Online Continual Learning For Visual Food Classification

Jiangpeng He
he416@purdue.edu

Fengqing Zhu
zhu0@purdue.edu

School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA

Abstract

Food image classification is challenging for real-world applications since existing methods require static datasets for training and are not capable of learning from sequentially available new food images. Online continual learning aims to learn new classes from data stream by using each new data only once without forgetting the previously learned knowledge. However, none of the existing works target food image analysis, which is more difficult to learn incrementally due to its high intra-class variation with the unbalanced and unpredictable characteristics of future food class distribution. In this paper, we address these issues by introducing (1) a novel clustering based exemplar selection algorithm to store the most representative data belonging to each learned food for knowledge replay, and (2) an effective online learning regime using balanced training batch along with the knowledge distillation on augmented exemplars to maintain the model performance on all learned classes. Our method is evaluated on a challenging large scale food image database, Food-1K[1], by varying the number of newly added food classes. Our results show significant improvements compared with existing state-of-the-art online continual learning methods, showing great potential to achieve lifelong learning for food image classification in real world.

1. Introduction

Food classification serves as the first and most crucial step for image-based dietary assessment [3], which aims to provide valuable insights for prevention of many chronic diseases. As shown in Figure 1, ideally food classification system should be able to update using each new recorded food image continually without forgetting the food class that has been already learned before. Achieving this goal would bring significant advantage for deploying such a system for automated dietary assessment and monitoring.

From the perspective of visual food classification, although recent works [41, 27, 32, 31] have been proposed using advanced deep learning based approaches to increase model performance, they use only static datasets for training and are not capable of handling sequentially available new food classes. Therefore, the classification accuracy could drop dramatically due to the unavailability of old data, which is also known as catastrophic forgetting [29]. Although retraining from scratch is a viable option, it is impractical to do whenever a new food is observed, which is time consuming and require high computation and memory resource especially for large scale food image datasets. For example, a model already learned $1,000$ food classes need to retrain from scratch for only 1 new observed food.

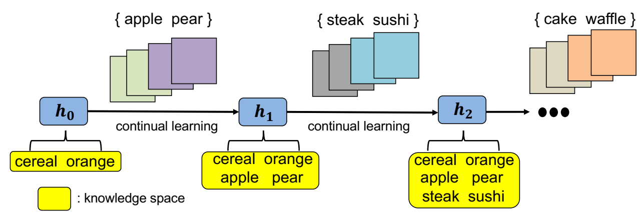

Figure 1: **Continual learning for food image classification.** The model h learns new food class sequentially overtime without accessing to already learned class data for each continual learning step. The updated model can classify all food classes seen so far.

From the perspective of continual learning, an increasing number of approaches [9, 5, 7, 34] have been proposed to address catastrophic forgetting and to learn new knowledge incrementally in online scenario. Compared to offline scenario where data can be used multiple epochs for training, online scenario is more challenging where each new data is observed only once by the model, but is more practical for real-life application such as food image classification system. Representative techniques to mitigate forgetting include (1) storing a small number of learned data as exemplars for replay [35], and (2) applying knowledge distillation [13] using a teacher model to maintain the learned performance. However, continual learning for food image classification is still lacking and there are two major ob-

[1] https://www.kaggle.com/c/largefinefoodai-iccv-recognition/data

stacles which make the above mentioned techniques less effective for food images. (i) Food images exhibit higher intra-class variation [27] compared with commonly seen objects in real life, which is due to different culinary culture and cooking style. Most existing continual learning methods [35, 4, 42, 43, 9, 15] apply herding algorithm [40] to select exemplars for each learned class based on class mean only, which is difficult to cover the diversity for food types within the same class. Therefore, catastrophic forgetting could become worse if stored exemplars are not good representations of learned classes. (ii) The distribution of future food classes is usually unpredictable and imbalanced due to the variance of consumption frequencies [23] among different food categories. Nevertheless, most online approaches only study continual learning on balanced datasets containing the same number of data per class such as CIFAR [18] and MNIST [19] without considering the class-imbalance problem that is common for food images. In addition, as indicated in [2], the knowledge distillation term becomes less effective if teacher model is not trained on balanced data.

In this work, we address the challenging problem of food image classification for online continual learning by first introducing a novel exemplar selection algorithm, which clusters data for each class based on visual similarity and then selects the most representative exemplars from each generated cluster based on cluster mean. We apply Power Iteration Clustering [22], which does not require the number of cluster beforehand. Therefore, our algorithm can adapt to different food categories, *i.e.*, food with higher variation will generate more clusters and vice versa. In addition, we propose an effective online learning regime by using balanced training batch for old and new class data and apply knowledge distillation loss between original and augmented exemplars to better maintain the model performance. Our method is evaluated on a large scale real world food database, Food-1K [32], and outperforms state-of-the-arts including ICARL [35], ER [5, 7], ILIO [9] and GDUMB [34], which are all implemented in online scenario and use exemplars for replay during continual learning.

The main contributions are summarized as follows.

- To the best of our knowledge, we are the first to study online continual learning for food image classification. We propose a novel clustering based exemplar selection algorithm and a new online training regime to address catastrophic forgetting.

- We conduct extensive experiments on a challenging class-imbalanced food image database to show the effectiveness for each component of our proposed method. We show that our method significantly outperforms existing approaches, especially for larger incremental step size.

