LayoutLM: Pre-training of Text and Layout for Document Image Understanding

Yiheng Xu[1*], **Minghao Li**[2*], **Lei Cui**[3], **Shaohan Huang**[3], **Furu Wei**[3], **Ming Zhou**[3]
[1]Harbin Institute of Technology
[2]Beihang University
[3]Microsoft Research Asia
yihengxu@hit.edu.cn
liminghao1630@buaa.edu.cn
{lecu,shaohanh,fuwei,mingzhou}@microsoft.com

Abstract

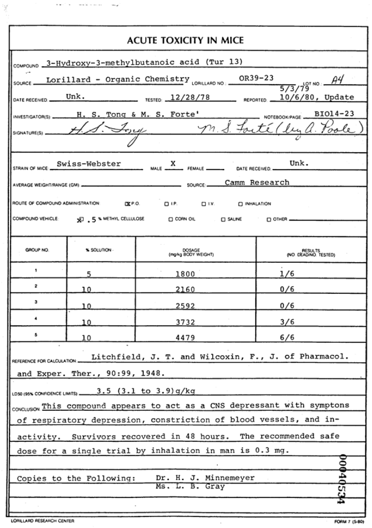
Pre-training techniques have been verified successfully in a variety of NLP tasks in recent years. Despite the widespread of pre-training models for NLP applications, they almost focused on text-level manipulation, while neglecting the layout and style information that is vital for document image understanding. In this paper, we propose the **LayoutLM** to jointly model the interaction between text and layout information across scanned document images, which is beneficial for a great number of real-world document image understanding tasks such as information extraction from scanned documents. Furthermore, we also leverage the image features to incorporate the visual information of words into LayoutLM. To the best of our knowledge, this is the first time that text and layout are jointly learned in a single framework for document-level pre-training. It achieves new state-of-the-art results in several downstream tasks, including form understanding (from 70.72 to 79.27), receipt understanding (from 94.02 to 95.24) and document image classification (from 93.07 to 94.42). The code and pre-trained LayoutLM models are publicly available at `https://github.com/microsoft/unilm/tree/master/layoutlm`.

1 Introduction

Document AI, or Document Intelligence[1], is a relatively new research topic that refers to the techniques to automatically read, understand and analyze business documents. Business documents are files that provide details related to a company's internal and external transactions, which is shown in Figure 1. They may be digital-born, occurring as electronic files, or they may be in scanned form that comes from written or printed on paper. Some common examples of business documents include purchase orders, financial reports, business emails, sales agreements, vendor contracts, letters, invoices, receipts, resumes and many others. Business documents are critical to a company's efficiency and productivity. The exact format of a business document may vary, but the information is usually presented in natural language and can be organized in a variety of ways from plain text, multi-column layouts, and a wide variety of tables/forms/figures. Understanding business documents is a very challenging task due to the diversity of layouts and formats, poor quality of scanned document images as well as the complexity of template structures.

Nowadays, many companies extract data from business documents through manual efforts that are time-consuming and expensive, meanwhile requiring manual customization or configuration. Rules and workflows for each type of document often need to be hard-coded and updated with changes to the specific format or when dealing with multiple formats. To address these problems, document AI models and algorithms are designed to automatically classify, extract and structuralize information from business documents, accelerating automated document processing workflows. Contemporary

*Equal contributions during internship at Microsoft Research Asia.
[1] `https://sites.google.com/view/di2019`



ACUTE TOXICITY IN MICE

COMPOUND 3-Hydroxy-3-methylbutanoic acid (Tur 13)
SOURCE Lorillard - Organic Chemistry LORILLARD NO. OR39-23 LOT NO. A4
DATE RECEIVED Unk. TESTED 12/28/78 REPORTED 5/3/79 10/6/80, Update
INVESTIGATOR(S) H. S. Tong & M. S. Forte' NOTEBOOK/PAGE BIO14-23
SIGNATURE(S) H. S. Tong M. S. Forte (by A. Poole)

STRAIN OF MICE Swiss-Webster MALE X FEMALE DATE RECEIVED Unk.
AVERAGE WEIGHT/RANGE (GM) SOURCE Camm Research
ROUTE OF COMPOUND ADMINISTRATION ☒ P.O. ☐ I.P. ☐ I.V. ☐ INHALATION
COMPOUND VEHICLE ☒ .5 % METHYL CELLULOSE ☐ CORN OIL ☐ SALINE ☐ OTHER

GROUP NO.	% SOLUTION	DOSAGE (mg/kg BODY WEIGHT)	RESULTS (NO. DEAD/NO. TESTED)
1	5	1800	1/6
2	10	2160	0/6
3	10	2592	0/6
4	10	3732	3/6
5	10	4479	6/6

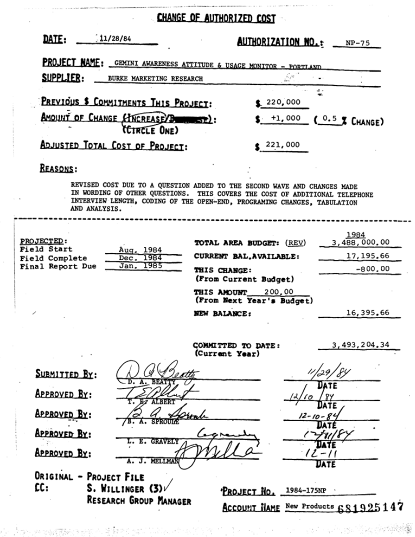
REFERENCE FOR CALCULATION Litchfield, J. T. and Wilcoxin, F., J. of Pharmacol. and Exper. Ther., 90:99, 1948.
LD50 (95% CONFIDENCE LIMITS) 3.5 (3.1 to 3.9)g/kg
CONCLUSION This compound appears to act as a CNS depressant with symptons of respiratory depression, constriction of blood vessels, and inactivity. Survivors recovered in 48 hours. The recommended safe dose for a single trial by inhalation in man is 0.3 mg.

