

Deep Learning-based Aesthetic Assessment of Food Images

Dian Zhang *

Computer Science and Technology, Beijing Jiaotong University, Beijing, 100000, China

* Corresponding Author Email: 21722118@bjtu.edu.cn

Abstract. In recent years, the significance of food aesthetics has increasingly been recognized within the food industry and on social media platforms. Accurate evaluation of the aesthetic quality of food images is crucial for enhancing consumer experience and improving marketing outcomes. This study aims to assess the performance of different deep learning models in food aesthetics evaluation to identify the most suitable model for this task. Two mainstream convolutional neural network models, including MobileNetV2 and ResNet50, were employed to score the aesthetics of food images. These models were evaluated based on loss values, mean absolute error (MAE), and aesthetic scores. The experimental results demonstrate that the MobileNetV2 model exhibits higher fitting accuracy and precision in capturing image features and predicting aesthetic scores, with its MAE and loss values both lower than those of the ResNet50 model. Furthermore, the scores generated by MobileNetV2 were closer to the reference results from a benchmark model, further validating its effectiveness in practical applications. Future research could explore the optimization of model structures using Vision Transformers (ViT), which can integrate global and local features for a more comprehensive aesthetic judgment.

Keywords: Deep learning; food aesthetic assessment; convolutional neural network.

1. Introduction

Image aesthetic assessment, a process involving computer vision and image processing techniques, quantitatively and qualitatively assesses visual elements, color harmony, contrast, and composition in images [1]. Widely applicable, this technology enables systems to automatically select optimal cropping options or apply suitable filters, thereby enhancing the visual appeal of images. Employed extensively in advertising and marketing, image aesthetic analysis assists advertisers in selecting visually appealing materials, thus increasing the attractiveness of advertisements and the effectiveness of marketing campaigns. Through evaluating the aesthetic quality of images, significant improvements in visual appeal can be achieved, benefiting a multitude of applications. Since the mid-2000s, the importance of image aesthetic analysis has significantly increased, with Datta et al.'s research laying the groundwork for subsequent developments. In 2006, a computational approach to evaluate the aesthetic quality of photographic images was proposed by Datta et al. This pioneering work demonstrated the potential of algorithmic methods for assessing visual aesthetics and established a framework for future exploration. Further advancement in the field was driven by Datta and Wang in 2010, who developed a real-time automatic image aesthetic quality inference engine, marked a breakthrough in real-time aesthetic evaluation, opening new avenues for the practical application of image aesthetic assessment [2].

Over time, image aesthetic assessment methods have evolved from mathematical models to deep learning-based approaches. Particularly important in computational aesthetics have become convolutional neural networks (CNNs), known for their outstanding performance in image processing tasks [3]. The advancement of deep learning technologies has significantly enhanced the capability to automatically predict human aesthetic judgments.

Despite advancements, the simulation of human subjective aesthetic assessment remains a challenge. This study focuses on the application of image aesthetics in the food domain, comparing different models to determine the most suitable one for assessing the aesthetics of food images. In recent years, the importance of food aesthetics has garnered increasing attention due to its significant impact on individuals' appetite and eating behavior [4]. For instance, color and plating significantly influence perceptions of a food's taste and quality [5]. Moreover, the visual presentation of food and

the ambiance in dining settings significantly affect the overall dining experience of customers [6]. Therefore, accurately evaluating the aesthetic attributes of food holds great importance for the food industry, the culinary sector, and related research fields.

Traditional methods for evaluating food aesthetics have numerous limitations. Relying on trained assessors, traditional sensory assessment methods are highly subjective and often lack consistent standards [7]. Additionally, these methods are typically costly and time-consuming, with results that may vary due to the subjective differences among evaluators [8]. Research indicates significant discrepancies between the descriptions and evaluations of trained assessors and untrained consumers for the same food samples [9]. These limitations highlight the inadequacy of traditional methods in meeting the demands of modern food aesthetic assessment. In contrast, deep learning models, with their powerful feature extraction capabilities and outstanding performance, have increasingly become a research focus in the field of food aesthetic assessment.

