DUBLIN - Document Understanding By Language-Image Network

Kriti Aggarwal[*], **Aditi Khandelwal**[*], **Kumar Tanmay**[*], **Owais Mohammed Khan**, **Qiang Liu**,
Monojit Choudhury, **Subhojit Som**, **Vishrav Chaudhary**, **Saurabh Tiwary**
Microsoft Turing
{kragga, t-aditikh, t-ktanmay, owais.mohammed, qiangliu}@microsoft.com,
{monojitc, subhojit.som, vchaudhary, satiwary}@microsoft.com

Abstract

Visual document understanding is a complex task that involves analyzing both the text and the visual elements in document images. Existing models often rely on manual feature engineering or domain-specific pipelines, which limit their generalization ability across different document types and languages. In this paper, we propose DUBLIN, which is pretrained on webpages using three novel objectives that leverage the spatial and semantic information in the document images: Masked Document Content Generation Task, Bounding Box Task, and Rendered Question Answering Task. We evaluate our model on several benchmarks, such as Web-Based Structural Reading Comprehension, Document Visual Question Answering, Key Information Extraction, Diagram Understanding, and Table Question Answering. We show that our model achieves competitive or better results than the state-of-the-art models on these tasks. In particular, we show that DUBLIN is the first pixel-based model to achieve an EM of 77.75 and F1 of 84.25 on the WebSRC dataset. We also show that our model outperforms the current pixel-based SOTA models on DocVQA and AI2D datasets by significant margins, 2% and 21% increase in performance, respectively. Also, DUBLIN is the first ever pixel-based model which achieves comparable to text-based SOTA methods on XFUND dataset for Semantic Entity Recognition showcasing its multilingual capability. Moreover, we create new baselines for text-based datasets by rendering them as document images and applying this model.

1 Introduction

In today's digital era, the availability of vast amounts of information in various document formats has grown exponentially (Dong et al., 2014). These documents encompass a wide range, including scientific research papers, official reports, online articles and webpages, and PowerPoint presentations, holding valuable knowledge and insights, both in textual and visual forms. A broad range of abilities is necessary for a thorough understanding of such a diverse range of documents, including the capacity to recognize text, comprehend language, and take into account various visual contexts (Lee et al., 2022).

Visual Document Understanding (VDU) models strive to enable machines to comprehend and interpret documents in a way that parallels human understanding. Humans naturally integrate various types of information, such as text, tables, charts, figures, and diagrams, to form a coherent understanding when reading a document or examining a webpage. They effortlessly adjust their comprehension strategies based on the specific content, regardless of its format or structure. Developing models that can autonomously discern and process various elements and structures within a document without relying on explicit human-defined rules opens up new opportunities for intelligent information extraction, knowledge synthesis, and data-driven decision-making. This advancement in VDU improves search and retrieval capabilities for document images, facilitating easier access to relevant information within documents (Li et al., 2022). VDU enhances accessibility for individuals with visual or reading impairments. It also facilitates task automation, including data entry, document verification, and document summarization, streamlining processes, and improving operational efficiency (Majumder et al., 2020; Tito et al., 2021). In addition, VDU can be applied to various industries, such as legal document analysis, historical document analysis, customer service automation, insurance claim processing, and more (Antonacopoulos et al., 2011; Lombardi and Marinai, 2020; Schweighofer and Merkl, 1999).

However, most current document understanding models often rely on manual feature engineering or domain-specific processing pipelines, which limits their capability to handle diverse document types

[*] equal contribution

effectively. Traditionally, visual document understanding has relied on optical character recognition (OCR) techniques to extract textual information from document images (Xu et al., 2020, 2022; Sage et al., 2020; Katti et al., 2018; Majumder et al., 2020; Hwang et al., 2021b,c). However, OCR techniques can be error-prone, particularly for handwritten or non-Latin scripts, and fail to adequately capture the visual information present in the document images, such as graphics, tables, and charts (Taghva et al., 2006; Hwang et al., 2021a; Rijhwani et al., 2020). In contrast, VDU can provide a more comprehensive understanding of the document's content by analyzing both the textual and visual information in the document images.

