Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22705

Research on Image Classification Based on Deep Learning

Yiming Gou

School of Computer and Network Security, Chengdu University of Technology, Chengdu, China
gouyiming_email@163.com

Abstract: This study aims to validate the performance differences between Convolutional Neural Networks and Transformer architectures in image classification. Based on their distinct feature extraction mechanisms, three algorithms—Resnet101, Resnet152, and Vision Transformer (ViT)—were trained and tested on the mini-Imagenet dataset. The results show that ViT achieved a top-1 accuracy of 92.79%, surpassing Resnet101's 90.98% and Resnet152's 91.71%. This largely demonstrates ViT's superiority over traditional convolutional neural networks in classification accuracy, as its unique Transformer architecture can more effectively capture global features and contextual information. Compared to the limitations of conventional CNN algorithms, ViT evidently exhibits enhanced performance in complex image classification tasks through its self-attention mechanism.

Keywords: Deep Learning, Convolutional Neural Networks (CNN), Vision Transformer (ViT), Image Classification

1. Introduction

In the evolution of computer vision, image classification tasks have been closely intertwined with artificial intelligence algorithms. Traditional machine learning algorithms once dominated this field, but with rapid technological advancements, the emergence of deep learning algorithms has opened a more efficient path for image classification research. Compared to conventional machine learning approaches, deep learning algorithms exhibit significantly enhanced performance and potential, driving unprecedented breakthroughs in image classification tasks.

The development of image classification algorithms has undergone continuous adjustments, transitioning from traditional machine learning to deep learning. In traditional machine learning, feature extraction and manual feature design were critical steps. Commonly used handcrafted features include Haar-like features (Haar)[1], Histogram of Oriented Gradients (HOG)[2], Scale-Invariant Feature Transform (SIFT)[3], and Local Binary Patterns (LBP)[4]. These features capture texture, shape, and edge information for subsequent classification tasks. For instance, Haar features, adept at capturing edges and linear patterns, were widely applied in face detection. HOG features, computed as histograms of oriented gradients over local image regions, effectively describe shape and texture characteristics and achieved notable success in pedestrian detection. After extracting handcrafted features, classifiers such as Support Vector Machines (SVM)[5], k-Nearest Neighbors (k-NN)[6], Decision Trees[7], and Random Forests [8] were typically employed for final category determination.

However, with the rise of deep learning, Convolutional Neural Networks (CNNs)[9]and their improved variants have become mainstream in image classification. CNNs, based on multi-layered

DOI: 10.54254/2755-2721/2025.TJ22705

neural architectures, automatically extract image features through convolutional layers, pooling layers, and fully connected layers. Unlike traditional approaches, CNNs eliminate manual feature engineering, significantly enhancing generalization and performing better on large datasets. Throughout CNN's evolution, architectures such as VGGNet (Visual Geometry Group Network)[10] and Residual Networks (ResNet)[11] have demonstrated exceptional efficacy. VGGNet improves performance by increasing network depth, utilizing small 3×3 convolutional kernels and strides to stack convolutional and pooling layers. ResNet addresses deep network degradation through residual learning and shortcut connections, enabling more effective training of deeper networks.

As deep learning advances and researchers worldwide explore new methodologies, the Vision Transformer (ViT)[12] has introduced novel paradigms for image classification. ViT applies the Transformer architecture to image tasks by dividing images into fixed-size patches, treating them as sequential inputs. Leveraging its self-attention mechanism, ViT globally processes image information, enabling efficient feature extraction and scalability to large datasets. To further enhance generalization and performance, the Semantic Cluster Vision Transformer (SCViT)[13] was proposed. SCViT integrates convolutional operations with Transformers, employing conditional positional encoding and semantic fairness-clustered self-attention modules to improve local detail and semantic information capture, achieving more comprehensive and effective feature extraction than ViT. This study focuses on exploring the application of deep learning in image classification.

