DocFormerv2: Local Features for Document Understanding

Srikar Appalaraju[1] *, Peng Tang[1], Qi Dong[1], Nishant Sankaran[1], Yichu Zhou[2] †, R. Manmatha[1]

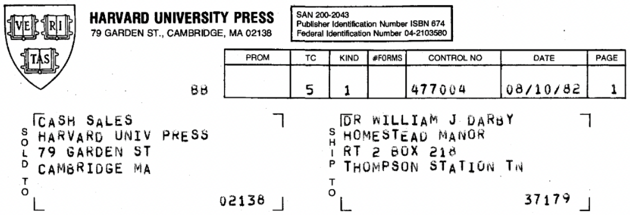
[1]AWS AI Labs, [2]School of Computing at University of Utah
{srikara, tangpen, qdon, nishsank, manmatha}@amazon.com, flyaway@cs.utah.edu

Abstract

We propose DocFormerv2, a multi-modal transformer for Visual Document Understanding (VDU). The VDU domain entails understanding documents (beyond mere OCR predictions) e.g., extracting information from a form, VQA for documents and other tasks. VDU is challenging as it needs a model to make sense of multiple modalities (visual, language and spatial) to make a prediction. Our approach, termed DocFormerv2 is an encoder-decoder transformer which takes as input - vision, language and spatial features. DocFormerv2 is pre-trained with unsupervised tasks employed asymmetrically i.e., two novel document tasks on encoder and one on the auto-regressive decoder. The unsupervised tasks have been carefully designed to ensure that the pre-training encourages local-feature alignment between multiple modalities. DocFormerv2 when evaluated on *nine* challenging datasets shows state-of-the-art performance over strong baselines e.g. TabFact (4.3%), InfoVQA (1.4%), FUNSD (1%). Furthermore, to show generalization capabilities, on three VQA tasks involving scene-text, DocFormerv2 outperforms previous comparably-sized models and even does better than much larger models (such as GIT2, PaLi and Flamingo) on these tasks. Extensive ablations show that due to its novel pre-training tasks, DocFormerv2 understands multiple modalities better than prior-art in VDU.

Introduction

Documents have become ubiquitous carriers of information, including forms, tables, invoices, and other structured documents. Many such documents require visual and layout understanding to make sense (just the text string is insufficient). Visual Document Understanding (VDU) is the task of leveraging machine learning techniques to comprehend such scanned documents, such as PDFs or images. Popular VDU tasks include Document and Tables VQA (Mathew et al. 2020; Chen et al. 2019), sequence labeling for key-value identification in forms (Jaume, Ekenel, and Thiran 2019), entity extraction (Seunghyun et al. 2019), and document classification (Harley, Ufkes, and Derpanis). While modern deep-learning based OCR models (Litman et al. 2020) have proven to be effective in extracting text from documents, the naive approach of linearizing the OCR-text and feeding it to a language model is sub-optimal. This is because the content of a document is presented according to a visual layout and structure that must be taken into account for accurate understanding. Naively linearizing the text from left-to-right will result in sub-optimal performance as the semantic meaning alters based on layout, as shown in Figure 1 - Table 5,4 has experiments demonstrating this. Instead, VDU requires a multi-modal approach that can comprehend text and visual features in the context of a document's 2D layout.



HARVARD UNIVERSITY PRESS
79 GARDEN ST., CAMBRIDGE, MA 02138

SAN 200-2043
Publisher Identification Number ISBN 674
Federal Identification Number 04-2103580

	PROM	TC	KIND	#FORMS	CONTROL NO	DATE	PAGE
88		5	1		477004	08/10/82	1

SOLD TO
CASH SALES
HARVARD UNIV PRESS
79 GARDEN ST
CAMBRIDGE MA
02138

SHIP TO
DR WILLIAM J DARBY
HOMESTEAD MANOR
RT 2 BOX 218
THOMPSON STATION TN
37179

Figure 1: **Visual Document Understanding**. Snippet of a document receipt from DocVQA (Mathew, Karatzas, and Jawahar 2021). VDU tasks could include a model asked to predict "SOLD TO" address (VQA) or predict all relations ("SOLD TO" → <address>, "SHIP TO" → <address>) or asked to infer info from table (at the top).

Multi-modal training in general entails feature alignment. Specific to vision-language learning this means aligning a piece of text with an arbitrary span of pixels in visual space (Ho et al. 2022; Kim, Son, and Kim 2021; Radford et al. 2021; Wang et al. 2022b; Alayrac et al. 2022; Biten et al. 2022; Appalaraju et al. 2021; Hao et al. 2023; Appalaraju et al. 2020; Li et al. 2022b; Chen et al. 2022b). How those features are aligned makes a big difference. In VDU, a majority of the tasks require *local and layout-relative* understanding of the document. For example, in document VQA, semantic labeling or entity extraction, a model needs to make sense of text in-relation to where the text is placed in a document. E.g.: "1" when placed at the top-right/bottom-left of a document is to be interpreted as a page-number vs as a number when placed anywhere else.

Based on this domain understanding of VDU and its challenges, we present DocFormerv2 (DFv2) which is an encoder-decoder multi-modal transformer. In this work, we meticulously devise two novel unsupervised pre-training

*Corresponding author.

†Work conducted during an internship at Amazon.

Model	Year	Conf.	Arch.	Input Mod.
LayoutLMv1 (Xu et al. 2020a)	2020	KDD	E	T + S
DocFormerv1 (Appalaraju et al. 2021)	2021	ICCV	E	T + V + S
LayoutLMv2 (Xu et al. 2020b)	2021	ACL	E	T + V + S
SelfDoc (Li et al. 2021c)	2021	CVPR	E	-
LayoutLMv3 (Huang et al. 2022)	2022	ACM	E	T + V + S
BROS (Hong et al. 2020a)	2022	AAAI	E	T + S
XYLayoutLM (Gu et al. 2022b)	2022	CVPR	E	T + V + S
FormNet (Lee et al. 2022a)	2022	ACL	E	-
ERNIE-Layout (Peng et al. 2022)	2022	EMNLP	E	T + V + S
LiLT (Wang, Jin, and Ding 2022)	2022	ACL	E	T + S
XDoc (Chen et al. 2022a)	2022	EMNLP	E	T
TILT (Powalski et al. 2021)	2021	ICDAR	E + D	T + V + S
DocFormerv2 (ours)	2023	-	E + D	T + V + S

Table 1: **VDU Related Work**: In this table, a summary of VDU prior art is presented with their architecture (E: Encoder, D: Decoder), the input (T: text, V: vision, S: spatial features), the vision features branch and core idea behind the work.

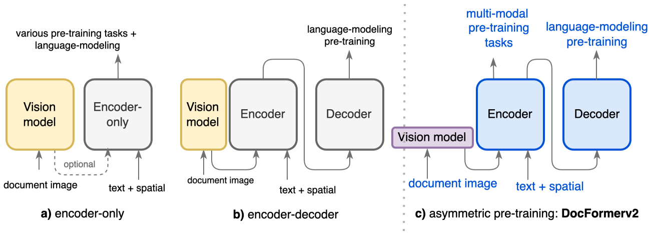


Figure 2: **VDU Paradigms:** Broad state of Visual Document Understanding (VDU) approaches. In **a)** E-only LayoutLM (Xu et al. 2020a) and variants. **b)** E+D but only language-task TILT (Powalski et al. 2021). **c)** Ours

tasks with the objective of incorporating local semantic information of a document into the model. These pre-training tasks impart the ability to the model to accurately locate relevant information within the document. We also depart from VDU prior-art(Powalski et al. 2021; Tang et al. 2022) as we introduce a novel asymmetrical method of pre-training. i.e., multi-task pre-training on encoder (two tasks) and decoder (one task). We propose two novel pre-training tasks on encoder with the intent to enrich the encoder with local semantic information. The tasks aid in by fusing and aligning multi-modal input and generating efficient representations for the decoder. We show that these pre-training tasks are necessary for effective VDU (see §). Furthermore, we demonstrate that a simplified linear visual layer is sufficient to encapsulate visual features, simplifying the architecture from previous VDU research (Xu et al. 2020b; Li et al. 2021c; Powalski et al. 2021) which required specific visual encoders (Dosovitskiy et al. 2020; Liu et al. 2021; He et al. 2016).

Experimentally we demonstrate that DocFormerv2 achieves state-of-the-art performance on five VDU tasks. In addition, we demonstrate the versatility of DocFormerv2 by utilizing its pre-trained model and fine-tuning it on text-VQA tasks from a completely different domain. Our approach yields superior performance on three distinct text-VQA datasets, surpassing comparable models and in some datasets much bigger models like GIT2 (Wang et al. 2022b), PaLi (Chen et al. 2022b) and Flamingo (Alayrac et al. 2022). Therefore, the primary contributions of this paper are as follows:

- Asymmetrical method of pre-training for VDU: Two novel tasks on the encoder which encourage local multi-modal feature collaboration (*Token-to-Line* task and *Token-to-Grid* task) and one on the decoder §.
- Simplified Visual branch: DocFormerv2 is end-to-end trainable and it does not rely on a pre-trained object detection network for visual features simplifying its architecture. On five varied downstream VDU tasks, DocFormerv2 achieves state of the art results §.
- We also show DocFormerv2 versatility by fine-tuning it on a totally different domain - text-VQA datasets without changing the pre-training. DocFormerv2 beats strong baselines and achieves state-of-the-art numbers on three text-VQA datasets amongst similar model sizes. Selectively, on Text-VQA it out-performs much larger models like PaLi-3B +6.8%, PaLi-15B +1.5% and Flamingo(Alayrac et al. 2022) (+9.9%) (106x DocFormerv2 size in the num. of parameters) by absolute accuracy §.

Furthermore, we conducted comprehensive ablation experiments to demonstrate the advantages of our pre-training tasks, the model's resilience to input noise, and the efficacy of the simplified visual branch.

