Journal of Engineering, Science and Mathematics

Volume 02, Number 01, Pages 11-31, Year 2023
https://jesm.in/archives/

Research Paper

CNN Based Self Attention Mechanism for Cross Model receipt Generation for Food Industry

Ismail Keshta[1], Mukesh Soni[2]

[1]Computer Science and Information Systems Department, College of Applied Sciences, Al Maarefa University, Riyadh, Saudi Arabia

[2]Department of CSE, University Centre for Research & Development, Chandigarh University, Mohali, Punjab-140413, India

Correspondence should be addressed to Ismail Keshta; imohamed@mcst.edu.sa

Handling Editor: Mohammad Shabaz

Diet management requires keeping track of what you eat. The researchers presented a recipe retrieval technique based on food photos that retrieves the related recipes from the taken images and creates nutritional information accordingly, making recording more convenient. The retrieval of recipes is an example of a cross-modal retrieval challenge. Still, as compared to other challenges, the main challenge is that recipes explain a succession of modifications from raw ingredients to completed goods rather than immediately apparent characteristics. As a result, the model must have a thorough understanding of the raw materials processing process. Current recipe retrieval research, on the other hand, uses a linear approach to text processing, which makes it difficult to capture long-range relationships during recipe processing. A cross-modal recipe retrieval model that is based on the self-attention mechanism is currently being developed in order to overcome this difficulty. The model makes use of the Transformer model's self-attention mechanism to effectively capture long-distance interactions in recipes. Additionally, the model improves upon the attention mechanisms of prior techniques in order to mine the semantics of recipes more effectively. The approach enhances the recall rate of the recipe retrieval task by 22% over the baseline strategy, according to experimental data.

Keywords: *Blockchain, CNN, Attention layer, Cross Model, receipt Generation, Food Industry*

1. Introduction

People's daily eating habits have a significant impact on people's health. With the continuous improvement of living standards, people pay increasingly attention to a healthy diet. As a result, the record of daily dietary intake, as a critical part of dietary management, has received increasingly attention [1-2]. Furthermore, for some patients with chronic diseases (such as chronic kidney disease), monitoring nutritional intake is even more crucial due to their need for disease treatment [3-4]. The basic process of traditional dietary intake recording tools is to predict food categories and raw material types through user food pictures and further estimate their nutrient content [5-6]. However, due to the variety of styles and appearances of people's homemade foods, the actual results are insufficient to support the inference of nutritional content from food pictures [7-8]. With the development of the Internet, increasingly people are uploading food pictures and recipes to social networks and public recipe sites. This makes it possible to collect food pictures and their corresponding recipes on a large scale and provide new ideas for solving the problem of dietary intake records. With the support of large-scale recipe data, a new research direction has emerged: recipe retrieval, that is, to retrieve the corresponding recipes in large-scale recipes through a picture of food [9-11]. The recipes contain food categories, raw material types, and dosages. On this basis, other related work, such as raw material identification and nutrient content estimation, can be further performed to achieve the purpose of dietary intake records. In addition to the label information used in traditional prediction methods, recipes also contain information on the operation steps of food production, which contains richer semantic information than category labels, which can help the model better understand the food images in the picture. Semantics and achieve higher accuracy [10]. This is a classic cross-modal challenge, in which the data of one modality (such as images) are used to locate matching samples in the database of another modality (such as text). It's usual for cross-modal retrieval to have issues with photos and their basic text explanations, for example. The difficulty of the recipe retrieval issue is mostly due to the intricacy of the relationship between the words in the recipe and the images of the dish. In contrast to pictures of cooked food, recipes explain the transformation of raw ingredients into

finished products rather than explicitly expressing the qualities of the finished product [12]. There is a great deal of interplay and superposition going on here. The look of the completed product may be comparable to that of other raw materials that have undergone a variety of treatments, yet the appearance of the raw material may be completely different. In order to retrieve recipes with picture characteristics, the model has to have a thorough understanding of the processing of the basic ingredients in the recipe. In recipes, however, the texts of multiple processing steps for the same raw material are not always close together. The processing steps in recipes can generally be represented as a tree structure according to their timing dependencies [13]. The processing steps for different raw materials belong to other branch processes without timing dependencies, and finally, the multiple processing flows of various raw materials are summarized. Come together. However, the format of most recipes requires that the original tree-structured process be recorded as a linear structure, and the methods of processing different raw materials are intertwined with each other, so the text between two adjacent steps of processing a specific raw material may be far apart. As shown in the recipe in Figure 1, steps 1 and 4 in italics are processing steps for the same raw material (potatoes), and their dependencies are significantly higher than their adjacent steps.

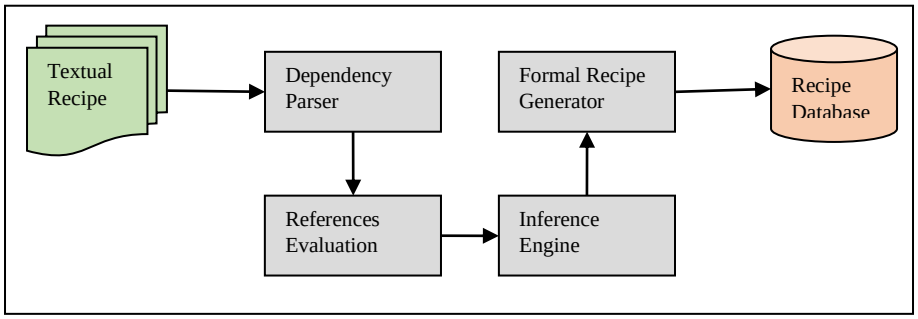

Figure 1. Distant dependency in the recipe

The current work on recipe retrieval uses traditional text processing models (RNN (recurrent neural network) or its variants LSTM (long short-term memory), GRU (gated recurrent unit), etc.) to process recipe text. However, these models linearly process

textual data and can capture long-range dependencies, making it difficult to fully understand how ingredients appearing in pictures are processed in recipes. Research works also try to use the attention mechanism to capture the dependencies in recipes [10]. Still, they select the same context vector for all recipes in the attention mechanism and do not consider the differences between recipes and uniqueness, resulting in poor accuracy. To make the trained model more in line with the eating habits of the domestic people, this paper obtains data from popular domestic recipe websites and constructs a new large-scale Global recipe dataset (LCR). The LCR dataset contains more than 170,000 Global recipes. Each recipe includes the title, raw materials used, and operating steps. The dataset also consists of the final product image corresponding to the formula.

