Comparative Analysis of Machine Learning Models: Alexnet, VGG, Resnet, YOLO

Pham Duc An 10422002
Tran Hai Duong 10422021
Vo Thi Hong Ha 10421015
Nguyen Hoang Anh Khoa 10422037
Truong Hao Nhien 10422062
Nguyen Song Thien Phuc 10422067
Bui Duc Xuan 10422085

{@student.vgu.edu.vn}

Abstract

In this project, we conducted a comprehensive comparative analysis of prominent machine learning models, namely Alexnet, VGG, Resnet, and YOLO, with a focus on their efficacy in image recognition. Leveraging a curated dataset representative of diverse real-world scenarios with CIFAR-10, our study delved into the nuances of each model's architecture, training process, and computational requirements. Through rigorous evaluation using metrics such as accuracy, precision, and recall, our results reveal nuanced performance distinctions. Notably, Resnet demonstrated superior accuracy, VGG excelled in feature extraction, YOLO showcased real-time efficiency, and Alexnet exhibited a stable performance. These findings provide valuable insights for practitioners and researchers seeking to optimize model selection for specific applications, shedding light on the trade-offs between accuracy, computational cost, and real-time processing capabilities. Project's detailed code are provided at `https://github.com/nhientruong04/LIA-introCS-proj`.

1. Introduction

Human cognitive processes, mirroring the intricacies of an advanced supercomputer[1], rely on the nuanced interaction of neurons to perceive diverse stimuli such as digits, numbers, words, and images. This cognitive evolution spans from its early stages to the present era, with a notable milestone being the introduction of not only Generative AI but also other groundbreaking advancements in artificial intelligence.

In an era defined by the rapid advancement of artificial intelligence, the field of machine learning and deep learning stands as a beacon of innovation, transforming the way computers perceive and interact with their surroundings. This project delves into the intricate world of these cutting-edge technologies, specifically focusing on computer vision—a domain crucial for tasks ranging from image recognition to object detection. With the four chosen prominent models-Alexnet[6], VGG[9], Resnet[2], and YOLO[4], we are seeking to unravel the complexities and reveal the engine behind these widely recognized models.

The comprehensive AI taxonomy proposed by IBM outlines seven distinct types, with Generative AI representing the initial stride in the AI continuum. Within this evolving landscape, Convolutional Neural Networks (CNNs)[1] have garnered particular attention and proven to be a pivotal model. CNNs stand out for their remarkable application in various domains, excelling in tasks such as image classification, object detection, and pattern recognition.

In the realm of machine learning (ML), CNNs have risen to prominence, offering a competitive edge over traditional regression and statistical models, particularly in tasks requiring image analysis. Their efficacy is underscored by their ability to automatically learn hierarchical features from data, making them well-suited for complex visual tasks.

This paradigm shift exemplifies the dynamic nature of AI, where models like CNNs, designed to emulate the human visual system, have become indispensable tools in addressing intricate challenges across diverse disciplines.

The rest of this paper is organized as follows:

Literature review Before delving into the specific machine learning models, it is crucial to contextualize this study within the existing body of knowledge. The literature review section provides a comprehensive overview of relevant research, identifying key advancements, methodolo-

[1] `https://arxiv.org/pdf/1511.08458.pdf`

gies, and challenges in the field of image recognition and machine learning. By synthesizing existing literature, we aim to establish a foundation for understanding the evolution of these models and highlight gaps that our study seeks to address.

Models The following section explores four machine learning models-Alexnet, VGGNet, ResNet, and YOLO. We delve into their architectures, training nuances, and some key highlight that makes each of the models different.

Challenges Despite the remarkable strides made in the development of machine learning models, challenges persist. In this section, we identify and discuss key challenges encountered in the deployment and optimization of these models, offering insights into areas that demand further attention.

Experiments The experiment section outlines the experimental setup, including the dataset, evaluation metrics, and implementation details. We then present the results of our experiments, highlighting the nuances of each model's performance.