2. Related Work

2.1. Food Classification

Food classification refers to the task of labeling image with food category, which assumes each input image contains only one single food item. Earlier work [14] use fusion of features including SIFT [25], Gabor, and color histograms for classification. Later, the modern deep learning models have been widely applied as backbone network to extract more class-discriminative features as in [17, 37, 28, 32, 10, 8, 31, 36], which significantly improves the performance. Recent works [41, 27] leveraging hierarchy structure based on visual information are able to achieve further improvements. However, all these methods use static food image datasets for training and none of them is capable of learning from sequentially available data, making it difficult to apply in real life applications as new foods are observed over time.

2.2. Continual Learning

The major challenge for continual learning is called catastrophic forgetting [29], where the model quickly forgets already learned knowledge due to the unavailability of old data. Below, we review and summarize existing knowledge-preserving techniques that are most relevant to our proposed method.

Replay-based methods store a small number of representative data from each learned class as exemplars to perform knowledge rehearsal during the continual learning. Herding dynamic algorithm [40] is first applied in ICARL [35] to select exemplars that are closer to the class mean. It has gradually became a common exemplar selection strategy that is being used in most existing methods [35, 4, 42, 43, 9, 15], where ICARL adopts a nearest class mean classifier [30] while others use softmax classifier for classification. In addition, reservoir sampling [39] along with random retrieval is applied in Experience Replay (ER) based methods [5, 7], which ensures each incoming data point has the same probability to be selected as exemplar in the memory buffer. A greedy balancing sampler with random selection is recently used in GDUMB [34] to store as much data as memory allowed, which also achieves competitive performance.

Regularization-based methods restrict the impact of learning new tasks on the parameters that are important for learned tasks. Knowledge distillation [13] is a popular representative technique, which makes the model mimic the output distribution for learned classes from a teacher model to mitigate forgetting during continual learning [21, 35, 4, 42, 15, 20, 11]. For most recent work, He *et al.* proposed ILIO [9], which applies an accommodation ratio to generate a stronger constraint for knowledge distillation loss to achieve improved performance.

However, among these methods, only a few [35, 9, 34,

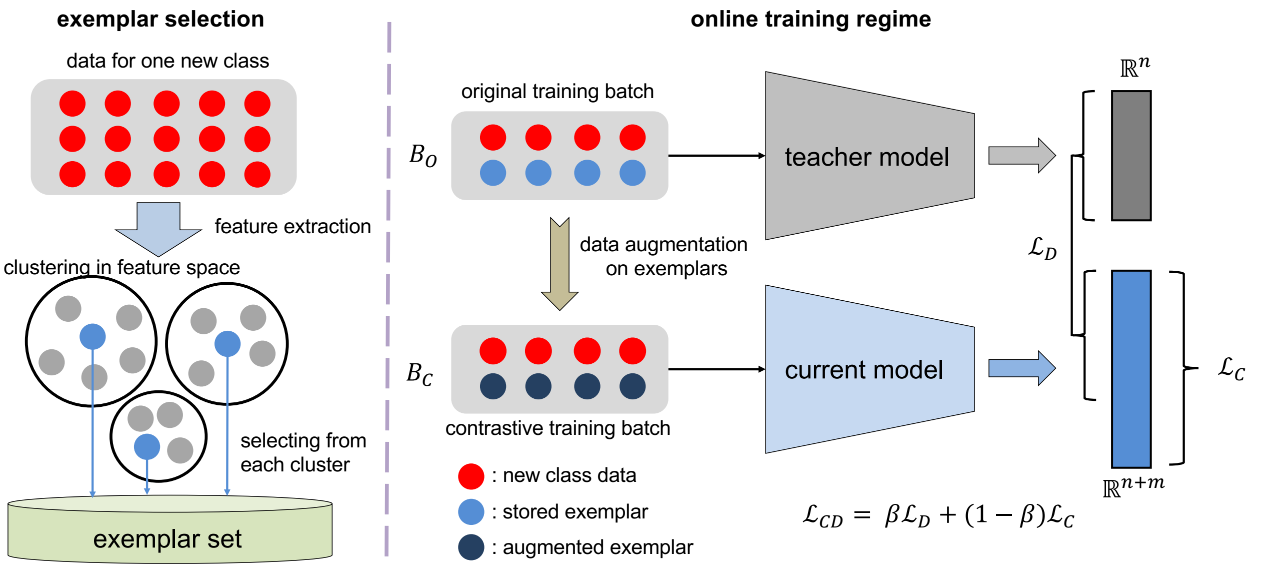

Figure 2: **Overview of proposed method.** The left side shows our exemplar selection algorithm, which selects the most representative data from center of each cluster generated based on visual similarity in feature space. Right part shows our online learning regime where each new class data is paired with one randomly selected exemplar to produce the original balanced training batch B_O. We perform data augmentation on selected exemplars to generate contrastive training batch B_C and the distillation loss $\mathcal{L}_D$ is applied between the output of the teacher model using B_O and the output of the current model using B_C. n and m denote the number of already learned classes and new added classes, respectively. β is a hyper-parameter to combine $\mathcal{L}_D$ with cross-entropy loss $\mathcal{L}_C$. (Best viewed in color)

5, 7] are feasible in online scenario to use each data only once for training. In addition, none of the existing methods focus on food images and as introduced in Section 1, the high intra-class variance and imbalanced data distribution make both exemplar and distillation based techniques less effective to address catastrophic forgetting. Therefore, we propose a novel exemplar selection algorithm to select exemplars from each generated cluster based on visual similarity to adapt to the variability of different food categories. Besides, we propose an effective online learning regime using balanced training batch and apply distillation on augmented exemplars to better maintain performance on learned classes, which is described in details in Section 4.

