Dean&Francis

ISSN 2959-6157

A Comparative Study of Inception V3 and InceptionResNetV2 for Bathroom Item Classification with and without ImageNet Pretraining

Shaozhe Chen

College of Information Science and Engineering, Hohai University, Changzhou, China
2262910122@hhu.edu.cn

Abstract:

Accurate classification of bathroom items presents notable challenges due to the objects' small sizes, high visual similarity, and frequent background clutter. The investigation focuses on the influence of convolutional neural network architecture and pretraining strategy on the performance of image classification models in such fine-grained scenarios. Two widely used architectures, InceptionV3 and InceptionResNetV2, were selected for evaluation under two training regimes: training from scratch and transfer learning via ImageNet pretraining. A curated dataset containing ten categories of common bathroom items was used for training and testing. Model performance was quantitatively assessed using overall accuracy, macro-averaged precision, recall, and F1-score, alongside qualitative analysis through confusion matrices. Experimental results demonstrate that ImageNet pretraining can significantly enhance model performance across all metrics. InceptionResNetV2 with ImageNet weights achieved the highest accuracy of 96.19%, while models trained from random initialization showed unstable convergence and poor generalization, often collapsing into predicting a dominant class. The superior performance of pretrained models is attributed to the reuse of domain-invariant features learned from large-scale datasets, which serve as effective initializations for downstream tasks with limited labeled data. These findings confirm the effectiveness of transfer learning in small-sample visual classification and highlight the additional benefit of residual connections in deeper architectures when fine-tuning on domain-specific tasks.

Keywords: Inception V3; InceptionResNetV2; ImageNet pretraining; bathroom item dataset.

1. Introduction

In daily life, the human eye can rapidly identify observed objects, whereas this remains a highly challenging task for machines. Effectively processing images with computers constitute a crucial research topic in the field of computer vision [1]. Among these efforts, image classification stands as a core task, with applications spanning autonomous driving, medical diagnostics, and smart home systems. Image classification refers to the process of effectively distinguishing images from different domains by associating them with their actual inherent characteristics. Specifically, it involves inputting an image into a computer system, which then determines the most appropriate category or label for the image from a predefined set of classes [2]. For instance, when presented with a collection of images containing animals, landscapes, and objects, a computer must classify each received image into a specific object category. Although image recognition may appear straightforward, it involves numerous challenges that require resolution. For example, the accuracy of computer-based image recognition can be affected by factors such as background clutter, lighting conditions, and rotational angles. This is particularly evident in classifying bathroom items (e.g., toothbrushes, soap bars, and towels), where small size, occlusions, and cluttered backgrounds present unique visual challenges. Accurately recognizing such items can significantly improve smart home functionalities, such as automated inventory tracking, elderly care assistance, or hygiene monitoring, making the system more responsive and user-friendly.

With the rise of deep Convolutional Neural Networks (CNNs), models such as ResNet, Inception, and EfficientNet have achieved state-of-the-art performance on large-scale datasets like ImageNet. These architectures, when combined with transfer learning, have been widely applied to downstream tasks involving limited data. In particular, the Inception family, including Inception V3 and InceptionResNetV2, balances model depth and computational efficiency, making them suitable candidates for real-world deployment. For instance, Revathi et al. utilized Inception V3 to extract deep features for medical image classification and retrieval based on similarity matching [3]. Deng et al. applied a fine-tuned Inception ResNet V2 model to accurately classify cigarette combustion cone images, achieving 97.22% accuracy and demonstrating strong robustness in real-world detection scenarios [4]. Yulita et al. employed Inception V3 as an image embedding extractor to capture visual features of garbage images, which were then classified using XGBoost, achieving strong performance in handling class imbalance [5].

Despite the proven effectiveness of CNNs on large datasets, less attention has been given to evaluating their performance on domain-specific, small-scale datasets—especially in visually similar object categories like bathroom items. Furthermore, while transfer learning using ImageNet-based weight is often used by default, few studies systematically compare its impact across different model architectures. These gaps underscore the importance of understanding the extent to which pretraining contributes to classification performance, as well as how architectural complexity influences model effectiveness in constrained data environments. To address these questions, this paper presents a comparative study between Inception V3 and InceptionResNetV2 for bathroom item classification under two conditions: with and without ImageNet pretraining. All models were trained and evaluated on a curated dataset of ten bathroom item categories. Classification performance was measured using accuracy, precision, recall, F1-score, and confusion matrices, enabling a detailed analysis of model behavior and generalization.

