Journal of Engineering Issues and Solutions

# Enhancing Handwritten Text Recognition Performance with Encoder Transformer Models

**Dipesh Bhattarai[1], Pawan Kumar Sharma[2]**

*[1,2]Texas College of Management and IT, Kathmandu, Nepal*

*Submitted: April 13, 2025; Revised: June 15, 2025; Accepted: June 21, 2025*

*https://doi.org/10.3126/joeis.v4i1.81606*

## Abstract

Handwritten Text Recognition (HTR) is a critical area in computer vision and natural language processing, aiming to convert handwritten content into machine-readable text. The task poses significant challenges due to the inherent variability in handwriting styles, stroke patterns, character spacing, and writing instruments. Traditional HTR techniques, often based on statistical models or shallow neural networks, frequently struggle to generalize across diverse handwriting samples, leading to suboptimal performance in real-world applications.

This improvement demonstrates a relative gain of approximately 10-15% over traditional RNN and LSTM-based models on the same dataset.

**Keywords:** Handwritten Text Recognition, Transformers, BERT, CNN, Deep Learning, OCR

## Introduction

Handwritten Text Recognition (HTR), a subfield of Optical Character Recognition (OCR), involves transforming handwritten input into machine-encoded text. The wide variability in personal handwriting styles complicates this task. Traditional HTR models using Hidden Markov Models (HMMs) [1] or Recurrent Neural Networks (RNNs) [2] often struggle with long-range dependencies and contextual nuances. Transformers, particularly Bidirectional Encoder Representations from Transformers (BERT) [3], offer a promising solution due to their powerful self-attention mechanisms and contextual modeling. This study proposes a CNN + BERT hybrid model tailored for HTR.

Despite advances in HTR, existing models often lack sufficient generalization across handwriting styles due to limited contextual modeling. This research addresses this gap by leveraging encoder-based transformers, offering a scalable and more accurate recognition solution. The key contributions of this paper are: (1) development of a CNN + BERT hybrid architecture for HTR, (2) extensive hyperparameter tuning for performance optimization, and (3) demonstration of improved metrics over baseline models.

## Related Work

Earlier models for HTR relied heavily on HMMs for sequential modeling [1]. With the advent of deep learning, CNNs [4] and Long Short-Term Memory (LSTM) networks [2] became dominant, especially when combined with Connectionist Temporal Classification (CTC) loss [5] for unsegmented sequence data. BERT [3], although originally developed for NLP tasks, has recently been applied to visual sequence modeling, with promising results [6, 7]. Transfer learning approaches [8] and multilingual CNN architectures [9] have also shown effectiveness in boosting HTR accuracy.

## Methodology

Below is a conceptual illustration of the proposed hybrid architecture. It shows how CNN layers are used for spatial feature extraction, followed by a BERT encoder for sequence modeling, and finally a dense classification layer.

