A Novel CNN Approach for Digital House Number Detection

Protiva Das
BRAC University
protiva.das@g.bracu.ac.bd
Annajiat Alim Rasel
BRAC University
annajiat@gmail.com

***Abstract*—Within the vast domain of computer vision, accurately distinguishing house numbers from street view images represents a substantial objective with applications spanning multiple sectors. This academic endeavor presents a meticulously constructed deep learning framework that has been customized to effectively navigate the intricacies of this undertaking. By integrating Convolutional Neural Networks (CNN), and Multi-Layer Perceptron (MLP), our methodology enhances its resilience against obstacles including obstructed pathways, varying styles, and unforeseeable illumination conditions that are characteristic of these numerals. The orchestrated phases comprise the architectural design of our framework: judicious feature extraction, initial data pre-processing, and concluding with elucidating sequential learning. With a certain level of proficiency, the CNN displays the extraction of complex features using a visual canvas of images, and the MLP makes the final judgment and provides the definitive classification.**

***Index Terms*—Computer Vision, CNN, MLP, Accuracy, Street view image, NLP, Numeric Cognition.**

I. Introduction

House number identification in street view images is important for a multitude of applications, including address verification, navigation, and urban planning. The main objective of this research is to find out a comprehensive deep learning framework that effectively tackles the challenge of residential number recognition from street view photographs. The architectural design uses Convolutional Neural Networks (CNNs), and Multi-Layer Perceptrons (MLPs) in an effort to use their advanced image recognition capability by optimizing training time and processing resources while maintaining a high level of accuracy. To simulate real-world situations, the dataset offers photographs of house numbers that has a huge amount of variety in scale, orientation, illumination, and occlusion. Convolutional Neural Networks (CNNs) are utilized to autonomously extract critical data by capturing complex structures and patterns in the images tackling these obstacles. The outcomes show how CNNs, and MLPs effectively address the difficulties created by a variety of street view images. This research paper presents a novel deep learning framework that increases accuracy in house number identification while testing the limits of the models by tweaking model size which also affects their training time. Thus the research endeavors to improve the precision and effectiveness of house number recognition.

With the growth of e-commerce platforms online shopping has become a day-to-day way of shopping with the convenience of not leaving the house. However, the whole e-commerce ecosystem needs a huge amount manpower, specially for the logistics of the parcels, which we can easily replace with automation. But if we consider current technological barrier, map service providers like Google Maps can not provide exact number of a house in a given street. If we send a self automated bot to deliver a parcel it can go to the specific street the house is located following the map. But it can not find the exact house to deliver the parcel in consumer's doorstep. In order to do that it has scan the whole street to find the specific house which is impossible without proper number recognition system for houses.

II. Existing Works

In recent times, a range of approaches have been developed that use deep Convolutional Neural Networks (CNNs) to identify and categorize house numbers and numerical representation present in Street View images. In their research, Haoqi et al. [1] used a CNN that had been trained on the SVHN dataset to detect house numbers with an excellent precision of 92.32%. Zang et al. [2] introduced an algorithm known as DAS, which uses street views to improve the accuracy of sampling and forecasting urban changes. By fine-tuning a deep CNN, Goodfellow et al. [3] obtained a 96% accuracy rate in identifying multi-digit numbers in Street View images. Luan et al. [4] conducted research on several techniques, such as data augmentation, spatial transformer networks, and the integration of Gabor filters, with which they tried to improve the feature representation and resistance of CNNs to transformations. In brief, the usage of deep convolutional neural network architectures, data enhancements, and specialized methodologies has shown remarkable efficacy in precisely identifying numerical values within street view visuals.

Researchers have tested a wide range of advanced learning strategies, such as convolutional neural networks (CNNs), to

determine how well objects and numbers within photos are classified and identified. On the SVHN digit dataset, a CNN model created by Sermanet et al. [5] shown an accuracy of 94.85%. Through accurate crop identification in Street View pictures, Ringland et al. [6] generated agricultural maps using the VGGNet model. A combination of CNNs and compressed sensing is used in Wang et al.'s [7] EHDCS-Net to improve the quality of HD photos. Ullah et al. [8] proposed a fruit identification and classification model that shown a 96% precision rate. Elshwai et al. [9] conducted a research on popular deep learning frameworks for a variety of workloads, including TensorFlow and MXNet. In their research, Shorten et al. [10] tried to find out how different data augmentation methods in deep learning affected the model accuracy. To sum up, the picture identification field has shown significant success with the combination of deep convolutional neural networks (CNNs) and focused techniques. It is important to understand that these networks' performance is significantly impacted by the choice of optimization techniques and framework.

