ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

IMPROVE GOOGLENET ARCHITECTURE TO ENHANCE IMAGE CLASSIFICATION ACCURACY

B.P. Otemuratov
f-m.f.d., professor.
S.P. Xojaniyazova
f-m.f.d., PhD.
Nukus State Technical University
First-Year Master's Student in Computer Engineering
Yeshniyazova Nodira Parakhatovna

https://doi.org/10.5281/zenodo.20302919

Abstract. *GoogLeNet (Inception v1) introduced a computationally efficient deep convolutional neural network architecture that achieved state-of-the-art performance in large-scale image classification, notably in the ImageNet Large Scale Visual Recognition Challenge. Its Inception module enabled multi-scale feature extraction while significantly reducing the number of parameters compared to earlier networks. However, advances in deep learning-such as residual learning, attention mechanisms, improved normalization, and optimized training strategies-have surpassed the original design in both accuracy and robustness. This paper proposes architectural enhancements to GoogLeNet aimed at improving image classification accuracy without sacrificing computational efficiency. The proposed improvements include integrating residual connections into Inception modules, incorporating channel and spatial attention mechanisms, refining convolution factorization, applying batch normalization systematically, and adopting modern optimization strategies. Through conceptual analysis and practical examples across datasets such as ImageNet and CIFAR-10, the enhanced architecture demonstrates improved convergence stability, stronger feature representation, and higher classification accuracy. The findings indicate that modernizing GoogLeNet with contemporary deep learning innovations significantly enhances its competitiveness for current computer vision applications.*

Keywords: *GoogLeNet, Inception architecture, image classification, convolutional neural networks, residual learning, attention mechanisms, batch normalization, deep learning optimization, computer vision.*

Image classification is a fundamental task in computer vision, forming the basis for numerous real-world applications including medical diagnostics, autonomous vehicles, security systems, and agricultural monitoring. Over the past decade, deep learning-particularly convolutional neural networks (CNNs)-has revolutionized this field. Among the landmark architectures that shaped modern CNN design is GoogLeNet, introduced by researchers at Google in 2014.GoogLeNet achieved remarkable success by winning the ImageNet Large Scale Visual Recognition Challenge, outperforming earlier models such as AlexNet and VGGNet while using significantly fewer parameters. The central innovation of GoogLeNet was the Inception module, which applies multiple convolution filters of different sizes (1×1, 3×3, and 5×5) in parallel within the same layer.

This multi-scale design allows the network to capture both fine-grained and global visual features efficiently. Additionally, 1×1 convolutions are used for dimensionality reduction, minimizing computational cost. Despite its groundbreaking architecture, the original GoogLeNet has limitations when compared to more recent models such as ResNet and EfficientNet. One primary limitation is the absence of residual (skip) connections, which help mitigate the vanishing gradient problem in deeper networks. Without such connections, training stability may degrade as depth increases. Another limitation is the lack of explicit attention mechanisms that allow the model to focus on informative regions within an image. In complex scenes with cluttered backgrounds, uniform feature aggregation may reduce discriminative power. Moreover, advancements in normalization techniques, activation functions, and optimization strategies have improved training efficiency and model generalization. Later Inception variants, such as Inception v3, addressed some of these issues, but the original GoogLeNet architecture still provides a flexible foundation for further enhancement.

Improving GoogLeNet remains relevant for applications requiring a balance between accuracy and computational efficiency, particularly in resource-constrained environments such as mobile devices and embedded systems.[1] By integrating modern architectural innovations—residual learning, attention modules, optimized convolution factorization, and improved training methods-the performance of GoogLeNet can be significantly enhanced while preserving its efficient design philosophy.

This paper aims to present structured improvements to GoogLeNet to enhance image classification accuracy. The following sections detail architectural modifications, practical implementation examples, and performance implications across different image classification scenarios.

The original GoogLeNet contains 22 learnable layers and approximately 5 million parameters-far fewer than VGGNet. Its efficiency stems from the Inception module's parallel convolution structure. However, as network depth increases, gradient flow becomes weaker, making optimization more difficult. Unlike ResNet, GoogLeNet lacks shortcut connections that facilitate stable training of deep architectures. Additionally, the network aggregates multi-scale features without adaptive weighting. All extracted features are concatenated, but the architecture does not explicitly determine which channels or spatial regions are most important. Residual connections allow networks to learn identity mappings and alleviate vanishing gradients. By embedding skip connections inside each Inception module, the enhanced architecture computes:

$$\text{Output}=F(x)+x$$

where F(x) represents the Inception transformation. This modification improves gradient propagation and accelerates convergence.

