https://doi.org/10.2352/EI.2023.35.7.IMAGE-269

Self-Supervised Visual Representation Learning on Food Images

Andrew Peng, Jiangpeng He, Fengqing Zhu

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

Abstract

Food image analysis is the groundwork for image-based dietary assessment, which is the process of monitoring what kinds of food and how much energy is consumed using captured food or eating scene images. Existing deep learning-based methods learn the visual representation for downstream tasks based on human annotation of each food image. However, most food images in real life are obtained without labels, and data annotation requires plenty of time and human effort, which is not feasible for real-world applications. To make use of the vast amount of unlabeled images, many existing works focus on unsupervised or self-supervised learning of visual representations directly from unlabeled data. However, none of these existing works focus on food images, which is more challenging than general objects due to its high inter-class similarity and intra-class variance.

In this paper, we focus on the implementation and analysis of existing representative self-supervised learning methods on food images. Specifically, we first compare the performance of six selected self-supervised learning models on the Food-101 dataset. Then we analyze the pros and cons of each selected model when training on food data to identify the key factors that can help improve the performance. Finally, we propose several ideas for future work on self-supervised visual representation learning for food images.

Introduction

Poor diet choices are linked to several health conditions such as cancer, heart diseases, and diabetes, some of the leading preventable causes of death. Additionally, the CDC reports that 9 in 10 Americans consume too much sodium, which may cause high blood pressure, heart disease, and strokes. Furthermore, nearly $173 billion is spent annually on health care for obesity [1]. However, it is difficult to accurately assess the dietary intake of a person, as traditional methods [2, 3] are based on self-reported information which may include errors due to recall or bias. On the other hand, image-based dietary assessment technologies [4] utilize eating occasion images captured by participants to determine their dietary intake. Due to the reduced amount of human input, such technologies can greatly improve the accuracy and reliability of a person's dietary information.

Nowadays, the vast majority of image-based dietary assessment technologies leverage deep learning for food recognition [5, 6, 7, 8], segmentation [9] and portion size estimation [10, 11, 12]. One of the major challenges of existing supervised methods, however, is their requirement for large amounts of annotated training data. Since most food images in real world are captured without labels, an additional step for data annotation is needed, which would be expensive and time-consuming. On the other hand, unsupervised and self-supervised learning models [13, 14, 15] can learn visual representations directly from unlabeled data to perform downstream tasks. As shown in Fig. 1, unsupervised learning trains a feature encoder from unlabeled images to classify the image into a certain category. We will focus on self-supervised learning as it concentrates on downstream tasks while unsupervised learning is more for clustering and dimensionality reduction. Though numerous deep learning approaches have been developed for self-supervised learning of general tasks, none of them have been tested specifically on food images, which is known to be more challenging due to their intra-class diversity and inter-class similarity [5].

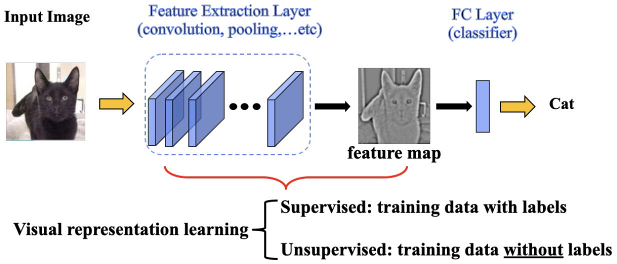
Input Image
Feature Extraction Layer
(convolution, pooling,...etc)
FC Layer
(classifier)
Cat
feature map
Visual representation learning
Supervised: training data with labels
Unsupervised: training data without labels

Figure 1: An overview of supervised and unsupervised visual representation learning.

In this paper, we aim to explore the performance of existing self-supervised learning methods on food images and provide insightful potential directions on improving the performance in future works. We select six state-of-the-art self-supervised learning methods including SimCLR [16], SwAV [17], BYOL [18], SimSiam [19], MoCo v2 [20, 21], and DINO [22], which are representative contrastive based, non-contrastive based and vision transformer based methods, respectively. We evaluate and analyze the self-supervised learning performance on Food-101 dataset [23], which contains 101 different foods with 1000 images each. Specifically, we first show that self-supervised learning on food images is more challenging by comparing the performance between Food-101 and reported results on ImageNet [24]. Then, we analyze the pros and cons of each selected model based on their performance on Food-101 to identify several insights and propose possible future steps for increasing the accuracy and efficiency of self-supervised learning on food images. The contribution of this work can be summarized as the following.

