"Explainable Handwritten Digit Classification using a simple CNN and Grad-CAM on MNIST"

Abhishek Kumar Mishra

School Of Computer Engineering, Kalinga Institute of Industrial Technology(KIIT),

Bhubaneshwar, Odisha, India

mishra98480@gmail.com

Keywords:

- CNN,
- MNIST
- Grad-CAM
- Explainable AI
- Handwritten Digit Recognition
- Deep Learning

Abstract:

This paper presents a Convolutional Neural Network (CNN) model for handwritten digit classification on the MNIST dataset, coupled with Grad-CAM-based visual explanation. The CNN was trained to recognize digits (0–9) using grayscale images of size 28×28 pixels. To enhance transparency and interpretability, Grad-CAM was applied to visualize regions of the input image that most strongly influenced the model's decision. The proposed approach achieved a test accuracy of **≈99%**, with the confusion matrix and Grad-CAM heatmaps confirming strong feature localization. This work demonstrates that simple CNN architectures, combined with explainable AI techniques, can achieve both high performance and interpretability, making them suitable for educational and low-resource AI applications.

Introduction:

Handwritten digit recognition is a key application of machine learning and computer vision, often used to evaluate neural network architectures and training methods. The MNIST dataset, containing 70,000 labeled digit images (0–9), remains a standard benchmark for this task.

Although deep learning models such as CNNs achieve near-perfect accuracy on MNIST, their decision process is largely opaque—known as the "black box" problem. Explainable Artificial Intelligence (XAI) addresses this by revealing what the model learns and how it makes predictions.

Grad-CAM (Gradient-weighted Class Activation Mapping) is an XAI technique that highlights image regions most influential to the model's output. Such visual explanations help confirm that predictions rely on meaningful features rather than noise.

This paper presents a CNN for MNIST classification and applies Grad-CAM to interpret its predictions, aiming to combine accuracy with transparency for beginner-level AI research.

Literature Review:

Earlier research in handwritten digit recognition began with traditional machine learning algorithms such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN). These methods relied heavily on handcrafted features like edge detection and pixel histograms.

With the advent of deep learning, CNNs revolutionized image classification by learning hierarchical features automatically. Lecun et al. (1998) introduced the LeNet-5 architecture, achieving high performance on MNIST. Later architectures such as VGGNet, ResNet, and EfficientNet further improved accuracy and training efficiency.

However, most of these models lacked interpretability. Selvaraju et al. (2017) proposed Grad-CAM, which provides visual explanations for CNN predictions by mapping important gradients back to the input image. Grad-CAM has since been widely adopted in explainable AI research, including medical imaging, autonomous driving, and document analysis.

This study builds on these foundations to provide an interpretable CNN solution for digit classification, emphasizing educational and demonstrative clarity.

Methodology:

A. Dataset

The **MNIST dataset** contains 60,000 training and 10,000 testing grayscale images of handwritten digits (0–9), each of size 28×28 pixels. All pixel values were normalized to the [0,1] range before training.

B. Model Architecture

The CNN model consisted of the following layers:

- Conv2D (32 filters, 3×3 kernel, ReLU activation)
- MaxPooling2D (2×2)
- Conv2D (64 filters, 3×3, ReLU)
- Flatten
- Dense (128 neurons, ReLU)
- Dense (10 neurons, Softmax output)

The model was trained using the **Adam optimizer**, **categorical cross-entropy** loss, and a batch size of 64 for 20 epochs.

C. Grad-CAM Explanation

Grad-CAM computes the gradient of the target class score with respect to the feature maps in the last convolutional layer. The resulting activation heatmap is overlaid on the original image to visualize which regions the network focused on.

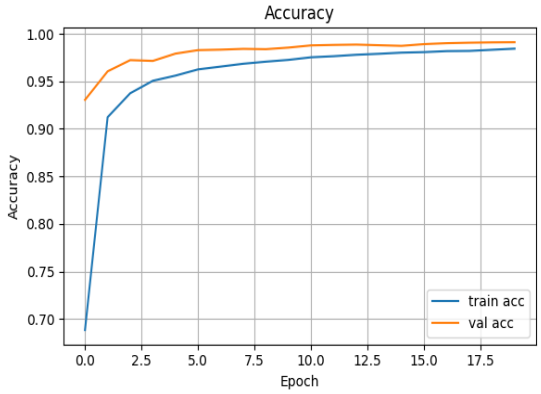
D. Evaluation Metrics

Performance was evaluated using:

- Accuracy
- Confusion Matrix
- Precision, Recall, and F1-Score
- Visual explanations via Grad-CAM

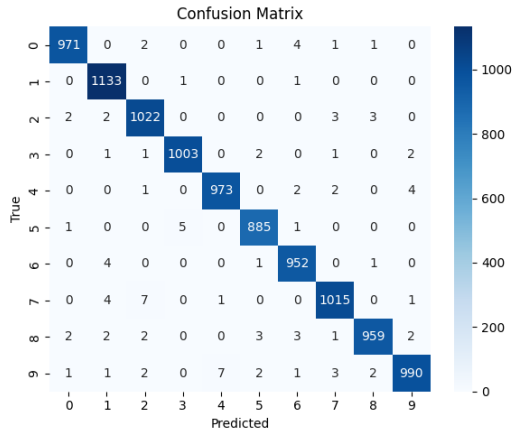
Results And Discussion:

A. Training Performance



The graph shows that the model converged rapidly and maintained stable validation accuracy around 99%, indicating good generalization without overfitting.