3. Problem Statement For Online Continual Learning

Continual learning has been studied under different scenarios. In general, it can be divided into (1) task-incremental (2) class-incremental and (3) domain-incremental as discussed in [16]. Methods for task-incremental problem use a multi-head classifier [1] for each independent task while task index is not available in class-incremental problem, which applies a single-head classifier [26] on all learned classes. Domain-incremental aims to learn the label shift instead of new classes. In addition, depending on whether each data is allowed to use more than once to update model, it can be categorized into (1) online learning that use each data once and (2) offline learning with no epoch restriction. In this work, we focus on online continual learning under class-incremental setting, which is more related to real life applications. The objective is to learn new class from data stream using each data once and to classify all classes seen so far during inference.

Specifically, the online class-incremental learning problem $\mathcal{T}$ can be formulated as learning a sequence of N tasks $\{\mathcal{T}^1, ..., \mathcal{T}^N\}$ corresponds to N incremental learning steps with model updating from h^0 to h^N, where the initial model h^0 is assumed to be trained on $\mathcal{T}^0$ before continual learning begins and h^N should be able to perform classification on test data belonging to $\{\mathcal{T}^0, \mathcal{T}^1 ..., \mathcal{T}^N\}$. Each task $\mathcal{T}^i \in \mathcal{T}$ for $i = \{0, 1, ...N\}$ contains fixed M non-overlapped new classes, which is defined as incremental step size. Let $\{D^0, D^1, ..., D^N\}$ denotes training data corresponds to N incremental steps plus the initial step D^0, where $D^i = \{(\mathbf{x}_1^i, y_1^i)...(\mathbf{x}_{n_i}^i, y_{n_i}^i)\}$, $\mathbf{x}$ and y represent the data and the label respectively, and n_i refers to the number of total training data in D^i. In online scenario, the new class data for each incremental learning step becomes available sequentially and one does not need to wait until all data has arrived to update the model as in offline case. The online learner observes each data $(\mathbf{x}^i, y^i) \in D^i$ only once for incremental step i.

4. Our Method

An overview of our proposed method is illustrated in Figure 2, including a novel exemplar selection method and an effective online training regime. Specifically, instead of selecting exemplars based on class mean as in herding [40], we first generate clusters based on similarity and then select exemplars from each cluster using the corresponding cluster mean. During the continual learning phase, each new class data from data stream is paired with one randomly selected exemplar from exemplar set to produce balanced training batch B_o that contains the same number of original new and old class samples. Then we apply data augmentation on selected exemplars in B_o to generate a contrastive training batch B_c and the knowledge distillation term is applied between the teacher output of B_o and the current model output of B_c to maintain the already learned knowledge. Details of each component is described in the remaining section.

4.1. Exemplar Selection From Clusters

The main challenge of existing exemplar selection methods is that they cannot adapt to the intra-class variation especially for food images due to its high variability. For example, the images in apple category may contain many types such as green apple, red apple, sliced apple, diced apple, whole apple and etc. Therefore, selecting from class mean as in Herding [40] will not work well when there exists more than one main types within that food class. Our proposed method addresses this problem by first clustering the data for each class based on visual similarity and then select exemplars from each generated cluster. We consider Power Iteration Clustering (PIC) [22] as our clustering approach, which is a graph based method and shown to be effective even in large scale database [6]. But other clustering methods are also feasible such as K-means [24]. One advantage of PIC is the number of generated clusters are not set beforehand, so there is more clusters if one class contains more main types and vice versa.

Given n_c images $\{(\mathbf{x}_1, y), ...(\mathbf{x}_{n_c}, y)\}$ for one new class c, we first generate nearest neighbor graph by connecting to their 10 neighbor data points in the Euclidean space using extracted feature embeddings. Let $f(\mathbf{x}_i)$ denotes the extracted feature for the i-th image, we apply the sparse graph matrix $G = \mathbb{R}^{n_c \times n_c}$ with zeros on the diagonal and the remaining elements of G are defined by

$$e_{i,j} = exp^{-\frac{||f(\mathbf{x}_i) - f(\mathbf{x}_j)||^2}{\sigma^2}}$$

where σ denotes the bandwidth parameter and we empirically use $\sigma = 0.5$ in this work. Then, we initialize a starting vector $s^{n_c \times 1} = [\frac{1}{n_c}, ..., \frac{1}{n_c}]^T$ and iteratively update it using Equation 1

$$s = L_1(\alpha(G + G^t)s + (1 - \alpha)s) \tag{1}$$

where $\alpha = 0.001$ refers to a regularization parameter and $L_1(\bullet)$ denotes the L-1 normalization step. The generated clusters are given by the connected components of a directed unweighted subgraph of G denoted as $\tilde{G}$. We set $\tilde{G}_{i,j} = 1$ if $j = argmax_j e_{i,j}(s_j - s_i)$ where s_i refers to the i-th element of the vector. Note that there is no edge starts from i if $\{\forall j \neq i, s_j \leq s_i\}$, *i.e.* s_i is a local maximum.