Convolutional neural networks (CNNs) are one of the most prominent deep learning techniques that scientists have used to address a variety of problems involving image and street view data. Bansal et al. [11] tried a number of methods for getting over deep learning challenges brought on by a lack of data. With the CIFAR-10 dataset, Abu et al. [12] combined TensorFlow and Convolutional Neural Networks (CNNs) for picture categorization. To differentiate between pictures and artworks, Sharma et al. [13] made modifications to the VGG16 model. Sun et al. [14] and Yang et al. [15] used the Faster R-CNN algorithm to identify urban buildings and air conditioners in street photos. Chen et al. [16] used DenseNet161 for feature extraction and SVR to infer the ages of structures from street views. The overall research generally indicates that deep convolutional neural networks (CNNs) are quite good at handling difficult visual recognition tasks, especially when sparse data and real-world images are included. However, these networks' capacity to adjust to different datasets and their processing requirements are the real problem.

In summary, new researches shows that deep convolutional neural networks (CNNs) can achieve very high accuracy even in challenging picture identification tasks when there is a limited amount of input data. Convolutional Neural Networks (CNNs) have been used by researchers to detect objects and numerical values by utilizing transfer learning from big datasets and by trying out specific approaches like data augmentation and compressed sensing to enhance overall performance. Despite these advances, limitations remain in terms of processing requirements, models' adaptability to different datasets, and efficiency optimization for specific use cases. The main objective of some of current researche is to increase the accuracy, efficiency, and flexibility of deep convolutional neural networks (CNNs) across several problem domains, such as unstructured visual data and street scenes. To fully utilise CNNs' powers is the main goal of this work.

III. Proposed Methodology

For this study, we obtained the dataset of house numbers using street view and then carried out a thorough preparation phase for the dataset, which included preprocessing and splitting. The dataset was split into three parts using a 7:2:1 ratio, with 70% of the data allocated for training, 20% for model validation, and the remaining 10% for testing. After preparing the dataset, we proceeded to configure the models and then conducted training, validation, and testing phases to evaluate the performance of each model.

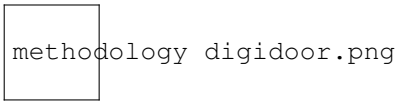
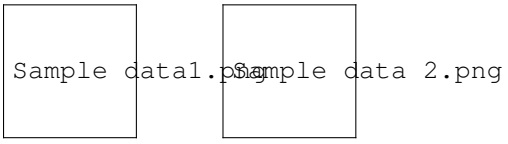


Fig. 1: Proposed Methodology

A. Dataset

The SVHN dataset [17] is comprised of over 600,000 genuine 32x32 RGB images of residential numbers that were acquired via Google Street View. Its objective is to facilitate the development of automatic number recognition algorithms. The dataset exhibits a heterogeneous distribution of classes and comprises images of diverse quality levels. SVHN provides an extensive collection of digit images spanning from 0 to 9. These images can be utilized to assess the efficacy of machine learning models, notwithstanding challenges such as extraneous background information and noise. SVHN is an effective dataset for training and evaluating the development of digit recognition systems in real-world computer vision.



(a) (b)

Fig. 2: Sample Data

B. Used Architectures

The architectures utilized in this work include Multilayer Perceptrons (MLP) and Convolutional Neural Networks (CNN), both of which were precisely customized to meet the specific requirements of our research. The architectures were intended to effectively tackle the specific issues presented by the different characteristics of the dataset. The goal was to improve the overall performance and efficacy of our models in identifying subtle patterns and correlations within the house numbers.

1) Convolutional Neural Networks (CNN): CNN is a widely used neural network for grid like data, for example, images. A CNN's convolutional layer is its core. Each layer can be structured differently with its own shape and filter size.

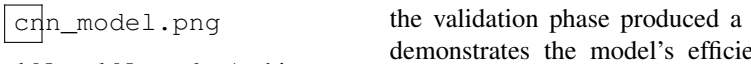


Fig. 3: Proposed Convolutional Neural Networks Architecture

For the Convolutional approach, we developed a customized CNN model considering our SVHN dataset where we increased the number of layers and filters. We trained the model with a batch size of 64 and ran it through 50 epochs. In our model, we used a 2d convolution layer with a filter size of 128. The model also consists of a dropout layer where 30% of the neurons in dropout layers will be set to zero randomly. Besides, we used 256 fully connected neurons in our Dense layer. Figure 3 illustrates an overview of our customized CNN architecture.

2) Multi-Layer Perceptrons (MLP): MLP is a feedforward neural network where data flows from the input layer to the output layer in one direction. In order to serve our purpose of showing higher accuracy and lower loss in our SVHN dataset, we developed a custom-built MLP model. We trained the model with a batch size of 1024 and we ran it through 50 epochs. Here we used the dense layers in a funnel-type architecture where it started with 512 neurons and then gradually we decreased the neuron size to 256 and 128. Before reaching the 128 neuron layer we flow our data through a dropout layer with the value of 0.3 meaning 30% of the neurons in this layer will be set to zero randomly. Figure 4 Illustrate the overview of our used MLP architecture.