2. Methods

2.1 Dataset Preparation

The dataset used in this study was obtained from the Kaggle platform [6]. It consists of ten commonly seen bathroom item categories, namely: bath towel, comb, curtain, mat, mirror, sink, soap bar, toothbrush, toothpaste, and wastebasket. For each category, 800 images were selected, resulting in a total of 8,000 images. All images are in RGB format, and their sizes vary, with no uniform resolution or aspect ratio across the dataset.

As for the data preprocessing procedures, all images were rescaled by a factor of 1/255 to normalize pixel values to the [0, 1] range. The dataset was then split into training and validation sets using an 80/20 ratio. During loading, all images were resized to 299×299 pixels to match the input size required by the Inception-based architectures. A categorical label encoding was applied to support multi-class classification. The Fig. 1 shows representative images from each class in the bathroom item dataset.

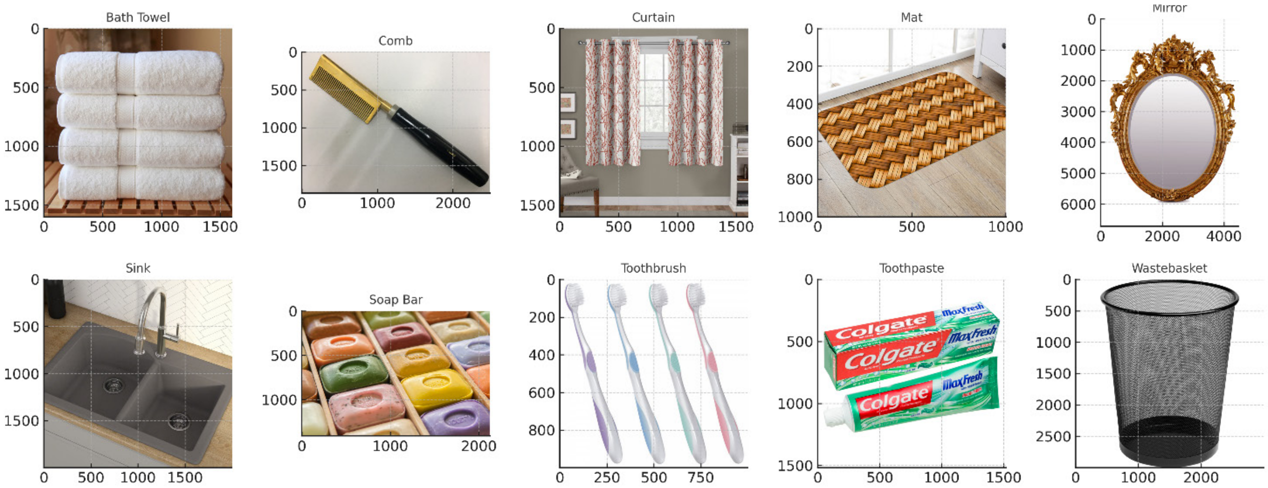


Fig. 1 Representative images from each class in the bathroom item dataset [6]

2.2 Convolutional Neurual Networks-Based Classification

2.2.1 Inception V3

Inception V3 is a convolutional neural network architecture proposed by Szegedy et al. to enhance both the computational efficiency and classification performance of deep neural networks [7]. It is part of the broader Inception family, which is characterized by multi-branch convolutional modules that allow for richer feature extraction across multiple spatial scales.

The core idea of the Inception module is to perform several convolutions with different kernel sizes (e.g., 1×1, 3×3, 5×5) in parallel, and then concatenate the outputs along the channel dimension. Inception V3 refines this structure by introducing factorization of convolutions, where larger convolutions (e.g., 5×5) are decomposed into smaller ones (e.g., two 3×3 convolutions), and asymmetric convolutions such as 1×7 followed by 7×1 to further reduce computational cost.

Another key innovation in Inception V3 is the use of auxiliary classifiers. These are small classification heads connected to intermediate layers of the network during training, which help combat the vanishing gradient problem and improve convergence. The architecture also incorporates batch normalization extensively to stabilize training, and label smoothing to improve generalization.

Compared with earlier Inception models and traditional deep CNNs, Inception V3 achieves a better trade-off between accuracy and computational demand. It significantly reduces the number of parameters and floating-point operations, making it well-suited for applications requiring a balance between speed and performance.

