IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2019, accepted December 18, 2019, date of publication January 8, 2020, date of current version January 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964836

Hybrid Attention-Based Prototypical Network for Unfamiliar Restaurant Food Image Few-Shot Recognition

GEGE SONG, ZHULIN TAO, XIANGLIN HUANG, GANG CAO, WEI LIU, AND LIFANG YANG

School of Computer Science and Cybersecurity, Communication University of China, Beijing 100024, China

Corresponding author: Xianglin Huang (huangxl@cuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61401408 and Grant 61772539, and in part by the Fundamental Research Funds for the Central Universities under Grant CUC2019B021 and Grant 3132017XNG1742.

ABSTRACT As eating-out became an indispensable part of our daily lives, demand for the food recognition of unfamiliar restaurant increased significantly due to health-care. Although there are many researches on generic food recognition, there are relatively fewer studies on restaurant food image recognition. Meanwhile, it becomes extremely challenging for restaurant food image recognition due to insufficient food image. Prototypical network is common utilized to address the such a task in recent years. Although the methods based on prototypical network achieve impressive results in capturing similarities feature of the same food category, it fails to highlight important information on feature and instance level. Toward this end, we propose an effective food image recognition scheme by incorporating hybrid attention mechanism into prototypical network in this paper. Specifically, the image feature is first captured by convolutional neural network (CNN). Then the image attention weights yielded by instance-based attention mechanism are used to modulate the image feature of CNN for constructing class prototypes. And feature-based attention mechanism is employed to grasp important information of image for enriching image representation. Extensive experimental results on the large food image dataset verify that the performance of our proposed classification scheme outperforms the state-of-the-art ones.

INDEX TERMS Restaurant food image recognition, unfamiliar restaurant, prototypical network, hybrid attention mechanism.

I. INTRODUCTION

Various restaurants and gourmet foods are filled with every corner of our lives. Tasting the diet of different restaurants has been becoming a life pattern in modern living, and people are more concerned now about the daily nutrition intake for health-care. Food is composed of ingredients which show diverse appearances in various cutting and cooking conditions. Restaurants use to highlight the uniqueness of a dish with a name reflecting its ingredients and cooking styles. Although each restaurant may have its own unique features, the appearances and names of a same dish cannot wildly vary between restaurants. Therefore, this allows people to check their intake status in a new unfamiliar restaurant by using food images of known restaurant for query and identification.

Several prior food image recognition methods have been proposed in the literature. The handcrafted features are extracted and fed into SVM model for food recognition [1]–[4]. As the development of deep learning, deep learning has recently achieved remarkable performance in many computer vision tasks [5]–[12], and convolutional neural network (CNN) has proven to be effective models [13]–[17]. The accuracy of generic food image recognition can be as high as 80% on the benchmark datasets such Food101 [1], FoodCam-256 [18] and VIREO Food-172 [19]. Such methods are mostly focused on recognizing a pre-defined set of food categories, ranging from 100 to 256 categories, known as generic food recognition [1]–[4], [19]–[22]. Restaurant information corresponding to food image is lacked for generic

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

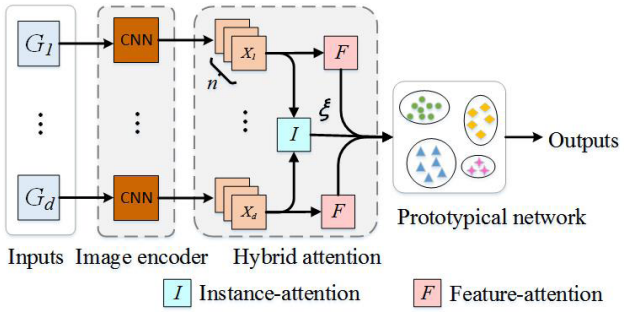


FIGURE 1. Architectures of our hybrid attention-based prototypical network.

food recognition. Nevertheless, restaurant information is very important for restaurant food image recognition. In other words, food category and restaurant information are fundamental to restaurant food image recognition. Meanwhile, extending the existing generic food recognition models to recognize all of the restaurant dishes is infeasible, since collecting sufficient image samples for all the restaurant food is intractable.

