Research of Convolutional Neural Networks for Text Recognition in Traffic Scene Imagery

Quanqi Liu

School of Integrated Circuits, University of Electronic Science and Technology of China, Chengdu, China
2024310204033@std.uestc.edu.cn

Abstract. In the era of rapid development of intelligent transportation systems (ITS), the ability to accurately recognize text in traffic scene imagery is of utmost significance. Traffic signs, license plates, and road information in these images are crucial for traffic management, law enforcement, and navigation. For autonomous driving, it is a fundamental requirement for vehicle decision - making. However, previous text recognition methods, which depend on hand - crafted features, struggle in complex traffic scenarios. This paper conducts an in - depth exploration of Convolutional Neural Networks (CNNs) in traffic scene text recognition. It first details the theoretical principles of CNNs, including the functions of each component in the network structure, the importance of image preprocessing, and the processes of text detection and feature recognition. Then, it showcases CNN - based research in the transportation field, particularly in traffic sign recognition. It presents the comparison of YOLOv5 and SSD models in terms of accuracy and speed, and highlights a Chinese traffic sign detection algorithm. Finally, the paper proposes future research directions such as dataset expansion and new model testing, aiming to strengthen the performance of text recognition in traffic scenes and play a role in promoting the evolution of intelligent transportation.

Keywords: Convolutional Neural Networks (CNNS), Traffic Scene Imagery, Intelligent Transportation Systems (ITS).

1 Introduction

In the global technological landscape of the contemporary era marked by rapid technological evolution, the development of intelligent transportation systems (ITS) has become a crucial frontier. Therefore, at the core of ITS lies the ability to accurately interpret and utilize the text information embedded within traffic scene imagery. a task that is fundamental to ensuring seamless traffic flow, enhancing road safety, and enabling the widespread adoption of autonomous driving technologies.

Traffic scene imagery contains a wealth of text - based information that is of utmost importance for various aspects of transportation management. Traffic signs, for example, are the visual language of the roads, communicating critical instructions and warnings to drivers. A simple "Stop" sign is instantly recognizable to motorists due to

J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_35

its distinctive octagonal shape and red color, requiring vehicles to halt at intersections. Speed limit signs, on the other hand, regulate the pace of traffic, preventing accidents caused by excessive speed. License plates, too, play a vital role in law enforcement, traffic monitoring, and toll collection systems. They serve as unique identifiers for vehicles, enabling authorities to track and manage traffic violations effectively. Moreover, other text elements such as street names and directional signs are essential for navigation, guiding drivers through complex urban landscapes and unfamiliar routes. In the context of autonomous driving, where vehicles rely on sensors and algorithms to perceive and interpret their surroundings, accurate text recognition in traffic scene imagery is not just important but a necessity. It directly influences the decision - making process of self - driving cars, determining actions like when to stop, turn, or adjust speed.

However, traditional methods for text recognition in traffic scenes have proven to be inadequate in the face of real - world challenges. These methods typically rely on hand-crafted features to identify and classify text, such as extracting color patterns and geometric shapes from traffic signs. While these features can be effective in ideal conditions such as clear daylight and unobstructed views, they struggle to perform consistently in complex and dynamic traffic environments.

For example, under low - light conditions, such as at dawn, dusk, or during heavy rain or fog, the color and contrast of traffic signs may be severely degraded, making it difficult for traditional algorithms to accurately detect and interpret them. Furthermore, occlusions caused by other vehicles, pedestrians, or road infrastructure compound the problem. A partially blocked traffic sign can lead to misinterpretation, potentially resulting in dangerous situations. Additionally, the wide variety of text styles and fonts used in traffic scene imagery, especially in license plates from different regions or countries, poses a significant challenge for traditional methods, which lack the flexibility to adapt to such diversity.

The arrival of CNNs which are based on deep - learning has transformed the field of text recognition from images, offering a more robust and adaptable solution. CNNs are designed to automatically learn hierarchical features from images, eradicating the need for manually - engineered features. They are capable of effectively seizing not only low - level features such as edges and corners but also high - level semantic features, enabling them to recognize text even in the presence of complex backgrounds, varying lighting conditions, and occlusions. This paper is aimed at providing a thorough exploration of CNNs' application in text recognition within traffic scene imagery. By delving into the theoretical basis of CNNs and reviewing the latest research achievements in this area, it seeks to provide valuable insights and references for researchers and practitioners alike. The ultimate goal is to strengthen the performance of text recognition systems in traffic scenes. This enhancement will make contributions towards the development and progress of intelligent transportation.