Copies to the Following: Dr. H. J. Minnemeyer
Ms. L. B. Gray

00040534

LORILLARD RESEARCH CENTER FORM 7 (5/80)

(a)



CHANGE OF AUTHORIZED COST

DATE: 11/28/84 AUTHORIZATION NO.: NP-75
PROJECT NAME: GEMINI AWARENESS ATTITUDE & USAGE MONITOR - PORTLAND
SUPPLIER: BURKE MARKETING RESEARCH

PREVIOUS $ COMMITMENTS THIS PROJECT: $ 220,000
AMOUNT OF CHANGE (INCREASE/[redacted]): $ +1,000 (0.5 % CHANGE)
(CIRCLE ONE)
ADJUSTED TOTAL COST OF PROJECT: $ 221,000

REASONS:

REVISED COST DUE TO A QUESTION ADDED TO THE SECOND WAVE AND CHANGES MADE IN WORDING OF OTHER QUESTIONS. THIS COVERS THE COST OF ADDITIONAL TELEPHONE INTERVIEW LENGTH, CODING OF THE OPEN-END, PROGRAMING CHANGES, TABULATION AND ANALYSIS.

PROJECTED:
Field Start Aug. 1984
Field Complete Dec. 1984
Final Report Due Jan. 1985

TOTAL AREA BUDGET: (REV) 1984 3,488,000.00
CURRENT BAL.AVAILABLE: 17,195.66
THIS CHANGE: (From Current Budget) -800.00
THIS AMOUNT 200.00 (From Next Year's Budget)
NEW BALANCE: 16,395.66

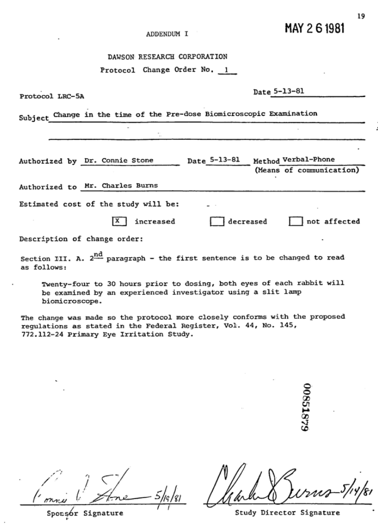
COMMITTED TO DATE: (Current Year) 3,493,204.34

SUBMITTED BY: D. A. BEATTY 11/29/84 DATE
APPROVED BY: T. E. ALBERT 12/10/84 DATE
APPROVED BY: B. A. SPROULE 12-10-84 DATE
APPROVED BY: L. E. GRAVELY 12/11/84 DATE
APPROVED BY: A. J. MELLMAN 12-11 DATE

ORIGINAL - PROJECT FILE
CC: S. WILLINGER (3)
RESEARCH GROUP MANAGER

PROJECT NO. 1984-175NP
ACCOUNT NAME New Products 681925147

(b)



19

MAY 26 1981

ADDENDUM I

DAWSON RESEARCH CORPORATION
Protocol Change Order No. 1

Protocol LRC-5A Date 5-13-81

Subject Change in the time of the Pre-dose Biomicroscopic Examination

Authorized by Dr. Connie Stone Date 5-13-81 Method Verbal-Phone (Means of communication)

Authorized to Mr. Charles Burns

Estimated cost of the study will be:

☒ increased ☐ decreased ☐ not affected

Description of change order:

Section III. A. 2nd paragraph - the first sentence is to be changed to read as follows:

Twenty-four to 30 hours prior to dosing, both eyes of each rabbit will be examined by an experienced investigator using a slit lamp biomicroscope.

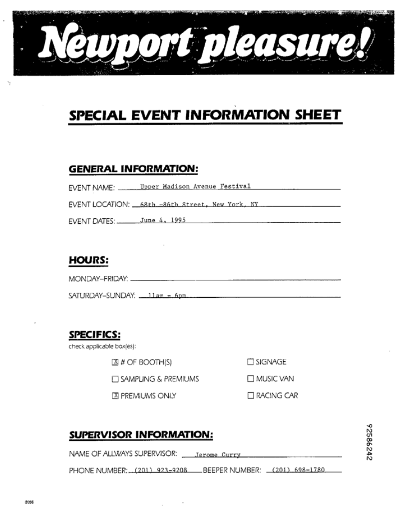
The change was made so the protocol more closely conforms with the proposed regulations as stated in the Federal Register, Vol. 44, No. 145, 772.112-24 Primary Eye Irritation Study.

00851879

Connie L. Stone 5/19/81 Charles Burns 5/14/81

Sponsor Signature Study Director Signature

(c)



Newport pleasure!

SPECIAL EVENT INFORMATION SHEET

GENERAL INFORMATION:

EVENT NAME: Upper Madison Avenue Festival
EVENT LOCATION: 68th -86th Street, New York, NY
EVENT DATES: June 4, 1995

HOURS:

MONDAY–FRIDAY:
SATURDAY–SUNDAY: 11am - 6pm

SPECIFICS:
check applicable box(es):

☒ # OF BOOTH(S) ☐ SIGNAGE
☐ SAMPLING & PREMIUMS ☐ MUSIC VAN
☒ PREMIUMS ONLY ☐ RACING CAR

SUPERVISOR INFORMATION:

NAME OF ALLWAYS SUPERVISOR: Jerome Curry
PHONE NUMBER: (201) 923-9208 BEEPER NUMBER: (201) 698-1780

92586242

3056

(d)

Figure 1: Scanned images of business documents with different layouts and formats

approaches for document AI are usually built upon deep neural networks from a computer vision perspective or a natural language processing perspective, or even a combination of them. Early attempts usually focused on detecting and analyzing certain parts of a document such as tabular areas. Hao et al. (2016) first proposed a table detection method for PDF documents based on Convolutional Neural Networks (CNN). After that, (Schreiber et al., 2017; Soto & Yoo, 2019; Zhong et al., 2019) also leveraged more advanced Faster R-CNN model (Ren et al., 2015) or Mask R-CNN model (He et al., 2017) to further improve the accuracy of document layout analysis. In addition, Yang et al. (2017) presented an end-to-end, multimodal, fully convolutional network for extracting semantic structures from document images, taking advantage of text embeddings from pre-trained NLP models. More recently, Liu et al. (2019a) introduced a Graph Convolutional Networks (GCN) based model to combine textual and visual information for information extraction from business documents. Although these models have made significant progress in the document AI area with deep neural networks, most of these methods confronted two limitations: (1) They only relied on a few human-labeled training samples, yet did not fully explore the possibility of using large-scale unla-