For restaurant food image recognition, there are relatively fewer studies on restaurant food image recognition compared with generic food recognition. Early works mostly focus on training image classifiers with single or multiple handcrafted low-level image features such as SIFT and HOG, and GPS which leverages as context information to discard unlikely candidates [23]–[25]. For example, Beijbom *et al.* [23] leverage GPS information to find the nearby restaurants and then map the food image to the menu item of nearby restaurants. Different to these early works, recent works mostly utilize deep features [26] for superior performance in recognition. In [26], deep features of restaurant food images are extracted to train a probabilistic model that connects dish, restaurants, and locations for dish recognition. The restaurant food recognition is tackled as a problem of image retrieval, and utilizes triplet network that considers fine-grained image similarity to learn features for image retrieval [27]. All of these methods have one thing in common, i.e., using GPS to limit the location and restaurant. People often go to a unfamiliarity place for tasting different flavors of food. Although GPS can locate automatically now, it still needs further manual selection to locate accurately. The inaccurate of GPS positioning likely to affect directly the recognition results. Therefore, unlike the above works, the proposed system will discard the location information only through restaurant food pictures for recognition, so that users can get food information when they are in a strange restaurant.

Aiming at the aforementioned problems, the prototypical network [28] and attention mechanism are introduced in our proposed scheme for restaurant food image recognition in this paper. The prototypical network is employed to address the overfitting of neural networks since data is severely limited. Attention mechanism is utilized to capture the important information. The proposed scheme is shown in Figure 1. Firstly, an embedding module generates representation of the image using a neural network. Secondly, in order to learn the better prototype of a class, a hybrid attention consisting of an instance-level attention and a feature-level attention is employed in our proposed model. The instance-level attention module is able to select more informative instances in the support set and denoise those noisy instances during training. The feature-level attention module can highlight important dimensions in the feature space and formulate specific distance functions for different relations, which enables our model to alleviate the problem of feature sparsity. With the hybrid attentions making the model focus more on those important instances and features, the proposed model becomes more effective and robust. Finally, classification is performed for an embedded query point by simply finding the nearest class prototype. The squared Euclidean distance is employed to compute distances between images. Dish models are trained by the dishes of different restaurants for dish recognition in new restaurants. Experiments show that the proposed scheme not only improve the recognition accuracy but also the efficiency.

The rest of this paper is organized as follows. Section II introduces the proposed framework. Experimental results and discussion are reported in Section III, which is followed by the conclusion drawn in Section IV.

II. THE PROPOSED SCHEME

We introduce the overall framework of our proposed hybrid attention-based prototypical networks in detail. The proposed model consists of three parts: image encoder, hybrid attention and prototypical networks.

A. IMAGE ENCODER

Given a food image taken in the restaurant, our goal is to retrieve the images that are in the same dish type with the query image. Through mapping the query image to the restaurant image database, the name of the query image can be obtained. Since the image recognition is posed as a retrieval problem, an essential step is to learn features effective for similarity measurement.

Deep learning has recently achieved remarkable performance in many computer vision tasks [6], [8]–[12], [29], and convolutional neural network (CNN) has proven to be effective models [14], [16], [17], [30]. CNN is a data-driven feed-forward and end-to-end multilayered neural network. By stacking a series of convolutional layers interleaved with non-linearities and downsampling, CNN is capable of capturing hierarchical patters with global receptive fields as powerful image descriptions. In this way, CNN not only can gain more reasonable discriminative features, but also reduce the errors incurred by artificial selection feature. Excellent models, such as AlexNet [14], Network-in-Network [31], VGG [32] and ResNet [7], have been developed. In this paper, ResNet-50 model is utilized to learn the food image features. The softmax layer of ResNet-50 is replaced with a fully connected layer to get the new image embedding features that are able to preserve the fine-grained image similarity.

