

Enhancing Document Information Analysis with Multi-Task Pre-training: A Robust Approach for Information Extraction in Visually-Rich Documents

Tofik Ali, Partha Pratim Roy

Abstract—This paper introduces a deep learning model tailored for document information analysis, emphasizing document classification, entity relation extraction, and document visual question answering. The proposed model leverages transformer-based models to encode all the information present in a document image, including textual, visual, and layout information. The model is pre-trained and subsequently fine-tuned for various document image analysis tasks. The proposed model incorporates three additional tasks during the pre-training phase, including reading order identification of different layout segments in a document image, layout segments categorization as per Pub-LayNet, and generation of the text sequence within a given layout segment (text block). The model also incorporates a collective pre-training scheme where losses of all the tasks under consideration, including pre-training and fine-tuning tasks with all datasets, are considered. Additional encoder and decoder blocks are added to the RoBERTa network to generate results for all tasks. The proposed model achieved impressive results across all tasks, with an accuracy of 95.87% on the RVL-CDIP dataset for document classification, F1 scores of 0.9306, 0.9804, 0.9794, and 0.8742 on the FUNSD, CORD, SROIE, and Kleister-NDA datasets respectively for entity relation extraction, and an ANLS score of 0.8468 on the DocVQA dataset for visual question answering. The results highlight the effectiveness of the proposed model in understanding and interpreting complex document layouts and content, making it a promising tool for document analysis tasks.

Index Terms—Document Information Analysis, Document Classification, Entity Relation Extraction, Document Visual Question Answering, Deep Learning, Transformer Models, Pre-training Tasks.

I. INTRODUCTION

A. Background

Document information analysis, a pivotal component of Document AI, entails extracting visual information, often termed visual information extraction (VIE). This area of research has garnered significant attention from academic researchers and industry professionals. It primarily focuses on understanding and interpreting visually rich documents (VrDs), such as forms and receipts. These interpretations involve the semantic entities recognition (SER) and the subsequent relations extraction (RE). How document layouts are geometrically represented plays a crucial role in this endeavour. Recent advancements have seen pre-training techniques revolutionize the Document AI domain, yielding significant

improvements in document comprehension tasks. Such pre-trained models can dissect the layout and pinpoint essential data from diverse documents, ranging from scanned forms to scholarly articles. This capability has notable implications in both industrial applications and academic investigations.

Transformer-based models [1], [2], [3], rooted in deep learning paradigms, aspire to encapsulate all facets of information present in a document image, encompassing the textual, visual, and layout dimensions. Once this information is encoded within the pre-trained model, it is subsequently fine-tuned for various document image analytical tasks. The prowess of these models is evident in their performance in VIE assignments, especially in SER tasks. However, there's a notable disparity regarding the RE task, which targets identifying relationships between semantic entities within documents. This particular task poses inherent challenges and needs to be more explored.

B. Motivation

While the aforementioned transformer-based models have exhibited commendable results in many areas, recent findings indicate a mismatch in relationships inferred by these models when tasked with RE within document images [4]. The challenges with RE stem from the objective gap between the pre-training and fine-tuning phases of these models. Although some works have proposed specific pre-training tasks to bridge this gap, results on datasets like FUNSD have not consistently demonstrated expected improvements. This implies potential hidden issues and emphasizes the importance of more refined layout representation in pre-trained models.

Moreover, the realization that current models might not efficiently comprehend the geometric layout nuances of documents makes it clear that there's an urgent need to investigate more discriminative methods of understanding document layouts. Incorporating explicit modelling of geometric relationships during the pre-training phase emerges as a promising solution to these challenges.

C. Objectives

The primary objective of this paper is to present a model tailored to address the challenges mentioned earlier. By incorporating innovative pre-training tasks and methodologies, the proposed model significantly enhances RE task performance. Emphasis has been placed on discerning the reading sequence of layout segments within document images, categorizing

T. Ali and P.P. Roy are with the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, India, e-mail: tali@cs.iitr.ac.in, e-mail: proy.fcs@iitr.ac.in

these layout segments according to standards such as PubLayNet, and generating text sequences within identified layout segments or text blocks.

D. Contributions

The salient contributions of this research are threefold:

- A novel collective pre-training strategy is introduced. This strategy incorporates the losses from all tasks into each parameter update step during pre-training. By leveraging datasets like RVL-CDIP [5], FUNSD [6], CORD [7], SROIE [8], Kleister-NDA [9], DocVQA [10], and PubLayNet [11], we not only improve the quality of the pre-training data but also manage to reduce its size from 11 million (IIT-CDIP) to just one million, resulting in less pre-training overall computation.
- Integration of three distinct tasks during the model's pre-training phase. These tasks involve: 1) categorizing layout segments as per PubLayNet, 2) determining the reading order amongst varied layout segments in document images, and 3) producing text sequences within specified layout segments or text blocks.
- Enhancement of the RoBERTa [12] network architecture by integrating additional encoder and decoder blocks. These blocks are pivotal for generating results across all tasks while minimizing disruption to the core RoBERTa network. The proposed model's pre-training phase fully utilizes these blocks, updating the relevant parameters in line with their respective losses.

Building on the findings of previous studies [3], [4], this paper introduces the contributions as mentioned above to advance the field of Document AI, particularly in relation extraction tasks.

II. RELATED WORK

A. Progress in Multimodal Self-Supervised Pre-training for Document Intelligence

Document Intelligence, a specialized segment of artificial intelligence aimed at distilling valuable knowledge from textual documents, has immensely benefited from multimodal self-supervised pre-training methodologies. These approaches synthesize text, visuals, and layout data to deepen the comprehension of document frameworks.

The pioneering efforts in this domain can be credited to LayoutLM [1], [2], [3] and its further iterations. These models were trailblazers in amalgamating layout representation by assimilating spatial coordinates of text [1], [13], [14], [15]. These methodologies hinge on the premise that the strategic positioning of text and visuals in a document offers indispensable insights into its content and structure, facilitating processes such as entity recognition and document classification.

Recognizing the potential of visual data, an innovative fusion of convolutional neural networks (CNNs) and transformer attention mechanisms was explored [16]. Initial endeavours sought to harness CNN grid features [2], [17] or capitalize on region features via object detection tools [1], [18], [19], [20].

Nevertheless, these ventures often encountered computational hurdles or required specific regional directives.

The evolution in natural image analytics, especially in the domain of vision-and-language pre-training (VLP), paved the way for a paradigm shift from region features to grid features [21], [22], [23], [24]. This adjustment sought to counteract the limitations associated with preset object categories and region-centric guidance. Vision Transformers (ViT) [25] championed the concept of leveraging image embeddings, sidestepping the traditional CNNs. This concept introduced a fresh array of VLP techniques. Despite several methodologies adhering to distinct self-attention mechanisms, ViLT [26] marked a turning point by deploying a consolidated linear layer for visual feature extraction, which translated to a notable reduction in model size and processing duration. Building on this foundation, LayoutLMv3 [3] was conceptualized, standing out as the avant-garde multimodal model in Document AI, uniquely operating without the dependency on CNNs.

