ConvNeXt Fine-Tuning for Accurate Classification of 300 Cooking Ingredients

Nhan Thanh Tran[0009−0002−7200−2454], Vuong Quoc Le[0009−0005−5225−6061], Thinh Ba Le[0009−0008−9711−7593], Duc Le Tai Nguyen[0009−0000−5371−7118], Long Phi Nguyen[0009−0001−3132−0184], and Khanh Hong Vo[0009−0000−2881−0943]

Department of Information Technology, FPT University, Can Tho, Vietnam
{nhanttce171358, vuonglqce171000, thinhlbce171590, ducnltce170499, longnpce170669}@fpt.edu.vn, and khanhvh@fe.edu.vn

Abstract. The MEAL300 dataset, comprising 300 food ingredient classes, poses a significant challenge for image classification models. In this study, we explore various Convolutional Neural Network (CNN) architectures and optimization techniques to improve classification accuracy on MEAL300. We evaluate baseline CNN models and apply transfer learning with fine-tuning strategies to adapt pre-trained models to the dataset. By optimizing fine-tuning methodologies and incorporating regularization techniques, we achieve significant performance improvements. Our proposed model, a fine-tuned ConvNeXtXLarge, attains a state-of-the-art accuracy of 93.79%, outperforming other architectures such as MobileNetV3Large (85.29%) and EfficientNetV2B0 (83.56%). These findings demonstrate the effectiveness of transfer learning and fine-tuning for large-scale food ingredient classification and contribute to advancing automated food recognition systems with applications in nutrition tracking, food safety, and smart kitchen technologies.

Keywords: Food Ingredients · Image classification · Transfer learning · Fine-tuning.

1 Introduction

Food ingredient classification, a key computer vision task, supports applications like dietary assessment and automated ingredient recognition. Unlike dish-level classification, it faces challenges from high intra-class variation and visual similarities across ingredients. The MEAL300 dataset [1], with 300 ingredient classes, is a robust benchmark for this task. Convolutional Neural Networks (CNNs) excel in image recognition but require optimized architectures, data augmentation, and transfer learning for best results. This study evaluates CNN architectures and techniques on MEAL300 to identify effective configurations for ingredient classification, highlighting their strengths and limitations.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

2 Related Work

Food ingredient classification has advanced with machine learning and computer vision, moving beyond labor-intensive manual inspection. DeepFood [2], a deep-learning framework trained on the MLC-41 dataset (41 ingredient categories, 100 images each), combines ResNet features, Information Gain selection, and SMO classification to achieve 87.78% accuracy. Similarly, the Meal300 dataset, with 300 ingredient categories, supports models like the Tree Adaptation Network (TAN) [1], which uses transfer learning to improve ingredient identification and quality assessment for catering, outperforming traditional methods. Transfer learning also enhances dish recognition, as seen in a framework using EfficientNet on the UEH-VDR dataset, achieving 92.33% accuracy for Vietnamese dishes [3], with applications in culinary tourism. However, ingredient-level classification on Meal300, particularly addressing class imbalance and large-scale CNN optimization, remains underexplored.

3 Methodology

3.1 Dataset

In this study, we utilize the MEAL300 dataset, which is specifically curated for research in food ingredient recognition. It includes a diverse range of food ingredient classes (Figure 1), with the number of images per ingredient varying significantly, as shown in Figure 2.

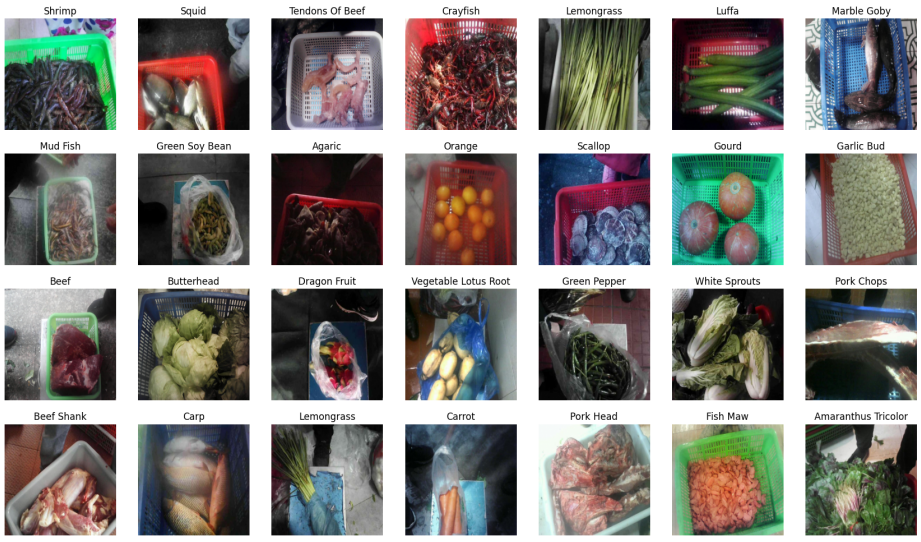

Fig. 1. A preview of the MEAL300 dataset's food ingredient classes.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

