Fusion Learning using Semantics and Graph Convolutional Network for Visual Food Recognition

Heng Zhao, Kim-Hui Yap, and Alex, Chichung Kot

School of Electrical and Electronic Engineering, Nanyang Technological University
50 Nanyang Ave, Singapore 639798
{zhao0248,ekhyap,eackot}@ntu.edu.sg

Abstract

Food-related applications and services are essential for the health and well-being of people. With the rapid development of social networks and mobile devices, food images captured by people can offer rich knowledge about the food and also necessary dietary assistance for people that require special care. Known food recognition frameworks and approaches in computer vision have heavy reliance on many-shot training of a deep network on existing large-scale food datasets. However, it is common for many food categories that it is difficult to collect enough images for training. Traditional few-shot learning is unable to properly address the problem due to the complex characteristics and large variations of food images, and most few-shot frameworks cannot perform classification for many-shot and few-shot categories at the same time. In this paper, we propose a new fusion learning framework for food recognition. It unifies many-shot and few-shot under a single framework, by leveraging on extracted image representations and context sensitive semantic embeddings. Further, considering food categories are often correlated to each other for many commonalities such as same ingredients, cooking methods, the fusion learning framework utilizes a Graph Convolutional Network (GCN) to capture the inter-class relations between both image representations and semantic embeddings of different food categories. The final output fusion classifier will be more robust and discriminative. Comprehensive experimental results on two popular food benchmarks have shown the proposed framework achieves the state-of-the-art fusion performance.

1. Introduction

Food-related study gains increasing popularity for its importance in people's life. Understandings of daily food intakes can greatly benefit the health of people and also help in personal dietary management. Over the past few years, great progress has been made in food recognition using just food images, thanks to the rapid advances in the development of powerful deep learning networks.

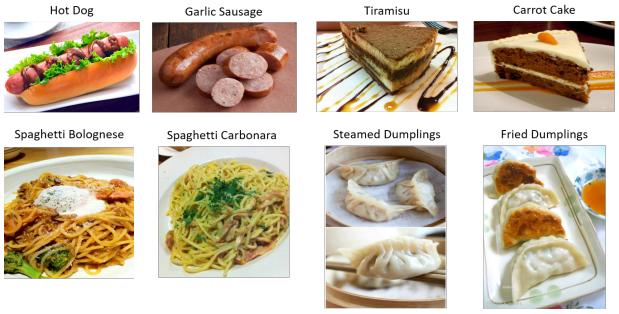

Figure 1. Some food image examples with class labels.

Traditional food recognition approaches capture handcrafted global and local image features, such as SIFTs, Local Binary Patterns (LBP) [27]. K-Nearest Neighbor (k-NN) [12], Support Vector Machine (SVM) [27] and random forest techniques [2] are amongst the commonly used classifiers. In contrast to the traditional approaches, simple and automatic food recognition approaches using deep networks give better performance in general [30, 10, 40]. It can be applied for various food applications, such as mobile visual food recognition, food logging and nutrition analysis services.

Visual food recognition using images, just like other types of fine-grained recognition tasks, often suffers from the difficulty to get discriminative image representations with good generalization ability for each class. Food recognition tasks are especially difficult to be addressed since many categories of food do not have distinctive appearance or obvious layout structure. A common strategy to alleviate the problem is by introducing additional information during

training and testing. For example, ingredient information associated with the food images, cooking recipes from Internet, and geographical locations of restaurants [44, 46]. Although it is proven the food recognition performance can be improved by incorporating such information, these approaches are often considered as less cost-effective given the extra human labors and other costs required to collect the additional information, and also sometimes, the sources may not be available.

To make the issue worse, it is common that many food categories cannot be collected with enough image data to train the deep networks. Few-shot learning is designed to address the issue of limited training data by either fully exploiting the features can obtained from the few sample images or introducing additional side information [18, 45, 17, 33]. Unfortunately, existing few-shot learning frameworks often under-utilize the possible inter-class relations between different fine-grained categories such as in food recognition. As a result, they struggle to properly classify food images.

In order to address the abovementioned issues, in this work, we propose a fusion learning framework that unifies both many-shot and few-shot classification for visual food recognition, by leveraging on image representations and context sensitive semantic embeddings. It conducts a two-stage fusion learning which fuses many-shot and few-shot learning under a single framework, where the final classifier can perform food recognition on both many-shot and few-shot categories. The first stage of fusion learning leverages on extracted image representations from a state-of-the-art deep network and semantic embeddings based on class labels. The motivation to incorporate class label semantic embeddings is that the labels of food names are the easiest accessible information which contain important characteristics of the food, such as dedicated food terminologies, common ingredients and cooking methods,etc.