Related Work

VDU research has attracted considerable attention over the past few years (Wang et al. 2022c; Xu et al. 2020a; Fujinuma et al. 2023; Xu et al. 2020b; Appalaraju et al. 2021; Li et al. 2021c; Powalski et al. 2021; Li et al. 2021b; Huang et al. 2022; Appalaraju et al. 2023; Hong et al. 2020b; Gu et al. 2022a; Tang et al. 2023b; Gu et al. 2022b; Lee et al. 2022a; Wang, Jin, and Ding 2022; Chen et al. 2022a; Tang et al. 2022; Łukasz Borchmann et al. 2021; Peng et al. 2022; Li et al. 2021a; Tang et al. 2023a). Prominent published research papers in this area are catalogued in Table 1 - the research focus has been lopsided towards encoder-only models. While TILT (Powalski et al. 2021) proposed a encoder-decoder transformer for VDU, they only train it on one pre-training task (masked language modeling) and also use a bulky visual CNN. Our approach DocFormerv2 , simplifies the architecture by not using a separate visual module (CNN or Transformer based) and has multiple unsupervised pre-training tasks. See supplemental for more on prior art.

Approach

Architecture

DocFormerv2 (DFv2) is a multi-modal encoder-decoder transformer architecture (see fig. 3). Three variations of DFv2 are designed - small, base and large variants (see supplemental material for details). DFv2 takes multi-modal inputs, the image of the document I, text T extracted by an OCR model along with OCR bounding box co-ordinates as spatial features $\overline{S}$. DFv2 has a unified multi-modal encoder

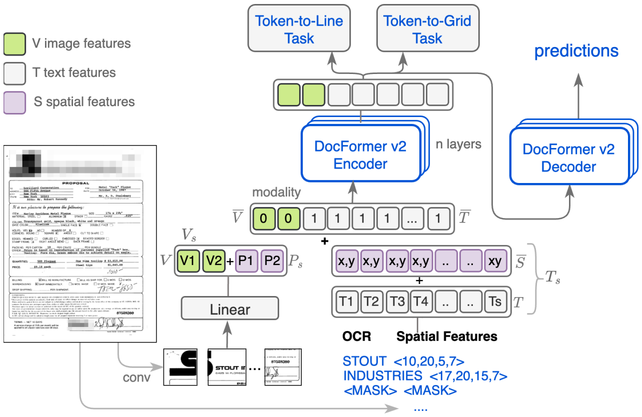


Figure 3: **DocFormerv2 Pre-train Architecture**. After pre-train, the two prediction heads (token-to-line and grid) on encoder are removed, rest of the architecture remains the same for down-stream tasks. Read section for more details on T_s and V_s. All components are end-to-end trainable. Best viewed in color.

where the multi-modal features fuse and align with the help of novel pre-training tasks (see §).

Visual features: DFv2 has a simplified visual branch contrary to most VDU prior-art (fig. 2). DFv2 consumes a flattened image sequence as visual input. Specifically, let $v \in \mathbb{R}^{3 \times h \times w}$ be the image of a document. A simple $V = linear(conv_{2\times2}(v))$ is used to create an image embedding. The weights are randomly initialized for pre-training. As documents tend to have lots of white-space, the linear down-sampling layer gives an opportunity for the model to only keep relevant visual features. Based on our ablation experiments (see supplemental material), this simple approach gives better results than using expensive image encoders such as Swin, ViT (Liu et al. 2021; Dosovitskiy et al. 2020; Ronneberger, Fischer, and Brox 2015) or bulky object-detection networks like FRCNN variants (Ren et al. 2015) as was used in VDU prior-art (Powalski et al. 2021; Appalaraju et al. 2021; Xu et al. 2021). Since transformer layers are permutation-invariant, a learnable 2D-positional encoding P_s is also computed. Finally, $V_s = V + P_s$

Language features: Let t be the predicted text extracted via an OCR model for a document image. DFv2 uses a sentence-piece sub-word tokenizer (Kudo and Richardson 2018) to get tokens t_{tok}. A maximum sequence limit s is applied during training and testing, so if the number of OCR tokens is greater than s, the rest is ignored. If the sequence length is less than s, the sequence is padded. The OCR tokens t_{tok} are sent to a learnable embedding layer W_t to create a text embedding $T = W_t(t_{tok})$.

Spatial features: For each OCR word t_i, the OCR model predicts its bounding-box location in the normalized form $b_i = (x_1, y_1, x_3, y_3)$. This information is encoded using four learnable spatial embedding layers - W_x for encoding a word horizontal spatial information x_i, W_y for the vertical coordinate y_i, W_h for word height h_i and W_w for the width w_i. The spatial features not only encode the location of a word in the document but also provides cues about a word's font-size and thereby its importance in a document (via h_i and w_i). Specifically, spatial features $\overline{S} = W_x(x_1, x_3) + W_y(y_1, y_3) + W_h(y_3 - y_1) + W_w(x_3 - x_1)$. Finally, $T_s = T + \overline{S}$.

Other features: T_s and V_s features are different modalities (fig. 3). As the model has no idea it is being fed multi-modal input, another learnable embedding W_m is used to provide cues to the model about the multi-modal input. A modality-embedding W_m learns nuances of different modalities, which generates M_v embedding for visual modality and M_t for text. Finally, $\overline{T} = T_s + M_t$ and $\overline{V} = V_s + M_v$. $\overline{T}$ and $\overline{V}$ are concatenated in the sequence dimension to form the input sequence to the DFv2 encoder.

Unsupervised Document Pre-training

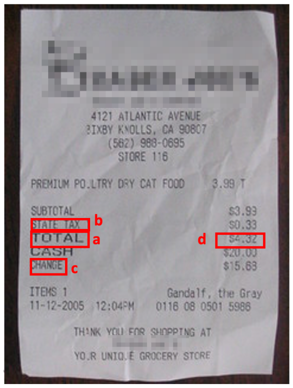
In DocFormerv2 we follow the now well established practice of unsupervised pre-training followed by downstream task fine-tuning. Furthermore, with the intent of eliciting the maximum benefit from unsupervised pre-training, we designed the pre-training tasks as a close proxy for downstream tasks. We now describe the two novel pre-training tasks employed on the encoder and the language modeling task on decoder. All three tasks are performed at the same time and the final loss is a linear combination of all three losses for each iteration.

Encoder Token-to-Line Task: We share the intuition that for VDU tasks local feature semantic alignment is important. Most of the related information for key-value prediction in a form or VQA is either on the same line or adjacent lines of a document e.g., see fig. 4, in order to predict the value for "`TOTAL`" (box a), the model has to look in the same line (to its right - "`$4.32`" box d). We teach the model the relative position information between tokens. For implementation, we randomly pick two language tokens and ask the model to predict the number of lines between them. Furthermore, as a document could have an arbitrary number of lines of text, the task is quantized. i.e., there are only three labels: `{0, 1, 2}`. All token pairs that are more than 2 lines apart are labelled as 2 because distant tokens are not likely related and the model should learn to ignore them. Assume that a, b, c, d (fig. 4) are lines. Let F be the DFv2 encoder head function trying to predict a label for this task. then:

$$F(a, d) = 0; \; F(a, b) = 1; \; F(b, c) = 2 \qquad (1)$$

Based on the ablation (table 9), this task gives +2.2% benefit on DocVQA task. The loss for this task is tracked as L_{tol}.

Encoder Token-to-Grid Task: Different semantic information is concentrated in different regions of the document. For example, a) In a financial document, the top block contains the header, the middle block contains information to be filled and the bottom block typically contains footer elements/instructions. b) Page numbers are typically at the top or the bottom. c) In a receipt/invoice the company name is typically at the top. The content of a document is presented according to a visual layout and structure that must be taken into account for accurate understanding. Based on this intuition, this task pairs language semantics with the location (visual,



[redacted]

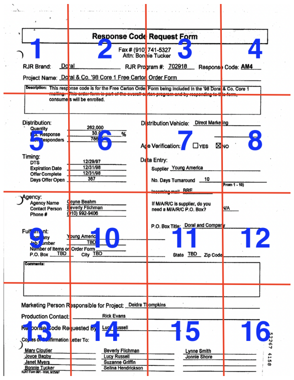
4121 ATLANTIC AVENUE
BIXBY KNOLLS, CA 90807
(562) 988-0695
STORE 116

PREMIUM PO.LTRY DRY CAT FOOD	3.99 T
SUBTOTAL	$3.93
STATE TAX	$0.33
TOTAL	$4.32
CASH	$20.00
CHANGE	$15.68

ITEMS 1 Gandalf, the Gray
11-12-2005 12:04PM 0116 08 0501 5986

THANK YOU FOR SHOPPING AT
[redacted]
YO.R UNIQ.E GROCERY STORE

Figure 4: **Token-to-Line**



Response Code Request Form

Fax # (910) 741-5327
Attn: Bonnie Tucker

RJR Brand: Doral RJR Program #: 702918 Response Code: AM4

Project Name: Doral & Co. '98 Core 1 Free Carton Order Form

Description: This response code is for the Free Carton Order Form being included in the '98 Doral & Co. Core 1 mailing. This order form is part of the overall carton program and by responding to this form, consumers will be enrolled.

Distribution:
Quantity 262,000
Est. Response 30.0 %
Est. Responders 786

Distribution Vehicle: Direct Marketing

Age Verification: □ YES ☒ NO

Timing:
DTS 12/29/97
Expiration Date 12/31/98
Offer Complete 12/31/98
Days Offer Open 367

Data Entry:
Supplier Young America
No. Days Turnaround 10 (From 1 - 10)
Incoming mail BRE

Agency:
Agency Name Coyne Beahm
Contact Person Beverly Flichman
Phone # (910) 992-9406

If M/A/R/C is supplier, do you need a M/A/R/C P.O. Box? N/A

P.O. Box Title: Doral and Company

Fulfillment:
Company Young America
Job Number TBD
Number of Items on Order Form
P.O. Box TBD City TBD State TBD Zip Code

Comments:

Marketing Person Responsible for Project: Deidra Thompkins
Production Contact: Rick Evans
Response Code Requested By: Lucy Russell
Copies of Confirmation Letter To:

Mary Cloutier	Beverly Flichman	Lynne Smith
Joyce Bagby	Lucy Russell	Jonnie Shore
Janet Myers	Suzanne Griffin	
Bonnie Tucker	Selina Hendrickson	

RJRT Form 991 - WWL 9/25/97

52267 4158

Figure 5: **Token-to-Grid** 4x4

spatial or both) in a document. Specifically, the document is virtually divided into a m x n grid. Each OCR token is assigned a grid number and DFv2 is tasked with predicting the grid number for each token. For each OCR token t_i, its top-left location (x_1, y_1) is used to determine its grid-number g_i. Grids are in raster-scan order, so if a particular token falls on the boundary of multiple grids, the scan-order is used to disambiguate. If a token falls on the boundary of normalized image co-ordinates, they are ignored for prediction. See fig. 5 for viz. Specifically, we have:

$$g_i = \lfloor \frac{x_1}{\Delta_x} \rfloor + \lfloor \frac{y_1}{\Delta_y} \rfloor \cdot m,$$

where Δ_x and Δ_y are the widths and heights of each grid, respectively, and m is the number of grids in a row. The loss L_{tog}.

