SDL: NEW DATA GENERATION TOOLS FOR FULL-LEVEL ANNOTATED DOCUMENT LAYOUT

Nguyen Truong Son
School of Environment and Society
Tokyo Institute of Technology
Tokyo, Japan
nguyen.s.ab@mot.titech.ac.jp

June 30, 2021

ABSTRACT

We present a novel data generation tool for document processing. The tool focuses on providing maximal level of visual information in a normal type document, ranging from character position to paragraph-level position. It also enables working with a large dataset on low-resource languages as well as providing a mean of processing thorough full-level information of documented text. The data generation tools come with a dataset of 320000 Vietnamese synthetic document images and an instruction to generate a dataset of similar size on other languages. The repository can be found at: https://github.com/tson1997/SDL-Document-Image-Generation

1 Introduction

Optical Character Recognition (OCR) is the problem of reading and extracting information from an image. In order to extract correct information in the image, multiple level detection were made. They range from character-level detection to paragraph level detection. Such set of document structuring problems is called Document Layout Analysis (DLA). The goal of DLA is to decompose an image of document into multiple parts, whose information are then further processed or filtered based on their usefulness.

With the power of deep learning, DLA can solve the detection problems by using a good object detection model, such as Faster RCNN [1], or Mask RCNN [2]. Hence, the problems left is how to obtain a good dataset to train such object detection model. [3] [4] proposes ways to generate synthetic scene text image. Such synthetic datasets yield good variety of scene text image with text blending into background. However, we cannot use those datasets directly to train a model for decomposing a document image into level component that are bigger than word-level, since there are differences in distribution between document text and scene text data.

On the other hand, [5] provided a way to process documented data from the Internet and provided their resulted PublayNet dataset of more than 300000 images and corresponding label. However, PublayNet dataset has the limitation on label-level (only big component such as paragraph, table, list), and limitation on language domain (only in English).

Despite the fact that there are many open-source for both scene-text dataset, at the time of this paper, there is no dataset that covers a full level of label from character up to big page component like paragraph. There is a need for such data providing an end-to-end solution, where all components ranging from smallest like character to biggest like paragraph would be covered thoroughly. In addition, there is also a need for data providing flexibility in languages, especially low-resource language where collecting big dataset is not feasible.

In this work, we are going to provide a new method to generate synthetic data, with the main focus on multi-level text component, and possible to generate in multiple languages. In short, our main contributions are:

- First algorithm to generate synthetic visual dynamic document layout,
- An algorithm that fits in multiple languages document, even low-resource language,
- First big document dataset in Vietnamese with biggest number of images and biggest number of component levels, ranging from character-level to paragraph-level.
2 Related Work

There are two main approaches on gathering big datasets for document layout or any text document dataset in general. The first set of approach is to generate synthetic data that is close to real data that we are going to use. The second set of approach is to crawl real data and either label it manually or automatically.

2.1 Synthetic Text Dataset

Throughout the years, as deep learning becomes the mainstream in many tasks including optical character recognition, large amount of text images data for training deep model has become more and more important. However, there are many low-resource languages where we cannot crawl much of the data from open source like the Internet. Hence, synthetically generating text dataset comes naturally as an inevitable option for those languages. As far as we know, there are several attempts to generate synthetic data for optical character recognition. [3] provided the first synthetic scene text dataset, with most number of images. The datasets help a lot in pre-training a good model for further fine-tuning on many text detection dataset. In addition, it is well noted that many people used [3] open-source code to generate their own custom dataset in multiple languages.

Another approach by [4] combined the power of 3D Unreal Engine in generating persuasive text in the wild data. The data were surreal and can be useful in task like scene-text recognition. It also provides code that people can generate their own custom data and use for many languages.

There are also frameworks that focuses on generating synthetic text document. However, to the best of our knowledge, there has been no visual synthetic dataset made for document prior to ours.

2.2 Document Layout Dataset

There are several open source dataset on document layout. They range from detecting multiple components of the document at the same level like paragraph, table, list, etc. to small text level such as word. For example, [6] provided AR IIIT13k focusing on table detection, but contain several other document components. [author] provided DocBank, claiming the largest dataset that contains both document visual information and textual information. PublayNet [reference] is another dataset that is widely used for pretraining document layout detection due to its utility and usage of real data.

There are, however, disadvantages behind such data collecting approaches. First, the data is domain limited. For
example, DocBank and Publaynet works well only on
dataset that are close to medical document or research
document when they do zero-shot detection. Second,
crawling data is not possible if the targeted language and
dataset are either low-resource or private resources. It
means that for data whose visual are far from English
like Arabic or Thai, those datasets are not very effective.

