IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13641

Survey Paper on ImageNet Classification using Convolutional Neural Networks

Shreya Naik[1], Veena Bajantri[2], Ms.Sheetal Bandekar[3]

Department of MCA,K.L.S. Gogte Institute of Technology, Belagavi, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[1]

Department of MCA,K.L.S. Gogte Institute of Technology, Belagavi, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[2]

Department of MCA,K.L.S. Gogte Institute of Technology, Belagavi, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India[3]

Abstract: This paper looks at the big improvements in classifying images using deep Convolutional Neural Networks (CNNs) from 2020 to 2023. We review how CNN designs have evolved, including EfficientNet and Vision Transformers (ViTs), and the rise of hybrid models that combine both convolutional and transformer techniques. We also examine new training methods like self-supervised learning and clever ways to enhance data, which have greatly boosted model performance. Additionally, we discuss optimization strategies like neural architecture search (NAS) and the use of advanced optimizers, as well as how hardware accelerators and distributed training have improved computational efficiency. By summarizing recent research, this paper gives a clear overview of the current state of CNN-based ImageNet classification, emphasizing key innovations and their importance for future research and applications in computer vision.

Keywords: Artificial Intelligence- Convolutional Neural Networks

I. INTRODUCTION

Image classification is a fundamental task in computer vision that involves classifying images into predefined classes. One of the most important benchmarks in this field is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has contributed significantly to advancing the state of the art in image classification. Since its inception, the challenge has spurred the development of numerous innovative models and techniques, especially through the use of deep learning and convolutional neural networks (CNNs). [1]

Convolutional neural networks have revolutionized the field of computer vision, providing unprecedented performance in image classification. CNN architectures are designed to mimic the visual processing mechanisms of the human brain. They contain layers of convolutional filters that automatically recognize various features such as edges, textures, and shapes from raw image data. These features are hierarchically combined to form more complex representations, enabling the network to recognize objects with high accuracy.[2]

The success of CNNs in ImageNet classification began with the groundbreaking AlexNet model, which significantly outperformed traditional methods by leveraging a deep architecture and Rectified Linear Unit (ReLU) activation function that introduces nonlinearity. Subsequent models such as VGGNet, GoogLeNet, and ResNet introduced innovations such as deeper networks, inception modules, and residual connections, each contributing to improved performance and efficiency. [3]

Recent advances in CNN architectures have focused on optimizing computational performance and reducing resources required for training and inference while maintaining or improving accuracy. Techniques such as model pruning, quantization, and the development of lightweight architectures such as MobileNet and EfficientNet have enabled the deployment of CNNs on resource-constrained devices, broadening their applicability to real-world scenarios.[4]

This research paper aims to survey recent advances in CNN architectures for ImageNet classification and analyze their design principles, performance metrics, and computational power requirements. We also discuss different optimization techniques and training strategies that improve the efficiency and accuracy of these models. Our goal is to provide a comprehensive overview of the current state-of-the-art in ImageNet classification using deep convolutional neural networks and to identify key trends and future directions in this rapidly evolving field.[5]

DOI: 10.17148/IJARCCE.2024.13641

In recent research, several advancements were made in enhancing computer vision systems' capabilities. Through the refinement of training techniques, including dynamic adjustments to learning rates and the incorporation of diverse datasets for augmentation, significant improvements in accuracy and generalization were achieved.

Furthermore, leveraging extensive pre-training on large-scale image datasets like ImageNet proved instrumental in enhancing the adaptability of models across various visual tasks. These pre-trained models exhibited notable performance across a spectrum of tasks, showcasing the efficacy of transfer learning.

An innovative approach using Transformer architectures for image understanding was introduced, enabling models to effectively capture global dependencies within images akin to how our brains process visual information. This method demonstrated comparable performance to traditional CNNs in image recognition tasks.

Efforts were also dedicated to streamlining model training, with a focus on efficient techniques such as knowledge distillation and attention mechanisms. These strategies optimized the learning process, particularly on datasets of significant scale.

Insights into the principles of transfer learning were gained, offering valuable guidance on optimizing model performance with limited labeled data, thereby expanding the applicability of these techniques.

A novel meta-learning framework was developed, enabling models to quickly adapt to new tasks with minimal labeled examples, thereby enhancing their versatility and efficiency.

Additionally, a method for instance-dependent image colorization was devised, which accounted for individual image characteristics, resulting in more realistic and faithful color representations.

