IMG2LATEX: Extract mathematical formula in English and Korean into LaTeX Form

Hyeonwook Jeon Junho Wang, Dongsu Park, Yurim Jang, Jaehyun Park, Younggyu Park, Seungheon Song

Abstract: Optical Character Recognition (OCR) has taken a large part of modern Computer Vision (CV). However, despite the rapid growth of the CV technologies, OCR systems adhere to the traditional method, text detection and text recognition. This method must be improved in line with the advent of new technologies such as Transformer. In this work, we focus on building a system extracting text with formula by only using DONUT, a pure transformer-based free-OCR model. We used pre-trained donut and fine-tuned the model with mathematical datasets and achieved competitive performance with other OCR systems, resulting accuracy of 94.2 for F1-score and 92.7 for Levenshtein distance score in both Korean and English. For the convenience of the users, we also built the function to classify levels with extracted formulas and provide explanations through BERT and ChatGPT-API. We fine-tuned BERT for classifying level of formula, which achieved 95 accuracy score and used ChatGPT for providing explanations. We have also created a web service to provide these services to users, and plan to upgrade all of them further.

1. Introduction

Optical Character Recognition (OCR) takes large part of modern Computer Vision (CV) technology for decades. And, as Computing resource has been grown up rapidly since 2012, new and old Deep Learning (DL) method keeps emerging on the surface. But the OCR Technology seems to stop its development compared to other CV methods although new DL structure keeps invented. Its problem is that it is largely depend on traditional OCR method which is slow and inaccurate. So the extracting handwritten letters from image is considered as hard job until state of the art model OCR-free Document Understanding Transformer (DONUT; Geewook Kim at al, 2022) shows up.

This paper utilize DONUT and take OCR to the next level. And it also can detect mathematical formula and turns it to the LaTeX symbol even if it is written in handwritten form. To do that we present specific algorithm to extract formula well and we also implement estimate its level based on the regular course of Korean education. We also use ChatGPT API so that the user could get explanation and fully understand the formula leading to study the formula in the easier way.

This paper is structured as follows. Section 2 explains what is OCR, LaTeX symbol and DONUT model. Section 3 provides Performance score based on F1 Score and Levenshtein distance. Section 4 presents experiment how we can implement and serves model through web, so that we can make people utilize our model for their own. Lastly, Section 5 describes what is the motivations to invent this kind of model.

2. Related Work

2.1. Optical Character Recognition

OCR is a technology extracting texts from image enabling the conversion of printed or handwritten text into machine-readable text. OCR has grown its portion in CV. As machine translation often require OCR technology to translate languages into other.

Traditional OCR method largely rely on CV technology as it requires optical devices so it can utilize them. And as DL methods has largely impacted on machine learning methods OCR has also got effected by DL methods such as Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs) which have demonstrated superior performance handling complex and varied textual content.

But the fact that it is heavily relying on DL methods especially Computer Vision (CV) limits its speed. as it can't be implemented in an end-to-end structure. And the fact that it doesn't use Natural Language Processing (NLP) model limits its accuracy, because it doesn't understand its context so the unrelated words to the context can be show up. These facts make hard to extract letters from image. Especially when it comes to handwritten forms or formula.

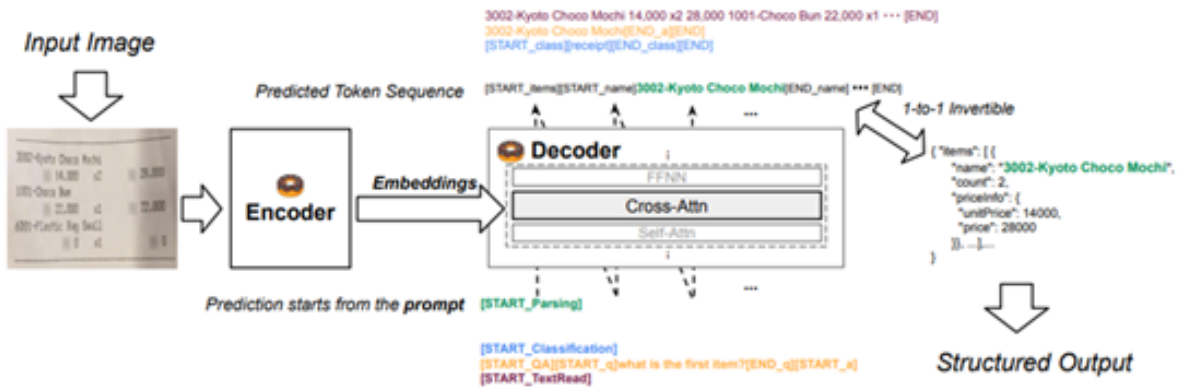

Fig. 1. The structure of the DONUT model

2.2. *LaTeX symbol*

LaTeX is a powerful typesetting system and markup language widely used for creating documents with complex formatting, mathematical equations, and professional layouts. It has become the standard in academic and technical publishing due to its ability to produce high-quality documents that are aesthetically pleasing and easily reproducible. LaTeX is now considered as a standard language to write formula through typing.