This research proposes a novel cross-modal recipe retrieval approach framework based on self-attention. The paper presents a recipe text encoder based on the transformer network and improves the environment vector selection method of the traditional method's attention mechanism. In summary, the two points improve the model's ability to understand the complex text of recipes. Furthermore, the effectiveness of the cross-modal recipe retrieval method framework in this paper is verified on the LCR dataset, and the retrieval accuracy (Recall@1) is 22% higher than the existing state-of-the-art baseline methods in the cross-modal recipe retrieval field.

2. Different modal retrieval

2.1 Cross-modal retrieval Model

Cross-modal retrieval refers to retrieving results related to a provided query sample using data from different modalities. Common cross-modal retrieval problems include retrieving associated descriptions through images or retrieving images through keywords. The main difficulty of the cross-modal retrieval problem is how to measure the similarity between data of different modalities [15. A common approach is to map the elements of various modalities into a shared latent space, where details that match each other are aligned, so that data from different modalities in the latent area can be processed. Compare. A classical approach in the field of cross-modal retrieval is canonical correlation analysis (CCA). Then CCA has developed many variants, such as kernel CCA (kernel CCA) [17]. Later, with the development of deep neural networks, many works also began to

use deep neural networks for latent space mapping [18]. Compared with common cross-modal retrieval problems, such as using images to retrieve descriptive text or vice versa [19], not all text information in recipe retrieval is directly reflected in photos, and there is a complex relationship between text and image relations.

2.2 Food-related research Model

Food-related research has always been a concern and valued by researchers. Common research directions include food identification [20], raw material identification [8], and nutrient composition estimation. The original recipe retrieval work can be seen as an extension of the above research, focusing on predicting food attributes, such as food categories, raw materials, etc. from food images, and further finding the corresponding recipes in the candidate set based on these attributes, as shown in Figure 2(a) show. Authors proposed a multitask learning method to simultaneously predict food categories and raw materials and adjust the expected results through a conditional random field (CRF). Afterwards, authors believed that more prosperous food attributes could be added and considered processing steps in addition to categories and raw materials [21]. The model proposed above is trained based on the labels corresponding to food images. The more attributes considered, the more complex the required labels and the higher the data requirements [22]. For example, the literature requires images to have fine-grained processing step labels, which are usually difficult to obtain. Authors took the lead in introducing the classic latent space alignment idea of cross-modal retrieval into recipe retrieval, and their model framework is shown in Figure 2. Such a structure enables retrieval not to be limited by fixed food attributes and requires less manual extraction of labels during training. Authors [10] introduced an attention mechanism based on the above model, making the model focus on the part of the recipe that has a more significant impact on the appearance of the food, and capturing the implicit causality. At the same time, a ranking loss (rank loss) is introduced for the retrieval problem, which speeds up the training speed. Based on the ranking loss, the author [11] proposed a double-triple-t loss function better to utilize the different levels of semantic information in recipes. At the same time, it also offers a new gradient descent and back propagation method to improve the training effect.

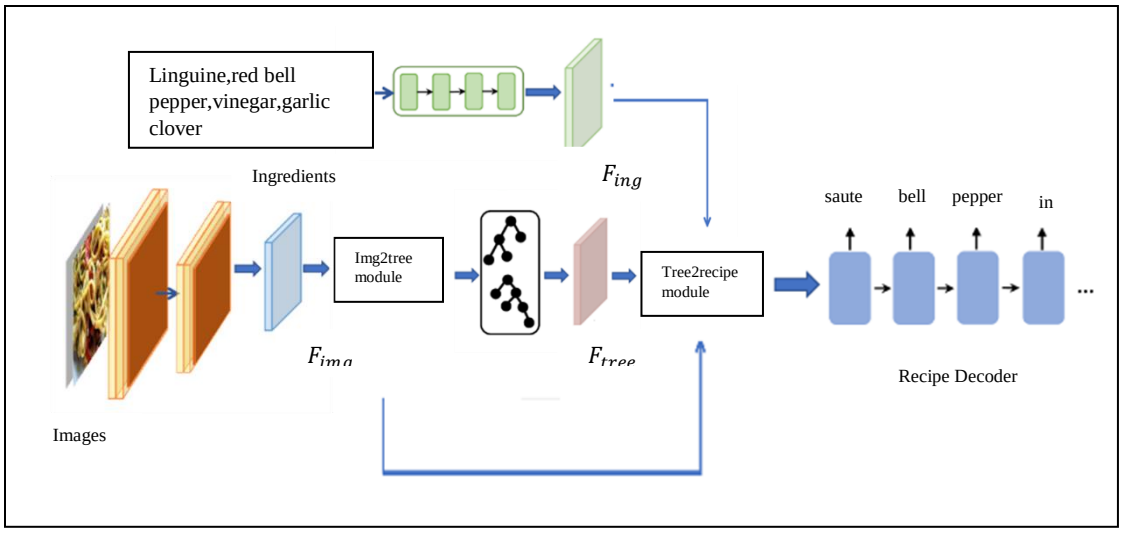

Figure 2. Two frameworks in recipe retrieval

However, these methods still have some shortcomings. However, in the above techniques, linear text processing models such as recurrent neural network (RNN) or its variants (LSTM, GRU) are used to encode nonlinear recipe text. This work offers a self-attention recipe text encoder that can manage long-distance interstep dependencies.

