COM322: Computer Vision
Food Image Classification

Derin Gezgin

Computer Vision & Food

Computer vision is a wide and rapidly expanding area that includes various sub-fields, such as object detection, segmentation, image acquisition, image classification, etc. Similarly, computer vision applications extend across many domains, such as medicine, ecology, environmental science, and product development. As time passes, new areas and innovative applications of computer vision come out as a result of the advancements in technology and the increasing amount of available data.

Food detection and classification is also another popular area in computer vision. Its applications range from simple food image classification [1][2][3] to recipe generation from food images [4][5]. Another sub-field of food recognition is fruit and vegetable recognition [6][7]. This paper will explore the classic food image classification problem where you have to label dishes solely based on dish names, compared to other examples involving ingredient detection. This paper will also review the available food image datasets and discuss prior work in this area, including their methodologies, results, and common challenges.

Available Datasets

As image classification requires a large amount of data, many prior works in this area introduced their dataset as part of their presentation. Depending on the scope of the study, some datasets focus on a specific cuisine, while some focus on a general palette of cuisines. The following table presents the food image datasets, including the paper that introduced them, whether they are publicly available, their class and total image counts, and their cuisine information. These datasets are selected by their ease of finding and accessibility for future research. There is a much greater variety of available datasets from different cuisines [2], but as they are not easily accessible, I wanted to exclude them from this table for the sake of simplicity.

Dataset Name	Class Count	Total	Cuisine	Availability	Paper
Food2k	2000	1,036,564	Mixed	✓	[8]
ISIA Food-500	500	399,726	Mixed	✓	[9]
Food524DB	524	247,636	Mixed	✓	[10]
Food-475	475	247,536	Mixed	✓	[11]
CNFOOD-241	241	191,786	Chinese	✓	[12]
ChineseFoodNet	207	185,628	Chinese	✓	[13]
FoodX-251	251	158,846	Mixed	✓	[14]
VireoFood-172	172	110,241	Mixed	By request	[15]
ETHZ Food-101	101	101,000	Mixed	✓	[16]
FFoCat	156	58,962	Mixed	✓	[17]
UEC Food-100	256	31,397	Mixed	✓	[18]
Food-11	11	16,643	Mixed	✓	N/A
THFood-50	50	15,770	Thai	✓	[19]
UEC Food-256	100	9,060	Mixed	✓	[20]
KenyanFood13	13	8,174	Kenyan	✓	[21]
FoodSeg103	104	7,118	Mixed	✓	[22]
Indian Food Image Dataset	N/A	5,000	Indian	Paid	N/A
Food-5k	2	5,000	Mixed	✓	N/A

Table 1: Selected food image datasets

Related Work

Food image classification is a problem tackled by computer vision researchers using a wide range of methodologies. This section will explore the classification methodologies in depth.

(Deep) Convolutional Neural Networks

Over many years since their introduction [23], (deep) convolutional neural networks have been the first go-to of computer vision researchers. Food image classification also has its share of (D)CNN solutions.

Kawano and Yanai [24] introduced one of the first examples of the usage of deep learning in food classification. They utilized a pre-trained DCNN on the ILSVRC 1000-class dataset for feature extraction. They achieved 72.56% top-1 accuracy and 92,00% top-5 accuracy with the UEC Food-100 dataset. Similarly, Yanai and Kawano [25] fine-tuned a

DCNN model pre-trained on the ImageNet dataset. This network achieved 78.77% top-1 accuracy for the UEC Food-100 dataset and 67.57% top-1 accuracy for the UEC Food-256 dataset.

In 2016, Hassanejad et al. [26] fine-tuned Google's Inception architecture [27] -a DCNN with 54 layers- for the classification of food images from ETH Food-101, UEC Food-100, and UEC Food-256 datasets. This architecture got results that were better than any of the published work on these datasets. It achieved 88.28% top-1 / 96.88% top-5 accuracy in ETH Food-101, 81.45% top-1 / 97.27% top-5 accuracy in UEC Food 100, and 76.17% top-1 / 92.58% top-5 accuracy in UEC Food 256 datasets. Moreover, this methodology requires fewer computational resources than its competitors, considering it has fewer parameters and low computational complexity.

