RECIPE GENERATION FROM FOOD IMAGES USING DEEP LEARNING

Rutuja Solanke[1], Priya Mane[2], Snehal Dane[3] Sejal Pande[4]

[1-4]*Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, India.*

Abstract - The research presented here is centered on the emergence of deep learning processes related to food, with a particular focus on map generation and cross-modal food retrieval. These tasks are aimed at enhancing their performance through unsupervised learning using map tree models. Creating recipes involves assembling food from images, which are obtained using a slicer. To enable this, different foods must be associated with images that resemble recipes based on queries. The challenge of translating directly from images to text led to the incorporation of visual objects in supporting these images, which ultimately improves performance. The Recipe1M dataset serves as the foundation of this research, featuring images of recipes along with ingredient descriptions and cooking instructions. However, there are significant differences from typical image-text cross-modal datasets like Flickr and MS-COCO, mainly due to multi-sentence cooking instructions and the lack of ingredient images.

To address these challenges, a Structure-Aware Generative Network (SGN) is introduced, combining tree data with training methods to generate maps. This involves utilizing an RNN for map tree creation based on food images and integrating tree models into recipe generation with the help of GAT, ensuring detailed food descriptions. The study also highlights the efficacy of unsupervised learning using map tree models in cross-modal food retrieval. Various modules like 2tree, img2tree, and tree2recipe are employed to handle tree structures, contributing to the recipe generation process. The contributions of this research include the proposal of a 2-tree map, the introduction of img2tree and tree2recipe modules for creating and utilizing tree models, and the demonstration of performance improvements over existing methods. Additionally, experiments beyond the Recipe1M dataset are conducted to validate the effectiveness of the proposed approach. In summary, this study addresses the gap in knowledge related to long-cooked foods by employing unsupervised sentence-level tree learning. This approach integrates recipes and enhances cross-modal food retrieval, showcasing the potential to transform cooking-related tasks and achieving state-of-the-art performance on the Recipe1M dataset. The research includes imaging and ablation studies to assess its effectiveness thoroughly

Key Words- Deep Learning, Unsupervised Learning, Cross-Modal Food Retrieval, Recipe Generation, Structure-Aware Generative Network, Map Tree Models, Recipe1M Dataset.

1.INTRODUCTION

In this paper, our research focuses on the emergence of a deep learning process regarding food, based on the importance of food in people's lives. In particular, we delve into two important tasks: map generation and cross-modal food retrieval, aiming to improve their performance through unsupervised learning of map tree models. Creating recipes involves creating recipes that are assembled from foods using a slicer to obtain images, while different foods need to be given images similar to recipes based on questions. We realize that it is difficult to translate directly from images to text to accomplish this task, so we use the best way to support beautiful images with visual objects. Additionally, learning the art of visual objects can improve performance on these tasks.

The Recipe1M dataset forms the basis of our research and includes images of recipes, descriptions of ingredients and cooking instructions. However, there are significant differences between these data and image-text cross-modal datasets such as Flickr and MS-COCO. First, cooking instructions are often multi-sentence descriptions, making it

difficult to represent the entire cooking process in a single food image. Secondly, the food pictures do not include instructions due to the mixing of ingredients during cooking. This special situation creates a need for new solutions.

To solve these problems, we ask subjects to pay no attention to tree structure and then place them differently on the production and food supply map to improve their performance. Inspired by advances in speech parsing, we introduce the Recipe2tree module, an extension of the ON-LSTM architecture designed for sentence-level model trees. Fast theory is used to train the ONLSTM continuum by creating a structure tree that represents the entire cooking process.

We introduce Structure-Aware Generative Network (SGN) that combines tree data with training methods to generate maps. RNN is used to create a map tree based on food images, and GAT helps incorporate the tree model into the creation of recipes. This ensures that the description of the food produced is detailed, thus increasing production efficiency .We also demonstrate the effectiveness of unsupervised learning of the map tree model on cross-modal food intake. The tree representation in the feature map plays an important role in the process. We use the 2tree module to capture the low-level tree map and the img2tree module to create the tree map from the menu image. The Tree2recipe module encodes the extracted tree structure, which is then used by the graph network to develop the recipe function.

Our contributions include a proposal for a 2-tree map, use of the img2tree and tree2recipe modules to create and use tree models, and demonstration of improvements that fix competition work. Further experiments to verify the effectiveness of our method beyond the capabilities of the original Recipe1M dataset. Our research also includes imaging and ablation studies evaluating vascular access .In summary, one of the main goals of this study is to address the lack of appropriate knowledge on long-cooked foods. Through an unsupervised approach to sentence-level tree learning, we aim to fill this gap, enabling integration of recipes and improving cross-modal access across foods. Our proposed model demonstrates the state-of-the-art performance of the Recipe1M dataset, demonstrating its potential to transform cooking-related task.