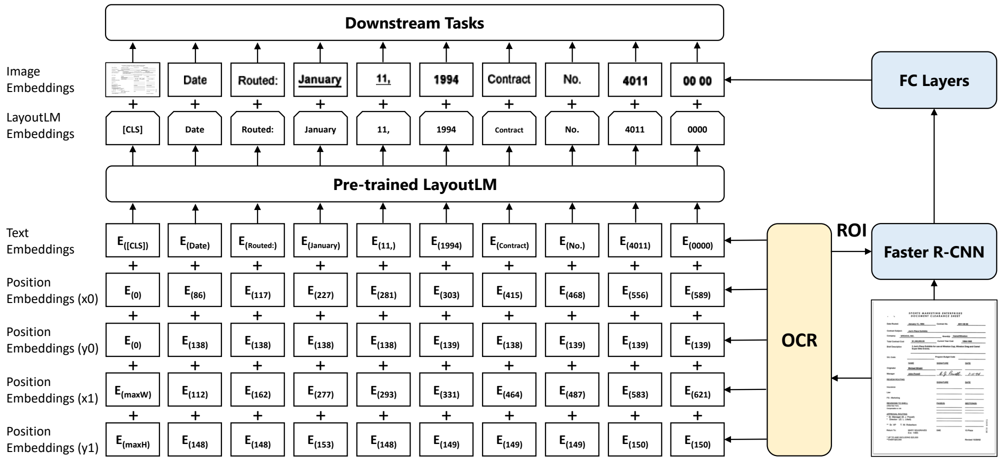


Figure 2: An example of LayoutLM, where 2-D layout and image embeddings are integrated into the original BERT architecture. The LayoutLM embeddings and image embeddings from Faster R-CNN work together for downstream tasks.

beled training samples. (2) They usually leveraged either pre-trained CV models or NLP models, but did not consider a joint training of textual and layout information. Therefore, it is indispensable to investigate how self-supervised pre-training of text and layout may help in the document AI area.

To this end, we propose LayoutLM, a simple but effective pre-training method of text and layout for document image understanding tasks. Inspired by the BERT model (Devlin et al., 2019), where input textual information is mainly represented by text embeddings and position embeddings, LayoutLM further adds two types of input embeddings: (1) a 2-D position embedding that denotes the relative position of a token within a document; (2) an image embedding for scanned token images within a document. The architecture of LayoutLM is shown in Figure 2. We add these two input embeddings because the 2-D position embeddings can capture the relationship among tokens within a document, meanwhile the image embedding can capture some appearance features such as font directions, types and colors. In addition, we adopt a multi-task learning objective for LayoutLM including a Masked Visual-Language Model (MVLM) loss and a Multi-label Document Classification (MDC) loss, which will further enforce the joint pre-training for text and layout. In this work, our focus is the document pre-training based on scanned document images, while digital-born documents are less challenging because they can be considered as a special case where OCR is not required, thus they are out of the scope of this paper. The LayoutLM is pre-trained on the IIT-CDIP Test Collection 1.0[2] (Lewis et al., 2006), which contains more than 6 million scanned documents with 11 million scanned document images. The scanned documents are in a variety of categories, including letter, memo, email, filefolder, form, handwritten, invoice, advertisement, budget, news articles, presentation, scientific publication, questionnaire, resume, scientific report, specification and many others, which is ideal for large-scale self-supervised pre-training. We select three benchmark datasets as the downstream tasks to evaluate the performance of the pre-trained LayoutLM model. The first is the FUNSD dataset[3](Jaume et al., 2019) that is used for spatial layout analysis and form understanding. The FUNSD dataset contains 199 fully annotated forms with 31,485 words and 9,707 semantic entities. The second is the SROIE dataset[4] for Scanned Receipts Information Extraction. The SROIE dataset contains 626 annotated receipts for training and 347 receipts for testing. The third is the RVL-CDIP dataset[5] (Harley et al., 2015) for document image classification, which consists of 400,000 grayscale images in 16 classes, with 25,000 images per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images. Experimental results illustrate that the pre-trained LayoutLM model significantly outperforms several SOTA pre-trained

[2] https://ir.nist.gov/cdip/
[3] https://guillaumejaume.github.io/FUNSD/
[4] https://rrc.cvc.uab.es/?ch=13
[5] https://www.cs.cmu.edu/~aharley/rvl-cdip/

models on these benchmark datasets, demonstrating the enormous advantage for pre-training of text and layout information in document image understanding tasks.

The contributions of this paper are summarized as follows:

- For the first time, textual and layout information from scanned document images is pre-trained in a single framework. Furthermore, image features are also leveraged to achieve new state-of-the-art results.
- LayoutLM uses the masked visual-language model and the multi-label document classification as the training objectives, which significantly outperforms several SOTA pre-trained models in document image understanding tasks.
- The code and the pre-trained LayoutLM models are publicly available at `https://github.com/microsoft/unilm/tree/master/layoutlm` for more downstream tasks.

2 LayoutLM

In this section, we briefly review the BERT model, and introduce how we extend to jointly model text and layout information in the LayoutLM framework.

2.1 The BERT model

The BERT model is an attention-based bidirectional language modeling approach. It has been verified that the BERT model shows effective knowledge transfer from the self-supervised task with large-scale training data. The architecture of BERT is basically a multi-layer bidirectional Transformer encoder. It accepts a sequence of discrete tokens and stacks multiple layers to produce final representations. In detail, given a set of tokens processed using WordPiece, the input embeddings are computed by summing the corresponding word embeddings, position embeddings, and segment embeddings. Then, these input embeddings are passed through a multi-layer bidirectional Transformer that can generate contextualized representations with an adaptive attention mechanism.

There are two steps in the BERT framework: pre-training and fine-tuning. During the pre-training, the model uses two objectives to learn the language representation: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), where MLM randomly masks some input tokens and the objective is to recover these masked tokens, and NSP is a binary classification task taking a pair of sentences as inputs and classifying whether they are two consecutive sentences. In the fine-tuning, task-specific datasets are used to update all parameters in an end-to-end way. The BERT model has been successfully applied in a set of NLP tasks.

