Classification and Separation of Images Received in File Sharing Applications Using Machine Learning

Koripelli Kartheesh Reddy(✉), Sai Akhil Podduturi, and G. Vanitha

Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India
kartheeshreddy22@gmail.com, gvanitha_cse@cbit.ac.in

Abstract. Nowadays, applications like WhatsApp, Telegram etc. are widely used for sharing images by students during their academics. The images that are shared through these applications include printed notes, handwritten notes, question papers, circulars, marksheets and many more. In due course of time, the storage capacity of mobile phones gets filled up. All these images are stored in a single media folder and get mixed up. Since there are many categories, searching for a particular kind of image would be tedious. To reduce the task of manual deletion of images and to sort the images according to their category, this paper describes the design of a model that will extract the image into its respective folder.

Keywords: Handwritten notes · printed notes · circulars · convolutional neural networks (CNN) · image classification · deep learning · artificial neural network

1 Introduction

There is a growing need for automated approaches to classify and separate various types of documents, particularly in the context of digitalization and data management. Handwritten notes, printed notes, circulars, question papers, and mark sheets are just a few examples of many types of images that need to be organized and managed in both academic and professional settings. Manual sorting and organizing of these documents can be tedious, error-prone, and time-consuming, especially when the volume of documents is large.

To address this issue, we can employ machine learning methods like Convolutional Neural Networks (CNNs). CNNs, a type of artificial neural network, excel at image classification tasks and have seen success in fields including object recognition, facial detection, and medical image analysis.

In this paper, we propose a method for automatically classifying and separating handwritten, printed notes, circulars, question papers, and mark sheets using CNN. Our approach involves collection and preprocessing a dataset of images of these documents, designing, and training a CNN on the dataset, and evaluating the performance of the trained model on a separate test set. We demonstrate the effectiveness of our approach through a series of experiments and provide a thorough analysis of the results.

B. Raj et al. (Eds.): ICETE 2023, AER 223, pp. 355–360, 2023.
https://doi.org/10.2991/978-94-6463-252-1_39

Table 1. Different Approaches for Image Classification

1	Unsupervised Classification	Pattern recognition and image clustering are the common image classification methods. Example: K-Mean, ISODATA
2	Supervised Classification	Here already classified reference samples are used to train the classifier and then use the classifier to predict the new samples.
3	Deep Learning	1. Using custom build CNN Model 2. Using pre-trained CNN Models like VGGNet, ResNet etc.

Overall, our work contributes to the development of automated approaches for document management and has the potential to significantly improve the efficiency and accuracy of this important task.

2 Related Work

Recent advancements in deep learning approaches for image classification have been the subject of intense research in computer vision. This paper discusses the techniques of deep learning for image recognition. The recognition of the characters written by hand, face recognition and features, the recognition of images related to medical field are some of the areas where the image classification is needed [1].

There has been a significant amount of research on the use of machine learning techniques for document classification and organization tasks. Convolutional neural networks (CNNs) have been widely used for image classification tasks and have been shown to be effective for document image classification as well. As the number of training samples increases, the classification accuracy is increased. To increase the number of data samples, one of the approaches that can be used is Data Augmentation [2].

Documents like screenshots, brochures, notes can be separated and classified using CNN. Every category has some patterns and formats that differentiate from other categories. When a model is trained, it extracts those patterns and separates the images. One way for classifying the image is based on the text present in it. Another way is to separate the document based on the format of it. Documents usually have a hierarchical structure formed by specific patterns, such as characters forming words which in turn for sentences. Similarly for tables formed by cells and rows.

Table 1 shows the different types of algorithms that can be used for image classification. This paper demonstrates the use of custom CNN Model. As the number of images in the training dataset increases and when the number of categories are more, it is better to use pre-trained CNN Model.

