


Advances and Analysis in Convolutional Neural Networks: A Comparative Study of AlexNet and ResNet

Jiawei Chen ^a

School of Computer Engineering, Guangzhou City University of Technology, Guangzhou, China

Keywords: Convolutional Neural Networks, AlexNet, ResNet, Deep Learning.


Abstract: Deep learning, particularly through Convolutional Neural Networks (CNNs), has significantly impacted various fields and is integral to many aspects of daily life. This study focuses on CNNs, with a specific emphasis on two foundational models: ImageNet Classification with Deep Convolutional Neural Networks (AlexNet) and Residual Network (ResNet). CNNs excel in image processing and recognition but require substantial data for effective training. AlexNet, a pioneer in the deep learning revolution, employs Rectified Linear Unit (ReLU) activation functions and Dropout techniques to mitigate overfitting but struggles with irregular and large datasets. On the other hand, ResNet introduces residual connections to address the vanishing gradient problem, although it still faces challenges related to overfitting. The paper provides a detailed comparison of CNN principles with traditional Neural Networks (NNs), highlighting the strengths and weaknesses of AlexNet and ResNet. It also explores the current challenges in deep learning, outlining potential areas for future research and development. This study offers insights into the evolution of CNN technologies and suggests directions for overcoming existing limitations and enhancing future advancements.

1 INTRODUCTION

As one of the most trending researched topics in computer vision and machine learning, biometrics technology has been widely applied in a variety of security and identity verification systems owing to the swift development of information technology and artificial intelligence. Compared with other biometric technologies thesis can see in daily life, such as iris (Nguyen et al., 2024), fingerprint (Kortli et al., 2020), and voice recognition, facial recognition has gradually emerged as the top choice result of its natural, contactless, and convenient features (Liu et al., 2023). Through analyzing and matching faces automatically which are based on the unique characteristics of a living person, this technology can enhance the user experience as well as improve significantly security. Utilizing facial recognition technology has become a common scene in life, such as in security (Nan et al., 2022; Li et al., 2021), finance, and safe driving (Jeong and Ko, 2018). It is significant for researchers who need more study and development of related technologies in light of the

vast potential of facial recognition technology in real-world applications.

As an organ possessed from birth, human beings have been identified by using faces almost since childhood, resulting in humans being talented in recognizing known faces, but once the number of faces becomes large and unknown, people's recognition ability will deteriorate. There are a number of researchers who have played a significant role in advancing facial recognition technology over the past few decades. In order to identify, in its early stages, facial recognition primarily depended on extracting a large number of facial characteristic coordinates and comparing them with stored facial data in a database (Wang and Deng, 2021). However, this method would show limitations when facing complex environmental conditions such as lighting conditions, posture changes, and dynamic backgrounds (Beham and Roomi, 2013). Additionally, the development of sensors has been shown to be able to obtain two-dimensional and three-dimensional information at the same time, leading some researchers to mix the two types of information into a system that can improve analytical

^a <https://orcid.org/0009-0007-4940-0813>

capabilities (Kortli et al., 2020). By the late 20th century, advancements in computer vision and computational power paved the way for techniques that could reduce data dimensionality while retaining key facial features, greatly enhancing the accuracy of facial recognition. The 21st century brought with it a seismic shift in facial recognition thanks to advances in deep learning and computer vision. Deep learning models based on Convolutional Neural Network (CNN) have pushed the boundaries in facial recognition. Practically, Google's FaceNet, introduced in 2015, achieved remarkable performance on multiple public data sets (Qinjun et al., 2023). In today's era, the advancement and application of deep learning have not only greatly enhanced recognition accuracy but demonstrated exceptional performance in managing complex scenarios and large-scale data.

The research aims to explore and develop the fundamental concepts and core technologies of facial recognition. The primary objective of this study is to first organize and summarize the concepts and background of facial recognition technology. It then proceeds to analyze and discuss the structure and logic underlying the implementation of core facial recognition technologies. Furthermore, the study introduces the strengths and weaknesses of the methods employed within the internal logic of facial recognition systems, while also providing insight into potential future developments in the field.