2 Technical Underpinnings and Practical Applications of CNNs in Traffic Scene Imagery Text Recognition

2.1 Theoretical principles and operational mechanisms of CNN

2.1.1 Network Structure.

The main framework of a CNN is comprised of an input layer, convolutional layers, pooling layers, fully - connected layers, and finally an output layer. The input layer is responsible for receiving the original image data presented in the form of a pixel matrix. Its dimensions include the image height, width, and the number of color channels. For example, for an RGB color image with 224×224 pixels, the dimension at the input layer is [224, 224, 3].

As the core of a CNN, the convolutional layer extracts local image features through operations with convolutional kernels. A convolutional kernel is a small weight matrix. By sliding on the image and performing weighted summation, it generates feature maps. Different convolutional kernels can extract different features like textures and edges. For instance, a 3×3 convolutional kernel slides to generate a feature map, and weight sharing reduces the number of parameters and computational complexity [1].

The pooling layer performs downsampling on the feature maps to reduce computation and memory usage. Common methods include max - pooling and average - pooling. For example, with a 2×2 pooling window, it can halve the width and height of the feature map, reducing the amount of computation and memory usage while enhancing the model's robustness [2].

The fully - connected layer is at the back - end of the network. It maps the features into the output space through a linear transformation for classification or regression tasks. It transforms the multi-dimensional features generated by the pooling layer into a one-dimensional vector. The neurons within this layer are all interconnected with the neurons of the preceding layer. Due to the substantial quantity of parameters it possesses, this layer is highly susceptible to overfitting. In the context of an image classification task, the output dimension corresponds to the number of classes, and it produces the probability values for each individual class [3].

In classification tasks, the softmax function, which turns the fully-connected layer's output into a probability distribution, predicts the class with the highest probability. In regression tasks, the output layer directly outputs the predicted values. Figure 1 is a simple neural network consisting of three hidden layers.

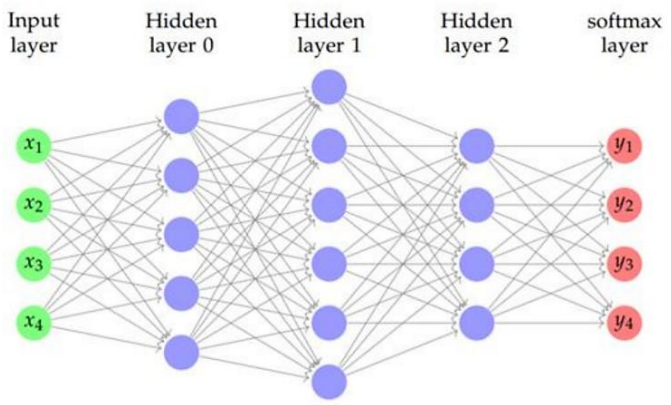

Fig. 1. The structure of the Neural Network [4]

Figure.2 illustrates different components of a CNN. To achieve progress in the architecture of CNNs, it is of great significance to have an understanding of the diverse components of CNNs and their applications.

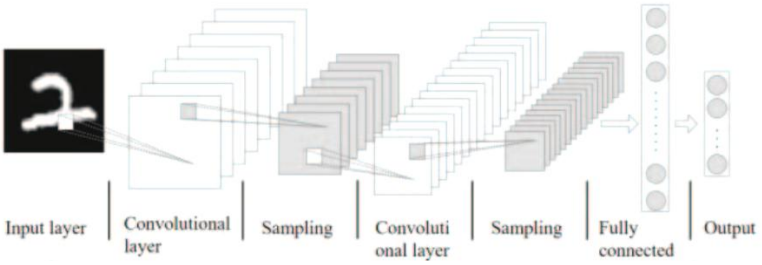

Fig. 2. Typical structure of convolutional neural network [5]

2.1.2 Image Preprocessing.

Image preprocessing (including binarization, denoising, and skew correction) enhances image quality and facilitates subsequent processing, serving as the foundational step in graphical text recognition. Color or grayscale images are converted into binary (black-and-white) images through binarization, which amplifies the contrast between text and background. Common methods include global thresholding (e.g., the Otsu algorithm for automatic threshold determination) and adaptive thresholding (dynamically determining thresholds based on local grayscale to accommodate uneven illumination).

During image acquisition, noise interference, like salt - and -pepper noise or Gaussian noise, is inevitable, degrading recognition accuracy. Denoising can be achieved via filtering algorithms: Gaussian filtering mitigates Gaussian noise, while median filtering effectively removes salt-and-pepper noise while preserving image details [6]. In practical scenarios, text in images may exhibit skew. Skew correction

involves estimating the tilt angle through projection analysis or detecting line orientations via the Hough transform, followed by adjusting the image to a horizontal state using rotation algorithms. These preprocessing operations collectively optimize image quality, ensuring robustness for subsequent recognition tasks.