Figure 1 shows some commonly seen food categories but that are not easy to distinguish each pair using only image features. For example, "Hot Dog" is a dedicated food name, with the dominant component of "Sausage". In order to correctly classify it without confusing with other food with "Sausage", we can generate such context sensitive embedding based on the class label "Hot Dog", so that it provides additional semantic meaningful information. Here the context sensitive refers that the embedding is generated based on context, in this example, it is in the context of food, not the animal. Other examples in the figure also show that the additional information given by the class label can be used in food recognition together with image representations. As a result, we propose to generate context sensitive semantic embeddings for each food category using the class labels. Therefore, the performance of the trained network can be improved by injecting such side information.

Furthermore, in order to capture possible inter-class relations and correlations of different food categories, the second stage of fusion learning of the proposed framework incorporates Graph Convolutional Network (GCN) as a mean to map the relations in terms of the similarity of the image representations and semantic embeddings for different food categories. The final resulting GCN representations are more robust and discriminative, which further boost the performance of food recognition. The contributions of this paper can be summarized as follows:

- We propose a fusion learning framework for visual food recognition, which is a two-stage fusion unifies both many-shot and few-shot learning. Its performance is enhanced by incorporating image representations from a state-of-the-art deep network and context sensitive semantic embeddings of class labels during training. A final robust classifier can be produced by further capturing the inter-class relations using GCN.

- We conduct comprehensive experiments on two popular food benchmark datasets, with different configurations and ablation study. The experimental results show that our approach can achieve state-of-the-art performance on both datasets.

- We design the fusion learning framework to be compatible with different types of backbone Convolutional Neural Networks (CNNs), we have demonstrated the effectiveness of the proposed framework using ResNet [11] and EfficientNet [39]. It could be further expanded to real-life food applications

2. Related Work

2.1. Visual food recognition

Visual food recognition using deep learning framework emerges as one of the most popular food-related studies. In recent years, automatic food recognition by simply scanning or capturing food images allows many health applications to be developed. For example, nutrition analysis [28], diet management [3, 21], food recommendation [6]. A comprehensive survey on food computing is provided by Min et al. [24]. Most existing methods deploy popular CNN architectures such as AlexNet [16], GoogLeNet [38], ResNet [11], and perform food classification by using the image features obtained from the networks. Aguilar et al. [1] evaluated fusion of different networks/classifiers for food recognition. Martinel et al. [19] proposed to use wide-slice ResNet for food recognition by leveraging on the vertical traits in some food.

Combinations of deep visual features and other useful information have also been explored. Some recent works have formulated food recognition as a multi-task classification problem where the recognition of food is attempted

along with other attribute recognition, such as GPS location, ingredients, types of cuisine, cooking recipes [41, 25]. Jiang et al. [13] designed a Multi-Scale Multi-View Feature Aggregation (MSMVFA) for food recognition, which combines features and semantics at different detail levels. Incorporation of such contextual information is proven to improve the food recognition accuracy. However, these additional attributes require huge extra human labors to collect and construct, and they may not always be available.

Different from existing works, we propose to generate context sensitive semantic embeddings based on food class labels, where the labels are easily accessed information during training. In the proposed method, both image representations obtained from state-of-the-art deep networks and the context sensitive semantic embeddings are used to train the framework together.

2.2. Few-shot learning

Few-shot learning, the ability to recognize few-shot/novel categories given only a few sample images, is a hot deep learning research topic for its wide applications. Few-shot learning can be the potential solution for many visual recognition tasks such as food recognition, where scarcity of data is usually a problem.

Few-shot learning is developed from the early-stage of learning good model initialization, typical works like Model-Agnostic Meta-Learning (MAML) [7] and Latent Embedding Optimization (LEO) [33], so that the classifiers for novel classes can be learned with a limited number of labeled examples and a small number of gradient update steps. Metric learning based few-shot frameworks become more popular for its easier implementation and superior performance. Typical works such as Prototypical Networks by Snell et al. [36] and RelationNet by Sung et al. [37] utilize the distance or similarity between trained images, so it can classify an unseen input image with the labeled instances based on distance metric learning. Few-shot learning with additional semantics has also been explored. Schwartz et al. [35] proposed a few-shot learning approach using multiple rich semantics, such as labels, attributes, and natural language descriptions. TAFE-NET proposed by Wang et al. [42] addressed the image representation adaptation to other tasks by learning task-aware feature embeddings. However, most existing few-shot learning approaches fail to capture inter-class correlations that are important for recognizing novel categories.