2. Methodology

2.1. Convolutional neural network

A Convolutional Neural Network (CNN) is a deep learning model specifically designed for processing grid-structured data (e.g., images, audio), inspired by biological visual systems. The concept originated from Hubel[14] and Wiesel's studies in the late 1950s and early 1960s on cat visual cortex cells, which identified simple and complex cells responsive to localized regions (receptive fields) of visual stimuli. This discovery laid the foundation for the notions of local receptive fields and hierarchical feature extraction in CNNs. The core principle of CNNs involves automatically extracting features through convolutional operations, reducing data dimensionality via pooling, and ultimately performing classification or other tasks through fully connected layers. The foundational architecture is illustrated in Figure 1.

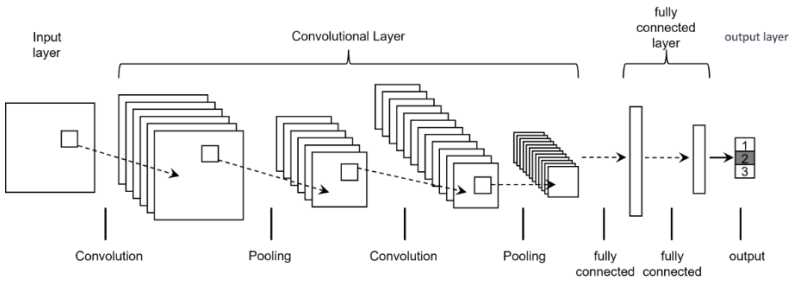

Figure 1: Basic model of convolutional neural network

The convolutional neural network comprises four critical steps: convolutional layer, activation function, pooling layer, and fully connected layer.

First, a convolution operation is applied to the input image (using a 3-channel example). The formula is as follows:

$$\text{Output}(i, j, c) = \sum_{k=1}^{3} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} \text{Kernel}(m, n, k, c) \times \text{Input}(iS + m - P, jS + n - P, k) + \text{Bias}(c) \quad (1)$$

DOI: 10.54254/2755-2721/2025.TJ22705

Output(i,j,k) denotes the pixel value at position (i,j) in channel k of the output image. Kernel(m,n,k,c) represents the weight value of convolution kernel c at position (m,n) in channel k. Bias(c) is the bias term for convolution kernel c. S indicates the stride, and P denotes the padding.。

Next, an activation function introduces nonlinearity to determine whether data is propagated to subsequent layers. Four common activation functions are: ReLU(Rectified Linear Unit, ReLU);Sigmoid;Tanh(Hyperbolic Tangent, Tanh); Softmax (converts outputs into probability distributions, often used for classification).

A pooling operation (including max pooling and average pooling) is then applied, aiming to reduce data volume and enhance computational efficiency. Finally, the fully connected layer maps features to the output space for classification. The functional process is:

$$y = \sigma(W^T x + b) \tag{2}$$

Here:x is the input vector with a dimension of n. W is the weight matrix with dimensions n×m. b is the bias vector with a dimension of m. σ denotes the activation function for introducing nonlinearity. y is the output vector with a dimension of m.

2.2. ResNet (residual neural network)

After introducing the fundamental convolutional neural network (CNN) architecture, researchers Kaiming He et al. from Microsoft Research proposed a novel deep neural network architecture in 2015: the Residual Neural Network (ResNet). Building upon CNNs, ResNet innovatively introduces residual learning and shortcut connections to mitigate the vanishing gradient problem. A structural comparison between ResNet and the traditional VGG network is illustrated in Figure 2.

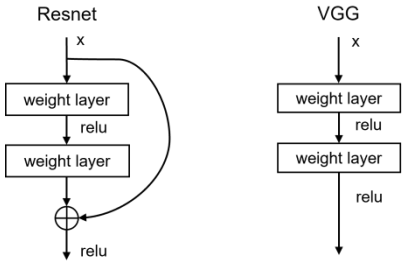

Figure 2: Comparison of partial architectures: ResNet vs VGG

The figure clearly demonstrates that ResNet incorporates residual learning blocks to alleviate gradient vanishing, enabling training of significantly deeper networks compared to traditional VGG. The core idea of residual learning is to train ultra-deep networks (e.g., ResNet-152) using skip connections. The implementation workflow is shown in Figure 3.