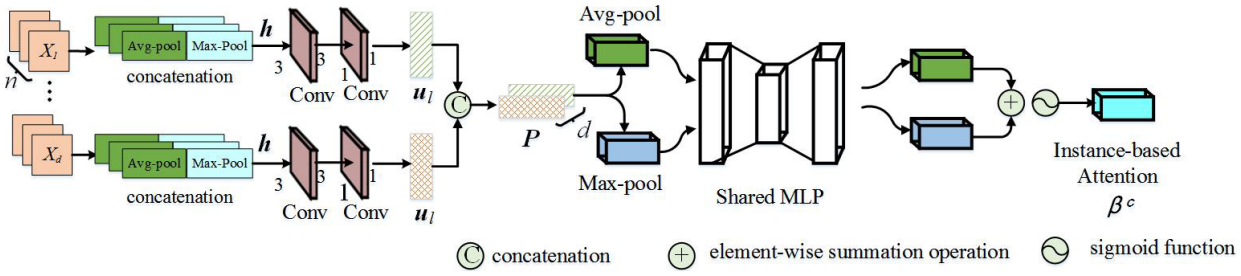


FIGURE 2. Illustration of proposed instance-attention mechanism. Conv denotes a convolutional layer.

B. HYBRID ATTENTION

Hybrid attention contains two parts: the instance- and feature-level attention mechanism. The prototype representations of each class can be calculated by capturing the key image yielding by the instance-attention mechanism. The feature-attention mechanism can highlight important information for generating better image feature.

1) INSTANCE-ATTENTION MECHANISM

The prototype representations of each class are the most important thing for the original prototypical network. Each food image is closely connected with the class prototype of its corresponding category. Hence, the generated prototype representation has a huge deviation from the expected prototype when data is noisy or relations cover diverse semantics. Meanwhile, it can be failed to recognize the unknown food image due to the poor robustness of the prototype. Therefore, the instance-attention mechanism can be designed to filter food images for capturing the key image and improving the performance of model. The structure of the instance-attention extractor is shown in Figure 2.

The food image feature $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ is extracted by image encoder, where n signifies the number of feature map, and $\boldsymbol{x}_i \in \boldsymbol{R}^z$ denotes each feature map. Suppose $\boldsymbol{F} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d]$ represents a certain type of image feature, where d signifies the number of images.

Average-pooling and max-pooling operations are firstly conducted to aggregate spatial information of a feature map. And two feature descriptors are generated, i.e., $\boldsymbol{x}^c_{avg}$ and $\boldsymbol{x}^c_{max}$, which signifies average-pooled features and max-pooled features respectively. Two feature descriptors are concatenated to built new feature maps $\boldsymbol{h}$ which remain the same number of maps, i.e.,

$$\boldsymbol{h} = [AvgPool(\boldsymbol{x}); MaxPool(\boldsymbol{x})] = [\boldsymbol{x}^c_{avg}; \boldsymbol{x}^c_{max}] \tag{1}$$

Then the concatenation feature $\boldsymbol{h}$ is conducted by convolution operation, we have:

$$\boldsymbol{u}_l = f^{1\times1}(ReLu(f^{3\times3}(\boldsymbol{h}))) \tag{2}$$

where $f^{3\times3}(\cdot)$ represents convolution operation with the filter size of 3×3, which can further capture the high-level representation. The number of filter reduces to the half of the original. The cross-channel information is integrated by the convolution operation with the filter size of 1×1 and the number of filter of 1, i.e., $f^{1\times1}(\cdot)$.

The features $\boldsymbol{u}_l$ of all images with the same category are integrated into a vector $\boldsymbol{P}$ with the number of maps of d. Subsequently, the average-pooling and max-pooling operations are operated again for the feature $\boldsymbol{u}_l$ for generating the descriptors $\boldsymbol{P}^c_{avg}$ and $\boldsymbol{P}^c_{max}$. Both feature descriptors are fed into a shared network for yielding the instance attention map $\boldsymbol{\beta^c} \in \boldsymbol{R}^{d\times1\times1}$. The multi-layer perceptron (MLP) with one hidden layer constitutes the shared network. Finally, element-wise summation operation is utilized to merge the output feature vectors.

$$\begin{aligned}\boldsymbol{\beta}^c &= \sigma(MLP(AvgPool(\boldsymbol{P})) + MLP(MaxPool(\boldsymbol{P}))) \\ &= \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{P}^c_{avg})) + \boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{P}^c_{max})))\end{aligned} \tag{3}$$

where σ denotes the sigmoid function, and $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ are parameters to be learned.