This study evaluates the effectiveness of different classification models in the task of food aesthetic assessment, specifically comparing MobileNetV2 and ResNet50. MobileNetV2, being a lightweight convolutional neural network, is suitable for operation on resource-constrained devices, whereas ResNet50, a more complex deep residual network, offers stronger feature extraction capabilities. By training and testing these two models on the same dataset, this study compares their performance and effectiveness in predicting food aesthetic scores.

2. Method

2.1. Dataset and Preprocessing

The Gourmet Photography Dataset (GPD) represents the first large-scale dataset designed for the aesthetic assessment of food photographs. Comprising 12,000 food images with human-annotated labels indicating positive or negative aesthetics, this dataset provides a robust foundation for computational aesthetic analysis. The images, retrieved from social media platforms, cover a wide range of food categories and geographical information. To enhance the complexity and balance of the dataset, additional images were sourced from food classification datasets. The preprocessing steps involved the removal of irrelevant samples, such as duplicate images, collages, and images with obvious artificial modifications. Further data cleaning, including the removal of unnecessary borders and rotation adjustments, resulted in a final collection of 16,400 food images [10].

Formally, the visual aesthetic assessment of food images in the GPD was classified as a binary classification problem. Aesthetic labels were assigned to the food images using Amazon's Mechanical Turk (AMT), where annotators could skip samples, they were not confident in evaluating. The dataset was randomly divided into two parts: 9,600 images for training and the remaining images for testing [10]. The dataset was derived from the Yelp Open Dataset Repository, focusing on restaurants in the Las Vegas metropolitan area, known for its high tourism activity. The data filtering steps included the removal of closed restaurants and those with fewer than 10 images, ultimately collecting 50,018 food images from 577 restaurants [11].

For this experiment, 1,200 positively labeled and 1,200 negatively labeled images were randomly selected from the GPD for training and testing purposes. Due to factors such as image clarity, the actual dataset used comprised 2,359 images. These images were randomly divided, with 1,887 used for training and 472 for testing. To ensure data diversity, food images from various cuisines and cultures were included, encompassing dishes from different countries for a comprehensive assessment. Initially, the dataset was loaded and subjected to preprocessing. The utilized image dataset included image filenames along with their corresponding aesthetic scores. To ensure consistency during model training, the following image transformation operations were applied: (1) Images were resized to 224x224 pixels. (2) Conversion of images to tensors was performed. (3) Normalization of images using mean and standard deviation values was conducted. Subsequently, the dataset was randomly divided into training and validation sets, with 80% allocated to training and 20%

to validation. Data loading was accomplished using PyTorch's DataLoader class, with the batch size for both the training and validation sets set to 32.

2.2. Model Architecture

In this study, two different model configurations are compared within the proposed neural network structure. Initially, the network architecture employed a pre-trained MobileNetV2 as the feature extractor, followed by a global average pooling layer and fully connected layers for aesthetic scoring. However, in the modified model, the feature extractor was replaced with a ResNet50-based structure instead of MobileNetV2.

The specific network structure is as follows:

Firstly, the pre-trained MobileNetV2 model was utilized as the feature extractor. By removing its original classification layer, the pre-trained MobileNetV2 model was adapted to output high-quality features suitable for aesthetic assessment. Subsequently, a global average pooling layer (GAP) was applied to reduce the feature map to a 1x1 dimension, effectively decreasing the number of parameters and enhancing computational efficiency. The feature map was then flattened into a one-dimensional vector, ready for processing by the fully connected layers.

This network comprises three fully connected layers: the first fully connected layer transforms the 1280-dimensional input into 1024 dimensions, the second fully connected layer converts the 1024-dimensional input into 512 dimensions, and the third fully connected layer reduces the 512-dimensional input to a single-dimensional output. After each fully connected layer, a ReLU activation function was applied to enhance the non-linear representation capability of the network. The final output layer provides a single aesthetic score for evaluating the input image's aesthetics.

