understandings on the text.

REFERENCES

[1] A. Salvador, M. Drozdzal, X. Giro-i Nieto, and A. Romero, “Inverse cooking: Recipe generation from food images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 453–10 462.

[2] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, “Learning cross-modal embeddings for cooking recipes and food images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3020–3028.

[3] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, ´ and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” arXiv preprint arXiv:1504.00325, 2015.

[4] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-tophrase correspondences for richer image-to-sentence models,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[5] Y. Shen, S. Tan, A. Sordoni, and A. Courville, “Ordered neurons: Integrating tree structures into recurrent neural networks,” arXiv preprint arXiv:1810.09536, 2018.

[6] L. Logeswaran and H. Lee, “An efficient framework for learning sentence representations,” arXiv preprint arXiv:1803.02893, 2018.