Finally, the prototype representations $\boldsymbol{\xi}$ that summarizes all the information of a category is obtained via a weighted sum of the image feature based on the weights:

$$\boldsymbol{\xi} = \sum_{i=1}^{d} \beta^c_i \boldsymbol{X}_i \tag{4}$$

The more key food images which have noteworthy category feature can be allocated higher weights via instance-attention mechanism. In such way, the prototypes generated by such images are similar to expected prototypes.

2) FEATURE-ATTENTION MECHANISM

Generally, the most existing methods based on deep learning are driven by a large number of data to learn class features. However, the number of image is extremely scarce for few-shot learning. Hence how to capture key features through a few number of images is very important. Feature-level attention mechanism is a common method in image processing. Some dimensions which are more discriminative for classifying in the feature space can be extracted by feature-attention mechanism. The structure of the feature-attention extractor is shown in Figure 3.

Firstly, the each map attention weights $\boldsymbol{\phi} = [\varphi_1, \ldots, \varphi_n]$ of image feature $\boldsymbol{X}$ are calculated by the equation (3). The importance of feature and weight value are in the direct ratio

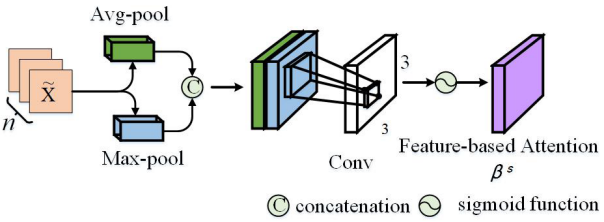


FIGURE 3. Illustration of proposed feature-attention mechanism. Conv denotes a convolutional layer.

and the important maps can be highlighted by weighting. The new image representation $\tilde{\boldsymbol{X}}$ can be computed by:

$$\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}_1}, \ldots, \tilde{\boldsymbol{x}_n}], \ where$$
$$\tilde{\boldsymbol{x}_i} = \varphi_i \boldsymbol{x}_i \tag{5}$$

Then, average-pooling and max-pooling operations are applied along the channel axis for the new features $\tilde{\boldsymbol{X}}$ and the generated features are concatenated to built an efficient feature descriptor. Meanwhile, a convolution layer is employed to generate a feature attention map $\boldsymbol{\beta}^s \in \boldsymbol{R}^{H\times W}$ on the concatenated feature descriptor. H and W are the height and weight of input image, respectively. It can be described as:

$$\begin{aligned}\boldsymbol{\beta}^s &= \sigma(([AvgPool(\tilde{\boldsymbol{X}}); MaxPool(\tilde{\boldsymbol{X}})]\boldsymbol{W}_2 + b_2)) \\ &= \sigma(([\tilde{\boldsymbol{X}}^s_{avg}; \tilde{\boldsymbol{X}}^s_{max}]\boldsymbol{W}_2 + b_2))\end{aligned} \tag{6}$$

where σ denotes the sigmoid function, and $\boldsymbol{W}_2$ and b_2 are parameters to be learned.

Finally, each image can be represented by $\boldsymbol{\psi}$:

$$\boldsymbol{\psi} = \sum_{i=1}^{n} \beta_i^s \boldsymbol{x}_i \tag{7}$$

where n denotes the number of feature maps and x_i represents the feature map of image.

C. PROTOTYPICAL NETWORK

The main idea of prototypical networks is to use one vector, also named prototype, to represent each relation. The prototype representation of each category $\boldsymbol{\xi}$ and the feature representation of image $\boldsymbol{\psi}$ are generated by instance- and feature-attention mechanism in our proposed method, respectively.