- To best of our knowledge, we are the first to systematically study the existing representative self-supervised methods on food images.
- We conduct extensive experiments on Food-101 to identify the challenges behind learning on food images compared to general tasks such as ImageNet.
- By analyzing the results, we provide insightful directions to

potentially improve the performance in future works.

Related Work

Self-supervised learning (SSL) and unsupervised learning are two types of models that are trained on completely unlabeled data. Unsupervised learning methods mainly focuses on finding patterns and clustering the data based on similar features. On the other hand, SSL methods attempt to solve tasks by augmenting the unlabeled data, such as rotating the images or taking two different augmentations of the same image.

In this work, we analyze three main categories of SSL models: *Contrastive Based Learning*, *Non-Contrastive Based Learning* and *Vision Transformers (ViTs) Based*. We selected these three categories because of their proven effectiveness on the ImageNet dataset, and because they are three of the most common types of self-supervised image classification models. Furthermore, both Contrastive and Non-contrastive based models that we examine are Siamese models, which are models that compare two augmentations of the same image to learn visual representations. Below we summarize and illustrate each category in detail.

(1) **Contrastive-based learning** is a common self-supervised learning algorithm that takes two augmentations of the same image, called positive pairs, and maximizes the agreement between the two, while also minimizing the agreement between two augmentations of different images, called negative pairs. One common drawback of contrastive models, however, is their necessity for larger batch sizes, since they require both positive and negative pairs. We selected two most popular methods in this category including SimCLR [16] and MoCo v2 [20, 21], which has a similar structure and comparable performance. Additionally, SimCLR and MoCo v2 are based on similar ideas, with MoCo coming out first and later revised after SimCLR's publish to MoCo v2 [21]. SimCLR follows the straightforward contrastive learning framework of comparing both positive and negative pairs, with an additional custom optimizer. MoCo instead uses a memory bank to store negative pairs, with MoCo v2 utilizing multi-crop to further increase the performance.

(2) **Non-contrastive based** learning models also utilize the idea of positive pairs while excluding the negative pairs. They utilize additional techniques to improve performance and also prevent the model from collapsing [25], which is a common failure in Siamese models when the encoder outputs a constant representation regardless of input. On the other hand, contrastive-based learning models do not collapse since they also contrast negative pairs. For this category, we selected SwAV [17], BYOL [18], and SimSiam [19], three models that each have their own unique components. SwAV incorporates online clustering within a Siamese model. Similar to MoCo v2, this model also uses multi-crop. Alternatively, BYOL employs two neural networks, called the online and target network, to predict each other's representation of the same image, along with a momentum encoder. Finally, SimSiam [19] is a simplistic network that only uses positive pairs and a stop-gradient operation, which prevents certain parts of the model from being updated to prevent collapsing.

(3) **Vision Transformers** [26] are another type of framework based on the self-attention-based Transfomer [27], which is the main method used in Natural Language Processing (NLP). For image tasks, attention-based models have historically underperformed compared to convolutional models such as CNNs and ResNet [28] due to inefficiencies, but are recently improving in accuracy and computational speed due to modifications on the attention portion of ViTs. One of the main discrepancy between the CNN and ViTs is that Vision Transformers [26] lack certain inductive biases that CNNs are able to produce, such as locality. However, ViTs' accuracy scale up based on the amount of image data, making them ideal for self-supervised image classification where models are typically trained on millions of images. For this category, we selected DINO [22], a state-of-the-art ViT-based model that also utilizes knowledge distillation [29]. We selected DINO because of its high performance and uniqueness and also we hope to explore other self-supervised learning approaches outside of contrastive and non-contrastive learning to see if they would perform better on food images.

Method

In this section, we illustrate the selected methods in detail from the perspective of each main category.

Contrastive based Learning

SimCLR and MoCo v2 both exhibit all the features of contrastive learning, using both positive and negative pairs. A brief overview of contrastive learning framework is shown in Fig. 2, which include four main components: (1) a data augmentation module that randomly generates the two different views of the same image x, denoted as x_i and x_j, (2) the common ResNet [28] encoder, represented as $f(\bullet)$, that extracts representation features from the image, (3) a projection head $g(\bullet)$ that calculates contrastive loss, defined as an multilayer perceptron (MLP) with one hidden layer for SimCLR and two hidden layers for MoCo v2 with ReLU [30], and (4) a contrastive loss function used to maximize the similarity between positive pairs and minimize the similarity between negative pairs.