Our work is motivated by the desire to bridge this gap and empower models to analyze documents and webpages with the same level of adaptability as humans. This paper presents a transformer-based encoder-decoder model known as DUBLIN. It serves as a versatile multilingual document understanding model that can effectively perform various downstream tasks. The model is trained on web pages, which provide a diverse range of data for training. To enhance the model's capabilities, it undergoes pretraining on three novel objectives. First, there is the Bounding Box Task, where the model predicts the location of a bounding box based on the word/sentence within the document and vice versa. Second, the Rendered Question Answering Task focuses on answering questions about the document given a specific query. Lastly, the Masked Document Content Generation Task involves generating the textual content of an entire document using a masked image of the document as context. Alongside these objectives, the model also undergoes standard multimodal pretraining tasks such as masked autoencoding. Our contributions are fourfold:

1. We introduce three new pretraining tasks: Bounding Box Task, Rendered Question Answering Task, and Masked Document Content Generation Task for effective pretraining of multimodal document understanding models.

2. We demonstrate robust performance comparable to the state-of-the-art (SOTA) on a wide range of tasks, including Web-Based Structural Reading Comprehension, Document Visual Question Answering, Key Information Extraction, Diagram Understanding, and Table Question Answering.

3. We showcase strong multilingual performance on the XFUND dataset for the Semantic Entity Recognition task, covering a total of seven languages.

4. We create the baselines for text-based datasets by rendering them, setting a precedent as the first to undertake such an approach for Squad1.1 and WikiSQL datasets.

2 Related Works

The transformer architecture has been highly successful in the field of document understanding. The LayoutLM family of models (Xu et al., 2020), starting with LayoutLMv1, aimed to bring the success of transformer models like BERT in natural language processing to the visual domain of documents. LayoutLMv1 was pre-trained in a similar way to BERT (Devlin et al., 2019) but included 2D spatial position information. Later versions like BROS (Hong et al., 2022), TILT (Powalski et al., 2021), and LayoutLMv2 (Xu et al., 2022) improved the architecture by adding visual tokens and introducing spatially biased attention to learn the cross-modality interaction between the visual and textual information. LayoutLMv3 is pre-trained with a crossmodal alignment objective that makes it learn to match text and image modalities by identifying if the image patch that corresponds to a text word is missing. LayoutXLM introduces a multimodal pre-trained model for multilingual document understanding and the XFUND dataset, which is a form understanding benchmark dataset. (Xu et al., 2021).

TILT and DocFormer(Appalaraju et al., 2021) use only visual features extracted near the text tokens spatially, making them blind to areas of the form without text. In contrast, LayoutLMv2 extracts visual tokens across the entire document.

Models in the LayoutLM family have been evaluated without taking text recognition into account, despite the fact that text recognition is an essential task. Docreader is an end-to-end neural network model for document information extraction that can be trained without bounding-box annotations, using only the images and the target values of the fields to be extracted using CNN-RNN approach (Klaiman and Lehne, 2021). Ernie-Layout paper introduces a correlative pre-training task, reading order prediction, to learn the proper reading order of documents (Peng et al., 2022). Other end-to-end models like Donut (Kim et al., 2022) and Dessurt

(Davis et al., 2022) are image-to-text models that do not use OCR at inference time. Pix2struct is a pretrained image-to-text model for purely visual language understanding that can be fine-tuned on tasks containing visually-situated language (Lee et al., 2022). It has been pre-trained by learning to parse masked screenshots of web pages into simplified HTML and enables resolution flexibility to a variety of visual language domains. Matcha was proposed pretrainig objectives to enhance the mathematical reasoning and chart derendering capability of visual language models (Liu et al., 2022). Models like Dessurt and Donut use Swin transformer (Liu et al., 2021) as an image encoder and BART as Text Decoder architecture. PIXEL (Rust et al., 2023) proposed that language modeling can be used as a visual recognition task. They trained Masked Autoencoding Visual Transformer (ViT-MAE) to reconstruct the pixels in masked image patches on a rendered version of the English Wikipedia and the Bookcorpus dataset.