2.2 The LayoutLM Model

Although BERT-like models become the state-of-the-art techniques on several challenging NLP tasks, they usually leverage text information only for any kind of inputs. When it comes to visually rich documents, there is much more information that can be encoded into the pre-trained model. Therefore, we propose to utilize the rich information from document layouts and align them with the input texts. Basically, there are two types of features which can significantly improve the language representation in a visually rich document, which are:

Document Layout Information It is evident that the relative positions of words in a document contribute a lot to the semantic representation. Taking form understanding as an example, given a key in a form, e.g. Passport ID:, its corresponding value is much more likely on its right or below instead of on the left or above. Therefore, based on the self-attention mechanism within the Transformer, embedding 2-D position features into the language representation will better align the layout information with the semantic representation.

Visual Information Compared with the text information, the visual information is another significantly important feature in document representations. Typically, documents contain some visual signals to show the importance and priority of document segments. The visual information can be represented by image features and effectively utilized in document representations. For document-level visual features, the whole image can indicate the document layout, which is an essential feature for document image classification. For word-level visual features, styles such as bold, underline, and italic, are also significant hints for the sequence labeling tasks. Therefore, we believe that combining the image features with traditional text representations can bring richer semantic representations to documents.

2.3 Model Architecture

To take advantage of existing pre-trained models and adapt to document image understanding tasks, we use the BERT architecture as the backbone and add two types of new input embeddings: a 2-D position embedding and an image embedding.

2-D Position Embedding Unlike the position embedding that models the word position in a sequence, 2-D position embedding aims to model the relative spatial position in a document. To represent the spatial position of elements in scanned document images, we consider a document page as a coordinate system with the top-left origin. In this setting, the bounding box can be precisely defined by (x0, y0, x1, y1), where (x0, y0) corresponds to the position of the upper left in the bounding box, and (x1, y1) represents the position of the lower right. So we add four position embedding layers with two embedding tables. The embedding layers which represent the same dimension share the same embedding table, which means that we look up the position embedding of x0 and x1 in the embedding table X and look up y0 and y1 in the table Y.

Image Embedding To utilize the image feature of a document and align the image feature with the text, we add an image embedding layer to represent image features in language representation. In more detail, with the bounding box of each word from OCR results, we split the image into several pieces, and they have a one-to-one correspondence with the words. We generate the image region features with these pieces of images from the Faster R-CNN (Ren et al., 2015) model as the token image embeddings. For the `[CLS]` token, we also use the Faster R-CNN model to produce embeddings using the whole scanned document image as the Region of Interest (ROI) to benefit the downstream tasks which need the representation of the `[CLS]` token.

2.4 Pre-training LayoutLM

Task #1: Masked Visual-Language Model Inspired by the masked language model, we propose the Masked Visual-language Model (MVLM) to learn the language representation with the clues of 2-D position embeddings and text embeddings. During the pre-training, we randomly mask some of the input tokens but keep the 2-D position embeddings and other text embeddings, then the model is trained to predict the masked tokens given the contexts. In this way, the LayoutLM model not only understands the language contexts, but also utilizes the corresponding 2-D position information, thereby bridging the gap between the visual and language modalities.

Task #2: Multi-label Document Classification For document image understanding, many tasks require the model to generate high-quality document-level representations. As the IIT-CDIP Test Collection includes multiple tags for each document image, we also use a Multi-label Document Classification (MDC) loss during the pre-training phase. Given a set of scanned documents, we use the document tags to supervise the pre-training process so that the model can cluster the knowledge from different domains and generate better document-level representation. Since the MDC loss needs the label for each document image that may not exist for larger datasets, it is optional during the pre-training and may not be used for pre-training larger models in the future. We will compare the performance of MVLM and MVLM+MDC in Section 3.

2.5 Fine-tuning LayoutLM

The pre-trained LayoutLM model is fine-tuned on three document image understanding tasks, including a form understanding task, a receipt understanding task as well as a document image classification task. For the form and receipt understanding tasks, LayoutLM predicts {B, I, E, S, O} tags for each token and uses sequential labeling to detect each type of entity in the dataset. For the document image classification task, LayoutLM predicts the class labels using the representation of the `[CLS]` token.

3 EXPERIMENTS

3.1 PRE-TRAINING DATASET

The performance of pre-trained models is largely determined by the scale and quality of datasets. Therefore, we need a large-scale scanned document image dataset to pre-train the LayoutLM model. Our model is pre-trained on the IIT-CDIP Test Collection 1.0, which contains more than 6 million documents, with more than 11 million scanned document images. Moreover, each document has its corresponding text and metadata stored in XML files. The text is the content produced by applying OCR to document images. The metadata describes the properties of the document such as the unique identity and document labels. Although the metadata contains erroneous and inconsistent tags, the scanned document images in this large-scale dataset are perfectly suitable for pre-training our model. In addition, we exclude the RVL-CDIP portion from the IIT-CDIP dataset during the pre-training step as the RVL-CDIP is used for evaluation.

3.2 FINE-TUNING DATASET

The FUNSD Dataset We evaluate our approach on the FUNSD dataset for form understanding in noisy scanned documents. This dataset includes 199 real, fully annotated, scanned forms with 9,707 semantic entities and 31,485 words. These forms are organized as a list of semantic entities that are interlinked. Each semantic entity comprises a unique identifier, a label (i.e., question, answer, header or other), a bounding box, a list of links with other entities, and a list of words. The dataset is split into 149 training samples and 50 testing samples. We adopt the word-level F1 score as the evaluation metric.

The SROIE Dataset We also evaluate our model on the SROIE dataset for receipt information extraction (Task 3). The dataset contains 626 receipts for training and 347 receipts for testing. Each receipt is organized as a list of text lines with the bounding boxes. Meanwhile, each receipt is labeled with four type of entities which are {company, date, address, total}. The evaluation metric is the exact match of the entity recognition results in F1 score.