2.2.2 InceptionResNetV2

InceptionResNetV2 is a hybrid deep convolutional neural network architecture that combines the multi-path feature extraction of Inception modules with the identity-based residual connections of ResNet. Originally proposed by Szegedy et al. in 2016, this architecture was designed to overcome limitations in training very deep models, such as vanishing gradients and slow convergence, while preserving the high representational capacity of Inception-style designs [8].

At its core, InceptionResNetV2 incorporates modified Inception modules—such as Inception-ResNet-A, -B, and -C—where the outputs of parallel convolutional branches are concatenated and then projected through a 1×1 convolution. These projections are then added to the module's input via a residual shortcut connection. This design enables the network to benefit from both multi-scale feature extraction (a key strength of Inception) and efficient training via residual learning (a hallmark of ResNet). Unlike traditional Inception modules, which treat each block as an independent transformation, the residual connections enforce a learning dynamic that encourages refinement over radical transformation, thus stabilizing the gradient flow even in very deep architectures.

InceptionResNetV2 is deeper than earlier Inception networks, comprising over 55 million parameters. To manage this depth without sacrificing computational efficiency, the architecture utilizes factorized convolutions (e.g., 1×n followed by n×1 instead of n×n) and dimensionality reduction techniques, which significantly reduce computational load while maintaining accuracy. Additional modules such as reduction blocks (Reduction-A and Reduction-B) are introduced at key stages to downsample feature maps and expand receptive fields efficiently.

Furthermore, InceptionResNetV2 incorporates a variety of regularization and optimization enhancements. Batch normalization is used extensively to stabilize training, dropout is applied to mitigate overfitting, and label smoothing is optionally used to improve calibration of probabilistic predictions. The model is also trained using RMSProp optimizer with learning rate decay and gradient clipping to further stabilize the update dynamics. Thanks to these architectural innovations, InceptionResNetV2 achieves state-of-the-art performance on standard benchmarks like ImageNet, while being more parameter-efficient than other models with comparable depth.

2.3 Implementation Details

The models were implemented using TensorFlow. For both InceptionV3 and InceptionResNetV2, the input images were resized to 299×299 pixels, consistent with the expected input size of these architectures. During training, a batch size of 32 was used, and the number of epochs was set to 10. The Adam optimizer was selected due to its adaptive learning rate capabilities and robustness across various tasks. The learning rate was not manually specified, allowing Adam to use its default settings. The loss function was categorical crossentropy, suitable for multi-class classification problems with one-hot encoded labels. Model performance was monitored using the accuracy metric during training. After training, evaluation was conducted on the validation set using four key metrics: precision, recall, F1-score, and accuracy. Additionally, confusion matrices were generated and visualized as heatmaps to analyze class-wise performance, providing a more detailed understanding of the model's predictive behavior.

To investigate the impact of transfer learning, this paper considered two training settings: with and without ImageNet pretraining. ImageNet is a large-scale image database containing over 14 million annotated images across more than 20,000 categories, and it has served as a benchmark dataset for visual recognition tasks. Pretrained weights on ImageNet enable convolutional neural networks to leverage rich, generalized features learned from diverse visual patterns, often accelerating convergence and improving performance on downstream tasks.

In this paper, each model was initialized in two ways. In the pretrained setting, the model weights were loaded from networks previously trained on ImageNet, allowing the models to start with transferable knowledge. In the from-scratch setting, the models were initialized with random weights and trained solely on the bathroom item dataset. This contrast enabled a controlled comparison of how prior knowledge affects performance in a domain-specific, limited-data scenario. All other training configurations, including input size, optimizer, batch size, and number of epochs, were kept identical across both settings to ensure fairness.

3. Results and Discussion

3.1 Classification Performance of Different Models

To evaluate the classification capability of different deep learning models on the bathroom item dataset, this study compared four experimental configurations: InceptionV3 and InceptionResNetV2 architectures, each trained both from scratch and with ImageNet pretraining. The comparative results are illustrated in Fig. 2.