2 METHODOLOGIES

2.1 Dataset Description

Image processing based on CNNs requires the collection of a great number of human face pictures with different angles and expressions for computer learning. This study reviews the commonly used datasets in several different papers, including their sources, quantity, and content. The Operations Research Laboratory (ORL) dataset provided by AT&T Laboratories at Cambridge University was captured between April 1992 and April 1994 (AT et al., 2017). The dataset consists of 40 different subjects, with each subject containing 10 images. Specifically, the images of each subject exhibit variations in facial brightness and expressions due to different capture times. The backgrounds of these images are all black, and they are taken from a front-facing, upward angle. The thumbnails of all images are shown in Figure 1. Each image is 92×112 pixels,

with 256 Gray levels per pixel. An enlarged image of each subject is shown in Figure 2.

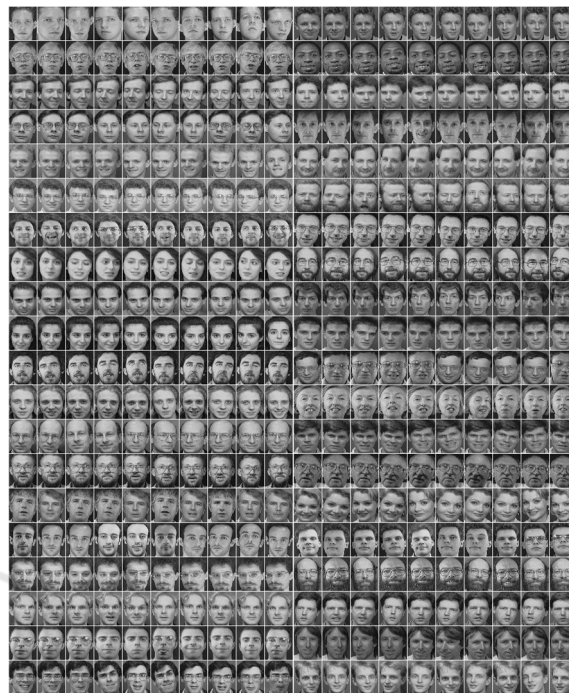


Figure 1: The ORL face database (AT et al., 2017).



Figure 2: The set of ten images for one subject (AT et al., 2017).

2.2 Proposed Approach

The purpose of this study is to examine the evolution of CNN-based face recognition technology and assess the benefits and drawbacks of various neural network approaches (see in Figure 3). Initially, the paper introduces CNNs, along with two significant variants includes ImageNet Classification with Deep Convolutional Neural Networks (AlexNet) and Residual Network (ResNet), which have achieved notable advancements in the field. It provides an overview of CNN fundamentals and highlights how AlexNet and ResNet have innovated beyond



Figure 3: The pipeline of the study (Picture credit: Original).

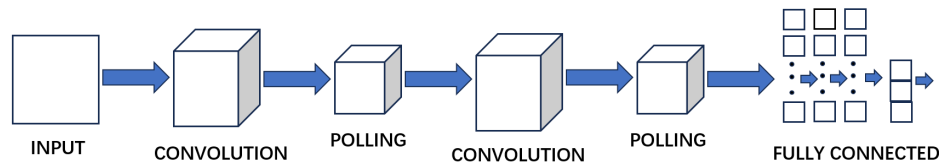


Figure 4: The structure of CNN (Picture credit: Original).

conventional convolutional neural networks. The study then analyzes three CNN methodologies, exploring their respective advantages and limitations. CNNs have notably addressed challenges such as the inefficient processing of large image datasets, incomplete feature extraction, and low recognition accuracy. AlexNet excels in handling large-scale datasets but requires substantial computational resources and extended training times. In contrast, ResNet overcomes issues like gradient vanishing and model degradation, although it demands considerable training data and is susceptible to overfitting. This comprehensive examination underscores the progress made in face recognition technology and identifies potential areas for future research and development.