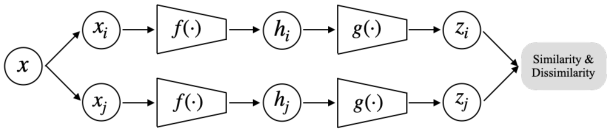

Figure 2: An overview of contrastive learning framework.

SimCLR and MoCo v2 use different contrastive loss functions. SimCLR uses *NT-Xent* (the normalized temperature-scaled cross-entropy loss), defined as in Eq. (1).

$$\ell_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\Sigma_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)} \tag{1}$$

where $sim(u,v) = u^T v/|u||v|$, N is the batch size, $\mathbb{1}_{[k \neq i]} \in 0,1$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature hyper-parameter. MoCo v2, on the other hand, uses InfoNCE, defined as in Eq. (2).

$$\mathscr{L}_q = -\log \frac{\exp(q \cdot k_i)/\tau}{\Sigma_{i=0}^{k} \exp(q \cdot k_+)/\tau} \tag{2}$$

where q is the query, k_+ is the positive key, k_i are the other keys, and τ also denotes a temperature hyper-parameter. These equations represent the final step of Fig. 2, and both of these functions aim to maximize the similarity between positive pairs and minimize the similarity between negative pairs. In this work, we select these two methods to represent the contrastive-based models and further evaluate the performance on food images, which can be more challenging due to the intra-class diversity and inter-class similarity.

Non-Contrastive based Learning

SwAV, BYOL, and SimSiam are non-contrastive based models as they do not incorporate negative pairs, but they are all Siamese models because they compare positive pairs. Additionally, since all three models are structured the same conceptually, we will examine SimSiam in detail as it is more representative compared to BYOL and SwAV, which both have additional unique features. SimSiam also demonstrates that they can learn visual representations with only a stop-gradient operation. In Fig. 3, an overview of non-contrastive learning is shown, which demonstrates the fundamental framework of our three models. Similar to SimCLR, $f(\bullet)$ represents an encoder while $h(\bullet)$ denotes the MLP head.

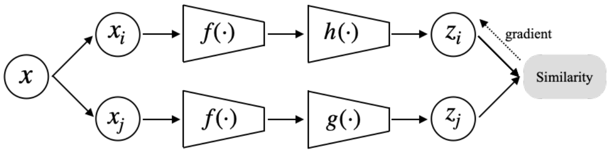

Figure 3: An overview of non-contrastive learning framework.

SimSiam utilizes a simplistic loss function defined as in Eq. (3):

$$\mathscr{L} = \frac{1}{2}\mathscr{D}(a_1, \text{stopgrad}(b_2)) + \frac{1}{2}\mathscr{D}(a_2, \text{stopgrad}(b_1)) \qquad (3)$$

where $a_1 = h(f(x_1))$ and $b_2 = f(x_2)$ and $\mathscr{D}(a_1, b_2)$ represents the negative cosine similarity: $-\frac{a_1}{\|a_1\|_2} \cdot \frac{b_2}{\|b_2\|_2}$, where $\|\cdot\|_2$ is the l_2-normalized vector. Since both SwAV and BYOL are different from SimSiam, we will consider their unique features, represented by $g(\bullet)$. SwAV utilizes online clustering with Sinkhorn-Knopp transform [31] and BYOL directly predicts the output of one view from another view using a momentum encoder. Both these methods are used to prevent collapsing and improve accuracy. Although non-contrastive methods do not utilize negative pairs, they are all still Siamese networks. Therefore, we evaluate their performance to see if visual representations of food images will be learned well by these methods.