In our works, we utilize the labels of food categories to generate context sensitive semantic embeddings, that does not require any rich semantics like attributes and language descriptions. We also manage to create a fusion system that fuses both many-shot and few-shot learning under a single framework, which leverages on both robust image representations and class label embeddings. We capture the inter-class relations in term of the similarity of the image representations and semantic embeddings between different food categories, by constructing and training a GCN.

3. Proposed Methodology

3.1. Overview of framework

As shown in Figure 2, our proposed framework has three major components: (i) generation of class label embedding: given a set of image-label pairs that contains both many-shot categories and few-shot categories, class label of each food category is used to generate context sensitive semantic embeddings ($\boldsymbol{t}$) by Bidirectional Encoder Representations from Transformers (BERT) [5]; (ii) fusion of many-shot and few-shot learning: images of many-shot categories are used to train a CNN, which can be divided into a feature extractor and a classifier. In addition to the normal classification loss, a new semantic loss based on the generated class label embeddings of many-shot categories is introduced during training. The parameters of the classifier can be extracted as classification weights ($\boldsymbol{w}^m$), which are used to classify many-shot categories. After the training, the feature extractor will be used to generate the prototypical vector ($\boldsymbol{v}$) for each few-shot category based on image feature aggregation; (iii) inter-class relation learning using GCN : after obtaining the overall classification weights ($\boldsymbol{w}$) and semantic embeddings, both representations will be used as input to construct a GCN in order to capture the inter-class relations. New class representation ($\boldsymbol{c}$) can be obtained after training the GCN. Finally, the classifier is finetuned with classification loss and a new matching loss between $\boldsymbol{w}$ and $\boldsymbol{c}$.

3.2. Generation of class label embedding

Semantic embedding usually refers to the text/word embedding generated by various nerual networks. It is a popular strategy to solve Natural Language Processing (NLP) tasks by learning good text embedding using Pre-Trained Models (PTMs). Based on definitions in the survey [31], the first generation of such PTMs which includes Skip-Gram [23], word2vec [22] and GloVe [29], are usually shallow in network architecture. They are also not context sensitive, which often fail to capture higher-level concepts and correlations between words in context. The second generation PTMs has deeper architectures and better training techniques, which improve their ability to learn contextual word embeddings, such as CoVe [20], OpenAI GPT [32] and BERT [5].

Although some recent works explored the possibility of enhancing image classification and retrieval using word embeddings, they either are based on traditional many-shot training using large-scale datasets [14] or simply applying word embeddings from pre-trained model without considering the rich dependency between semantic embeddings with

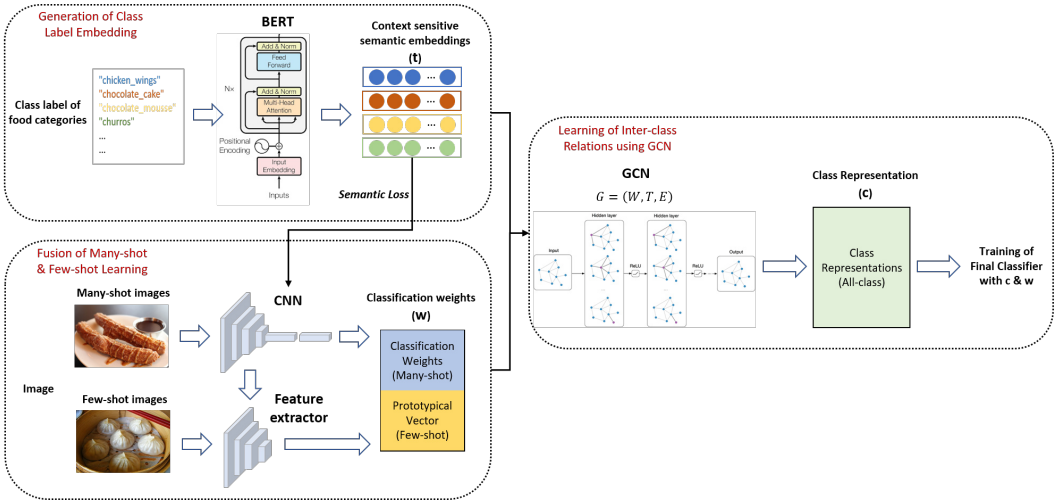

Figure 2. Overview of our proposed fusion learning framework. It consists of three branches, one for generation of context sensitive semantic embeddings, one for many-shot & few-shot fusion learning and the last is inter-class relation learning using Graph Convolutional Network (GCN).

image features [26].