Decoder Language Modeling: Since VDU predictions are in the language domain, language understanding forms an important component of DFv2 pre-training. We do the denoising masked language modeling popularized by T5 (Raffel et al. 2019). During pre-training, not only are the input tokens randomly MASKED it's spatial features (mentioned in §) are also masked. Masking the spatial features $\overline{S}$ for the masked tokens makes the grid prediction and line prediction hard because the model does not have 2D-position information of the masked tokens. It has to infer from other available context. The loss for this operation is denoted L_{dlm}.

Final pre-training loss: The final loss is a linear combination of all three pre-training losses i.e., $L_{final} = k * L_{tol} + l * L_{tog} + m * L_{dlm}$, where k, l, m are empirically determined.

Downstream Tasks: Once pre-training is done, we remove the token-to-line and token-to-grid linear prediction heads. The rest of the pre-trained model is fine-tuned on the respective downstream train data.

Experiments

Implementation details: Following prior-art (Appalaraju et al. 2021; Powalski et al. 2021; Biten et al. 2022; Xu et al. 2020a, 2021; Huang et al. 2022) we use the Industrial Document Library (IDL)[1] dataset for pre-training. The IDL is a collection of industry documents hosted by UCSF. It hosts millions of documents publicly disclosed from various industries like tobacco, drug, food etc. The data from the website amounts to about 13M documents, translating to about 70M pages of various document images. We further extracted OCR for each document. Data was cleaned and about 6M documents were pruned, the resulting 64M document images and OCR-text (with spatial co-ordinates) is used for unsupervised pre-training. The data distribution for IDL 64M is presented in supplemental section.

Downstream experiments: The model is fine-tuned on the provided training set and numbers are reported on the corresponding validation/test set. No dataset specific hyper-parameter tuning is done. This is an advantage of our approach and we believe that the numbers may be higher if dataset specific fine-tuning is done. Details about fine-tuning datasets are in the supplemental section. We used Pytorch (Paszke et al. 2019) and the Huggingface library (Thomas et al. 2019).

Evaluation Metrics: A dataset specific evaluation metric is adopted. For DocVQA(Mathew et al. 2020), InfoVQA(Mathew et al. 2022), ST-VQA(Biten et al. 2019b), Average Normalized Levenshtein Similarity (ANLS) (Biten et al. 2019a) is used. ANLS measures the similarity between the predicted results and ground truth and ranges from (0,100). For FUNSD(Jaume, Ekenel, and Thiran 2019), CORD(Seunghyun et al. 2019) F1-score is used. For TextVQA (Singh et al. 2019) and OCR-VQA(Mishra et al. 2019) accuracy is used. In all metrics, higher the better.

Table VQA

WikiTable and **TabFact** (Chen et al. 2019; Łukasz Borchmann et al. 2021): These datasets study table understanding and fact verification with semi-structured evidence over tables collected from Wikipedia. Entailed and refuted statements corresponding to a single row or cell were prepared by the authors of TabFact. This task poses challenges due to the complex linguistic and spatial reasoning involved. In Table 2, we can see that DocFormerv2 out-performs prior art by a large margins (+1.1%) and (+4.3%) resp.

Model	**WikiTable** Acc. (%)	**TabFact** Acc. (%)
methods based on only text / (text + spatial) features:		
T5$_{\text{large}}$(Raffel et al. 2019)	33.3	58.9
T5$_{\text{large}}$+U (Łukasz Borchmann et al. 2021)	38.1	76.0
T5$_{\text{large}}$+2D (Łukasz Borchmann et al. 2021)	30.8	58.0
T5$_{\text{large}}$+2D+U (Łukasz Borchmann et al. 2021)	43.3	78.6
methods based on image + text + spatial features:		
LayoutLMv3$_{\text{large}}$ (Huang et al. 2022)	45.7	78.1
UDOP (Tang et al. 2022)	<u>47.2</u>	<u>78.9</u>
DocFormerv2$_{\text{large}}$	**48.3**(+1.1%)	**83.2** (+4.3%)

Table 2: **Comparison on Table VQA Datasets**: Our work, DocFormerv2 outperforms the previous state of the art on WikiTableQuestions and TabFact table VQA datasets. **bold** is SOTA and <u>underline</u> indicates the previous SOTA. See supplemental for viz.

[1] https://www.industrydocuments.ucsf.edu/

Document VQA

DocVQA (Mathew et al. 2020) and InfographicsVQA (Mathew et al. 2022) are datasets for the document VQA task. DocVQA (Mathew et al. 2020) focuses on VQA for real-world industry documents and requires that the model understand images, texts, tables, forms, . InfographicsVQA (Mathew et al. 2022) focuses on VQA for infographics and requires that the model understand plots/graphs, texts, layout, figures. A model needs to reason multi-modally to generate an answer for this data. Please see the supplemental for data statistics and samples.

Model	**DocVQA** test ANLS (%)	**InfoVQA** test ANLS (%)
methods based on only image:		
Donut$_{\text{base}}$ (Kim et al. 2021)	67.5	11.5
Pix2Struct$_{\text{large}}$ (Lee et al. 2022b)	76.6	40.0
methods based on only text / (text + spatial) features:		
T5$_{\text{large}}$(Raffel et al. 2019)	70.4	36.7
T5$_{\text{large}}$+U (Łukasz Borchmann et al. 2021)	76.3	37.1
T5$_{\text{large}}$+2D (Łukasz Borchmann et al. 2021)	69.8	39.2
T5$_{\text{large}}$+2D+U (Łukasz Borchmann et al. 2021)	81.0	46.1
methods based on image + text + spatial features:		
LayoutLMv3$_{\text{large}}$ (Huang et al. 2022)	83.4	45.1
UDOP (Tang et al. 2022)	84.7	<u>47.4</u>
LayoutLMv2$^{\dagger}_{\text{large}}$ (Xu et al. 2020b)	86.7	-
TILT$^{\dagger}_{\text{large}}$ (Powalski et al. 2021)	<u>87.05</u>	-
DocFormerv2$_{\text{large}}$	87.2	-
DocFormerv2$^{\dagger}_{\text{large}}$	**87.84** (+0.79%)	**48.8** (+1.4%)

Table 3: **Comparison on Document VQA Datasets**: Our work, DocFormerv2 outperforms the previous state of the art. † indicates training with extra document VQA data.

Following common practice (Łukasz Borchmann et al. 2021; Powalski et al. 2021; Xu et al. 2020b), we train DocFormerv2 on the combination of the training and validation sets and do evaluation on the test set for each dataset. In addition, we also follow (Powalski et al. 2021; Xu et al. 2020b) to train DocFormerv2 on an extra document VQA dataset with 850k question-answer pairs and then fine-tune on DocVQA/InfographicsVQA for higher accuracy.

DocFormerv2 outperforms (Table 3) the previous state of the art for document VQA even without using any extra document VQA pre-training data. After pre-training on the extra data, DocFormerv2 surpasses the previous state of the art by 0.79% on DocVQA and 1.4% on InfographicsVQA, which confirms the effectiveness of our approach.

Sequence Labeling Task

We study the performance of DocFormerv2 on the semantic entity-labeling task (i.e., group tokens which belong to the same class). We test the model on FUNSD dataset (Jaume, Ekenel, and Thiran 2019), which is a forms dataset containing 199 noisy documents (149 images for train, 50 images for test). There are four classes: *question, answer, header*, and *other*. We measure entity-level performance using F1 score (Table 4). The input sequence to Docformerv2 includes individual texts as prompts and all document texts as context, and the decoder sequence contains the entity texts and predicted labels. Docformerv2 achieves 88.89% F1 score (Table 4), and outperforms the existing methods without using entity box priors in pretraining and finetuning (grayed models in the table).

Model	Precision	Recall	F1
methods based on only image:			
Dessurt$_{\text{base}}$ (Davis et al. 2022)	-	-	65.0
methods based on only text / (text + spatial) features:			
BERT$_{\text{base}}$ (Devlin et al. 2018)	54.69	61.71	60.26
RoBERTa$_{\text{base}}$ (Liu et al. 2019)	63.49	69.75	66.48
UniLMv2$_{\text{base}}$ (Bao et al. 2020)	63.49	69.75	66.48
LayoutLMv1$_{\text{base}}$ (Xu et al. 2020a)	76.12	81.55	78.66
BROS$_{\text{base}}$ (Hong et al. 2020a)	80.56	81.88	81.21
BERT$_{\text{large}}$ (Devlin et al. 2018)	61.13	70.85	65.63
RoBERTa$_{\text{large}}$ (Liu et al. 2019)	67.80	73.91	70.72
UniLMv2$_{\text{large}}$ (Bao et al. 2020)	67.80	73.91	70.72
LayoutLMv1$_{\text{large}}$ (Xu et al. 2020a)	75.36	80.61	77.89
StructuralLM$_{\text{large}}$ (Li et al. 2021a)	83.52	86.81	85.14
FormNet (Lee et al. 2022a)	85.21	84.18	84.69
methods based on image + text + spatial features:			
LayoutLMv1$_{\text{base}}$ (Xu et al. 2020a)	76.77	81.95	79.27
LayoutLMv2$_{\text{base}}$ (Xu et al. 2020b)	80.29	85.39	82.76
LayoutLMv2$_{\text{large}}$ (Xu et al. 2020b)	83.24	85.19	84.20
DocFormer$_{\text{base}}$ (Appalaraju et al. 2021)	80.76	86.09	83.34
DocFormer$_{\text{large}}$ (Appalaraju et al. 2021)	82.29	86.94	84.55
SelfDoc (Li et al. 2021c)	-	-	83.36
UDoc (Gu et al. 2022a)	-	-	<u>87.93</u>
StrucTexT (Li et al. 2021d)✢	85.68	80.97	83.09
LayoutLMv3$_{\text{base}}$ (Huang et al. 2022)❄	77.39	81.65	79.46
LayoutLMv3$_{\text{large}}$ (Huang et al. 2022)❄	81.35	83.75	82.53
LayoutLMv3$_{\text{base}}$ (Huang et al. 2022)❍	89.55	91.65	90.29
LayoutLMv3$_{\text{large}}$ (Huang et al. 2022)❍	92.19	92.10	92.08
UDOP (Tang et al. 2022)❍	-	-	91.62
DocFormerv2$_{\text{base}}$	89.15	87.6	88.37
DocFormerv2$_{\text{large}}$	89.88	87.92	**88.89** (+1.0%)

Table 4: **FUNSD comparison**: DocFormerv2 does better than models its size and compares well with even larger models. ✢ does not use standard train/test split, and the results are not directly compared with others. ❍ use OCR lines (not word box) as 2D position for words, and use entity boxes as 2D position for each word during finetuning and test, and thus the results are not directly comparable. ❄ are results by using the word boxes as 2D position for each word as other competitors do.