Vision Transformer-based Learning

DINO is the representative model we selected in Vision Transformer based category. It is a self-supervised learning approach that utilizes *self-distillation with no labels*. In addition to SSL, it utilizes knowledge distillation [29], which involves training a student network's probability distribution based on an input image to match the output of a teacher network. By maximizing the similarity between their predictions and propagating the information to update the networks, the model is able to learn visual representations of different images. An overview of DINO is shown in Fig. 4. A positive pair is passed into the two networks represented by g, which are composed of a backbone ViT and a projection MLP head similar to the one used in SwAV. The loss function is defined as $\min H(P_t(x), P_s(x))$, which takes the cross-entropy loss of probability distributions of the teacher and student network. $H(a,b) = -a\log b$ and P_s is defined as in Eq. (4):

$$P_s = \frac{\exp(g_{\theta_s}(x)/\tau_s)}{\Sigma_{k=1}^{K}\exp(g_{\theta_s}(x)/\tau_s)} \qquad (4)$$

with τ_s as a hyperparameter. Additionally, a stop-gradient operator (SG) is applied to propagate gradients only through the student, while the teacher parameters are updated with an exponential moving average (EMA), defined by the formula $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$, where θ_t and θ_s are parameters, and λ follows a cosine schedule from 0.996 to 1. We chose this Vision Transformer-based model because it was fundamentally different from the other methods while also achieving high performance on the ImageNet dataset. Additionally, DINO also uses knowledge distillation, which is another technique we hope to explore on food images.

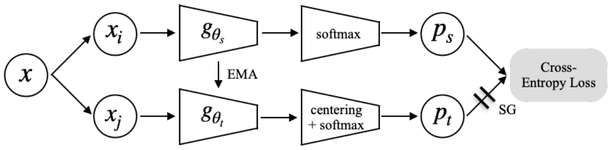

Figure 4: An overview of the DINO model.

Experiments

In this section, we first evaluate the selected six self-supervised method by comparing the performance on both general object dataset and food image dataset. Then, we specifically analyze the results on food data and summarize the pros and cons of each selected model. Finally, we provide insights as future work to further improve the performance on food images.

Experimental Setup

Datasets: we used the Food-101 dataset, which includes 101 different food classes with $1,000$ images each, providing a total of $101,000$ food images. Each food class is further divided into 750 training images with 250 test images. Additionally, some of the images are purposely uncleaned with a certain amount of noise, such as intense colors and mislabeled images. We selected this dataset as it is one of the most well-known food datasets used for various downstream tasks.

While focusing on food images, we also leverage ImageNet dataset containing images of general objects as a reference compared to Food-101. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012-2017 contains 1000 image classes and over 1.2 million images, $50,000$ validation images and $100,000$ test images, which is commonly used as a benchmark to evaluate the model performance especially on image classification tasks.

As shown in Fig. 5, the images in Food-101 is of higher intra-class diversity and inter-class similarity compared to ImageNet datasets, making it more challenging to learn the visual representation from unlabeled data.

Evaluation metric: we use the widely applied *Linear Evaluation* as the evaluation metric, which trains a supervised linear classifier on frozen features learned by self-supervised visual representation learning. Specifically, a fully-connected layer followed by softmax is trained on the test images, and the gradients are not propagated back to the frozen features which ensure the feature extractor does not learn anything from supervised labels.

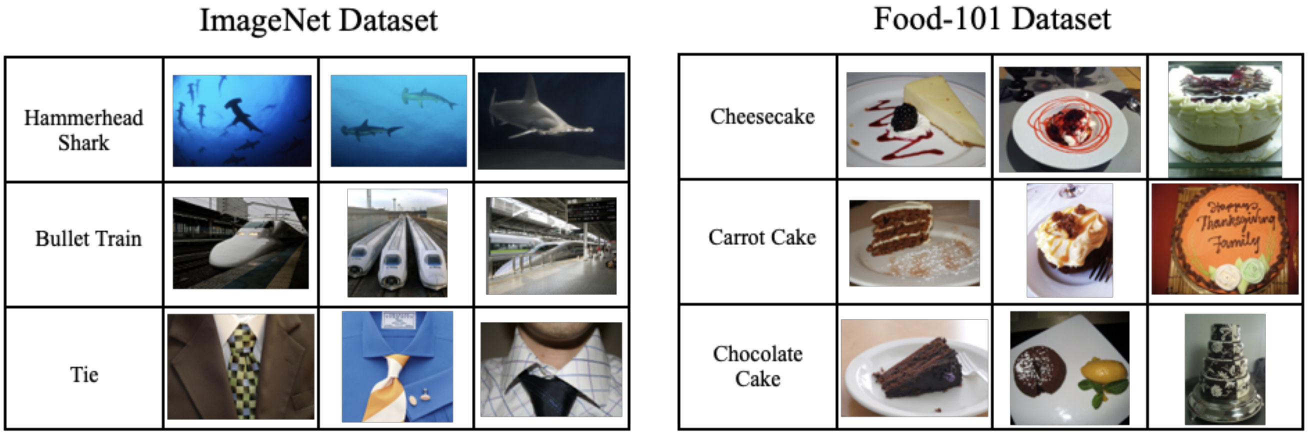

Figure 5: Intra-class diversity and inter-class similarity in Food-101 compared to ImageNet.