The RVL-CDIP Dataset The RVL-CDIP dataset consists of 400,000 grayscale images in 16 classes, with 25,000 images per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images. The images are sized so their largest dimension does not exceed 1,000 pixels. The 16 classes include {letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo}. The evaluation metric is the overall classification accuracy.

3.3 DOCUMENT PRE-PROCESSING

To utilize the layout information of each document, we need to obtain the location of each token. However, the pre-training dataset (IIT-CDIP Test Collection) only contains pure texts while missing their corresponding bounding boxes. In this case, we re-process the scanned document images to obtain the necessary layout information. Like the original pre-processing in IIT-CDIP Test Collection, we similarly process the dataset by applying OCR to document images. The difference is that we obtain both the recognized words and their corresponding locations in the document image. Thanks to Tesseract[6], an open-source OCR engine, we can easily obtain the recognition as well as the 2D positions. We store the OCR results in hOCR format, a standard specification format which clearly defines the OCR results of one single document image using a hierarchical representation.

[6] `https://github.com/tesseract-ocr/tesseract`

3.4 Model Pre-training

We initialize the weight of LayoutLM model with the pre-trained BERT base model. Specifically, our BASE model has the same architecture: a 12-layer Transformer with 768 hidden sizes, and 12 attention heads, which contains about 113M parameters. Therefore, we use the BERT base model to initialize all modules in our model except the 2D position embedding layer. For the LARGE setting, our model has a 24-layer Transformer with 1,024 hidden sizes and 16 attention heads, which is initialized by the pre-trained BERT LARGE model and contains about 343M parameters. Following (Devlin et al., 2019), we select 15% of the input tokens for prediction. We replace these masked tokens with the `[MASK]` token 80% of the time, a random token 10% of the time, the unchanged token 10% of the time. Then, the model will predict the corresponding token with the cross-entropy loss.

In addition, we also add the 2D position embedding layers with four embedding representations (x_0, y_0, x_1, y_1), where (x_0, y_0) corresponds to the position of the upper left in the bounding box, and (x_1, y_1) represents the position of the lower right. Considering that the document layout may vary in different page size, we scale the actual coordinate to a "virtual" coordinate: the actual coordinate is scaled to have a value from 0 to 1,000. Furthermore, we also use the ResNet-101 model as the backbone network in the Faster R-CNN model, which is pre-trained on the Visual Genome dataset (Krishna et al., 2016).

We train our model on 8 NVIDIA Tesla V100 32GB GPUs with a total batch size of 80. The Adam optimizer is used with an initial learning rate of 5e-5 and a linear decay learning rate schedule. The BASE model takes 80 hours to finish one epoch on 11M documents, while the LARGE model takes nearly 170 hours to finish one epoch.

3.5 Task-specific Fine-tuning

We evaluate the LayoutLM model on three different document image understanding tasks: **Form Understanding**, **Receipt Understanding**, and **Document Image Classification**. We follow the typical fine-tuning strategy and update all parameters in an end-to-end way on task-specific datasets.

Form Understanding This task requires extracting and structuring the textual content of forms. It aims to extract key-value pairs from the scanned form images. In more detail, this task includes two sub-tasks: semantic labeling and semantic linking. Semantic labeling is the task of aggregating words as semantic entities and assigning pre-defined labels to them. Semantic linking is the task of predicting the relations between semantic entities. In this work, we focus on the semantic labeling task while semantic linking is out of the scope. To fine-tune LayoutLM on this task, we treat the semantic labeling task as a sequence labeling problem. We pass the final representation into a linear layer followed by a softmax layer to predict the label of each token. The model is trained for 100 epochs with a batch size of 16 and a learning rate of 5e-5.

Receipt Understanding This task requires filling several pre-defined semantic slots according to the scanned receipt images. For instance, given a set of receipts, we fill specific slots like company, address, date, and total. Different from the form understanding task that requires labeling all matched entities and key-value pairs, the number of semantic slots is fixed with pre-defined keys. Therefore, the model only needs to predict the corresponding values using the sequence labeling method.

Document Image Classification Given a visually rich document, this task aims to predict the corresponding category for each document image. Distinct from the existing image-based approaches, our model includes not only image representations but also text and layout information using the multimodal architecture in LayoutLM. Therefore, our model can combine the text, layout, and image information in a more effective way. To fine-tune our model on this task, we concatenate the output from the LayoutLM model and the whole image embedding, followed by a softmax layer for category prediction. We fine-tune the model for 30 epochs with a batch size of 16 and a learning rate of 2e-5.

3.6 RESULTS

Form Understanding We evaluate the form understanding task on the FUNSD dataset. The experiment results are shown in Table 1. We compare the LayoutLM model with two SOTA pre-trained NLP models: BERT and RoBERTa (Liu et al., 2019b). The BERT BASE model achieves 0.603 and while the LARGE model achieves 0.656 in F1. Compared to BERT, the RoBERTa performs much better on this dataset as it is trained using larger data with more epochs. Due to the time limitation, we present 4 settings for LayoutLM, which are 500K document pages with 6 epochs, 1M with 6 epochs, 2M with 6 epochs as well as 11M with 2 epochs. It is observed that the LayoutLM model substantially outperforms existing SOTA pre-training baselines. With the BASE architecture, the LayoutLM model with 11M training data achieves 0.7866 in F1, which is much higher than BERT and RoBERTa with the similar size of parameters. In addition, we also add the MDC loss in the pre-training step and it does bring substantial improvements on the FUNSD dataset. Finally, the LayoutLM model achieves the best performance of 0.7927 when using the text, layout and image information at the same time.

Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	0.5469	0.671	0.6026	110M
	$\text{RoBERTa}_{\text{BASE}}$	0.6349	0.6975	0.6648	125M
	$\text{BERT}_{\text{LARGE}}$	0.6113	0.7085	0.6563	340M
	$\text{RoBERTa}_{\text{LARGE}}$	0.678	0.7391	0.7072	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.665	0.7355	0.6985	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.7377	0.782	0.7592	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.7597	0.8155	0.7866	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7076	0.7695	0.7372	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.7194	0.7780	0.7475	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.7171	0.805	0.7585	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	0.7536	0.806	0.7789	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7101	0.7815	0.7441	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**0.7677**	**0.8195**	**0.7927**	160M

Table 1: Model accuracy (Precision, Recall, F1) on the FUNSD dataset

In addition, we also evaluate the LayoutLM model with different data and epochs on the FUNSD dataset, which is shown in Table 2. For different data settings, we can see that the overall accuracy is monotonically increased as more epochs are trained during the pre-training step. Furthermore, the accuracy is also improved as more data is fed into the LayoutLM model. As the FUNSD dataset contains only 149 images for fine-tuning, the experiment results confirmed that the pre-training of text and layout is effective for scanned document understanding especially the low resource setting.