3 Proposed Methodology

Automated approaches to classify and separate various types of documents, such as handwritten notes, printed notes, question papers etc. can significantly improve the efficiency and accuracy of document management tasks. In this paper, we propose a method

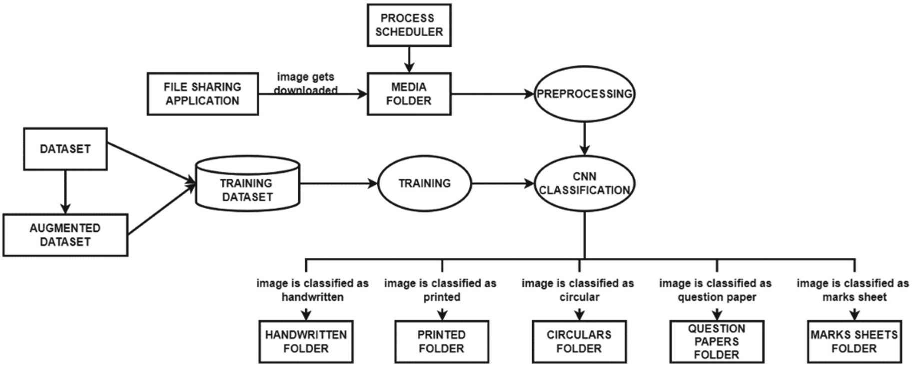

Fig. 1. Data Flow Diagram

for classifying and separating these types of documents using a Convolutional Neural Network (CNN) (Fig. 1).

3.1 Data Preparation

To prepare the dataset, we manually gathered many images of handwritten notes, printed notes, circulars, question papers, and mark sheets from various sources. Then the images are preprocessed by resizing them to a uniform size and converting them into grayscale. Data Augmentation techniques such as random cropping and horizontal flipping is also applied to increase the density of the training set and improve the generalization ability of the model. Around 620 images are present in the dataset. 80% of the images are used for training and 20% of the images are used for testing.

3.2 CNN Model Construction

CNN Model is constructed with several convolutional layers and pooling layers, followed by flatten, fully connected and output layer. During the training, Adam Optimization algorithm and categorical cross-entropy are used. The images that are collected along with the images which are augmented are together given as training data to the model. The model then trains with the images provided and predicts the category for the new image. Random Forest Regressor is used as a meta-learner because it is a powerful and robust algorithm that can handle a variety of data distributions and feature interactions. To use Random Forest Regressor as a meta-learner, the base learners' predictions are used as input features to the random forest regressor. The Random Forest Regressor then learns to combine these predictions to make a final prediction (Fig. 2).

4 Results and Evaluation

After classification, the images are classified into their respective categories. Since there are different types of handwriting, the model works well when trained with lot of data. Accuracy achieved for this trained model for categories printed notes, handwritten notes,

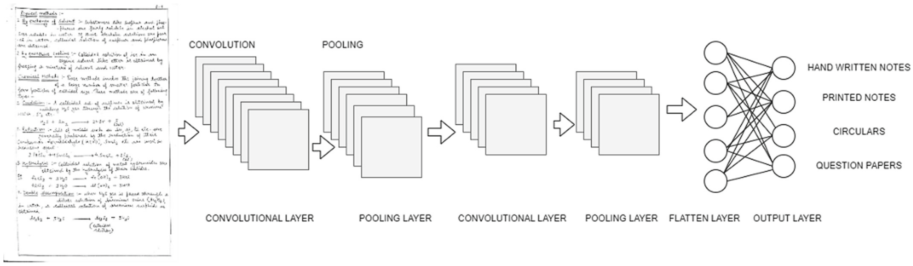

Fig. 2. CNN Model

circulars, question papers and marksheets is around 85%. The metrics that are calculated are mentioned in Table 1. Figure 3 shows the CNN training graph in which x-axis represents the training epoch and y-axis represents the accuracy and loss values. Figure 4 shows the confusion matrix in which x-axis represents the predicted labels and y-axis represents the true labels (Table 2).