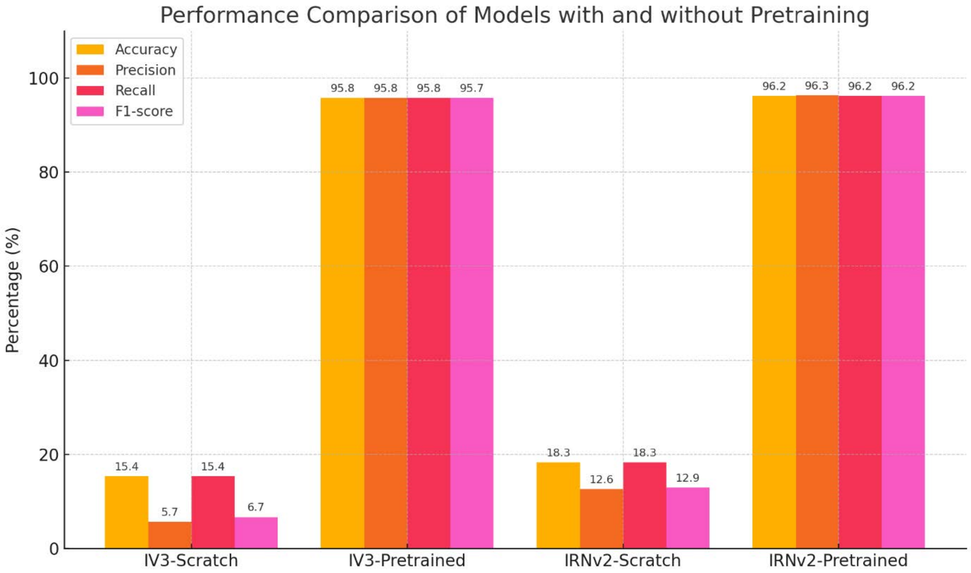


Fig. 2 Performance comparison of InceptionV3 and InceptionResNetV2 with and without pretraining (Picture credit : Original)

As shown in Fig. 2, models trained with ImageNet pretraining consistently outperform those trained from scratch by a wide margin across all metrics. The pretrained InceptionResNetV2 model achieves the best overall performance, with an accuracy of 96.19%, a precision of 96.26%, a recall of 96.19%, and an F1-score of 96.18%. The pretrained InceptionV3 model closely follows, also delivering above 95% on all metrics.

In stark contrast, both models trained from scratch demonstrate poor performance. InceptionV3 trained from scratch yields the lowest scores, with only 15.38% accuracy and 6.67% F1-score, while InceptionResNetV2 trained from scratch performs slightly better at 18.31% accuracy and 12.93% F1-score. This discrepancy clearly highlights the critical role of transfer learning when dealing with limited and domain-specific datasets.

Overall, these results demonstrate that model architecture and pretraining are both crucial factors influencing classification performance. While InceptionResNetV2 offers a marginal improvement over InceptionV3 due to its deeper residual structure, pretraining has the most substantial impact, elevating model effectiveness from nearly random guessing to highly reliable performance.

3.2 Confusion Matrix and Class-wise Performance Analysis

To further understand the behavior of each model beyond aggregate metrics, confusion matrices were generated for all four configurations. These matrices visualize the model's ability to distinguish between each of the ten bathroom item categories. The confusion matrices are shown in Figs. 3–6.