2.2.1 Introduction of CNN

CNN is a feedforward neural network and a deep-learning neural network. It mainly consists of three parts, respectively convolution layer, pooling layer, and fully connected layer. The convolution layer plays a crucial role in CNN and is responsible for executing numerous calculations to extract local features in the image. During the computation, a specific convolution kernel, such as a sliding window with fixed weights, is required, and then multiplies with the image and is summed to produce a set of convolved data. After the calculation is completed, CNN requires a pooling layer to simplify the data. There are two main types of pooling layers, namely maximum pooling and average pooling. Figure 4 illustrates the structure. The primary function of the pooling layer is to perform downsampling and feature selection for different regions, reducing the number of features and thus simplifying the model parameters. This approach may compromise the integrity of the data but significantly reduces data complexity, thereby improving processing efficiency.

The role of the fully connected layer is to integrate the local features extracted by the previous two layers to form global features. Then, using linear transformations and activation functions, the global features are filtered, and the final prediction result is output.

2.2.2 Introduction of AlexNet

AlexNet is a milestone of CNN in the field of computer vision, which has attracted people's attention to the technology of CNN. AlexNet adopts a deeper network architecture with five convolutional layers and three fully connected layers, which significantly improve image classification performance. In order to improve the problem of calculation speed, AlexNet uses multiple graphics processing units (GPU) for training, with each GPU handling part of the computation, which significantly speeds up the training process. In addition, AlexNet used the non-saturating Rectified Linear Unit (ReLU) function as its activation function, which is simpler to calculate. The ReLU function can not only significantly accelerate convergence during training but also better mitigate the vanishing gradient problem compared to the traditional Sigmoid function. AlexNet implemented overlapping pooling, where the stride (step size) is smaller than the window size, leading to overlapping regions during the pooling process. Additionally, AlexNet used data augmentation and dropout techniques to prevent overfitting and enhance recognition accuracy. Dropout refers to randomly removing certain neurons during training to reduce the dependency between multiple neurons, thereby improving generalization and introducing uncertainty. By deepening the network structure, AlexNet demonstrated the potential of deep neural networks and triggered a surge of interest in deep learning.

2.2.3 Introduction of ResNet

ResNet is a deep residual network structure that utilizes residual modules and residual connections to build a network. Although increasing the depth of a neural network can enhance its feature extraction capabilities, going beyond a certain depth may lead to issues like vanishing or exploding gradients, which weaken this ability and make training more difficult, ultimately reducing accuracy. Additionally, excessively deep networks are prone to overfitting, which compromises the model's generalization performance on new data. It is important to note that the emergence of this problem is not caused by overfitting but by the degradation of the deep network. To solve the network degradation problem, ResNet introduced the concept of residual learning, incorporating residual blocks through a shortcut mechanism that allows the network to learn residual mappings. This shortcut mechanism is essentially an identity mapping. There are various designs for residual blocks for networks of different depths. Figure 5 shows the basic residual blocks. ResNet greatly increases the depth of network learning and solves the challenge of training deep CNN modules. Furthermore, ResNet has greatly influenced the way researchers think about deep learning, highlighting the importance of network depth and architecture in model performance.

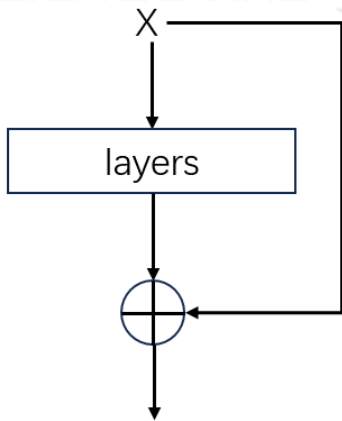


Figure 5: The basic structure of residual blocks (Picture credit: Original).