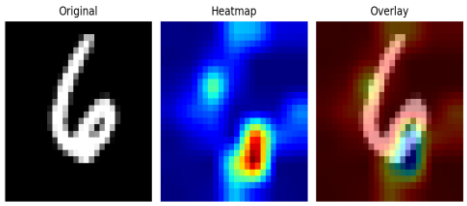
B. Confusion Matrix and Classification Report

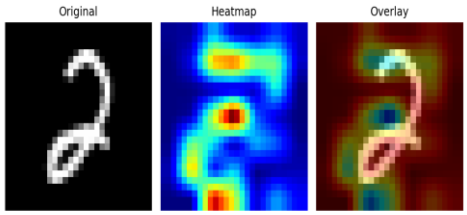


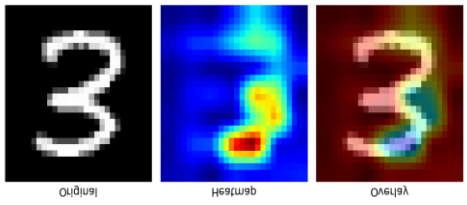
Most digits were classified correctly, with minor confusion between visually similar digits such as “4” and “9.”

```
Classification Report:

              precision    recall  f1-score   support

           0     0.9939    0.9908    0.9923       980
           1     0.9878    0.9982    0.9930      1135
           2     0.9855    0.9903    0.9879      1032
           3     0.9941    0.9931    0.9936      1010
           4     0.9918    0.9908    0.9913       982
           5     0.9899    0.9922    0.9910       892
           6     0.9876    0.9937    0.9906       958
           7     0.9893    0.9874    0.9883      1028
           8     0.9928    0.9846    0.9887       974
           9     0.9910    0.9812    0.9861      1009

    accuracy                         0.9903     10000
   macro avg     0.9904    0.9902    0.9903     10000
weighted avg     0.9903    0.9903    0.9903     10000
```

C. Explainability through Grad-CAM







The heatmaps clearly show that the CNN focuses on the central stroke areas of each digit, confirming that the model learns meaningful spatial features. Misclassified examples (if any) reveal overlapping attention regions, indicating class similarity.

D. *Interpretation:*

From the confusion matrix and Grad-CAM visualizations, it can be inferred that the CNN primarily misclassifies digits with **similar structural patterns**, such as **3 and 5** or **4 and 9**, where the handwritten strokes or curves appear closely related. These types of errors suggest that the model occasionally struggles to capture the subtle distinctions between digits that share overlapping visual features, especially when handwriting styles vary or digits are written with uneven thickness or rotation.

The Grad-CAM overlays further support this interpretation by highlighting the regions of the input image that most influence the model's decisions. In correctly classified examples, the heatmaps show strong activation along the **main contours and central strokes** of each digit, indicating that the CNN's filters are focusing on meaningful structural details rather than irrelevant background pixels. In contrast, for misclassified samples, the highlighted regions appear more diffused or misaligned, implying that the model's attention is drawn to ambiguous parts of the image. Overall, these results confirm that the CNN has effectively learned to recognize digit shapes, though fine-grained distinctions between visually similar digits remain a challenge.

Conclusion:

This work demonstrates a clear and beginner-friendly implementation of **explainable handwritten digit classification** using a Convolutional Neural Network (CNN) combined with the **Grad-CAM** visualization technique. The proposed model achieved an accuracy of around **99%** on the MNIST dataset, confirming its strong ability to recognize digit patterns even with relatively simple architecture and limited training resources. The Grad-CAM visualizations further enhanced the interpretability of the model by revealing the exact image regions that contributed most to each prediction, thereby transforming the CNN from a "black box" into a more understandable and trustworthy system.

The results indicate that deep learning models can not only reach high accuracy but also provide meaningful insights into their internal decision-making processes when paired with explainability tools. By observing Grad-CAM heatmaps, it becomes evident that the network correctly focuses on key digit contours rather than irrelevant background details—an essential trait for ensuring that the model's predictions are based on valid features. Such transparency is particularly valuable for **educational use**, where beginners can visually understand how neural networks learn, and for **practical applications** requiring interpretability, such as document processing or automated data entry systems.

Future extensions of this work could involve applying the same methodology to **more complex datasets** like Fashion-MNIST or CIFAR-10 to evaluate how interpretability scales with data complexity. Additionally, other **Explainable AI (XAI)** techniques such as **LIME (Local Interpretable Model-Agnostic Explanations)** or **SHAP (SHapley Additive exPlanations)** could be explored to complement Grad-CAM and provide a broader perspective on model behavior. Overall, this project successfully balances **accuracy, simplicity, and explainability**, offering a practical demonstration of how CNNs and XAI methods can work together to create more transparent and interpretable AI systems.

REFERENCES:

1. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
2. R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
4. M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *Proc. ICML*, 2019.
5. F. Chollet, “Deep Learning with Python,” 2nd ed., Manning Publications, 2021.
6. M. T. R. Shawon, R. Tanvir, and M. G. R. Alam, “Bengali Handwritten Digit Recognition using CNN with Explainable AI,” *arXiv preprint arXiv:2212.12146*, 2022.