WISeR, a wide-sliced residual network, outperformed the Inception architecture in 2016 [28]. This is a two-branch network with a branch of a deep-residual network and another branch of a slicing convolutions. The residual network detects the general visualization of images, while the slicing convolutions detect the vertical food layers. This slicing kernel captures the vertical layer structure of dishes compared to normal square convolutional kernels, which capture local information. The vertical slicing significantly improved the accuracy. This architecture achieved 89.58% top-1 / 99.23% top-5 accuracy in UEC Food-100, 83.15% top-1 / 95.45% top-5 accuracy in UEC Food-256, and 89.58% top-1 / 99.23% top-5 accuracy in Food-101 datasets.

Matarat [29] utilized a fine-tuned CNN model to classify Thai food images. This is an example study as it specifically works on the Thai food classification compared to other studies on more general datasets. It used a pre-trained Conv2D layer and MobileNet model to extract the crucial features and trained the classification layers. This structure achieved over 80% accuracy in the classification of Thai food images.

Transformer Models and Attention

The introduction of the transformer architecture [30], followed by their usage in computer vision with the vision transformers starting from 2020 [31], marked a significant breakthrough in the computer vision field. Considering the newness of the vision transformers,

there are not many applications of them in the food image classification sub-field.

Gao et al. [32] used the LayerScale method [33] with the original vision transformer architecture. This addition is crucial to improve the training in the deep networks. Moreover, this paper utilized the feature local enhancement to catch the important features of the images (the CNNs important skill) while also looking at long-term relationships between the features, an important part of the vision transformers. This was done by integrating convolutional layers into the multi-layer-perceptron (MLP) structure. This model achieved 95.17% accuracy in the Food-101 dataset and 94.29% accuracy in the Food-172 dataset.

Xiao et al. [34], utilized a Swin-Transformer [35], that was pre-trained on the ImageNet-22k dataset. This pre-trained model was fine-tuned using Foodx-251 and UEC Food-256 datasets. Similar to the previous paper, this paper proposes a deep convolutional neural network for enhanced feature representation on top of the global feature extraction using the Swin-Transformer. This architecture achieved 81.07% accuracy in Foodx-251 and 82.77% accuracy in the UEC Food-256 dataset.

Transfer Learning

Transfer learning is a commonly used methodology in machine learning, data mining, and their applications [36]. As a field where the amount of data can be limited in some cases, computer vision is also one of the areas that heavily benefit from transfer learning [37].

