2017 IEEE Third International Conference on Multimedia Big Data

Learning Food Image Similarity for Food Image Retrieval

Wataru Shimoda Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi Tokyo 182-8585 JAPAN
Email: {shimoda-k,yanai}@mm.inf.uec.ac.jp

***Abstract*—For food application, recipe retrieval is an important task. However, many of them rely on only text query. Food image retrieval has relation to recipe retrieval so that similar food images are expected that they have similar recipes. Rising image retrieval performance is desired for recipe retrieval. On the other hand, to learn similarity by Siamese Network or Triplet Network are known as an effective method for image retrieval. However, there are no research for food image retrieval using similarity learning with Convolutional Neural Network as far as we know. Food recognition is known as one of fine-grained recognition tasks. Therefore it is unclear that how effective similarity learning methods based on CNN in food images. In our work, we trained some networks for feature similarity, and evaluated their effectiveness in food image retrieval.**

***Keywords*-deep learning, Siamese network, triplet network, food image recognition**

I. Introduction

Food image recognition is one of the promising applications of visual object recognition, since it will help estimate food calories and analyze people's eating habits for healthcare. Therefore, many works have been published so far [1]–[7]. To make food recognition more practical, increase of the number of recognizable food is crucial.

Meanwhile, recently the effectiveness of Convolutional Neural Network (CNN) have been proved for large-scale object recognition at ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012. Krizhevsky et al. [8] won ILSVRC2012 with a large margin to all the other teams who employed a conventional hand-crafted feature approach. Regrading food image recognition, the classification accuracy on the UEC-FOOD100 dataset [4] was improved from 59.6% [3] to 72.26% [9] by replacing Fisher Vector and liner SVM with DCNN.

In recent years, CNN have also achieved large progress in wide image recognition tasks. it was shown CNN is also effective. Especially image retrieval methods which deal an image as query and search images by similarity from a query image is receiving a lot of attention. In general, the similarity is calculated by distance of image features. In this image retrieval task, how generate features which are specific on visual appearance plays an important role.

Siamese Network [10] and Triplet Network [11] are methods to learn feature similarity by optimizing feature distance of image pairs or triplet. These learning methods provide better features in similarity. Especially Triplet Network is known as the state-of-the-art methods on image retrieval. These network was used in some application fields such like face authentication and clothes recognition and so on [12], [13].

For food application, recipe retrieval is important paradigm. But many of them rely on only inputting text. Food image retrieval has relation to recipe retrieval so that similar food images are expected that they have similar recipes. Rising image retrieval performance is desired for recipe retrieval. However, there are no research for food image retrieval using similarity learning such like Siamese Network and Triplet Network as far as we know. Food recognition is known as one of the fine-grained recognition task. Therefore it is unclear that how effective similarity learning methods based on CNN in food images. In our work, we trained some network for feature similarity and evaluated the potential in image retrieval task.

II. Method

Our purpose is to obtain good features for image retrieval. In this paper, we mainly tested three type networks as following:

- Simple Classification Network
- Siamese Network
- Triplet Network

For simply, these network construction is based on Alex Net. We illustrated the three type networks in Fig 1.

A. Similarity Learning Network

In this section we mention about Siamese Network and Triplet Network. These network goals are to learn image similarity models. We define the similarity of two images P and Q according to their Euclidean distance in the image embedding space:

$$D\left(f(P), f(Q)\right) = ||f(P), f(Q)||_2^2 \tag{1}$$

where $f(.)$ is the image embedding function that maps an image to a point in an Euclidean space, and$D(.,.)$ is

978-1-5090-6549-3/17 $31.00 © 2017 IEEE
DOI 10.1109/BigMM.2017.73

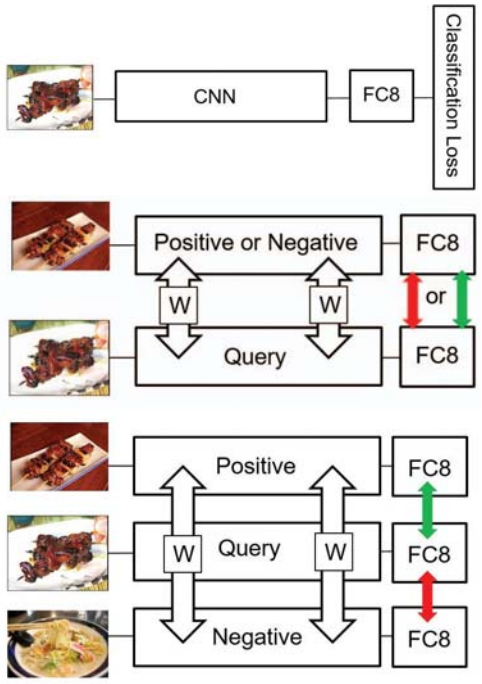

Figure 1. illustration for three types of networks. 1st network is general simple network for classification. 2nd network is Siamese Network. 3rd network is Triplet Network. Green means to minimize distance of features, while red means maximize distance of features.

the Euclidean distance in this space. The distance is smaller means image P and Q are more similar. These network solve this minimization problem by optimizing CNN parameters using Stochastic Gradient Decent (SGD).