B. Evolution through Reconstructive Pre-Training Paradigms

The landscape of representation learning has witnessed transformative changes owing to reconstructive pre-training paradigms. These paradigms endeavour to encapsulate the intrinsic nature of data, thereby enhancing model adaptability across a spectrum of tasks. In the realm of natural language processing (NLP), "masked language modelling" (MLM) marked a pivotal shift. This strategy, propelled by BERT [27], aimed at bidirectional learning by obscuring certain input words and prompting the model to deduce them from the surrounding context. This approach set new benchmarks for several language comprehension tasks as a foundation for subsequent advancements [3].

In the domain of computer vision (CV), a parallel approach, termed Masked Image Modeling (MIM), surfaced. This approach took cues from NLP's MLM, intending to deduce concealed image portions based on the observable context. Vision Transformer (ViT) [25] epitomized this by estimating the average shade of concealed segments, augmenting its prowess on tasks like ImageNet categorization. BEiT [28] further expanded on this concept, zeroing in on visual token revival, and yielded notable outcomes in visual categorization and semantic demarcation. Shifting the focus to documents, Document Transformer (DiT) harnessed these principles for document image layout examination [29].

Drawing inspiration from the triumphant trajectories of MLM and MIM in NLP and CV, respectively, scholars blending vision and language embarked on probing reconstructive paradigms tailored for multimodal representation. While the foundational idea of obscuration and prediction persisted, the actual masking modalities were refined according to image embedding specifics. This evolution led to three distinct MIM variants: regional obscuration modelling (MRM), grid-based obscuration modelling (MGM), and segment-based obscuration modelling (MPM). Among these, MRM demonstrated prowess in regaining original regional attributes or discerning labels of obscured regions. In contrast, MGM adeptly deduced visual lexicon linkages for obscured grids. When it came to

segment-level representations, platforms like ViLT [26] and METER forged paths reminiscent of ViT [25] and BEiT [28], underscoring the technique’s promise, even if they sometimes faced hurdles in certain assignments.

The advent of LayoutLMv3 [3] signified a noteworthy evolution. Inspired by ViLT [26], it emerged as Document AI’s pioneer in harnessing image embeddings sans the use of CNNs. This shift not only optimized computations but also enhanced the nuances of document layout portrayals. Moreover, it ratified the potential of MIM in handling linear segment image embeddings, casting light on promising avenues for further exploration [3], [4].

To sum up, reconstructive pre-training paradigms have charted a transformative course in representation learning across disciplines, such as audio signal processing, medical imaging, social network analysis, and e-commerce product categorization. Bridging the capabilities of NLP and CV, they have paved the way for innovative models adept at comprehending and managing multimodal content, contributing substantially to intricate endeavours such as document layout examination [3], [4].

C. Extraction of Visual Information from Document Images

Deriving Visual Information (VDI) from document images is pivotal for automating the comprehension of documents. This mainly addresses tasks such as Semantic Entity Identification (SEI) and Entity Relationship Analysis (ERA)[30], [6], [31]. Earlier VDI approaches majorly employed Graph Neural Networks (GNN)[32], [33], focusing on acquiring features that represent text and layout components for immediate VDI tasks.

The introduction of pre-training approaches has profoundly expanded the scope of document comprehension. Researchers have ingeniously proposed several pre-training tasks aimed at enhancing textual and visual features while ensuring a solid alignment for a sturdy multimodal document interpretation [31], [34], [35], [1], [2]. While significant strides have been made in SEI, ERA is still a less explored domain [30], [31], [36]. Significantly, BROS [30] integrated spatial text positional data into the BERT architecture [27], amplifying layout interpretation. This paper emphasizes harnessing pre-training approaches to garner enhanced features for document examination.

An essential aspect of VDI is the spatial details within document structures, which often act as innate indicators for document layout interpretation. As an illustration, Liu et al.[37] deployed 2D spatial positions with GNN, while GraphNEMR[32] combined geometry proximities and distance metrics for SEI. Innovations such as SPADE [38] revamped the self-attention mechanism, embedding spatial vectors that incorporate coordinates, distances, and angular data. StrucText [31] introduced a task during pre-training to determine the geometric orientation between text blocks. Yet, there’s an evident shortfall as most techniques focus solely on pair-level spatial connections. Our approach seeks to expand these connections to multiple pairs and groups, ensuring a comprehensive investigation.

Furthermore, as pinpointed by [4], mismatches between the pre-training and task-specific tuning stages often pose intricate

issues. Many recent endeavors [39], [40], [41], [42], [43] are directed at bridging this inconsistency. For example, Hu et al.[42] discerned gaps in the training design and task knowledge, ingeniously synchronizing the subsequent ranking task more seamlessly with a pre-training design. The emergence of prompt-centric models has forged a route, allowing models to adapt across diverse contexts by converting specific tasks to associated prompts, aligning seamlessly with the pre-training design[43]. Drawing from these advancements, our research infuses spatial tasks during the pre-training stage, ensuring superior incorporation of spatial understanding. This bolsters the model’s adaptability, especially in relation extraction, by capitalizing on extensive pre-training resources.

D. Aligning Vision and Language in Multimodal Frameworks

Integrating visual and textual information in multimodal models poses distinct challenges, primarily regarding the formation of a cohesive understanding from both modalities. Pioneering work in this arena has spanned both overarching and intricate mechanisms for vision-language (VL) alignment.

At the broader scale, VL alignment seeks to comprehend overarching relationships between visual elements and their associated text. One key approach in this realm has been matching images with their corresponding textual descriptions, serving to deepen our grasp on the synergy between visual and textual contexts [3]. Such general alignments form the foundational understanding for more refined alignment techniques.

Digging deeper, intricate VL alignment targets the intricate mappings between individual textual units and specific regions in images. Document images, distinct from typical pictures, often present clear correlations between their text and visual sections. UNITER [21], for instance, implements a word-region correlation mechanism via optimal transports, calculating the least resource-intensive way to relate image-based contextual representations with individual words [3]. ViLT [26] then broadens this goal, aiming for patch-level image contextualizations.

Recognizing the idiosyncrasies of document images, UDoc leveraged techniques like contrastive learning and distilled similarities to sharpen the correlation of text and image sectors within the same segments [3]. In a related vein, LayoutLMv2 honed this by introducing a masking technique on particular textual sections in images, prompting the model to anticipate the hidden textual components.