However, extracting formula from image to text is a challenging job especially in handwritten form. Because, extracting texts from handwritten lettered image is already a hard job to be done. And the fact there are few datasets about LaTeX and formula gives difficulty to the job compared to normal OCR process. So, we had to utilize from limited datasets and resources as we can raise its performance to the usable level.

2.3. *DONUT Model Explained*

Traditional Visual Document Understanding (VDU) is run based on traditional OCR form. So, the cost was high and accuracy and speed is not good enough. But understanding the context is critical to our job because extracting texts from handwritten form of image is too much job to be done only with OCR. So we used DONUT the OCR-free Document Understanding Transformer (Geewook Kim at al, 2022)

To put it simply DONUT model uses Swin Transformer method as encoder so that create embedding which is dividing an image into a several different times and build a stack like pyramid to take examine about the image. They use them because Swin Transformer method has strong points to jobs like object detection, segmentation, and classification. Once done that, Textual Encoder playing role as a decoder process embedding through multilingual BART model, so that it can reproduce vocabulary token in one-hot vector for GPT-3. Model takes sets of one-hot vector as inputs and process it for GPT-3 and it can reproduce words as prompt engineering tells what to do like parsing, classification, QnA, Textread. such downstream tasks are much accurate than traditional OCR because it can understand context. Finally, GPT-3 reproduces Json file as an output. so that users can easily enable to utilize the output as they want.

3. Performance and Computing Resource

3.1. *Performance and Evalutation*

F1 Score: 94.2%
handwritten letters F1 Score: 92.6%
printed letters F1 Score: 98.9%
Levenshtein Score: 92.7%

3.2. *Computing resources*

- Model Server AWS g5.8xlarge Instance
OS: Ubuntu 20.04
Python Version: 3.10.6
GPU: Nvidia A10G (24GB)
Web Framework: Flask

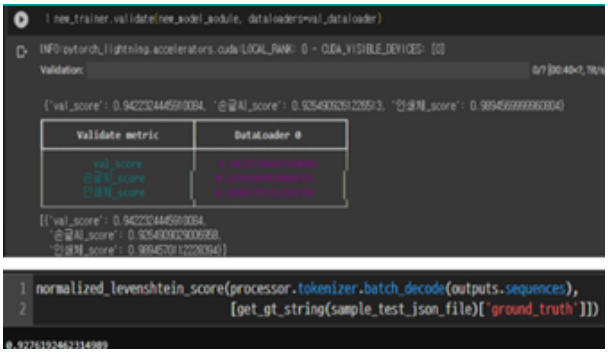

Fig. 2. Performance and Evalutation Score

DB: Sqlite3

- Web Server:
AWS Free-tier Instance
OS: Amazon Linux
Python Version: 3.10.6
Web Framework: Django
DB: MariaDB (AWS RDS)

3.3. Algorithm

At first, we just use donut model in training since we believed that no error could happen. But, because DONUT model use generative artificial intelligence there was an hallucination problem to the output when we set the max length too long. the output was disappointing as it repeatedly reproduce useless token or words in the front part of the sentence. so we have to set the max length more accurately to the problem.

To solve the problem, we used EasyOCR to roughly guess the the number of line of output text. But the output text includes LaTeX token which makes appropriate number of line longer compared to the number of line that EasyOCR guessed. So, we use the fact that there are end-line tokens at the ends of the lines.

The idea is that when we train the model, we made model guess the number of line of the texted image so that we can use it to guess the number of line of the output text. As said before, the estimated length that EasyOCR guessed is shorter than final output's one. So before end line token is same as the trained model estimating numbers of line, we repeat the process for donut model to do the OCR while increasing the max length of the output text. For example, there is an image with 5-line text and we made donut model guess the number of lines of the text in the image. So, we know the output's line's length because the number of lines is same as input's one. we made Donut model do the OCR work repeatedly while increasing the max length of the model when one OCR work is done. When the OCR output's end-line token become same as DONUT model's guessed lines of text which is 5. we end the loop and show the final output of the OCR.

By that, we can get rid of hallucination problem by limiting the max-length to the same amount of actual output text.

4. Implementation

4.1. Main Model

Firstly, we use data preprocessing process to input images. We resized the image into 480 x 480 pixels. and then we convert image to RGB and to pixel values so that we can put image into CUDA. When we train we made start line token 'line ', end line token '/line ' and the number of the line of texts 'len ' token which indicates the number of line of the texts so that we could guess final output's numbers of lines.
and then we used pretrained model 'donut-base' and Adam optimizer to train the model. the epoch value was 3, validation check interval, check validation every n epoch, gradient clip validation value were all 1. learning rate was 0.0004 and learning scheduler type was linear. the number of warmup steps was 100, seed was 42, and log steps was 200, and finally batch size was 2. The reason of epoch is low is that when we tried the higher epochs the handwritten letter's OCR accuracy becomes less accurate. So did the batch size.
when the OCR is done, we delete special tokens and put the output into json form.