1) Siamese Network: Siamese Network is proposed by Bromley et al. [10] in 1994. This network employs input as pair images and learns similarity in pair images. As function $f(.)$ Siamese Network extracts features from each pair images by CNN which weights are shared. When pair images are similar this network minimizes Euclid distance in pair images and when pair images aren't similar this network maximize Euclid distance in pair images after L2normalization. Suppose we have a set of images P and $p_i \in \mathbb{P}$. We can define the following loss function for a pair with label Y:$s_i = (p_i^1, p_i^2, Y_i)$:

$$\begin{aligned} l(p_i^1, p_i^2, Y_i) = Y_i \cdot D\left(f(p_i^1), f(p_i^2)\right) + \\ (1 - Y_i) max(0, C - D\left(f(p_i^1), f(p_i^2)\right), \\ \forall p_i^1, p_i^2, Y_i \in \{0, 1\} \end{aligned} \tag{2}$$

where C is a margin value and Y is a label for pair images are similar or not. If images are similar Y is 1, while if images are not similar Y is 0.

2) Triplet loss: Triplet Network is proposed by Wang et al. [11] in 2014. This network employs input as triplet images and learns similarity in triplet images. As function $f(.)$ Triplet Network extracts features from each triplet images by CNN which weights are shared. In triplet images we regards a base image as query and other one as a positive image and rest as a negative image. Suppose we have a set of images P and $p_i \in \mathbb{P}$, and $r_{(i,j)} = r(p_i, p_j)$ is a pairwise relevance score which states how similar the image $p_i \in \mathbb{P}$ and $p_j \in \mathbb{P}$ are. We can define the following loss for a triplet:$t_i = (p_i, p_i^+, p_i^-)$:

$$\begin{aligned} l(p_i, p_i^+, p_i^-) = max\left\{0, D\left(f(p_i), f(p_i^+)\right) - D\left(f(p_i), f(p_i^-)\right)\right\} \\ \forall p_i, p_i^+, p_i^- \text{such that } \mathrm{r(p_i, p_i^+)} > \mathrm{r(p_i, p_i^-)} \end{aligned} \tag{3}$$

Positive images should be same category, while negative images are separated to two patterns. The first pattern is different categories from query image category p_i. The second type is in-class negative samples, which are the negative samples that are in the same category as p_i but is less relevant to p_i^+. In order to ensure robust ordering between p_i^+ and p_i^- in a triplet $ti = (p_i, p_i^+, p_i^-)$, we also require that the margin between the relevance score $r(p_i, p_i^+)$ and $r(p_i, p_i^?)$ should be larger than the margin C as following:

$$\begin{aligned} r(p_i, p_i^+) - r(p_i, p_i^-) \geq C \\ \forall t_i = (p_i, p_i^+, p_i^-) \end{aligned} \tag{4}$$

B. Combine Classification Loss Term

For image retrieval with a query image, results sometimes include similar images to a query but different from query image class. In recipe retrieval which is one of the purpose of food image retrieval, since recipes of different category food are different, food category of retrieved images should be same to query image category. To solve this problem, we implement multitask network by combining similarity learning loss and classification loss. We illustrate our multi task network in Fig 2.

We suppose similarity loss term is L_s and classification loss term is L_{cl}. Over all loss L is expressed as following:

$$L = L_s + \lambda \cdot L_{cl} \tag{5}$$

λ is a parameter. In this paper, we simply set λ to 1.

III. Experiment

A. Dataset

In the experiments, we use the UEC-FOOD256 dataset [4]. UEC-FOOD256 dataset [4] consist of 256 class food category and each category have round 100 images.

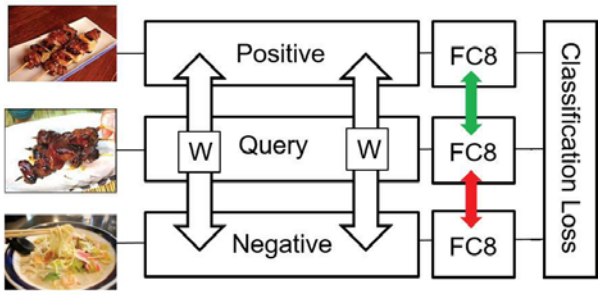

Figure 2. Illustration for combination of classification loss and similarity loss.

For testing, we use "val 0" which is one of five validations split list of UEC-FOOD256 dataset [4]. Of course we exclude validation images from training images.

B. Implementation Detail

We used the open source Chainer package to train our models. The initial learning rate is set to 0.001, and mini-batch size is 64. We multiply the learning rate by 0.1 at every 10000 iterations.

We used Alex Network as base Network models. In fine-tuning we initialize parameters excepting of last layer by publicly available pre-trained models.

