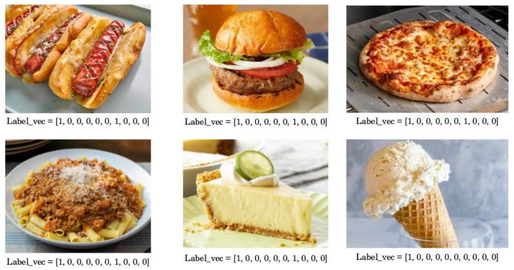
Benchmarking Deep Learning Model Architectures for Multi-Label Big-9 Allergen Classification in Food Images

Calvin J. Lomax
Luddy School of Informatics
Data Science B.S.
cjlomax@iu.edu

Introduction

Multi-label detection of the "**Big-9**" allergens (milk, egg, peanut, tree nuts, soy, wheat, fish, shellfish, sesame [7]) is a difficult vision task due to **occlusion**, **overlapping ingredients**, **mixed preparation states**, and severe **class imbalance.** These challenges are repeatedly noted in food-ingredient recognition research [1, 2].

Food Images w/ Corresponding Allergen Vectors



This study applies standard multi-label learning methods [3], imbalance-aware stratification [4], and large-scale visual-recognition practices modeled on ImageNet pipelines [5]. Deep CNN and Transformer architectures are trained on the same curated dataset to isolate architectural effects, using cross-entropy–based multi-label objectives consistent with recent analyses of loss-function behavior in deep learning [6].

Objective

The goal of this study is to determine which **model class**, either a **DCNN (viz. ResNet-18)** or a **Multi-head Latent Attention (MLA) Transformer**, offers the **strongest overall performance** across several general evaluation metrics, including **AUROC**, **mAP**, and **validation loss**.

By comparing these architectures under **identical training conditions**, the study seeks to clarify whether **traditional computer-vision frameworks** or **modern Transformer-based models** are better suited for **"Big-9" allergen recognition** and, more broadly, for **multi-label visual classification** tasks involving complex, overlapping ingredient features.

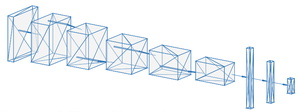
Methods

BINARY MODEL:

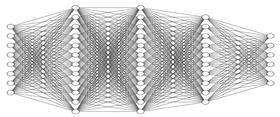
ResNet-10 trained on small subset w/ binary 'presence' annotations

MULTI-LABEL MODELS:

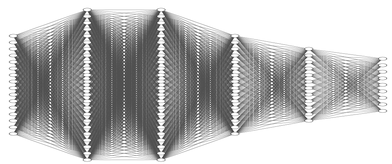
ResNet-18 (DCNN, three runs, 11,181,642 parameters)



Small MLA (Transformer, two runs, 3,429,130 parameters)



Large MLA (Transformer, three runs, 34,198,026 parameters)



DATASET & TRAINING:

EdiSet97k: a 97k 1:1 image-vector subset built from Recipe1M+ [4,8]

Environment: IU Quartz GPU-accelerated Supercomputer

Transfer learning via ImageNet w/ Loss Function [6]**:**

$$\mathcal{L}(z, y) = -\sum_{k=1}^{K} \left(y_k \log\left(\sigma(z_k)\right) + (1 - y_k) \log\left(1 - \sigma(z_k)\right) \right)$$

where:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

NOTE: Batch sizes varied from 32 to 512, Epochs from 15 to 40

Metrics: AUROC, mAP, Val. loss, Loss gap, Improvement slopes [3].

EVALUATION METRICS

- **AUROC:** Threshold-independent measure of how well the model separates positive and negative cases.
- **mAP**: Evaluates ranking quality by averaging precision across recall levels for each allergen.
- **Validation loss:** Tracks overall fit and optimization behavior, including signs of convergence or overfitting [4].
- **Train–validation gap:** Quantifies overfitting by comparing training and validation loss divergence over time.