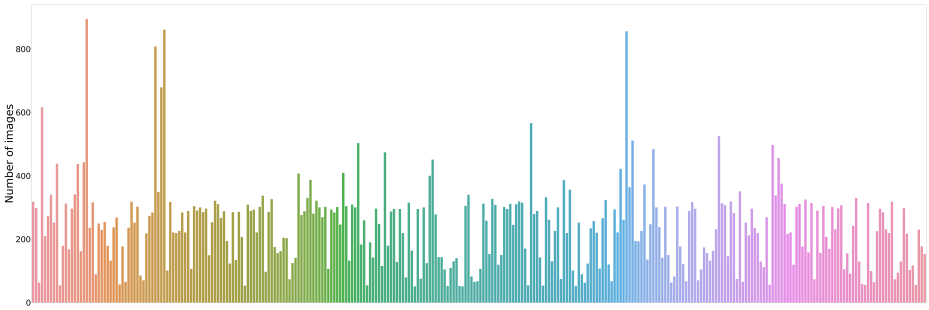

Fig. 2. Distribution of images across ingredient classes.

To ensure a balanced training dataset and reduce bias during both training and evaluation, we limited each ingredient class to a maximum of 100 images. This threshold was selected because certain classes contained as few as 50 images; allowing significantly more samples for other classes would introduce distributional imbalance, potentially leading to biased learning and unreliable performance metrics. The final class distribution after this filtering step is illustrated in Figure 3.

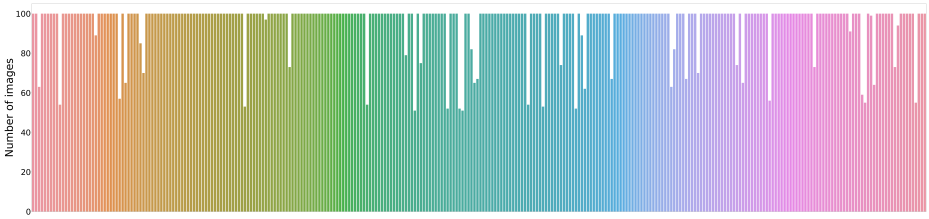

Fig. 3. Image distribution per class after selecting a maximum of 100 images per class.

To facilitate effective model evaluation, the dataset is partitioned into three subsets as follows:

- Training set (70%): Used for learning and optimizing model parameters.
- Validation set (20%): Utilized for hyperparameter tuning and model selection.
- Test set (10%): Reserved for assessing final model performance.

The MEAL300 dataset includes 300 food ingredient classes with varying image counts. To balance the dataset, we capped each class at 100 images, resulting in 20,037 training, 5,753 validation, and 2,835 test images. Images were resized to 224×224 pixels and converted to RGB for compatibility with CNN models.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

3.2 Transfer learning, fine-tuning and hyperparameter tuning

Transfer learning uses a pre-trained model on a large dataset, reusing early layers while adjusting later layers to recognize new patterns. It powers models like ResNet [4], VGG [5], EfficientNet [6], and ConvNeXt [7], supporting tasks like MEAL300 classification by reducing computing costs and improving efficiency.

Fine-tuning modifies specific layers of a pre-trained model, retaining general knowledge while learning dataset-specific patterns. It's useful when the new dataset differs from the original, refining the model without full retraining.