DUE Benchmark evaluates how well multimodal models can handle different document tasks such as Visual Question Answering, Key Information Extraction, and Machine Reading Comprehension. It covers various document domains and layouts that include tables, graphs, lists, and infographics (Łukasz Borchmann et al., 2021).

3 Approach/Method

In this section, we present an in-depth exploration of DUBLIN, describing its architecture, the diverse set of pre-training objectives it is trained upon, and the extensive array of pre-training datasets utilized to bolster its learning process.

3.1 Model Architecture

DUBLIN is a novel end-to-end framework that combines the Bletchley (Mohammed et al., 2023) image encoder and the text decoder of InfoXLM (Chi et al., 2021). Bletchley is a model that leverages a bootstrapping mechanism to train image and text encoders that can handle different modalities. InfoXLM is a model that learns a universal language representation that can handle different languages. We use Bletchley's image encoder and InfoXLM's text decoder as the initial weights for our model and then further pre-train them on various datasets using a combination of multi-task pre-training objectives and curriculum learning. The pretraining datasets include CCNews 200M, Google NQ Dataset, Bing QA, Rendered InfoXLM EN, Wikipedia and Synthetic TableQNA, elaborated further in Section 3.3. We also use cross-attention layers between the image encoder and the text decoder to model the interaction between the visual and textual modalities. This enables the decoder to attend to relevant regions in the image based on the query or context. Figure 1 shows the diagram of our pretraining framework. During pretraining, we resize the input image to a resolution of 224×224 and split it into fixed patches of 14×14, which are fed to the image encoder of our framework. We use curriculum learning and increase the resolution of the images to 896×896. We use the same image resolution (896×896) while finetuning our model.

3.2 Pretraining Objectives

We present a novel pretraining framework with four different levels of objectives for effectively capturing the complex structures of visual documents. These objectives include tasks at the language, image, document structure, and question answering levels. By focusing on multiple levels, our proposed framework is designed to enhance the model's ability to comprehend and reason about visual documents in a holistic manner. The language-level tasks involve language modeling and predicting masked words within the text of the document, whereas image-level tasks aim to recognize visual features and patterns within the images. Document structure tasks focus on modeling the hierarchical and relational structures present within documents while question-answering tasks aim to train the model to answer natural language questions based on the information in the document. Overall, our proposed pre-training objectives provide a comprehensive and effective approach for training models to understand and reason about visual documents. Now, we describe in detail the pre-training objectives used in our model.

Masked Document content Generation Task In this pre-training approach, we randomly mask out some parts of the input image before feeding it into the image encoder. The model is then trained to predict the entire document tokens given this masked image as a context using Masked Language modeling. In this way, DUBLIN learns the ability to understand the language given the visual information only, thereby bridging the gap between the language and visual modalities.

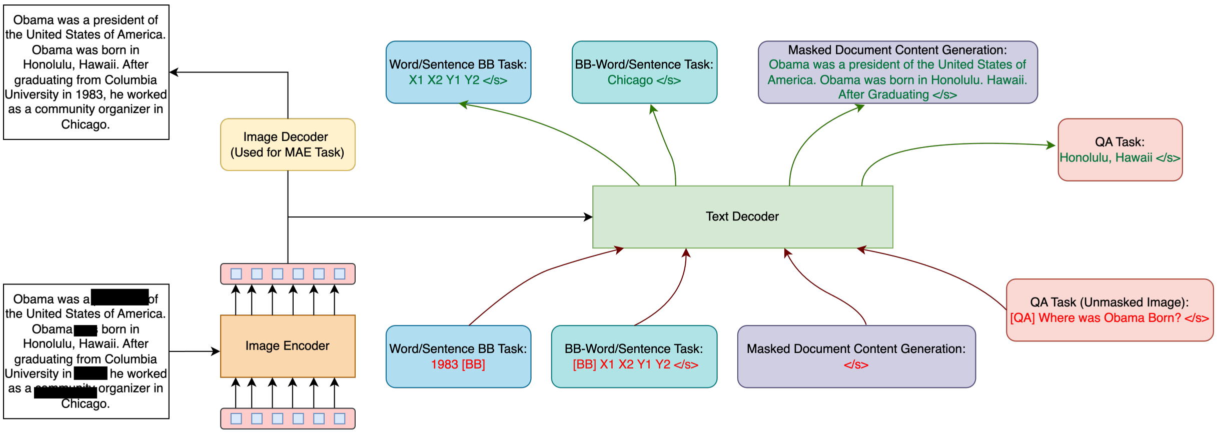
Obama was a president of the United States of America. Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.
Image Decoder (Used for MAE Task)
Image Encoder
Obama was a [redacted] of the United States of America. Obama [redacted] born in Honolulu, Hawaii. After graduating from Columbia University in [redacted] he worked as a [redacted] organizer in Chicago.
Word/Sentence BB Task: X1 X2 Y1 Y2 </s>
BB-Word/Sentence Task: Chicago </s>
Masked Document Content Generation: Obama was a president of the United States of America. Obama was born in Honolulu. Hawaii. After Graduating </s>
QA Task: Honolulu, Hawaii </s>
Text Decoder
Word/Sentence BB Task: 1983 [BB]
BB-Word/Sentence Task: [BB] X1 X2 Y1 Y2 </s>
Masked Document Content Generation: </s>
QA Task (Unmasked Image): [QA] Where was Obama Born? </s>

