Enhanced Optical Character Recognition by Optical Sensor Combined with BERT and Cosine Similarity Scoring (Student Abstract)

Woohyeon Moon, Sarvar Nengroo, Taeyoung Kim, Jihui Lee, Seungah Son, Dongsoo Har

Korea Advanced Institute of Science and Technology, Daejeon 34051, South Korea
moonstar, sarvar, jihui, seungahson, dshar@kaist.ac.kr

Abstract

Optical character recognition(OCR) is the technology to identify text characters embedded within images. Conventional OCR models exhibit performance degradation when performing with noisy images. To solve this problem, we propose a novel model, which combines computer vision using optical sensor with natural language processing by bidirectional encoder representations from transformers(BERT) and cosine similarity scoring. The proposed model uses a confidence rate to determine whether to utilize optical sensor alone or BERT/cosine similarity scoring combined with the optical sensor. Experimental results show that the proposed model outperforms approximately 4.34 times better than the conventional OCR.

Introduction

Optical character recognition (OCR) (Smith 2007; Kim et al. 2020) technology recognizes text within a digital image that is typed, handwritten, or printed. This includes two primary fields, text detection and text recognition. Text detection detects text tokens and deals with the location of text tokens within an image, either in a word or sentence. On the other hand, text recognition attempts to comprehend the content of text images and convert visual signals into natural language tokens.

However, despite the significant advancements in OCR techniques, challenges persist, particularly when dealing with noisy images. To address these challenges and improve the robustness of OCR systems, we propose a bidirectional encoder representations from transformers(BERT) based OCR (BERTOCR) model. Unlike the conventional OCR methods, the proposed approach integrates contextual information from the surrounding words by using the BERT model. By incorporating visual and contextual information, the proposed model aims to enhance the accuracy and reliability of text recognition, especially when dealing with noisy images.

Proposed Method

Figure 1 illustrates the architecture of proposed model. This model involves a sequence of steps, commencing by providing input to the OCR model. The output from the OCR model is optionally used for the BERT and cosine similarity scoring.

OCR

The OCR model provides the text estimated from text image. The probability of accuracy is indicated by this output text. The model uses the hyperparameter K to determine whether an image is clear or noisy. We set the value of K to 0.5 for this work. If the probability of the text is greater than K, the output is considered to have good performance, and if the output is less than K, it is treated as potentially erroneous one. The potentially erroneous one is used for the BERT model.

BERT

The BERT model is trained to predict the masked tokens, e.g., words. To do this, we replace the noisy word with a masked token. The BERT model gives predetermined number of candidate words with different prediction probabilities. For example, the sentence,'was voted top selling animespecial effect DVD' in Fig 2 is changed to 'was voted top [MASK] animespecial effect DVD', as the word 'selling' is noisy. The '[MASK]' represents masked token.

Cosine Similarity Scoring

The BERT model predicts the masked token using other context tokens. However, the BERT model considers the context only, not the image similarity. From this viewpoint, the cosine similarity scoring is added to consider image similarity. The output from the BERT is fed to the 'Cosine Similarity' scoring process. For this, words with high prediction probability (greater than the average probability) are chosen. In the next step, these words are converted into text images to measure the similarity between the predicted words and the original noisy words. The predicted images and the original noisy images are compared by Cosine Similarity. Finally, the arithmetic operation of the final score is taken to get the result. For this, the confidence rate (generated by BERT) and image similarity score (produced by the cosine similarity scoring) are multiplied. In evaluating the scores, the optimal selection is determined by using argmax function.