2. Body of Paper

This research delves into the burgeoning field of deep learning processes pertaining to food, focusing specifically on map generation and cross-modal food retrieval. The objective is to elevate the performance of these tasks through the application of unsupervised learning using innovative map tree models. Crafting recipes involves orchestrating various ingredients from images obtained through a slicer. To facilitate this, associating different foods with images resembling recipes based on queries becomes essential. Overcoming the challenge of translating images directly to text is addressed by incorporating visual objects, thereby enhancing overall performance.

The research builds upon the Recipe1M dataset, comprising images of recipes along with detailed ingredient descriptions and cooking instructions. Distinct from conventional image-text cross-modal datasets like Flickr and MS-COCO, Recipe1M poses challenges with multi-sentence cooking instructions and the absence of ingredient images.

To tackle these challenges, we introduce the Structure-Aware Generative Network (SGN), a novel paradigm combining tree data with advanced training methods to generate maps. Leveraging an RNN, the SGN creates a map tree based on food images. Integration of tree models into recipe generation is achieved with the assistance of Graph Attention Networks (GAT), ensuring nuanced food descriptions. Key modules, namely 2tree, img2tree, and tree2recipe, are introduced to handle tree structures, significantly contributing to the recipe generation process. This research pioneers the concept of a 2-tree map, while the img2tree and tree2recipe modules play pivotal roles in creating and utilizing tree models. Contributions of this research encompass advancements in the proposal of a 2-tree map, the introduction of img2tree and tree2recipe modules, and a demonstrable enhancement in performance compared to existing methodologies. The efficacy of the proposed approach is validated through experiments extending beyond the Recipe1M dataset.

Table -1 Deep Literature Survey of Current Technologies

Sr. No.	Paper Title Publication Details	Pre-Processing	Feature Extraction and Classification	Accuracy	Post Processing	Research Gap Identified
1.	AutoChef: Automated Generation of Cooking Recipes.	**Extract ingredients, actions, and cooking instructions from existing recipes.**	Feature extraction: Extract features from the food image, such as color, texture, and shape. This can be done using a Pre-trained CNN. Classification: Once the feature has been extracted then it is classified into different recipes. This can be done using algorithms such as SVM and Random Forests.	82%	AutoChef is the process of refining and improving the generated recipes to make them more accurate, complete, and user-friendly. This can be done in a variety of ways, such as: Correcting errors in the ingredients and instructions. Adding additional details, such as cooking times and temperatures. Reformatting the recipes to make them easier to read and follow. Customizing the recipes to meet the user's preferences	Need for a larger and more diverse training dataset to improve the accuracy and creativity of the recipe generation model. Need for better methods to handle complex recipes and recipes with multiple steps. Need for better methods to incorporate user preferences and dietary restrictions into the recipe generation process.

					and dietary restrictions.	
2.	Swasth: An inverse cooking Recipe Generation From Food images	**Resize, crop, normalize, and extract features from food images.**	Feature extraction: T he feature extraction component extracts a set of visual features from the food image. These features represent the appearance of the food, such as its color, textur e, and shape. Classification : The classification component takes the extracted visual features as input and predicts the ingredients and cooking steps for the recipe.	76.4%	Filtering out irrelevant ingredients and instructions fro m the generated recipe. Correcting the order of instructions to ensure that the recipe is logical and easy to follow. Adding missing details to the recipe, such as cooking times and temperatures. Generating a human-readable and informative recipe that is easy to understand and follow.	The system is not able to generate recipes for complex dishes. The system is not able to generate recipes for dishes that are not in the training dataset. The system is not able to generate recipes that are tailored to the user's preferences.
3.	Instant Recipe Generation from food images	Resize the image to 224 × 224 pixels. Normalize the image by	Feature extraction: A convolutiona l neural network (CNN) is used to extract	75%	Recipe filtering: The generated recipe is filtered to remove any invalid or	Need for more efficient and accurate object detection and segmentatio