	SimCLR	SwAV	BYOL	SimSiam	MoCo v2	DINO
Batch Size	256	256	128	128	256	64
Epochs	100	100	100	100	100	100
Backbone	ResNet-50	ResNet-50	ResNet-50	ResNet-50	ResNet-50	ViT-S
Optimizer	LARS	SGD	SGD	SGD	SGD	AdamW

Table 1: Implementation Details

	SimCLR	SwAV	BYOL	SimSiam	MoCo v2	DINO
Accuracy (%)	51.0	54.7	47.7	44.5	53.9	61.4
Training Time	2 days	2 days	3 days	2 days	3 days	2 days
Memory Size	107M	217M	283M	292M	305M	672M

Table 2: Experimental results on Food-101

Implementation details: In Table 1, we summarize the implementation details of each selected model. We selected 100 epochs for all the models and ResNet-50 for contrastive and non-contrastive-based models. Additionally, we chose reasonable batch sizes of 128/256 for the contrastive and non-contrastive models, which were the maximum allowed by our computational resources. For DINO, we chose 64 batch size due to it being a ViT-based model, which claims to require less batch size.

Results on Food-101

The experimental results of six selected self-supervised methods on Food-101 are summarized in Table 2. We include the top-1 linear evaluation accuracy (%) along with training time and how much memory the model parameters use to show that our trained models require a similar amount of computational resources. Only DINO takes up significantly more memory, which is due to ViTs requiring storing more memory after training. From the results, we observe that DINO performed the best while BYOL and SimSiam performed the worst. We expected slightly lower accuracy in SimSiam because it had no unique method to improve accuracy, but BYOL's lower accuracy was unexpected. Additionally, we notice that the other three models, SimCLR, SwAV, and MoCo v2, have similar accuracy, showing that each model's unique methods increased their accuracy. The performance difference between the best and worst model is approximately 18.5%, which is quite significant.

Comparison Between Food-101 and ImageNet

In Fig. 6, we included the results on Food-101 side-by-side with the results on the ImageNet dataset. These results were obtained using the same number of epochs and backbone encoder.

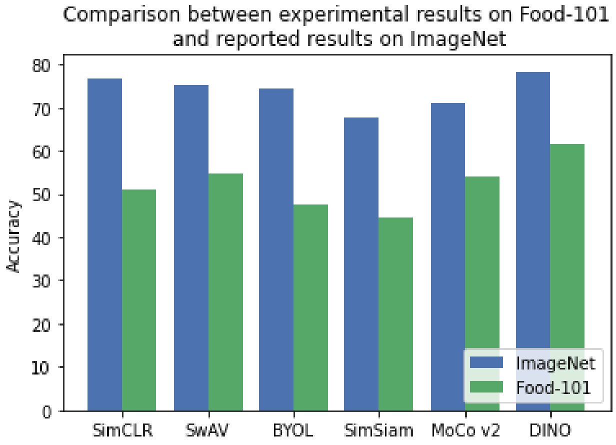

Figure 6: Linear evaluation results on ImageNet and Food-101.

Our experimental results on Food-101 are much lower than the results on ImageNet. This observation also shows that the visual representation on food images can be more challenging than general objects in real life due to higher intra-class diversity and inter-class similarity, resulting in the lower accuracy. Another reason for the lower accuracy could be that the Food-101 contains fewer images than ImageNet, which hurts the performance of existing self-supervised models as they rely on massive amounts of data for training.