Entity Extraction Task

We evaluate DocFormerv2 for the entity extraction task on the CORD dataset. CORD (Seunghyun et al. 2019) consists of 1000 receipts (800/100/100 images for train/val/test). It defines 30 fine-grained fields under 4 coarse-grained categories. To extract all entities, in the input sequence, we add a question of "*What are entities of <CLASS>?*" in front of all text context tokens. The output of the decoder includes all entities which are separated by a separator token. Following the standard evaluation metric for entity extraction, we measure entity-level performance using F1 score. Docformerv2 (Table 5) achieves 97.7% F1 score, and outperforms existing methods. Docformerv2 enables multiple entities decoding in an auto-regressive way which shows that the model is able to learn both intra-entity and inter-entity structures. Note that it is unfair to directly compare Docformerv2 with LayoutLMv3(LaMv3), because LaMv3 uses segment-level layout positions, while the other works use word-level lay-

Model	Precision	Recall	F1
methods based on only text / (text + spatial) features:			
BERT$_{base}$ (Devlin et al. 2018)	88.33	91.07	89.68
UniLMv2$_{base}$ (Bao et al. 2020)	89.87	91.98	90.92
SPADE (Hwang et al. 2020)	-	-	91.50
LayoutLMv1$_{base}$ (Xu et al. 2020a)	94.37	95.08	94.72
BROS$_{base}$ (Hong et al. 2020a)	95.58	95.14	95.36
BERT$_{large}$ (Devlin et al. 2018)	88.86	91.68	90.25
RoBERTa$_{large}$ (Liu et al. 2019)	-	-	93.80
UniLMv2$_{large}$ (Bao et al. 2020)	91.23	92.89	92.05
LayoutLMv1$_{large}$ (Xu et al. 2020a)	94.32	95.54	94.93
FormNet (Lee et al. 2022a)	98.02	96.55	97.28
methods based on image + text + spatial features:			
LayoutLMv2$_{base}$ (Xu et al. 2020b)	94.53	95.39	94.95
LayoutLMv2$_{large}$ (Xu et al. 2020b)	95.65	96.37	96.01
TILT$_{base}$ (Powalski et al. 2021)❍	-	-	95.11
TILT$_{large}$ (Powalski et al. 2021)❍	-	-	96.33
DocFormer$_{base}$(Appalaraju et al. 2021)	96.52	96.14	96.33
DocFormer$_{large}$ (Appalaraju et al. 2021)	97.25	96.74	96.99
UDoc (Gu et al. 2022a)	-	-	96.86
LayoutLMv3$_{base}$ (Huang et al. 2022)❄	92.92	94.31	93.61
LayoutLMv3$_{large}$ (Huang et al. 2022)❄	96.78	96.78	96.78
LayoutLMv3$_{base}$ (Huang et al. 2022)❍	-	-	96.56
LayoutLMv3$_{large}$ (Huang et al. 2022)❍	-	-	97.46
UDOP (Tang et al. 2022)❍	-	-	97.58
DocFormerv2$_{base}$	97.51	96.10	96.80
DocFormerv2$_{large}$	97.71	97.70	**97.70**

Table 5: **CORD dataset comparison**. We present entity-level Precision, Recall, F1 on test set. ❍ use OCR lines (not word box) as 2D position for words, and use entity boxes as 2D position for each word during fine-tuning and testing, and thus the results are not directly comparable. ❄ are results by using the word boxes as 2D position for each word as the other competitors do.

out positions [2]. More importantly, the task studied in Table 5 is entity extraction: predicting words and classes of all entities, against this problem setting if one uses segment-level boxes as inputs.

Generalization Experiments - Scene-Text VQA

In this section, we show the strength of DocFormerv2 on a different task - Text-VQA. Unlike document understanding which focuses on document images, the Text-VQA task answers questions for natural images with scene text. We fine-tune our *document* pre-trained models on three Text-VQA datasets. We emphasize that no image-text pre-training was performed on DocFormerv2, it was merely fine-tuned on the respective Text-VQA training dataset. Three popular Text-VQA datasets are used - OCR-VQA (Mishra et al. 2019), TextVQA (Singh et al. 2019) and ST-VQA (Biten et al. 2019b), each with strong baselines from the vision-language community (as is standard practice by Text-VQA we mean any scene text VQA dataset while TextVQA refers to a specific dataset). Please see the supplemental for a dataset breakdown. For OCR-VQA, we fine-tune our models on the training set and do evaluation on the validation and test sets.

Model	Val Acc. (%)	Test Acc. (%)
Blk+CNN+W2V	-	48.3
M4C (Hu et al. 2020)	63.5	63.9
LaAP (Han, Huang, and Han 2020)	63.8	64.1
LaTr$_{base}$ (Biten et al. 2022)	67.5	67.9
GIT$_{base}$	57.3	57.5
GIT$_{large}$	62.4	62.9
GIT	67.8	68.1
GIT2✢ (Wang et al. 2022b)	-	70.3
DocFormerv2$_{base}$	69.7	70.3
DocFormerv2$_{large}$	71.1	**71.5** (+3.4%)

Table 6: **Comparison on OCR-VQA**: DocFormerv2 is better than the previous SOTA by (+3.4%). **Bold** indicates best and underline indicates the previous state of the art. GIT2 ✢: uses extra VQA data (aggregation of 8 VQA datasets).

For TextVQA and ST-VQA, following the previous state-of-the-art methods (Biten et al. 2022; Yang et al. 2021), we fine-tune our models on the combination of the TextVQA and ST-VQA training sets and do evaluation on the validation and test sets of each dataset. Tables 6, 7, 8 show that our large size model outperforms the comparably sized previous state-of-the-art method LaTr (Biten et al. 2022) by +3.4%, +2.4% and +2.2% on the OCR-VQA, TextVQA, and

Model	Val Acc. (%)	Test Acc. (%)
M4C (Hu et al. 2020)	47.8	-
LaAP (Han, Huang, and Han 2020)	41.0	41.4
SA-M4C (Kant et al. 2020)	45.4	44.6
SMA	44.6	45.5
SceneGate (Luo et al. 2022)	42.4	44.0
SC-Net (Fang et al. 2022)	44.8	45.7
LOGOS (Lu et al. 2021)	51.5	51.1
TAP + TAG (Wang et al. 2022a)	53.6	53.7
TAP (Yang et al. 2021)	54.7	54.0
TAP Two-Stage (Li et al. 2022a)	55.9	55.3
Flamingo-80B★ (Alayrac et al. 2022)	57.1	54.1
PreSTU (Kil et al. 2022)	-	56.3
LaTr-0.3B$_{base}$ (Biten et al. 2022)	58.0	58.9
LaTr-0.3B$^{\dagger}_{base}$(Biten et al. 2022)	59.5	59.6
LaTr-0.85B$^{\dagger}_{large}$(Biten et al. 2022)	61.0	61.6
GIT-0.13B$_{base}$❍	18.8	-
GIT-0.4B$_{large}$❍	37.5	-
GIT-0.7B❍	59.9	59.8
GIT2-5.1B✢ (Wang et al. 2022b)	68.4	67.3
PaLi-3B❍ (Chen et al. 2022b)	58.8	-
PaLi-15B❍ (Chen et al. 2022b)	64.1	-
PaLi-17B❍ (Chen et al. 2022b)	70.5	73.1
DocFormerv2-0.2B$^{\dagger}_{base}$	61.6	60.0
DocFormerv2-0.75B$^{\dagger}_{large}$	**65.6**	**64.0** (+2.4%)

Table 7: **Comparison on TextVQA**: † indicates the model used the combination of ST-VQA and TextVQA training sets to train the model. GIT2 ✢: extra data used (aggregation of 8 VQA datasets) ★: video-text data. ❍: proprietary image-text data. Grey rows shows models which are much bigger (# parameters $\geq$ 3x DFv2 $_{large}$ parameters) and use large amounts of external data. DFv2 $_{large}$ still outperforms Flamingo (+9.9%), Pali-3B (+6.8%) and Pali-15B (+1.5%) models.

[2] LaMv3 (Huang et al. 2022) highlighted that using segment-level positions may benefit the semantic entity labeling task, so the two types of work are not directly comparable.

Model	Val ANLS (%)	Test ANLS (%)
M4C (Hu et al. 2020)	47.2	46.2
LaAP (Han, Huang, and Han 2020)	49.7	48.5
SA-M4C (Kant et al. 2020)	51.2	50.4
SceneGate (Luo et al. 2022)	52.5	51.6
LOGOS (Lu et al. 2021)	58.1	57.9
TAP (Yang et al. 2021)	59.8	59.7
TAP + TAG (Wang et al. 2022a)	62.0	60.2
PreSTU (Kil et al. 2022)	-	65.5
LaTr-0.3B$_{base}$ (Biten et al. 2022)	67.5	66.8
LaTr-0.3B$^{\dagger}_{base}$(Biten et al. 2022)	68.3	68.4
LaTr-0.85B$^{\dagger}_{large}$(Biten et al. 2022)	70.2	69.6
GIT-0.13B$_{base}$	20.7	-
GIT-0.4B$_{large}$	44.6	-
GIT-0.7B	69.1	69.6
DocFormerv2-0.2B$^{\dagger}_{base}$	70.1	68.4
DocFormerv2-0.75B$^{\dagger}_{large}$	**72.9**	**71.8** (+2.2%)

Table 8: **Comparison on ST-VQA**: On ST-VQA DocFormerv2 outperforms comparable sized models like GIT and LaTr but large margin (+2.2%) in-spite of being pre-trained on less data. † indicates the combination of the ST-VQA and TextVQA training sets is used.