		subtracting the mean and dividing by the standard deviation of the training dataset. Detect and segment the food items in the image. This can be done using a variety of object detection and segmentation algorithms, such as Faster R-CNN and Mask R-CNN, respectively. Extract features from the food items. This can include color, texture, and shape features. Augment the data by randomly cropping, flipping, rotating, and adjusting the	features from the food image. The CNN learns to extract features that are relevant to the task of recipe generation. Classification: A machine learning classifier is used to classify the extracted features into different categories. The categories can correspond to different types of food items, different cooking methods, or different cuisines.		infeasible instructions. Recipe summarization: The generated recipe is summarized to make it more concise and easier to read. Recipe personalization: The generated recipe can be personalized to meet the user's preferences and dietary restrictions.	n algorithms to handle complex food images with multiple food items. Need for more robust feature extraction methods that are invariant to changes in lighting, scale, and pose. Need for better methods to incorporate user preferences and dietary restrictions into the recipe generation process. Need for more realistic and informative evaluation metrics to assess the quality of generated recipes.

		brightness and contrast of the images.				
4.	Recipe Generation From Food Images Using Deep Learning	The preprocessing steps involved in the Recipe Generation From Food Images Using Deep Learning (IRJET) paper can be summarized as: **image resizing, normalization, object detection, segmentation, and feature extraction**.	Feature extraction: The paper uses a convolutional neural network (CNN) to extract features from the food images. The CNN learns to identify high-level features in the images, such as the shapes, textures, and colors of the food items. Classification: The extracted features are then used to classify the food images into different categories, such as main courses, side dishes, and desserts. This classification information is then used to generate a	90%	Filtering out unrealistic or infeasible ingredients and instructions. Generating a natural language description of the recipe.	**Need for more robust and diverse training datasets** to improve the accuracy and creativity of the recipe generation model. This research gap can be addressed by developing new methods for collecting and curating training data, and by developing new methods for augmenting existing training datasets.

			recipe for the food items in the image.			
5.	Image-to-Recipe Translation with Deep Convolutional Neural Networks	Resize and normalize the image. Extract features from the image using a pre-trained DCNN. Reduce the dimensionality of the extracted features	Feature extraction: A pre-trained CNN is used to extract deep visual features from the food image. Classification: A classifier is trained on the extracted visual features to predict the corresponding recipe category.	82%	Filtering out unrealistic or infeasible ingredients and instructions. Generating a natural language description of the recipe. Clustering similar recipes together.	**Need for better methods to handle complex food images with multiple food items.** **Need for better methods to incorporate user preferences and dietary restrictions**
6.	Inverse Cooking Recipe From Food Images And Cuisine Classification	**Image resizing and normalization.** **Object detection and segmentation.** **Feature extraction.**	Feature Extraction: Visual features are extracted from the food image using a CNN. . Textual features are extracted from the recipe ingredients using an RNN.	85.20 %	Filtering **out unrealistic or infeasible ingredients and instructions.** **Generating a natural language description of the recipe.** **Cuisine classification**	Previous systems have focused on retrieving recipes from a database based on the similarity of the image to the recipes in the database. However, these systems are not able to generate new recipes or to adapt

		Cuisine classification.	Classification : The extracted visual and textual features are concatenated and fed to a classifier. The classifier predicts the cuisine of the dish.		**Recipe recommendation.**	recipes to the user's preferences.
7.	PIXEL TO PLATE: Generating Recipes From Food Images Using CNN.	**Image resizing and normalization.** **Object detection and segmentation.** **Feature extraction.** Data Augmentation.	Feature Extraction: Visual features are extracted from the food image using a CNN. Classification : The extracted visual and textual features are concatenated and fed to a SOFTMAX classifier. The softmax classifier is trained on the same dataset of food images and recipes as the CNN.	85%	Filtering **out unrealistic or infeasible ingredients and instructions.** **Generating a natural language description of the recipe.** Formatting the Recipe.	**Need for** **better methods to handle complex food images with multiple food items.** **Need for better methods to incorporate user preferences and dietary restrictions**
8.	Image to Recipe Prediction System	In this system, preprocessing involves:	Feature extraction: E xtract features from the	79.4%	Post-processing involves refining and structuring the predicted	**Need for more robust and diverse training datasets**