3. Cross-Model Framework

3.1 Different parts of recipe text

The model framework of this paper is shown in Figure 3, which consists of three parts: the recipe text encoding module, the image encoding module, and the joint embedding module. The data of the two modalities of text and pictures respectively obtain corresponding feature representations through their respective encoding modules [23]. The recipe text is divided into three parts: title, raw material, and operation steps, encoded by three text encoders with different parameters. And finally, the feature representation of the recipe text is obtained by splicing, while a deep convolution neural network processes the food image coding [24]. Then the feature vectors of the two modalities enter the joint embedding module to learn the embedding representation in a shared latent space so that the similarity between the matched text and the image is as high as possible. The entire model is trained end-to-end. Next, the different

modules in the model will be introduced separately.

3.2 Different parts of Cross-Model

3.2.1 Transformer

The Transformer model is a model proposed by Google in 2017 [16]. Its original purpose was to replace the recurrent neural network to solve the task of seq2seq in natural language processing. Compared with the recurrent neural network, the Transformer solves two problems: one is to get rid of the serial calculation order and improve its parallel ability; the other is to solve the problem of the recurrent neural network processing long text because the text is too long" Forgetting" phenomenon, which makes it challenging to deal with the issue of long-distance dependence.

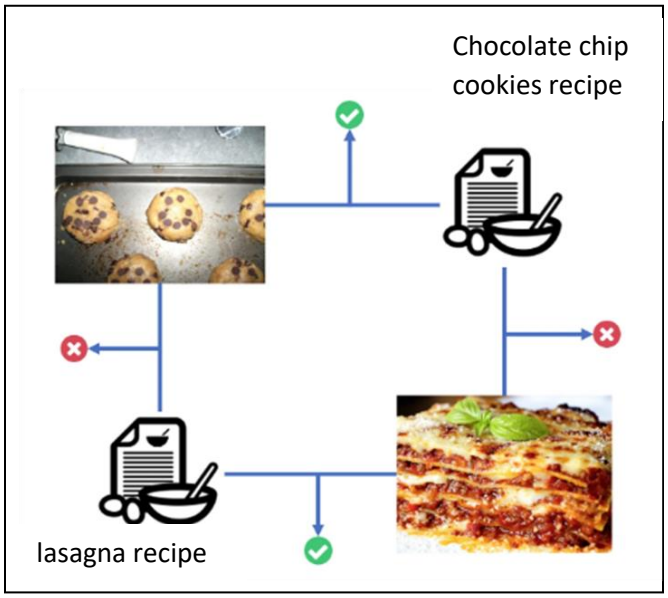

Figure 3. Model framework

The ability of the transformer model to deal with long-distance dependencies mainly depends on the self-attention mechanism it adopts [25]. Unlike the recurrent neural network (RNN), which can only learn the previous information through the hidden layer of information transmitted in the last time slice, with the help of the self-attention mechanism, the Transformer model can "globally browse" the input data and find a higher correlation with it.

In the remainder of this article, this network is noted as:

$$x = \text{Transformer}(y) \quad (1)$$

Among them, x is the input word vector; y is the hidden layer vector after the model encodes the word vector.