3 Methods

We aimed at filling the document with 5 forms of doc-
ument components: "Title", "Paragraph", "Table", "Im-
age", "Formula". Thus, we divided our algorithm into a
2-step pipeline. First, we generate random math formula
and crawl image data for "Formula" and "Image" class
of component. Second, we use one of the two algorithms
discussed later to divide a whole page into multiple re-
gions to fill document components into. Finally, we use
Pillow library to fill text and fill image component into
a background paper. It would result in a large number of
document. We provided two options for data generation.
The first option is to generate document with layout that
are divided into specific number
of columns. This was inspired by the style that many
research document was presented in. By training the de-
tection algorithm on this generated dataset, we can tell
the algorithm to separate paragraph in vertical way even
though some of the text columns can be close to each
other. In addition, such dataset distribution would have a
similar distribution to PublayNet. Thus, we can be use it
as an alternative version of PublayNet on low-resource
languages.

3.1 Fixed columns layout algorithm

Algorithm 1: Fixed layout algorithms

Result: Images and Labels
Initialize numColumns;
while generating do
  randomize n in range(1,numColumns);
  divide a page into n columns;
  for each column do
    randomize column break position;
    divide the column into 2 parts along the
direction;
    fill text component into divided regions;
  end
end

With this option, we focus on the range of document
whose layouts are only divided into specific number
of columns. This was inspired by the style that many
research document was presented in. By training the de-
tection algorithm on this generated dataset, we can tell
the algorithm to separate paragraph in vertical way even
though some of the text columns can be close to each
other. In addition, such dataset distribution would have a
similar distribution to PublayNet. Thus, we can be use it
as an alternative version of PublayNet on low-resource
languages.

3.2 Flexible layout generation algorithm

The second option is to develop an algorithm that can
freely divide a page into multiple sections/ regions, and fill many information in such regions. Due to

the wide variety of document data structure in the
wild, this option helps model to explore a wider set of
data, thus can represent document model in a bet-
ter manner. Experiments on this was done and the
differences of pre-training models on flexible layout
SDL and pre-training models on other limited do-
main dataset would be shown in the next section.

Algorithm 2: Flexible layout algorithm

Result: Images and Labels
Initialize minimal area of components;
while component area not too small do
  randomize direction to divide;
  divide component area into 2 parts along the
direction;
  for each divided region do
    fill text component into the region;
end

Implementation details

For the Vietnamese synthetic
document dataset, we use fontsize ranging from 18 to
31, with the height and width of the image ranging from
1500 to 2500. The text color in the dataset are mainly
black with a small variance. The text are filled into
5 different categories: Paragraph, Title, Table, Figure,
and Formula. For Paragraph, Title and Table compo-
nent, text annotations are divided into 4 different levels:
component-level, line-level, word-level and character-
level.

4 Future work

There are several ways that we would like to proceed
with our data generation tool. First, multi-level text de-
tection models such as TextFuseNet [7] or CRAFT [8]
can be implemented on our dataset to get a better weight
on document model. Second, we can develop a light-
weight segmentation model that would take advantages
of a multi-level annotations in our dataset. Finally, we
would like to have a collaboration on generating dataset
in multiple languages other than Vietnamese or English,
which would bring benefits to more people who are us-
ing low-resource languages.

References

[1] Shaoqing Ren, Kaiming He, Ross B. Girshick, and
Jian Sun. Faster r-cnn: Towards real-time object
detection with region proposal networks. In Corinna
Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi
Sugiyama, and Roman Garnett, editors, NIPS, pages
91–99, 2015.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár,
and Ross Girshick. Mask r-cnn, 2017. cite
arxiv:1703.06870Comment: open source; appendix
on more results.
Figure 2: Layout generation algorithm illustration. Up: fixed column; Down: Flexible layout

[3] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] Shangbang Long and Cong Yao. Unrealtext: Synthesizing realistic scene text images from the unreal world. *arXiv preprint arXiv:2003.10608*, 2020.

[5] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep. 2019.

[6] Ajoy Mondal, Peter Lipps, and C. Jawahar. *IIIT-AR13K: A New Dataset for Graphical Object Detection in Documents*, pages 216–230. 08 2020.

[7] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. Textfusenet: Scene text detection with richer fused features. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 516–522. International Joint Conferences on Artificial Intelligence Organization, 2020.

[8] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.