Figure 1: An illustration of the pretraining framework of DUBLIN. The input is a document image that is masked with random noise. The output consists of two decoders: a text decoder and an image decoder. The text decoder predicts the textual content of the document image, while the image decoder reconstructs the original document image. The text decoder is trained on three novel objectives: Bounding Box Task, Rendered Question Answering Task, and Masked Document Content Generation Task. These objectives enable the model to learn the spatial and semantic relationships between the textual and visual elements in the document image. The image decoder is trained on a standard objective: Masked Autoencoding Task. This objective enables the model to learn the visual features of the document image.

Masked Autoencoding Task Inspired by PIXEL, We used the MAE task, which is based on pixel regeneration. We masked out some patch tokens of the image randomly in a similar fashion as was suggested in BERT, and the model is trained to generate the unmasked image patch tokens. This way, the model learns how to recognize rich visual content from these document/webpage images in a self-supervised manner.

Bounding Box Task We use this objective to predict the coordinates of the bounding box that encloses a given the word/sentence in the document image and the word/sentence given the bounding bound coordinates enclosing it. This way, the model learns the layout structure of the document image in a supervised fashion.

Rendered Question Answering Task We introduce this task as an intermediate pretraining step for document image question answering. We employ a proprietary Bing QA dataset comprising 1 Billion question-answer pairs. We render the passage containing the answer as an image and also render the question on top of this image. This image is then input to the image encoder; the question is used as a prefix for the text decoder. The decoder then generates the answer to the question. This task helps the model to answer questions given the context image in a supervised setting.

3.3 Pretraining Data

CCNews 200M We use this dataset to obtain document images, texts, and bounding box coordinates in various web domains and languages. This is done by scraping the URLs from the CCNews 200M dataset (Crawl, 2016) using the method outlined in CCNet (Wenzek et al., 2020) followed by rendering the HTML pages as screenshots and storing the document texts and their corresponding bounding boxes with the help of the Selenium library. Prior to pre-training, we split the images into smaller chunks in two different resolutions: 224 × 224 and 896 × 896. The model is then trained with the lower resolution images in the initial steps, followed by pretraining on higher resolution images using multi-task pre-training objectives comprising of MAE, Masked Document Content Generation, Bounding Box Prediction for words/sentences, Word/Sentence generation given the bounding box coordinates.

Google NQ Dataset This is a publicly available dataset (Kwiatkowski et al., 2019) based on open domain question answering. It contains around 307k training samples. We scrape the webpage content using the HTML URLs. The webpage content is rendered as an image with the question added at the top. The question will also be used as a prefix for the decoder. We train our model on this dataset as a Rendered Question Answering task.

Bing QA Dataset We leveraged Bing to obtain question-answer pairs in English. We randomly sample question-answer pairs from Bing and render the passages and questions in a similar way as we did for the Google NQ dataset. In order to make our model generalization ability better over different kinds of texts, we render the text with random font size, color, and style obtained from the Google Fonts library. The model is then trained on Rendered Question Answering task.