2.1.3 Text Detection and Localization.

Text detection and localization aim to identify the position of text in an image, providing target regions for subsequent recognition. Common approaches include connected component analysis, edge detection, and deep learning-based methods. Connected component analysis relies on image morphology, applying dilation and erosion to enhance the connectivity of fragmented or isolated text regions. Connected components are then labeled and filtered based on features like area and aspect ratio to identify text regions [7].

Edge detection employs operators such as Canny or Sobel to detect areas with significant grayscale changes, thereby outlining text contours. Subsequent contour analysis and morphological operations are used to localize text [8]. Over the last several years, text detection methods grounded in deep learning have witnessed remarkable progress. For example, the EAST (Efficient and Accurate Scene Text detector) model uses a fully convolutional network to directly predict text region boundaries from images. CTPN (Connectionist Text Proposal Network) combines convolutional and recurrent neural networks to detect curved text lines, making it suitable for complex scenarios [9].

2.1.4 Feature Extraction and Recognition.

After completing text detection and localization, feature extraction and recognition of the text region images are required. CNNs play a vital role in this process. The text region images are fed into the CNN, undergoing operations through convolutional and pooling layers. The convolutional layers employ convolutional kernels to extract local features such as text strokes and corners, with different convolutional layers learning features at varying hierarchical levels. The pooling layers perform downsampling on feature maps, effectively reducing computational complexity while preserving critical information. The processed feature maps are then input into fully connected layers, where linear transformations map the features into a classification space. For text recognition, the softmax function converts the fully connected layer outputs into probability distributions. The class which possesses the highest probability value is then classified as the recognition result.

Additionally, the integration of Recurrent Neural Networks (RNNs) and its various derivatives like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) is carried out to handle sequential text information can significantly enhance recognition accuracy, particularly for handwritten or irregular text where contextual dependencies are crucial.

2.2 Recent research on CNN in the field of transportation

2.2.1 Research on the Application of CNNs in terms of Traffic Sign Recognition.

In the course of the evolution of intelligent transportation systems, Traffic Sign Recognition (TSR) assumes a crucial function in ensuring road safety and efficient traffic management. Traditional TSR methods depend on visual features like color and shape to identify traffic signs. Nevertheless, they are easily interfered by factors like lighting, occlusion, and vehicle speed, and perform poorly in multi - target detection and real - time performance. With the rise of deep learning, CNNs have been widely applied in the TSR field, effectively overcoming the shortcomings of traditional methods [10]. In their research, Yanzhao Zhu and Wei Qi Yan constructed a dataset consisting of 2,182 images in 8 categories, covering common traffic signs. They processed and labeled the images using software. Subsequently, the dataset was partitioned into a training set and a test set with a proportion of 8 to 2. They trained and tested the YOLOv5 and SSD models using the Google Collaboratory platform with specific hardware configurations [11].

The research focused on the YOLOv5 and SSD models. YOLOv5 comprises an input layer, a backbone network, a neck, and a prediction layer, and it uses various techniques to enhance data, extract features, and optimize predictions. SSD is composed of a modified VGG - 16 network and newly added convolutional layers, and it optimizes the model through localization loss and confidence loss [11]. The experiments showed that the mAP@0.5 of YOLOv5 reached 97.70%, demonstrating consistently high accuracy across all traffic sign categories. The mAP@0.5 of SSD was 90.14%, and the accuracy of some categories was relatively low due to the influence of the number of samples. In terms of speed, YOLOv5 had a significant advantage and was more suitable for real - time traffic scenarios [11]. This research serves as a reference for deep-learning-based TSR. Future studies could expand the dataset, test emerging models, and employ advanced evaluation metrics to further improve system performance.

2.2.2 Detection of traffic signs under deep CNNs.

In the paper 'A Traffic Sign Detection Algorithm Based on Deep CNN', the research team developed an algorithm used in the detection of Chinese traffic signs using a CNN and Region Proposal Network (RPN). The core objective was to effectively detect seven major categories of Chinese traffic signs [12]. Traffic sign detection is a key component of intelligent transportation systems, ensuring accurate and timely recognition of road signs. Currently, however, most existing detection methods are restricted to specific categories of traffic signs. Chinese traffic signs are more complex due to the presence of numerous Chinese characters. Coupled with the lack of a universal dataset, relevant research encounters multiple difficulties. In view of these issues, Xiong Changzhen et al. proposed this algorithm [12]. In the field of object detection, deep CNNs have witnessed singular development. Networks such as RCNN and SPPnets have been continuously iterated and optimized. Nevertheless, the currently available public traffic sign detection datasets mainly revolve around European traffic signs, and there is a severe shortage of data resources for Chinese traffic signs [12].