ST-VQA test sets respectively. These results show that our method generalizes beyond document understanding tasks.

Analysis: Surprisingly, on OCR-VQA, DocFormerv2$_{large}$ even performs better than GIT2 (Wang et al. 2022b) which is a 5.1B size model (750M for DocFormerv2$_{large}$) and uses 12.9B data for pre-training (64M for DocFormerv2$_{large}$). On TextVQA, DocFormerv2 does better than several vision-language models which are much bigger and have been pre-trained on much more data. On the test set, it is (+9.9%) better than Flamingo (which at 80B has 106x the number of parameters as ours). On the validation set, it is better than PaLi-3B and 15B (+2.2%, +6.8%) respectively. GIT2 and PaLi-17B do perform better than it. (GIT2 also uses 8 VQA datasets to train). DocFormerv2 gets this performance without any natural image-text pre-training. We present this as evidence that DocFormerv2 is a good approach to solving this problem with a much smaller model and much less data.

Ablation Experiments

Ablation of DFv2 novel pre-training tasks: Table 9 shows DFv2 ablation on the proposed novel pre-training tasks and multi-modal training. The denoising language modeling task and spatial features mentioned in § are applied to all architectures. Note, this ablation was performed on DFv2 -small with 1M doc pre-training.

Pre-training Impact or Better Approach? DFv2 was pre-trained with 64M documents whereas prior-art like LayoutLMv2 (Xu et al. 2020b) was pre-trained with only 11M documents. In order to see if DFv2 benefits come from more pre-training data or a better approach, we ablate. Table 10 shows that DFv2 $_{base}$ is superior to LayoutLMv2$_{base}$ when pre-trained on the same quantity of data. We also see DFv2 improve in performance as more pre-training data is provided (64M). The table shows that the novel DFv2 asymmetric pre-training approach is a superior VDU approach.

Model Ablation	Datasets DocVQA (ANLS)
baseline B	69
B + V	70.5 (+1.5)
B + V + L	71.2 (+2.2)
B + V + G	71.7 (+2.7)
B + V + L + G	73.0 (+4.0)

Table 9: **DocFormerv2 Pre-training Tasks Ablation**: Impact of three pre-training tasks on four downstream tasks over baseline. **B:** baseline, **V:** only with Visual features §, **L:** with Token-to-Line prediction pre-training §, **G:** with Token-to-Grid prediction pre-training §.

Model	# pre-train data	Datasets	
		FUNSD	CORD
LayoutLMv2$_{base}$	11M	82.7	94.9
DocFormerv2$_{base}$	11M	86.1 (+3.4%)	96.2 (+1.3%)
DocFormerv2$_{base}$	64M	87.9 (+5.2%)	96.8 (+1.9%)
DocFormerv2$^{\dagger}_{base}$	64M	88.3 (+5.6%)	96.8 (+1.9%)

Table 10: **DocFormerv2 Pre-training Data Ablation**: Impact of training with different # of pre-training data on various down-stream tasks. The F1 scores are reported. † indicates the combination of the ST-VQA and TextVQA training sets is used.

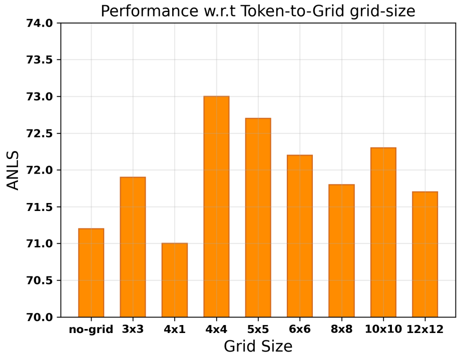


Figure 6: **Token-to-Grid Ablation**. How different grid sizes used for the Token-to-Grid pre-training task affects model performance on DocVQA. 4x4 seems best and was used for all final pre-training.

Correct grid size for Token-to-Grid pre-training? In §, we presented the novel Token-to-Grid pre-training task. In this pre-training ablation §9 this task was observed to provide benefits. Here the appropriate virtual grid-size is empirically determined. From Fig. 6, 4x4 grid seems optimal. Smaller or asymmetric grid structures (4x1) seem to cause harm. On the other end, if the grid is too granular (12x12, 8x8), the performance seems to hurt as well. All models pre-trained on DFv2 $_{small}$ and 1M documents from IDL, with the Vision and Token-to-line enabled.

More ablations Please find more ablation experiments in supplemental [3] highlighting more experiments of our ap-

[3] https://arxiv.org/abs/2306.01733

proach.

Conclusion

Our work DocFormerv2 highlights the importance of two novel pre-training tasks and the efficacy of enriching encoder representations with local semantic information via pre-training tasks. We perform experiments on eight varied datasets (five on VDU and three on scene-text VQA) achieving state-of-the-art numbers on all datasets. Based on ablations, we also show the various design choices and its impact on downstream performance

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv*, abs/2204.14198.

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 993–1003.

Appalaraju, S.; Tang, P.; Dong, Q.; Sankaran, N.; Zhou, Y.; and Manmatha, R. 2023. DocFormerv2: Local Features for Document Understanding - Full Paper and Supplemental. *arXiv preprint arXiv:2306.01733*.

Appalaraju, S.; Zhu, Y.; Xie, Y.; and Fehérvári, I. 2020. Towards Good Practices in Self-supervised Representation Learning. *Neural Information Processing Systems (NeurIPS Self-Supervision Workshop 2020)*.

Bao, H.; Dong, L.; Wei, F.; Wang, W.; Yang, N.; Liu, X.; Wang, Y.; Piao, S.; Gao, J.; Zhou, M.; and Hon, H.-W. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. arXiv:2002.12804.

Biten, A. F.; Litman, R.; Xie, Y.; Appalaraju, S.; and Manmatha, R. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16548–16558.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Mathew, M.; Jawahar, C.; Valveny, E.; and Karatzas, D. 2019a. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1563–1570.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; and Karatzas, D. 2019b. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4291–4301.

Chen, J.; Lv, T.; Cui, L.; Zhang, C.; and Wei, F. 2022a. XDoc: Unified Pre-training for Cross-Format Document Understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; LI, S.; Zhou, X.; and Wang, W. Y. 2019. TabFact: A Large-scale Dataset for Table-based Fact Verification. *ArXiv*, abs/1909.02164.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A. J.; Padlewski, P.; Salz, D. M.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A. V.; Bradbury, J.; Kuo, W.; Seyedhosseini, M.; Jia, C.; Ayan, B. K.; Riquelme, C.; Steiner, A.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2022b. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *ArXiv*, abs/2209.06794.

Davis, B. L.; Morse, B.; Price, B.; Tensmeyer, C.; Wigington, C.; and Morariu, V. I. 2022. End-to-end Document Recognition and Understanding with Dessurt. In *ECCV Workshops*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fang, C.; Zeng, G.; Zhou, Y.; Wu, D.; Ma, C.; Hu, D.; and Wang, W. 2022. Towards Escaping from Language Bias and OCR Error: Semantics-Centered Text Visual Question Answering. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06.

Fujinuma, Y.; Varia, S.; Sankaran, N.; Appalaraju, S.; Min, B.; and Vyas, Y. 2023. A Multi-Modal Multilingual Benchmark for Document Image Classification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14361–14376. Singapore: Association for Computational Linguistics.

Gu, J.; Kuen, J.; Morariu, V. I.; Zhao, H.; Barmpalios, N.; Jain, R.; Nenkova, A.; and Sun, T. 2022a. Unified Pre-training Framework for Document Understanding. *ArXiv*, abs/2204.10939.

Gu, Z.; Meng, C.; Wang, K.; Lan, J.; Wang, W.; Gu, M.; and Zhang, L. 2022b. XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4573–4582.

Han, W.; Huang, H.; and Han, T. 2020. Finding the Evidence: Localization-aware Answer Prediction for Text Visual Question Answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3118–3131.

Hao, X.; Zhu, Y.; Appalaraju, S.; Zhang, A.; Zhang, W.; Li, B.; and Li, M. 2023. MixGen: A New Multi-Modal Data Augmentation. In *IEEE WACV 2023 - Pre train Workshop*, volume abs/2206.08358.

Harley, A. W.; Ufkes, A.; and Derpanis, K. G. ???? Evaluation of Deep Convolutional Nets for Document Image

Classification and Retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Ho, C.-H.; Appalaraju, S.; Jasani, B.; Manmatha, R.; and Vasconcelos, N. 2022. YORO-Lightweight End to End Visual Grounding. In *European Conference on Computer Vision - ECCV CAMP Workshop*.

Hong, T.; Kim, D.; Ji, M.; Hwang, W.; Nam, D.; and Park, S. 2020a. BROS: A Pre-trained Language Model for Understanding Texts in Document. *under review https://openreview.net/references/pdf?id=uCz3OR6CJT*.

Hong, T.; Kim, D.; Ji, M.; Hwang, W.; Nam, D.; and Park, S. 2020b. BROS: A Pre-trained Language Model for Understanding Texts in Document. *https://openreview.net/forum?id=punMXQEsPr0*.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9992–10002.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv preprint arXiv:2204.08387*.

Hwang, W.; Yim, J.; Park, S.; Yang, S.; and Seo, M. 2020. Spatial Dependency Parsing for Semi-Structured Document Information Extraction. arXiv:2005.00642.

Jaume, G.; Ekenel, H. K.; and Thiran, J.-P. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2: 1–6.

Kant, Y.; Batra, D.; Anderson, P.; Schwing, A.; Parikh, D.; Lu, J.; and Agrawal, H. 2020. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, 715–732.

Kil, J.; Changpinyo, S.; Chen, X.; Hu, H.; Goodman, S.; Chao, W.-L.; and Soricut, R. 2022. PreSTU: Pre-Training for Scene-Text Understanding. *ArXiv*, abs/2209.05534.

Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; and Park, S. 2021. OCR-Free Document Understanding Transformer. In *European Conference on Computer Vision*.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.

Kudo, T.; and Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Conference on Empirical Methods in Natural Language Processing*.