Synthetic Table Structure QA: Continue Pretraining In order to teach the model how to understand the table structure, we curated this dataset by randomly selecting 1 million URLs that contain tables and using Selenium to extract those HTML table elements. To further enhance our training dataset, we perform data augmentation by employing five different CSS styles for rendering the HTML representation of each table as an image. These styles encompass various attributes such as border, font size, table separators, background, and text color. We devise this task of training the model to recognize table structures in document images. During the training process, for each table, we randomly select one out of the five available styles. This ensured a diverse range of table appearances for our model to learn from. To generate synthetic questions and answers, we developed fifteen distinct templates. These templates, reminiscent of SQL-like queries, were designed to reflect the content and format of the tables. An example template is as follows: "What is the value in the cell in the `[column_name]` column, where the row contains `[row_content]`?" Further elaboration on the templates and additional details can be found in the Appendix.

We pretrain our model for a total of 600k steps with different objectives and resolutions. We start with 50k steps of combined MAE and document content generation tasks on low-resolution images (896×224) on the CCNews 200M, GoogleNQ, and Rendered InfoXLM EN datasets. Then in a curriculum learning fashion, we add the Rendered Question Answering objective using the Google NQ and Bing QA datasets for 350k steps. Next, we increase the resolution to (896×896) for 55k steps with the same objectives. Finally, we add the bounding box prediction objectives and train for another 150k steps on high-resolution images (896×896). Next, we delve into the process of fine-tuning for various downstream tasks, shedding light on the precise steps taken to adapt DUBLIN for specific tasks. In addition, we meticulously describe the datasets used for fine-tuning, enabling DUBLIN to excel in a variety of real-world scenarios.

4 Finetuning

To capture the model's ability to generalize over diverse varieties of documents, we perform extensive experimentation on different kinds of documents like handwritten, typewritten, born-digital, scanned documents, Infographics, diagrams, tables, forms, and webpages. We also explore how our model is performing on multilingual documents. Table 1 summarizes the datasets used in this paper, along with their brief descriptions.

We use some datasets curated by the DUE-Benchmark (Łukasz Borchmann et al., 2021) for evaluating the performance of our model on visual document understanding. Datasets made specifically for Visual Question Answering (VQA), like WebSRC and DocVQA, are simple to use because the input image and output text can be used directly. We directly render the question as a header at the top of the original image, following the methodology used in Pix2Struct (Lee et al., 2022). Additionally, the question is fed to the text decoder as a prefix. Our model uses the visual modality to process the query and the image simultaneously. To evaluate DUBLIN's performance on various tasks, we use the datasets that are described in Table 1.

4.1 Data

DocVQA DocVQA dataset (Mathew et al., 2021b) focuses on question-answering tasks using single-page excerpts from real-world industry documents that include printed, handwritten and digital documents. The questions in this dataset often require understanding and processing various elements such as images, free text, tables, lists, forms, or a combination of these components.

InfographicsVQA The InfographicVQA dataset (Mathew et al., 2021a) contains questions that are specifically targeted at Infographics that can be found online. The inclusion of large images with extreme aspect ratios is one distinguishing feature of this dataset. Answering questions about visualized data found in a variety of Infographics is part of the task. The information needed to answer these questions can be presented using a variety of elements, including text, plots, graphs, or infographic layout components.

Dataset	Dataset Type	Description
DocVQA	Single page Documents	VQA over scanned documents.
InfographicsVQA	Long Documents With Text, Graphics, Charts etc.	VQA over high-res Infographics.
XFUND	Multilingual Form Documents	Key-Information Extraction
WebSRC	Documents	VQA over webpage documents
DeepFORM	Documents on Election Spending	Key Information Extraction
Rendered Squad	Documents	Visual QA over passages
WikiTable	HTML tables From Wikipedia	VQA over tables
TabFact	Tables from Wikipedia	Visual fact verification over tables
WikiSQL	Rendered Tables	VQA over tables
AI2D	Illustrations	VQA over science diagrams

Table 1: A summary of the datasets used for experiments in this paper. The datasets cover a variety of document types and languages and are used to evaluate the performance of DUBLIN on different downstream tasks.