Furthermore, we compare different initialization methods for the LayoutLM model with BERT and RoBERTa. The experiment results in Table 3 show that the $\text{LayoutLM}_{\text{BASE}}$ model initialized with $\text{RoBERTa}_{\text{BASE}}$ outperforms $\text{BERT}_{\text{BASE}}$ by 2.1 points in F1. For the LARGE setting, the $\text{LayoutLM}_{\text{LARGE}}$ model initialized with $\text{RoBERTa}_{\text{LARGE}}$ further improve 1.3 points over the $\text{BERT}_{\text{LARGE}}$ model. We will pre-train more models with RoBERTa as the initialization in the future, especially for the LARGE settings.

Receipt Understanding We evaluate the receipt understanding task using the SROIE dataset. The results are shown in Table 4. As we only test the performance of the Key Information Extraction task in SROIE, we would like to eliminate the effect of incorrect OCR results. Therefore, we pre-process the training data by using the groundtruth OCR and run a set of experiments using the baseline models (BERT & RoBERTa) as well as the LayoutLM model. The experiment results show that the LayoutLM LARGE model trained with 11M document images achieve an F1 score of 0.9524, which is significantly better than the first place in the competition leaderboard. This result also verifies that the pre-trained LayoutLM not only performs well on the in-domain dataset (FUNSD) but also outperforms several strong baselines on the out-of-domain dataset like SROIE.

# Pre-training Data	# Pre-training Epochs	Precision	Recall	F1
500K	1 epoch	0.5779	0.6955	0.6313
	2 epochs	0.6217	0.705	0.6607
	3 epochs	0.6304	0.718	0.6713
	4 epochs	0.6383	0.7175	0.6756
	5 epochs	0.6568	0.734	0.6933
	6 epochs	0.665	0.7355	0.6985
1M	1 epoch	0.6156	0.7005	0.6552
	2 epochs	0.6545	0.737	0.6933
	3 epochs	0.6794	0.762	0.7184
	4 epochs	0.6812	0.766	0.7211
	5 epochs	0.6863	0.7625	0.7224
	6 epochs	0.6909	0.7735	0.7299
2M	1 epoch	0.6599	0.7355	0.6957
	2 epochs	0.6938	0.759	0.7249
	3 epochs	0.6915	0.7655	0.7266
	4 epochs	0.7081	0.781	0.7427
	5 epochs	0.7228	0.7875	0.7538
	6 epochs	0.7377	0.782	0.7592
11M	1 epoch	0.7464	0.7815	0.7636
	2 epochs	**0.7597**	**0.8155**	**0.7866**

Table 2: LayoutLM$_{\mathrm{BASE}}$ (Text + Layout, MVLM) accuracy with different data and epochs on the FUNSD dataset

Initialization	Model	Precision	Recall	F1
BERT$_{\mathrm{BASE}}$	LayoutLM$_{\mathrm{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299
RoBERTa$_{\mathrm{BASE}}$	LayoutLM$_{\mathrm{BASE}}$ (1M, 6 epochs)	0.7173	0.7888	0.7514
BERT$_{\mathrm{LARGE}}$	LayoutLM$_{\mathrm{LARGE}}$ (11M, 1 epoch)	0.7536	0.8060	0.7789
RoBERTa$_{\mathrm{LARGE}}$	LayoutLM$_{\mathrm{LARGE}}$ (11M, 1 epoch)	**0.7681**	**0.8188**	**0.7926**

Table 3: Different initialization methods for BASE and LARGE (Text + Layout, MVLM)

Document Image Classification Finally, we evaluate the document image classification task using the RVL-CDIP dataset. Document images are different from other natural images as most of the content in document images are texts in a variety of styles and layouts. Traditionally, image-based classification models with pre-training perform much better than the text-based models, which is shown in Table 5. We can see that either BERT or RoBERTa underperforms the image-bases approaches, illustrating that text information is not sufficient for this task and it still needs layout and image features. We address this issue by using the LayoutLM model for this task. The experiment results show that, even without the image features, LayoutLM still outperforms the single model of the image-based approaches. After integrating the image embeddings, the LayoutLM achieves the accuracy of 94.42%, which is significantly better than several SOTA baselines for document image classification.

Furthermore, we generate the confusion matrix for our best result (94.42%) on the test set, which is shown in Figure 3. From this matrix, we can see that our model performs best on the ”email” category while performs worst on the ”form” category. We will further investigate how to take advantage of both pre-trained LayoutLM and image models, as well as involve the image information in the pre-training step for the LayoutLM model.

[7] `https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=3`

[8] `https://medium.com/@jdegange85/benchmarking-modern-cnn-architectures-to-rvl-cdip-9dd0b7ec2955`

Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\mathrm{BERT}_{\mathrm{BASE}}$	0.9099	0.9099	0.9099	110M
	$\mathrm{RoBERTa}_{\mathrm{BASE}}$	0.9107	0.9107	0.9107	125M
	$\mathrm{BERT}_{\mathrm{LARGE}}$	0.9200	0.9200	0.9200	340M
	$\mathrm{RoBERTa}_{\mathrm{LARGE}}$	0.9280	0.9280	0.9280	355M
Text + Layout MVLM	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (500K, 6 epochs)	0.9388	0.9388	0.9388	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	0.9380	0.9380	0.9380	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (2M, 6 epochs)	0.9431	0.9431	0.9431	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 2 epochs)	0.9438	0.9438	0.9438	113M
Text + Layout MVLM+MDC	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	0.9402	0.9402	0.9402	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 1 epoch)	0.9460	0.9460	0.9460	113M
Text + Layout MVLM	$\mathrm{LayoutLM}_{\mathrm{LARGE}}$ (1M, 6 epochs)	0.9416	0.9416	0.9416	343M
	$\mathrm{LayoutLM}_{\mathrm{LARGE}}$ (11M, 1 epoch)	**0.9524**	**0.9524**	**0.9524**	343M
Text + Layout + Image MVLM	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	0.9416	0.9416	0.9416	160M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 2 epochs)	0.9467	0.9467	0.9467	160M
Baseline	Ranking 1^{st} in SROIE[7]	0.9402	0.9402	0.9402	-