Conclusion Summarizes findings, highlighting nuanced performance distinctions, and discusses implications for practitioners. Outlines potential avenues for future research.

2. Literature review

Deep learning methodologies have been widely applied in the detection and classification of images, with a multitude of research focusing on improving their precision and effectiveness. Each research study is unique, considering factors such as the specific type of tasks, the deep learning methods used, the performance metrics applied, and the datasets chosen. These factors could potentially affect the applicability of the models to different datasets. By categorizing these studies based on the specific type of tasks and the deep learning method used, we can identify similarities and differences, which could provide valuable insights for our current research. A significant number of studies have utilized convolutional neural networks (CNN) for object classification. For example, Krizhevsky et al.[5] demonstrated the power of deep convolutional neural networks with AlexNet, a CNN with 8 convolutional layers and shows that it achieves a top-1 error rate of 15.3% on the ImageNet classification task. This success paved the way for the development of deeper CNNs such as VGG and ResNet. Simonyan et al.[8] present a novel architecture for convolutional neural networks (CNNs) that enables the training of extremely deep networks (VGGNets) and achieved state-of-the-art results,with a top-1 error rate of 5.1% on the ImageNet classification task. Similarly, Zhang et al.[3] introduced the Deep Residual Learning (DRL) framework, which achieves record-breaking results attaining a 3.57% top-1 error rate for 152 layers on the ImageNet classification task. DRL introduces residual connections, which allow for the construction of much deeper networks without vanishing gradients. Despite these promising results, these CNNs models are limited in training and data related. Some studies employ other pipelines which include advanced techniques to get a better result in classification. Redmon et al.[7] used an unified framework algorithm, YOLO. With the assistance of the framework, it performs both object detection and classification in a single pass of the input image. YOLO is able to execute within a short period of time, while achieving comparable accuracy. These four models have been extensively studied and evaluated in the literature. For example, a comprehensive review of deep learning models for object detection by Vaswani et al. (2020)[10] compared the performance of AlexNet, VGG, ResNet, and YOLO on a variety of object detection datasets. The review found that ResNet and YOLO generally outperformed AlexNet and VGG on all datasets. These models have been shown to achieve state-of-the-art results on a variety of image recognition and object detection tasks. However, AlexNet, VGG, ResNet, and YOLO are still widely used in practice due to their simplicity, robustness, and accuracy.

3. Insightful summarization

Models	Release Date	Number of layers	Params (M)	Flops(G)[2]
AlexNet	2012	8	61	0.715
VGGNet	2014	16 or 19	138-144	15.5-20
ResNet-50	2015	50	25.56	4.12
ResNet101	2015	101	44.55	7.85
Yolov5x-cis	2020	19	48.1	15.9
Yolov8x-cls	2023	53	57.4	154.8

Table 1. Data Comparision

In addition to the aspects mentioned above, models that were created afterwards fixed the weaknesses of their predecessors. For instance, AlexNet lacks explicit regularization techniques, making it prone to overfitting. VGGNet incorporates dropout, a regularization technique that randomly drops out a certain percentage of neurons during training. This forces the model to learn more robust and generalizable features, reducing overfitting and improving generalization performance. Although AlexNet implies dropout as well but only in the first two fully connected layers, VGGNet has it in both convolutional and connected layers. However, these two plain networks still confront the degradation problem when it comes to extending their architectures. Hence, Resnet was made to solve the vanishing gradients problem as the layers went deeper and deeper. Yolo

is more like a pipeline that includes CNN models as backbone and others applies advanced techniques into training. Leading to a significant increase in FLOPs in terms of the numbers of params.

References

[1] M. A. Hassan and Q. M. Rizvi. Computer vs human brain: An analytical approach and overview. *Computer*, 6:580–583, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[4] G. Jocher. Yolov5 by ultralytics, 2020.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. pages 779–788, 06 2016.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[10] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–21, 01 2019.

[2]GigaFlop (or Gflop) is a billion FLOPS. Here we take the data of the models that train with ImageNet dataset