A recurring tactic across these methods has been the application of masking, exemplified by practices in Masked Image Modeling (MIM), which helps pinpoint both correlated and non-correlated pairs [3]. This underlines the importance of flexibility in VL alignment strategies. Especially concerning document images, tailoring alignment methods to the dataset’s particularities is imperative.

Straying from models that predominantly focus on general images, specialized architectures such as Luo’s GeoLayoutLM address challenges specific to document images [3]. Given the inherent structured relationships in document images, they open doors to pioneering alignment techniques like contrastive learning. This fosters a more seamless connection between the visual and textual aspects.

Conclusively, the interconnection of vision and text in multimodal models is a thriving domain. The continual emergence of innovative methodologies and richer data sources indicates that alignment methods will evolve in response to the multifaceted requirements of diverse visual-textual contexts.

E. Bridging the Gap between Pre-training and Fine-tuning

Achieving seamless progression from the pre-training phase to the fine-tuning phase is fundamental in today's deep learning landscape, especially in the context of document intelligence. Misalignment between these two stages can hinder a model's efficacy in subsequent tasks, as the model might find it challenging to apply its pre-acquired knowledge to the specialized requirements of fine-tuning tasks.

Recent studies spotlight two predominant chasms: discrepancies in the training objectives and variances in the knowledge needed for distinct tasks, as pointed out by [4]. The former relates to the incongruities in goals and architectures across the pre-training and fine-tuning phases. The latter pertains to the specific insights needed for a target task which might not be sufficiently grasped during pre-training. These challenges have birthed innovative methodologies. For example, Hu et al. [42] aimed to bridge the objective discrepancy by aligning the goal of a subsequent ranking task with a pre-training one. Their findings revealed that such a harmonized training approach fostered more effective knowledge transfer.

Moreover, the emergence of prompt-based models has made the transition from pre-training to fine-tuning more seamless [4]. By shaping downstream tasks into compatible prompts consistent with the pre-training objectives, these models enhance coherence across both stages. This results in increased adaptability and sets the stage for enhanced generalization for an array of tasks.

In the domain of document layout analysis, [4] shed light on a formidable challenge when employing transformer-based models for RE tasks. The complex task of amalgamating textual, visual, and layout cues emphasized a marked objective misalignment during the RE's two phases. To counteract this, certain geometric-centric tasks were introduced to endow the model with pivotal geometric insights, aiming at better representation generalization harnessed from extensive pre-training datasets.

The emphasis on geometric comprehension has proven invaluable. Prior endeavours, such as StrucText [31], integrated geometric orientation details of textual segments in the pre-training phase. Yet, these primarily centred on exploring dyadic geometric relations. The avant-garde approach of extending these relations to encompass multi-pair and triplet configurations [4] offers a more comprehensive perspective of document layouts.

To encapsulate, establishing a seamless bridge between the pre-training and fine-tuning stages is paramount for the success of models in diverse document analysis endeavours. As contemporary research showcases, refinements in training paradigms, task-oriented prompts, and an amplified emphasis on geometric insights can yield models endowed with resilience and adaptability across various document analysis challenges.

III. PROPOSED METHOD

The proposed method involves a deep learning-based model that utilizes transformer-based models to encode all the information in a document image. This model is pre-trained and subsequently fine-tuned for various document image analysis tasks.

A. Proposed Model Architecture

The architecture of the proposed model is influenced by the LayoutLMv3 and GeoLayoutLM models, with alterations made to better accommodate the tasks at hand. The architecture is designed to encode all the information in a document image, including textual, visual, and layout information.

1) *Overview of the Proposed Model:* The proposed model is a deep learning-based model that integrates additional tasks during the pre-training phase. These tasks include identifying the reading order of different layout segments in a document image, categorizing layout segments as PubLayNet, and generating the text sequence within a given layout segment (text block). The model also incorporates a collective pre-training scheme where losses of all the tasks under consideration, including pre-training and fine-tuning tasks with all datasets, are considered. Additional encoder and decoder blocks are added to the RoBERTa network to generate results for all tasks (refere the Figure 1 for visual insights.).

2) *Detailed Model Architecture:* The model architecture consists of an independent text-embedding layer, patch embedding layer, position encoding, and multi-head attention (MHA) based encoder and decoder layers. The input of the proposed model is the document images and the text-line level OCR toolkit output. The input is first processed by corresponding text and patch embedding layers and converted into a latent feature vector. This vector is further added with respective encoding and then processed by the sequence of MHA layers.

3) *Text Embedding:* The OCR generates the text-line (Note: text with sufficient gap are considered different text-lines) level output with bounding box coordinates (coordinates are rescaled such that the longer side of the document image becomes 512). Each text line is first tokenized into a sequence of tokens, and then for each token, a latent feature vector is created by the text embedding layer (same as RoBERTa); we call these vectors a text latent feature vector and represented as V_T .

4) *Patch Embedding:* The document image is first resized such that the longer side of the document image becomes 512. The resized image is then divided into non-overlapping patches of size 32×32 . These patches are converted into vectors by flattening them. Then, a linear transformation (linear dense layer) is applied to convert them into latent feature vectors; we call these vectors visual latent feature vectors and represent them as V_V .

5) *Position Embedding:* There are two position embeddings: 1) Segment box embedding and 2) token sequence embedding within a segment box.

A segment's box is represented by $BOX_{seg} = x_1, y_1, x_2, y_2$ where x_1, y_1 are the coordinates of the left-top corner point of the box whereas the x_2, y_2 is the right-bottom corner point.