In the modified network architecture, ResNet50 was leveraged as the feature extractor. This alteration resulted in the convolutional layers of the original MobileNetV2 being replaced by those of ResNet50. Consequently, the dimensionality of the feature map for the first fully connected layer was adjusted to 2048, given that ResNet50 typically outputs feature maps of this size. Failure to make this adjustment would have resulted in a mismatch. A significant modification in the revised model was the application of the sigmoid activation function in the output layer. This ensured that the output values were constrained within the [0, 1] range, which is particularly important for normalized outputs in regression tasks.

In summary, the modifications to the network structure encompassed changes in the selection of the feature extractor, the input dimensions of the fully connected layers, and the activation function in the output layer. These adjustments potentially impact the model's performance and its suitability for various application scenarios.

2.3. Assessment Criteria

In this study, mean absolute error (MAE) and loss were utilized as the primary metrics to evaluate the model's fitting degree and performance. The MAE and loss metrics offer a direct quantitative assessment of the model's prediction accuracy and training effectiveness, thereby facilitating a comprehensive understanding of the model's validity and reliability.

MAE represents the average of the absolute differences between predicted and actual values. The specific calculation formula is as follows:

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (1)$$

Where y_i is the predicted value, x_i is the actual value, n is the sample size. MAE directly reflects the magnitude of the prediction deviation of the model. A smaller MAE value indicates that the model's predictions are closer to the actual values, suggesting higher prediction accuracy. Conversely, a larger MAE value signifies greater prediction error, implying the need for further model adjustment or improvement.

Ideally, the MAE and Loss values for both the training and validation sets should be low and similar, indicating that the model performs well on both the training data and unseen validation data,

without severe overfitting or underfitting. If the MAE and Loss values for the training set are significantly lower than those for the validation set, overfitting may be present. This scenario implies that the model performs well on the training set but poorly on the validation set, indicating insufficient generalization ability. Conversely, high MAE and Loss values on both the training and validation sets may suggest the need for further optimization of the model structure or an increase in training data to enhance model performance.

3. Experiments and Results

3.1. Training Details

The model training process encompasses several distinct stages: (1) Data Loading and Preprocessing: Training data is loaded and subjected to preprocessing procedures. (2) Model Training: Iterative training of the model is conducted on the training data. During each epoch, model parameters are updated through backpropagation. Within each batch, both the loss value and the MAE are computed, with progress updates provided. (3) Model Validation: Model performance is assessed on the validation set to detect potential overfitting issues. (4) Model Saving: At the end of each epoch, the state of the model is saved.

Specifically, the training process of this study is outlined as follows: Twenty epochs were employed for training, with the model set to training mode at the start of each epoch. Model parameters were updated through traversal of the training set. The loss value for each batch was computed and accumulated into the running loss, while the MAE was also aggregated. Every ten batches, the current training progress was reported, including data loading time, batch processing time, remaining time, current loss, and MAE. At the conclusion of each epoch, the model's state was saved.

The experiments in this study were conducted on the Google Colab platform, with hardware featuring an NVIDIA Tesla K80 graphics processing unit. The software environment included Python 3.7, PyTorch 1.8.1, torchvision 0.9.1, and scikit-learn 0.24.2.

3.2. Experimental Result

To assess the effectiveness of different models in food aesthetics assessment, two distinct criteria were employed. Criterion 1: Scores, this work tested seven images and compared the results produced by two different models. This comparative analysis facilitates an assessment of each model's performance in accurately assessing food aesthetics, offering insights into their practical application efficacy. The seven images are presented in Fig. 1.



Fig 1. Images leveraged for testing (Figure Credits: Original).

The following Table 1 shows the result of seven images tested on different models. It could be observed that the MobileNetV2 performs better.

Table 1. Aesthetic scores of images from different models.