Prototypical network learns a metric space in which classification can be performed by computing distances to prototype representations of each class. Hence, distance function can significantly affect the capacity of prototypical network. As in standard setting [28], Euclidean distance is regarded as the metrics in our proposed model. We can compute the probabilities of the relations for the query instance x_q as follows,

$$p(y = i|x_q) = \frac{exp(-D(\psi_{x_q}, \xi_i))}{\sum_{j=1}^{d} exp(-D(\psi_{x_q}, \xi_j))} \tag{8}$$

where $D(\cdot, \cdot)$ is the Euclidean distance function for two given vectors.

TABLE 1. Few-shot results of proposed model with different setting of image encoder. Numbers are in percentage.

Models	5-shot (%)			10-shot (%)		
	Acc.	hit@2	hit@5	Acc.	hit@2	hit@5
AlexNet	42.98	44.26	55.65	43.92	45.16	56.73
VGG-16	44.55	46.56	56.41	46.08	47.76	58.11
GoogLeNet	47.00	49.77	59.96	48.27	51.42	63.01
ResNet-50	49.32	52.17	63.17	51.98	54.86	65.22

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASET, EVALUATION METRICS

Our proposed method is evaluated on a large restaurant database, which was crawled from two popular online restaurant review websites Burpple [1] and Openrice [2] which provide restaurant names and dish reviews with food images and tags. The dataset is a image collection of 314478 food images that cover 7,452 restaurants in Singapore. The goal of our proposed method is to obtain the dish names of a strange restaurant through the dish images and names of familiar restaurants. Therefore, the dataset is filtered by selecting at least 5 restaurants of each food category and each restaurant contains at least 15 pictures. This eventually yield a dataset of 9774 images collected elaborately from 31 food classes.

In order to thoroughly measure our model and the baselines, the accuracy (Acc.) and hit rate at top K (hit@K) are employed as evaluation metric. For each food image, we generate the highest ranked labels and compare the generated labels to the ground truth labels. The accuracy is the number of correctly annotated labels divided by the number of generated labels. Hit@K calculates the fraction of times that a correct food image is found in top-K retrieved candidates.

B. EXPERIMENTAL DETAILS

All models in our experiment are trained and tested based on Tensorflow and conducted on a 64-bit Ubuntu 16.04 Server with Intel Xeon(R) CPU E5-1603 v3 at 2.80 GHz 4 on 20 GB RAM and NVIDIA GeForce GTX Tesla GPU support. The network is trained from scratch using the annotated training images. ReLU is used as the activation function. For all models, we selected Adam as the optimizer. The weights of all network parameters are initialized by Gaussian distribution with standard deviation as 0.005. We train our model using stochastic gradient descent with momentum of 0.9, and weight decay of 0.0005. The learning rate is initialized at 0.001. To perform fair comparison with state-of-the-art methods, we trained our model and the baselines over the same training set and verified them over the same testing one. And we repeated the experiment ten times for each model and reported the average testing results.

[1] www.burpple.com/sg
[2] sg.openrice.com

TABLE 2. Few-shot results of proposed model with different setting of attention mechanism. Conv denotes convolutional layer. Numbers are in percentage.

Models		5-shot			10-shot		
		Acc.	hit@2	hit@5	Acc.	hit@2	hit@5
Without instance- and feature-attention mechanism		43.78	46.86	57.77	45.98	49.06	62.55
Only with instance-attention	Conv(1 layer)	43.66	46.98	58.01	46.01	49.33	63.02
	Conv(2 layer)	45.31	47.89	59.21	48.06	50.39	65.46
	Conv(3 layer)	44.98	47.29	58.99	47.67	48.74	64.02
Only with feature-attention	Conv(1 layer)	45.33	49.02	60.13	48.09	54.22	65.03
	Conv(2 layer)	44.64	48.23	59.61	46.97	52.68	63.92
	Conv(3 layer)	44.67	48.11	59.47	47.00	53.01	63.88
With instance- and feature-attention mechanism		49.32	52.17	63.17	51.98	54.86	65.22

C. PERFORMANCE COMPARISON WITH IMAGE ENCODER

The influence of image encoder on food image recognition is investigated. We compare four commonly used networks:

- AlexNet: The architecture of AlexNet contains eight layers with weights. The first five layers are convolutional and the remaining ones are fully connected.
- VGG-16: The architecture of VGG contains sixteen layers with weights. The first thirteen layers are convolutional and the remaining ones are fully connected. Unlike the AlexNet, it only uses convolutional filters of size 3×3.
- GoogLeNet: The architecture of GoogLeNet contains twenty-two layers with weights. The fully connected layer is replaced by the average pooling layer.
- ResNet-50: The architecture of ResNet contains fifty layers with weights. The deep residual learning framework was introduced to address the degradation problem.

