Comparative Analysis of CNN Architectures for Image Classification Tasks

Lucas Teixeira Rocha - m11813
Computer Science and Engineering - 2nd Cycle Degree
Universidade da Beira Interior
Covilhã, Portugal
lucas.rocha@ubi.pt

***Abstract*—This report provides a comprehensive evaluation of several Convolutional Neural Network (CNN) architectures for image classification tasks, specifically focusing on identifying individuals, recognizing facial expressions, determining gender, and detecting the presence of glasses. Utilizing a robust dataset of images with diverse features, we compared the performance of a tailor-made CNN, VGG16, and ResNet50 architectures. Our approach involved training models from scratch as well as applying fine-tuning techniques to pre-trained networks. The results reveal intriguing insights into the architectures' efficiencies, with particular attention to loss metrics, accuracy, and training duration. These findings offer valuable implications for the selection and optimization of CNNs in practical image classification scenarios.**

I. Introduction

Machine learning, especially the use that involves image processing, has been making huge strides in a variety of fields. Among the tools used for this, Convolutional Neural Networks (CNNs) stand out. They're really good at understanding images, and thanks to resources like Aurélien Géron's book "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" [1], more people can now build and use these tools for their projects.

In this report, we take a closer look at three types of CNNs: a custom-made one, the well-known VGG16, and the more recent ResNet50. We trained these models to recognize different features in images, like who's in the picture, their facial expression, their gender, and whether they're wearing glasses. The goal was to see how well each model did in terms of accuracy and how long they took to train. We also wanted to see if starting from scratch or tweaking models that were already trained on imagenet images (a process known as 'fine-tuning') made any difference. The findings from this study should help us figure out better ways to use CNNs for classifying images.

II. Methodology

The methodology employed in this study was meticulously designed to evaluate the performance of various CNN architectures for image classification. The dataset used, referred to as "AR.zip", consists of 3315 images of 136 subjects with various lighting conditions, facial expressions, and accessories, providing a comprehensive set of features for analysis.

A. Dataset Preparation

The images were first standardized to a uniform size and color space to facilitate consistent processing. We then divided the dataset into training, validation, and testing subsets. This split was crucial to not only train and fine-tune the models but also to independently evaluate their performance.

B. Model Implementation

Three distinct CNN architectures were implemented:

- A custom CNN (MyCNN), developed from the "src_deep_learning_keras.py" script provided.
- The VGG16 architecture, a well-known model with a deep structure renowned for its performance on the ImageNet dataset.
- The ResNet50 model, characterized by its residual learning framework to ease the training of deeper networks.

All models were trained using both 'from scratch' and 'fine-tuning' approaches. 'From scratch' involved random initialization of weights, while 'fine-tuning' leveraged weights from models pre-trained on the ImageNet dataset.

C. Training Process

During training, the 'train_in_batch()' function or data generators were used to handle the large volume of data efficiently. For each model and task, we employed a consistent set of hyperparameters and training epochs to ensure comparability.

D. Evaluation Metrics

The effectiveness of each model was measured using three metrics:

- Loss: The mean squared error between the predicted and actual values.
- Accuracy: The proportion of correct predictions to total predictions.
- Training Time: The total time taken to train the model, measured in seconds.

E. Statistical Analysis

Finally, we performed a statistical analysis of the results to determine the significance of the differences observed between the models and training methods.

The methodology outlined ensured a fair and rigorous evaluation of each CNN's ability to classify images based on the given features. It also allowed for the analysis of the impact of different training paradigms on the model's learning efficiency and predictive performance.

III. Results

The results indicate that our custom CNN generally performed well, especially in gender and glasses detection tasks. ResNet50 showed significant improvement with fine-tuning, particularly in ID prediction and expression recognition. VGG16's performance was enhanced slightly by fine-tuning but was not as notable.