Table 4: Model accuracy (Precision, Recall, F1) on the SROIE dataset

Modality	Model	Accuracy	#Parameters
Text only	$\mathrm{BERT}_{\mathrm{BASE}}$	89.81%	110M
	$\mathrm{RoBERTa}_{\mathrm{BASE}}$	90.06%	125M
	$\mathrm{BERT}_{\mathrm{LARGE}}$	89.92%	340M
	$\mathrm{RoBERTa}_{\mathrm{LARGE}}$	90.11%	355M
Text + Layout MVLM	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (500K, 6 epochs)	91.25%	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	91.48%	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (2M, 6 epochs)	91.65%	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 2 epochs)	91.78%	113M
Text + Layout MVLM+MDC	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	91.74%	113M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 1 epoch)	91.78%	113M
Text + Layout MVLM	$\mathrm{LayoutLM}_{\mathrm{LARGE}}$ (1M, 6 epochs)	91.88%	343M
	$\mathrm{LayoutLM}_{\mathrm{LARGE}}$ (11M, 1 epoch)	91.90%	343M
Text + Layout + Image MVLM	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (1M, 6 epochs)	94.31%	160M
	$\mathrm{LayoutLM}_{\mathrm{BASE}}$ (11M, 2 epochs)	**94.42%**	160M
Baselines	VGG-16 (Afzal et al., 2017)	90.97%	-
	Single model (Das et al., 2018)	91.11%	-
	Ensemble (Das et al., 2018)	92.21%	-
	InceptionResNetV2[8] (Szegedy et al., 2016)	92.63%	-
	LadderNet (Sarkhel & Nandi, 2019)	92.77%	-
	Single model (Dauphinee et al., 2019)	93.03%	-
	Ensemble (Dauphinee et al., 2019)	93.07%	-

Table 5: Classification accuracy on the RVL-CDIP dataset

4 RELATED WORK

The research of Document Analysis and Recognition (DAR) dates to the early 1990s. The mainstream approaches can be divided into three categories: rule-based approaches, conventional machine learning approaches and deep learning approaches.

4.1 RULE-BASED APPROACHES

The rule-based approaches (Lebourgeois et al., 1992; O'Gorman, 1993; Ha et al., 1995b; Simon et al., 1997) contain two types of analysis methods: bottom-up and top-down. The bottom-up meth-

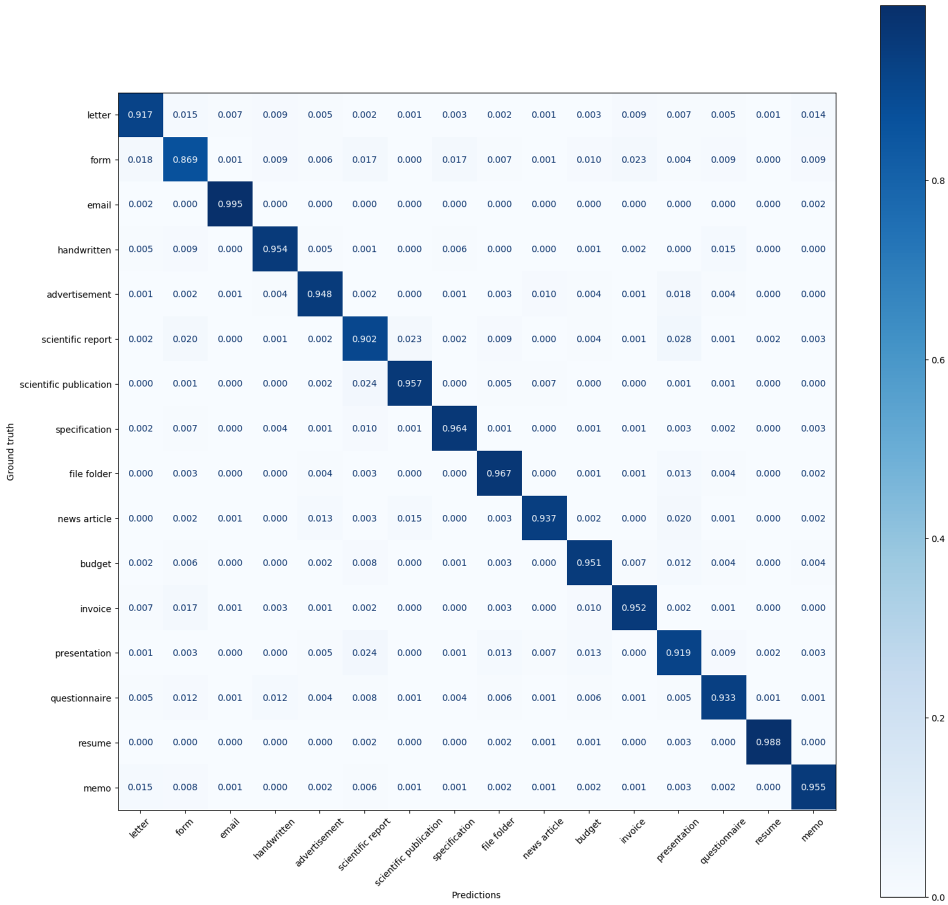


Figure 3: Confusion matrix for the best result (94.42%) on the test set

ods (Lebourgeois et al., 1992; Ha et al., 1995a; Simon et al., 1997) usually detected the connected components of black pixels as the basic computational units in document images, and the document segmentation process is to combine them into higher-level structures through different heuristics and label them according to different structural features. Docstrum algorithm (O'Gorman, 1993) is among the earliest successful bottom-up algorithms that is based on connected component analysis. It groups connected components on a polar structure to derive the final segmentation. Simon et al. (1997) used a special distance-metric between different components to construct a physical page structure. They further reduced the time complexity by using heuristics and path compression algorithms.