Lee, C.-Y.; Li, C.-L.; Dozat, T.; Perot, V.; Su, G.; Hua, N.; Ainslie, J.; Wang, R.; Fujii, Y.; and Pfister, T. 2022a. FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction. In *Annual Meeting of the Association for Computational Linguistics*.

Lee, K.; Joshi, M.; Turc, I.; Hu, H.; Liu, F.; Eisenschlos, J. M.; Khandelwal, U.; Shaw, P.; Chang, M.-W.; and Toutanova, K. 2022b. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. *ArXiv*, abs/2210.03347.

Li, B.; Wang, J.; Zhao, M.; and Zhou, S. 2022a. Two-Stage Multimodality Fusion for High-Performance Text-Based Visual Question Answering. In *Asian Conference on Computer Vision*.

Li, C.; Bi, B.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; and Si, L. 2021a. StructuralLM: 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)*, 6309–6318.

Li, C.; Fehérvári, I.; Zhao, X.; Macêdo, I.; and Appalaraju, S. 2022b. SeeTek: Very Large-Scale Open-set Logo Recognition with Text-Aware Metric Learning. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 587–596.

Li, J.; Xu, Y.; Cui, L.; and Wei, F. 2021b. MarkupLM: Pre-training of Text and Markup Language for Visually Rich Document Understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Li, P.; Gu, J.; Kuen, J.; Morariu, V. I.; Zhao, H.; Jain, R.; Manjunatha, V.; and Liu, H. 2021c. SelfDoc: Self-Supervised Document Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5648–5656.

Li, Y.; Qian, Y.; Yu, Y.; Qin, X.; Zhang, C.; Liu, Y.; Yao, K.; Han, J.; Liu, J.; and Ding, E. 2021d. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Litman, R.; Anschel, O.; Tsiper, S.; Litman, R.; Mazor, S.; and Manmatha, R. 2020. SCATTER: selective context attentional scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11962–11972.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Lu, X.; Fan, Z.; Wang, Y.; Oh, J.; and Rosé, C. P. 2021. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2631–2639.

Luo, S.; Cao, F.; Núñez, F. W.; Wen, Z.; Poon, J.; and Han, C. 2022. SceneGATE: Scene-Graph based co-Attention networks for TExt visual question answering. *ArXiv*, abs/2212.08283.

Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. 2022. InfographicVQA. In *Proceedings of* *the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1697–1706.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2200–2209.

Mathew, M.; Karatzas, D.; Manmatha, R.; and Jawahar, C. V. 2020. DocVQA: A Dataset for VQA on Document Images. arXiv:2007.00398.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Peng, Q.; Pan, Y.; Wang, W.; Luo, B.; Zhang, Z.; Huang, Z.; Hu, T.; Yin, W.; Chen, Y.; Zhang, Y.; et al. 2022. ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. *arXiv preprint arXiv:2210.06155*.

Powalski, R.; Borchmann, Ł.; Jurkiewicz, D.; Dwojak, T.; Pietruszka, M.; and Pałka, G. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, 732–747.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Raffel, C.; Shazeer, N. M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv*, abs/1910.10683.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 1137–1149.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*, abs/1505.04597.

Seunghyun, P.; Seung, S.; Bado, L.; Junyeop, L.; Jaeheung, S.; Minjoon, S.; and Hwalsuk, L. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Tang, P.; Appalaraju, S.; Manmatha, R.; Xie, Y.; and Mahadevan, V. 2023a. Multiple-Question Multiple-Answer Text-VQA. *arXiv preprint arXiv:2311.08622*.

Tang, P.; Zhu, P.; Li, T.; Appalaraju, S.; Mahadevan, V.; and Manmatha, R. 2023b. DEED: Dynamic Early Exit on Decoder for Accelerating Encoder-Decoder Transformer Models. *arXiv preprint arXiv:2311.08623*.

Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.-Y.; and Bansal, M. 2022. Unifying Vision, Text, and Layout for Universal Document Processing. *ArXiv*, abs/2212.02623.

Thomas, W.; Lysandre, D.; Victor, S.; Julien, C.; Clement, D.; Anthony, M.; Pierric, C.; Tim, R.; Rémi, L.; Funtowicz, M.; et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wang, J.; Gao, M.; Hu, Y.; Selvaraju, R. R.; Ramaiah, C.; Xu, R.; JaJa, J. F.; and Davis, L. S. 2022a. TAG: Boosting Text-VQA via Text-aware Visual Question-answer Generation. *arXiv preprint arXiv:2208.01813*.

Wang, J.; Jin, L.; and Ding, K. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022b. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*.

Wang, Z.; Zhou, Y.; Wei, W.; Lee, C.-Y.; and Tata, S. 2022c. A Benchmark for Structured Extractions from Complex Documents. *ArXiv*, abs/2211.15421.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020a. 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*, 1192–1200.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. 2020b. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *arXiv preprint arXiv:2012.14740*.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. 2021. 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)*, 2579–2591.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8751–8761.

Łukasz Borchmann; Pietruszka, M.; Stanisławek, T.; Jurkiewicz, D.; Turski, M. P.; Szyndler, K.; and Gralinski, F. 2021. DUE: End-to-End Document Understanding Benchmark. In *NeurIPS Datasets and Benchmarks*.

DocFormerv2: Local Features for Document Understanding - Supplemental

Srikar Appalaraju[1] *, Peng Tang[1], Qi Dong[1], Nishant Sankaran[1], Yichu Zhou[2] †, R. Manmatha[1]

[1]AWS AI Labs, [2]School of Computing at University of Utah
{srikara, tangpen, qdon, nishsank, manmatha}@amazon.com, flyaway@cs.utah.edu

Supplemental

This is the supplemental material for the main DocFormerv2 paper (Appalaraju et al. 2023). Please read the main paper for model formulation, performance numbers on various datasets and further analysis and ablation. DocFormerv2 achieves state-of-the-art performance on eight datasets each with strong baselines. Specifically, on Table Fact (Chen et al. 2019; Łukasz Borchmann et al. 2021) (+4.0%), InfoVQA (Mathew et al. 2022) (1.4%), FUNSD (Jaume, Ekenel, and Thiran 2019) (1%). On three VQA tasks involving scene-text, DocFormerv2 outperforms previous comparably-sized models and even does better than much larger models (such as GIT2, PaLi and Flamingo) on some tasks.

Implementation Details

We present all the hyper-parameters in Table 1 used for pre-training and fine-tuning DocFormerv2 . We follow the standard practice of unsupervised pre-training followed by supervised fine-tuning (Xu et al. 2020a,b; Hong et al. 2020; Appalaraju et al. 2021; Biten et al. 2022; Ho et al. 2022). We emphasize that even for Text-VQA datasets, the same document pre-trained model was used. We only performed supervised fine-tuning on provided Text-VQA train data.

Specifically, FUNSD (Guillaume Jaume 2019) datasets were fine-tuned for 300 epochs. CORD (Seunghyun et al. 2019) for 200 epochs.

For pre-training, the DocFormerv2 is initialized with T5 (Raffel et al. 2019) pre-trained weights for the language branch and keep the visual weights randomly initialized. During unsupervised pre-training, the DocFormerv2 learns to fuse and utilize multi-modal features (vision, language and spatial) to minimize the three objectives - Token-to-Grid, Token-to-Line and denoising language modeling task.

DUE Benchmarks(Łukasz Borchmann et al. 2021): Is a collection of seven datasets on document understanding task. We use four datasets from the DUE benchmarks - DocVQA (Mathew et al. 2020), InfoVQA (Mathew et al. 2022), TabFact (Chen et al. 2019) and WikiTableQuestions (Pasupat and Liang 2015). They were chosen due to their diversity and to show the versitility of DocFormerv2 . The other three datasets DeepForm , KleisterCharity and PWC are similar to VQA setting and in the interest of time, we do not evaluate on them.

Hyper-Parameter	Pre-training	Fine-tuning
Epochs	3	varies
Learning rate	5E-05	1E-05/2.5E-05
Warm-up	1000 iters	0
Gradient Clipping	1.0	1.0
Gradient agg.	False	False
Optimizer	AdamW	AdamW
Lower case	True	True
Sequence length	512	varies
Encoder layers	12/24 (B/L)	12/24 (B/L)
32-bit mixed precision	True	True
Batch size	9/1 per GPU (B/L)	9/1 per GPU (B/L)
GPU hardware	A100 (40GB)	A100 (40GB)
Training Num. Samples	64M	varies

Table 1: **Implementation Details**: Hyper-parameters used for pre-training DocFormerv2 and fine-tuning for down-stream tasks. Training epochs vary for down-stream tasks. **B**: DocFormerv2 $_{\text{base}}$ and **L**: DocFormerv2 $_{\text{large}}$

Ablations

Similar to the main paper ablations, here too, ablation was performed on DFv2-small model with 1M doc pre-training followed by 40 epoch fine-tuning on DocVQA (Mathew et al. 2020) dataset.

Robustness to OCR errors. DFv2 consumes OCR text which can have errors. Since it also has a generative decoder, in theory it is robust to certain distortions and noise from OCR-text. To quantify the degree of robustness, we conduct a study using the FUNSD dataset and artificially introduce noise/typographical errors to the input words, simulating OCR errors. Specifically, for every character in the text, we randomly replace it with an erroneous character with a probability p, limiting to a maximum of 1 character error per word. We then evaluate the performance of the DocFormerv2$_{\text{base}}$ and LayoutLMv2$_{\text{base}}$ models on the error injected text to observe their resilience to such noise [1]. From Figure 1, we see that for increasing amount of injected OCR

*Corresponding author.

†Work conducted during an internship at Amazon.

[1] Note that both LayoutLMv2 and ours use visual features, and thus it is fair to compare the robustness in multi-modality context.

errors. Specifically, 20% OCR errors only decreases the performance by -1.68% whereas an encoder-only model decreases by a wide margin -9.84%. This shows the benefit of our approach (in having a generative decoder).

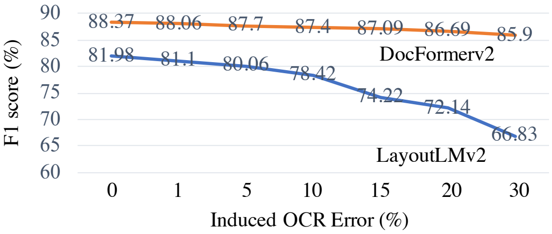


Figure 1: **Induced OCR Error Ablation**. F1 score performance evaluated on FUNSD for varying orders of injected OCR errors.