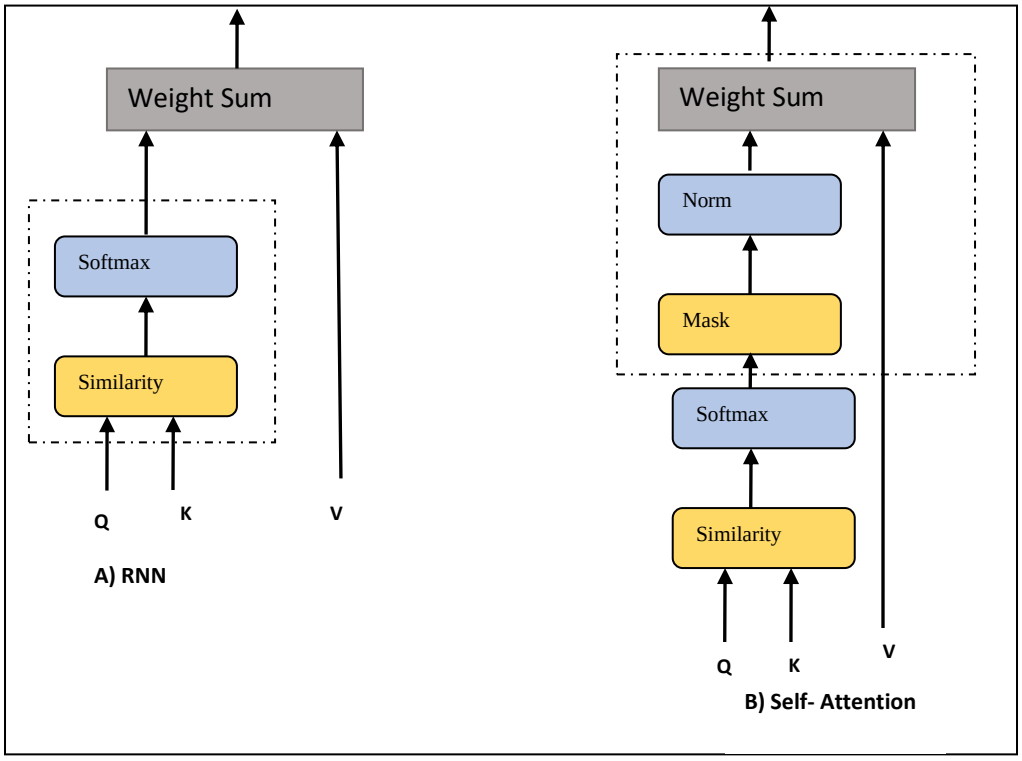

Figure 4. Comparison between RNN and self-attention mechanism

3.2.2 Primary cooking method

Each recipe contains a title. The title of a recipe is usually a high-level summary of the recipe. It usually includes the primary raw materials used in the recipe and the primary cooking method (such as scrambled eggs with tomatoes), and some also include flavors or cuisine (such as spicy shredded potatoes, Sichuan-style spareribs) Wait). For example, I was given a recipe title containing U words x_u,u∈ [1, U]. First, it is converted into n-dimensional word vectors by the word2vec algorithm, and then the T-word vectors are input into a text encoder. Each vector can obtain an n-dimensional hidden layer representation ht:

$$y_u = \text{word2vec}(x_u), u \in [1, U] \quad (2)$$

$$i_u = \text{Transformer}(x_u), u \in [1, U] \quad (3)$$

Then this paper adopts the attention mechanism to calculate the encoding of the entire title. The principle of the attention mechanism is to calculate the weights for different hidden layer representations and obtain the final feature representation by weighted summation, to reflect the importance of other text parts. In this paper, the attention mechanism is implemented by a single-layer multilayer perceptron (MLP) and a SoftMax layer:

$$v_u = \tanh(X_{att} i_u + c_{att}) \quad (4)$$

$$\alpha_u = \frac{\exp(v_u^S v_d)}{\sum_u \exp(v_u^S v_d)} \quad (5)$$

Among them, X_{att} and c_{att} are the parameters of the perceptron layer, which can be trained by back propagation. The final weight, α_u obtained by the attention mechanism, is measured by the similarity between v_u and v_d. $v_d$is is called the context vector and represents the preference for the hidden layer vector. The selection method of the environment vector is detailed in Section 3.2.5.

Finally, the encoded subtitle of the title (64 dimensions) is obtained by the weighted average of the hidden layer representations of each word in the title:

$$emb_{title} = \sum_t \alpha_u i_u \quad (6)$$

3.2.3 Material encoding

Recipes usually have a section listing the various raw materials used, visible (e.g., tomatoes, potatoes, etc.) and invisible (e.g., salt, sugar, etc.).

The raw material encoding emb_{ingr} is the same way as the title encoding. First, the hidden layer representation is obtained by formula (2) and formula (3). Then, the weight is calculated by formula (4) and formula (5), and finally, the final encoded emb_{ingr} (64 dimensions) is obtained by procedure (6).

3.2.4 Different operation steps

The operation steps describe the food preparation process in detail and are the most semantically informative part of the recipe but also the most complex part. Usually, the section is made up of multiple long sentences. The description of the operation steps usually has no direct correspondence with the appearance of the finished product. For example, the degree of "putting in the pot and stir-frying" does not

directly describe the changes in the color and shape of the raw materials after processing but requires the model to understand itself. When encoding the step information, this paper does not choose a hierarchical encoding method similar to [10]: first, through an encoding model, the word vector is converted into sentence-level encoding, and then through the second encoding model, all sentences are obtained using the overall code. This paper does not use the hierarchical encoding method but uses a single transformer model to get the final general representation of the word vector directly [26]. The reasons for this are as follows: first, the use of the hierarchical transformer model consumes a lot of computing resources; secondly, the stacking of multiple sublayers of the transformer model itself has played a certain degree of abstraction (this article also found two in the actual test. The accuracy of these methods is the same). To sum up, when encoding step information, consider it as a lengthy document and still use equations (2) to (6) to obtain the final step encoding ambient (256 dimensions).

3.2.5 Environment vector

A similar attention mechanism to this paper is used in [10]. This paper improves the v_d(environment vector) selection method of the attention mechanism in [10]. In [10], Vdis were randomly initialized, continuously adjusted during training, and shared among all recipes. This method is not the optimal way to select the environment vector, and the randomly initialized and shared Vdis is challenging to reflect the uniqueness of different recipes. This paper chooses to use the maximum pooling of all hidden layer vectors of the title as v_d; the reason is that the title of the recipe can be regarded as the overall summary of the formula provided by the author of the procedure, from which a better comprehensive representation of the recipe can be obtained, thus helps the attention mechanism to give higher weights to more critical parts of the recipe. The experimental feature also confirms that the method adopted in this paper can obtain higher accuracy.

3.2.6 Encoding process

The overall encoding of the recipe text is obtained by splicing three parts of the encoding:

$$emb_{\text{recipe}} = [emb_{title}, emb_{ingr}, emb_{inst}] \quad (7)$$

3.3 Processing pictures

In this paper, the 50-layer residual network version ResNet50 is used as the image encoding module, and the network parameters are initialized with the parameters pretrained on the ImageNet dataset. After the picture is input into the residual neural network, this paper selects the penultimate layer of the network (i.e., removes the last layer of the SoftMax classification layer) as the encoding image of the picture data, with a dimension of 2 048 dimensions.

3.4 Neural Network Encoding

In this paper, the corresponding information in the text and the picture is mapped to the same latent space through a layer of fully connected neural network encoding, which are denoted as ϕR and ϕv, respectively. The latent space dimension is 1 024 dimensions. The fully connected network uses the tanh activation function because it is desirable to use cosine similarity in this latent space to measure the similarity between sample representations:

$$\emptyset_S = \tanh(X_S emb_{\text{recipe}} + c_S) \quad (8)$$

$$\emptyset_x = \tanh(X_x emb_{\text{image}} + c_x) \quad (9)$$

3.5 Different loss function

This paper uses the same loss function as the literature [11], including the retrieval loss M_{retr} and the semantic loss M_{sem}. The total loss function is:

$$M = M_{retr} + \lambda M_{sem} \quad (10)$$

Among them, λ is a hyper parameter that controls the relative size of retrieval loss and semantic loss.