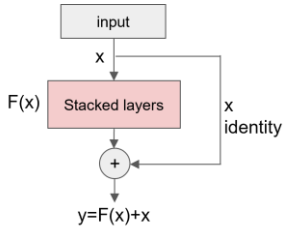

Figure 3: Basic workflow of ResNet

Similar to traditional CNNs, the ResNet model follows the basic forward propagation sequence: input → convolutional layers → activation functions → pooling → fully connected layers →

DOI: 10.54254/2755-2721/2025.TJ22705

output. Here, "Stacked Layers" represent repeated sequences of convolutional layers, activation functions, and pooling layers, with the number of layers determined by training objectives. The innovation of ResNet lies in introducing residual mapping between input and output, rather than direct mapping, by incorporating skip connections at each stage. For a single residual block, the output can be expressed as:

$$x_{l+1} = x_l + F(x_l, W_l) \tag{3}$$

where x_l is the input to the l layer, $F(x_l,W_l)$ denotes the residual mapping (typically composed of convolutional layers, batch normalization layers, and activation functions), and x_{l+1} represents the output of the l+1 layer.

For multiple residual blocks, this can be generalized as:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{4}$$

The ResNet algorithm excels in image recognition and classification, enabling the learning of complex, high-level features from input images. Its depth allows the model to extract highly abstract and nuanced features. However, ResNet's convolutional operations primarily focus on local feature extraction, with relatively limited capability to capture global features. This limitation motivated the development of the Vision Transformer (ViT) algorithm.

2.3. Vision transformer (ViT) algorithm

As previously mentioned, convolutional neural networks (CNNs) have achieved significant practical success in computer vision. However, their inherent focus on local features imposes limitations in image recognition and classification. Concurrently, the Transformer architecture (Figure 4) achieved groundbreaking advancements in natural language processing (NLP), prompting researchers to explore its application in computer vision. In 2020, the Vision Transformer (ViT) was introduced, breathing new life into image classification tasks.

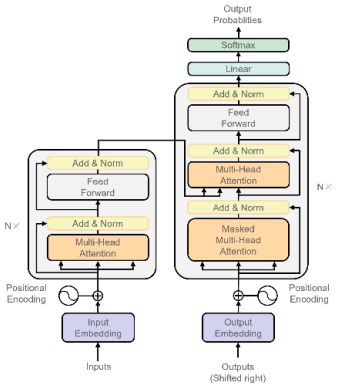

Figure 4: Basic framework of transformer

The ViT algorithm primarily comprises four components: patch module, positional encoding, multi-head attention structure, and multilayer perceptron (MLP) (Figure 5). These components are elaborated below:

DOI: 10.54254/2755-2721/2025.TJ22705

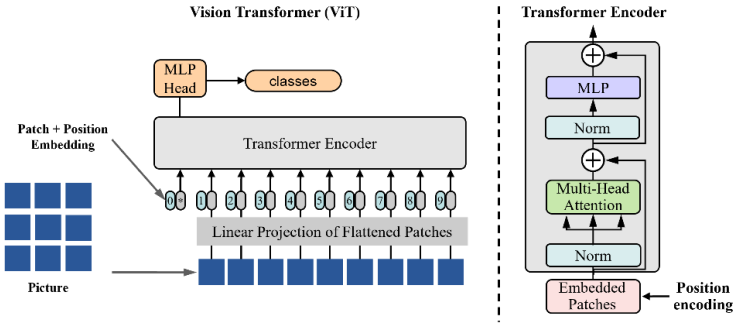

Figure 5: Workflow of the ViT framework

2.3.1. Patch module

In this module, the input image is partitioned into fixed-size image patches, which are linearly embedded into a high-dimensional space. The embedding formula is:

$$\mathrm{E} = \mathrm{Linear}(\mathrm{X}) \tag{5}$$

where Linear denotes a linear transformation mapping each patch to an embedding dimension D, and E represents the sequence of embedded vectors.

2.3.2. Positional encoding

This component assigns positional information to each image patch by incorporating it into the Embedded Patches, enabling the model to leverage spatial relationships between patches. The formula is:

$$E\mathrm{pos} = E + P \tag{6}$$

where P is the positional encoding matrix, which can be generated using various encoding methods.