Image	Reference model	MobileNetV2	ResNet50
Image1	0.503099	0.561869	0.546891
Image2	0.221175	0.392540	0.164386
Image3	0.990786	0.752544	0.562732
Image4	0.990868	0.410619	0.711977
Image5	0.834099	0.000237	0.004474
Image6	0.797402	0.233445	0.823548
Image7	0.034906	0.035732	0.370385

MAE used as a quantitative measure of model accuracy, MAE evaluates the average absolute difference between predicted and actual aesthetic scores. By integrating these two criteria, the comprehensive performance of each model in the context of food aesthetics assessment is thoroughly examined. The MAE values for two models are demonstrated in Table 2.

Table 2. MAE results of different batches across models.

Batch	MobileNetV2	ResNet50
1	0.0682	0.0658
2	0.0677	0.0666
3	0.0634	0.0661
4	0.0650	0.0662
5	0.0659	0.0668

By comparing the MAE and loss values, it is evident that the first model exhibits a higher degree of fit. In summary, the MobileNetV2 model is better suited for the assessment of food image aesthetics.

3.3. Comparison on Different Learning Rates

In this study, adjustments were made to the learning rate, specifically reducing it to 0.0001. As a result, both the loss values and MAE values decreased, indicating improved model fit. Moreover, the overall aesthetic scores achieved with this modified learning rate were found to be closer to the results obtained from the baseline code. This change underscores the enhanced accuracy and alignment with benchmark outcomes achieved through the refined learning rate.

Table 3. Aesthetic scores of with learning rate 0.0001.

Image	MobileNetV2
Image1	0.531966
Image2	0.399120
Image3	0.926834
Image4	0.743591
Image5	0.752810
Image6	0.774135
Image7	0.063072

Table 4. MAE results of different batches with learning rate 0.0001.

Batch	MobileNetV2
1	0.0322
2	0.0335
3	0.0362
4	0.0375
5	0.0361

As illustrated in Table 3 and Table 4, Image3 achieved the highest score of 0.9268, which is in close proximity to the reference model's score of 0.9908 and is notably closer compared to the score obtained prior to adjusting the learning rate. Nevertheless, the scores for not all images aligned consistently with those from the reference model, indicating that there remains room for further refinement of the model.

4. Discussion

In this study, comparative experiments were conducted using MobileNetV2 and ResNet50 convolutional neural network models for food aesthetic assessment. The training process spanned 20 epochs, with metrics such as loss and MAE recorded within each epoch to assess model fitting and

prediction accuracy. Additionally, a secondary assessment criterion was employed, comparing the proximity of the highest image scores to reference results.

The results of the experiments demonstrated that MobileNetV2 exhibited superior fitting and accuracy in the task of food aesthetic assessment. Evaluating the models based on the two criteria, firstly, MobileNetV2's scores were closer to the reference results, highlighting its reliability and precision in practical applications. Secondly, in terms of Loss and MAE, MobileNetV2 outperformed ResNet50 significantly; notably, during multiple epochs, the MAE and Loss values for MobileNetV2 consistently decreased, indicating its enhanced ability to capture image features and predict aesthetic scores, thereby further validating its effectiveness and stability in aesthetic assessment tasks.

In summary, the experimental results underscore the significant advantages of the MobileNetV2 model in food aesthetic assessment. Its lower MAE and Loss values, coupled with scores that are more aligned with reference results, substantiate its effectiveness and accuracy in practical applications. Thus, MobileNetV2 is recommended as the preferred model for food aesthetic assessment. Future research may focus on optimizing the model structure to enhance performance on larger datasets and exploring the potential of other lightweight models in aesthetic assessment.

Despite the superior performance of MobileNetV2 in food aesthetic assessment as demonstrated by this study, several areas warrant further exploration and improvement. Although CNNs excel at extracting localized features, they often fall short when it comes to understanding images that require a complex global context. Traditional CNNs process images by breaking them down into smaller, local regions and then combining these local features to form an understanding of the whole image. While this approach proves effective for many tasks, it can be inadequate for tasks that necessitate an understanding of global context, such as aesthetic judgment. To address these limitations, future models can leverage the Vision Transformer (ViT) [12].