Given a labeled dataset of K categories that contains K^m many-shot categories and K^f few-shot categories, each image-label pair in the dataset can be denoted as (x^m, y^m) for many-shot categories and (x^f, y^f) for few-shot categories, where x is the image and y is the corresponding label with different word length u. In this work, we aim to encode label information to semantic embeddings to better measure the semantic relations of different food category. The obtained semantic embeddings are used to train the fusion learning framework. We use the architecture of BERT, which is designed to pre-train deep bidirectional representations from text by jointly conditioning on both left and right context in all layers. BERT is trained with two unsupervised text tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP) on plain corpus of BooksCorpus and English Wikipedia, where MLM is used to enhance its efficiency at predicting masked tokens and NSP is used to generate a sequence prediction rather than a token prediction. As compared with other embedding methods, BERT based embedding is context sensitive, where the representations of the phrase or word are closely related to the context.

Let $\boldsymbol{t}$ denote the resulting embedding of projecting the label y to the embedding space ($\boldsymbol{t} \in \mathbb{R}^{1\times d}$). We can obtain a sequence of tokens $\{y_1, y_2, ..., y_a\}$ as input to the Transformer encoder, where a denotes number of word in y. Class label embedding is given by the sequence embedding of y from pre-trained BERT model:

$$\boldsymbol{t} = \frac{1}{a}\sum_{i}^{a} \boldsymbol{e}(y_i) \tag{1}$$

where $\boldsymbol{e}$ denotes the token value in the second last hidden layer in the pre-trained model.

3.3. Fusion of many-shot and few-shot learning

A standard CNN architecture can be split into two main components: a feature extractor and a classifier. Given an input image x that belongs to a dataset of K classes, the feature extractor will generate a d-dimensional feature $\boldsymbol{z}$ ($\boldsymbol{z} \in \mathbb{R}^{1\times d}$) which then the classifier will compute the raw classification scores before softmax $\{s_1, s_2, ..., s_K\} = \{\boldsymbol{z}^T\boldsymbol{w}_1, \boldsymbol{z}^T\boldsymbol{w}_2, ..., \boldsymbol{z}^T\boldsymbol{w}_K\}$, where the parameters $\boldsymbol{w} = \{\boldsymbol{w}_i\}_{i=1}^{K}$ are the classification weights that are used for classifying different categories. Common approaches in many-shot learning use dot-product based classifier and SGD to update the weights in the layers progressively. However, few-shot learning cannot use conventional classification weights due to lacking of data. Instead, we generate the prototypical vectors using the trained feature extractor for each few-shot category by utilizing the feature vectors of the few training images. Similar classification score is obtained by measuring the similarity between the query image and the prototypical vector of each category.

Let $\boldsymbol{v}_k$ denote the prototypical vector for k-th few-shot category by averaging input image feature vectors, where N_k is the number of images available for the few-shot category:

$$\boldsymbol{v}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{z}_i \tag{2}$$

The magnitude of the prototypical vectors depend on the input image feature vectors, and the raw classification scores of many-shot and few-shot are different in magnitude. To create a unified fusion framework of many-shot and few-shot, we adopt cosine similarity based feature classifier

which is inspired by [8]. The raw classification score for k-th many-shot category $\boldsymbol{s}_k$ is calculated as:

$$\overline{s_k} = \epsilon \cdot \overline{\boldsymbol{z}_i}^T \overline{\boldsymbol{w}_k} \tag{3}$$

where $\overline{\boldsymbol{z}_i}$ and $\overline{\boldsymbol{w}_k}$ are l_2-normalized image feature vector and classification weight for the k-th many-shot category, respectively. ϵ is a learnable scalar value that is introduced to adjust the range of cosine similarity to fit softmax function. The prototypical vector $\boldsymbol{v}$ can also take both positive and negative values by removing the ReLU layer. As a result, the prototypical vectors of few-shot categories are no longer affected by the magnitude of image features. To summarize, the final classification weights $\boldsymbol{w_i}$ for any category in the dataset can be unified and represented as:

$$\boldsymbol{w}_k = \begin{cases} \boldsymbol{w}_k, & \text{if } k \in K^m \\ \boldsymbol{v}_k, & \text{if } k \in K^f \end{cases} \tag{4}$$