Do we need a separate Image encoder? We investigate if DocFormerv2 's way of consuming visual features is optimal for VDU. Instead of using linear features, we use Swinv2 (Liu et al. 2021) and pre-train this setup on 1M documents. When fine-tuned on DocVQA we observe that the setup with Swinv2 as visual backbone, substantially underperforms our approach by (+4.3%). For this task, the complex visual features from Swinv2 are less beneficial than our simple linear features.

Model	image encoder	DocVQA eval ANLS (%)
baseline	-	69.0
DocFormerv2$_{\text{small}}$	Swinv2$_{\text{small}}$	66.2
DocFormerv2$_{\text{small}}$	Linear (ours)	70.5 (+4.3%)

Table 2: **Image Encoder Ablation**: All models pre-trained on 1M docs from IDL. Swinv2 too was pre-trained and fine-tuned.

Model	Setting	DocVQA eval ANLS (%)
DocFormerv2$_{\text{small}}$	No Mask BBox	71.9
DocFormerv2$_{\text{small}}$	Mask BBox	72.4 (+0.5%)

Table 3: **Grid Bounding box Masking**: We clearly see the benefit of this design choice where masking the bounding box information (along with the tokens) helps the model generalize better.

Varying Image Tokens. Concatenating the image tokens along with the text tokens is a simple and intuitive approach for the model to learn to jointly capture multi-modal information. However, since we are limited on the total number of tokens we can use, it begs the question - what is a suitable proportion of vision-to-text tokens to be used for this design? We perform ablations in this regard to measure the performance obtained by finetuning the model on FUNSD dataset with varying ratios of vision tokens to text tokens. Figure 2 shows that 128 image tokens appears to provide the best performance compared to the other settings.

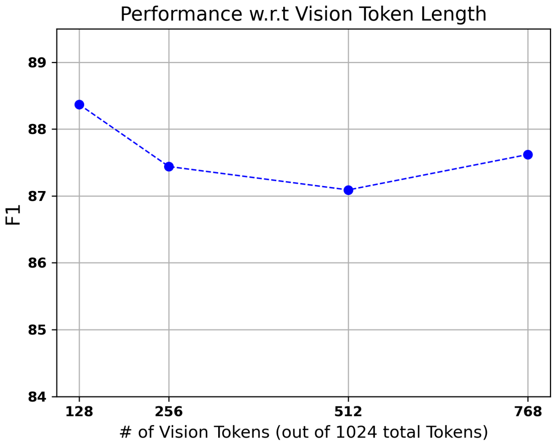


Figure 2: **Image token length ablation**. Effect on model performance w.r.t variation of the proportion of vision tokens to text tokens provided as input to the model. 128 works best and was used as the final model design.

Grid Bounding Box Masking In DocFormerv2 , for the de-noising language modeling task, along with the masked tokens, we also mask out the corresponding bounding box. Masking the bounding boxes make the grid prediction and line prediction task harder. This is because the model will not have the position information of the masked tokens. It has to infer from the context. In this ablation, we show the effectiveness of this design choice on DocVQA (Mathew et al. 2020) dataset. See Table 3.

Datasets

In this section we describe the pre-training and fine-tuned datasets used in DocFormerv2 paper.

Pre-training Dataset

Following prior-art (Biten et al. 2022; Appalaraju et al. 2021; Tang et al. 2022; Xu et al. 2020a,b; Huang et al. 2022) we pre-train DocFormerv2 on Industrial Document Library (IDL) dataset [2]. As documented in the main paper, the IDL is an electronic repository of documents generated by industries that have an impact on public health, and is hosted by the University of California, San Francisco Library. It encompasses millions of publicly disclosed documents from a range of industries such as pharmaceuticals, chemicals, food, and fossil fuels. The IDL dataset, which contains approximately 13 million documents, equating to about 70 million pages (64 million usable) of various document images, was crawled from the website. Various document types, such as forms, tables, and letters, with diverse layouts are available on IDL. To extract text from the documents, OCR model was run for each document. OCR predictions could be noisy (depending on the quality of the document image). Before pre-training, the crawled and OCR'ed IDL data was pre-processed. Pre-processing was done to

[2] https://www.industrydocuments.ucsf.edu/

discard documents based on quality. After pre-processing and pruning, 64 million documents (approximately 6 million documents were discarded) were retained for pre-training.

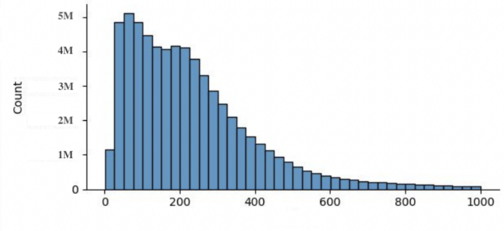


Figure 3: **IDL data distribution**: Frequency plot for predicted OCR tokens in IDL dataset used for DocFormerv2 pre-training

The detected OCR word distribution across all the 64M documents is presented in Fig. 3, indicating a right-skewed normal distribution with the majority of documents containing 20 to 400 words per document. We note that IDL dataset contains more words on average than image-text datasets like CC and LAION datasets, which is particularly advantageous for pre-training in Text-VQA tasks. Fig. 4 illustrates representative examples from the IDL dataset.

Document Understanding Datasets

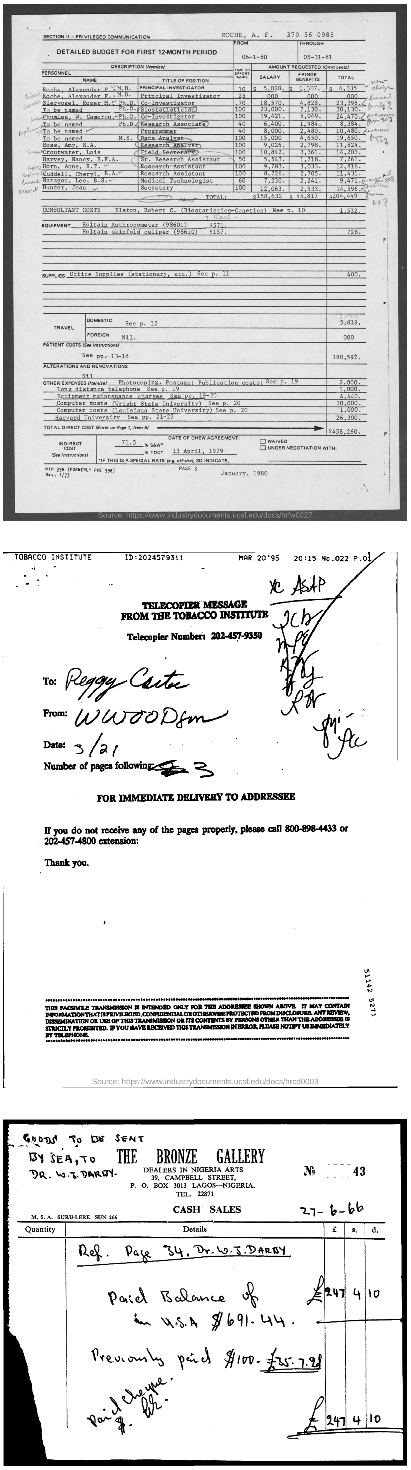
Dataset	Size (k documents)			Type	Metric
	Train	Dev	Test		
TabFact	13.2	1.7	1.7	Table NLI	Acc.
WikiTableQuestions	1.4	0.3	0.4	Table QA	Acc.
DocVQA	10.2	1.3	1.3	Visual QA	ANLS
InfoVQA	4.4	0.5	0.6	Visual QA	ANLS
FUNSD	0.15	-	0.05	Entity Labeling	F1
CORD	0.8	0.1	0.1	Entity Extraction	F1

Table 4: **Document Understanding Datasets Details**: Details of datasets types and their train/val/test split distributions for TabFact (Chen et al. 2019), WikiTableQuestions (Pasupat and Liang 2015), DocVQA (Mathew et al. 2020), InfoVQA (Mathew et al. 2022), FUNSD (Jaume, Ekenel, and Thiran 2019) and CORD (Seunghyun et al. 2019).

TabFact (Chen et al. 2019) contains 16k Wikipedia tables as the basis for 118k human annotated questions and statements sourced through Amazon Mechanical Turk (AMT). The dataset focuses on the task of Natural Language Inference (NLI) based on reasoning with the provided tabular data. The goal is to recognize which statements are validated/refuted by the tabular content associated with it.

WikiTableQuestions (Łukasz Borchmann et al. 2021; Pasupat and Liang 2015) is a dataset consisting of 2,108 semi-structured HTML tables from Wikipedia and 22,033 question-answer pairs about the content in the tables. AMT workers gathered trivia questions about the tables that covered a wide range of operations such as table lookup, aggregation, joins, etc.