4.2. Sub Model

BERT (Bidirectional Encoder Representations from Transformers) stands out as a state-of-the-art model that builds upon and refines a series of developed transformer methods (Vaswani et al., 2017; Radford et al., 2018). In our research, we employed three distinct tokenizers: the main model tokenizer, a Korean-based tokenizer (bert-kor-base), and a Korean-English tokenizer (ke-t5-base).
Among these, the main model tokenizer, referred to as the "Donut" model, showcased superior performance, achieving an accuracy of 0.95. In a recent project, we extended the capabilities of BERT to classify LaTeX formulas specific to Korean mathematics curriculum, ranging from elementary school 4th grade to high school 3rd grade. This endeavor bridges the gap between advanced NLP techniques and educational content, providing a robust tool for educators and researchers. BERT employs a unique training strategy where it masks certain tokens in a sentence and predicts them, leveraging both left and right context. This bidirectional nature sets it apart from previous models like OpenAI's GPT (Radford et al., 2018) which only used a unidirectional approach. Recent studies, such as Liu et al. (2019) and Lan et al. (2020), have observed that while BERT performs exceptionally well on balanced datasets, its performance can degrade on imbalanced or niche datasets. This observation is crucial in real-world scenarios where data imbalance is a common challenge. For addressing these challenges in NLP tasks, a variant of BERT, named BERT-Balance, has been proposed. In BERT-Balance, the training data is post-processed to ensure that the class distribution of tokens is more uniform. This approach is particularly useful when there's no prior knowledge about the distribution of classes in the training data

4.3. Server

5. Motivations and Expected Effect

5.1. Motivation for Selecting the Topic and Expected Effect

Our teammates met deep learning bootcamp. As course is online, we had common problem that machine learning foumula is hard to note down through computer. So we decided to solve the problem on our own. it was also good opportunity to review what we learned. As there was no Korean version of LaTeX OCR program. it gave our project originality too.
we implemented not only OCR function we also developed level estimation of formula and simple explanatrion using ChatGPT. we also optimize the model using trained model to serve service through website (https://kr-img2latex.site/fileUpload/) You can also see other people's work so that people can share the link and see each other's work with co-workers.
we hope that students like us get help through our work even if we can't maintain server for too long because of financial reason.

6. Conclusion

In this paper, we have developed and fine-tuned the DONUT model, a Transformer-based pretrained model, to effectively extract text data from images containing a combination of Korean and English text along with LaTeX symbols. The model's specialization in mathematical contexts holds great promise for applications within the field. Fine-tuned DONUT model did generate high quality outputs of the images, but still had a problem of hallucination which the model generates unpredictable tokens to fill the max length of sequence. So, we applied our unique algorithm using EasyOCR, we could handle DONUT's hallucination problem and bring out satisfying result, 94.2% for F1-score and 92.7% for Levenshtein score.
And by incorporating a BERT-based submodel, we successfully introduced an education-level-classification system that aligns with the Korean mathematics curriculum. fine-tuned BERT model classifies the level of the given formulas and uses ChatGPT to provide explanations about it. We achieved 95 accuracy score for level classification and found the best prompt for the explanation in ChatGPT. This development lays the foundation for the DONUT model's potential to provide differentiated services in the domain of Korean mathematics education.
Furthermore, the absence of a LaTeX symbol tokenizer during our research posed challenges in text-data preprocessing. The envisioned creation of such a tokenizer has the potential to significantly reduce the time required for text data refinement, while also enhancing the overall performance of the model.
In conclusion, the DONUT model's inception from transformer-based pretrained models enables effective text extraction from images featuring a mixture of Korean and English text with LaTeX symbols. Its potential for advancement within the mathematical domain, coupled with the prospect of a LaTeX symbol tokenizer, suggests a pathway to streamlined processing and heightened performance. This work contributes to the growing landscape of specialized applications for language models, particularly in the field of Korean mathematics education

References

1. Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park. *OCR-free Document Understanding Transformer.* (arXiv:2111.15664).
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* (arXiv:1810.04805).
3. ZeLiu, YutongLin, YueCao, HanHu, YixuanWei, Zheng Zhang, Stephen Lin, Baining Guo. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.* (arXiv:19010.13461).
4. Mike Lewis,Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* (arXiv:19010.13461).
5. https://openai.com/ (OpenAI-ChatGPT API)
6. https://www.kaggle.com/code/nbroad/donut-train-benetech (kaggle notebook - donut train benetech)
7. https://github.com/clovaai/donut
8. https://github.com/google-research/bert
9. https://github.com/JaidedAI/EasyOCR
10. https://github.com/kiyoungkim1/LMkor