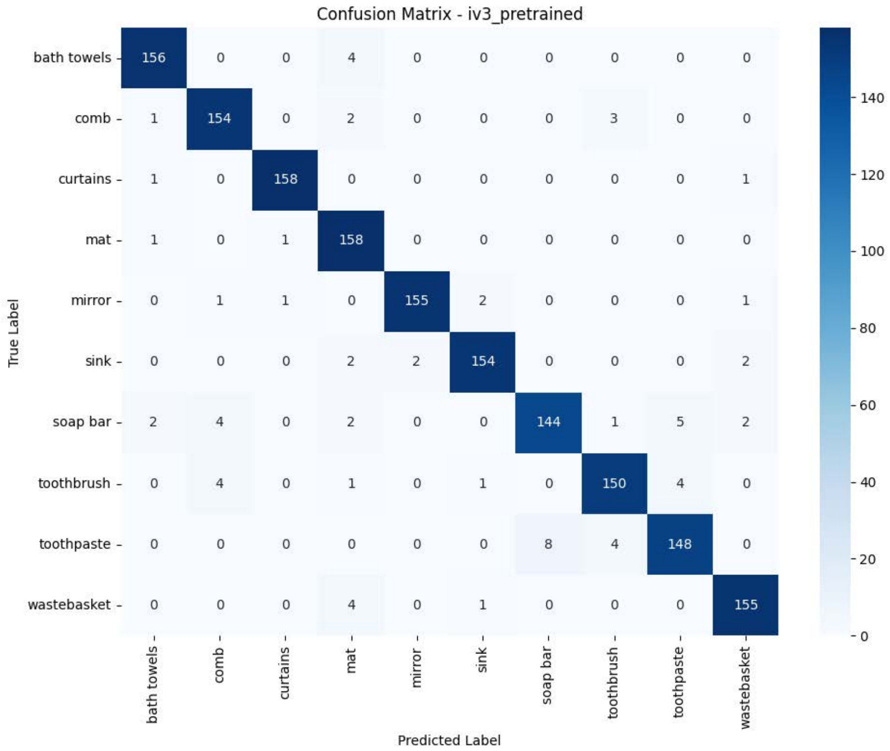


Fig. 3 Confusion Matrix – InceptionV3 (Pretrained) (Picture credit : Original)

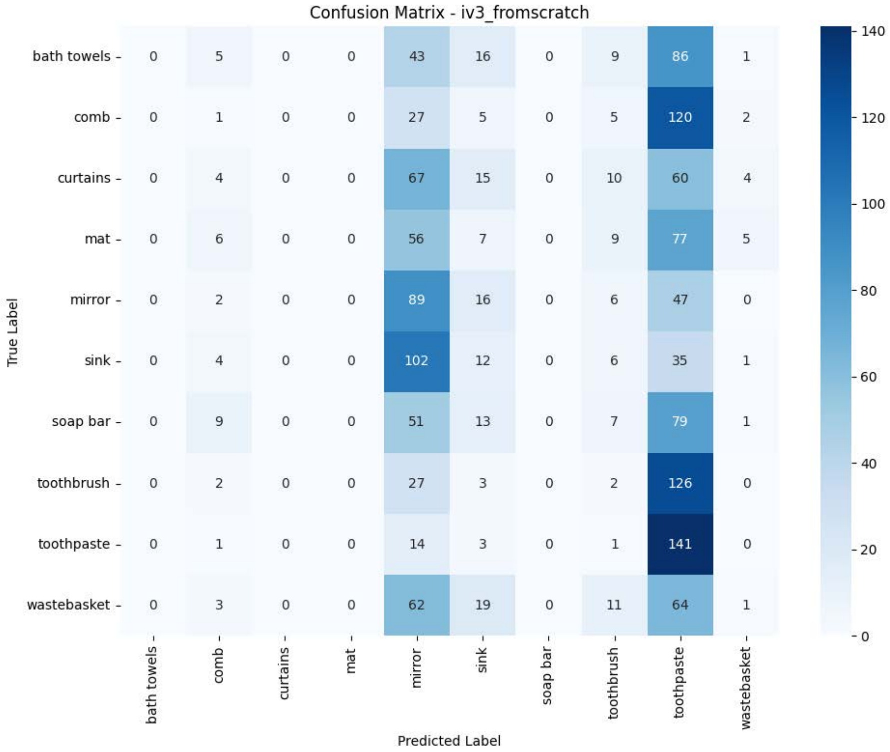


Fig. 4 Confusion matrix – InceptionV3 (Scratch) (Picture credit : Original)