In addition, it is necessary to well train a feature extractor since the classification performance is based on the derived classification weights and the obtained image features. We propose to adopt classification training with softmax loss and a new semantic loss on K^m many-shot categories:

$$L_{train} = \alpha L_{cls} + \beta L_{semantic} \tag{5}$$

The probability of k-th many shot category given image-label pair (x^m, y^m) can be obtained from:

$$p(k|x^m) = \frac{exp(s_k)}{\sum_{i=1}^{K^m} exp(s_k)} \tag{6}$$

where s_k is the raw classification score for k-th many-shot category. The softmax cross-entropy loss is given by:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} O(k) \log(p(k|x^m)) \tag{7}$$

where $O(k)$ is true if $k = y^m$ and N is the total number of many-shot training images. To fully utilize the semantic information during the training, we design a new semantic loss which measures the similarity between the image feature and its respective class label semantic embedding. The motivation is that the distance of image features that belong to different categories should be close to the distance of their corresponding class label semantic embeddings. No semantic loss is calculated for images that belong to the same class. For two image-label pairs (x_i, y_i) and (x_j, y_j), we can obtain their image feature $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ after passing the feature extractor. Their class label embeddings can be calculated using Eq. 1 as $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$.

$$L_{semantic} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (dist(\boldsymbol{z}_i, \boldsymbol{z}_j) - dist(\boldsymbol{t}_i, \boldsymbol{t}_j))^2 \tag{8}$$

After training the network using L_{train}, the classification weights for both many-shot and few-shot categories can be calculated using Eq. 4.

3.4. Learning of inter-class relations using GCN

Classification weights and class label embeddings contain rich information that defines inter-class relationships between food categories. However, such information cannot be captured or leveraged using conventional approaches of CNNs. The concept of Graph Neural Network (GNN) was first introduced in [34], which is based on CNNs and graph embeddings. GNNs are good at aggregating information from locally connected structure. Each node in GNNs is defined by its own features and features from its related nodes. In the survey paper [47], GNNs are classified into different types, such as by propagation step. Graph Convolutional Networks (GCNs) are one of the most popular variants of GNNs. GCNs are designed with convolution and readout function that can be trained to perform node classification or graph classification with task-specific loss.

Let $G = (V, E)$ represent a standard GCN with multiple layers (L), where the nodes $V = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_K\}$ is the set of classification weights that each node corresponds to the food class in the dataset (Eq. 4), and E is the edge connectivity between neighboring classes. In this work, we design the edges based on cosine similarity between each node in terms of $\boldsymbol{w}_i$ and $\boldsymbol{t}_i$ and the edge connection of each node is based on its pre-defined first-order neighboring nodes $j \in \mathbb{N}(i)$. For example, E_{ij} is the edge between class/node i and class/node j with edge strength $q_{ij} \in [0, 1]$ after applying softmax operation over the cosine similarity of both $\boldsymbol{w}$ and $\boldsymbol{t}$. Please note the edges are independent of direction.

$$q_{ij} = Softmax(\theta(\boldsymbol{w}_i, \boldsymbol{w}_j) + \theta(\boldsymbol{t}_i, \boldsymbol{t}_j)) \tag{9}$$

where θ denotes the cosine similarity function.

Neighborhood aggregation using Jumping Knowledge. Denote $\boldsymbol{h}_i^{(l)}$ as the representation for node i at layer l. It is updated as:

$$\boldsymbol{h}_{\mathbb{N}(i)}^{(l+1)} = \rho(\Theta_l \cdot AGGREGATE(\{\boldsymbol{h}_j^{(l)}, \forall j \in \mathbb{N}(i)\})) \tag{10}$$

$$\boldsymbol{h}_i^{(l+1)} = COMBINE(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_{\mathbb{N}(i)}^{(l+1)}) \tag{11}$$

where ρ is a point-wise non-linearity which in our case is LeakyReLU, and Θ is the trainable filter parameters at layer l. AGGREGATE is defined as follows:

$$\boldsymbol{h}_{\mathbb{N}(i)}^{(l+1)} = \sum_{j \in \mathbb{N}(i)} q_{ij} F(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}) \tag{12}$$

where F is a fully-connected network that is used to perform non-linear combination of node features from different nodes. COMBINE is based on node feature concatenation. For final representation for node i, we adopt the concatenation solution in Jumping Knowledge Networks [43] to connect all layers' representations to the final output. In this way, the model can learn to selectively exploit information from all intermediate layers:

$$\boldsymbol{h}_i^{final} = \varphi([\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_i^{(2)}, ..., \boldsymbol{h}_i^{(L)}]) \tag{13}$$

where φ is a linear transformation to obtain the final GCN output $\boldsymbol{c}_i$.

4. Experiments and Results

4.1. Datasets

To evaluate the performance of the proposed fusion learning framework, we conduct comprehensive experiments on two popular datasets: Food-101 [4] and UECFood-256 [15]. In order to train the GCN, we apply similar technique of training episodes used in [9].