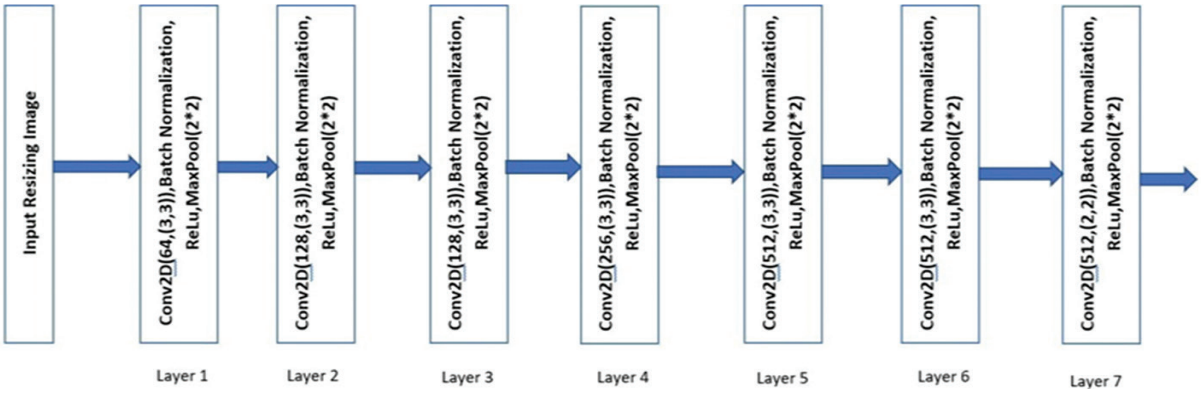


**Figure 1:** CNN Model Architecture

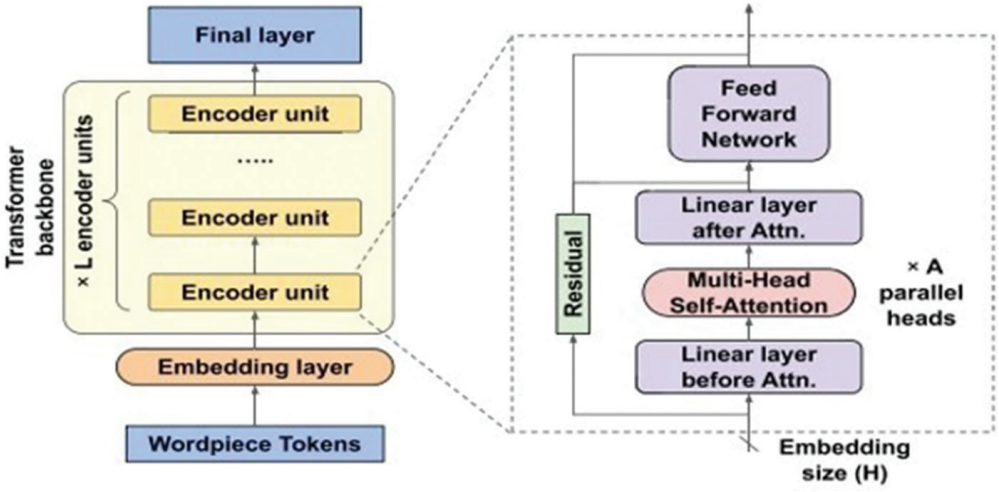


**Figure 2:** BERT flowchart

We used the IAM dataset consisting of 115,320 grayscale images. The images were resized to 32x128 pixels, normalized, and padded to a uniform format. The dataset was split in an 8:1:1 ratio for training, validation, and testing.

The model architecture includes CNN layers for spatial feature extraction and a BERT encoder for contextual interpretation. Dense layers with Softmax activation perform final classification. Key hyperparameters tuned include optimizer (Adamax), learning rate (0.0001), number of transformer blocks (6), head size (8), and number of attention heads (10).

Evaluation metrics include Accuracy, Precision, Recall, F1-score, and Character Error Rate (CER), where CER = (Substitutions + Insertions + Deletions) / Total Characters.

CNN layers extract spatial features from input images, which are reshaped into sequences compatible with transformer input. These sequences are then passed into the BERT encoder to model contextual dependencies. BERT was fine-tuned specifically on the image-derived embeddings.

Basic augmentations such as rotation, noise addition, and horizontal shifts were applied to improve robustness. The choice of 6 transformer blocks and 10 attention heads was based on balancing computational efficiency with model depth. Training was performed using an NVIDIA RTX 3060 GPU with 12GB VRAM and 16GB system RAM.

## Results and Discussion

The model was trained with early stopping enabled, halting after 18 epochs. The training accuracy reached 90%, and validation accuracy stabilized at 86.20%.

Character Error Rate was evaluated on test samples such as 'for', 'way', and 'have', showing CERs of 0 and 0.25, respectively. The final performance metrics were: Precision = 88.94%, Recall = 83.86%, and F1-score = 0.8674.

The hybrid architecture effectively leverages CNN's local feature extraction and BERT's deep sequence modeling capability, providing a robust and generalizable HTR solution.

Despite strong overall performance, minor misclassifications occurred, particularly on visually similar characters (e.g., 'a' and 'o'). Low CER values across tests demonstrate practical reliability in real-world digitization tasks.

Table 1 below compares the proposed model's performance with baseline methods on the IAM dataset. This clearly highlights the significant improvement offered by our approach.

**Table 1:** Comparison of Model Performance on IAM Dataset

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| HMM | 65.3% | 66.1% | 64.0% | 65.0% |
| RNN | 72.5% | 74.0% | 71.2% | 72.6% |
| LSTM-CTC | 78.6% | 80.1% | 77.3% | 78.7% |
| Proposed CNN+BERT | 86.2% | 89.39% | 84.35% | 86.74% |

## Applications

The system can be applied in diverse domains such as:

- Document digitization and archival
- Assistive technology for visually impaired users

- Automated form processing in government and healthcare
- Postal and logistics services for address recognition

For instance, postal services can use the system to automatically extract handwritten addresses from scanned envelopes, reducing sorting errors and improving delivery speed.

## Future Work

Future enhancements include the incorporation of multilingual datasets (e.g., Hindi, Nepali), advanced data augmentation techniques, pre-trained domain-specific transformer models, and deployment on edge devices for real-time applications.

Multilingual support could be realized via multilingual BERT models or by applying transfer learning from large-scale pre-trained handwriting datasets in diverse scripts. Edge deployment challenges include memory constraints and inference latency, which can be addressed through model quantization, pruning, and ONNX/TFLite conversion.

## References

L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proc. 23rd Int. Conf. on Machine Learning (ICML), 2006, pp. 369–376.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 20, 2009, pp. 577–584.

P. Nayak and S. Chandwani, "Improved offline optical handwritten character recognition: A comprehensive review using TensorFlow," Int. J. Eng. Res. Technol. (IJERT), vol. 10, no. 11, pp. 2278–0181, 2021.

P. Ganesh et al., "Compressing large-scale transformer-based models: A case study on BERT," Trans. Assoc. Comput. Linguist. (TACL), vol. 9, pp. 1061–1080, 2021.

J. C. Aradillas Jaramillo, J. J. Murillo-Fuentes, and P. M. Olmos, "Boosting handwriting text recognition in small databases with transfer learning," in 16th Int. Conf. Front. Handwriting Recognit. (ICFHR), IEEE, 2018, pp. 429–434.

J. Bai, Z. Chen, B. Feng, and B. Xu, "Image character recognition using deep convolutional neural network learned from different languages," in IEEE Int. Conf. Image Process. (ICIP), 2014, pp. 2560–2564.