Xiaoming Huang
Jun 11, 2024
CSCI 184

Machine Learning for Alphanumeric Character Classification

Introduction:

Handwritten character recognition is an important area of pattern recognition and computer vision that aims to convert handwritten text into a machine-readable format. This capability has a wide range of applications, including document digitization, automatic form processing, and educational tools. Different people have unique handwriting styles, which can be difficult to accurately recognize. This project aims to employ various machine learning algorithms to classify handwritten characters, with a focus on converting handwritten math notes into printed versions to aid in educational settings.

Background Studies:

1. "Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey":

In the first research paper, the authors described that Handwritten text recognition (HTR) plays a vital role in the process of digitizing historical documents, converting them into accessible digital formats. Their investigation addresses key challenges and recent developments at the HTR, with a particular focus on French documents such as the Balfour Civil Birth Register. The records date back to the French Revolution. They are valuable for historical research but have been difficult to digitize due to varied handwriting, overlapping characters and marginalia. The survey classifies various HTR systems based on the technology employed, data set, year of publication, and level of identification. Early HTR methods mainly used Hidden Markov Models (HMM), but had limitations such as memorylessness and the necessity of manual feature selection.

To overcome these issues, more advanced methods are employed to combine HMMs with Gaussian mixtures, convolutional neural networks (CNN), and recurrent neural networks (RNN). These show significant improvement on recognition accuracy. Contemporary HTR systems utilize sophisticated machine learning techniques to analyze and identify complex document layouts, letters, lines of text, paragraphs, and complete documents. Techniques such as CNNs, convolutional recurrent neural networks (CRNN), and gated CNNs are shown to be particularly effective.

However, there are still challenges remaining due to inherent bias in handwriting style, document quality, noise, and text alignment issues. This article highlights the historical importance of the Balfour Civil Register and highlights the need for efficient transcription methods. Manual transcription is laborious and expensive, while automated transcription using advanced HTR systems offers a more feasible solution. The survey also examines the various datasets used to train and evaluate HTR systems, highlighting datasets presented at major

conferences such as the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR).

We can see that this research provides an up-to-date overview of the state of the art in HTR, identifies key challenges, and suggests directions for future research. The ultimate goal is to increase the accuracy and efficiency of HTR systems, and as a result improve the digitization of handwritten documents. This advancement can therefore be critical to preserving historical records and promoting scholarly research.

2. "Handwritten Character Recognition Using Machine Learning":

The second paper discusses the specific application of machine learning techniques for the complex task of handwritten character recognition. Handwritten character recognition involves identifying and classifying handwritten characters to convert them into digital formats. It is also emphasized here, again, that it is essential for applications like document digitization, automated form processing, and intelligent handwriting analysis.

Traditional character recognition methods have largely been supplanted by machine learning algorithms capable of learning from large datasets and making accurate predictions. The authors present a methodology that utilizes CNN. CNNs can automatically learn features from images, making them particularly suitable for recognizing handwritten characters. The paper details the dataset preparation process, which includes resizing, normalizing, and preprocessing images of handwritten characters to improve recognition accuracy. The study compares the performance of various machine learning models, including Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN), but primarily focuses on the superior performance of CNNs. It shows that the CNN model achieved an impressive accuracy of 98.6% on the standard MNIST dataset, which contains images of handwritten digits. The authors discuss the architecture of the CNN model, specifically on the use of convolutional layers for feature extraction and pooling layers for dimensionality reduction.

The paper also points out the importance of preprocessing steps such as noise removal, binarization, and morphological operations to enhance image quality and recognition accuracy. It also highlights the use of libraries and tools like TensorFlow, Pandas, and NumPy for implementing and optimizing the machine learning models.

Data Preprocessing:

The dataset for this project consists of images of handwritten English characters, extracted from a compressed .7z file. It comes with a csv file with the one feature of the image filename and a target of the corresponding label. To prepare the images for model input, each image is resized to 64x64 pixels, ensuring uniformity and normalized to scale the pixel values between 0 and 1. And each sample is turned into numerical format: an 64 by 64 array. This preprocessing step is crucial as it enhances the consistency and performance of the machine learning models. Subsequently, the dataset is divided into training and testing sets to ensure robust evaluation of the models on unseen data, thereby preventing overfitting and enabling accurate performance assessment.

Experimental Design:

The primary objective of this project is to evaluate the efficacy of various machine learning algorithms—k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN)—in recognizing handwritten characters.

1. k-Nearest Neighbors (k-NN): This non-parametric algorithm classifies data points based on the majority vote of their nearest neighbors. The choice of the parameter 'k', which denotes the number of neighbors, is critical as it directly influences the model's accuracy and generalization ability. A smaller 'k' may lead to overfitting, while a larger 'k' may smooth out the decision boundary too much.

2. Support Vector Machines (SVM): SVMs are supervised learning models that find the optimal hyperplane in a high-dimensional space to separate different classes. The effectiveness of SVMs heavily relies on the selection of the kernel function, with options including linear, polynomial, and radial basis functions (RBF). Each kernel can handle different types of data structures, with RBF often being effective in handling non-linear separations.

3. Convolutional Neural Networks (CNN): CNNs are specialized deep learning architectures particularly adept at image recognition tasks. They consist of multiple layers, including convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. CNNs automatically learn spatial hierarchies of features from input images, making them exceptionally powerful for complex pattern recognition tasks.

