

Deep learning 1 Progress Report (week 12)




Group 6 - Sujeong Oh, Kyumhi Lee, Joan Vasco

“OCR-free Document Understanding Transformer”

Data loading status:

In this project we have to use data which is image of CORD, Ticket, Business Card, Receipt and also corresponding json file which contains the data recorded based on the information written in the image with the location of it. It has four languages English, Chinese Japanese and Korean. We decided to use English data only due to the large huge number of image dataset.

The paper required us to deal with the data with “hugging face”, web that allow us to do deep learning related things. The data looks like below. There are images and the corresponding text information together.

image (image)	ground_truth (string)
	<code>{"gt_parse": {"text_sequence": "Dares Wins Vol. 5 Tommy's Heroes Vol . 6: For Tomorrow Vol. 7: Closin g Time miniserie s.Clark Kent is being inter viewed about Su perm an 's connectio n to notorious kill..."}</code>
	<code>{"gt_parse": {"text_sequence": "nary rule does not apply to evidence foun d due to neglige nce regarding a gover nment database"}}"</code>
	<code>{"gt_parse": {"text_sequence": "ns. Like civilian pr ofessors they seek academi c promotion to the rank of associate professo r and professor. However they are not eligible for tenure. is an aca demic..."}</code>

We can use this dataset through the api of “hugging face”. The code can should use in our environment is “curl -X GET \...” if we use in our colab environment, 500,000 images with English language will upload in our environment.

Task division:

We have three team members, so we divided the project into three tasks. Our main mission was to transfer the code to our environment and make it executable, as well as thoroughly understanding the content of this paper and creating a PowerPoint presentation to deliver it. Vasco, who came from a foreign country and has a good command of English, took on the role of reading and summarizing the paper. Student OhSujeong took the role of dealing with the code. Especially, set up a deep learning environment on her computer and make the code executable in her environment. Lastly, student Lee,Kyumhi took the role that collaborate with Vasco to read the paper in detail and based on that, start to creating the PowerPoint presentation early. He also required to assisting with necessary parts of the code. It was decided that Vasco and Lee,Kyumhi, who thoroughly understood the paper, will do the presentation.

Paper understanding:

To extract useful information from such document images, Visual Document Understanding (VDU) is the main task and main theme of this paper. “OCR-free Document Understanding Transformer” from its name we can understand that in this research, they didn’t use the optical character recognition (OCR) method which is common method to deal with scanning documents image to understanding documents. The reason why it didn’t use OCR is because OCR-dependent approach has critical problems. First is using OCR as a pre-processing method is expensive. The computational cost for inference would be expensive for high-quality OCR results. Second, the off-the-shelf OCR methods rarely have flexibility dealing with different languages or domain changes, which may lead to poor generalization ability. Third, OCR errors would propagate to the VDU system and negatively influence subsequent processes. Upon these problems in the current models, authors proposed model based on Transformer-only architecture, referred to as Document understanding transformer (Donut), which has a huge success in vision and language.

What the authors introduce of their research:

1. We propose an innovative OCR-free approach for Visual Document Understanding (VDU). This method is the first to utilize an OCR-free Transformer trained end-to-end.
2. We introduce a simple pre-training scheme that allows for the use of synthetic data. Using our generator, SynthDoG, we demonstrate that our approach, Donut, can easily be extended to handle multiple languages, unlike conventional methods that require retraining an existing OCR engine.
3. Through extensive experiments on both public benchmarks and private industrial datasets, our proposed method not only achieves state-of-the-art performance but also provides practical advantages such as cost-effectiveness in real-world applications.

There are three abilities that this transformer model can do.

First is Document Classification, classifying which type of document the image is through the information and the location of the words.

Second Is, Document Information Extraction. This is the process of automatically extracting relevant information and data from documents. It involves identifying and capturing specific data elements such as names, dates, addresses, keywords, or structured information from unstructured or semi-structured documents.

Third Is Document Visual Question Answering. It refers to the task of answering questions related to the content of a document, taking into account both textual and visual information present in the document. The task involves understanding the textual and visual elements of the document and generating accurate and relevant answers to questions based on the information contained within.

Deal with immediate problems and errors:

Immediately, we struggled with running the codes. Several problems occurred in the process of learning the model of the GitHub repository in the Colab environment. First, there was a conflict between the version of the dependency library in the repository and the package already installed in the Colab environment. Because Colab is basically installing many packages, the repository could conflict with the required packages and versions. This resulted in a failure to properly install the required library and a code execution failure. Secondly, it was necessary to import the example datasets provided by the repository into Colab and convert them into the appropriate format. The data set of the repository was uploaded to the hugging face and provided. Therefore, the dataset had to be imported into Colab and converted so that it could be used for model learning. Third, the file path hardcoded to the code or configuration file in the repository did not match the Colab environment, which caused the problem. In a Colab environment, the path hardcoded in the code or configuration file could be incorrect because the path of the file system might be different. This resulted in an error in which the file could not be found, requiring modification of the path to resolve the problem. Due to the above problems, the process of conducting model learning at Colab was more complicated and difficult than expected. However, we are trying to overcome these problems and successfully proceed with model learning.