The general process to select exemplars for a new class c after incremental step k is illustrated in Algorithm 1, where we select the same number of exemplars q_e from each cluster generated using PIC based on cluster mean. Note that for the special situation when a cluster C_i contains very few data with $|C_i| < q_e$, we store all data from that small cluster at first and then evenly select from the remaining clusters.

Algorithm 1 Selecting exemplars for a new class c

Input: New class data: $\{(\mathbf{x}_1, y), ...(\mathbf{x}_{n_c}, y)\} \in \mathcal{T}^k$
Require: Clustering algorithm : Θ
Require: Number of exemplars per class : q
Output: Exemplar set for new class : E_c

$E_c \leftarrow \emptyset$ ▷ initialization of exemplar set for new class c
$f \leftarrow h^k$ ▷ use current model as feature extractor
$C_1, ...C_n \leftarrow \Theta(f(\mathbf{x}_1), ...f(\mathbf{x}_{n_c}))$ ▷ generated clusters
$q_e \leftarrow floor(\frac{q}{n})$ ▷ number of exemplar for each cluster
for i = 1, 2, ... n **do**
  $\mu_i = \frac{1}{|C_i|}\sum_{\mathbf{x} \in C_i} f(\mathbf{x})$ ▷ cluster mean
  for j = 1, 2, ... q_e **do**
    $v_j \leftarrow argmin_{\mathbf{x} \in C_i}||\mu_i - f(\mathbf{x})||_2$
    $E_c \leftarrow E_c \cup v_j$
    $C_i \leftarrow C_i - v_j$ ▷ remove stored exemplar from cluster

4.2. Online Learning Regime

Since future food class distribution is usually unpredictable and imbalanced, it becomes more challenging to maintain the learned knowledge due to potential class-imbalanced problem. However, almost all existing online continual learning methods use balanced datasets such as MNIST [19] and CIFAR [18] which contain the same number of training data for each class. In addition, the knowledge distillation term also becomes less effective when the teacher model is not trained on balanced data [2]. Therefore, we propose a more effective online learning regime, which consists of two main parts: using balanced training batch and applying knowledge distillation on augmented exemplars.

Suppose the model is already trained on n classes and the data stream $\{(\mathbf{x}_1^k, y_1^k)...\} \in D^k$ for incremental step k contains m newly added classes where $y^k \in \{n +$

$1, n+2, ..., n+m\}$. We pair each new class data $(\mathbf{x}_i^k, y_i^k)$ with a randomly selected exemplar $(\mathbf{v}_j, y_j) \in E^{k-1}$ where E^{k-1} denotes exemplar set containing stored exemplars for classes $\{1, 2, ..., n\}$ belonging to $\{\mathcal{T}^0, ..., \mathcal{T}^{k-1}\}$. Therefore, each training batch B contains exactly $\frac{b}{2}$ new class data and $\frac{b}{2}$ augmented old class exemplars given batch size $b = |B|$.

To make the distillation term more effective, instead of using the identical training batch for both current model and teacher model as done in existing approaches, we propose to apply data augmentation on selected exemplars in original training batch B_o to generate its corresponding contrastive training batch B_c where B_c and B_o are used as input to current model and teacher model, respectively.

The output logits of the current model is denoted as $p^{(n+m)}(B_c(\mathbf{x})) = (o^{(1)}, ..., o^{(n)}, o^{(n+1)}, ...o^{(n+m)})$, the teacher's output logits is $\hat{p}^{(n)}(B_o(\mathbf{x})) = (\hat{o}^{(1)}, ..., \hat{o}^{(n)})$ where $B_c(\mathbf{x})$ and $B_o(\mathbf{x})$ denote the data in augmented and original training batch. The knowledge distillation loss [13] is formulated as in Equation 2, where $\hat{p}_T^{(i)}$ and $p_T^{(i)}$ are the i-th distilled output logit as defined in Equation 3

$$\mathcal{L}_D(B_c(\mathbf{x}), B_o(\mathbf{x})) = \sum_{i=1}^{n} -\hat{p}_T^{(i)}(B_o(\mathbf{x})) log[p_T^{(i)}(B_c(\mathbf{x}))] \tag{2}$$

$$\hat{p}_T^{(i)} = \frac{\exp{(\hat{o}^{(i)}/T)}}{\sum_{j=1}^{n}\exp{(\hat{o}^{(j)}/T)}}, \; p_T^{(i)} = \frac{\exp{(o^{(i)}/T)}}{\sum_{j=1}^{n}\exp{(o^{(j)}/T)}} \tag{3}$$

$T > 1$ is the temperature scalar used to soften the distribution, which forces the network to learn more fine grained knowledge. The cross entropy loss to learn new classes can be expressed as in Equation 4

$$\mathcal{L}_C(B_c(\mathbf{x})) = \sum_{i=1}^{n+m} -\hat{y}^{(i)} log[p^{(i)}(B_c(\mathbf{x}))] \tag{4}$$

where $\hat{y}$ is the one-hot label for input data x. The overall cross-distillation loss function is formed as in Equation 5 by using a hyper-parameter β to tune the influence between two components.

$$\mathcal{L}_{CD}(B_c(\mathbf{x})) = \beta\mathcal{L}_D(B_c(\mathbf{x}), B_o(\mathbf{x})) + (1-\beta)\mathcal{L}_C(B_c(\mathbf{x})) \tag{5}$$

In this work, we set $T = 2$ and $\beta = 0.5$. We also notice that using stronger random data augmentation techniques to generative contrastive training batch can achieve better performance to maintain the knowledge for learned classes. Therefore our data augmentation pipeline includes *random flip*, *random color distortions* and *random Gaussian blur*.