The input to the retrieval loss contains two triplets ϕ_x,ϕ_S+,ϕ_S- and (ϕ_S,ϕ_x+,ϕ_x). The first element of the triplet is the feature representation of the image (ϕ_x) or text (ϕ_S), and the sample is called the anchor sample; the last two elements are the matching sample (positive example) in another modality and the non- Representation of matched samples (negative examples). Then the retrieval loss is defined as follows, where α is the margin, a number between 0 and 1:

$$M_{retr} = \frac{\sum_{(\emptyset_x,\emptyset_S,\emptyset_S)} M((\emptyset_x,\emptyset_S,\emptyset_S))}{\sum_{(\emptyset_x,\emptyset_S,\emptyset_S)} sgn(M((\emptyset_x,\emptyset_S,\emptyset_S)))} + \frac{\sum_{(\emptyset_x,\emptyset_S,\emptyset_S)} M((\emptyset_S,\emptyset_x,\emptyset_x))}{\sum_{(\emptyset_S,\emptyset_x,\emptyset_x)} sgn(M((\emptyset_S,\emptyset_x,\emptyset_x)))} \quad (11)$$

Among them, M(a,p,n)=max(0,α-cos(a,p)+cos(a,n)) .

The goal of retrieval loss is to make the similarity between the anchor sample and the positive example at least a positive number α higher than the similarity between it and the negative example.

The calculation method of the semantic loss is the same as that of the retrieval loss; the difference is that when the triplet is constructed, the positive example and the negative example, respectively, select the representation of the same and different samples as the anchor sample in another modality, denoted as($\emptyset_{\mathrm{x}}, \emptyset_S, \emptyset_S$) , $(\emptyset_{\mathrm{S}}, \emptyset_x, \emptyset_x)$

$$M_{sem} = \frac{\sum_{(\emptyset_{\mathrm{x}}, \emptyset_S, \emptyset_S)} M((\emptyset_{\mathrm{x}}, \emptyset`_S, \emptyset`_S))}{\sum_{(\emptyset_{\mathrm{x}}, \emptyset`_S, \emptyset`_S)} sgn(M((\emptyset_{\mathrm{x}}, \emptyset`_S, \emptyset`_S)))} + \frac{\sum_{(\emptyset_{\mathrm{x}}, \emptyset_S, \emptyset_S)} M((\emptyset_{\mathrm{S}}, \emptyset`_x, \emptyset`_x))}{\sum_{(\emptyset_{\mathrm{S}}, \emptyset`_x, \emptyset`_x)} sgn(M((\emptyset_{\mathrm{S}}, \emptyset`_x, \emptyset`_x)))} \quad (12)$$

Semantic loss aims to maximize the similarity between samples of the same category in different modalities and reduce the similarity between representatives of various types.

3.6 Learning parameters

Since the overall model is large and has many parameters, learning, parameters of the two encoding modules of text and pictures simultaneously may cause shocks in the training results. Therefore, this paper draws on the idea commonly used when training models in cross-modal domains and trains the model in stages. The training process is divided into three phases: In the first stage, the parameters of the image encoder module are fixed. Then, during back propagation, only the parameters of the text encoding module and the joint representation learning module are updated until the accuracy converges on the validation set. In the second stage, the parameters of the text encoding module and the joint representation learning module are fixed. During back propagation, only the parameters of the picture encoding module are updated until the accuracy converges on the validation set. In the third stage, the parameters of all modules are updated at the same time. After the first two stages of training, the parameters of all modules have been trained, and this stage only further finetunes the model parameters.

4. Different Stages of Training

4.1 Home eating habits

To make the model more in line with the domestic eating habits, this paper downloads data from domestic popular public recipe websites and constructs a large-scale Global food recipe dataset LCr dataset. The LCR

dataset includes 177,048 recipes and their corresponding finished pictures. During the experiment, this paper randomly selected 15,000 recipes as the validation set, 15,000 as the test set, and the remaining 147,048 as the training set. Each recipe in the LCR dataset includes three parts: title, raw materials, and operation steps. According to statistics, the average title length of each recipe is 6.83 words, and each formula contains an average of 6.89 raw materials and 7.34 operating steps. At the same time, using the method of reference [9] in this paper, through binary/triple matching, the recipes in the LCR dataset are divided into 1 026 categories, among them, the 0th class is the background class, representing the recipe whose type cannot be determined exist among all recipes, recipes belonging to the background category accounted for about 52%.

4.2 Different Model

4.2.1 Pytorch framework

The code in this article is implemented using the Pytorch framework. The model was trained using the Adam optimizer with a learning rate of 10-4, an interval α of 0.3, λ of 0.1, and a batch size of 50. The above hyper parameters are selected on the validation set. When computing the loss function, we select positive and negative examples for the anchor samples in each batch. Specifically, all samples in a collection are regarded as anchor samples. For each anchor sample, all set's qualified positive and negative examples are found to construct triples to calculate retrieval loss and semantic loss.

4.2.2 Median rank

In this experiment, two tasks of using images to retrieve recipes (im2recipe) and using recipes to retrieve images (recipe2im) were selected to measure the model's accuracy. In addition, they were tested on three alternative sets of 1 000/5 000/10 000 sizes. As indications of accuracy, the experiments presented in this work employ both the median rank (MedR) and the top K recall rate (recall rate at top K, R@K). The ordinal is the median of the ordinal numbers that correspond to all recovered samples in the retrieval results. It is referred to as the median of the ordinal numbers of the retrieved samples. The lower the value, the higher the accuracy of the model; the top K recall rate, for example, the top in the im2recipe task 5 recall rate refers to the ratio of the recipes corresponding to the pictures appearing in the top 5 of the retrieval results

when using photographs to retrieve recipes the higher the value, the higher the accuracy of the model. The above evaluation metrics are independently calculated 30 times on the candidate set.