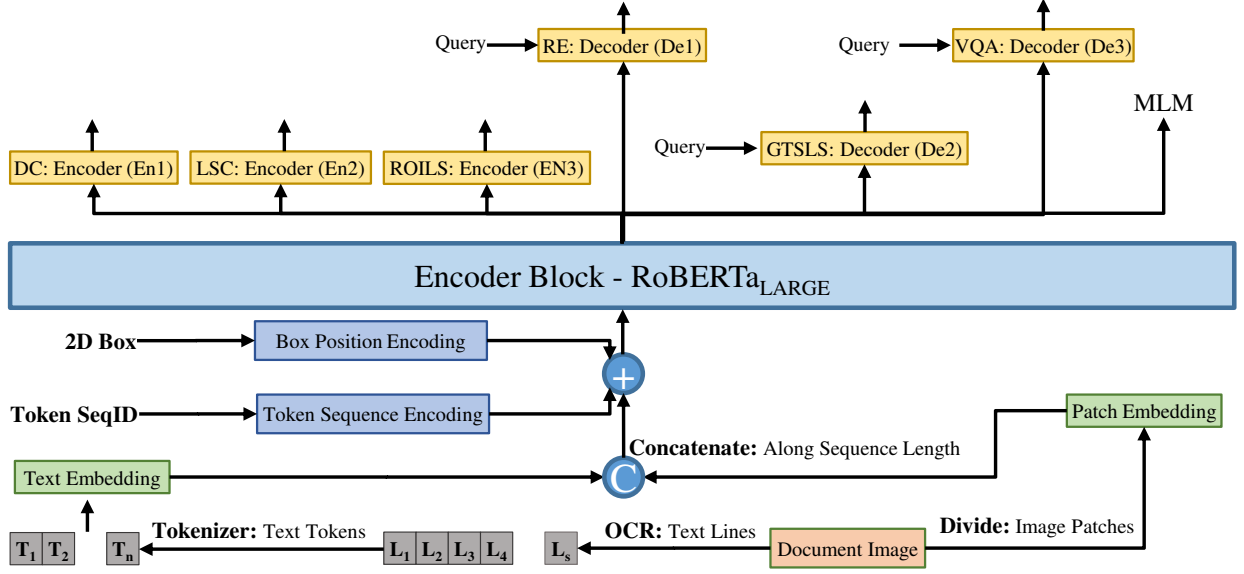


Fig. 1. A schematic of the suggested model showcasing various encoder and decoder units. Initially, text lines and their bounding boxes are identified via OCR. These lines undergo tokenization using a RoBERTa-based tokenizer. Subsequently, embeddings are produced for each token. Simultaneously, the input image is segmented into 32×32 patches, flattened and converted into patch embedding vectors. Both text and patch embeddings are combined with position encodings. The resulting vector sequence is then processed by the RoBERTa framework, passing through the designated encoder or decoder units, culminating in the final outputs.

The coordinates $BOX_{seg} = x_1, y_1, x_2, y_2 \in [1, 512]$ thus a box (0,0,0,0) is not a valid box and utilized only when box coordinates are masked or not available. The 1D encoding vectors for the $x_1, y_1, x_2, and y_2$ values are concatenated to form the segment box position vector. We use P_B symbol to represent this vector.

A token sequence id is represented by $T_{seqid} \in [1, 512]$; thus, a $T_{seqid} = 0$ is not a valid token sequence id and is utilized only when there is no sequence available. The 1D encoding vector for the T_{seqid} is used as the token sequence position vector. We use P_t symbol to represent this vector.

The P_B and P_t position vectors are added together to form the final position vector, which is further added to the latent vectors V_T or V_V . The resultant vectors are then processed by the sequence of MHA layers.

6) *Addition of Encoder and Decoder Blocks to RoBERTa Network*: Additional encoder and decoder blocks are added to the RoBERTa network to generate results for all tasks. The pre-training of the proposed model utilizes all these encoder and decoder blocks and updates the corresponding parameters as their respective losses. Different blocks are added according to different tasks; refer to section III-B "Pre-training Tasks" for better insights.

B. Pre-training Tasks

The proposed model incorporates several pre-training tasks to enhance its performance. These tasks include:

1) *Masked Language Modeling (MLM)*: All datasets are used for this task in the pre-training phase of the model parameter learning. The masked language modelling inspires the MLM task in BERT [27]. A percentage of text tokens are masked with a span masking strategy, and the pre-training objective is to maximize the log-likelihood of the correct

masked text tokens based on the contextual representations of corrupted sequences of image tokens and text tokens.

2) *Document Categorization (DC)*: The RVL-CDIP dataset [5] is employed for the document image classification task. We have added a single MHA (cross attention) encoder layer (En1) with [CLS] token as the query vector with $BOX_{seg} = 0, 0, 0, 0$ and $T_{seqid} = 0$. Further, a linear transformation layer is added, followed by softmax activation for the document categorization.

3) *Layout Segment Categorization (LSC)*: The PubLayNet dataset [11] is used for the layout segment categorization. We have added a single MHA (cross attention) encoder layer (En2) with [CLS] token as the query vector with $BOX_{seg} = layout_segment(x_1, y_1, x_2, y_2)$, $T_{seqid} = 0$. Further, a linear transformation layer is added, followed by softmax activation for the document categorization.

4) *Reading Order Identification of Layout Segments (ROILS)*: The PubLayNet dataset [11] is used for the reading order identification of a given set of layout segments. We have added a single MHA (cross attention) encoder layer (En3) with a sequence of segment boxes as the query vector (each segment is encoded as $BOX_{seg} = layout_segment(x_1, y_1, x_2, y_2)$, $T_{seqid} = 0$). This layer generates the output vector corresponding to each layout segment. These output vectors are further used in their reading order identification as introduced in the ERNIE[44] model. This task helps the model to understand the flow of information in a document.

5) *Relation Extraction (RE)*: There are 4 different datasets, FUNSD [6], CORD [7], SROIE [8], Kleister-NDA [9], used for this task. The relation extraction task involves identifying the relationships between different entities in the document. The output sequence length is also not fixed; therefore, we employed a decoder block for this task.

We have added a block of two MHA (casual and cross attention) decoder layers (De1) with the description text of the required relation as query (each token of this text has $BOX_{seg} = (0, 0, 0, 0)$). This block generates the next text token until the [EOS] token is generated.

6) *Generation of Text Sequence within Layout Segments (GTSLS)*: The PubLayNet dataset [11] is used for this task. Due to the same reason, this task also requires a decoder block. Therefore, We have added a block of two MHA (casual and cross attention) decoder layers (De2) with [SOS] token as the query vector with $BOX_{seg} = layout_segment(x_1, y_1, x_2, y_2)$, $T_{seqid} = 0$. This block generates the next text token until the [EOS] token is generated.

7) *Visual Question Answering (VQA)*: In the visual question-answering task, the model is trained to generate the answer for a question according to the information available inside the document image. This task helps the model understand the document’s content in a more detailed and comprehensive manner. The DocVQA [10] dataset is used for this task. We have added a block of three MHA (casual and cross attention) decoder layers (De3) with the question text as query (each token of this text has $BOX_{seg} = (0, 0, 0, 0)$). This block generates the next text token until the [EOS] token is generated.

IV. EXPERIMENTAL SETUP

It is important to note that the experimental setup, which includes the datasets and evaluation metrics, is in line with the most recent models, such as LayoutLMv3 and GeoLayoutLM [3], [4]. This makes sure that our experiments are measured against the most recent progress in the field. We are utilising 4 GPUs, V100, for the pre-training and fine-tuning the proposed model.