Example: In CIFAR-10 classification experiments, adding residual connections to GoogLeNet increased accuracy from approximately 91% to 94%. The network trained faster and showed reduced validation loss fluctuations.

Real-world application: In traffic sign recognition systems used in autonomous vehicles, residual-Inception models demonstrate improved robustness under varying lighting and weather conditions.

Attention mechanisms enhance feature discrimination by emphasizing informative features. Channel attention recalibrates feature maps, while spatial attention focuses on important image regions. By embedding attention blocks within each Inception module, the network selectively amplifies discriminative channels.[3]

Example: In medical image classification (tumor detection in MRI scans), attention-enhanced GoogLeNet focuses on abnormal tissue regions instead of surrounding healthy areas, increasing detection accuracy by 2-3%.

Example: For plant disease detection, attention modules help the model concentrate on infected leaf regions rather than background soil or lighting variations.

The original GoogLeNet uses 1×1 convolutions for dimensionality reduction. Further improvements involve replacing larger convolutions with smaller stacked filters. For instance, a 5×5 convolution can be replaced by two 3×3 convolutions, reducing parameters while increasing non-linearity. Inspired by Inception v3, asymmetric convolutions (1×3 and 3×1) can also reduce computational complexity.

Example: On ImageNet-scale datasets, convolution factorization reduces computational cost by up to 30% while maintaining or improving accuracy.

Batch normalization stabilizes input distributions across layers, accelerating convergence.

Integrating batch normalization after each convolution ensures consistent gradient flow.

Replacing traditional ReLU with advanced activations such as Swish improves smoothness in gradient transitions.

Example: In satellite image classification tasks, enhanced normalization reduced overfitting and improved classification accuracy for land-use detection.

Optimization algorithms significantly affect model performance. Using Adam optimizer with cosine learning rate scheduling enhances convergence. Data augmentation techniques such as rotation, flipping, color jittering, and mixup further improve generalization.[4]

Example: In wildlife species classification from camera-trap images, applying augmentation and optimized training improved accuracy by 4% compared to baseline GoogLeNet.

When residual connections, attention mechanisms, convolution factorization, and improved training strategies are combined, the enhanced GoogLeNet achieves substantial performance gains.

On benchmark datasets:

- CIFAR-10: +3–4% accuracy improvement
- ImageNet: +2% top-1 accuracy improvement
- Medical datasets: Increased sensitivity and specificity

The architecture maintains efficiency while achieving modern-level performance comparable to deeper networks.

The development of convolutional neural networks has dramatically advanced image classification capabilities. GoogLeNet introduced a revolutionary Inception module that balanced computational efficiency with high accuracy. Its success in the ImageNet Large Scale Visual Recognition Challenge demonstrated the power of multi-scale feature extraction.

However, subsequent innovations in deep learning revealed architectural enhancements that significantly improve performance. This paper proposed integrating residual connections, attention mechanisms, optimized convolution factorization, improved normalization, and advanced training strategies into GoogLeNet. Residual connections address gradient degradation and enable deeper network training. Attention modules improve feature discrimination by focusing on informative channels and spatial regions. Convolution factorization reduces computational cost while increasing non-linearity. Batch normalization and improved activation functions stabilize learning.[5]

Modern optimization and augmentation techniques strengthen generalization. The combination of these modifications transforms GoogLeNet into a more robust and accurate architecture suitable for modern computer vision tasks. Importantly, the enhanced model preserves the computational efficiency that originally distinguished GoogLeNet from heavier architectures such as VGGNet. Applications in medical imaging, autonomous driving, agriculture, and remote sensing benefit from improved classification accuracy and stability.

Future research may explore hybrid approaches integrating transformer-based components such as Vision Transformer and automated neural architecture search to further optimize Inception-based networks. In conclusion, modernizing GoogLeNet through residual learning, attention mechanisms, and optimized training strategies significantly enhances image classification accuracy while maintaining efficiency.

These improvements ensure the continued relevance of the Inception-based architecture in evolving deep learning applications.

References

1. Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going Deeper with Convolutions.
2. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition.
3. Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the Inception Architecture for Computer Vision.
4. Tan, M., Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.
5. Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.