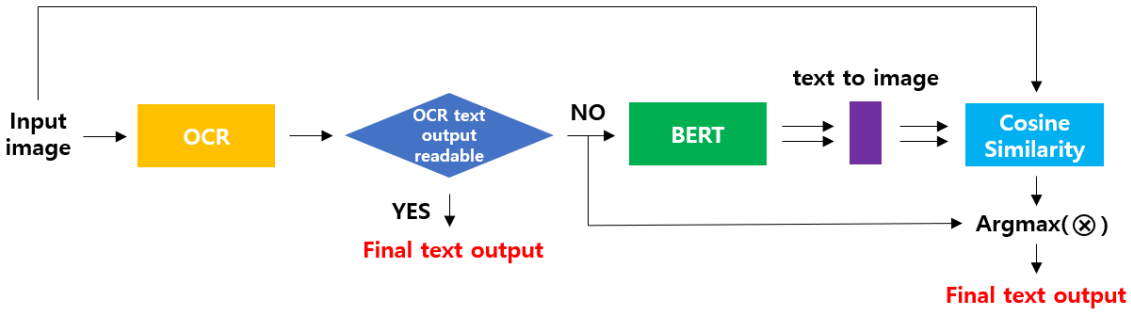

Figure 1: Schematic of the proposed model consisting of OCR, BERT, and cosine similarity scoring. OCR alone represents conventional OCR and OCR combined with BERT and cosine similarity scoring represents the proposed model BERTOCR.

Figure 2: Example of noisy text (two gray noise lines on top of the word 'selling').

Model name	Dataset Wiki	CNN	Fiction	Average accuracy
conventional OCR	0.0625	0.0588	0.0570	0.0594
OCR and BERT	0.3641	0.1979	0.1838	0.2486
BERTOCR	0.3719	0.2086	0.1923	0.2576

Table 1: Experimental results of the 'conventional OCR', 'OCR and BERT', and BERTOCR method.

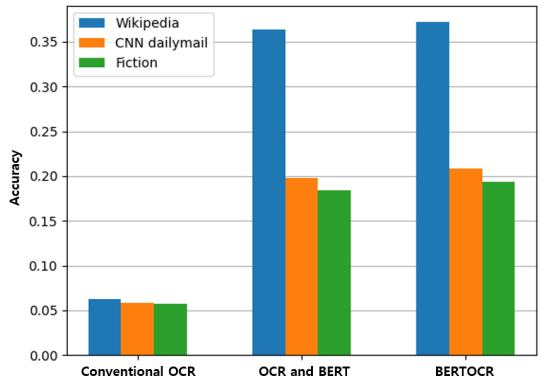

Figure 3: Comparison results obtained from 'conventional OCR', 'OCR and BERT', and BERTOCR method.

Experimental Results

In this work, 7.8 million sentences from Wikipedia, 300 thousand sentences from CNN dailymail, and 2000 sentences from Fiction data are used for testing. To prove the effectiveness of the proposed model, artificial noise like the one in Fig. 2 is introduced to the Wikipedia, CNN dailymail, and Fiction data. The performance of different models is evaluated considering the accuracy score, as graphically shown in Fig. 3 and tabulated in Table 1. The conventional OCR model yields an average accuracy score of 0.0594. In contrast, the combination of the OCR and BERT model significantly improves the performance, achieving an average accuracy score of 0.2486. The proposed model surpasses both the conventional OCR model and the OCR model combined with BERT model, attaining an average accuracy score of 0.2576. The results demonstrate that the proposed method outperforms the conventional OCR model by an impressive factor of 4.3367 and was better than the OCR and BERT model by 1.0362 times(3.6%).

Conclusion

This paper proposes a novel model that unifies computer vision using optical sensors with sophisticated natural language processing by BERT and cosine similarity scoring. The performance analysis of the proposed method is compared with the conventional OCR model. The proposed model outperforms the conventional OCR model by 4.3367 times. This work opens up possibilities for further advancements in OCR technology, particularly in challenging noisy input scenarios by using the NLP model.

Acknowledgments

This work was supported by the Institute for Information Communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No.2020-0-00440, Development of Artificial Intelligence Technology that continuously improves itself as the situation changes in the real world).

References

Kim, T.; Vecchietti, L. F.; Choi, K.; Lee, S.; and Har, D. 2020. Machine learning for advanced wireless sensor networks: A review. *IEEE Sensors Journal*, 21(11): 12379–12397.

Smith, R. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, 629–633. IEEE.