TABLE I
STATISTICS OF DATASETS

Dataset	# of keys or categories	# of examples (train/dev/test)
RVL-CDIP [5]	16	320K/4K/4K
FUNSD [6]	4	149/0/50
CORD [7]	30	800/100/100
SROIE [8]	4	626/0/347
Kleister-NDA [9]	4	254/83/203
DocVQA [10]	—	39K/5K/5K
PubLayNet [11]	5	335703/11245/11405

A. Datasets

The statistics of different datasets used in all of our experiments regarding pre-training and fine-tuning of the proposed model is listed in Table I. The descriptions of different datasets according to their targeted tasks are as follows:

1) *Dataset for Layout Segment Analysis*: The PubLayNet dataset [11] stands out as a substantial resource for research on document layout analysis, mainly when using deep learning approaches. The dataset encompasses a comprehensive collection of research paper images, each meticulously annotated with bounding boxes and polygonal segmentations. The main

objective behind this annotation is to distinctly categorise various segments of the document layout, which have been broadly divided into five categories: text, title, list, figure, and table.

The dataset’s structure and organisation have been designed to facilitate both training and evaluation processes. The official partitioning of the dataset includes a training set comprising 335,703 images, a validation set with 11,245 images, and a test set with 11,405 images. The division ensures that the models can be thoroughly trained on vast data and then validated and tested for performance consistency.

Moreover, the inclusion of different layout structures in PubLayNet aligns well with the proposed work’s objective of reading order identification, layout segment categorisation, and generating the text sequence within specific layout segments. Utilising this dataset for pre-training tasks, as intended in the proposed model, could bridge the objective gap observed between the pre-training and fine-tuning phases.

Note: This dataset is utilised only for the pre-training phase of the model parameter learning as we do not have the required components to get the layout segment bounding box from the given document image (an object detection network or a segmentation-based network). A pictorial depiction of the utilisation of this dataset in different pre-training tasks is given in Figure 2.

2) *Dataset for Document Classification*: The RVL-CDIP dataset [5] is employed for the document image classification task. This dataset is a subset of the IIT-CDIP collection, containing document images labelled with 16 distinct categories. The RVL-CDIP dataset comprises 400,000 document images distributed into 320,000 training images, 40,000 validation images, and 40,000 test images. Text and layout information are extracted using the Microsoft Read API [3]

3) *Dataset for Entity Relation Extraction*: For entity relation extraction, we used two primary datasets:

FUNSD Dataset [6]: The FUNSD dataset is a collection of noisy scanned forms for document understanding and analysis. It consists of 199 documents that provide comprehensive annotations for 9,707 semantic entities. The objective is to label each semantic entity as “question”, “answer”, “header”, or “other”. The dataset is split into 149 training and 50 test samples.

CORD Dataset [7]: CORD is dedicated to the extraction of key information from receipts. It is segmented into 800 training, 100 validation, and 100 test samples and houses 30 semantic labels distributed under four categories.

SROIE Dataset [8]: The SROIE dataset, unveiled by Huang et al. during the ICDAR2019 Competition for Scanned Receipt OCR and Information Extraction, encompasses 1000 entire scanned receipt images along with annotations. The dataset is divided into a training set of 626 documents and a testing set of 347 documents, categorised under four distinct classes. It was crafted for a contest centred on OCR (Optical Character Recognition) and the extraction of crucial information from scanned receipts. The SROIE dataset is a valuable asset for the progress, assessment, and refinement of OCR and key information extraction methodologies, particularly in scanned receipts.

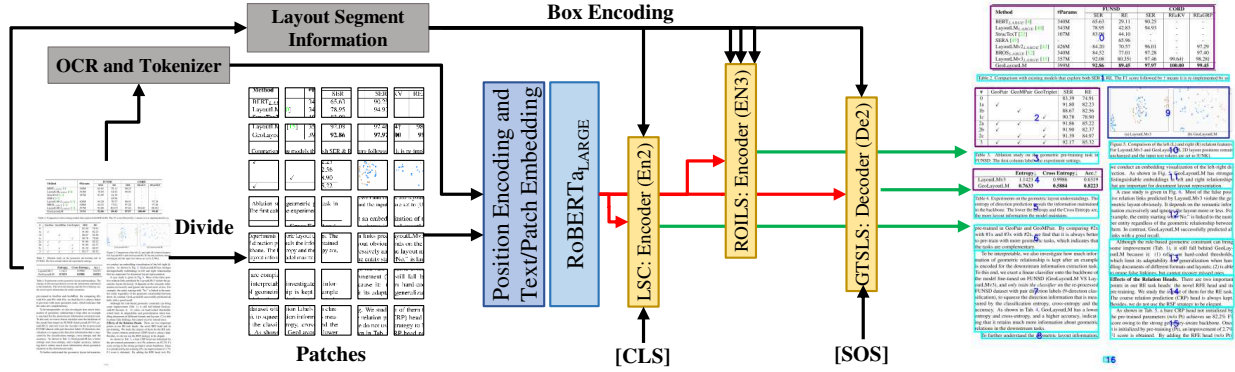


Fig. 2. Illustration of utilising the PubLayNet dataset for pre-training tasks in the proposed model. The dataset, containing a rich collection of document layout images, is employed to pre-train the model parameters for subsequent layout segment analysis tasks. The figure depicts how the dataset facilitates reading order identification, layout segment categorisation, and text sequence generation within specific layout segments, aligning with the objectives of the proposed work.

Kleister-NDA [9] The “Kleister-NDA” dataset, tailored for Key Information Extraction (KIE) tasks, includes a blend of scanned and born-digital formal English documents, mainly Non-disclosure Agreements (NDAs). It houses 540 NDAs (training set: 254, validation set: 83, test set: 20), encompassing 3,229 unique pages and 2,160 entities for extraction. Initially sourced from the Electronic Data Gathering, Analysis and Retrieval system (EDGAR) via Google, these documents were converted to PDFs and annotated for entity extraction through a two-tier process involving three annotators and a super-annotator.

4) Dataset for Document Visual Question Answering: For the document visual question answering (DocVQA) task, the DocVQA dataset [10] is employed. This dataset is designed for visual question answering over document images and includes a partition of 10,194 training images, 1,286 validation images, and 1,287 test images. Additionally, there are 39,463 questions for the training set, 5,349 for the validation set, and 5,188 for the test set. The objective is to input a document image alongside a question and expect the model to produce an accurate answer.

B. Evaluation Metrics

1) Metrics for Document Classification: The main evaluation metric used for document classification is accuracy, which measures the proportion of correctly classified documents against the total documents.

2) Metrics for Entity Relation Extraction: The F1 score is employed as the primary metric for the entity relation extraction task. The F1 score provides a harmonic mean of precision (ratio of correctly predicted positive observations to the total predicted positives) and recall (ratio of correctly predicted positive observations to all observations in actual class). It is especially useful when the dataset has imbalanced classes.

3) Metrics for Document Visual Question Answering: For document visual question answering, the commonly-used edit distance-based metric ANLS (also known as Average Normalized Levenshtein Similarity) is reported.