5. Experimental Results

In this section, we first compare our proposed online continual learning method with existing approaches including **ICARL** [35], **ER** [5, 7], **GDUMB** [34] and **ILIO** [9], which all have already been discussed in Section 2. We also include **Fine-tune** and **Upper-bound** for comparison. **Fine-tune** use only new class data and apply cross-entropy loss for continual learning without considering the previous task performance, *i.e.*, neither exemplar set nor distillation loss is used and it can be regarded as the lower-bound. **Upper-bound** trains a model using all the data seen so far for each incremental learning step using cross-entropy loss in online scenario. Results are discussed in Section 5.3.

In the second part of this section, we conduct ablation study to show the effectiveness of each component of proposed method including exemplar selection algorithm and online training regime, which is illustrated in Section 5.4.

5.1. Datasets

In this work, we use **Food1K** to evaluate our method, which is a recently released challenging food dataset consisting of $1,000$ selected food classes from Food2K [32]. The dataset is originally divided as 60%, 10% and 30% for training, validation and testing, respectively. Note that no class label is given in test set so we use images in validation set as testing data. In addition, we also construct a subset of Food1k using 100 randomly selected food classes denoted as **Food1K-100** for experiment. Specifically, for **Food1K-100**, we randomly arrange 100 classes into the splits of 1, 2, 5, 20 as step size (number of new class added for each step) and for **Food1K** we perform large scale continual learning using 100 new classes for each incremental step.

5.2. Implementation Details

Our implementation is based on Pytorch [33]. We use ResNet-18 as our backbone network by following the setting suggested in [12] with input image size 224×224. We use stochastic gradient descent optimizer with fixed learning rate of 0.1 and weight decay of 0.0001. We store $q = 20$ exemplars per class in exemplar set as suggested in [35] and the batch size is set as 32 (with 16 new class data paired with 16 randomly selected exemplars). For all experiments, each data (except stored exemplars) is used only once to update the model in online scenario.

Evaluation protocol: after each incremental learning step, we evaluate the updated model on test data belonging to all classes seen so far and we use Top-1 accuracy for Food1K-100 and Top-5 accuracy for Food1K. Besides, we also report average accuracy (*Avg*) and last step accuracy (*Last*) for comparison where *Avg* is calculated by averaging the accuracy for all incremental steps to show the overall performance for entire continual learning process and *Last* accuracy shows the final performance on the entire dataset after the last step of continual learning. We repeat each experiment 5 times using different random seeds to arrange class and the average results are reported.

Datasets	Food1K-100										Food1K	
Step size	1		2		5		10		20		100	
Accuracy	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last
Fine-tune	0.043	0.009	0.081	0.029	0.182	0.018	0.379	0.134	0.497	0.233	0.265	0.099
Upper-bound	0.805	0.759	0789	0.752	0.807	0.743	0.827	0.749	0.813	0.744	0.788	0.805
ICARL [35]	0.619	0.539	0.694	0.615	0.581	0.502	0.729	0.603	0.769	0.660	0.573	0.474
ER [5, 7]	0.645	0.586	0.612	0.582	0.528	0.520	0.694	0.599	0.728	0.633	0.533	0.428
GDUMB [34]	0.606	0.430	0.612	0.441	0.573	0.507	0.591	0.456	0.754	0.623	0.506	0.289
ILIO [9]	**0.695**	**0.670**	0.681	**0.643**	0.501	0.452	0.703	0.633	0.708	0.596	0.515	0.428
Ours	0.692	0.661	**0.702**	0.641	**0.643**	**0.563**	**0.762**	**0.669**	**0.786**	**0.699**	**0.612**	**0.504**

Table 1: **Average accuracy and Last step accuracy** with step size 1, 2, 5, 10, 20 on Food1K-100 and step size 100 on Food-1K. Best results (except upper-bound) are marked in bold.

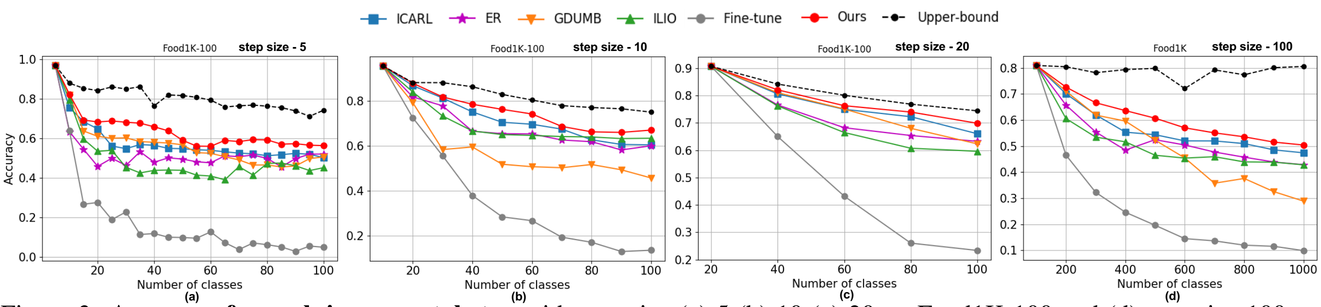

Figure 3: **Accuracy for each incremental step** with step size (a) 5 (b) 10 (c) 20 on Food1K-100 and (d) step size 100 on Food-1K. (Best viewed in color)