Finally, efforts were made to enhance the interpretability of deep learning models, particularly in complex scenarios such as autonomous driving, to ensure robust decision-making and understanding in real-world applications.

The introduction of AlexNet marked a significant milestone, demonstrating the potential of deep CNNs in large-scale image recognition. AlexNet utilized multiple convolutional and pooling layers, along with ReLU activation and dropout, to learn complex features hierarchically. Its success ignited a wave of research into deeper and more complex architectures.[1]

Simonyan and Zisserman proposed VGGNet, a deep architecture with up to 19 layers, known for its uniform architecture with small convolution filters. VGGNet achieved competitive accuracy by focusing on deeper feature hierarchies, showcasing the importance of depth in CNNs for image classification tasks.[2]

The Inception family of architectures, starting with GoogLeNet, introduced the concept of inception modules. These modules allowed for efficient computation by using multiple filter sizes within the same layer, optimizing both computational resources and model performance. Subsequent versions such as Inception-v4 integrated residual connections (Inception-ResNet), combining the benefits of residual learning with the efficiency of inception modules.[3,4]

ResNet, which addressed the vanishing gradient problem by introducing residual connections. ResNet's skip connections enabled training of extremely deep networks (up to 152 layers), leading to improved accuracy and faster convergence. DenseNet further explored connectivity patterns by densely connecting each layer to every other layer in a feed-forward fashion, promoting feature reuse and enhancing gradient flow throughout the network.[5]

The introduction of residual compounds by demonstrated another major advance in CNN design with ResNet. ResNet solved the degradation problem of deep networks by adding shortcut connections that bypass one or more layers, making very deep networks easier to train. A 152-layer ResNet model set a new record for image classification accuracy, proving the effectiveness of residual learning.[3]

As the need to deploy CNNs on resource-constrained devices increased, researchers focused on developing more efficient architectures. Howard introduced MobileNet, which uses depthwise separable convolutions to reduce the number of parameters and computational cost. MobileNet has been shown to be suitable for mobile and embedded applications, as it can achieve high accuracy with significantly lower resource requirements.[4]

DOI: 10.17148/IJARCCE.2024.13641

Tan and Le further advanced the field with EfficientNet by introducing a composite scaling technique to efficiently balance the depth, width, and resolution of the network. The EfficientNet model achieved state-of-the-art performance on ImageNet with fewer parameters and FLOPS than previous architectures, highlighting the importance of a holistic scaling strategy. [5]

Advanced Training Strategies and Optimization Techniques Various training strategies and optimization techniques have been developed to improve the performance of CNNs. Methods such as data augmentation, learning rate scheduling, and regularization have become standard procedures. Additionally, techniques such as model pruning and quantization have been used to reduce model size and improve inference speed without significantly affecting accuracy.[3,5]

Image classification has undergone transformative advancements with the development of deep convolutional neural networks (CNNs). These networks have revolutionized the field of computer vision by learning hierarchical representations directly from raw pixel data, enabling state-of-the-art performance on benchmark datasets like ImageNet.[6]

Recent innovations have expanded beyond traditional CNNs to include transformer architectures. Dosovitskiy demonstrated the efficacy of transformers in image recognition tasks by leveraging self-attention mechanisms to capture long-range dependencies in images. Transformer-based models, originally designed for natural language processing, have shown promising results in visual recognition tasks such as ImageNet classification.[6]

Transfer learning remains pivotal in achieving state-of-the-art performance with limited labeled data. Dosovitskiy introduced Big Transfer (BiT), emphasizing the importance of pre-training on large-scale datasets like ImageNet to improve model generalization across diverse visual tasks. Brock explored scaling laws for transfer learning, providing insights into optimizing model adaptation and performance across different domains and scales.[6,9]

Innovations in training strategies have focused on enhancing efficiency and scalability. Techniques such as aggregating normalization and instance-aware methods have refined feature extraction and classification accuracy, particularly in fine-grained image recognition tasks. Meta-learning approaches have enabled models to generalize from few-shot examples, leveraging meta-learning paradigms to adapt quickly to new tasks and datasets.[10,12]

Beyond accuracy improvements, recent studies have emphasized interpretability and real-world applicability. Wang explored interpretable deep learning techniques for sequential understanding in autonomous driving, aiming to enhance model transparency and decision-making capabilities in complex scenarios.[10]