WebSRC WebSRC, also known as Web-based Structural Reading Comprehension, is a dataset consisting of 440,000 question-answer pairs (Chen et al., 2021). These pairs were collected from a diverse collection of 6,500 web pages. Each entry in the dataset includes not only the questions and answers but also the HTML source code, screenshots, and metadata associated with the respective web page. Answering questions in the WebSRC dataset requires a certain level of understanding of the structure of the web page. The answers can take the form of specific text excerpts, Key Information Extraction (KIE), or table question answering. To assess the performance on this dataset, we use metrics such as Exact Match (EM) and F1 score (F1). The training and development datasets are obtained using the official split provided by the authors. However, it's important to note that the authors have not released the testing set, so the results are solely based on the development set.

DeepForm We make use of the Key Information Extraction (KIE) dataset DeepForm (Svetlichnaya, 2020), which includes important election finance-related documents. The goal of this dataset is to extract crucial data from advertising disclosure forms submitted to the Federal Communications Commission (FCC), such as contract numbers, advertiser names, payment amounts, and air dates. Instead of the query, we provide the "Key" to the text decoder for the model to extract information from the image.

SQuAD1.1 To evaluate our model's extractive question-answering performance, we fine-tune it on the SQuAD dataset (Rajpurkar et al., 2016). We render this dataset as images on the fly, choosing a random font text, font style, etc., for each data point to maintain diversity and to test that, at inference time, the model is not biased toward answering questions from documents that all look a certain way but rather diverse in their fonts, styles, etc. The SQuAD dataset consists of over 100,000 question-answer pairs for over 500 articles. Given a question and its corresponding context paragraph, the task is to extract the span of text that contains the answer to the question. We follow the standard evaluation metrics for this dataset, including Exact Match (EM) and F1 score (F1), which measure the model's ability to output an answer that exactly matches the ground truth and its overlap with the ground truth, respectively. By evaluating this widely used benchmark, we can compare the performance of our model against the state-of-the-art approaches in extractive question answering.

WikiTable WikiTableQuestions dataset (Pasupat and Liang, 2015) utilized in this study focuses on question answering using semi-structured HTML tables obtained from Wikipedia. The authors specifically aimed to provide challenging questions that require multi-step reasoning on a series of entries within the given table, involving operations such as comparison and arithmetic calculations. We use the table images provided by the DUE Benchmark.

TabFact TabFact dataset includes entailed and refuted statements corresponding to a single row or cell to investigate fact verification using semi-structured evidence from clean and straightforward tables sourced from Wikipedia (Chen et al., 2020). Despite the task's binary classification nature, it presents challenges that go beyond simple categorization. The task requires sophisticated linguistic and symbolic reasoning to achieve high accuracy. We pass the table image to the image encoder and expect a binary output from the text decoder for this table fact verification task.

WikiSQL WikiSQL is a large crowd-sourced dataset consisting of 80,654 meticulously annotated examples of questions and corresponding SQL queries (Zhong et al., 2017). These exam-

	Metrics	Sota with Special Pipelines	Pixel Only		Ours
			Pix2Struct	Donut	
DocVQA	ANLS	88.41 (Ernie-Layout)	76.6	67.5	**78.2**
InfoVQA	ANLS	46.1 (T52DU)	**40.0**	11.6	36.82
WebSRC	EM/F1	81.66/86.24 (TIE)	-	-	**77.75/84.25**
DeepForm	F1	83.3 (T52DU)	-	-	62.23
AI2D	EM	42.6 (Matcha)	42.1	30.8	**51.11**

Table 2: The performance of DUBLIN on different visual document understanding datasets. The table lists the metrics, External SOTA on these datasets with specialized pipelines for the tasks associated with these datasets, performance of pixel-only models-Pix2Struct (Large) and Donut, and the score of our model. The scores are compared with the state-of-the-art pixel-based models only, and the best scores are highlighted in bold.

