

An Investigation of Cross-dataset Model Generalization of Convolutional Neural Network

Zhengping Guan *

College of Art and Science, Santa Clara University, Santa Clara, California, 95053, United States

* Corresponding Author Email: ZGuan3@scu.edu

Abstract. Transfer learning has become increasingly important as a method to leverage pre-trained models on new tasks, potentially saving significant training time and computational resources. Understanding how well these models generalize to new datasets is critical when people are going to use them to solve real-world problems. This research investigates the transfer learning performance of pre-trained models. Specifically, this work evaluates whether these models trained on dataset Canadian Institute for Advanced Research (CIFAR)-10 can still perform well when transferred to another dataset Self-Taught Learning (STL)-10. Both datasets share the same classes, ensuring a meaningful comparison. The models were trained for five epochs on CIFAR-10 and subsequently evaluated on STL-10. Residual Neural Network (ResNet)18 achieved a maximum accuracy of 41.54% on STL-10, while Visual Geometry Group (VGG)16 reached up to 53.36%. These results show the moderate generalization capabilities of the models and suggest that even though transfer learning is not completely ineffective, there are challenges in achieving high performance on new datasets without further fine-tuning. This study aids in comprehending model generalization and provides insight into the potential and limitations of transfer learning in real-world applications.

Keywords: Transfer learning; model generalization; convolutional neural network.

1. Introduction

Cross-dataset model generalization is a critical challenge in the field of computer vision, particularly when models trained on one dataset are applied to a different, unseen dataset. This challenge is closely tied to the broader task of image classification, which involves categorizing images into predefined classes. Accurate image classification serves as a crucial role in diverse applications, including satellite image analysis, biometric authentication, environmental monitoring, and robotics. The ability of a model to generalize across datasets is essential for these applications to function efficiently and effectively in diverse real-world scenarios. For example, in the field of healthcare, models that can effectively generalize across different types of medical image datasets, such as MRI scans from patients with varying stages of cancer, can assist doctors in identifying hard-to-detect tumors. This capability is crucial for early detection and treatment, potentially saving lives [1].

The rise of deep learning has greatly enhanced the ability of models to generalize across different datasets in the field of image classification. In particular, convolutional neural networks (CNNs) have achieved impressive results in accurately identifying and categorizing images [2]. These networks excel at automatically extracting and learning hierarchical features from raw image data, so that they can handling complex image classification challenges. The effectiveness of deep learning in realizing model generalization has led to its widespread adoption in various domains, including security, entertainment, and industrial automation.

Although deep learning models are powerful, they still often face challenges related to generalization [3, 4]. Generalization is the model's capability to maintain high performance on previously unseen data that may differ significantly from the training set [5]. A common pitfall is overfitting, where a model achieves outstanding results on the training data but fails to replicate that success on test data or in real-world applications. This issue poses significant challenges to the broader use of deep learning models, as they may struggle to deliver consistent performance when applied to new datasets or varied environments.

Transfer learning has gained attention as an effective strategy for improving model generalization [6]. This technique involves taking a model that has been pre-trained on one dataset and fine-tuning it for use on another, related dataset. By reusing the knowledge acquired during the initial training, transfer learning minimizes the requirement for large amounts of new training data and extensive computational power for the subsequent task. This method proves especially beneficial when working with datasets that are limited in size or challenging to annotate.

Despite the potential of transfer learning, its effectiveness can vary depending on the similarity between the source and target datasets. If the datasets are significantly different, the pre-trained model may struggle to adapt, resulting in suboptimal performance. Therefore, it is crucial to evaluate how well pre-trained models generalize to new datasets with different characteristics.

This article investigates the transfer learning performance of two widely used pre-trained models, Residual Neural Network (ResNet)18 and Visual Geometry Group (VGG)16 [7,8]. Specifically, this work assesses whether these models, initially trained on the Canadian Institute for Advanced Research (CIFAR)-10 dataset [9], can maintain their performance when transferred to the Self-Taught Learning (STL)-10 dataset [10]. Both datasets consist of the same 10 classes, ensuring a meaningful comparison. By analyzing the training loss on CIFAR-10 and the evaluation accuracy on STL-10, this work aims to understand the generalization capabilities of these models and the effectiveness of transfer learning across datasets with different resolutions and sample sizes.