C. Model Pre-training

The proposed model leverages the rich information in collecting datasets for different document analysis tasks. The collection utilised in the proposed work is consist of the dataset as RVL-CDIP [5], FUNSD [6], CORD [7], SROIE [8], Kleister-NDA [9], DocVQA [10], and PubLayNet [11]. The collective training set of all these datasets is one million (1M). Building upon the foundational works of LayoutLMv3 [3] and GeoLayoutLM [4]. The matched layers of the proposed model are initialised from the weightings of RoBERTa [12] and the remaining are initialised from random distribution, same as LayoutLMv3 [3]. This hybrid approach seamlessly merges textual and visual understanding, providing a more holistic representation of documents. The Adam optimiser is used to pre-train the model with a sizeable batch size of 32 over 2,500,000 steps. We have samples from all datasets for each batch as RVL-CDIP: 8, FUNSD: 2, CORD: 2, SROIE: 2, Kleister-NDA: 2, DocVQA: 8, PubLayNet: 8 samples.

D. Model Fine-tuning

Each task necessitates a unique fine-tuning approach:

- **Document Classification:** Based on the experiments conducted on the RVL-CDIP [5] dataset, the model is adjusted over 50,000 steps. This fine-tuning employs a batch size of 32 with a learning rate of 2×10^{-5} .
- **Entity Relation Extraction:** Four datasets, FUNSD [6], CORD [7], SROIE [8], Kleister-NDA [9], are considered for this task. The model undergoes a fine-tuning with batch size 32, epochs 100, and the learning rate is adjusted to 3×10^{-5} for each dataset separately.
- **Document Visual Question Answering:** This task, based on the DocVQA dataset, requires the model to interpret both a document image and a related question, aiming to output a relevant answer. The base model is adjusted over 500,000 steps using a batch size of 32, a learning rate of 2×10^{-5} , and a warmup ratio of 0.05.

V. RESULT AND ANALYSIS

This section presents the results obtained from the proposed model and provides a detailed analysis. The results are cate-

TABLE II
COMPARATIVE CLASSIFICATION ACCURACIES ON THE RVL-CDIP DATASET. NOTE: "R/G/P" IMPLIES THE "REGION/GRID/PATCH" IMAGE EMBEDDING TYPES.

Modality	Model	Image Embedding	Pre-Training		Accuracy	#Parameters
			Dataset Size	Epochs		
Text only	BERT _{BASE}	None	-	-	89.81%	110M
	RoBERTa _{BASE}	None	-	-	90.06%	125M
	BERT _{LARGE}	None	-	-	89.92%	340M
	RoBERTa _{LARGE}	None	-	-	90.11%	355M
Text + Layout + Image	DocFormer _{BASE} [17]	(G) ResNet-50	-	-	96.17%	183M
	LayoutLMv1 _{BASE} [1]	(R) ResNet-101	11M	2	94.42%	160M
	LayoutLMv2 _{BASE} [2]	(G) ResNeXt101-FPN	11M	5	95.25%	200M
	LayoutLMv3 _{BASE} [3]	(P) Linear	11M	90	95.44%	133M
	DocFormer _{LARGE} [17]	(G) ResNet-50	-	-	95.50%	536M
	LayoutLMv2 _{LARGE} [2]	(G) ResNeXt101-FPN	11M	20	95.25%	426M
	LayoutLMv3 _{LARGE} [3]	(P) Linear	11M	90	95.44%	368M
	Our Proposed Model BackBone network RoBERTa _{LARGE}	(P) Linear	1M	80	95.87%	390M

gorized based on the tasks performed by the model.

A. Analysis of Document Classification Performance

In analyzing the results from our experiments on the RVL-CDIP dataset (Table II), the proposed model registered an impressive accuracy of 95.87%. This achievement places our model at a competitive standing amidst leading-edge models in the document classification domain.

Several observations and takeaways can be gleaned from these outcomes:

- **The Power of Hybrid Training:** One of the most distinguishing features of our model's success was its ability to harness information from textual, visual, and layout sources. It is evident from Table II that models leveraging a combination of these elements consistently outperformed text-only models.
- **Intrinsic Value of Enhanced Pre-training:** The meticulous incorporation of specialized pre-training tasks, such as reading order identification and layout segment categorization, arguably fortified the model's capability to discern document layouts and their inherent content. This intrinsic understanding was particularly pivotal in accurately classifying complex documents with sophisticated layouts.
- **Efficiency in Parameter Utilization:** Despite having a parameter count of 390M, which is considerably fewer than some counterparts, our model was able to maintain competitive performance. This speaks to its efficiency and optimization in parameter utilization.
- **Comparative Analysis:** In direct comparison with the widely recognized LayoutLM series and DocFormer models, our proposed model stands out due to its superior understanding of document layout and content and its unique pre-training schemes and architecture adjustments.
- **Image Embedding Insights:** The models that employ diverse image embeddings showcase varied performance. For instance, the "patch" type of image embedding used in our proposed model and LayoutLMv3 seems to strike a harmonious balance between accuracy and model complexity.

In conclusion, the advancements we integrated into the model's architecture and training regimen have borne fruit in the form of enhanced classification accuracy. This reinforces the significance of tailored pre-training tasks and architectural refinements in deep learning-based document analysis endeavors.

B. Analysis and Interpretation of Entity Relation Extraction Results

The proposed model was subjected to rigorous testing on four distinct datasets for the entity relation extraction task, namely FUNSD, CORD, SROIE, and Kleister-NDA. As depicted in Table III, the model achieved impressive F1 scores across all datasets, with 0.9306 on FUNSD, 0.9804 on CORD, 0.9794 on SROIE, and 0.8742 on Kleister-NDA.

These high F1 scores are indicative of the model's superior performance in accurately extracting relations between entities in complex documents. This can be attributed to the model's comprehensive understanding of the document structure and content, which directly results from incorporating three additional tasks during the pre-training phase. These tasks include reading order identification of different layout segments in a document image, layout segments categorization as per PubLayNet, and generation of the text sequence within a given layout segment (text block).

The proposed model outperforms other models in the same category, such as GeoLayoutLM [4] and LayoutLMv3_{LARGE} [3], in terms of F1 scores on the FUNSD and CORD datasets. This suggests that the proposed model is more effective in handling complex documents with intricate layouts and multiple entities.

On the SROIE dataset, the proposed model achieved a comparable F1 score to LayoutLMv2_{LARGE} [2], indicating its effectiveness in extracting key information from scanned receipts.

On the Kleister-NDA dataset, the proposed model achieved an F1 score of 0.8742, significantly higher than other models in the same category. This demonstrates the model's ability to accurately extract entities from formal English documents, such as Non-disclosure Agreements (NDAs).