Fig. 4: Proposed Multi-Layer Perceptrons Architecture

IV. Result Analysis

This section provides a thorough assessment of the model's performance and its process of acquiring knowledge. The learning process is closely monitored by analyzing accuracy and loss curves throughout the training and validation phases. Through successive epochs, these curves provide a graphical representation of the models' ability to adapt to the dataset. Moving on to performance metrics, The study thoroughly explores widely recognized performance indicators, including the confusion matrix, accuracy, precision, recall, F1 score, and ROC curves. These measures evaluate the models' ability to accurately classify instances, offering a detailed comprehension of their overall efficiency.

A. *Convolutional Neural Networks (CNN)*

The accuracy and loss curves depicted in Figure 5 provide a visual representation of the model's performance during the training and validation phases. These curves offer insights into how the model learns from the training data and applies its knowledge to the validation data. Within the scope of our investigation, the accuracy score of 0.9159 given during the validation phase indicates that the proposed CNN architecture attained an accuracy of 91.59%. This metric measures the model's accuracy in correctly classifying instances within the SVHN dataset throughout the validation phase. Similarly, the validation phase produced a loss score of 0.3214, which demonstrates the model's efficiency in minimizing training errors. A decrease in the loss value indicates that the model has effectively acquired and adjusted to the patterns in the training data, showcasing its capacity to apply this knowledge to unfamiliar data.

Fig. 5: Accuracy & Loss Curves for proposed CNN architectures

Limitations: As the graph shows the validation loss and validation accuracy is fluctuating which may indicate a potential overfitting situation

Prediction Results:

Figure 7: Custom CNN ROC Curve

Here we can see that our CNN model is performing pretty well for some classes like class 2, 5, 8 and 9 where the AUC score is respectively 0.51, 0.62, 0.55, and 0.63. However, for other classes the AUC score below 0.5 threshold.

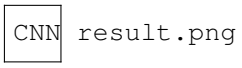


Figure 8: Custom CNN Predictions

Here we can see our model performed very well predicting the right digit in the given test cases every single time.

Figure 9: Custom CNN Confusion Matrix

B. *Multi-Layer Perceptrons (MLP)*

Our custom MLP model performed very well with our SVHN dataset with an accuracy score of 0.7802 and a loss score of 0.8350. The loss and accuracy graph our custom MLP model is given below:

(a) MLP Loss

(b) MLP Accuracy

Fig. 6: Accuracy & Loss Curves for proposed MLP architectures

Limitations: In this model the validation loss and validation accuracy is fluctuating constantly over the epochs. Continuous fluctuation of validation loss and validation accuracy signifies there may be an issue of overfitting lies beneath the grounds.

Prediction Results:

Figure 11: Custom MLP ROC Curve

Here we can see that our MLP model performed pretty well for the class 5, 8, and 9 with AUC of 0.72, 0.63, and 0.65 respectively. But for other classes the AUC is very low specially for class 6 and 7 the AUC is significantly low.

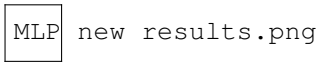


Figure 12: Custom MLP Predictions

This picture shows our MLP model successfully predicted 4 classes everytime in a given test case. But with rest of the test cases the prediction is not that much good.

Figure 13: Custom MLP Confusion Matrix

Model Comparison:

In the end we can draw a difference between the CNN and the MLP model by showing the Accuracy, Precision, Recall and F1 score table for comparison.

Figure 14: Model Comparison Table

V. Future Work Plan

In future works the main concern should be fixing serious overfitting flaws found in SVHN dataset testing. Training MLP and CNN models reduced loss, although overfitting worsened it after a few epochs. This project requires thorough hyperparameter optimization. We carefully tweak learning rates, batch sizes, and network depths to prevent overfitting and optimize generalization. Model learning training data noise is prevented by dropout and L2 regularisation.

In the future, the models have to be evaluated extensively. Thorough cross-validation tests the model's performance, overfitting, and generalization is needed. A comprehensive technique to rectify overfitting and improve model performance on the SVHN dataset will promote research with machine learning solutions in this industry.