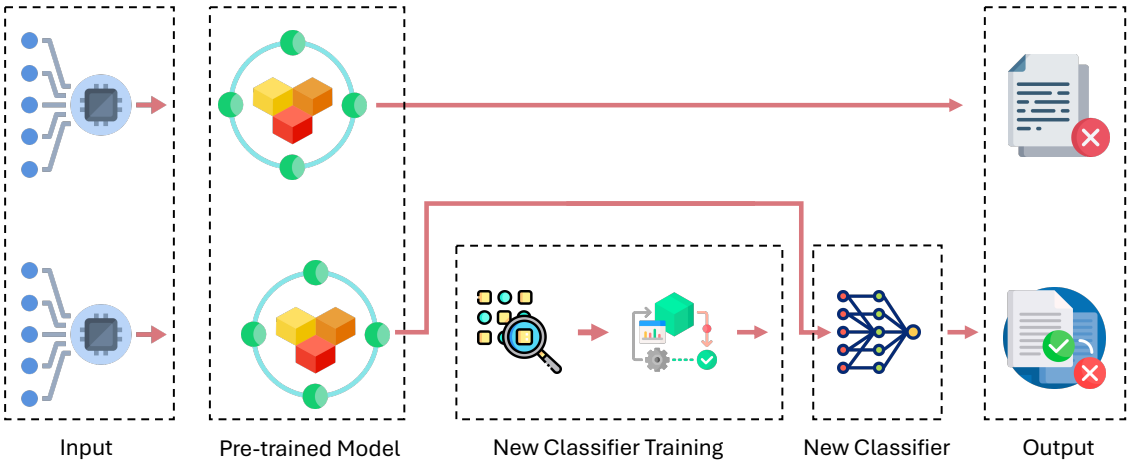

Figure 1: Transfer learning visualization [37]

Transfer learning is a machine learning technique that uses the knowledge gained

from one field in another field. It can be used when a large enough dataset is not available in an area.

As is shown in previous examples, the majority of the studies in food image classification leveraged transfer learning [24][25][26][29][34]. Transfer learning is widely used in this area, considering -except a few- food image datasets are not large enough to train a model from scratch. Similarly, as one of the important areas in food recognition is feature extraction, pre-trained models trained on many images are already valuable for feature extraction. Directly using their feature extraction capability is significant enough that we can only train the last few linear classification layers to have good enough results. In cases where this feature extraction is insufficient, some extra feature extraction techniques like local feature enhancement can be significantly helpful.

Common Challenges

Among all these mentioned papers, some common challenges are worth noting for future research in this area.

Need for More Data and its Drawbacks

Like any computer vision task, the dataset size can be crucial in training an image classification model. Many papers in food classification also mentioned this as a crucial problem and proposed data augmentation solutions for this [32]. While data augmentation can be a solution for this, it can create other problems like over-fitting when done a lot. At that point, some alternative methods like stable diffusion can be used [38]

Generalization of Classification

While with large datasets and proper data augmentation, it is possible to have a nice classification algorithm, making this classifier useful in the different pictures of the same food is still a challenge. As the same food can be done differently and can vary in shape, size, and color, on top of lacking a rigid structure, it is very hard to have a generalizable and smart enough classifier for food images. Some papers propose different convolution techniques to overcome this issue [28][39][40]. Classic data augmentation can also be helpful for this issue.

Conclusion & Future Work

Food classification is one of the newly explored areas of computer vision. With the rise of convolutional neural networks and deep learning, many solutions with these methodologies have also been offered for food classification. Recently, with the introduction of the transformer architecture and vision transformers, a few solutions leveraging these architectures have also been offered.

While CNN/DCNN and vision transformer solutions have proven to work in this area, some improvements are possible considering the new methodologies and datasets. At the same time, multi-cuisine models can be developed for more robust classifiers. Lastly, the models can be made more generalized to detect different types of the same dishes, including their slight variations.

References

[1] L. Zhu, P. Spachos, E. Pensini, and K. N. Plataniotis, "Deep learning and machine vision for food processing: A survey," *Current Research in Food Science*, vol. 4, pp. 233–249, 2021.

[2] Y. Zhang, L. Deng, H. Zhu, *et al.*, "Deep learning in food category recognition," *Information Fusion*, vol. 98, p. 101 859, 2023.

[3] W. Min, S. Jiang, L. Liu, Y. Rui, and R. C. Jain, "A survey on food computing," *ACM Computing Surveys (CSUR)*, vol. 52, pp. 1–36, 2018.

[4] P. Chhikara, D. Chaurasia, Y. Jiang, O. Masur, and F. Ilievski, *Fire: Food image to recipe generation*, 2024.

[5] J. Chen and C.-w. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM International Conference on Multimedia*, Association for Computing Machinery, 2016, pp. 32–41.

[6] H. B. Muresan and M. Oltean, "Fruit recognition from images using deep learning," *Acta Universitatis Sapientiae, Informatica*, vol. 10, pp. 26–42, 2017.