3 RESULT AND DISCUSSION

Compared to traditional Neural Networks (NNs), CNNs offer several key advantages that make them

particularly effective for image processing tasks. CNNs reduce the number of redundant training parameters and lower model complexity and the risk of overfitting by sharing convolutional kernel parameters. This parameter sharing allows the network to capture spatial hierarchies while maintaining efficiency. Their ability to capture local features in images through strong local perceptual capabilities allows CNNs to extract detailed information with high precision. Additionally, the convolution operation nature means CNNs are relatively invariant to image translations, enhancing their robustness in feature extraction. However, CNNs face limitations in handling irregular or non-grid-like data, which can be challenging for tasks that require processing such data structures. Furthermore, training CNNs, especially on large-scale datasets, demands significant computational resources and time. The advent of AlexNet marked a pivotal moment in deep learning, as it replaced the traditional Sigmoid activation function with the ReLU, thereby alleviating the vanishing gradient problem. AlexNet also implemented Dropout in the fully connected layers to reduce overfitting and utilized GPUs to accelerate computations by distributing tasks across multiple GPUs. As can be seen from Table 1, after ResNet introduced the residual block and solved the gradient vanishing problem caused by too many network layers, the model depth was increased from 8 layers in 2012 to 152 layers in 2015, and the accuracy rate also increased from 84.6% to 96.43%.

Table 1: Accuracy for ImageNet challenge various models.

Model	Year	Accuracy (%)	Layers
AlexNet	2012	84.6	8
VGGNet	2014	92.7	19
ResNet	2015	96.43	152

ResNet further advanced the field by addressing issues related to vanishing or exploding gradients in very deep networks. The introduction of residual blocks enabled ResNet to maintain high accuracy despite an increased number of layers. Despite these advancements, ResNet also demands extensive training data and computational resources due to its large number of parameters. Looking ahead, future research in deep learning will likely focus on developing methods for efficient model training with limited data and enhancing the performance and efficiency of model training. Innovations in this area will be crucial for advancing CNN technology and

expanding its applicability to a broader range of tasks and datasets.

4 CONCLUSIONS

This paper introduces CNNs and explores two seminal models: AlexNet and ResNet. CNNs, a type of feedforward neural network, consist of convolutional layers, pooling layers, and fully connected layers, and represent a significant advancement in deep learning algorithms. AlexNet pioneered the use of ReLU and Dropout, addressing the vanishing gradient problem and reducing overfitting. ResNet further advanced the field by introducing residual blocks, allowing the effective training of much deeper networks. The development of CNNs has led to substantial progress and broad applications across various fields. However, challenges remain, such as the need for extensive data and substantial computational resources for training. These limitations highlight areas for future research and technological development. Future work will focus on exploring more foundational CNN models, gaining a deeper understanding of their mechanisms, and addressing their strengths and weaknesses. Additionally, research will investigate how deep learning can integrate insights from other fields to innovate and transform everyday life.

REFERENCES

- AT&T., 2017. Database of Faces: ORL face database. Retreid from: <http://cam-orl.co.uk/facedatabase.html>.
- Beham, M.P., Roomi, S.M.M., 2013. A review of face recognition methods. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(04), 1356005.
- Jeong, M., Ko, B.C., 2018. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12), 4270.
- Kortli, Y., Jridi, M., Al Falou, A., et al. 2020. Face recognition systems: A survey. *Sensors*, 20(2), 342.
- Li, Z., Zhang, T., Jing, X., et al. 2021. Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes. *Alexandria Engineering Journal*, 60(1), 1411-1420.
- Liu, F., Chen, D., Wang, F., et al. 2023. Deep learning based single sample face recognition: a survey. *Artificial Intelligence Review*, 56(3), 2723-2748.
- Nan, Y., Ju, J., Hua, Q., et al. 2022. A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, 61(6), 4435-4444.
- Nguyen, K., Proença, H., Alonso-Fernandez, F., 2024. Deep learning for iris recognition: A survey. *ACM Computing Surveys*, 56(9), 1-35.
- Qinjun, L., Tianwei, C., Yan, Z., et al. 2023. Facial Recognition Technology: A Comprehensive Overview. *Academic Journal of Computing & Information Science*, 6(7), 15-26.
- Wang, M., Deng, W., 2021. Deep face recognition: A survey. *Neurocomputing*, 429, 215-244.