Food-101. The dataset consists of 101,000 images with 101 categories and each category has 1,000 images, most of which are Western cuisines. We partition each category into 70% for training 10% for validation and 20% for testing. For many-shot and few-shot partition, the 101 classes are divided into 60 many-shot categories and 41 few-shot categories.

UDCFood-256. The dataset consists of 31, 397 images with 256 categories. These categories are collected from different cuisines, such as Japanese, Chinese, Western, etc. Although small in size, the large inter-class variation can be used to test the robustness of the framework. We partition each category into 70% for training 10% for validation and 20% for testing. For many-shot and few-shot partition, the 256 classes are divided into 156 many-shot categories and 100 few-shot categories. Images from both datasets are resized to 256×256.

4.2. Implementation details

Semantic embedding To obtain context sensitive class label embedding, we use a pretrained BERT model to generate the embeddings. The BERT model was trained using uncased English vocabulary with WordPiece masking, which has 12 hidden layers and outputs $d = 512$ dimensional features.

Architectures of feature extractor Feature extractor is trained using many-shot categories, which later is used to generate the prototypical vectors for few-shot categories based on aggregation of image features. For the following experiments, we use an EfficientNet-b0 [39] as the feature extractor. The output feature dimension is $d = 512$. We have also tried ResNet-10 [9] as the feature extractor for comparison.

Architecture of GCN We choose a standard GCN with skip connections as the base structure for our GCN architecture. We adopt the concatenation strategy that used in JK-Net [43] as the aggregation function on the final layer. The final output class representation $\boldsymbol{c}$ is used to train a new classifier using classification loss and matching loss based on episodic training.

For network training, we use SGD optimizer with momentum of 0.9 and weight decay of $1e-5$. We evaluate our model by performing 5-way-1-shot and 5-way-5-shot experiments on both datasets, 5-way-1-shot means we sample 5 random novel classes from the dataset, from each class we sample 1 random sample. For fusion evaluation, equal number of random classes and random images are selected for both many-shot and few-shot categories.

4.3. Comparisons with related work

Due to the reason that most food recognition work cannot handle few-shot recognition and there is no existing food dataset based fusion results for comparison, we have selected the few-shot performance from popular few-shot frameworks as the comparison metrics. The results of these few-shot frameworks are based on our implementations.

In Table 1 and Table 2, we compare the 5-way few-shot performance of our proposed method against related prior work on both Food-101 and UECFood-256. "Conv-4-64" stands for a simple network with 4 convolutional layers followed by a fully-connected layer resulting in a 64 output feature size. For Few-shot with DAE [9], we also change the feature extractor to EfficientNet-b0 for better comparison. We can observe our proposed fusion learning framework achieves superior performance on both datasets. To be more specific, our proposed method gives an improvement of 8.27% regarding 5-way-1-shot performance against Few-shot with DAE on Food-101, which is one of the state-of-the-art prior work. Our proposed method also achieves consistent improvement on UECFood-256 of 9.63% and 6.96% in terms of 1-shot and 5-shot performance. The results show that the few-shot recognition performance can be significantly enhanced by incorporating necessary inter-class correlations based on both image and semantic/text representations.

4.4. Ablation study of the proposed framework

In this section, we provide extra experiments with 4 different settings to validate the contributions of different components of our proposed framework. The 4 different settings are shown in Table 3.

Here, we examine different experimental configurations based on two major aspects: Implementation of "Fusion learning with Semantic Loss" and "GCN training", which are basically the first and second half of fusion learning in

Table 1. Few-shot performance comparisons of average accuracy on Food-101 (%)

		5-Way			
		1-shot		5-shot	
Model	Feature Extractor	Top-1	Top-5	Top-1	Top-5
MAML [7]	Conv-4-64	47.31	-	64.19	-
Prototypical Networks [36]	Conv-4-64	48.03	-	64.40	-
RelationNet [37]	Conv-4-64	50.11	-	65.78	-
Dynamic few-shot [8]	ResNet-10	53.21	86.45	68.82	90.79
Few-shot with DAE [9]	ResNet-10	53.70	86.67	70.29	91.13
Few-shot with DAE [9]	EfficientNet-b0	58.60	88.99	75.06	92.14
Ours	EfficientNet-b0	**61.97**	**92.25**	**77.72**	**94.09**

Table 2. Few-shot performance comparisons of average accuracy on UECFood-256 (%)