[7] V. Meshram and K. Patil, "Fruitnet: Indian fruits image dataset with quality for machine learning applications," *Data in Brief*, vol. 40, p. 107 686, 2022.

[8] W. Min, Z. Wang, Y. Liu, *et al.*, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 9932–9949, 2021.

[9] W. Min, L. Liu, Z. Wang, *et al.*, "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*, 2020.

[10] G. Ciocca, P. Napoletano, and R. Schettini, "Learning cnn-based features for retrieval of food images," in *New Trends in Image Analysis and Processing – ICIAP 2017*, S. Battiato, G. M. Farinella, M. Leo, and G. Gallo, Eds., 2017, pp. 426–434.

[11] G. Ciocca, P. Napoletano, and R. Schettini, "Cnn-based features for retrieval and classification of food images," *Computer Vision and Image Understanding*, vol. 176-177, pp. 70–77, 2018.

[12] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, "Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning," *arXiv preprint arXiv:2402.15761*, 2024.

[13] X. Chen, H. Zhou, and L. Diao, "Chinesefoodnet: A large-scale image dataset for chinese food recognition," *ArXiv*, vol. abs/1705.02743, 2017.

[14] P. Kaur, K. Sikka, W. Wang, S. J. Belongie, and A. Divakaran, "Foodx-251: A dataset for fine-grained food classification," *ArXiv*, vol. abs/1907.06167, 2019.

[15] C.-w. N. Jing-jing Chen, "Deep-based ingredient recognition for cooking recipe retrival," *ACM Multimedia*, 2016.

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

[17] I. Donadello and M. Dragoni, "Ontology-driven food category classification in images," in *ICIAP (2)*, ser. Lecture Notes in Computer Science, vol. 11752, Springer, 2019, pp. 607–617.

[18] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.

[19] C. Termritthikun, P. Muneesawang, and S. Kanprachar, "NU-InNet: Thai food image recognition using convolutional neural networks on smartphone," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 2-6, pp. 63–67, 2017.

[20] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[21] K. Wang, M. Jalal, S. Jefferson, Y. Zheng, E. O. Nsoesie, and M. Betke, "Scraping social media photos posted in kenya and elsewhere to detect and analyze food types," *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, 2019.

[22] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S. C. Hoi, and Q. Sun, "A large-scale benchmark for food image segmentation," in *Proceedings of ACM international conference on Multimedia*, 2021.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 589–593.

[25] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015, pp. 1–6.

[26] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, 2016, pp. 41–49.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2015.

[28] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 567–576, 2016.

[29] K. Matarat, "Enhancing thai food classification: A cnn-based approach with transfer learning," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 6, pp. 1633–1640, 2024.

[30] A. Vaswani, N. M. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[32] X. Gao, Z. Xiao, and Z. Deng, "High accuracy food image classification via vision transformer with data augmentation and feature augmentation," *Journal of Food Engineering*, vol. 365, p. 111 833, 2024.

[33] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 32–42.

[34] Z. Xiao, G. Diao, and Z. Deng, "Fine grained food image recognition based on swin transformer," *Journal of Food Engineering*, vol. 380, p. 112 134, 2024.

[35] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[37] A. Panda, D. Panigrahi, S. Mitra, S. Mittal, and S. Rahimi, "Transfer learning applied to computer vision problems: Survey on current progress, limitations, and opportunities," *ArXiv*, vol. abs/2409.07736, 2024.

[38] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," *ArXiv*, vol. abs/2302.07944, 2023.

[39] E. Aguilar, B. Nagarajan, and P. Radeva, "Uncertainty-aware selecting for an ensemble of deep food recognition models," *Computers in Biology and Medicine*, vol. 146, p. 105 645, 2022.

[40] J. Teng, D. Zhand, D.-J. Lee, and Y. Chou, "Recognition of chinese food using convolutional neural network," *Multimedia tools and applications*, 2019.