5.3. Comparison With Existing Methods

Table 1 summarizes the average accuracy (*Avg*) and last step accuracy (*Last*) for all incremental step sizes. Overall, we notice that the online continual learning performance vary a lot for different step sizes. Given fixed total number of classes to learn, smaller step size will produce more incremental steps so catastrophic forgetting appears more frequently. On the other hand, for larger step size, although there will be less incremental steps, learning more classes for each step is also a challenging task especially in online scenario to use each data only once for training. Specifically, we observe severe catastrophic forgetting problem by using *Fine-tune* where both *Avg* and *Last* accuracy are much lower compared with *Upper-bound* due to the lack of training data for learned tasks during the continual learning process. All existing methods achieve significant improvement compared with *Fine-tune* especially for *ILIO* [9], which works more effectively when step size is very small as their final prediction is given by the combination of outputs for both the teacher model and current model. Note that *ILIO* requires the teacher model for both training and inference phases which greatly increases the memory storage while other methods included ours only use teacher model during the training phase. However, as incremental step size increase, our method achieves best performance even for very large scale continual learning for $1,000$ classes in Food1K. We also show the accuracy evaluated after each incremental learning step with step size 5, 10, 20 and 100 in Figure 3. Our method outperforms state-of-the-art for all learning steps with smallest performance gap compared with *upper-bound*. Note that we did not provide the figures for step size 1 and 2 as they contain too many learning steps (100 and 50 respectively), which is difficult for visualization.

5.4. Ablation Study

In this part, we conduct ablation studies to analyze the effectiveness of (1) **component-1:** our proposed exemplar selection algorithm that selects representative data from clusters generated based on visual similarity and (2) **component-2:** our online training regime using balanced training data for new and old class, and contrastive training batch for knowledge distillation. Specifically, we consider the following methods for comparisons:

- **baseline**: removing both component-1 and component-2 from our method, *i.e.*, use herding [40] for exemplar selection instead and pair new class data in training batch with the random number of exemplars
- **baseline + our exp**: baseline + component-1
- **baseline + our training regime**: baseline + component-2

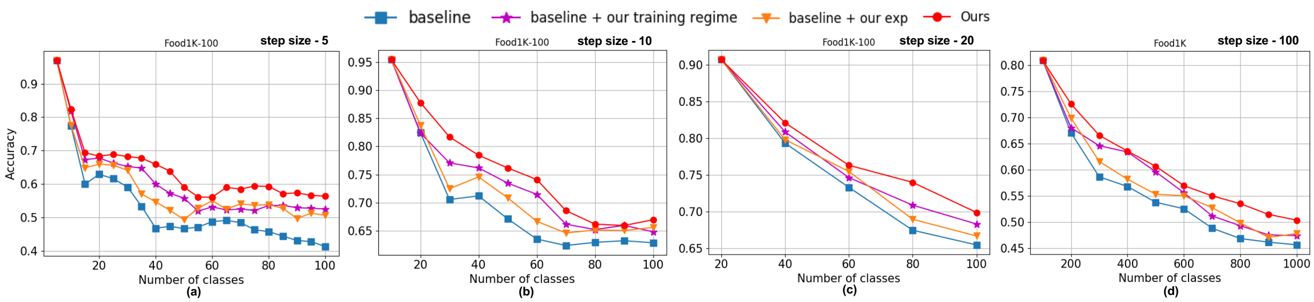

Figure 4: **Ablation study** with step size (a) 5 (b) 10 (c) 20 on Food1K-100 and (d) step size 100 on Food-1K. (Best viewed in color)