We also notice that BYOL performs worse than expected with the lowest performance on Food-101 out of all of the selected models, despite it achieving a high accuracy on ImageNet. One of the possible reason is that the BYOL reuiqres larger batch size where it achieves 74.3% accuracy after training on 512 TPUs with 4096 batch size, while our model was only trained on 128 batch size with ResNet-50. Therefore, we speculate that BYOL and other models could scale up in accuracy with larger batch size and computation resources. To prove our assumption, we ran intermediate experiments using SimCLR as shown in Table 3. The increased batch size results in a notable increase in accuracy even

Batch Size	64	128	256
Accuracy	41.8	48.0	51.0

Table 3: SimCLR accuracy on Food-101 with various batch sizes.

without increasing the number of epochs or changing the backbone network. However, the computation resource is one of the major constrains in deep learning especially for real world applications.

Additionally, we noticed that DINO performed much better than all the contrastive models. One possible reason is that ViTs perform better than ResNet as the backbone for visual representation learning. This could possibly be due to the fact that self-supervised learning uses a large amount of training data, which benefits ViTs more significantly due to their unique attention modules. Another possible reason is that the lower batch sizes hurt more performance of contrastive learning models, as they require both positive and negative pairs for training. Non-contrastive models are also impacted by lower batch sizes, although to a lesser extent. Therefore, we have demonstrated that batch sizes and, in general, computational resources are more impactful for contrastive and non-contrastive models, while ViTs do not depend as much on batch sizes when compared with contrastive and non-contrastive models.

Finally, we compare the performance of four Siamese models: SimCLR, MoCo v2, SwAV, and SimSiam. We exclude BYOL due to its lower-than-expected accuracy from our comparison. Firstly, SimSiam has a lower accuracy than the other three models, which all have very similar performance, although SwAV and MoCo v2 perform slightly better as they both adopted some ideas from SimCLR. This shows that the unique features in SimCLR, MoCo v2, and SwAV improved their accuracy compared to SimSiam's stop-gradient operation. Furthermore, both SwAV and MoCo v2 add an extra layer of complexity with their unique features, which are online clustering and momentum encoders, respectively. Through this comparison, we see that contrastive and non-contrastive based models perform similarly. Since both models achieved similar high performances, we can conclude that the Siamese learning framework is efficient in learning visual representations.

Insightful Directions for Future Work

Based on our experiments and analysis, we proposed three ideas on how to improve accuracy in the future:

- **Fine-Tuning or Transfer Learning.** This method involves training a model on a large dataset, for example ImageNet, and then transferring on Food-101. This approach could help resolve the issue of an insufficient amount of training data since the model will have learned visual representations on a larger dataset. Therefore, when fine-tuning on a smaller dataset, the model will not require a large batch size to learn visual representations from scratch.
- **Larger Computational Resources.** As already shown in Table 3, accuracy scales up with larger batch sizes and more training epochs. Therefore, we would expect better performance if the models are trained with larger computation resources.
- **Ensemble of Models.** We propose that combining certain models could improve accuracy. We observed that each method category had its own unique techniques which improved their accuracy, so combining some of the methods together may be a potential solution. For example, we researched pre-text tasks, which are unsupervised image-based problems solved to learn the visual representation of an image, such as colorizing a black-and-white image. We will be examining the rotation pre-text task, which predicts if an image is rotated 0°, 90°, 180°, and 270°, implemented by the model RotNet [32]. We propose that combining SimCLR with RotNet, for example, could further improve the accuracy because it learns more visual representations.

Conclusion

Overall, we explored the performance of 6 state-of-the-art models from 3 main categories on food images, specifically the Food-101 dataset. Our experimental results show that visual representation learning is more challenging on food images by comparing performance on ImageNet and Food-101. Additionally, all three categories of models show promising results on food data. The experimental results suggest that ViT models are worth exploring further for self-supervised image tasks, but contrastive and non-contrastive models should still be considered when working on self-supervised classification tasks. Finally, based on our analysis, we propose that there is also potential for transfer learning or combining models to help improve accuracy.

References

[1] "Poor nutrition," www.cdc.gov/chronicdisease/resources/publications/factsheets/nutrition.html, Sep 2022.

[2] C. J. Boushey, M. Spoden, F. M. Zhu, E. J. Delp, and D. A. Kerr, "New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods," *Proceedings of the Nutrition Society*, vol. 76, no. 3, p. 283–294, 2017.