As observed from Table 1, the model with ResNet-50 outperform AlexNet, VGG-16 and GoogLeNet with a clear margin in our experiments and the AlexNet-based models yield worse accuracies. There are significant gaps among the performances of models in term of hit@5. It can be explained by their network architecture, which enjoys more layers to capture more structure information with high-level semantics. Resnet-50 has a more elaborate structure compared with others. Deep residual learning is proposed to avoid the problem of "degradation" of network learning ability with the increase of network layers. Therefore, Resnet-50 is used to extract the features of the image in our proposed models.

D. PERFORMANCE COMPARISON WITH HYBRID ATTENTION

The influence of hybrid attention mechanism with different setting of convolutional layer on classification result is evaluate.

- Proposed without hybrid attention: The model without instance-attention and feature-attention, i.e., prototypical network.
- Proposed only with instance-attention: The proposed model only contains instance-attention and the number of convolutional layer is set to 1, 2 and 3. *Conv(1 layer)* denotes the instance-attention only has one convolutional layer with the filter size of 3×3. *Conv(3 layer)* denotes the instance-attention increases a convolutional layer with the filter size of 3×3.
- Proposed only with feature-attention: The proposed model only contains feature-attention and the number of convolutional layer is set to 1, 2 and 3. *Conv(1 layer)* denotes the feature-attention has one convolutional layer.
- Proposed with hybrid attention: The instance-attention of two convolutional layers and instance-attention of three convolutional layers are involved in the proposed model.

Hybrid attention contains two parts: instance- and feature-attention mechanism in our proposed model. Table 2 tabulates the classification results on the test set. Using any of the individual attention mechanism alone reports better performance than the model without hybrid attention mechanism. It proves the effectiveness of the both attention mechanisms. The model with instance-attention mechanism of two

TABLE 3. Few-shot results of proposed model with different setting of distance functions. Numbers are in percentage.

Distance	5-shot (%)			10-shot (%)		
	Acc.	hit@2	hit@5	Acc.	hit@2	hit@5
Euclidean	49.32	52.17	63.17	51.98	54.86	65.22
Manhattan	42.41	45.92	53.25	44.97	47.61	56.97
Chebychev	39.87	43.41	50.58	43.01	46.27	52.49
Mahalanobis	39.92	43.12	50.89	42.74	46.16	53.19

TABLE 4. Few-shot results of different models. Numbers are in percentage.

Models	5-shot			10-shot		
	Acc.	hit@2	hit@5	Acc.	hit@2	hit@5
Prototypical Network [28]	43.78	46.86	57.77	45.98	49.06	62.55
RN [43]	44.25	50.56	60.01	47.36	51.27	63.07
our (Tanh)	48.46	51.22	62.11	51.13	53.07	65.10
Our (ReLu)	**49.32**	**52.17**	**63.17**	**51.98**	**54.86**	**65.22**

TABLE 5. Few-shot results of different models in term of different way and different shot. Numbers are in percentage.

Models	5-way Acc.(%)			10-way Acc.(%)			31-way Acc.(%)		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Prototypical Network [28]	53.98	56.86	59.36	46.00	50.21	56.17	36.98	43.78	45.98
RN [43]	55.55	56.56	61.20	47.72	52.46	57.37	38.55	44.25	47.36
Our	**56.47**	**60.33**	**63.82**	**50.89**	**54.26**	**59.17**	**38.90**	**49.32**	**51.98**

convolutional layers and the model with feature-attention mechanism of one convolutional layers achieve the best results, respectively. The attention mechanism based on instance can generate better class prototypes, and the more effective image features can be extracted by the feature-attention mechanism.