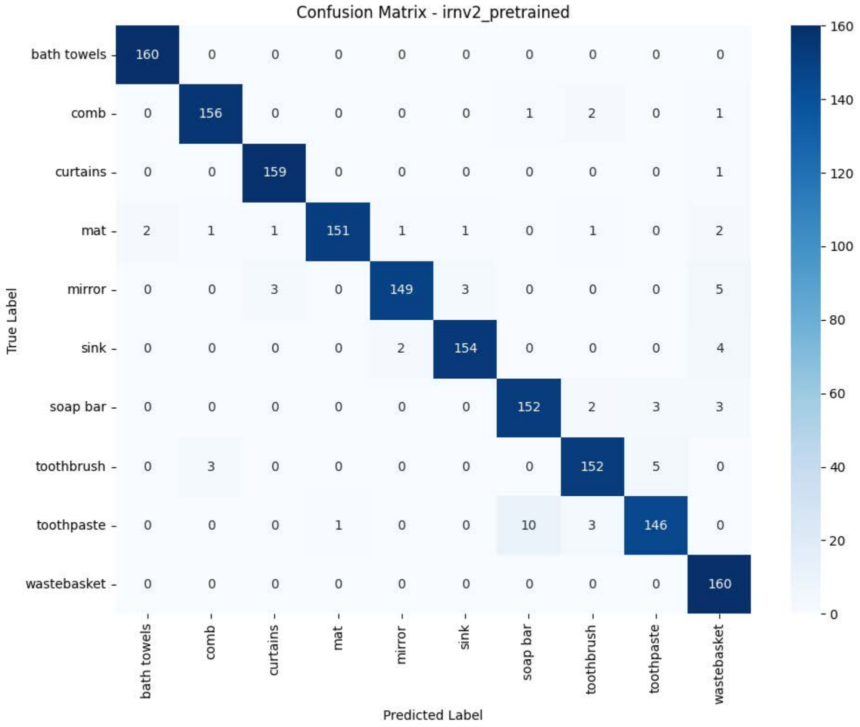
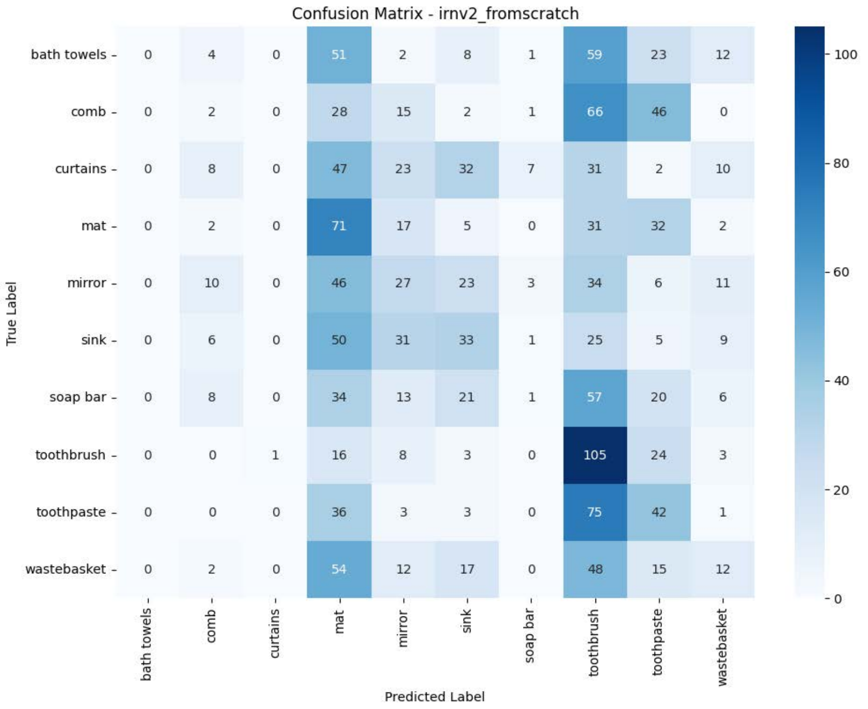


Fig. 5 Confusion matrix – InceptionResNetV2 (Pretrained) (Picture credit : Original)



True Label \ Predicted Label	bath towels	comb	curtains	mat	mirror	sink	soap bar	toothbrush	toothpaste	wastebasket
bath towels	0	4	0	51	2	8	1	59	23	12
comb	0	2	0	28	15	2	1	66	46	0
curtains	0	8	0	47	23	32	7	31	2	10
mat	0	2	0	71	17	5	0	31	32	2
mirror	0	10	0	46	27	23	3	34	6	11
sink	0	6	0	50	31	33	1	25	5	9
soap bar	0	8	0	34	13	21	1	57	20	6
toothbrush	0	0	1	16	8	3	0	105	24	3
toothpaste	0	0	0	36	3	3	0	75	42	1
wastebasket	0	2	0	54	12	17	0	48	15	12

Fig. 6 Confusion Matrix – InceptionResNetV2 (Scratch) (Picture credit : Original)