2.3.3. Multi-head attention

This module employs a multi-head attention mechanism to project inputs linearly into multiple feature subspaces. Parallel processing via independent attention heads (Figure 6) allows the model to attend to different positions in the input sequence simultaneously, capturing richer information. The vectors are then concatenated and mapped to the final output.

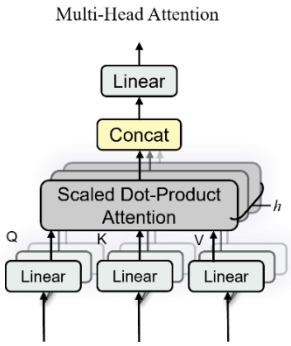

Figure 6: Multi-Head Attention Mechanism

The entire process can be expressed as:

$$MSA(Q,K,V) = Concat(head_1, \ldots, head_h)W^o \tag{7}$$

DOI: 10.54254/2755-2721/2025.TJ22705

Where $Q=XW^Q$, $K=XW^K$, $V=XW^V$,$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

2.3.4. Multilayer perceptron (MLP)

The MLP, functionally equivalent to a feedforward neural network (FNN), applies nonlinear transformations to the embedded vectors at each position, enhancing the model's expressive power. The process is formulated as:

$$\text{MLP}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \tag{9}$$

where W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors, and GELU denotes the Gaussian Error Linear Unit activation function.

3. Experiments

To validate the performance evaluations of the aforementioned algorithms, experiments were conducted on a public dataset. This section details the dataset description and experimental parameter settings, followed by a comprehensive comparison between a CNN-based network model and a Transformer-based (ViT) neural network.

3.1. Dataset description

To evaluate the performance of the two algorithms, the mini-ImageNet dataset was employed. Compared to the traditional CIFAR-10 dataset, mini-ImageNet is more complex yet significantly smaller than the full ImageNet dataset, making it ideal for model training, evaluation, and experimental research.

The original mini-ImageNet dataset contains 100 classes with 60,000 RGB images, each sized 84×84. For this experiment, after data filtering and adjustments, the dataset was refined to include 99 distinct object classes and 60,000 images, each resized to 224×224 pixels. The dataset covers a broad range of image contents with diverse categories.

3.2. Experimental setup

Prior to experimentation, the dataset was partitioned into training and test sets. 75% of the full dataset was allocated for training, and 25% for testing. Within the training set, 60% was used for training and 15% for validation. During image preprocessing, all images were resized to 224×224 pixels for consistency. In the training phase, Total training epochs: 200,Batch size: 32,Optimizer: Momentum SGD[15] (initial momentum: 0.9, weight decay: 0.0005),Learning rate: Initialized to 0.01, with a minimum learning rate of 0.0001 (1% of the initial rate), following a cosine annealing schedule[16]. Checkpoints (including weights, training loss, and accuracy) were saved every 10 epochs. Data loading utilized multi-threading with num_workers=4. All experiments were executed on an NVIDIA RTX 4090 GPU with CUDA 11.3 and PyTorch 1.1.0. During validation, the evaluteTop1_5 function was invoked to compute evaluation metrics: Top-1 accuracy: The proportion of predictions where the single most probable class matches the ground truth label. Top-5 accuracy: The proportion of predictions where the ground truth label is among the top five most probable classes. Recall and precision were also calculated to assess model performance.

DOI: 10.54254/2755-2721/2025.TJ22705

3.3. Experimental result evaluation and analysis

To validate the performance of the algorithms discussed in this paper and compare the effectiveness of convolutional neural network (CNN) architectures versus Transformer architectures in image recognition, ResNet101[17], ResNet152[18], and ViT were employed for image classification tasks on the modified mini-ImageNet dataset.

Table 1: Classification results of three methods on the mini-ImageNet dataset

	Algorithm	top-1 accuracy	top-5 accuracy	mean Recall	mean Precision
CNN	Resnet101	90.98%	98.00%	90.98%	91.10%
	Resnet152	91.71%	98.07%	91.71%	91.81%
Transformer	ViT	92.79%	98.71%	92.79%	92.89%

When comparing the training-validation accuracies of ResNet101, ResNet152, and ViT, ResNet101 achieved a top-1 accuracy of 90.98%, ResNet152 attained 91.71%, and ViT reached 92.79%, indicating that ViT exhibits the highest classification accuracy among the three algorithms.