Meanwhile, the combination of instance- and feature-attention mechanism (i.e., with hybrid attention) reports the best performance by taking the advantages of both attention. In addition, Table 2 verifies that instance- and feature-attention mechanism are useful in accounting for the overall food image classification performance, since removing any one of them results in performance degradation. We note that feature-attention mechanism plays a more important role due to the larger drop in comparison to instance-attention mechanism.

E. PERFORMANCE COMPARISON WITH DISTANCE FUNCTIONS

The influence of distance functions on image classification is investigated. We compare four commonly used distance functions: Euclidean, Manhattan, Chebychev and Mahalanobis. The classification accuracies using different distance functions are shown in Table 3. The performance of model with euclidean distance is significantly better than others on test dataset, outperforming the closest result by more than 6.25%. This may be explained by the sparsity of extracted image features due to insufficient image.

F. PERFORMANCE COMPARISON WITH MODELS

In order to demonstrate the effectiveness of our proposed model, several existing state-of-the-art methods are compared as follows:

1) PROTOTYPICAL NETWORKS [28]

Prototypical network learn a metric space in which classification can be performed by computing distances to prototype representations of each class.

2) RN [34]

Relation Network (RN) learns to learn a deep distance metric to compare a small number of images within episodes, each of which is designed to simulate the few-shot setting.

The comparison of proposed method with those in the literature is presented in Table 4. Generally speaking, our method performs better than other methods with significant margins. In particular, our method can gain the respectful improvements in the dataset. Although the proposed model is slightly superior to prototypical network in term of hit@5 (10-shot), it achieves higher accuracy over prototypical network by at least 5% in other experiments. Moreover, compared with the relational network, the performance of our method is improved by at least 2%. The performance improvements of proposed method are mainly due to the effectiveness of the proposed model. Meanwhile, it also further proves the effectiveness of the hybrid attention mechanism. The attention mechanism based on instance can generate better class prototypes, and the more effective image features can be extracted by the feature-attention mechanism.

In order to investigate the setting of activation function, we compare different models where activation function is set as Tanh and ReLu respectively. As observed from Table 3, the model with ReLu is superior to the model with Tanh. This is mainly because ReLu exhibits prominent ability of inhibiting varnishing gradient, making the training of deep network less difficult.

Moreover, for further investigating the performance of proposed method, we compare the state-of-the-art methods where the number of categories (i.e., the value of *way*) is set as 5, 10 and 31 respectively and the number of images (i.e., the value of *shot*) is set as 1, 5 and 10 respectively. As observed from Table 5, the accuracy of our proposed model outperforms other state-of-the-art baselines on all the experiments. The performances of proposed method prove the effectiveness of our model again.

IV. CONCLUSION

Few-shot learning becomes increasingly essential in the absence of actual data. In this paper, we propose a hybrid attention-based prototypical network for unfamiliar restaurant food photo few-shot recognition. Food image features are first obtained by convolutional neural network. And then the instance-based attention mechanism is used to get the class prototype of prototypical network, while feature-based attention mechanism can enable model to capture feature-level features more effectively. Image features are directly integrated into attention weights generated by feature-based attention mechanism for constructing better image representation. Evaluation results on the benchmark food image classification dataset demonstrate that our proposed scheme achieves higher accuracy and outperforms some state-of-the-art methods. As future work, more detailed experiments will be conducted to evaluate performance of the hybrid

attention-based prototypical networks including image feature extraction and the construction of attention mechanism.

REFERENCES

[1] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101Í mining discriminative components with random forests," in *Proc. ECCV*, 2014, pp. 446–461.

[2] T. Joutou and K. Yanai, "A food image recognition system with Multiple Kernel Learning," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 285–288.

[3] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. Adjunct Publication (UbiComp)*, Seattle, WA, Australia, 2014, pp. 589–593.

[4] D. T. Nguyen, Z. Zong, P. O. Ogunbona, Y. Probst, and W. Li, "Food image classification using local appearance and global structural information," *Neurocomputing*, vol. 140, pp. 242–251, Sep. 2014.

[5] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 838–846.

[6] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2956–2964.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV* , Amsterdam, The Netherlands, 2016, pp. 630–645.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[9] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS,* Montreal, QC, Canada, 2014, pp. 2204–2212.