Feature	Model	Loss	Accuracy	Time (seconds)
	ID Prediction			
	MyCNN	1.7360	53.73%	288.41
	VGG16 (From Scratch)	0.0394	1.55%	1547.62
	VGG16 (Fine Tuning)	0.0339	12.59%	1366.75
	ResNet50 (From Scratch)	0.0386	2.18%	760.58
	ResNet50 (Fine Tuning)	0.0123	80.13%	973.87
	Expression Prediction			
	MyCNN	0.4826	79.31%	287.64
	VGG16 (From Scratch)	0.3373	69.75%	1528.28
	VGG16 (Fine Tuning)	0.2271	68.93%	1344.19
	ResNet50 (From Scratch)	0.1643	76.18%	844.77
	ResNet50 (Fine Tuning)	0.538	93.99%	903.96
	Gender Prediction			
	MyCNN	0.0187	99.42%	288.25
	VGG16 (From Scratch)	0.6882	55.34%	1536.72
	VGG16 (Fine Tuning)	0.6893	54.49%	1284.25
	ResNet50 (From Scratch)	0.1602	94.87%	688.94
	ResNet50 (Fine Tuning)	0.0486	98.45%	771.06
	Glasses Prediction			
	MyCNN	0.0009	99.97%	316.85
	VGG16 (From Scratch)	0.5423	76.88%	1416.26
	VGG16 (Fine Tuning)	0.6202	76.40%	1266.38
	ResNet50 (From Scratch)	0.0582	98.76%	649.05
	ResNet50 (Fine Tuning)	0.1595	90.23%	719.71

TABLE I
MODEL PERFORMANCE FOR FEATURE PREDICTION TASKS

IV. Discussion

This study's exploration into the capabilities of different CNN architectures for image classification tasks has yielded illuminating results. The custom CNN model, designed specifically for this research, demonstrated remarkable effectiveness, particularly in glasses detection and gender classification tasks, where it achieved near-perfect accuracy. These results underscore the potential benefits of tailoring CNN architectures to the nuances of the dataset and task at hand [2].

Contrasting the performance of the widely-used VGG16 and ResNet50 models revealed expected trends and surprising deviations. VGG16's lower accuracy, especially when trained from scratch, indicates a possible overfitting to ImageNet features that do not generalize well to our dataset. This could be attributed to the dissimilarities between the image features in the AR dataset used for this project and those in datasets typically used to pre-train models like VGG16.

ResNet50 showed a marked improvement when fine-tuned, which suggests that its deeper architecture captures more generalizable features that, when fine-tuned, adapt effectively to new tasks. This aligns with the model's design philosophy, which emphasizes learning residual functions with reference to the layer inputs, allowing it to perform well even with a significant increase in depth [1].

The time efficiency of MyCNN also invites consideration of the trade-offs between training duration and accuracy. MyCNN's training times were substantially lower compared to VGG16 and ResNet50, raising important questions about the practical implications of using complex pre-trained models versus simpler, customized models, especially when computational resources are limited.

These findings contribute to the ongoing dialogue in machine learning about the balance between model complexity, training time, and performance. They also suggest that for specific applications, custom models can be engineered to outperform established architectures, both in terms of accuracy and efficiency. Future research may delve deeper into the architectural nuances that enable MyCNN to perform so well and explore the scalability of such custom models to other datasets and classification tasks.

V. Conclusion

In conclusion, this study's comparative analysis of CNN architectures for image classification has provided clear evidence that model selection and training approaches significantly influence performance. The custom CNN model, while less complex than its counterparts, displayed superior efficiency and effectiveness in several tasks, underscoring the value of task-specific model development. VGG16 and ResNet50, with their deep architectures, also demonstrated their potential, particularly when fine-tuned, which is in line with the principles outlined in recent literature on deep learning such as "Deep Learning" by Goodfellow, Bengio, and Courville [3].

The nuanced performance differences revealed between 'learning from scratch' and 'fine-tuning' methodologies highlight the complexity of model training in the field of machine learning. These insights not only enhance our understanding of CNN architectures but also guide practitioners in selecting and training models for image classification tasks.

As the field of machine learning continues to evolve, studies like this one are vital for informing best practices and driving innovation. Further research may explore the integration of emerging techniques, such as transfer learning and domain adaptation, to expand the versatility and applicability of CNNs across varied and challenging datasets.

References

[1] A. Géro, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd Edition, O'Reilly Media., 2019.

[2] A. Ng, "Machine learning," 2021. Coursera course.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.