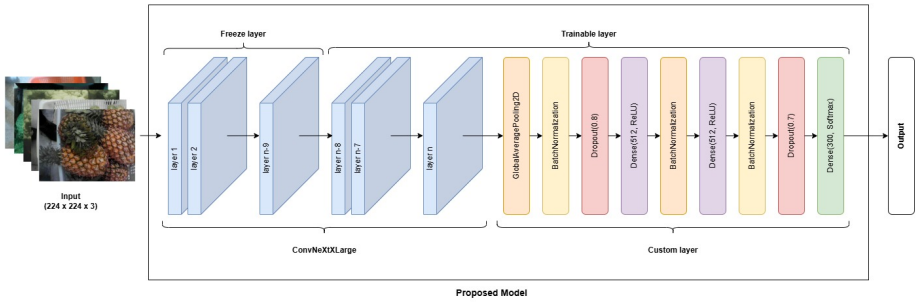

Fig. 4. Architecture of proposed model

To enhance food ingredient classification on the MEAL300 dataset, we applied transfer learning using ConvNeXt-XLarge pre-trained on ImageNet. In figure 4, instead of training from scratch, we froze all layers except the last nine, as empirical testing showed that this configuration balances adaptation to MEAL300-specific features while retaining general feature extraction capabilities from ImageNet pre-training. This finding is supported by the performance comparison in Table 1, where fine-tuning the last nine layers yields the highest accuracy. We also reduced the learning rate of Adam optimizer [8] to 3e-5 to ensure stable optimization and prevent catastrophic forgetting.

Table 1. Performance comparison between different numbers of layers at the end of the ConvNeXtXLarge that can be trained

Trainable Layers	0	1	2	3	5	7	**9**	11	13
Accuracy	0.9090	0.9160	0.9129	0.9185	0.9263	0.9326	**0.9379**	0.9333	0.9337
Precision	0.9117	0.9254	0.9175	0.9237	0.9311	0.9398	**0.9431**	0.9372	0.9377
Recall	0.9106	0.9164	0.9093	0.9208	0.9266	0.9315	**0.9405**	0.9344	0.9325
F1	0.9060	0.9138	0.9060	0.9148	0.9237	0.9301	**0.9370**	0.9309	0.9302

To improve generalization and prevent overfitting, we added Batch Normalization [9] and Dropout layers after the ConvNeXt-XLarge backbone. Specifi-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

cally, we applied Dropout(0.8) after feature extraction, followed by two Dense(512, ReLU) [10] layers with Batch Normalization, and a final Dropout(0.7) before the output layer. The classification layer consisted of 300 neurons with softmax activation.

We trained the model using categorical cross-entropy loss and the Adam optimizer with a learning rate of 3e-5. A lower learning rate was chosen to ensure stable convergence and prevent drastic weight updates, which is especially important when fine-tuning a large pre-trained model like ConvNeXt-XLarge. This approach helps retain valuable pre-trained features while allowing the model to adapt effectively to the new dataset. With this setup, our model achieved a state-of-the-art accuracy of 93.8%, outperforming other architectures.

4 Experiments and Results

Experiments were conducted on Kaggle Notebooks using an NVIDIA Tesla P100 GPU with TensorFlow/Keras. To prevent overfitting, we used early stopping with a patience of 15 epochs and evaluated performance using accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1 Score} = \frac{2TP}{2TP + FP + FN}$$

Table 2. Comparison of Models on Accuracy, Precision, Recall and F1

Model	Accuracy	Precision	Recall	F1
VGG16	0.6423	0.6521	0.6425	0.6298
ResNet50	0.7658	0.7783	0.7681	0.7587
EfficientNetV2B0	0.8356	0.8376	0.831	0.8247
MobileNetV3Large	0.8529	0.8609	0.8559	0.8497
ConvNeXtBase	0.9086	0.9131	0.9127	0.9051
ConvNeXtXLarge	0.9252	0.9295	0.9272	0.9215
Proposed Model	**0.9379**	**0.9431**	**0.9405**	**0.9370**

Our experiments evaluated multiple deep learning architectures on the MEAL300 dataset, focusing on classification accuracy, precision, recall, and F1-score. The results, summarized in Table 2, highlight the performance differences across various models.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