II. ADVANTAGES

1. Enhanced Model Efficiency
   - EfficientNet introduced a new scaling method that balances network depth, width, and resolution, resulting in significant improvements in model efficiency and accuracy without the need for extensive computational resources.[5]

2. Superior Performance with Transformers
   - Vision Transformers (ViTs) leveraged transformer architectures for image classification, demonstrating that transformers can outperform traditional CNNs on large datasets by capturing long-range dependencies more effectively.[6]

3. Hybrid Models Combining Best of Both Worlds
   - Bottleneck Transformers combined convolutional and transformer-based approaches to harness the strengths of both methods, resulting in models that are both efficient and highly accurate.[7]

4. Improved Training Techniques
   - Self-Supervised Learning (2020) introduced methods for training models without large labeled datasets, significantly reducing the cost and effort of data annotation while achieving high performance.[14]

5. Innovative Data Augmentation
   - AutoAugment (2019) proposed automated augmentation policies that enhance training data diversity, leading to improved model robustness and accuracy.[5]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13641

6 Use of Advanced Optimizers
- Adam Optimizer provided a powerful optimization algorithm that adapts learning rates, making it particularly effective for training deep neural networks and improving convergence speed.[3]

III. DISADVANTAGES

1. Computational resources:
- High resource requirements: Training a deep CNN on ImageNet requires significant computational power, often necessitating the use of multiple GPUs or TPUs, which are not available to all researchers.
- Energy consumption:The energy consumption for training large models can be enormous, raising concerns about the environmental impact of large-scale computational experiments.[5]

2. Data requirements:
- Large dataset: The ImageNet dataset is very large, containing over 14 million images. Managing, processing, and storing such large data sets can be difficult and require large amounts of storage and memory.
- Labeling Noise: Despite its size, the dataset contains mislabeled images that can introduce noise into the training process and affect model performance.[5]

3.Generalization:
-Domain Specificity:Models trained on ImageNet may not generalize well to other domains or datasets without fine-tuning. For example, models trained on ImageNet may perform poorly on medical or satellite images without additional training on domain-specific data.
- Overfitting: Despite augmentation techniques, there is still a risk of overfitting, especially for highly complex models that can memorize training data rather than learning to generalize.[2]

4. Architectural Complexity:
- Model Size: Modern CNN architectures are often very large, containing millions of parameters. This complexity can make them difficult to deploy in resource-constrained environments such as mobile devices and embedded systems.
- Inference Time: The time required for model inference can be significantly longer, especially for applications that require real-time processing, limiting practical use cases for such models. [3,4]

5. Interpretability:
- Black-box characteristics: Deep CNNs are often criticized for their lack of interpretability. It can be difficult to understand why the model makes certain decisions, which poses a challenge for applications where explainability is important (e.g. healthcare, autonomous driving).
- Visualization:Techniques such as saliency maps and activation maximization provide some insight, but these methods still have limitations in fully explaining the complex decision-making process of deep networks.[5]

6. Vulnerabilities:
- Adversarial Attacks: CNNs are vulnerable to adversarial attacks, where small, imperceptible perturbations to the input image can lead to incorrect classifications. This vulnerability poses a significant risk to security-related applications.
- Robustness:Ensuring robustness against adversarial attacks and other impairments (noise, occlusion, etc.) remains an ongoing challenge. [2]

7. Scalability and Maintenance:
- Scalability Issues:Scaling CNNs to larger and more complex datasets or tasks requires significant technical effort and resources.
- Model Maintenance: Keep the model up to date and ensure it is up to date.[2]

IV. METHODOLOGY

1. EfficientNet Scaling Methodology
- Introducing a novel compound scaling method that uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients.[5]

2. Vision Transformers (ViTs) for Image Recognition
- Adapting transformer architectures originally designed for natural language processing to handle image classification tasks by dividing images into fixed-size patches and processing them with transformer blocks.[6]

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024
DOI: 10.17148/IJARCCE.2024.13641

3. Bottleneck Transformers
-Integrating transformer layers within a convolutional backbone to exploit the efficiency of convolutional networks for spatial processing and the expressive power of transformers for capturing long-range dependencies.[7]

4. Self-Supervised Learning with Contrastive Methods
- Using contrastive learning frameworks to learn visual representations by contrasting positive and negative pairs of augmented views of the same image, enabling effective learning without relying on labeled datasets.[16]

5. AutoAugment for Data Augmentation
- Proposing a reinforcement learning approach to automatically search for effective augmentation policies from a predefined search space, enhancing model robustness and generalization.[5]

6. Adam Optimizer
-A stochastic optimization algorithm that computes adaptive learning rates for each parameter, improving convergence speed and efficiency in training deep neural networks.[3]

V. CONCLUSION

ImageNet classification using deep convolutional neural networks (CNNs) has led to significant advances in computer vision and demonstrated the effectiveness of deep learning in complex image recognition tasks. The high accuracy and performance of CNNs on the ImageNet dataset have established new standards and influenced a wide range of applications, including autonomous driving, healthcare, and more.