		5-Way			
		1-shot		5-shot	
Model	Feature Extractor	Top-1	Top-5	Top-1	Top-5
MAML [7]	Conv-4-64	42.10	-	57.43	
Prototypical Networks [36]	Conv-4-64	42.63	-	58.25	-
RelationNet [37]	Conv-4-64	43.03	-	59.43	-
Dynamic few-shot [8]	ResNet-10	44.68	75.84	60.49	86.67
Few-shot with DAE [9]	ResNet-10	45.22	76.19	62.12	87.80
Few-shot with DAE [9]	EfficientNet-b0	48.82	80.36	65.00	90.25
Ours	EfficientNet-b0	**54.85**	**86.08**	**69.08**	**92.36**

Table 3. Different settings in ablation study

Config	Fusion learning with $L_{semantic}$	GCN
Baseline	✗	✗
A	✓	✗
B	✗	✓
Ours (A+B)	✓	✓

Table 5. Top-1 few-shot/fusion accuracy on UECFood-256 (%)

		Baseline	A	B	Ours
1-shot	Few-shot	48.82	53.83	50.19	**54.85**
	Fusion	53.25	55.63	54.31	**56.73**
5-shot	Few-shot	65.00	68.49	66.22	**69.08**
	Fusion	57.63	59.94	58.02	**60.13**

our framework. For simplification, we denote them as A and B. Baseline refers to the case of using the classification weights and prototypical vectors to perform the classification, without implementations of semantics and GCN. Config A means the class label embedding is used during the fusion training but without the second stage of training GCN, Config B means GCN is used to capture inter-class relations between food categories based on both classification weights and class label embeddings but without involving semantic information during the first stage of fusion training. A+B is our complete proposed method.

Table 4. Top-1 few-shot/fusion accuracy on Food-101 (%)

		Baseline	A	B	Ours
1-shot	Few-shot	58.60	61.19	59.04	**61.97**
	Fusion	66.93	68.08	66.75	**68.76**
5-shot	Few-shot	75.06	76.94	75.95	**77.72**
	Fusion	78.66	79.96	79.38	**80.11**

From Table 4 and Table 5 show the performance comparison of few-shot and fusion results testing on Food-101 and UECFood-256 for 4 different experimental settings. Few-shot result is the average accuracy of 5-way-1-shot or 5-way-5-shot training, and fusion accuracy is obtained by randomly sampling equal number of testing images from both few-shot and many-shot categories.

Impact of semantic loss. Use of semantic information in this work consists of two parts, which are the semantic loss introduced during fusion learning of many-shot and few-shot that based on BERT class label embeddings, and the same class label embeddings are used during inter-class relations learning using GCN. Refer to Table 4 and Table 5, we can observe that by incorporating the class label embedding along during the fusion learning of many-shot and few-shot (Config A), an improvement of 2.59% and 5.01% is achieved for Top-1 1-shot recognition on Food-101 and UECFood-256 as compared to baseline, respectively. Since the introduction of the semantic loss enhances the feature extractor, hence improves the recognition accuracy of many-shot categories, the Top-1 fusion results are also improved by 1.15% and 2.38% on both datasets.

Impact of GCN. We use GCN to find inter-class relations based on both image representation (classification weights) as well as class label embedding. We believe both image and text based representations can be used to improve the discriminative ability of the final classifier. From Table 4 and Table 5, we can see consistent improvement of performance by introducing GCN (Config B) for few-shot accuracy on both datasets. For example, by using only GCN, it achieves 0.44% and 1.37% of improvements for 1-shot few-shot accuracy on Food-101, UECFood-256. The improvement is more significant on UECFood-256, which has more classes with less images per class, indicating the effectiveness of training GCN with both the classification weights and context sensitive class label embeddings.

Finally, by combining both A and B, which is our complete proposed method, the performance on both datasets is improved significantly, which surpasses both standalone A or B setting. Our proposed method achieve 3.37% and 6.03% of improvement for 1-shot few-shot accuracy on Food-101 and UECFood-256, respectively. Consistent improvement can be also observed for the fusion performance.

Table 6. Top-1 few-shot/fusion accuracy on Food-101 using ResNet-10 as feature extractor (%)

		Baseline	A	B	Ours
1-shot	Few-shot	53.70	56.44	54.12	**56.89**
	Fusion	62.66	63.93	62.91	**64.06**
5-shot	Few-shot	70.29	72.60	71.06	**73.51**
	Fusion	72.33	73.49	72.62	**73.82**

Table 7. Top-1 few-shot/fusion accuracy on UECFood-256 using ResNet-10 as feature extractor (%)

		Baseline	A	B	Ours
1-shot	Few-shot	45.22	49.58	46.69	**51.02**
	Fusion	52.57	53.78	52.84	**54.29**
5-shot	Few-shot	62.12	64.66	62.83	**65.15**
	Fusion	56.02	57.29	56.30	**57.46**

4.5. Ablation study of the architecture of feature extractor

To test the general robustness of our proposed method, we also include the experimental results of our proposed method with same settings as in Table 4 and Table 5 with the exception of using ResNet-10 as the feature extractor. From Table 6 and Table 7 show the performance comparison of few-shot and fusion results on Food-101 and UECFood-256.