VI. Conclusion

The efficacy of these methods in image classification is substantiated by our thorough study of the SVHN dataset, which adopted three distinct models— CNN, and MLP. The results underscore the significance of carefully choosing a suitable model, as it has a direct impact on the outcomes. As MLP has restricted ability to depict complex spatial relationships, the custom MLP model demonstrated average performance in terms of both accuracy and loss metrics compared to CNN. However comparing with the default model MLP improved significantly, though it may have some issues related to data overfitting. This proofs that the CNN is capable of extracting hierarchical information from images, resulting in an overall improvement in performance across all metrics. Our custom CNN model shows remarkable precision and negligible attrition in pattern recognition across the SVHN dataset, solidifying its position as the most effective methodology for image classification endeavors that prioritize visual information. The experiment that was performed without any doubt proves that the process of selecting models is a critical component of machine learning. The differences between our two custom model highlights the exceptional performance of the CNN in image-related tasks, underscoring the importance of deliberate model architecture selection, as the MLP's performance fell short of expectations. These observations helps in the progression of machine learning implementations by providing guidance to professionals in selecting the optimal model architecture for particular datasets.

References

[1] H. Yang and H. Yao, "Street View house number identification based on deep learning," *International Journal of Advanced Network, Monitoring, and Controls*, vol. 4, no. 3, pp. 47–52, 2019. https://doi.org/10.21307/ijanmc-2019-058

[2] G. Zhang, J. Yi, J. Yuan, Y. Li, and D. Jin, "DAS: Efficient Street View Image Sampling for urban prediction," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 2, pp. 1–20, 2023. https://doi.org/10.1145/3576902

[3] I. J. Goodfellow, "Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks," *arXiv.org*, December 20, 2013. https://arxiv.org/abs/1312.6082

[4] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018. https://doi.org/10.1109/tip.2018.2835143

[5] P. Sermanet, "Convolutional neural networks applied to house numbers digit classification," *arXiv.org*, April 18, 2012. https://arxiv.org/abs/1204.3968

[6] J. Ringland, M. Bohm, and S. Baek, "Characterization of food cultivation along roadside transects with Google Street View imagery and deep learning," *Computers and Electronics in Agriculture*, vol. 158, pp. 36–50, 2019. https://doi.org/10.1016/j.compag.2019.01.014

[7] T. Zeng, J. Wang, X. Wang, Y. Zhang, and B. Ren, "An efficient Deep Learning-Based High-Definition image compressed sensing framework for Large-Scene construction site monitoring," *Sensors*, vol. 23, no. 5, p. 2563, 2023. https://doi.org/10.3390/s23052563

[8] D. Hussain, I. Hussain, M. Ismail, A. Alabrah, S. S. Ullah, and H. M. Alaghbari, "A simple and efficient Deep Learning-Based framework for automatic fruit recognition," *Computational Intelligence and Neuroscience*, 2022, pp. 1–8. https://doi.org/10.1155/2022/6538117

[9] R. Elshawi, A. Wahab, A. Barnawi, and S. Sakr, "DLBench: a comprehensive experimental evaluation of deep learning frameworks," *Cluster Computing*, vol. 24, no. 3, pp. 2017–2038, 2021. https://doi.org/10.1007/s10586-021-03240-4

[10] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019. https://doi.org/10.1186/s40537-019-0197-0

[11] A. Bansal, R. Sharma, and M. Kathuria, "A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–29, 2022. https://doi.org/10.1145/3502287

[12] M. A. Abu, N. H. Indra, A. H. A. Rahman, N. A. Sapiee, and I. Ahmad, "A study on Image Classification based on Deep Learning and Tensorflow," *International Journal of Engineering Research and Technology*, vol. 12, no. 4, pp. 563–569, 2019. https://www.ripublication.com/irph/ijert19/ijertv12n4_16.pdf

[13] H. K. Sharma, T. Choudhury, S. N. Mohanty, S. Swagatika, and S. Swain, "Deep Learning based approach for Photographs and Painting Classification using CNN Model," https://ceur-ws.org/Vol-3283/Paper101.pdf

[14] H. Xu, H. Sun, L. N. Wang, X. Yu, and T. Li, "Urban Architectural Style Recognition and Dataset Construction Method under Deep Learning of Street View Images: A Case Study of Wuhan," *ISPRS International Journal of Geo-information*, vol. 12, no. 7, p. 264, 2023. https://doi.org/10.3390/ijgi12070264

[15] F. Yang and M. Wang, "Deep Learning-Based Method for Detection of External Air Conditioner Units from Street View Images," *Remote Sensing*, vol. 13, no. 18, p. 3691, 2021. https://doi.org/10.3390/rs13183691

[16] Y. Chen, A. Rajabifard, and M. Aleksandrov, "Estimating building age from Google street view images using deep learning (short paper)," in *10th international conference on geographic information science (GIScience 2018)*. https://drops.dagstuhl.de/opus/volltexte/2018/9368/pdf/LIPIcs-GISCIENCE-2018-40.pdf

[17] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng, "The street view house numbers (SVHN) dataset," Technical report, Accessed 2016-08-01.[Online], 2018. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=The+Street+View+House+Numbers+%28SVHN%29+Dataset&btnG=