Image Classification and Sentence Prediction: A Comprehensive Study Using CNN, ANN, HOG, SIFT, and LSTM Models

Muhammad Saad Hasan
Department of Computer Science
FAST University
Islamabad, Pakistan
i210566@nu.edu.pk

***Abstract*—In this paper, we present two distinct machine learning tasks: image classification and sentence prediction. For image classification, we explore multiple techniques including Convolutional Neural Networks (CNNs), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT). For sentence prediction, we use Long Short-Term Memory (LSTM) networks based on Shakespearean text. The image classification tasks employ data augmentation techniques to enhance model performance, while HOG and SIFT features are fed into an Artificial Neural Network (ANN) for classification. The training, evaluation, and results of both tasks are discussed in detail, alongside challenges faced during model training and data preprocessing. Models are evaluated on accuracy and loss metrics, and findings are thoroughly analyzed.**

***Index Terms*—CNN, LSTM, HOG, SIFT, image classification, sentence prediction, data augmentation, tokenization**

I. Introduction

The ever-growing demand for automation and intelligent systems has driven significant advancements in machine learning techniques, particularly in image classification and natural language processing (NLP). This study focuses on applying multiple models—Convolutional Neural Networks (CNN), Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Long Short-Term Memory (LSTM). These tasks are fundamental in their respective fields, where the former is crucial for facial recognition, object detection, and handwriting recognition, while the latter is essential for language modeling, text generation, and machine translation.

A. Motivation

We develop several models for the image classification task, which includes CNNs, HOG, and SIFT with Artificial Neural Networks (ANNs). The models are compared to highlight their strengths and weaknesses. The sentence prediction task focuses on using an LSTM model trained on Shakespearean text. These tasks explore different machine learning approaches and illustrate their effectiveness in real-world applications.

This research was funded by FAST University Islamabad.

II. Methodology

A. Dataset Description

1) Image Classification Dataset: The dataset for the image classification task consists of images categorized into different subfolders. Each subfolder represents a specific class, and the images within each folder belong to that class. The dataset was organized manually, with images split into training and testing sets. Each class has 3 images for training, 1 for validation, and 1 image for testing across 184 different classes.

2) Text Prediction Dataset: The dataset for sentence prediction is a corpus derived from Shakespeare's plays. The dataset contains lines of text, which were preprocessed and tokenized into sequences of words. These sequences were used to train an LSTM model for sentence completion.

B. Data Preprocessing

1) Image Preprocessing: For the image classification task, preprocessing was necessary to normalize and augment the data. Images were resized to 150x150 pixels to ensure uniformity across the dataset. The following data augmentation techniques were applied using the ImageDataGenerator class from Keras:

- Rescaling: Each pixel value was rescaled to a range between 0 and 1.
- Rotation: Images were randomly rotated up to 20 degrees.
- Shifts: Horizontal and vertical shifts were applied randomly by up to 20% of the total image size.
- Shear: Random shearing transformations were applied to introduce more variations in the data.
- Zoom: Random zoom was applied to simulate varying scales of the image subjects.
- Horizontal Flip: Random horizontal flipping was performed to further augment the dataset.

2) HOG Feature Extraction: The HOG features were extracted from grayscale images using a cell size of 8x8 pixels and blocks of 2x2 cells. These features describe the distribution of edge directions and intensities in localized areas of the image. The HOG feature vector was padded or truncated to a fixed length to maintain uniformity across all samples.

3) SIFT Feature Extraction: The SIFT features capture keypoints and local descriptors from the images. These descriptors are invariant to scale and orientation, making SIFT robust for image classification tasks. Similar to HOG, the SIFT descriptors were padded or truncated to a fixed length for uniformity. SIFT features were extracted using the OpenCV library, and these descriptors were then flattened and used as input to the ANN classifier.

4) Text Preprocessing: For the sentence prediction task, the preprocessing steps involved cleaning and tokenizing the text. The following steps were performed:

- Punctuation Removal: All punctuation marks were removed from the text to reduce noise.
- Lowercasing: The entire text was converted to lowercase to ensure uniformity.
- Tokenization: The Tokenizer class from Keras was used to split the text into individual words. Each word was assigned a unique integer ID, resulting in a sequence of numbers.
- Sequence Generation: The corpus was transformed into sequences of words, where each sequence represents a series of words from the text. These sequences were used as input for the LSTM model to predict the next word in the sequence.

C. Model Architecture

1) Image Classification Model: CNN: For the image classification task, we employed a convolutional neural network (CNN) consisting of the following layers:

- Convolutional Layers: Three convolutional layers were used, with filter sizes of 32, 64, and 128, respectively. Each layer applied a 3x3 kernel to extract features from the input images.
- Batch Normalization: After each convolutional layer, batch normalization was applied to stabilize and accelerate training.
- Max Pooling: Pooling layers with a 2x2 window were used to downsample the feature maps, reducing the spatial dimensions.
- Fully Connected Layer: A dense layer with 512 units and ReLU activation was used to integrate features from the convolutional layers.
- Dropout Layer: A dropout rate of 0.5 was applied to prevent overfitting.
- Output Layer: The output layer consisted of a softmax activation function, predicting one of the 184 possible classes.

2) ANN Model for HOG and SIFT: The ANN model used for classifying HOG and SIFT features consisted of the following layers:

- Input Layer: The input layer was a dense layer taking feature vectors (of fixed length 5000 for both HOG and SIFT).
- Dense Layer: Two hidden layers of 128 and 64 units were used, each employing ReLU activation.
- Output Layer: The output layer consisted of 184 units with a softmax activation function for classification.