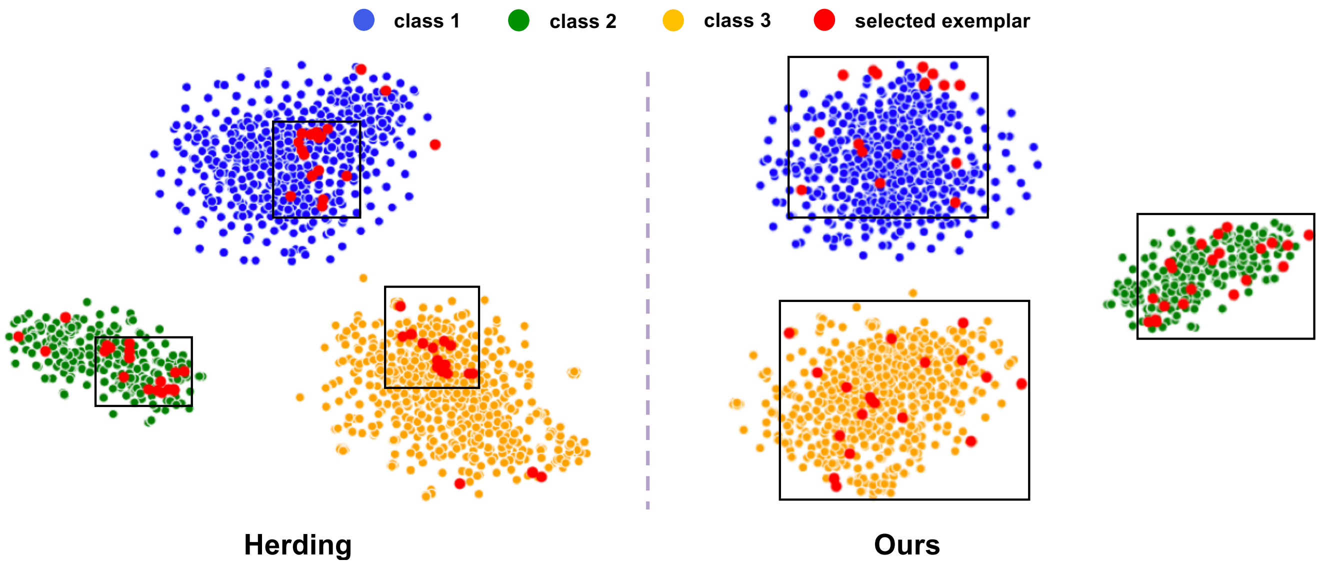

Figure 5: **A t-SNE [38] visualization** by comparing herding [40] with our proposed exemplar selection algorithm. We randomly select three classes from Food1K corresponds to three different colors and the red dots represent the selected exemplars. The black box indicates the area where most exemplars are located for each class. (Best viewed in color)

- **Ours**: baseline + component-1 + component-2

Figure 4 shows the results for each incremental step with step size 5, 10, 20 and 100. Compared with *baseline*, we observe performance improvement by incorporating each component of proposed method. The best performance is obtained when combining both components. In addition, we notice that our training regime using balanced training batch performs more effectively than our exemplar selection since severe class-imbalanced problem exists in this Food1K dataset, where the number of training data ranges from $[91, 1199]$ per food class.

5.4.1 Influence of Exemplar Size

For experiments in Section 5.3, we follow the protocol [35] to use 20 exemplars per class. In this part, we vary the number of exemplar stored for each class $q \in \{10, 50, 100\}$ and compare *baseline* + *our exp* using our proposed exemplar selection algorithm with *baseline* using Herding selection [40]. We use Food1K-100 with step size 5 and the average accuracy are shown in Table 2. In general, the performance becomes better for both methods when more exemplars are used. However, the memory storage capacity is one of the most important factors for continual learning especially in online scenario and we observe that our proposed approach is more efficient which outperforms *baseline* for a larger margin when using less exemplars.

Method	$q = 10$	$q = 50$	$q = 100$
baseline	0.486	0.629	0.697
baseline + our exp	0.527	0.651	0.706

Table 2: **Average accuracy on Food1K-100 with step size 5 by varying exemplar size**. Best results marked in bold.

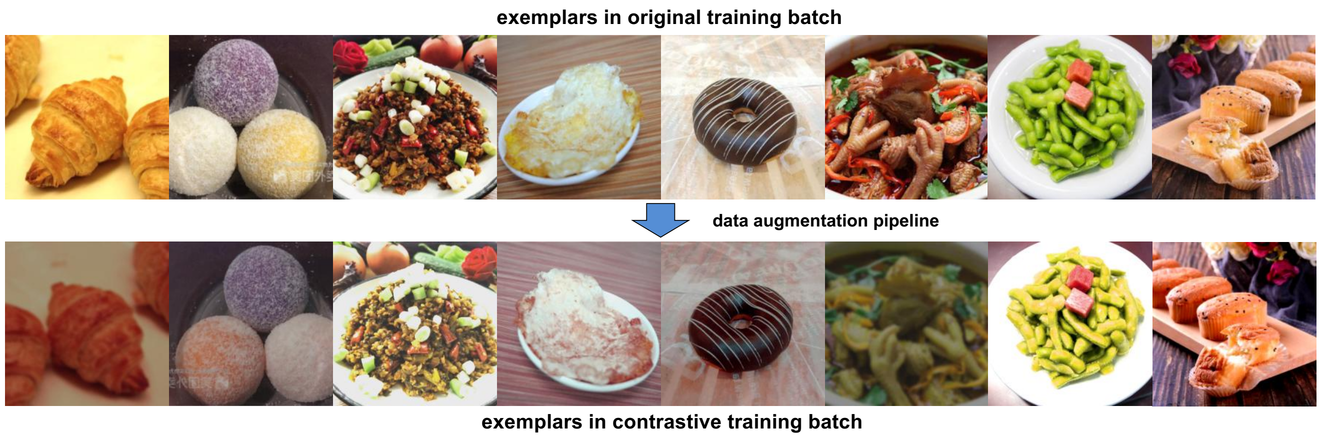

Figure 6: **Visualization of contrastive training batch** generated by our proposed data augmentation pipeline including *random flip*, *random color distortions* and *random Gaussian blur*. (Best viewed in color)

5.4.2 Visualization of Selected Exemplars

A t-SNE [38] visualization comparing herding [40] and our proposed exemplar selection method is shown in Figure 5 where we randomly select three food classes from Food1K as denoted by blue, green and orange dots, respectively and red dots refer to the selected exemplars. As shown in the left half of the figure, most exemplars selected by herding are concentrated in a small area for each class as indicated by the black box. Therefore, the model gradually forgets the knowledge outside the black box during the continual learning process, leading to catastrophic forgetting. Our method addressed this problem by performing clustering at first based on visual similarity and then select exemplars from all generated clusters to better represent the intra-class diversity for each food class as illustrated in Section 4.1. In the right half of this figure, we find that the exemplars selected by our method covers a wider region for each food class, which helps to produce higher quality classifiers to retain the learned knowledge due to better generalization ability of our selected exemplars as shown in Figure 4 by comparing **baseline** with **baseline + our exp**.