4.3 Different algorithms

This paper selects three models proposed in recipe retrieval in recent years as the benchmark algorithm and compares the accuracy with the proposed algorithm. They are: (1) JNE (joint neural embedding) [9]; (2) ATTEN [10]; (3) Adamine [11]. The experimental results are shown in Table 1. It can be seen from Table 1 and Fig 5 that the accuracy of the model in this paper is significantly better than the three benchmark algorithms under different tasks and different candidate set sizes. Among them, the top 1 recall of the im2recipe job on the 10000-candidate set is 22% higher than the best benchmark algorithm. All three benchmark algorithms use a recurrent neural network model to process text linearly in the text encoding module. The model in this paper better captures long-range dependencies in recipes through (1) self-attention mechanism; (2) improved environmental variables. As a result, the uniqueness of recipes is better captured, the semantic understanding ability of the model is improved, and higher retrieval accuracy is achieved.

Table 1. Comparison between proposed model and baseline

Size	Method	im2recipe				recipe2im			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1000	JNE	4.4	0.239	0.546	0.674	5.2	0.227	0.511	0.483
	ATTEN	4.3	0.253	0.555	0.69	4.1	0.257	0.56	0.688
	Adamine	2.6	0.352	0.672	0.771	2.5	0.361	0.675	0.774
	Proposed	2	0.407	0.733	0.828	2	0.413	0.739	0.834
5000	JNE	18.3	0.097	0.277	0.393	21.9	0.094	0.261	0.369
	ATTEN	17.1	0.105	0.3	0.408	16.7	0.109	0.298	0.411
	Adamine	8.9	0.162	0.405	0.534	8.5	0.169	0.413	0.541
	Proposed	6.5	0.194	0.462	0.598	6.1	0.203	0.475	0.606
10000	JNE	34.9	0.064	0.194	0.286	43	0.061	0.179	0.264
	ATTEN	33.9	0.068	0.202	0.304	33	0.071	0.21	0.304
	Adamine	16.7	0.109	0.295	0.412	16	0.114	0.304	0.419
	Proposed	12	0.133	0.344	0.471	11.5	0.139	0.358	0.483

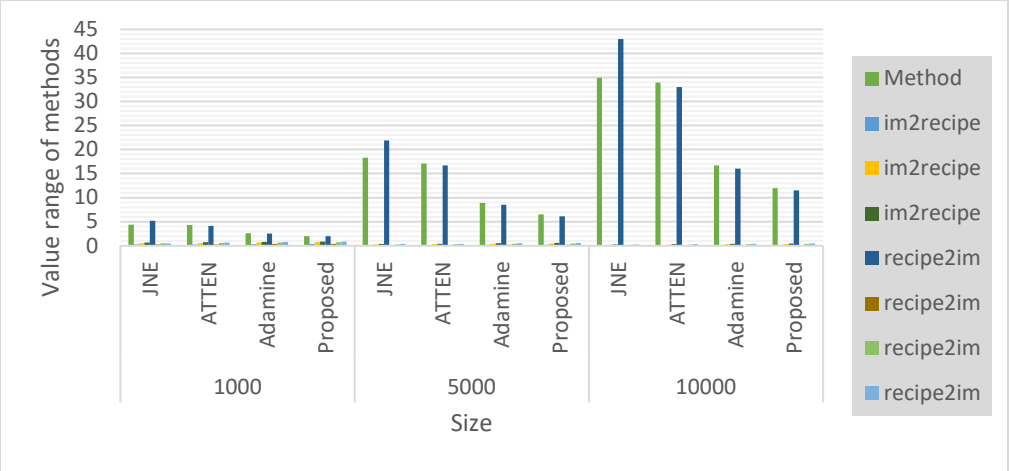

Figure 5. Comparison between proposed model and baseline

Table 2. Comparison between two attention methods

Size	Method	im2recipe				recipe2im			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1000	Uc_glob	2	0.4	0.733	0.833	2	0.41	0.734	0.836
	Uc_title	2	0.407	0.733	0.828	2	0.413	0.739	0.834
5000	Uc_glob	6.8	0.193	0.46	0.592	6.4	0.202	0.469	0.597
	Uc_title	6.5	0.194	0.462	0.598	6.1	0.203	0.475	0.606
10000	Uc_glob	12.2	0.132	0.345	0.468	12	0.138	0.354	0.476
	Uc_title	12	0.133	0.344	0.471	11.5	0.139	0.358	0.483

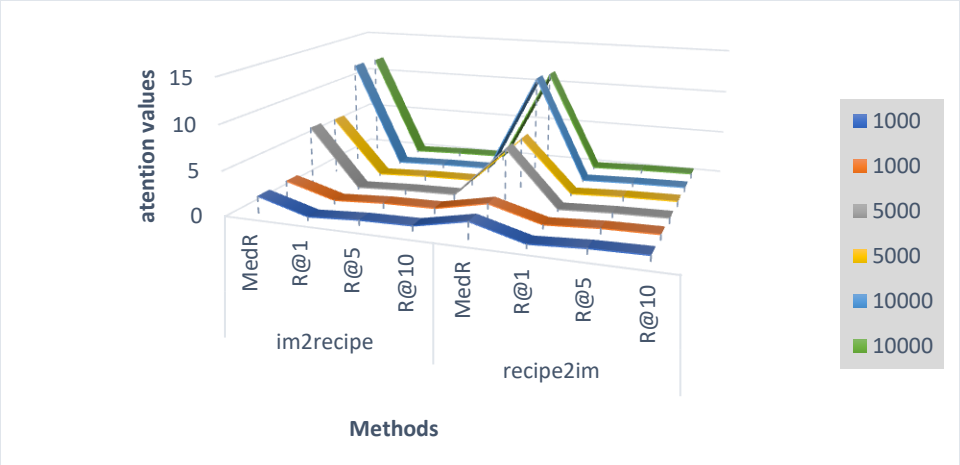

Figure 6. Comparison between two attention methods

4.4 Different mechanisms in the model

While comparing the accuracy, this paper also conducts some control experiments to verify the effectiveness of each module in this model by changing or deleting some modules in the model and comparing the accuracy.