Implementation:

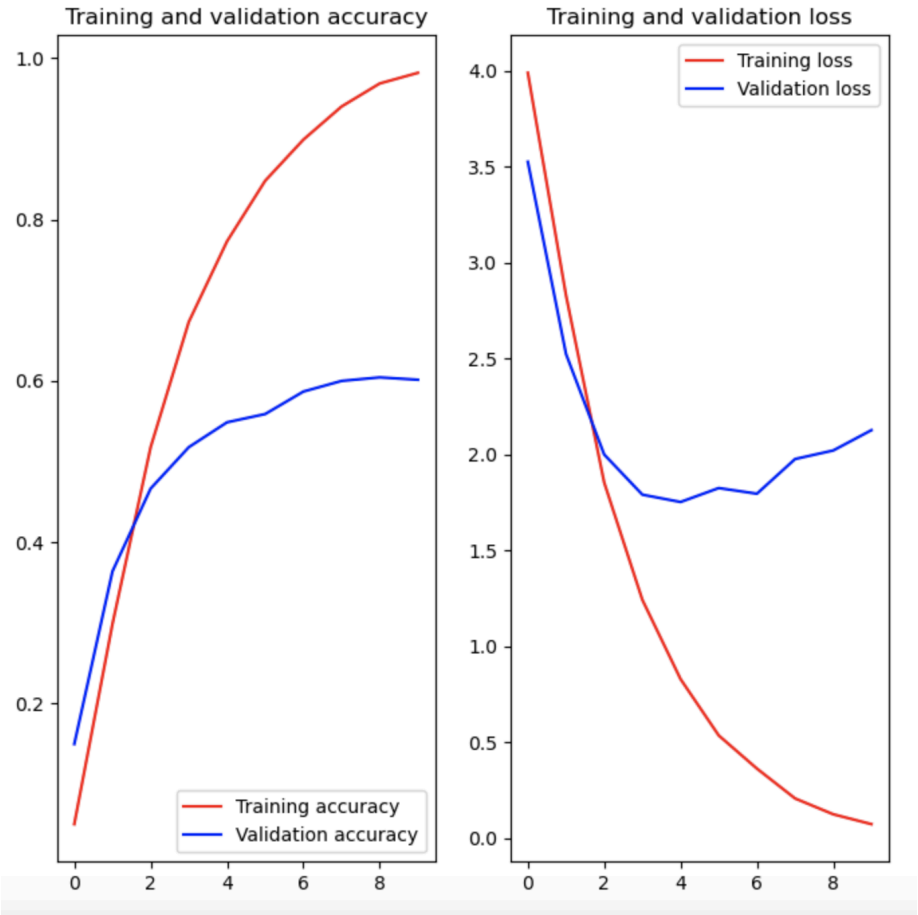
The implementation of each model involves specific steps tailored to their unique architectures:

1. SVM: The SVM model is trained using flattened image data, where each image is represented as a one-dimensional array of pixel values. Post-training, the model's performance is assessed using accuracy metrics, classification reports, and confusion matrices to understand the types of misclassifications. The SVM achieved an accuracy of approximately 39.7%.

2. k-NN: Similarly, the k-NN model is trained on flattened image data. The number of neighbors (k) is set to 3, and the model's performance is evaluated using accuracy and other relevant metrics. The choice of k is based on cross-validation results to optimize performance. The k-NN model attained an accuracy of around 35.2%.

3. CNN: The CNN model involves a more intricate architecture, including convolutional layers for extracting local patterns and pooling layers to reduce spatial dimensions. The images are reshaped into a four-dimensional array to fit the CNN input requirements. The model is trained to minimize categorical cross-entropy loss, and its performance is

visualized through training and validation accuracy and loss plots over epochs. The CNN model achieved an accuracy of approximately 60.1%.



Results and Evaluation:

The performance of each model is meticulously evaluated using various metrics:

1. SVM Results: The SVM model achieved an accuracy of approximately 39.7%. Detailed classification reports and confusion matrices reveal the types of misclassifications and provide insights into areas for potential improvement.

2. k-NN Results: The k-NN model attained an accuracy of around 35.2%. Similar to the SVM, detailed evaluation metrics offer a comprehensive analysis of the model's performance.

3. CNN Results: The CNN model stands out with an accuracy of approximately 60.1%. The training and validation accuracy and loss plots indicate a clear trend of improvement, showcasing the model's learning capability and effectiveness.

Analysis and Future Improvement:

The superior performance of the CNN model highlights the advantages of deep learning in pattern recognition tasks. Compared with traditional methods such as SVM and k-NN, CNN is able to automatically learn and extract hierarchical feature representations from images, achieving higher accuracy. This project highlights the importance of choosing the right algorithm based on the complexity of the task and the nature of the data. While SVM and k-NN work well for simpler tasks or smaller datasets, CNN excels in complex scenarios involving large amounts of image data. Although CNN achieves higher accuracy, the overall accuracy level of all models (39.7% for SVM, 35.2% for k-NN, and 60.1% for CNN) indicates that there is still much room for improvement.

Several factors may contribute to the relatively low accuracy of the models. First, the complexity and variability of handwritten characters poses a huge challenge. The style, spacing, and size of handwritten text can vary greatly, making it difficult for the model to generalize well. In addition, the size and quality of the dataset can also affect the performance. A limited dataset may not capture a wide range of handwriting styles, leading to overfitting and poor generalization. Insufficient preprocessing steps can also introduce noise and distortion, which can affect model performance.

Future improvements can focus on expanding and diversifying the dataset to cover a wider range of handwriting styles and conditions. Data augmentation techniques such as rotation, scaling, and adding noise to training images can help create more robust models. Advanced CNN architectures such as deeper networks or hybrid models that combine CNNs with other techniques such as recurrent neural networks (RNNs) can further improve performance. In addition, fine-tuning hyperparameters and exploring different preprocessing methods can achieve better feature extraction and overall model accuracy. By addressing these areas, the effectiveness of handwritten character recognition systems can be significantly improved, making them more reliable in real-world applications.