5.4.3 Visualization of Contrastive Training Batch

Figure 6 shows the exemplars for learned food classes in original and contrastive training batch using our proposed data augmentation pipeline including *random flip*, *random color distortions* and *random Gaussian blur*. By comparing results of **baseline** with **baseline + our training regime** as shown in Figure 4, we observe that using augmented data is more effective to help retain the already learned knowledge to achieve better performance. One explanation is that each exemplar stored in the exemplar set can be selected for more than once to pair with new class data during the online training phase, so the data augmentation step helps to improve the classifier's generalization ability to obtain higher accuracy on learned classes. In addition, the knowledge distillation term also becomes more efficient to maintain the performance for old classes by using balanced training batch for old and new class data and transferring the learned knowledge from teacher model using original training batch to the current model using contrastive training batch as formulated in Equation 2.

6. Conclusion

In summary, we studied online continual learning for food image classification in this work and proposed a novel exemplar selection algorithm that selected representative data from each cluster generated based on visual similarity to alleviate the high intra-class variation problem of food images. In addition, an effective online learning regime was introduced using balanced training batch for old and new class and we proposed to apply knowledge distillation using contrastive training batch to help retain the learned knowledge. Our method achieved promising results on a challenging food dataset, Food1K, with significant performance improvement compared with existing state-of-the-art especially when the number of new food classes added for each incremental step increased, showing great potential for large scale continual learning of food image classification in real life.

For future work, although achieving promising results, our method still require storing part of original learned data as exemplars for replay during continual learning, which may not be feasible in many real scenarios due to the privacy issue or memory constraint. One possible solution is to use class prototype as recently introduced in [44].

References

[1] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2020.

[2] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. *Proceedings of the IEEE International Conference on Computer Vision*, pages 583–592, 2019.

[3] CJ Boushey, M Spoden, FM Zhu, EJ Delp, and DA Kerr. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proceedings of the Nutrition Society*, 76(3):283–294, 2017.

[4] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. *Proceedings of the European Conference on Computer Vision*, September 2018.

[5] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[6] Matthijs Douze, Hervé Jégou, and Jeff Johnson. An evaluation of large-scale methods for image instance and class discovery. *Proceedings of the on Thematic Workshops of ACM Multimedia*, pages 1–9, 2017.

[7] Tyler L Hayes, Nathan D Cahill, and Christopher Kanan. Memory efficient experience replay for streaming learning. *Proceedings of the International Conference on Robotics and Automation*, pages 9769–9776, 2019.

[8] Jiangpeng He, Runyu Mao, Zeman Shao, Janine L. Wright, Deborah A. Kerr, Carol J. Boushey, and Fengqing Zhu. An end-to-end food image analysis system. *Electronic Imaging*, 2021(8):285–1–285–7, 2021.

[9] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13926–13935, 2020.

[10] Jiangpeng He, Zeman Shao, Janine Wright, Deborah Kerr, Carol Boushey, and Fengqing Zhu. Multi-task image-based dietary assessment for food recognition and portion size estimation. *arXiv preprint arXiv:2004.13188*, 2020.

[11] Jiangpeng He and Fengqing Zhu. Unsupervised continual learning via pseudo labels. *arXiv preprint arXiv:2104.07164*, 2021.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*, 2015.

[14] H. Hoashi, T. Joutou, and K. Yanai. Image recognition of 85 food categories by feature fusion. *Proceedings of 2010 IEEE International Symposium on Multimedia*, pages 296–301, Dec 2010.

[15] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[16] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.

[17] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. Food detection and recognition using convolutional neural network. *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1085–1088, 2014. Orlando, Florida, USA.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[20] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. *Proceedings of the IEEE International Conference on Computer Vision*, pages 312–321, 2019.

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[22] Frank Lin and William W Cohen. Power iteration clustering. *Proceedings of International Conference on Machine Learning*, 2010.

[23] Luotao Lin, Fengqing Zhu, Edward Delp, and Heather Eicher-Miller. The most frequently consumed and the largest energy contributing foods of us insulin takers using nhanes 2009–2016. *Current Developments in Nutrition*, 5:426–426, 2021.

[24] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[25] David G Lowe. Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision*, 2:1150–1157, 1999.

[26] Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019.

[27] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. Visual aware hierarchy based food recognition. *arXiv preprint arXiv:2012.03368*, 2020.

[28] N. Martinel, G. L. Foresti, and C. Micheloni. Wide-slice residual networks for food recognition. *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 567–576, March 2018.

[29] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989.

[30] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.

[31] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. *Proceedings of the 28th ACM International Conference on Multimedia*, pages 393–401, 2020.

[32] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *CoRR*, abs/2103.16107, 2021.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. *Proceedings of the Advances Neural Information Processing Systems Workshop*, 2017.

[34] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. *Proceedings of the European Conference on Computer Vision*, pages 524–540, 2020.

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[36] Zeman Shao, Shaobo Fang, Runyu Mao, Jiangpeng He, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu. Towards learning food portion from monocular images with cross-domain feature adaptation. *arXiv preprint arXiv:2103.07562*, 2021.

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

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[39] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.

[40] Max Welling. Herding dynamical weights to learn. *Proceedings of the International Conference on Machine Learning*, pages 1121–1128, 2009.

[41] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R Smith. Learning to make better mistakes: Semantics-aware visual food recognition. *Proceedings of the 24th ACM international conference on Multimedia*, pages 172–176, 2016.

[42] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[43] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020.

[44] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.