Datasets	Metrics	Text based Baselines	DUBLIN
Rendered Squad	EM/F1	90.6/95.7 (ANNA)	79.69/87.07
WikiTable	EM	62.5 (Dater)	25.69
TabFact	F1	93.0 (Dater)	71.98
WikiSQL	EM	89.2 (TAPEX)	75.26

Table 3: Our model's performance on different datasets rendered and tasks.

ples are derived from 24,241 tables extracted from Wikipedia. This dataset mainly focuses on translating text to SQL. However, given our model's focus on answering questions based on documents, we transformed the denotations of this dataset into question-answer pairs in a natural language format. We rendered the tables as images by converting the table's JSON to HTML and then obtaining their screenshots in a similar fashion as described for the synthetic table structure QA dataset.

AI2D AI2 Diagrams (AI2D) is a comprehensive dataset consisting of over 5000 science diagrams typically found in grade school textbooks, along with more than 150,000 annotations, including ground truth syntactic parses and over 15,000 corresponding multiple choice questions (Kembhavi et al., 2016). The diagrams cover a wide range of scientific topics, such as geological processes, biological structures, and more. The multiple-choice questions are based on the science diagrams and are designed to test students' comprehension of the content. The dataset provides only train and test splits, with 1 percent of the train split set aside for validation.

XFUND XFUND is a comprehensive multilingual benchmark dataset that focuses on form understanding. It includes human-labeled forms with key-value pairs in seven different languages, namely Chinese, Japanese, Spanish, French, Italian, German, and Portuguese. With this dataset, researchers and practitioners can evaluate and improve the performance of form understanding models across different languages, making it a valuable resource for advancing the field (Xu et al., 2021).

5 Analysis and Discussion

The comparison of DUBLIN and other pixel-based models on various downstream tasks is presented in Table 2. The current state-of-the-art (SOTA) approach for DocVQA, known as Ernie Layout (Peng et al., 2022), achieves an ANLS of 88.41 by using both the training and development sets for fine-tuning. Other models like LayoutLMv3 achieved ANLS of 83.4 (Huang et al., 2022) by incorporating additional in-domain pretraining data from the IIT-CDIP scanned documents corpus. In contrast, our model relies exclusively on visual input and achieves a state-of-the-art ANLS of 78.2 among pixel-based models. One of the limitations or future directions of our current model is handling the extreme aspect ratio or multi-page documents in some of the datasets, such as InfographicsVQA and DeepForm. These datasets contain images that have an extreme aspect ratio. When these images are resized to 896 × 896, the model has difficulty reading the text within the images due to the loss of resolution and detail. Our model is the first pixel-based model to tackle question answering on the WebSRC dataset, which encompasses three types of question-answering tasks: Extractive QA, K-V Pair QA, and Table Question Answering. We have achieved impressive performance on this dataset, comparable to the current state-of-the-art (SOTA) model called TIE (Zhao et al., 2022). TIE, the existing SOTA model, employs a specialized pipeline specifically designed for the WebSRC dataset. It combines the Graph Attention Network (GAT) and a Pre-trained Language Model (PLM) to leverage the topological information of logical structures and the spatial structures in the dataset. Our results demonstrate a strong baseline and establish a

	FUNSD(en)	ZH	JA	ES	FR	IT	DE	PT	AVG.
LayoutXLM	0.8068	**0.9155**	**0.8216**	0.8055	0.8384	0.8372	0.8530	**0.8650**	0.8429
LILT	0.8574	0.9047	0.8088	0.8340	0.8577	**0.8792**	**0.8769**	0.8493	**0.8585**
DUBLIN	**0.8754**	0.8645	0.7819	**0.8413**	**0.8727**	0.8703	0.8760	0.8570	0.8550