TABLE III
ENTITY-LEVEL F1 SCORES OF ENTITY EXTRACTION TASKS ON THE FOUR DATASETS: FUNSD, CORD, SROIE AND KLEISTER-NDA. **NOTE:** "R/G/P" IMPLIES THE "REGION/GRID/PATCH" IMAGE EMBEDDING TYPES.

Modality	Model	Image Embedding	FUNSD	CORD	SROIE	Kleister-NDA	#Parameters
Text only	BERT _{BASE}	None	0.6026	0.8968	0.9099	0.7790	110M
	UniLMv2 _{BASE} [45]	None	0.6648	0.9092	0.9459	0.7950	125M
	BERT _{LARGE}	None	0.6563	0.9025	0.9200	0.7910	340M
	UniLMv2 _{LARGE} [45]	None	0.7072	0.9205	0.9488	0.8180	355M
Text + Layout	BROS _{BASE} [14]	None	0.8305	0.9573	0.9548	-	110M
	BROS _{LARGE} [14]	None	0.8452	0.9740	-	-	340M
Text + Layout + Image	DocFormer _{BASE} [17]	(G) ResNet-50	0.8334	0.9633	-	-	183M
	LayoutLMv1 _{BASE} [1]	(R) ResNet-101	0.7866	0.9472	0.9438	0.8270	160M
	LayoutLMv2 _{BASE} [2]	(G) ResNeXt101-FPN	0.8276	0.9495	0.9625	0.8330	200M
	LayoutLMv3 _{BASE} [3]	(P) Linear	0.9029	0.9656	-	-	133M
	DocFormer _{LARGE} [17]	(G) ResNet-50	0.8455	0.9699	-	-	536M
	LayoutLMv1 _{LARGE} [1]	(R) ResNet-101	0.7895	0.9493	0.9524	0.8340	343M
	LayoutLMv2 _{LARGE} [2]	(G) ResNeXt101-FPN	0.8420	0.9601	0.9781	0.8520	426M
	LayoutLMv3 _{LARGE} [3]	(P) Linear	0.9208	0.9746	-	-	368M
	GeoLayoutLM [4]	(R) ConvNeXt-FPN	0.9286	0.9797	-	-	399M
	Our Proposed Model BackBone network RoBERTa _{LARGE}	(P) Linear	0.9306	0.9804	0.9794	0.8742	425M

In conclusion, the proposed model demonstrates superior performance across various documents in entity relation extraction tasks. The high F1 scores achieved on all datasets validate the effectiveness of the additional tasks incorporated during the pre-training phase. The model's ability to understand and interpret complex document layouts and content makes it a promising tool for document analysis tasks.

C. Document Visual Question Answering Results

TABLE IV
ANLS SCORE ON THE DocVQA DATASET (PM: PARAMETERS)

Modality	Model	ANSL	#PM
Text only	BERT _{BASE}	0.6354	110M
	UniLMv2 _{BASE} [45]	0.7134	125M
	BERT _{LARGE}	0.6768	340M
	UniLMv2 _{LARGE} [45]	0.7709	355M
Text + Layout + Image	LayoutLMv1 _{BASE} [1]	0.6979	160M
	LayoutLMv2 _{BASE} [2]	0.7808	200M
	LayoutLMv3 _{BASE} [3]	0.7876	133M
	LayoutLMv1 _{LARGE} [1]	0.7259	343M
	LayoutLMv2 _{LARGE} [2]	0.8348	426M
	LayoutLMv3 _{LARGE} [3]	0.8337	368M
Text + Layout + Image	Our Proposed Model BackBone RoBERTa _{LARGE}	0.8468	440M

The Document Visual Question Answering (DocVQA) task demands a delicate blend of visual, layout, and textual comprehension. Our experimentation reveals significant insights into the models' performance on this multifaceted task. Purely text-based models, such as BERT and UniLMv2, tend to have restricted performance capabilities, as evident from their ANLS scores in Table IV. While BERT, even in its large configuration, couldn't cross the 0.7 mark, UniLMv2_{LARGE} managed to achieve a score of 0.7709. The pronounced improvement in ANLS scores when transitioning from text-only models to those incorporating layout and image information is striking. LayoutLM versions, especially, show significant

strides in their performance. By leveraging the additional modalities, these models are better positioned to comprehend the complexities of document images holistically.

Our proposed model, with additional encoder-decoder layers, achieves an impressive ANLS score of 0.8468. Not only does this score surpass the benchmarks set by the other models, but it also underscores the effectiveness of our approach. The collective pre-training scheme adopted, integrating diverse datasets and tasks, is pivotal in refining our model's understanding. By leveraging this diverse knowledge base, our model excels in extracting intricate details from documents, making it proficient at answering nuanced queries. The results highlight the fundamental importance of multi-modality integration in the DocVQA task. The fusion of textual, visual, and layout information, combined with innovative pre-training strategies, has the potential to push the boundaries of what models can achieve in document analysis.

D. Ablation Study: Analysis of Pre-training Tasks

In the light of recent breakthroughs in document analysis, particularly the cutting-edge models presented in [3], [4], a noticeable performance improvement is observed for the BERT_{LARGE} and RoBERTa_{LARGE} models compared to their base versions, BERT_{BASE} and RoBERTa_{BASE}. Our study focuses on the RoBERTa_{LARGE} model, intertwined with various pre-training tasks, aiming to unravel and clarify these tasks' individual and combined contributions.

Our analysis reveals several key insights:

Reading Order Identification of Layout Segments (ROILS): This task is a cornerstone for the model's performance across all tasks. Its significant impact suggests that understanding the spatial sequence of text in a document is crucial. It serves as a roadmap for models to decode the logical flow, inherently enhancing contextual comprehension.

Layout Segments Categorization (LSC): The inclusion of LSC led to a considerable improvement in Document

TABLE V

ABLATION STUDY OF THE PROPOSED MODEL WITH DIFFERENT PRE-TRAINING TASKS. THE PERFORMANCE OF DIFFERENT DATASETS WITH RESPECT TO THEIR CORRESPONDING TASK IS LISTED HERE. THE DOCUMENT CLASSIFICATION TASK IS EVALUATED AS ACCURACY (%), THE RELATION EXTRACTION TASK AS F1 SCORE, AND THE VISUAL QUESTION ANSWERING TASK AS THE ANSL SCORE.

Task	Document Classification RVL-CDIP	Relation Extraction FUNSD CORD SROIE Kleister-NDA				Visual Question Answering DocVQA
MLM + DC + LSC	95.49%	0.8343	0.9452	0.9529	0.8530	0.7959
MLM + DC + LSC + RE + VQA	95.38%	0.8843	0.9552	0.9629	0.8603	0.8334
MLM + DC + LSC + RE + VQA + ROILS + GTSLS	95.87%	0.9306	0.9804	0.9794	0.8742	0.8468

Classification and Entity Relation Extraction tasks. This highlights the importance of identifying different layouts within a document, enabling a clearer delineation of segments and their hierarchical relationships.