This study contributes to the ongoing research in computer vision by exploring the challenges and possibilities of model generalization across different datasets. By focusing on the performance of models trained on CIFAR-10 and evaluated on STL-10, this work offers practical insights into how well pre-trained models can adapt to new data environments. The findings are intended to guide practitioners and researchers in refining their approaches to model selection and fine-tuning, thereby enhancing the overall effectiveness and reliability of image classification systems in varied real-world scenarios.

2. Method

2.1. Dataset

The datasets used in this study are CIFAR-10 and STL-10. CIFAR-10 consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class [9]. STL-10, on the other hand, consists of 96x96 color images in 10 classes, with 5,000 training images and 8,000 test images [10]. The class mapping between CIFAR-10 and STL-10 was adjusted to exclude classes that don't have direct equivalents.

2.2. Neural Network

This work utilized two pre-trained models: ResNet18 and VGG16.

ResNet18 is a member of the ResNet family, which tackled the challenge of vanishing gradients in deep networks by introducing residual learning [7]. This approach enables the network to learn identity functions, making the training of deeper models more feasible. Comprising 18 layers, ResNet18 includes convolutional layers, batch normalization layers, and crucially, residual blocks that help maintain effective gradient flow throughout the network. These residual blocks are key to its ability to capture complex features without performance degradation, contributing to the model's well-established reputation for robustness and high accuracy in image classification tasks.

VGG16 is another widely recognized convolutional neural network, known for its straightforward yet powerful design. The network is structured with 16 layers—13 convolutional and 3 fully connected—using small 3x3 filters across all convolutional layers. To manage spatial dimensions, max-pooling layers are strategically applied throughout the architecture. Despite its simplicity, VGG16 has demonstrated exceptional performance in numerous image classification benchmarks, thanks to its consistent use of convolution and pooling layers. This uniform design not only simplifies the network but also enhances its adaptability to a variety of tasks. The depth and structure of VGG16

enable it to effectively capture detailed patterns in images, which is a major factor in its high accuracy across different classification challenges [8].

2.3. Evaluation Indexes

The primary evaluation metric used in this study is accuracy. Additional metrics such as F1-score, precision, and recall were considered for a comprehensive evaluation but were not calculated due to the scope of this study. Future work may include these metrics for a more detailed analysis.

3. Experiments and Results

The models were trained on CIFAR-10 for five epochs and evaluated on STL-10. During each epoch, both the training loss and accuracy on CIFAR-10, as well as the evaluation loss and accuracy on STL-10, were recorded. The results are summarized in Tables 1 and 2.

Table 1. Generalization result comparison cross dataset on ResNet18.

Epoch	ResNet18 Loss on CIFAR-10	ResNet18 Accuracy on CIFAR-10	ResNet18 Loss on STL-10	ResNet18 Accuracy on STL-10
1	0.4584	84.418	3.3615	46.0833
2	0.1951	93.252	3.8657	39.7500
3	0.1048	96.374	4.0997	41.9306
4	0.0587	97.998	4.3163	44.4306
5	0.0380	98.720	4.6618	43.3056

Table 2. Generalization result comparison cross dataset on VGG16.

Epoch	VGG16 Loss on CIFAR-10	VGG16 Accuracy on CIFAR-10	VGG16 Loss on STL-10	VGG16 Accuracy on STL-10
1	0.3940	0.3867	3.7890	48.7083
2	0.1735	0.1741	4.4927	46.2500
3	0.1004	0.1014	5.0100	47.9583
4	0.0589	0.0575	5.6535	51.4722
5	0.0404	0.0370	5.9257	50.0833

These results highlight several key observations: Firstly, both ResNet18 and VGG16 showed a significant reduction in training loss over the epochs, indicating effective learning on CIFAR-10. VGG16 consistently achieved slightly higher accuracy on CIFAR-10 compared to ResNet18. Secondly, when evaluated on STL-10, VGG16 outperformed ResNet18, achieving a maximum accuracy of 53.36% compared to 41.54% for ResNet18. This suggests that VGG16's architecture may be more suitable for generalizing to higher-resolution images with fewer training samples. Thirdly, the moderate accuracy on STL-10 demonstrates the challenges of transfer learning. Despite high accuracy on CIFAR-10, both models faced difficulties adapting to the new dataset, highlighting the need for additional fine-tuning or data augmentation techniques to improve performance.