[3] Z. Shao, Y. Han, J. He, R. Mao, J. Wright, D. Kerr, C. Boushey, and F. Zhu, "An Integrated System for Mobile Image-Based Dietary Assessment," 2021. [Online]. Available: https://arxiv.org/abs/2110.01754

[4] Z. Shao, Y. Han, J. He, R. Mao, J. Wright, D. Kerr, C. J. Boushey, and F. Zhu, "An Integrated System for Mobile Image-Based Dietary Assessment," *Proceedings of the 3rd Workshop on AIxFood*, p. 19–23, 2021.

[5] R. Mao, J. He, Z. Shao, S. K. Yarlagadda, and F. Zhu, "Visual Aware Hierarchy Based Food Recognition," *arXiv preprint arXiv:2012.03368*, 2020.

[6] J. He, L. Lin, H. Eicher-Miller, and F. Zhu, "Long-tailed Food Classification," *arXiv preprint arXiv:2210.14748*, 2022.

[7] X. Pan, J. He, A. Peng, and F. Zhu, "Simulating Personal Food Consumption Patterns using a Modified Markov Chain," *arXiv preprint arXiv:2208.06709*, 2022.

[8] J. He and F. Zhu, "Online Continual Learning for Visual Food Classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2337–2346, October 2021.

[9] W. Shimoda and K. Yanai, "CNN-Based Food Image Segmentation Without Pixel-Wise Annotation," *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, pp. 449–457, 2015.

[10] Z. Shao, S. Fang, R. Mao, J. He, J. Wright, D. Kerr, C. J. Boushey, and F. Zhu, "Towards Learning Food Portion From Monocular Images With Cross-Domain Feature Adaptation," *arXiv preprint arXiv:2103.07562*, 2021.

[11] J. He, Z. Shao, J. Wright, D. Kerr, C. Boushey, and F. Zhu, "Multi-task Image-Based Dietary Assessment for Food Recognition and Portion Size Estimation," *2020 IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 49–54, 2020.

[12] J. He, R. Mao, Z. Shao, J. L. Wright, D. A. Kerr, C. J. Boushey, and F. Zhu, "An End-to-End Food Image Analysis System," *Electronic Imaging*, vol. 2021, no. 8, pp. 285–1, 2021.

[13] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised Visual Representation Learning by Context Prediction," 2015. [Online]. Available: https://arxiv.org/abs/1505.05192

[14] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative Unsupervised Feature Learning with Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[15] M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," 2016. [Online]. Available: https://arxiv.org/abs/1603.09246

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *Proceedings of the International Conference on Machine Learning*, vol. 119, pp. 1597–1607, 13–18 Jul 2020. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," 2020. [Online]. Available: https://arxiv.org/abs/2006.09882

[18] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised Learning," 2020. [Online]. Available: https://arxiv.org/abs/2006.07733

[19] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 745–15 753.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2019. [Online]. Available: https://arxiv.org/abs/1911.05722

[21] X. Chen, H. Fan, R. Girshick, and K. He, "Improved Baselines with Momentum Contrastive Learning," 2020. [Online]. Available: https://arxiv.org/abs/2003.04297

[22] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.14294

[23] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining Discriminative Components with Random Forests," *European Conference on Computer Vision*, 2014.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[25] A. C. Li, A. A. Efros, and D. Pathak, "Understanding Collapse in Non-Contrastive Siamese Representation Learning," 2022. [Online]. Available: https://arxiv.org/abs/2209.15007

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[30] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," 2018. [Online]. Available: https://arxiv.org/abs/1803.08375

[31] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[32] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," 2018. [Online]. Available: https://arxiv.org/abs/1803.07728

Author Biography

Andrew Peng is currently a senior at Henry M. Gunn High School in Palo Alto, CA. He conducted research over the summer and throughout 2022–2023 at VIPER Lab with the School of Electrical and Computer Engineering in Purdue University, West Lafayette, IN. His current research interests lie in Image Classification and Signal Processing using deep neural networks.

Jiangpeng He received his Ph.D. degree in Electrical and Electronic Engineering from Purdue University in August 2022. He is currently a postdoc research assistant at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include image processing, computer vision, image-based dietary assessment and deep learning.

Fengqing Zhu is an Associate Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. Dr. Zhu received the B.S.E.E. (with highest distinction), M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University in 2004, 2006 and 2011, respectively. Her research interests include image processing and analysis, video compression and computer vision. Prior to joining Purdue in 2015, she was a Staff Researcher at Futurewei Technologies (USA).