The results demonstrate that both ResNet101 and ResNet152 are deep CNNs based on residual networks, differing primarily in layer depth. The former has fewer layers than the latter. While residual connections mitigate the vanishing gradient problem in CNNs, extremely deep networks still face challenges such as training difficulty and overfitting, which degrade validation performance. Additionally, their local feature extraction approach may neglect certain low-level features, leading to the loss of effective discriminative information.

In contrast, ViT leverages the Transformer architecture to better capture global features and contextual information in images, resulting in superior classification accuracy. Its unique self-attention mechanism enables the model to consider relationships between all pixels during processing, allowing it to learn richer feature representations. Consequently, ViT achieves the highest accuracy among the three algorithms, primarily due to its ability to efficiently capture global features and contextual information, as well as its improved data utilization efficiency when trained on large-scale datasets.

The classification accuracies for specific object categories are partially listed below:

Table 2: Partial classification results by category

Type	Resnet 101 accuracy / %	Resnet 152 accuracy / %	ViT accuracy / %	Type	Resnet 101 accuracy / %	Resnet 152 accuracy / %	ViT accuracy / %
Ant	81	84	82	oboe	81	84	86
Arctic_fox	89	91	93	orange	97	93	97
ashcan	79	76	79	organ	94	92	97
barrel	87	87	87	photocopier	92	91	94
beer_bottle	88	92	90	poncho	86	90	93
bolete	94	96	99	reel	86	87	89
boxer	86	93	94	robin	94	94	98
cocktailshaker	91	89	94	rock_beauty	94	95	94
combination_lock	86	90	92	Saluki	92	89	87
consomme	98	93	97	school_bus	98	95	96
dome	93	93	98	scoreboard	93	93	96

DOI: 10.54254/2755-2721/2025.TJ22705

Table 2: (continued).

dugong	96	96	98	slot	95	97	100
Ear	86	87	86	snorkel	97	94	92
electric_guitar	93	89	92	solar_dish	89	93	93
golden_retriever	92	89	95	stage	85	87	85
goose	99	98	99	street_sign	91	92	89
hair_slide	82	88	86	tile_roof	93	93	91
holster	94	94	97	toucan	97	99	99
hotdog	90	90	95	triceratops	99	99	97
house_finch	94	94	96	trifle	95	96	98
iPod	89	90	93	unicycle	93	91	96
komondor	91	97	92	upright	93	92	92
ladybug	97	95	98	vase	88	92	92
lipstick	90	94	97	white_wolf	90	92	90
miniskirt	90	91	96	yawl	98	98	99

Based on the table, most categories exhibit a gradual increase in accuracy across ResNet101, ResNet152, and ViT. For instance, in the "slot game" category, ViT achieves 100% accuracy, while ResNet101 and ResNet152 attain only 95% and 97%, respectively. This highlights ViT's superior performance in classifying images with complex features or high intra-class variability.

4. Summary

This paper compares the image classification performance of CNN-based architectures (ResNet101 and ResNet152) and the Transformer-based architecture (ViT) on the mini-ImageNet dataset, validating the effectiveness of different algorithms in image classification tasks. Experimental results demonstrate that ViT outperforms traditional CNNs in classification accuracy. While CNN algorithms retain significant advantages in image recognition tasks, ViT's Transformer architecture excels in capturing global features and contextual information, enabling richer feature learning. Although ResNet mitigates vanishing gradients through residual connections, CNN frameworks still face challenges such as training complexity and limitations in local feature recognition. In conclusion, ViT exhibits superior performance and potential in image classification due to its unique architecture and mechanisms.