4. Discussion

Firstly, the observed performance gap between training on CIFAR-10 and evaluating on STL-10 is not surprising, given the inherent differences between the two datasets. CIFAR-10 images are 32x32 pixels, whereas STL-10 images are 96x96 pixels. Despite both datasets containing the same classes, the resolution difference requires the models to adapt to higher detail levels in STL-10 images. This challenge is intrinsic to the task and demonstrates that models need to handle variations in image resolution effectively.

Secondly, the STL-10 dataset, although sharing the same classes as CIFAR-10, has fewer training samples. This smaller dataset size can impact the model's ability to generalize well, as it has less data

to learn from. However, this limitation also highlights the importance of having robust models that can perform well even with limited training data. The performance of VGG16, which reached over 50% accuracy, suggests that certain architectures might be more adept at generalizing under these conditions.

Thirdly, the moderate performance suggests that the pre-trained models capture some universal features that are transferable between datasets with similar classes. However, the specific class representations and distributions in STL-10 might differ from those in CIFAR-10, which could affect performance. This aspect underscores the complexity of real-world image classification tasks, where datasets can vary widely in terms of visual characteristics and distributions.

To improve the performance of transfer learning models, several strategies can be considered: (1) Fine-Tuning: Adjusting the pre-trained models specifically for the target dataset (STL-10) can help the models align more closely with the new data. This involves retraining the model on the target dataset with a reduced learning rate, allowing it to better capture the unique features of the new data. (2) Data Augmentation: Implementing data augmentation methods to artificially expand and diversify the training data can enhance the model's generalization capabilities. Techniques like random cropping, rotation, and color adjustments can increase the model's resilience to variations within the input data. (3) Transfer Learning with Intermediate Datasets: Introducing an intermediate dataset that bridges the gap between CIFAR-10 and STL-10 can help. Training the model on an intermediate dataset with characteristics similar to both CIFAR-10 and STL-10 can improve the transition. (4) Advanced Architectures: Exploring advanced model architectures that are specifically designed to handle variations in image resolution and dataset sizes can be beneficial. Models like EfficientNet, which scale depth, width, and resolution, might offer better performance.

Overall, while the experiment demonstrates the potential of transfer learning, it also highlights the inherent challenges and areas for improvement. Future research should focus on developing techniques to enhance the robustness and adaptability of pre-trained models, ensuring they perform effectively across diverse and varied datasets.

5. Conclusion

This study highlights the capabilities and limitations of transfer learning in image classification. By evaluating the performance of ResNet18 and VGG16 models trained on CIFAR-10 and transferred to STL-10, this work observed that while transfer learning enables some level of generalization, there are significant challenges in achieving high accuracy on new datasets. The results underscore the need for further research into techniques to enhance pre-trained models' robustness and adaptability in diverse data environments. Future work should focus on fine-tuning strategies, data augmentation, and exploring advanced model architectures to improve transfer learning performance. Additionally, the inclusion of intermediate datasets that bridge the gap between source and target datasets could provide a smoother transition and better results. This research is helpful to accumulate computer vision knowledge and offers practical insights for developing more effective and adaptable machine learning models. Understanding these aspects is crucial for deploying reliable and efficient image classification systems in real-world applications.

References

- [1] Rawat Waseem, Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 2017, 29(9): 2352-2449.
- [2] Li Zewen, Liu Fan, Yang Wenjie, et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021, 33(12): 6999-7019.
- [3] Wang Haohan, Wu Xindi, Huang Zeyi, et al. High-frequency component helps explain the generalization of convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, 8684-8694.

- [4] Kabkab Maya, Emily Hand, Rama Chellappa. On the size of convolutional neural networks and generalization performance. International Conference on Pattern Recognition, 2016: 3572-3577.
- [5] Lin Shan, Jingwei Zhang. Generalization bounds for convolutional neural networks. arXiv preprint, 2019: 1910.01487.
- [6] Zhuang Fuzhen, Qi Zhiyuan, Duan Keyu, et al. A comprehensive survey on transfer learning. Proceedings of the IEEE, 2020, 109(1): 43-76.
- [7] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [8] Simonyan Karen, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014: 1409.1556.
- [9] CIFAR-10 dataset. URL: <https://www.cs.toronto.edu/~kriz/cifar.html>. Last Accessed: 2024/08/08.
- [10] Coates Adam, Andrew Ng, Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 215-223.