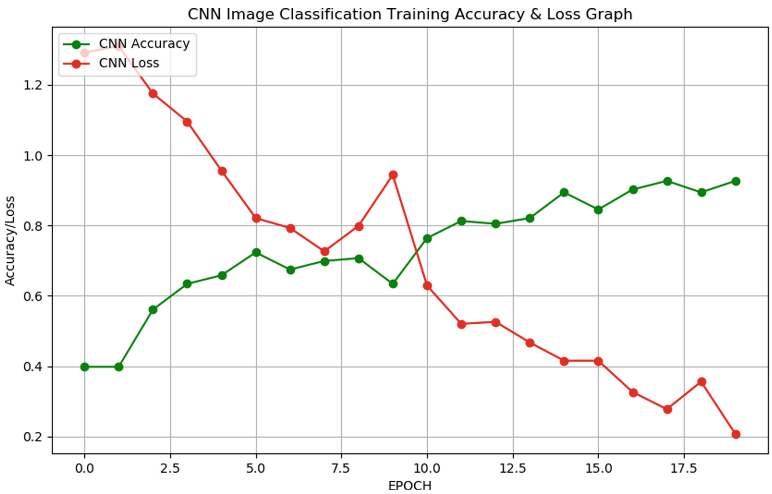

Fig. 3. CNN Training Graph

Table 2. Model Results

Type of Measurement	**Score**
Accuracy	89.4308
Precision	90.0617
Recall	81.2663
FSCORE	83.9428

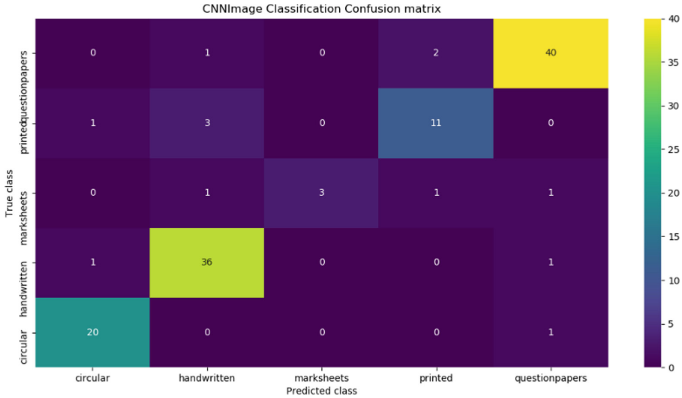

Fig. 4. Confusion Matrix

5 Conclusion and Future Scope

In this paper, we presented a method for classifying and separating handwritten notes, printed notes, circulars, question papers and mark sheets using a Convolutional Neural Network (CNN). Our approach involved collecting and preprocessing a dataset of images of these documents, designing, and training a CNN on the dataset, and evaluating the performance of the trained model on a separate test set.

In conclusion, the study we performed has advanced automated image classification and has the potential to significantly improve the task's effectiveness and accuracy. Our results indicate that the use of CNNs is a promising route for further research in this field, despite the fact that there are still certain restrictions, such as the requirement for high-quality input images.

Future scope could explore the use of other machine learning techniques or incorporation of additional features, such as document meta data, to improve the performance of the model. It would also be interesting to study the generalization ability of our CNN on larger and more diverse datasets of images of different categories. This can also be extended to documents other than images like pdf documents. As the number of categories and number of images in the dataset increases, it is better to use the updated algorithms which are pre-trained like VGG Net, ResNet etc. Using a pre-trained model provides more accuracy than using a custom build convolutional neural network.

References

1. Smys, S., and Wang Haoxiang. "Naïve Bayes and Entropy based Analysis and Classification of Humans and Chat Bots." Journal of ISMAC 3, no. 01 (2021): 40–49.
2. Tripathi, Milan. "Analysis of Convolutional Neural Network based Image Classification Techniques." Journal of Innovative Image Processing (JIIP) 3, no. 02 (2021): 100–117.
3. T. R. Sk, V. Komalatha, G. Sandeep, K. B. Rekha and S. Ravi Kishan, "Classification and Separation of Whatsapp Images Using Machine Learning," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1374–1379. https://doi.org/10.1109/ICOSEC51865.2021.9591866.

4. M. Z. Afzal et al., “Deepdocclassifier: Document classification with deep Convolutional Neural Network,” 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1111–1115. https://doi.org/10.1109/ICDAR.2015.7333933.
5. Pritpal Singh, Sumit Budhiraja, “Feature Extraction and Classification Techniques in O.C.R Systems for Handwritten Gurmukhi Script – A Survey,” IJERA, Vol. 1, Issue 4, pp. 1736–1739.
6. Sammed S. Admuthe, Hemlata P. Channe, “Document Image Classification using Visual and Textual Features,” Volume 10, Issue 09 (September 2021).