C. Evaluation metric

We evaluate food image retrieval accuracy by Average Precision which is commonly used in retrieval evaluation. We define a correct image that is same class from a query image. In below tables, Top $\mathbf{K}$ mAP means mean of 256 classes results of Average Precision(mAP) value for ranked top $\mathbf{K}$ images.

D. Baseline and Options

For comparing we have three network options "Simple Network","Siamese Network" and "Triplet Network" and some branches for taking some learning metrics. I list our options as below:

- Pre-trained model (PT)
- Siamese Network with Pre-trained model (SN-PT)
- Triplet Network with Pre-trained model (TN-PT)
- Fine-tuned model (FT)
- Siamese Network with Fine-tuned model (SN-FT)
- Triplet Network with Fine-tuned model (TN-FT)
- Multi Task Siamese Network with Pre-trained model (MT-SN-PT)
- Multi Task Triplet Network with Pre-trained model (MT-TN-PT)
- Multi Task Siamese Network with Fine-tuned model (MT-SN-FT)
- Multi Task Triplet Network with Fine-tuned model(MT-TN-FT)

"Pre-trained model" is a trained CNN by ImageNet. This is a most simplest baseline. "with Pre-trained model" means parameters are initialized by pre-trained model. "Fine-tuned model" is a trained CNN by UECFOOD256 and parameters are initialized by pre-trained model. "with Fine-tuned model" means parameters are initialized by fine-tuned model. "Multi Task" means containing classification loss.

E. Evaluation of Food Image Retrieval

First of all, we compare three types of network. For simply, we pick up only three patterns which are initialized by pre-trained model and we used last fully connected layer as features. The performance comparison is shown in Table I. Superiority for food image retrieval of three types of network is clear. Siamese Network performance is better than a pre-trained model result, while Triplet Network more over Siamese Network performance with large margin. This results show similarity learning is effective but Triplet Network is more powerful than Siamese Network.

In the next place, on the contrary to the previous comparison, we evaluate combination of similarity learning and classification factor. Table II shows the results for the comparison. Fine-tuned model achieved high mean average precision at food image retrieval even though being trained for classification. On the other hand, we used fine-tuned model as initialization parameters. For example, at method of "SN-FT", we initialized Siamese Network parameters. However, the result is lower than "FT" score. This means that even though start better parameters, similarity learning hasn't guarantee for getting better results. We also evaluate classification factor by combining losses. As a results, combining classification loss and similarity loss is very effective. "MT-TN-FT" achieved a best result. Although mAP of "TN-PT" is already higher score than "FT", multi task loss showed further improvement. Classification factor is important and classification task and similarity learning can be well combined. For more detail, initialization with fine-tuned model showed better results than initialization with pre-trained model even though combining classification loss and similarity loss. It is expected that training with cascade steps will be better. We show examples in Fig. 3.

Table I
COMPARISON OF THREE NETWORKS RESULTS

method	PT	SN-PT	TN-PT
Top 10 mAP	9.7	11.3	25.7
Top 20 mAP	9.7	12.5	23.3

Table II
EVALUATION OF EFFECTIVENESS CLASSIFICATION FACTOR

method	FT	SN-FT	MT-SN-PT	MT-SN-FT	TN-FT	MT-TN-PT	MT-TN-FT
Top 10 mAP	24.8	18.0	23.3	26.4	27.1	27.8	31.7
Top 20 mAP	22.6	16.8	21.4	24.1	24.4	25.0	28.8

Figure 3. Image retrieval result examples. (a) PT. (b) SN-PT. (c) TN-PT. (d) FT. (e) SN-FT. (f) MT-SN-PT. (g) MT-SN-FT. (h) TN-FT. (i) MT-TN-PT. (j) MT-TN-FT.

IV. CONCLUSIONS

In this paper, we analyze effectiveness of similarity learning for food image retrieval. We tested three types of CNN and it was turned out that Triplet Network was the most powerful network compared to others. We also showed the performance of Triplet Network can be improved by combining classification task.

REFERENCES

[1] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 - mining discriminative components with random forests," in *Proc. of European Conference on Computer Vision*, 2014.

[2] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. of ACM International Conference Multimedia*, 2014, pp. 1085–1088.

[3] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, pp. 1–25, 2014.

[4] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2012, pp. 1554–1564.

[5] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," in *SIGGRAPH Asia*, 2012.

[6] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *Proc. of IEEE International Conference on Image Processing*, 2011.

[7] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks.," in *Advances in Neural Information Processing Systems*, 2012.

[9] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proc. of ACM UbiComp Workshop on Workshop on Smart Technology for Cooking and Eating Activities (CEA)*, 2014.

[10] J. Bromley, I. Guyon, E. Sckinger, and R Shah, "Signature verification using a " siamese " time delay neural network," in *Advances in Neural Information Processing Systems*, 1994.

[11] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[12] Junlin. H, Jiwen. L, and Yap-Peng. T, "Discriminative deep metric learning for face verification in the wild," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[13] Qian. Y, Feng. L, Yi. S, Tao. X, Timothy. H, and Chen. L, "Sketch me that shoe," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016.