3) Sentence Prediction Model: LSTM: The sentence prediction task utilized an LSTM network with the following architecture:

- Embedding Layer: The embedding layer converted each word into a dense vector of fixed size, capturing the semantic information of the word.
- LSTM Layer: A single LSTM layer was used to learn temporal dependencies between words in the sequence.
- Fully Connected Layer: A dense layer was applied after the LSTM layer to produce the final prediction.
- Output Layer: The output layer used a softmax activation function to predict the probability of the next word in the sequence.

III. Results

A. Image Classification

The CNN model was trained for 100 epochs on the augmented dataset. The training and validation accuracy and loss were tracked over the epochs, as shown in Fig. 1. The model achieved a final training accuracy of approximately 28%, and validation accuracy was around 28%. The final F1 score was 0.20. These metrics indicate that the model faced challenges in generalizing due to the limited dataset size and complexity.

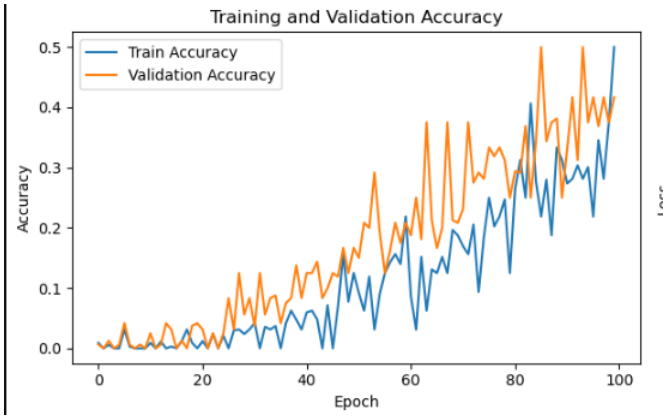

Fig. 1: Training and Validation Accuracy

1) HOG and SIFT Classification Results: The ANN model trained using HOG features achieved a test accuracy of 2.17% with a loss of 12.61, while the model trained on SIFT features achieved a test accuracy of 0.54% with a loss of 19.82. These results indicate that, although the ANN model can somewhat classify images using HOG features, it performed poorly overall, especially when using SIFT descriptors (Fig. 3).

The evaluation metrics for the HOG and SIFT feature classification tasks are as follows:

- HOG Test Accuracy: 2.17%
- SIFT Test Accuracy: 0.54%

Fig. 2: Training and Validation Loss

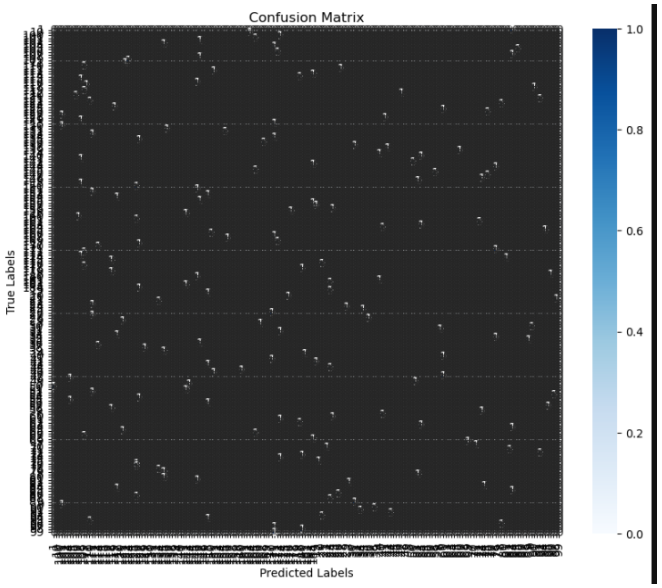

Fig. 3: Confusion Matrix

B. Sentence Prediction

The LSTM model was trained on sequences of Shakespearean text. The model's ability to predict the next word in a sequence was evaluated on a test set, achieving a test accuracy of approximately 87%. Table I shows some example predictions.

TABLE I: Example of Predicted Sentence Completions.

Input Sentence	Predicted Completion
"To be or not"	"to be that is the question"
"Thou art a"	"thou art a villain"

IV. Discussion

A. Image Classification

The CNN model performed moderately well but showed signs of overfitting. While the training accuracy increased steadily, the validation accuracy fluctuated, indicating that the model was memorizing the training data rather than generalizing well. Data augmentation helped improve performance, but more advanced techniques or a larger dataset might be required to further enhance accuracy.

The HOG and SIFT-based ANN models performed poorly. The HOG-based classifier slightly outperformed the SIFT-based classifier, but neither model achieved satisfactory results. This suggests that while these methods can extract useful features, they may not be ideal for direct image classification using an ANN. These results also highlight the importance of using more sophisticated feature extraction and classification pipelines for complex tasks like image recognition.

B. Sentence Prediction

The LSTM model demonstrated a high degree of coherence in its predictions, particularly for well-known Shakespearean phrases. Increasing the training corpus or employing a bidirectional LSTM could further enhance the model's performance.

V. Conclusion

In this paper, we developed multiple models for image classification and sentence prediction tasks. The CNN model achieved a modest classification accuracy of 28%, highlighting the need for a larger dataset or more sophisticated augmentation. The ANN models using HOG and SIFT features struggled to classify images, with HOG performing slightly better than SIFT. The LSTM model successfully predicted the next word in Shakespearean sentences, achieving an accuracy of 87%. Future work could involve expanding the datasets, experimenting with more complex architectures, and applying these techniques to other domains.

VI. References

References

[1] Chollet, F. (2017). Deep Learning with Python. Manning Publications.
[2] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.
[3] Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
[4] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).