The top-down methods often recursively split a page into columns, blocks, text lines and tokens. Ha et al. (1995b) proposed to replace the basic unit with the black pixels from all the pixels, and the method decomposed the document using the recursive the X-Y cut algorithm to establish a X-Y tree, which made complex documents decompose more easily. Although these methods perform well on some documents, they require extensive human efforts to figure out better rules, while sometimes failing to generalize to documents from other sources. Therefore, it is inevitable to leverage statistical machine learning approaches in the DAR research.

4.2 Machine Learning Approaches

With the development of conventional machine learning, statistical machine learning approaches (Shilman et al., 2005; Marinai et al., 2005) became the mainstream for document segmentation tasks during the past decade. Shilman et al. (2005) considered the layout information of a document as a parsing problem, and globally searched the optimal parsing tree based on a grammar-based loss function. They utilized a machine learning approach to select features and train all parameters during the parsing process. Meanwhile, artificial neural networks (Marinai et al., 2005) have been extensively applied to document analysis and recognition. Most efforts have been devoted to the recognition of isolated handwritten and printed characters with widely recognized successful results. In addition to the ANN model, SVM and GMM (Wei et al., 2013) have been used in document layout analysis tasks. For machine learning approaches, they are usually time-consuming to design manually crafted features and difficult to obtain a highly abstract semantic context. In addition, these methods usually relied on visual cues but ignored textual information.

4.3 Deep Learning Approaches

Recently, deep learning methods have become the mainstream and de facto standard for many machine learning problems. Theoretically, they can fit any arbitrary functions through the stacking of multi-layer neural networks and have been verified to be effective in many research areas. Yang et al. (2017) treated the document semantic structure extraction task as a pixel-by-pixel classification problem. They proposed a multimodal neural network that considers visual and textual information, while the limitation of this work is that they only used the network to assist heuristic algorithms to classify candidate bounding boxes rather than an end-to-end approach. Viana & Oliveira (2017) proposed a lightweight model of document layout analysis for mobile and cloud services. The model used one-dimensional information of images for inference and compared with the model using two-dimensional information, which achieved comparable accuracy in the experiments. Katti et al. (2018) made use of a fully convolutional encoder-decoder network that predicts a segmentation mask and bounding boxes, and the model significantly outperforms approaches based on sequential text or document images. Soto & Yoo (2019) incorporated contextual information into the Faster R-CNN model that involves the inherently localized nature of article contents to improve the region detection performance.

The existing deep learning approaches for DAR usually confronted two limitations: (1) The models often relied on limited labeled data while leaving a large amount of unlabeled data unused. (2) Current deep learning models usually leveraged pre-trained CV models or NLP models, but did not consider the joint pre-training of text and layout. LayoutLM addresses these two limitations and achieves much better performance over the previous baselines.

5 Conclusion and Future Work

We present LayoutLM, a simple but effective pre-training technique with text and layout information in a single framework. Based on the Transformer architecture as the backbone, LayoutLM takes advantage of multimodal inputs including token embeddings, layout embeddings and image embeddings. Meanwhile, the model can be easily trained in a self-supervised way based on large scale unlabeled scanned document images. We evaluate the LayoutLM model on three tasks: form understanding, receipt understanding and scanned document image classification. Experimental results show that LayoutLM substantially outperforms a set of SOTA pre-trained model in these tasks.

For the future research, we will investigate pre-training models with more data and more computation resources. In addition, we will also train LayoutLM using the LARGE architecture with text and layout, as well as involving image embeddings in the pre-training step. Furthermore, we will also explore other self-supervised training objectives and strategies that may further unlock the power of LayoutLM.

References

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:883–888, 2017.

Arindam Das, Saikat Roy, and Ujjwal Bhattacharya. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3180–3185, 2018.

Tyler Dauphinee, Nikunj Patel, and Mohammad Mehdi Rashidi. Modular multimodal architecture for document classification. *ArXiv*, abs/1912.04376, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. Document page decomposition by the bounding-box project. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pp. 1119–1122. IEEE, 1995a.

Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. Recursive xy cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pp. 952–955. IEEE, 1995b.

Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang. A table detection method for pdf documents based on convolutional neural networks. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp. 287–292, 2016.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 991–995, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL `http://arxiv.org/abs/1703.06870`.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:1–6, 2019.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4459–4469, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1476. URL `https://www.aclweb.org/anthology/D18-1476`.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL `https://arxiv.org/abs/1602.07332`.

Frank Lebourgeois, Z Bublinski, and H Emptoz. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems*, pp. 272–276. IEEE, 1992.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pp. 665–666, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148307. URL `http://doi.acm.org/10.1145/1148170.1148307`.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pp. 32–39, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-2005. URL `https://www.aclweb.org/anthology/N19-2005`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019b.

S. Marinai, M. Gori, and G. Soda. Artificial neural networks for document analysis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):23–35, Jan 2005. ISSN 1939-3539. doi: 10.1109/TPAMI.2005.4.

L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162–1173, Nov 1993. ISSN 1939-3539. doi: 10.1109/34.244677.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

Ritesh Sarkhel and Arnab Nandi. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3360–3366. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/466. URL `https://doi.org/10.24963/ijcai.2019/466`.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:1162–1167, 2017.

Michael Shilman, Percy Liang, and Paul Viola. Learning nongenerative grammatical models for document analysis. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pp. 962–969. IEEE, 2005.

Anikó Simon, J-C Pret, and A Peter Johnson. A fast algorithm for bottom-up document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):273–277, 1997.

Carlos Soto and Shinjae Yoo. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3462–3468, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1348. URL `https://www.aclweb.org/anthology/D19-1348`.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2016.

Matheus Palhares Viana and Dário Augusto Borges Oliveira. Fast cnn-based document layout analysis. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1173–1180, 2017.

H. Wei, M. Baechler, F. Slimane, and R. Ingold. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 1220–1224, Aug 2013. doi: 10.1109/ICDAR.2013.247.

Xiaowei Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4342–4351, 2017.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: largest dataset ever for document layout analysis. *ArXiv*, abs/1908.07836, 2019.