In both pretrained settings, the confusion matrices exhibit strong diagonal dominance, indicating highly accurate classification across most categories. In the pretrained InceptionResNetV2 model (Fig. 5), nearly all categories achieved perfect or near-perfect classification, with only minimal confusion observed between classes such as mirror and toothbrush or soap bar and toothpaste. Similarly, the InceptionV3 pretrained model (Fig. 3) produced highly clean matrices, though with slightly more off-diagonal noise compared to InceptionResNetV2. In contrast, the confusion matrices for the models trained from scratch (Figs. 4 and 6) are highly disordered and dispersed, lacking strong diagonal structures. The InceptionV3 scratch model (Fig. 4), for instance, frequently misclassified bath towels, soap bars, and comb, often labeling them as toothpaste, mirror, or even sink. The scratch-trained InceptionResNetV2 model (Fig. 6) performed marginally better but still failed to reliably distinguish many classes. Interestingly, even in low-performing models, some classes such as toothpaste and toothbrush retained moderately higher recall values, likely due to their distinct visual features (e.g., elongated shapes or color cues). However, items like mat, comb, and soap bar were highly prone to confusion under scratch training, reflecting the visual similarity and contextual overlap among bathroom items in cluttered scenes. These observations reaffirm the importance of transfer learning in visually dense and fine-grained classification tasks. The pretrained models not only achieve higher aggregate scores but also exhibit greater robustness across all individual categories, resulting in more reliable and generalizable performance.

3.3 Discussion

The significant performance advantage observed in models pretrained on ImageNet can be attributed to both the scale and the semantic diversity of the ImageNet dataset. With over 14 million images and 1,000 object categories, ImageNet contains many household items—such as towels, sinks, and mirrors—that visually resemble the bathroom items used in this study. As Huh et al. demonstrated, pretraining on such a large and diverse dataset enables networks to learn transferable visual features that generalize well to downstream tasks, even in different domains [9].

These pretrained models benefit from strong initialization: early convolutional layers capture low-level features like edges and textures, which are largely domain-invariant. This reduces the amount of task-specific data required to train the model effectively. In this study, this effect is clearly reflected in the faster convergence and higher accuracy of pretrained models compared to their randomly initialized counterparts. In contrast, models trained from scratch must learn both foundational and discriminative features simultaneously, which often leads to unstable optimization or prediction collapse. The slight advantage of InceptionResNetV2 over InceptionV3 may be explained by its residual connections, which help preserve gradient flow and improve training stability, particularly in deep networks. This observation aligns with the findings of He et al., who showed that residual learning enables deeper models to train more effectively without suffering from vanishing gradients [10]. Overall, these results affirm that the effectiveness of ImageNet pretraining lies in its broad visual coverage and feature transferability, which together provide a robust foundation for learning in data-scarce settings.

4. Conclusion

This paper presented a comparative study of InceptionV3 and InceptionResNetV2 models for bathroom item classification, evaluating the effect of ImageNet pretraining. Experimental results show that pretraining significantly enhances model performance, with pretrained InceptionResNetV2 achieving the highest accuracy and consistency across all metrics. In contrast, models trained from scratch exhibited poor generalization and frequent misclassification. The results confirm the importance of transfer learning in data-limited, visually fine-grained tasks. While InceptionResNetV2 offers slight performance gains over InceptionV3, it comes at the cost of increased computational demand. Future work may explore lightweight architectures and attention-based techniques to improve class discrimination and deployment efficiency.

References

[1] Hou X, Yang Y. Research and application of image classification based on convolutional neural networks. Electronic Components and Information Technology, 2022, 6(11): 93-97.

[2] Xiao Z, Wang X, Yang B, et al. Research on painting image classification based on convolutional neural networks. Journal of China University of Metrology, 2017, 28(2): 226-233.

[3] Revathi K, Kumar S. V. Development of medical image retrieval and classification using YOLOv7 segmentation and Inception V3 classifier. Proceedings of the 2024 9th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2024: 1169-1174.

[4] Deng G, et al. Image classification and detection of cigarette combustion cone based on Inception Resnet V2. Proceedings of the 2020 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, 2020: 395-399.

[5] Yulita I. N, Ardiansyah F, Sholahuddin A, Rosadi R, Trisanto A, Ramdhani M. R. Garbage classification using Inception V3 as image embedding and extreme gradient boosting. Proceedings of the 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 2024: 1394-1398.

[6] Kaggle. Common Objects in Bathroom Dataset. [EB/OL]. https://www.kaggle.com/datasets/mehantkammakomati/cob-common-objects-in-bathroom?select=sink

[7] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2818-2826.

[8] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).

[9] Huh M, Agrawal P, Efros A. A. What makes ImageNet good for transfer learning? arXiv preprint arXiv:1608.08614, 2016.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.