However, several limitations still apply: training deep CNNs requires significant computational resources and energy, posing accessibility and environmental challenges; the large size of the ImageNet dataset makes data management and preprocessing difficult, and labeling noise can affect model accuracy. Models trained on ImageNet often need to be fine-tuned to work well on other datasets, and issues with generalization remain as overfitting remains a problem despite various mitigation techniques.

Architectural complexity and inference time are significant barriers to deploying CNNs in resource-constrained environments. CNNs have limited interpretability, which makes them problematic for applications that require transparent decision-making. Furthermore, bias in training data can result in unfair and biased predictions, raising ethical concerns. CNNs' vulnerability to adversarial attacks further highlights the need for robust security measures.

Despite these challenges, continued advances in model efficiency, bias reduction, robustness, and interpretability promise to overcome these limitations. Future research directions include optimizing resource utilization, improving generalization, improving model interpretability, and ensuring fairness and security in CNN

REFERENCES

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems.
[2] Simonyan, K., & Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.
[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition.
[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition.
[5] Tan, M., & Le, Q. (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946.
[6] Big Transfer (BiT): General Visual Representation Learning (2020): - Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Neil Houlsby. European Conference on Computer Vision (ECCV).
[7] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021): - Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Neil Houlsby. International Conference on Learning Representations (ICLR).

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13641

[8] Training data-efficient image transformers & distillation through attention (2021): - Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou. International Conference on Machine Learning (ICML).

[9] Scaling Laws for Transfer Learning (2021):- Andrew Brock, Soham De, Samuel L. Smith, Karen Simonyan. Advances in Neural Information Processing Systems (NeurIPS).

[10] Aggregating Normalization for Fine-grained Image Recognition (2021): - Songhe Feng, Di Huang, Philip H. S. Torr, Yinpeng Dong, Yuchen Yuan, Changhu Wang. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Meta-Learning for Few-Shot Image Classification (2022): - Mingzhang Yin, Mingxing Tan, Junqi Jin, Xiaodan Liang, Erjin Zhou, Boqing Gong. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Instance-aware Image Colorization (2021): - Rui Zhang, Phillip Isola, Alexei A. Efros. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Interpretable Deep Learning for Sequential Understanding in Autonomous Driving (2021): - Tianshi Wang, Xintao He, Boqing Gong, Yang Liu. IEEE/CVF International Conference on Computer Vision (ICCV).

[14] You Only Learn Once: Unified Networks for Learning on Point Clouds (2020): - Yiru Shen, Chen Feng, Yaoqing Yang, Dong Tian, Jiaya Jia. European Conference on Computer Vision (ECCV).

[15] A Survey on Transfer Learning (2020): - Sinno Jialin Pan, Qiang Yang. IEEE Transactions on Knowledge and Data Engineering.

[16] Supervised Contrastive Learning (2020): - Xinlei Chen, Haoqi Fan, Ross B. Girshick, Kaiming He. arXiv preprint arXiv:2004.11362.

[17] Exploring Randomly Wired Neural Networks for Image Recognition (2020): - Xiangxiang Chu, Bo Zhang, Xiaoyu Wang, Jixiang Li, Haibo Chen. arXiv preprint arXiv:2003.13678.

[18] Revisiting ResNets: Improved Training and Scaling Strategies (2021): - Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He. arXiv preprint arXiv:2103.07579.

[19] Exploring Simple Siamese Representation Learning (2021): - Zeming Li, Shyamal Buch, Denny Zhou, Hartwig Adam, Jonathon Shlens, Zhenhua Liu. International Conference on Machine Learning (ICML).

[20] Learning Transferable Visual Models From Natural Language Supervision (2021): - R. P. Alexander, Michalis Papakostas, Razvan Pascanu, Yujia Li, Razvan Pascanu. NeurIPS.