[10] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1653–1660.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1725–1732.

[12] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. 2nd Int. Conf. Hum. Behav. Unterstanding*, Amsterdam, The Netherlands, 2011, pp. 29–39.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. NIPS*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[18] Y. Kawano and K. Yanai, "FoodCam-256: A large-scale real-time mobile food RecognitionSystem employing high-dimensional features and compression of classifier weights," in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA, 2014, pp. 761–762.

[19] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. ACM Multimedia Conf. (MM)*, Amsterdam, The Netherlands, 2016, pp. 32–41.

[20] K. Aizawa and M. Ogawa, "FoodLog: Multimedia tool for healthcare applications," *IEEE Multimedia Mag.*, vol. 22, no. 2, pp. 4–8, Apr. 2015.

[21] Z. Ming, J. Chen, Y. Cao, C. Forde, C. Ngo, and T. S. Chua, "Food photo recognition for dietary tracking: System and experiment," in *Proc. MMM*, Bangkok, Thailand, 2018, pp. 129–141.

[22] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Turin, Italy, Jun. 2015, pp. 1–6.

[23] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-match: Restaurant-specific food logging from images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2015, pp. 844–851.

[24] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2015, pp. 580–587.

[25] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain, "Geolocalized modeling for dish recognition," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015.

[26] L. Herranz, S. Jiang, and R. Xu, "Modeling restaurant context for food recognition," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 430–440, Feb. 2017.

[27] Z. Wei, J. Chen, Z. Ming, C.-W. Ngo, T.-S. Chua, and F. Zhou, "DietLens-Eout: Large scale restaurant food photo recognition," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, Ottawa, ON, Canada, 2019, pp. 399–403.

[28] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 4077–4087.

[29] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Nancy, France, Aug. 2015, pp. 991–995.

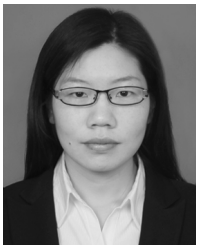
[30] C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Kyoto, Japan, Nov. 2017, pp. 388–393.

[31] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

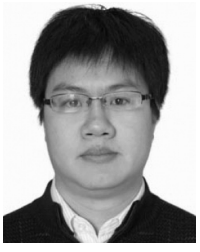
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[34] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1199–1208.



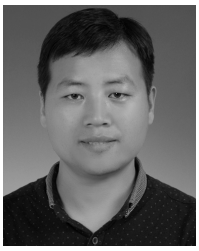
GEGE SONG received the B.S. degree from Shandong Technology and Business University, Shandong, China, in 2013, and the master's degree from the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China, in 2016, where she is currently pursuing the Ph.D. degree. Her current research interests include image processing and natural language processing.



ZHULIN TAO received the master's degree from the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China, where he is currently pursuing the Ph.D. degree. His current research interest includes image processing.

XIANGLIN HUANG received the B.S. and M.S. degrees from Jilin University, Changchun, Jilin, China, and the Ph.D. degree from the Beijing University of Technology, Beijing, China. He is currently a Professor with the School of Computer Science and Cybersecurity, Communication University of China. His research interests include document image retrieval, content-based image retrieval, compressed domain image processing, intelligent signal processing, and multimedia information processing.

GANG CAO received the B.S. degree from the Wuhan University of Technology, Hubei, China, in 2005, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2013. He was with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2010. He is currently an Associate Professor with the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China. His current research interests include digital forensics, data hiding, and multimedia signal processing.



WEI LIU received the B.S. degree in computer science and technology from Xinyang Normal University, Henan, China, in 2006, and the master's degree in computer application and technology from the China University of Geosciences, Wuhan, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China. His current research interests include multimedia content analysis and deep learning.

LIFANG YANG received the B.S. degree in electronic information from Qingdao University, Qingdao, Shandong, China, in 2005, and the M.S. and Ph.D. degrees in signal and information processing from the Communication University of China, Beijing, China. She is currently an Associate Professor with the Communication University of China. Her research interests are mainly in intelligent retrieval and high-dimensional index structure.