		Resize and normalize the image. Detect and segment food items. Extract colour, texture, and shape features. Apply data augmentation techniques.	food image, such as color, texture, and shape. Classification: Use a machine learning model to classify the features and predict the recipe.		recipe output for presentation and further use, like arranging ingredients and instruction in a user-friendly format	**and more effective methods for object detection, segmentation, feature extraction, recipe prediction, post-processing, and personalization.**
9.	Food Image to the Recipe Generator	Image Resizing and cropping, Colour Normalization, noise reduction, object detection, Feature Extraction.	Feature extraction: Extract features from the food image, such as color, texture, and shape. This can be done using a Pre-trained CNN. Classification: Once the feature has been extracted then it is classified into different recipes. This can be done	80%	Removing Duplicate ingredients and Instructions. Correct Grammatical Errors. Add Missing Information. Format the recipe. Test the Recipe.	Ingredient Ambiguity: Food image can be ambiguous and it can be difficult to identify ingredients in dish. Lack of cooking instructions: Food images typically do not include cooking instructions it make difficult to generate

			using algorithms such as SVM and Random Forests.			recipe from image. Limited training Data

Table -2 Algorithmic Survey of Research Studies

Sr No.	Paper Title	Algorithm used	Time Complexity	Space Complexity	Accuracy	Advantages/ Disadvantages
1	Inverse Cooking:Recipe generation from food images	Faster R-CNN	O(N+K*(R*C+C_cls+C_reg+k^2))	O(R*C)	55.47%	It detects objects and also recognizes them,making it suitable for tasks that require both localization and classification,such as object detection and image segmentation.
2	AutoChef:Automated Generation of Cooking Recipes	AutoChef :The Recipe Generator			82%	AutoChef extracts the data from existing recipes using natural language processing, learns the combination of ingredients, preparation actions and cooking instructions, and autonomously generates the recipes.
3	Model for Cooking Recipe Generation using Reinforcement	Reinforcement Learning				RL models can explore a wide range of recipes by trying various combinations of ingredients and steps, potentially

	Learning					discovering novel and unique dishes.
4	Food image to Cooking Instructions Conversion Through Compressed Embeddings Using Deep Learning	CNN,LSTM,Bi-Directional LSTM				The proposed model can he significantly useful for information retrieval system and it can also be effectively utilized in automatic recipe recommendation.
5	Food Recipe Alternation and Generation with Neural Language Processing Techniques	Natural Language Processing				In this research project, we investigated how to apply the state-of-the-art natural language processing techniques such as word embedding to help people choose alternative ingredients/recipes and build language models – N-gram and neural network model to generate new recipes with authentic flavor of certain cuisine style.
6	Reinforcement Learning for Logic Recipe	Reinforcement Learning,				The results indicate that the proposed system achieves a better
	Generation:Bridging Gaps from	LSTM				performance than other methods on both aspects of producing proper

	Images to Plans					ingredients and effective recipes.
7	Learning Structural Representations for Recipe Generation and Food Retrieval	RNN	O(Td2h+ Tdhdi)		89.50%	RNNs can generate recipes of variable length based on the complexity of the image, accommodating both simple and complex dishes .

3. CONCLUSIONS

In conclusion, this research has explored the applications of deep learning in the domain of food-related tasks, with a particular emphasis on map generation and cross-modal food retrieval. These tasks have been significantly enhanced through the use of unsupervised learning and the integration of map tree models. The creation of recipes, which involves assembling foods from images, presented challenges that were effectively addressed by incorporating visual objects to support image-based representations. The Recipe1M dataset served as a valuable foundation for this research, despite its unique characteristics compared to traditional image-text cross-modal datasets.

To overcome these challenges, the study introduced the Structure-Aware Generative Network (SGN), which combines tree data with innovative training methods to generate detailed maps. This approach incorporated tree structures into recipe generation, resulting in more detailed and efficient food descriptions. The study also demonstrated the effectiveness of unsupervised learning with map tree models for cross-modal food retrieval, emphasizing the crucial role played by the tree representation in feature maps. Various modules were developed to handle tree structures and enhance recipe generation. The contributions of this research include the proposal of a 2-tree map, the development of img2tree and tree2recipe modules for creating and utilizing tree models, and the demonstration of performance improvements over existing methods. Furthermore, the study conducted experiments that extended beyond the Recipe1M dataset to validate the effectiveness of the proposed approach.

In summary, this research has made substantial strides in addressing the lack of knowledge related to long-cooked foods. Through an unsupervised approach to sentence-level tree learning, it has bridged the gap, enabling the integration of recipes and significant improvements in cross-modal food retrieval. The proposed model demonstrates its potential to transform cooking-related tasks and showcases state-of-the-art performance on the Recipe1M dataset. The research has also been supported by imaging and ablation studies, further validating its effectiveness.

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] 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. JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 1

[3] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," arXiv preprint arXiv:1607.06215, 2016.

[4] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4613–4623.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156– 3164.

[6] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8415–8424.

[7] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 323– 10 332.

[8] Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7181–7189.

[9] 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.

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