Table 4: Results on Multilingual Dataset - XFUND

new state-of-the-art performance on the WebSRC benchmark for pixel-based models. In the case of WikiTable, our model's performance is not optimal. This is primarily due to the dataset containing questions that necessitate a deep comprehension of tables and the complex operations that can be performed on them. These operations include aggregation and multistep reasoning, which are beyond the scope of our current work. However, when it comes to Tabfact, we achieve a decent level of performance. This is noteworthy considering that the state-of-the-art model, Dater (Ye et al., 2023), is a text-based model that makes use of large language models (LLMs) as decomposers for effective table-based reasoning. Our model achieves 51.1(EM) on the AI2D dataset, surpassing previous models such as DQANet (Kembhavi et al., 2016) with 38.5(EM) and Pix2Struct with 42.1(EM). This demonstrates our model's ability to comprehend both the text and the diagram in the document. Regarding Rendered Squad and WikiSQL, we have made significant contributions. We are the pioneers in rendering these datasets as images and creating strong baselines specifically tailored for pixel-based models. This innovative approach allows us to establish new benchmarks and achieve impressive results. Table 3 summarizes the results on the rendered/text-based datasets. To test our model's generalizability in a multilingual setting, we finetuned it on a combined XFUND and FUNSD dataset for the Semantic Entity Recognition task and evaluated it on each test set for different languages. Our model outperforms existing SOTA models (LILT and LayoutXLM) in FUNSD (English), Spanish, and French languages and achieves comparable performances in other languages as well.

6 Conclusion

We have presented DUBLIN, a novel framework for visual document understanding. DUBLIN is a transformer-based encoder-decoder model that can analyze both the text and the visual elements in document images. DUBLIN is pretrained on webpages using three novel objectives that capture the spatial and semantic relationships between the document elements. We have evaluated DUBLIN on several downstream tasks and shown that it achieves competitive or superior results compared to the state-of-the-art models. We have also demonstrated its multilingual capabilities and created new baselines for text-based datasets by rendering them as document images. Our work shows that DUBLIN is a versatile and robust model for multimodal document understanding that does not rely on external OCR systems and can be finetuned easily in an end-to-end fashion. Our work opens up new possibilities for document analysis and understanding in various domains and applications.

References

A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. 2011. Historical document layout analysis competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1516–1520.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification.

Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training.

Common Crawl. 2016. News dataset available.

Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end document recognition and understanding with dessurt.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 601–610, New York, NY, USA. Association for Computing Machinery.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking.

Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. 2021a. Cost-effective end-to-end information extraction for semi-structured document images.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021b. Spatial dependency parsing for semi-structured document information extraction.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021c. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. Association for Computational Linguistics.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer.

Shachar Klaiman and Marius Lehne. 2021. Docreader: Bounding-box free training of a document information extraction model.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2struct: Screenshot parsing as pretraining for visual language understanding.

Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. 2022. Markuplm: Pre-training of text and markup language for visually-rich document understanding.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2022. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Francesco Lombardi and Simone Marinai. 2020. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10).

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504, Online. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images.

Owais Khan Mohammed, Kriti Aggarwal, Qiang Liu, Saksham Singhal, Johan Bjorck, and Subhojit Som. 2023. Bootstrapping a high quality multilingual multimodal dataset for bletchley. In *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pages 738–753. PMLR.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels.

Clément Sage, Alex Aussem, Véronique Eglin, Haytham Elghazel, and Jérémy Espinas. 2020. End-to-end extraction of structured information from business documents with pointer-generator networks. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 43–52, Online. Association for Computational Linguistics.

Erich Schweighofer and Dieter Merkl. 1999. A learning technique for legal document analysis. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, ICAIL '99, page 156–163, New York, NY, USA. Association for Computing Machinery.

Stacey Svetlichnaya. 2020. DeepForm: Understand Structured Documents at Scale — wandb.ai. `https://wandb.ai/stacey/\deepform_v1/reports/DeepForm-Understand-Structured-Documents/-at-Scale--VmlldzoyODQ3Njg`. [Accessed 15-May-2023].

Kazem Taghva, Russell Beckley, and Jeffrey Coombs. 2006. The effects of ocr error on the extraction of private information. In *Document Analysis Systems VII*, pages 348–357, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rubèn Tito, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2021. Icdar 2021 competition on document visualquestion answering.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. 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*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning.

Zihan Zhao, Lu Chen, Ruisheng Cao, Hongshen Xu, Xingyu Chen, and Kai Yu. 2022. Tie: Topological information enhanced structural reading comprehension on web pages.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning.

Łukasz Borchmann, Michal Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Gralinski. 2021. Due: End-to-end document understanding benchmark. In *NeurIPS Datasets and Benchmarks*.