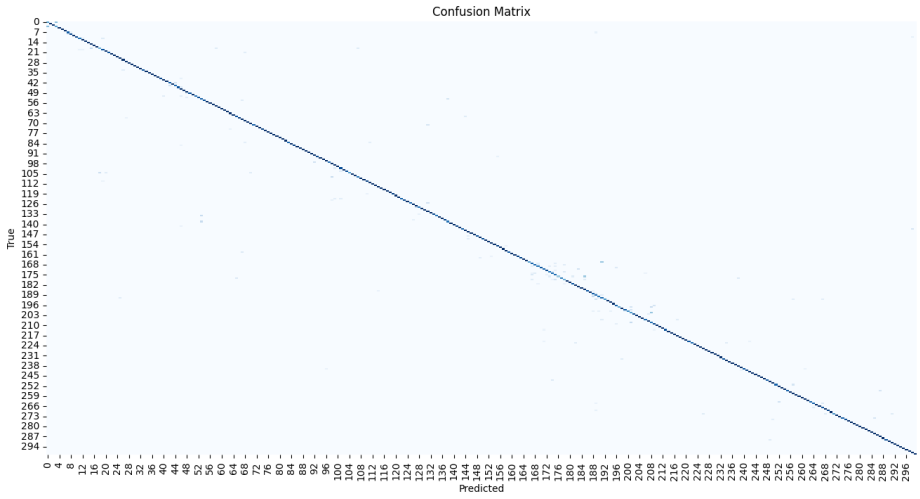

Fig. 5. The confusion matrix of the proposed model on Meal300 dataset

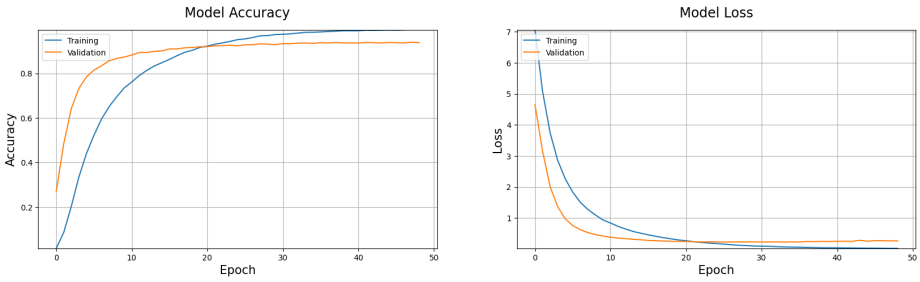

Fig. 6. Accuracy and loss graph of proposed model when training on Meal300 dataset

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

Figure 5 shows the confusion matrix, highlighting misclassifications primarily among visually similar classes. Figure 6 illustrates the training and validation accuracy/loss curves, demonstrating convergence after 20 epochs.

Our proposed model, which fine-tunes ConvNeXtXLarge, significantly outperformed all baselines, achieving a state-of-the-art accuracy of 93.79%, along with the highest precision (94.31%), recall (94.05%), and F1-score (93.70%). These improvements stem from selective fine-tuning, learning rate adjustments, and regularization techniques such as Batch Normalization and Dropout, which enhanced generalization and robustness.

The results indicate that transfer learning with ConvNeXtXLarge is highly effective for food classification tasks, with our enhancements pushing the model's performance beyond existing benchmarks. The high recall and F1-score further confirm the model's reliability in distinguishing between the 300 different food ingredient classes, making it well-suited for real-world applications in food recognition and dietary analysis.

The high accuracy and F1-score demonstrate the model's potential for real-world applications such as automated dietary tracking and food safety monitoring. However, the computational demands of ConvNeXt-XLarge may limit its use on resource-constrained devices, highlighting the need for model optimization or lightweight alternatives.

5 Conclusion and Future Work

We investigated food ingredient classification on the MEAL300 dataset with 300 classes. Using transfer learning and fine-tuning ConvNeXtXLarge, our model achieved 93.79% accuracy, outperforming MobileNetV3Large and EfficientNetV2B0. Optimized layer selection and regularization enhanced robustness. This demonstrates ConvNeXt's effectiveness for large-scale food classification, with applications in dietary assessment and nutrition analysis.

5.1 Limitations

Our model is limited by class imbalance and varying image quality in MEAL300, which may affect generalization. ConvNeXtXLarge's high computational cost hinders deployment on low-power devices, and single-label classification restricts its use for multi-ingredient images.

5.2 Future Work

Future research will explore self-supervised and contrastive learning to enhance feature extraction with fewer labeled samples, address class imbalance through data augmentation or re-weighting loss functions, develop lightweight architectures like distilled ConvNeXt or Vision Transformers (ViTs) [11] for real-time edge deployment, and expand the dataset with diverse real-world images to improve robustness in practical food recognition scenarios.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7

References

1. G. Xiao, Q. Wu, H. Chen, D. Cao, J. Guo, and Z. Gong, "A deep transfer learning solution for food material recognition using electronic scales," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2290–2300, 2019.
2. L. Pan, S. Pouyanfar, H. Chen, J. Qin, and S.-C. Chen, "Deepfood: Automatic multi-class classification of food ingredients using deep learning," in *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*. IEEE, 2017, pp. 181–189.
3. T. T. Tai, D. N. H. Thanh, and N. Q. Hung, "A dish recognition framework using transfer learning," *IEEE Access*, vol. 10, pp. 7793–7799, 2022.
4. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556
6. M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946
7. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545
8. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980
9. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167
10. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993
11. 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," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_7