4.4.1 Environment vector

This paper adopts a different mechanism from the literature [10] to select the environment vector in the attention mechanism (see Section 3.2.5 for details). This paper compares the two selection methods. The detailed results are shown in Table 2. Among them, Uc_glob is the selection method of the environment vector in [10], and Uc_title is the selection method of the model of this paper. Reference [10] uses the same end-to-end learned parameter vectors as the environment vector for all encoding vector as the environment vector, which better captures the difference between different recipes. In most cases, the results obtained are better than the selection method in the literature [10]. recipes. At the same time, this paper selects the maximum pooling of each recipe title.

4.4.2 Training model

Figure 7with Table 3shows the results of the training model on the data of each part of the recipe and its combination (reporting only the accuracy of the im2recipe task on the 10,000-sized candidate set), which can reflect the importance of each part in the recipe.

Table 3. Experiment results of using different recipe parts and their combinations

Size	Method	R@1	R@5	R@10
1000	JNE	0.239	0.546	0.674
	ATTEN	0.253	0.555	0.69
	Adamine	0.352	0.672	0.771
	Proposed	0.407	0.733	0.828
5000	JNE	0.097	0.277	0.393
	ATTEN	0.105	0.3	0.408
	Adamine	0.162	0.405	0.534
	Proposed	0.194	0.462	0.598
10000	JNE	0.064	0.194	0.286
	ATTEN	0.068	0.202	0.304
	Adamine	0.109	0.295	0.412
	Proposed	0.133	0.344	0.471

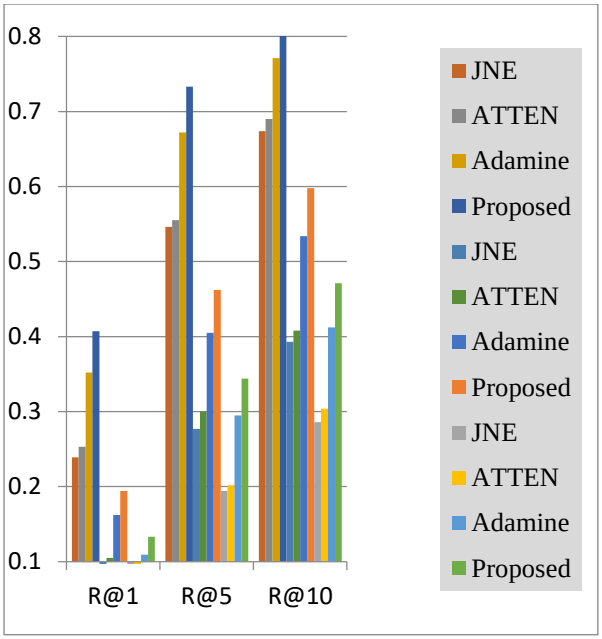

Figure 7. Experiment results of using different recipe parts and their combinations

It can be seen that when training on a specific part of the recipe alone, among the three features, the training using the data of "operation steps" has the highest accuracy, which reflects the importance of operation steps in the recipe. At the same time, this also shows that when the model understands the input recipe, in addition to the partially visible features contained in the title and raw materials, the model does learn richer implicit semantic information from the complex operation step information for retrieving the final finished image. At the same time, according to the experimental results, it can be seen that the results obtained by training the model using the title or raw material data alone are not good. Still, when the two parts are combined, the accuracy rate is even slightly better than using the operation step data alone. This indicates that the information related to the recipe retrieval task in the two parts of the data is complementary.

4.4.3 Hyper parameters

This paper also conducts a parameter sensitivity test on the two hyper parameters λ and α used in the model, where λ is a hyper parameter that controls the relative size of the semantic loss and retrieval loss. A is the margin in the loss function, indicating that this paper has the expected value of the similarity difference between positive and negative sample pairs. The detailed results are shown in Figure 8 (R@10 accuracy of im2recipe task on 10 000 candidate sets).

5. Conclusion

If you'd want to keep track of what you eat and how much you consume, this research presents a cross-modal recipe retrieval model that uses a self-attention mechanism to automatically retrieve the recipes that are most relevant to the input food images. The model in this paper uses the Transformer model's self-attention mechanism to capture better long-distance dependencies appearing in recipes and improves the semantic understanding

ability of the model. This research also improves the attention mechanism that is utilized in the classic recipe retrieval model, which better captures the uniqueness across various recipes and further boosts the model's accuracy. Both of these benefits come as a result of the publication of this work. Although the model in this paper has made noticeable progress compared with the traditional model, it has not yet reached the perfect point. Therefore, the current method only takes targeted consideration in the recipe text encoding module, while a more commonly used image processing model is used in the picture encoding module. In the future, the image encoding module will be adjusted and optimized for the recipe retrieval problem.

References

[1] H. Wang et al., "Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes With Semantic Consistency and Attention Mechanism," in IEEE Transactions on Multimedia, vol. 24, pp. 2515-2525, 2022, doi: 10.1109/TMM.2021.3083109.

[2] A. Salvador, E. Gundogdu, L. Bazzani and M. Donoser, "Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15470-15479, doi: 10.1109/CVPR46437.2021.01522.