Generation of Text Sequence within Layout Segments (GTSLS): This task plays a crucial role in enhancing the Document Visual Question Answering task. By allowing the model to generate textual sequences within specific layouts, it equips the model with a more detailed understanding of the content, facilitating more accurate information retrieval. However, its impact on the Entity Relation Extraction task was less pronounced, indicating potential areas for further improvement.

In conclusion, while each pre-training task contributed to enhancing the model’s performance to varying extents, their collective inclusion resulted in the most significant improvements. This underscores that a comprehensive understanding of a document, with its complex interplay of textual, visual, and spatial information, requires a multifaceted approach.

REFERENCES

- [1] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “Layoutlm: Pre-training of text and layout for document image understanding,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1192–1200.
- [2] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che *et al.*, “Layoutlmv2: Multi-modal pre-training for visually-rich document understanding,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2579–2591.
- [3] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, “Layoutlmv3: Pre-training for document ai with unified text and image masking,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4083–4091.
- [4] C. Luo, C. Cheng, Q. Zheng, and C. Yao, “GeoLayoutLM: Geometric Pre-training for Visual Information Extraction,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 7092–7101. [Online]. Available: <https://ieeexplore.ieee.org/document/10204221/>
- [5] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval,” in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 991–995.
- [6] G. Jaume, H. K. Ekenel, and J.-P. Thiran, “Funsd: A dataset for form understanding in noisy scanned documents,” in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2. IEEE, 2019, pp. 1–6.
- [7] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, “Cord: a consolidated receipt dataset for post-ocr parsing,” in *Workshop on Document Intelligence at NeurIPS 2019*, 2019.
- [8] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, “Icdar2019 competition on scanned receipt ocr and information extraction,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1516–1520.
- [9] T. Stanisławek, F. Graliński, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski, and P. Biecek, “Kleister: Key information extraction datasets involving long documents with complex layouts,” Berlin, Heidelberg: Springer-Verlag, 2021, p. 564–579. [Online]. Available: https://doi.org/10.1007/978-3-030-86549-8_36
- [10] M. Mathew, D. Karatzas, and C. Jawahar, “Docvqa: A dataset for vqa on document images,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.
- [11] X. Zhong, J. Tang, and A. J. Yepes, “Publaynet: largest dataset ever for document layout analysis,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1015–1022.
- [12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” July 2019. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [13] C. Li, B. Bi, M. Yan, W. Wang, S. Huang, F. Huang, and L. Si, “Structallm: Structural pre-training for form understanding,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6309–6318.
- [14] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, and S. Park, “Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10767–10775.
- [15] C.-Y. Lee, C.-L. Li, T. Dozat, V. Perot, G. Su, N. Hua, J. Ainslie, R. Wang, Y. Fujii, and T. Pfister, “Formnet: Structural encoding beyond sequential modeling in form document information extraction,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3735–3754.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “Docformer: End-to-end transformer for document understanding,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 993–1003.
- [18] J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, N. Barmpalios, A. Nenkov, and T. Sun, “Unidoc: Unified pretraining framework for document understanding,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 39–50, 2021.
- [19] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, “Selfdoc: Self-supervised document representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5652–5660.
- [20] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Palka, “Going full-tilt boogie on document understanding with text-image-layout transformer,” in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*. Springer, 2021, pp. 732–747.
- [21] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *European conference on computer vision*. Springer, 2020, pp. 104–120.
- [22] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “Vi-bert: Pre-training of generic visual-linguistic representations,” in *International Conference on Learning Representations*, 2019.
- [23] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder

- representations from transformers,” 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)*. Association for Computational Linguistics, 2019.
- [24] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, “Seeing out of the box: End-to-end pre-training for vision-language representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12976–12985.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2020.
- [26] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.
- [27] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.
- [28] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” in *International Conference on Learning Representations*, 2021.
- [29] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, “Dit: Self-supervised pre-training for document image transformer,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3530–3539.
- [30] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, and S. Park, “BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents,” pp. 10767–10775, June 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21322>
- [31] Y. Li, Y. Qian, Y. Yu, X. Qin, C. Zhang, Y. Liu, K. Yao, J. Han, J. Liu, and E. Ding, “Structext: Structured text understanding with multi-modal transformers,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1912–1920.
- [32] C. Luo, Y. Wang, Q. Zheng, L. Li, F. Gao, and S. Zhang, “Merge and recognize: a geometry and 2d context aware graph model for named entity recognition from visual documents,” in *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, 2020, pp. 24–34.
- [33] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, “Pick: processing key information extraction from documents using improved graph learning-convolutional networks,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4363–4370.
- [34] C. Luo, G. Tang, Q. Zheng, C. Yao, L. Jin, C. Li, Y. Xue, and L. Si, “Bi-vldoc: Bidirectional vision-language modeling for visually-rich document understanding,” *arXiv preprint arXiv:2206.13155*, 2022.
- [35] J. Wang, L. Jin, and K. Ding, “Lilt: A simple yet effective language-independent layout transformer for structured document understanding,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7747–7757.
- [36] Y. Zhang, Z. Bo, R. Wang, J. Cao, C. Li, and Z. Bao, “Entity relation extraction as dependency parsing in visually rich documents,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 2759–2768.
- [37] X. Liu, F. Gao, Q. Zhang, and H. Zhao, “Graph convolution for multimodal information extraction from visually rich documents,” in *Proceedings of NAACL-HLT*, 2019, pp. 32–39.
- [38] W. Hwang, J. Yim, S. Park, S. Yang, and M. Seo, “Spatial dependency parsing for semi-structured document information extraction,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 330–343.
- [39] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [40] X. Han, Z. Huang, B. An, and J. Bai, “Adaptive transfer learning on graph neural networks,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 565–574.
- [41] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 328–339.
- [42] X. Hu, S. Yu, C. Xiong, Z. Liu, Z. Liu, and G. Yu, “P3 ranker: Mitigating the gaps between pre-training and ranking fine-tuning with prompt-based learning and pre-finetuning,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1956–1962.
- [43] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [44] Q. Peng, Y. Pan, W. Wang, B. Luo, Z. Zhang, Z. Huang, Y. Cao, W. Yin, Y. Chen, Y. Zhang *et al.*, “Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 3744–3756.
- [45] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou *et al.*, “Unilmv2: Pseudo-masked language models for unified language model pre-training,” in *International conference on machine learning*. PMLR, 2020, pp. 642–652.