References

[1] *Xiaopei Yuan, Xiaofeng Chen, Ming Lian. Multi-feature fusion target detection algorithm based on Haar-like and LBP. Computer Science, 2021, 48(11): 219-225.*

[2] *Yufei Duan, Jiwei Sun, Geng Dong, et al. Automatic recognition method for Camellia oleifera shells and seeds based on HOG+LBP features. Modern Food Science and Technology, 2024, 40(10): 270-275. DOI:10.13982/j.mfst.1673-9078.2024.10.0864..*

[3] *Ziwen Yu, Ning Zhang, Yue Pan, et al. Heterogeneous image matching based on an improved SIFT algorithm. Laser & Optoelectronics Progress, 2022, 59(12): 214-225.*

[4] *Wu Xu, Han Gao, Xinda Wang, et al. Infrared facial expression recognition technology based on LBP feature matching algorithm. Laser Journal, 2023, 44(03): 158-162. DOI:10.14016/j.cnki.jgzz.2023.03.158..*

[5] *Haigang Zha, Xiangyang Qi, Huaitao Fan. HRRP ship target classification method for small-sample imbalanced data based on SVM. Modern Electronics Technique, 2024, 47(15): 109-114. DOI:10.16652/j.issn.1004-373x.2024.15.018.*

[6] *Wei Zhou, Danbo Yi. Weighted K-NN classifier and its applications. Acta Automatica Sinica, 1989, (02): 174-177.DOI:10.16383/j.aas.1989.02.013.*

DOI: 10.54254/2755-2721/2025.TJ22705

[7] Bo Zhang. Visual target tracking algorithm based on decision tree classification. Journal of Detection & Control, 2022, 44(06): 87-92.
[8] Ruijiao Wu. Masson pine recognition combining object-oriented CNN and random forest. Geomatics & Spatial Information Technology, 2024, 47(10): 50-53+58.
[9] Hongda Liu, Xuhui Sun, Yibin Li, et al. A survey of deep learning models for image classification based on convolutional neural networks. Computer Engineering and Applications, 1-29 [2025-03-15]. http://kns.cnki.net/kcms/detail/11.2127.TP.20250213.1223.013.html.
[10] Youjun Yue, Bokai Tian, Hongjun Wang, et al. Application of improved VGG model in apple appearance classification. Science Technology and Engineering, 2020, 20(19): 7787-7792.
[11] Jianfang Cao, Cunhe Peng, Zhiqiang Chen, et al. Ancient mural classification method based on improved ResNet deep learning. Electronic Measurement Technology, 2025, 48(01): 186-196. DOI:10.19651/j.cnki.emt.2416339.
[12] Xin Ying, Ning Zhang, Si Shen. Trousers silhouette recognition and classification based on Vision Transformer and transfer learning. Journal of Silk, 2024, 61(11): 77-83.
[13] Linglong Zhu, Yagang Wang, Yi Chen. Image classification algorithm integrating Transformer and CNN. Electronic Science and Technology, 1-11 [2025-03-15]. https://doi.org/10.16180/j.cnki.issn1007-7820.2025.10.012.
[14] Fanji Gu. David Hubel: Father of modern visual science. Chinese Journal of Nature, 2016, 38(04): 307-312.
[15] Tao Ju, Heting Kang, Shuai Liu, et al. A multi-step delayed parameter update parallel optimization method for deep neural networks. Journal of Harbin Institute of Technology, 1-17 [2025-03-19]. http://kns.cnki.net/kcms/detail/23.1235.t.20241231.1630.002.html.
[16] Zhenchao Tang, Wei Wei, Weiran Luo, et al. Remote sensing image semantic segmentation integrating cosine annealing and dilated convolution. Journal of Remote Sensing, 2023, 27(11): 2579-2592.
[17] Congwang Bao, Wei Jiang, Yongzhi Liu, et al. Gear defect detection based on improved ResNet101 network. Modular Machine Tool & Automatic Manufacturing Technique, 2024, (08): 145-148+153. DOI:10.13462/j.cnki.mmtamt.2024.08.028.
[18] Zhuyuan Qin, Haozhong Wu, Daiqing Tan, et al. Fine-grained recognition of Ophiopogon japonicus based on multi-scale ResNet fused with attention mechanism. Computer and Modernization, 2023, (07): 105-111.