3 Conclusion

CNNs have demonstrated wonderful potential and advantages in the field of text recognition in traffic scene imagery. Through in-depth analysis of its theoretical principles, CNNs show remarkable capabilities. These principles include network structure, image preprocessing, text detection and localization, and feature extraction and recognition. It can be seen that CNNs possess a powerful ability to automatically learn features. This enables them to effectively handle the challenges of text recognition in complex traffic environments and overcome the limitations of traditional methods. In relevant research in the transportation field, CNNs have achieved remarkable results in traffic sign recognition. For example, the comparative study shows that YOLOv5 achieves higher accuracy while SSD provides faster detection speed, providing a reference for model selection in practical applications. The research on Chinese traffic sign detection has successfully achieved high - accuracy and real - time detection of multiple types of Chinese traffic signs, solving, to some extent, the complex problems such as character variety and font diversity in Chinese traffic sign detection.

However, in current research, it is still possible to make further improvements. Future research can be carried out in multiple directions, with dataset expansion being one of the most significant aspects. By covering a more extensive range of traffic scene images that incorporate diverse scenarios such as urban intersections, highway stretches, and rural roads, various weather conditions like sunny days, rainy periods, and foggy mornings, different fonts on traffic signs and vehicle number plates, as well as distinct styles of traffic flow patterns, it becomes possible to effectively help enhance the generalization ability of the model. This expanded dataset allows the model to better adapt to the real - world complexity and variability, thus improving its performance across a wide array of real - life situations. Enabling it to stably and accurately recognize text in various real - world situations. On the other hand, continuously exploring and testing new model architectures, combined with emerging technologies such as Transformer, is expected to further optimize model performance and elevate the precision and velocity of text recognition. Moreover, the adoption of more comprehensive and advanced evaluation metrics can precisely measure the actual performance of the model in complex traffic scenes, providing stronger support for model improvement and application. By continuously promoting the above - mentioned research directions, it is expected to further enhance the performance of text recognition in traffic scene imagery, inject new vitality into the development of intelligent transportation systems, facilitate the widespread application of autonomous driving technology, improve road safety, and enhance the efficiency of traffic management, thus promoting the intelligent transportation field to a new stage of development.

References

1. Liu, H., Sun, X., Li, Y.: 'Review of Deep Learning Models for Image Classification Based on Convolutional Neural Networks', Computer Engineering and Applications, 2025, 1, 1-29

2. Devi, N., Borah, B.: 'Cascaded pooling for Convolutional Neural Networks'. Proc. Fourteenth Int. Conf. Information Processing (ICINPRO), Bangalore, India, December 2018, 1-5
3. Chen, S., Sun, W., Huang, L., Yang, X., Huang, J.: 'Compressing Fully Connected Layers using Kronecker Tensor Decomposition'. Proc. IEEE 7th Int. Conf. Computer Science and Network Technology (ICCSNT), Dalian, China, October 2019, 308-312
4. R. S, A. S. Bharadwaj, D. S K, M. S. Khadabadi and A. Jayaprakash, "Digital Implementation of the Softmax Activation Function and the Inverse Softmax Function," 2022 4th International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2022, pp. 64-67
5. H. Li, "Computer network connection enhancement optimization algorithm based on convolutional neural network," 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 281-284
6. Xiao, K., Wei, Z., Li, K., Jiang, J.: 'A Novel Adaptive Switching Median filter for laser image based on local salt and pepper noise density'. Proc. IEEE Power Engineering and Automation Conf., Wuhan, China, September 2011, 38-41
7. Bharadwaj, A.S., SK, D., Khadabadi, M.S., Jayaprakash, A.: 'Digital Implementation of the Softmax Activation Function and the Inverse Softmax Function'. Proc. 4th Int. Conf. Circuits, Control, Communication and Computing (I4C), Bangalore, India, October 2022, 1-6
8. Yu, N., Zhang, Z., Xu, Q., Firdaous, E., Lin, J.: 'An improved method for cloth pattern cutting based on Holistically-nested Edge Detection'. Proc. IEEE 10th Data Driven Control and Learning Systems Conf. (DDCLS), Suzhou, China, July 2021, 1246-1251
9. Kotagiri, S., Venkataramana, A., Kiran, G.: 'Blind Aid: State of the art for Scene Text Detector and Text to Speech'. Proc. Int. Conf. Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, March 2022, 1-8
10. Fan, L.: 'Lighting Optimization Strategy Based on Image Recognition in Intelligent Transportation Systems', China Lighting & Lighting Accessories, 2025, 1, 101-103
11. Zhu, Y., Yan, W.Q.: 'Traffic sign recognition based on deep learning', Multimedia Tools and Applications, 2022, 81, (13), 17779-17791
12. Changzhen, X., Cong, W., Weixin, M., Yanmei, S.: 'A traffic sign detection algorithm based on deep convolutional neural network'. Proc. IEEE Int. Conf. Signal and Image Processing (ICSIP), Beijing, China, August 2016, 676-679