Digital Restoration for Printed Text Manuscript

SPOORTHI. G. S[1], ADHYA T P[2], AMITKUMAR C S[3], CHETHAN KUMAR R L[4], D SAI VASANTH[5]
[1] *Assistant Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*
[2, 3, 4, 5] *B.E. Students, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*

Abstract*— Text recovery is a complex problem in natural language processing, involving the recovery of partially destroyed text. Deep learning techniques, like the Seq2Seq model, have demonstrated significant promise for text recovery tasks in the past couple of years. In this paper, we suggest the GloVe integrated Seq2Seq model for text recovery. This approach aims to exploit the semantic information contained in the GloVe embedding to translate partially destroyed text. In order to assess how successful the suggested approach is, on carrying out a sensitivity analysis to investigate how various hyperparameters affect the model's functionality. The analysis shows that the choice of hyperparameters, such as hidden layer size and learning rate, can significantly affect model performance, also by perform preprocessing steps such as data cleaning and augmentation to improve input data quality. GloVe embedding, which encodes semantic information of words and sentences in a dense vector space, is the primary strength of the suggested approach. This enables the model to interpret the input text's meaning even when it is entirely or partially distorted. The model's accuracy is further increased by using attention techniques, which enable the model to concentrate on pertinent segments of the input sequence.

I. INTRODUCTION

Text restoration is a challenging problem in natural language processing, which involves restoring partially destroyed or corrupted text. This problem arises in various scenarios such as natural disasters, digital document processing, and historical text preservation. In recent years, deep learning techniques such as Seq2Seq models have shown great promise for text restoration tasks.

Seq2Seq models are a type of neural network architecture that consists of two recurrent neural networks (RNNs), an encoder and a decoder. The encoder takes an input sequence and encodes it into a fixed-length hidden representation, which is then used by the decoder to generate an output sequence. Seq2Seq models have been successfully applied to various natural languages processing errands, such as machine interpretation, content summarization and question answering.

For text restoration, we present a Seq2Seq model with GloVe embeddings in this study. A method for learning dense vector representations of words and phrases that encode their semantic information is called GloVe (Global Vectors for Word Representation). It has been demonstrated that GloVe embeddings work well for a variety of natural language processing tasks, including word similarity and analogy tests. Our suggested method seeks to recover the partially destroyed text by utilizing the semantic information present in GloVe embeddings.

We propose using an encoder-decoder architecture that incorporates attention techniques. The partially destroyed text is fed into the encoder, which converts it into a hidden representation. The decoder uses this representation to produce the recovered text. Given the encoded input sequence, the decoder is taught to predict the most likely output sequence. In order to enable the model to selectively focus on pertinent segments of the input sequence during the decoding process, we also employ attention methods.

Two things our research has to offer. Firstly, we provide a unique method for text restoration based on GloVe embeddings in a Seq2Seq model. Secondly, we carry out in-depth tests to illustrate the efficacy of our methodology and juxtapose it with multiple reference techniques. Our method may find use in a number of situations where textual material is distorted or lost, including digital document processing, historical text preservation, and natural disasters.

II. LITERATURE SURVEY

Early approaches to text restoration focused on using rule-based techniques to correct spelling errors and grammatical mistakes in the text. However, these approaches were limited by their reliance on hand-crafted rules, which made them difficult to scale and adapt to new languages or domains.

The progression in text restoration has moved from Hidden Markov Models to Recurrent Neural Networks (RNNs), which include LSTM and GRU variants. Despite RNNs' achievements, they encounter challenges such as slow processing and a limited grasp of semantics. To overcome these, Attention mechanisms pinpoint vital sections of the input sequence. Moreover, integrating pre-trained embeddings like GloVe and Word2Vec enhances models by infusing semantic information and refining generalization.

"Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag & Nando de Freitas, "Restoring and attributing ancient texts using deep neural networks",09-March-2022"
Ithaca is a deep neural network that is revolutionary for attribution and restoration of ancient Greek inscriptions. It works faster and more accurately than conventional techniques. It supports historians in collaborative workflows with interpretable results, demonstrating efficacy in tasks related to chronological sequence, geographical attribution, and restoration. In a historical discussion, its effect on rearranging chronological information is clear. Ithaca provides cutting-edge research tools that broaden the field of ancient history, and its open-source interface makes it easier for scholars working with ancient texts to apply its findings across disciplinary boundaries.

"Chinmaya Misra, P K Swain, Jibendu Kumar Mantri, "Text Extraction and Recognition from Image using Neural Network", February-2012"
The paper proposes an image indexing and retrieval system using neural networks, focusing on text extraction. HSV-based color reduction is employed, and features from specific color planes are used for text detection. Identified text blocks undergo OCR, and the output is stored as keywords in a database. The conclusion highlights plans to enhance efficiency, address limitations in detecting non-horizontal text and complex backgrounds, improve text tracking with complex motion, and enhance OCR accuracy for better retrieval quality.

"Mayank Wadhwani, Debapriya Kundu, Deepayan Chakraborty, Bhabatosh Chanda, "Text Extraction and Restoration of Old Handwritten Documents", March-2021"
This chapter introduces two new approaches that use deep neural networks to reconstruct historical transcriptions. To help with preparation, a small dataset consisting of 26 historical letters is provided, together with semi-automated ground truth data era. The main method uses a Gaussian mixture model for foundation rebuilding and a neural network for text extraction. The current method uses a neural network for both extracting a closer image and reconstructing the foundation. These techniques are suitable for digital legacy preservation repositories with small sample sizes since they work effectively on severely deteriorated written by hand reports, even with a little dataset. Additionally, the tactics can be strengthened to restore printed reports that have deteriorated.

"N Shobha Rani, B J Bipin Nair, M Chandrajith, G Hemantha Kumar, Jaume Fortuny, "Restoration of deteriorated text sections in ancient document images using a tri-level semi-adaptive thresholding technique", 23-Febraury-2022"
Adaptive thresholding for binarizing deteriorated document images is investigated in this study. The authors suggest a tri-level semi-adaptive binarization method that uses local thresholds for post-enhancement and global thresholds for initial degradation removal. The method effectively removes noise and retains text in DIBCO samples, but challenges remain for palm leaf documents. Future studies may focus on improving text retention for such cases and evaluating the algorithm's speed compared to other methods.

Overall, the literature on text restoration has shown that deep learning techniques, such as RNNs and Attention mechanisms, are effective in restoring missing or corrupted text. The use of pre-trained embedding, such as GloVe and Word2Vec, has also

been shown to improve the performance of text restoration models. However, there is still much room for improvement in this area, and future work could investigate the use of more advanced techniques, such as Transformers and BERT, for text restoration tasks.

III. PROBLEMS IDENTIFIED

The current framework faces various challenges, requiring escalates endeavors to upgrade speed and execution. Past specialized changes, tending to these issues in compressed space handling requires a nuanced investigation of elective strategies. In non-horizontally arranged content discovery, the investigation of imaginative calculations and picture handling procedures may bridge holes and abdicate more comprehensive comes about. Progressing the system's capacity to track content in scenarios with complex movement includes refining existing strategies and considering progressed following calculations or machine learning models. This all encompassing approach upgrades versatility in energetic situations. Tending to destitute acknowledgment precision against complex foundations requires creating modern picture handling calculations, joining relevant data, and utilizing versatile sifting methods. Recognizing the system's unacceptability for computerized legacy conservation stores emphasizes the require for domain-specific adjustments, including collaboration with specialists in computerized conservation, authentic sciences, and social legacy. In conclusion, overcoming distinguished wasteful aspects requires a multifaceted approach including specialized optimizations, algorithmic advancements, and domain-specific adjustments, moving the framework to higher execution in different and complex scenarios.

IV. PROPOSED SYSTEM

An encoder/decoder design with an attention mechanism is the basis of our suggested methodology. Text that has been partially garbled is fed into the encoder, which converts it into a secret format. The decoder uses this to construct the restored text. Based on the encoded input sequence, the decoder is trained to predict the most likely output sequence. During the decoding phase, attention techniques are used to help the model focus on relevant portions of the input sequence.