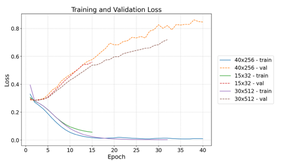
Results

GLOBAL BEST MODEL

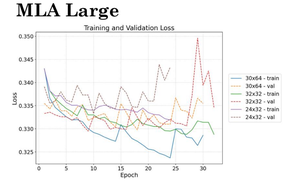
Resnet-18 w/ 40 epochs & 256 Batch Size; **AUROC:** 0.7900, **mAP:** 0.3840, **val_loss:** 0.2840, t**rain–val gap:** –0.0123

LOSS CURVES

Resnet-18

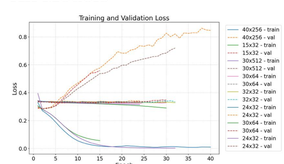


MLA Large



MLA Small

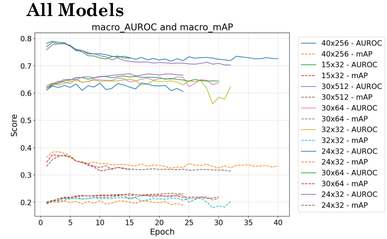
All Models



AUROC & MAP

AUROC is how well models separate positives from negatives across thresholds., while mAP is how precise true positives across recall levels. (i.e., overall separability vs. ranking quality)

All Models



- **Strongest AUROC/mAP** combination: **Resnet-18 40x256**, followed by the other two Resnet-18 models (15x32, then 30x512)
- The 'best' transformer model, surprisingly, was the **30x64 run of the small model**, which still scored **lower than all three Resnet-18 runs.**

MODEL BEHAVIOR AND LEARNING TRENDS

- **ResNet-18:** Fast early gains, early AUROC peak, mild overfitting later, overall best performance.
- **Transformers:** Stable but low AUROC/mAP, slow improvement, and weak final accuracy.

Discussion and Future Work

CONCLUSION

ResNet-18 is the strongest architecture in this study, yielding the highest AUROC, lowest validation loss, and most stable early-epoch generalization. The 60-epoch run is notable for plateauing and then improving again, suggesting a possible double-descent pattern.

Transformer models underperformed, likely due to dataset scale or limited regularization options imposed by the experimental design.

FUTURE WORK

Further investigation should focus on expanding dataset diversity, applying large-scale SSL pretraining (e.g., MAE, DINOv2), using class-balanced loss functions, and evaluating hierarchical allergen grouping.

References

[1] Gao, J., Chen, J., Fu, H., & Jiang, Y.-G. (2023). *Dynamic Mixup for Multi-Label Long-Tailed Food Ingredient Recognition.* IEEE Transactions on Multimedia, 25, 4764–4776.
[2] Chen, J., & Ngo, C.-W. (2016). *Deep-Based Ingredient Recognition for Cooking Recipe Retrieval.* ACM Multimedia (MM '16).
[3] Tsoumakas, G., & Katakis, I. (2007). *Multi-Label Classification: An Overview.* International Journal of Data Warehousing and Mining, 3, 1–13.
[4] Szymański, P., & Kajdanowicz, T. (2017). *A Network Perspective on Stratification of Multi-Label Data.* arXiv:1704.08756.
[5] Deng, J., Dong, W., Socher, R., et al. (2009). *ImageNet: A Large-Scale Hierarchical Image Database.* CVPR.
[6] Mao, A., Mohri, M., & Zhong, Y. (2023). *Cross-Entropy Loss Functions: Theoretical Analysis and Applications.* arXiv:2304.07288.
[7] Gupta, R. S., et al. (2019). *Prevalence and Severity of Food Allergies Among US Adults.* JAMA Network Open, 2(1), e185630.
[8] Marín, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., & Torralba, A. (2023). *Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 1–14.

Acknowledgements

- Jiangpeng He, PhD., AI for Food Computing Primary Investigator
- Jeffrey D. Lomax, Vice President, Product Marketing, Fizyr.AI