Fusing image feature and title semantic regularization with adversarial networks for cross-modal retrieval

Kai Zhou
School of Computer Sichuan University, Chengdu 610065, China
2019223045180@stu.scu.edu.cn

Abstract. With the improvement of living standard, people pay more and more attention to their health. Food is the foundation of human life, modern society places more emphasis on a balanced diet, which controls fat, protein, carbohydrate and vitamin intake. So food computing is more and more important. Food retrieval is one of the important research directions in food computing. On the basis of food retrieval, we can predict ingredients and instructions in each dish, according to the ingredients and instructions to speculate on the visual effect of cooking, which can guide human reasonable diet, analyse human diet structure and diet culture and so on. In this paper, we focus on cross-modal retrieval between food image and recipe. Firstly, we analyze the problems existing in the present method. Based on the problems existing in the existing method, a fusion image feature and title regularization with adversarial network is proposed, which uses the idea of generative adversarial to align the modes, fuses the local features and global features of the image, and adds the semantic regularity of title to improve the accuracy of the retrieval.

1. Introduction
Food is the foundation of human survival, food can not only provide us with energy, but also demonstrate our culture[1]. In digital times, people share what they eat on social media, slowly forming a food culture and generating massive food data. Food is closely related to people's customs, lifestyle, health and social life, so food computing is an important research topic. The domains of food computing are mainly divided into perception, recognition, retrieval, recommendation, prediction and monitoring[2]. Due to the variety of cooking methods, ingredients are often obscured, with different colour, forms and textures, which causes a high degree of intraclass variability, bringing great difficulties to various tasks of food computing. Thanks to the construction of rich large-scale food datasets[3,4,5], we can solve the problem about food computing by neural network.

There are three main types of retrieval: visual food retrieval, recipe retrieval and cross-modal recipe-image retrieval. We mainly research cross-modal recipe-image retrieval. We mainly research cross-modal recipe-image retrieval, where the main difficulty is the modality gap between image feature and text feature. The research work has used canonical correlation analysis (CCA)[6], the full connection layer maps the two modal features to the public space[5], generate adversarial network for modal alignment[7], latent variables to capture the interactions between modalities[8] and so on.

Despite of the progress, the above approaches all extract local feature, but global feature are ignored, which causes information loss. In addition, the above methods only use the features of ingredients and instructions, and do not use the title information about recipe.
2. Method

In order to solve the problems existing in the current methods, we made the following improvements on the basis of ACME\(^7\). We extracted the global feature of the image and fuse local feature and global feature by network. We use the title feature shared by the picture and the text as semantic regularization to let image feature and text feature tend to the title feature. The following figure 1 and figure 2 show the network architecture. Modality alignment, hard sample mining and translation consistency module can be found in ACME\(^7\).

Figure 1. The architecture of network.

2.1. Modality Alignment

Because there is a modal gap between image feature and text feature, it is necessary to align modal features before distance measurement. According to ACME, generative adversarial networks is used to align image features and text features. We take image feature extractor and text feature extractor as generator network, and train the network to make the discriminator network unable to distinguish whether feature representation is image feature or text feature. The objective \( L_{MA} \) is as follow:

\[
L_{MA} = \mathbb{E}_{i \sim p_{image}}[\log D_M(E_v(i))] + \mathbb{E}_{r \sim p_{recipe}}[\log (1 - D_M(E_r(r)))]
\]

\( E_v \) represents image feature extractor, \( D_M \) represents discriminator, \( E_r \) represents text feature extractor, \( i \) represents image feature, and \( r \) represents text feature.

2.2. Fusing local and global information sub-network

Convolutional neural network is locally perceptive, which can’t capture global information. Although the receptive field can be expanded through multi-layer convolutional neural network, it still extracts local information in essence. Non-local\(^9\) operation performs on image features, which can extract global feature. And its formula is as follow:

\[
y_i = \frac{1}{c(x)} \sum_j f(X_i, X_j)g(X_j)
\]
\(X_i\) is the channel vector of the image feature at a certain point in space. \(f(X_i, X_j)\) is the function to calculate the similarity between the two vectors, \(g(x)\) is the mapping function, and \(C(x)\) is the normalized coefficient. Here we set \(g(x)\) to be the \(1 \times 1\) convolution.

### 2.3. Hard sample mining

In order to let the model distinguish the samples that are most difficult to classify, a triple \((X_a, X_p, X_n)\) is selected, where \(X_a\) is anchor, \(X_p\) is the corresponding positive sample and \(X_n\) is the negative sample. And we choose the \(X_p\) that has the furthest distance with \(X_a\) and choose the \(X_n\) that has the closest distance with \(X_a\). The target \(L_{\text{Ret}}\) is as follow:

\[
L_{\text{Ret}} = \sum_v [d(V_a, R_p) - d(V_a, R_n) + \alpha] + \sum_{b} [d(R_a, V_p) - d(R_a, V_n) + \alpha]
\]

(3)

\(V_a\) represents image feature anchor, \(R_a\) represents text feature anchor and \(\alpha\) is the margin, \(d(X_i, X_j)\) is the metrics of distance.

### 2.4. Title semantic regularization

In order to prevent model overfitting and improve the semantic expression ability of the model, we add title semantic regularization. Firstly, we design the network to extract title feature to classify. After classification task, a pre-trained LSTM is used to extract title feature, then the title feature extracted are used to measure distance with image feature and text feature respectively. The title semantic regularization formula is as follow:

\[
L_{\text{titlle}} = d(X_{\text{title}}, X_{\text{image}}) + d(X_{\text{title}}, X_{\text{text}})
\]

(4)

Where \(d(X_i, X_j)\) is the distance between \(X_i\) and \(X_j\), \(X_{\text{title}}\) represent title feature, \(X_{\text{image}}\) represent image feature and \(X_{\text{text}}\) represent text feature.