The ViT model calculates average attention weights, which reveal the spatial relationships and local features of the image, thereby allowing the model to comprehend the importance of each patch relative to others. This encoding captures the spatial structure and local details of the image. The classification token in ViT, on the other hand, provides a global representation of the entire image. The Attention Weights for the classification token aggregate information from all patches, thereby capturing global features. By combining the average attention weights and the classification token weights, the ViT model achieves a comprehensive view of the image. This dual perspective ensures that the model focuses on fine-grained local details while maintaining an overall understanding of the image's context and spatial relationships.

Subsequently, the combined local and global features extracted from the ViT model are fed into a scoring module. This module, consisting of fully connected layers, processes the integrated features to output an aesthetic score for the image.

5. Conclusion

In summary, this project successfully compared the efficacy of two models, MobileNetV2 and ResNet50, for scoring food aesthetics. Utilizing GPD, this work trained both models for 20 epochs, recording metrics such as loss and MAE. MobileNetV2 demonstrated superior performance with consistently lower MAE and loss values, indicating its higher accuracy in aesthetic prediction. To address the limitations of CNNs in capturing global context, this work proposes future integration of ViT. ViT's ability to calculate average attention weights and utilize the classification token can provide a comprehensive view, enhancing the model's capability to understand both local and global features. This dual perspective promises more accurate and reliable aesthetic assessments, highlighting the potential for ViT to further improve food aesthetics scoring models. The findings underscore MobileNetV2's current effectiveness and pave the way for advanced methodologies incorporating ViT for even greater precision in aesthetic assessments.

References

- [1] Bhattacharya Subhabrata, Rahul Sukthankar, Mubarak Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. *Proceedings of the 18th ACM international conference on Multimedia*. 2010: 271-280.
- [2] Datta Ritendra, James Z. Wang. ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics. *Proceedings of the international conference on Multimedia information retrieval*. 2010: 421-424.
- [3] Liu Tsung-Jung, Lin Yu-Chieh, Lin, Weisi, et al. Visual quality assessment: recent developments, coding applications and future trends. *APSIPA Transactions on Signal and Information Processing*, 2013, 2: e4.
- [4] Spence Charles, Okajima Katsunori, and Cheok Adrian David, et al. Eating with our eyes: From visual hunger to digital satiation. *Brain and cognition*, 2016, 110: 53-63.
- [5] Akdeniz Defne, Erdem Temeloglu. How Color-Harmony on a Food Plate Affects Consumers' Perceptions? *International Journal of Gastronomy Research*, 2022, 1(1): 16-25.
- [6] Vermeir Iris, Gudrun Roose. Visual design cues impacting food choice: A review and future research agenda. *Foods*, 2020, 9(10): 1495-1500.
- [7] Jiménez María José, Tarrega, Amparo and Fuentes, Raul, et al. Consumer perceptions, descriptive profile, and mechanical properties of a novel product with chickpea flour: Effect of ingredients. *Food Science and Technology International*, 2016, 22(6): 547-562.
- [8] Lawless Harry T., Hildegarde Heymann. *Sensory evaluation of food: principles and practices*. Springer Science & Business Media, 2010.
- [9] Sharif Mian Kamran, Butt Masood Sadiq, Sharif Hafiz Rizwan, et al. Sensory evaluation and consumer acceptability. *Handbook of food science and technology*, 2017, 10: 362-386.
- [10] Li Zhaotong, Zeru Zhang, Song Gao. Classical learning or deep learning: a study on food photo aesthetic assessment. *Multimedia Tools and Applications*, 2024, 83(12): 36469-36489.
- [11] Sheng Kekai, Dong Weiming, Huang Haibin, et al. learning to assess visual aesthetics of food images. *Computational Visual Media*, 2021, 7: 139-152.
- [12] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, 2020: 2010.11929.