Experimental results in the tables show consistent improvement of recognition performance on both food datasets using ResNet-10. Our proposed method using ResNet-10 achieves 3.19% and 5.8% of improvement for 5-way-1-shot few-shot accuracy on Food-101 and UECFood-256, respectively. As compared to the baseline implementation, an improvement of 0.42% and 1.47% is obtained for 1-shot accuracy on Food-101 and UECFood-256 using Config B, which is the implementation of GCN only. An improvement of 2.74% and 4.36% is obtained for the 1-shot accuracy on both datasets using Config A, which implements the semantic loss during the first stage of fusion learning. Further, the accuracy of fusion results are also enhanced consistently, which shows the robust improvements contributed by each component in our proposed fusion framework using different feature extractors.

5. Conclusion

In this work, we proposed a two-stage fusion learning framework for visual food recognition which unifies both many-shot and few-shot learning, so it can classify images from both many-shot categories and few-shot categories together. Our framework is based on extracted classification weights as well as context sensitive embeddings of class label.

We have also introduced a second stage training using GCN, which captures the inter-class relations between different food categories based on image and text representations. The final performance of the framework is evaluated on two popular food datasets extensively. The results show a consistent improvement for both few-shot and fusion accuracy as compared with prior work. The experiments demonstrate the effectiveness of our proposed method of fusion learning with semantics and GCN on food recognition.

Acknowledgement

The research work was done at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University. This research is supported in part by the NTU-PKU Joint Research Institute, a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation.

References

[1] Eduardo Aguilar, Marc Bolaños, and Petia Radeva. Food recognition using fusion of classifiers based on cnns. In *International Conference on Image Analysis and Processing*, pages 213–224. Springer, 2017.

[2] Marios M Anthimopoulos, Lauro Gianola, Luca Scarnato, Peter Diem, and Stavroula G Mougiakakou. A food recognition system for diabetic patients based on an optimized bag-of-features model. *IEEE journal of biomedical and health informatics*, 18(4):1261–1271, 2014.

[3] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 844–851. IEEE, 2015.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] David Elsweiler, Christoph Trattner, and Morgan Harvey. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 575–584, 2017.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[8] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[9] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2019.

[10] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 41–49, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Ye He, Chang Xu, Nitin Khanna, Carol J Boushey, and Edward J Delp. Analysis of food images: Features and classification. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2744–2748. IEEE, 2014.

[13] Shuqiang Jiang, Weiqing Min, Linhu Liu, and Zhengdong Luo. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29:265–276, 2019.

[14] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. Graph-rise: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*, 2019.

[15] Yoshiyuki Kawano and Keiji Yanai. Foodcam-256: a large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 761–762, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7212–7220, 2019.

[18] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9258–9267, 2019.

[19] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 567–576. IEEE, 2018.

[20] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

[21] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1233–1241, 2015.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[24] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019.

[25] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2016.

[26] Pradyumna Narayana, Aniket Pednekar, A. Krishnamoorthy, Kazoo Sone, and Sugato Basu. Huse: Hierarchical universal semantic embeddings. *ArXiv*, abs/1911.05978, 2019.

[27] Duc Thanh Nguyen, Zhimin Zong, Philip O Ogunbona, Yasmine Probst, and Wanqing Li. Food image classification using local appearance and global structural information. *Neurocomputing*, 140:242–251, 2014.

[28] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 1–12, 2011.

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[30] Parisa Pouladzadeh and Shervin Shirmohammadi. Mobile multi-food recognition using deep learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s):1–21, 2017.

[31] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*, 2020.

[32] Alec Radford. Improving language understanding by generative pre-training. 2018.

[33] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

[34] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

[35] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Baby steps towards few-shot learning with multiple semantics. *arXiv preprint arXiv:1906.01905*, 2019.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[39] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[40] Ryosuke Tanno, Koichi Okamoto, and Keiji Yanai. Deepfoodcam: A dcnn-based real-time mobile food recognition system. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 89–89, 2016.

[41] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019.

[42] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2019.

[43] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[44] Ruihan Xu, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain. Geolocalized modeling for dish recognition. *IEEE transactions on multimedia*, 17(8):1187–1199, 2015.

[45] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2770–2779, 2019.

[46] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1124–1133, 2016.

[47] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.