### 2.5. Translation consistency

We hope the feature extracted can rebuild another modal feature, so we bring into generate tasks. For image feature, we let it generate ingredients and classify. For text feature, we let it generate image and classify. The objective \(L_{\text{Trans}}\) is as follow:

\[
L_{\text{Trans}} = L_{\text{r2i}} + L_{\text{i2r}}
\]

(5)

And \(L_{\text{r2i}}\) is as follow:

\[
L_{\text{r2i}} = \mathbb{E}_{i \sim P_{\text{image}}} [\log D_{\text{r2i}}(i)] + \mathbb{E}_{r \sim P_{\text{recipe}}} [\log(1 - D_{\text{r2i}}(G_{\text{r2i}}(r)))] + L_{\text{cr2i}}
\]

(6)

\(G_{\text{r2i}}\) is a generator that makes the text into image and \(L_{\text{cr2i}}\) is the text feature classify loss, and the same as \(L_{\text{i2r}}\).

### 3. Experiments

#### 3.1. Dataset

The dataset we used is Recipe1M[5], and the dataset distribution is shown below Table 1. Dataset is extracted from more than 20 popular cooking websites. It includes pictures, ingredients, instructions, titles, and summarized into 1048 categories. The dataset has a long tail, and according to statistics, the recipe consists of 9 ingredients and 10 instructions averagely. With the limitation of laboratory hardware, we randomly sample the data of each category under the condition that the category distribution is consistent, and the data volume is one-fifth of the original dataset.

| Table 1. The partition of Recipe1M. Number of samples in training, validation and test |
|-------------------------------|----------|----------|
| partition | recipe  | image    |
| training  | 720639  | 619508   |
| validation| 155036  | 133860   |
| test      | 154045  | 134338   |
3.2. Implementation Details
We use pre-trained ResNet-50 to be the image extractor, use LSTM to extract instruction feature and title feature, use bi-LSTM to extract ingredient feature considering that ingredients’ order doesn’t matter. The total loss is as follow:
\[
\mathcal{L}_{\text{total}} = \mathcal{L}_{MA} + \alpha \mathcal{L}_{R_{\text{Rec}}} + \beta \mathcal{L}_{\text{Trans}} + \gamma \mathcal{L}_{\text{title}}
\]  \hspace{1cm} (7)
Where \(\alpha, \beta\) and \(\gamma\) are hyper-parameter, and we set them 0.005, 0.002 and 0.005 respectively.

3.3. ACME sub-class result
In order to analyse ACME retrieval result of different categories, we select several categories and each category is randomly sampled 50 pairs, and then sample 950 pairs from other categories. The experimental result is shown as Figure 3.

![Figure 3. The results of ACME in the sub dataset.](image)

The effect of ACME retrieval on different categories is different, and it has little relationship with the proportion of categories. Through the analysis of the categories with poor effect, it is found that the food distribution is uniform and fine-grained, and the image feature extractor can not extract good features.

3.4. Ablation Studies
We make some ablation experiments to test our improvement effect, first of all, we test ACME’s effect on sub dataset, and then respectively join global and local information fusion sub-network and semantic regularization to assess whether the components have benefits. Finally we add two components together, take a look at how their effects. The following Table 2 shows the results of the ablation experiment.

| Methods   | Image to recipe retrieval | Recipe to image retrieval |
|-----------|---------------------------|---------------------------|
|           | medR | R@1 | R@5 | R@10 | medR | R@1 | R@5 | R@10 |
| ACME      | 5.1  | 24.5| 50.9| 62.9 | 5.1  | 25.1| 51.9| 63.0 |
| ACME+ FN  | 4.9  | 24.8| 51.7| 63.9 | 4.9  | 25.6| 51.8| 63.5 |
| ACME+TR   | 4.8  | 25.6| 52.1| 64.0 | 4.75 | 26.6| 52.4| 64.2 |
| ACME+FN+TR| 4.7  | 26.7| 52.9| 64.7 | 4.65 | 26.6| 52.9| 64.4 |

We can get result from Table 2, each of the modules we propose make sense. And we put two modules together get the best result.

4. Result
By analyzing the retrieval result of ACME, it can be concluded that ACME has poor retrieval effect on some categories those are fine-grained and relatively disorder. The addition of local and global information fusion sub-network can enable the image feature extractor to extract global information
and local information, and title semantic regularization can constrain image features and text features with title features. Both components improve the retrieval accuracy of the model.

References

[1] Fischler C, Food, Self and Identity[J]. Social Science Information, 1988, 27(2):275-292.

[2] Min W, Jiang S, Liu L, et al. A Survey on Food Computing[J]. 2018.

[3] Li J , Han F , Guerrero R , et al. Picture-to-Amount (PITA): Predicting Relative Ingredient Amounts from Food Images[J]. 2020.

[4] Ofli F , Aytar Y , Weber I , et al. Is Saki #delicious?: The Food Perception Gap on Instagram and Its Relation to Health[C]// International Conference. International World Wide Web Conferences Steering Committee, 2017.

[5] Salvador A , Hynes N , Aytar Y , et al. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[6] Harold Hotelling. Relations between two sets of variates. Biometrika, 28(3/4):321–377, 1936.

[7] Wang H , Sahoo D , Liu C , et al. Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images[J]. IEEE, 2019.

[8] Fu H , Wu R , Liu C , et al. MCEN: Bridging Cross-Modal Gap between Cooking Recipes and Dish Images with Latent Variable Model. 2020.

[9] Wang X , Girshick R , Gupta A , et al. Non-local Neural Networks[J]. 2017.