However, some limitations should be considered to further improve the performance of the model. Eventually, we should consider using a transfer learning approach for text recovery. Transfer learning uses pre-trained weights from a deep learning model to tune them

CONCLUSION

In summation, the optimization of the existing system necessitates the implementation of a comprehensive strategy, amalgamating technical enhancements with pioneering methodologies, notably the integration of advanced technologies such as machine learning. This strategic approach addresses intricacies within compressed domain processing, non-horizontally oriented text detection, and dynamic text tracking, thereby fortifying the system's adaptability in real-world scenarios.

Furthermore, the refinement of recognition accuracy amid complex backgrounds, achieved through sophisticated image processing techniques, contributes substantially to the system's overall versatility. Concomitantly recognizing constraints in the context of digital heritage preservation, collaborative endeavors with domain experts become imperative for tailoring the system effectively to repositories managing a substantial volume of samples. This bifold strategy positions the system as a more proficient solution, poised for enhanced performance and sustained efficacy across a spectrum of diverse applications, aligning with the rigorous standards of contemporary research.

REFERENCES

[1] Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag & Nando de Freitas, "Restoring and attributing ancient texts using deep neural networks",09-March-2022, https://www.nature.com/articles/s41586-022-04448-z

[2] Yannis Assael, Thea Sommerschield *, Jonathan Prag, “Restoring ancient text using deep learning: a case study on Greek epigraphy”, 15-October-2019, https://www.deepmind.com/publications/restoring-ancient-text-using-deeplearning-a-case-study-on-greek-epigraphy

[3] Amit Chauhan, “Text Detection and Extraction From Image With Python”, 13-March2022, https://medium.com/pythoneers/text-detection-and-extraction-from-image-with python-5c0c75a8ff14

[4] Chinmaya Misra, P K Swain, Jibendu Kumar Mantri, “Text Extraction and Recognition from Image using Neural Network”, February-2012, https://www.researchgate.net/publication/224890763_Text_Extraction_and_Recognition_from_Image_using_Neural_Network

[5] Mayank Wadhwani, Debapriya Kundu, Deepayan Chakraborty, Bhabatosh Chanda, “Text Extraction and Restoration of Old Handwritten Documents”, March-2021, https://www.researchgate.net/publication/350144825_Text_Extraction_and_Restoration_of_Old_Handwritten_Documents

[6] N Shobha Rani, B J Bipin Nair, M Chandrajith, G Hemantha Kumar, Jaume Fortuny, “Restoration of deteriorated text sections in ancient document images using a tri-level semi-adaptive thresholding technique”, 23-Febraury-2022, https://www.tandfonline.com/doi/full/10.1080/00051144.2022.2042462

[7] J. Xu, W. Ding and H. Zhao, "Based on Improved Edge Detection Algorithm for English Text Extraction and Restoration From Color Images," in *IEEE Sensors Journal*, vol.20,no.20,pp.11951-11958,15Oct.15, 2020,doi:10.1109/JSEN.2020.2964939https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8952767&isnumber=9199391

[8] Honggang Zhang, Kaili Zhao, Yi-Zhe Song, Jun Guo,” Text extraction from natural scene image: A survey” Neurocomputing 122, 310-323, 2013https://www.researchgate.net/publication/262205838_Text_extraction_from_natural_scene_image_A_survey

[9] Jian Liang, David Doermann, Huiping Li” Camera-based analysis of text and documents: a survey” International Journal of Document Analysis and Recognition (IJDAR) 7, 84-104, 2005 https://www.researchgate.net/publication/225238035_Camera-Based_analysis_of_text_and_documents_a_survey_Int_J_Doc_Anal_Recognit_72-384-104

[10] Yihong Gong, Xin Liu, “Generic text summarization using relevance measure and latent semantic analysis”, DOI:10.1145/383952.383955 https://www.researchgate.net/publication/220017549_Generic_Text_Summarization_Using_Relevance_Measure_and_Latent_Semantic_Analysis