[3] H. Wang et al., "Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes With Semantic Consistency and Attention Mechanism," in IEEE Transactions on Multimedia, vol. 24, pp. 2515-2525, 2022, doi: 10.1109/TMM.2021.3083109.

[4] Z. Xie, L. Liu, Y. Wu, L. Li and L. Zhong, "Learning TFIDF Enhanced Joint Embedding for Recipe-Image Cross-Modal Retrieval Service," in IEEE Transactions on Services Computing, doi: 10.1109/TSC.2021.3098834.

[5] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu and L. Herranz, "Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration," in IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 1100-1113, May 2017, doi: 10.1109/TMM.2016.2639382.

[6] Lokhande, M. P., Patil, D. D., Patil, L. V., & Shabaz, M. (2021). Machine-to-Machine Communication for Device Identification and Classification in Secure Telerobotics Surgery. In C. Chakraborty (Ed.), Security and Communication Networks (Vol. 2021, pp. 1–16). Hindawi Limited. https://doi.org/10.1155/2021/5287514

[7] J. Marín et al., "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 187-203, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2927476.

[8] B. Zhu, C. -W. Ngo and W. -K. Chan, "Learning From Web Recipe-Image Pairs for Food Recognition: Problem, Baselines and Performance," in IEEE Transactions on Multimedia, vol. 24, pp. 1175-1185, 2022, doi: 10.1109/TMM.2021.3123474.

[9] H. Wang, D. Sahoo, C. Liu, E. -p. Lim and S. C. H. Hoi, "Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11564-11573, doi: 10.1109/CVPR.2019.01184.

[10] A. Salvador et al., "Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3068-3076, doi: 10.1109/CVPR.2017.327.

[11] T. Ueda, M. Okada, N. Mori and K. Hashimoto, "A Method to Estimate Request Sentences using LSTM with Self-Attention Mechanism," 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), 2019, pp. 7-10, doi: 10.1109/IIAI-AAI.2019.00013.

[12] Yang, M., Kumar, P., Bhola, J., & Shabaz, M. (2021). Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. In International Journal of System Assurance Engineering and Management. Springer Science and Business Media LLC. https://doi.org/10.1007/s13198-021-01415-1

[13] R. Ye, W. Wang, Y. Ren and K. Zhang, "Bearing Fault Detection Based on Convolutional Self-Attention Mechanism," 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, 2020, pp. 869-873, doi: 10.1109/ICCASIT50869.2020.9368683.

[14] Lohani, T. K., Ayana, M. T., Mohammed, A. K., Shabaz, M., Dhiman, G., & Jagota, V. (2021). A comprehensive approach of hydrological issues related to ground water using GIS in the Hindu holy city of Gaya, India. In World Journal of Engineering. Emerald. https://doi.org/10.1108/wje-04-2021-0223

[15] Wu, C., Lu, P., Xu, F., Duan, J., Hua, X., & Shabaz, M. (2021). The prediction models of anaphylactic disease. In Informatics in Medicine Unlocked (Vol. 24, p. 100535). Elsevier BV. https://doi.org/10.1016/j.imu.2021.100535

[16] M. Kim, T. Kim and D. Kim, "Spatio-Temporal Slowfast Self-Attention Network For Action Recognition," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 2206-2210, doi: 10.1109/ICIP40778.2020.9191290.

[17] Y. Yang, H. Liang, Y. Yang and W. Xu, "Image Arbitrary Style Transfer via Self-Attention Mechanism Based on Feature Fusion," 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE), 2021, pp. 58-63, doi: 10.1109/ICAIE53562.2021.00019.

[18] Evaluation and Categorization of Handwriting Patterns reflecting Sentiments. (2019). In International Journal of Recent Technology and Engineering (Vol. 8, Issue 2, pp. 2475–2477). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijrte.b2081.078219

[19] Z. Zhang and R. Zhang, "Combined Self-attention Mechanism For Biomedical Event Trigger Identification," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 1009-1012, doi:

10.1109/BIBM47256.2019.8983274
.

[20] Li, C., Niu, H., Shabaz, M. et al. Design and implementation of intelligent monitoring system for platform security gate based on wireless communication technology using ML. Int J Syst Assur Eng Manag 13, 298–304 (2022). https://doi.org/10.1007/s13198-021-01402-6

[21] P. -Y. Wang, C. -T. Chen, J. -W. Su, T. -Y. Wang and S. -H. Huang, "Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism," in IEEE Access, vol. 9, pp. 55244-55259, 2021, doi: 10.1109/ACCESS.2021.3071306.

[22] J. Yang and J. Yang, "Aspect Based Sentiment Analysis with Self-Attention and Gated Convolutional Networks," 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), 2020, pp. 146-149, doi: 10.1109/ICSESS49938.2020.9237640.

[23] Gagandeep Kaur Saini, Harshit Chouhan, Sharlabh Kori, Akanksha Gupta, Mohammad Shabaz, Vishal Jagota, Bhupesh Kumar Singh. (2021). Recognition of Human Sentiment from Image using Machine Learning. Annals of the Romanian Society for Cell Biology, 1802–1808.

[24] J. Yang and J. Yang, "Aspect Based Sentiment Analysis with Self-Attention and Gated Convolutional Networks," 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), 2020, pp. 146-149, doi: 10.1109/ICSESS49938.2020.9237640.

[25] Phasinam, K., Kassanuk, T., & Shabaz, M. (2022). Applicability of Internet of Things in smart farming. Journal of Food Quality, 2022, 1–7. doi:10.1155/2022/7692922

[26] T. Li, Y. Guo and A. Ju, "A Self-Attention-Based Approach for Named Entity Recognition in Cybersecurity," 2019 15th International Conference on Computational Intelligence and Security (CIS), 2019, pp. 147-150, doi:10.1109/CIS.2019.00039