DocVQA (Mathew et al. 2020) consists of 50,000 questions defined on 12,767 document images. The task it defines



SECTION II – PRIVILEGED COMMUNICATION ROCHE, A. F. 370 56 0985

DETAILED BUDGET FOR FIRST 12-MONTH PERIOD FROM 06-1-80 THROUGH 05-31-81

DESCRIPTION *(Itemize)* — PERSONNEL — AMOUNT REQUESTED *(Omit cents)*

NAME	TITLE OF POSITION	TIME OR EFFORT %/HRS.	SALARY	FRINGE BENEFITS	TOTAL
Roche, Alexander F., M.D.	PRINCIPAL INVESTIGATOR	10	$ 5,028.	$ 1,307.	$ 6,335
Roche, Alexander F., M.D.	Principal Investigator	25	000	000	000
Siervogel, Roger M., Ph.D.	Co-Investigator	70	18,570.	4,828.	23,398.
To be named Ph.D.	Biostatistician	100	23,000.	7,130.	30,130.
Chumlea, W. Cameron, Ph.D.	Co-Investigator	100	19,421.	5,049.	24,470.
To be named Ph.D.	Research Associate	40	6,400.	1,984.	8,384.
To be named	Programmer	40	8,000.	2,480.	10,480.
To be named M.S.	Data Analyst	100	15,000	4,650.	19,650.
Ross, Amy, B.A.	Research Analyst	100	9,026.	2,798.	11,824.
Croutwater, Lois	Field Secretary	100	10,842.	3,361.	14,203.
Harvey, Nancy, B.F.A.	Sr. Research Assistant	50	5,543.	1,718.	7,261.
Horn, Anne, R.T.	Research Assistant	100	9,783.	3,033.	12,816.
Cuddell, Cheryl, B.A.	Research Assistant	100	8,726.	2,705.	11,431.
Naragon, Lee, B.S.	Medical Technologist	60	7,230.	2,241.	9,471.
Hunter, Joan	Secretary	100	12,063.	2,533.	14,596.
	new TOTAL:		$158,632	$ 45,817	$204,449

CONSULTANT COSTS Elston, Robert C. (Biostatistics-Genetics) See p. 10 + Reed -	1,532.
EQUIPMENT Holtain Anthropometer (98601) $571.	
Holtain skinfold caliper (98610) $157.	728.
SUPPLIES Office Supplies (stationery, etc.) See p. 11	400.
TRAVEL DOMESTIC See p. 12	~~5,819.~~ 3,500
TRAVEL FOREIGN Nil.	000
PATIENT COSTS *(See instructions)* See pp. 13-18	180,592.
ALTERATIONS AND RENOVATIONS Nil	
OTHER EXPENSES *(Itemize)* Photocoping, Postage; Publication costs; See p. 19	2,000.
Long distance telephone See p. 19	1,000.
Equipment maintenance charges See pp. 19-20	4,440.
Computer costs (Wright State University) See p. 20	30,000.
Computer costs (Louisiana State University) See p. 20	1,000.
Harvard University See pp. 21-22	26,300.
TOTAL DIRECT COST *(Enter on Page 1, Item 5)* →	$458,260.

INDIRECT COST *(See Instructions)* 71.5 % S&W* ____ % TDC* DATE OF DHEW AGREEMENT: 13 April, 1979 □ WAIVED □ UNDER NEGOTIATION WITH:

*IF THIS IS A SPECIAL RATE *(e.g. off-site)*, SO INDICATE.

NIH 398 (FORMERLY PHS 398) Rev. 1/73 PAGE 3 January, 1980

Source: https://www.industrydocuments.ucsf.edu/docs/hrfw0227

TOBACCO INSTITUTE ID:2024579311 MAR 20'95 20:15 No.022 P.01

xc ASAP

TELECOPIER MESSAGE FROM THE TOBACCO INSTITUTE

Telecopier Number: 202-457-9350

To: Peggy Carter

From: WWOODSm

Date: 3/21

Number of pages following: 3

fyi

FOR IMMEDIATE DELIVERY TO ADDRESSEE

If you do not receive any of the pages properly, please call 800-898-4433 or 202-457-4800 extension:

Thank you.

51142 5271

THIS FACSIMILE TRANSMISSION IS INTENDED ONLY FOR THE ADDRESSEE SHOWN ABOVE. IT MAY CONTAIN INFORMATION THAT IS PRIVILEGED, CONFIDENTIAL OR OTHERWISE PROTECTED FROM DISCLOSURE. ANY REVIEW, DISSEMINATION OR USE OF THIS TRANSMISSION OR ITS CONTENTS BY PERSONS OTHER THAN THE ADDRESSEE IS STRICTLY PROHIBITED. IF YOU HAVE RECEIVED THIS TRANSMISSION IN ERROR, PLEASE NOTIFY US IMMEDIATELY BY TELEPHONE.

Source: https://www.industrydocuments.ucsf.edu/docs/hrcd0003

GOODS TO BE SENT
BY SEA, TO
DR. W.J. DARBY.

THE BRONZE GALLERY
DEALERS IN NIGERIA ARTS
39, CAMPBELL STREET,
P. O. BOX 3013 LAGOS—NIGERIA.
TEL. 22871

№ 43

CASH SALES 27-6-66

M. S. A. SURU-LERE SUN 266

Quantity	Details	£	s.	d.
	Ref. Page 34. Dr. W.J. DARBY			
	Paid Balance of in U.S.A $691.44.	£247	4	10
	Previously paid $100. £35.7.2d			
	Paid cheque. $.	£247	4	10

Figure 4: **IDL sample documents visualization**

Dataset	Train Set	Val Set	Test Set
OCR-VQA (Mishra et al. 2019)	166K/801.7K	20.7K/100K	20.8K/100.4K
TextVQA (Singh et al. 2019)	21.9K/34.6K	3.2K/5K	3.3K/5.7K
ST-VQA (Biten et al. 2019)	17K/23.4K	1.9K/2.6K	3K/4.1K

Table 5: **Text-VQA Dataset Stats:** The number of images/questions of different text-VQA datasets. We report performance on Test-set in the main paper after training only on Train-set.

is to verify the capability of answering questions about document images and the textual content within. It encompasses typed and handwritten document samples and also requires the comprehension of layout, indicators (like ticks, etc) and style of the presented content to answer questions about it.

InfographicVQA (Mathew et al. 2022) is a collection of 5,485 images about infographics collected from 2,594 web domains. It poses 30,035 questions pertaining to these images grounded on tables, figures and visualizations. The questions involve basic numeric reasoning elements such as counting, arithmetic and sorting operations.

FUNSD (Jaume, Ekenel, and Thiran 2019) is a dataset that poses the challenge of understanding forms structure and content from noisy scanned images. It consists of 199 real, fully annotated scanned forms which are split into 149 for train and 50 for test. It has defines 3 tasks: (i) word grouping (clubbing words belonging to the same semantic entity), (ii) semantic entity labeling and (iii) entity linking (establish relations between entities).

CORD (Seunghyun et al. 2019) consists of 11,000 scanned receipts collected from shops and restaurants which contains annotations for OCR and multi-level semantic parsing tasks. The fine-grained classification task entails 8 superclasses and 54 subclass labels for the word entities within the receipts like menu name, quantity, total price, and so on.

Text-VQA datasets

TextVQA(Singh et al. 2019) contains 28k images from the Open Images dataset. The questions and answers are collected through Amazon Mechanical Turk (AMT). The workers are instructed to come up with questions that require reasoning about the scene text in the image using the following process - , 10 answers were collected for each question.

ST-VQA (Biten et al. 2019) is an aggregation of well-known computer vision datasets, namely: ICDAR 2013 , ICDAR2015 , ImageNet , VizWiz [7], IIIT Scene Text Retrieval , Visual Genome and COCO-Text . ST-VQA is collected through Amazon Mechanical Turk, asking workers to come up with questions so that the answer is always the scene text in the image.

OCR-VQA (Mishra et al. 2019) is composed of 207,572 images of book covers and contains more than 1 million QA pairs about these images. The questions are template-based, asking about information on the book such as title, author, year. The questions are all can be answered by inferring the book cover images.

Related Work Discussion

We add more discussions of the existing works in Visual Document Understanding domain.

Encoder-Only VDU models LayoutLMv1(Xu et al. 2020a) is a representative work using a Transformer encoder model to tackle the entity extraction and labelling task. Besides text level token embeddings, LayoutLMv1 adds spatial embeddings and OCR-patch visual embedding for individual word tokens. Specifically, it uses OCR box locations as 2D position embeddings, and uses a Faster-RCNN model to extract local visual features for each input word. The final embedding of each input tokens are the summation of text embedding, 2D position embedding, 1D position embedding and OCR patch visual embedding. To enhance the visual information for each token, LayoutLMv2(Xu et al. 2020b) uses one CNN to extract the global image features, and splits the image features into multiple visual tokens, and concatenates them with text tokens as the input sequence. It emphasises the visual information learning in attention layers. To further simplify the visual token representation, LayoutLMv3 (Huang et al. 2022) uses one Conv2D layer to split the original images into multiple visual tokens, rather than using an extra visual encoder to extract global image features, and leveraging visual features as visual tokens. These encoder-only models provide simple frameworks to handle VDU tasks, but limit the application scenarios due to the token classification formulation. For example, it is hard to obtain the complete entities by leveraging the predicted BIO tags for non-reading order input sequences and complex entity layouts.

Encoder-Decoder VDU models TILT (Powalski et al. 2021) designs a Transformer Encoder-Decoder model to tackle entity labelling and entity extraction tasks. It uses U-Net as visual encoder and uses T5 decoder to conduct the label prediction. While TILT (Powalski et al. 2021) proposed a encoder-decoder transformer for VDU, they only train it on one pre-training task (masked language modeling) and also use a bulky visual CNN. Our approach DocFormerv2 , not only simplifies the architecture by not using a separate visual module (CNN or Transformer based) and has multiple unsupervised pre- training tasks. Several other popular encoder-decoder VDU models (Tang et al. 2023b,a) were proposed recently (after this paper came out).

Qualitative Analysis

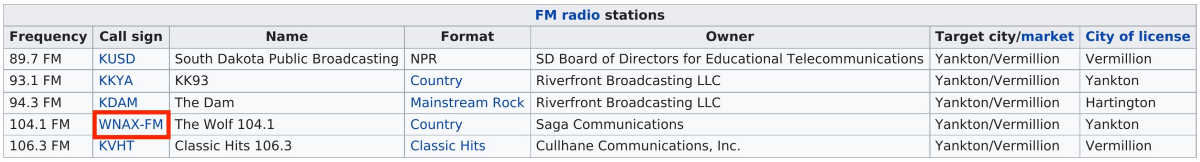
In this section, we present qualitative analysis of DocFormerv2 by visualizing its predictions on various datasets.

WikiTableQuestions Vizualizations

We present DocFormerv2 prediction visualizations in this section. See Fig. 5, 6, 7, 8.

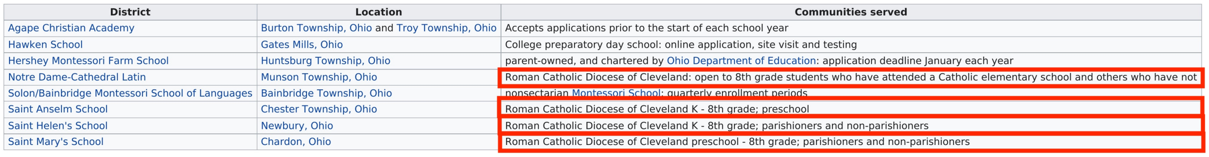
FUNSD Visualizations

DocFormerv2 achieves state-of-the-art performance of 88.89% F1-score (see Section 4.3 in the main file) on FUNSD (Guillaume Jaume 2019) dataset amongst other multi-modal models with same settings. In this sub-section



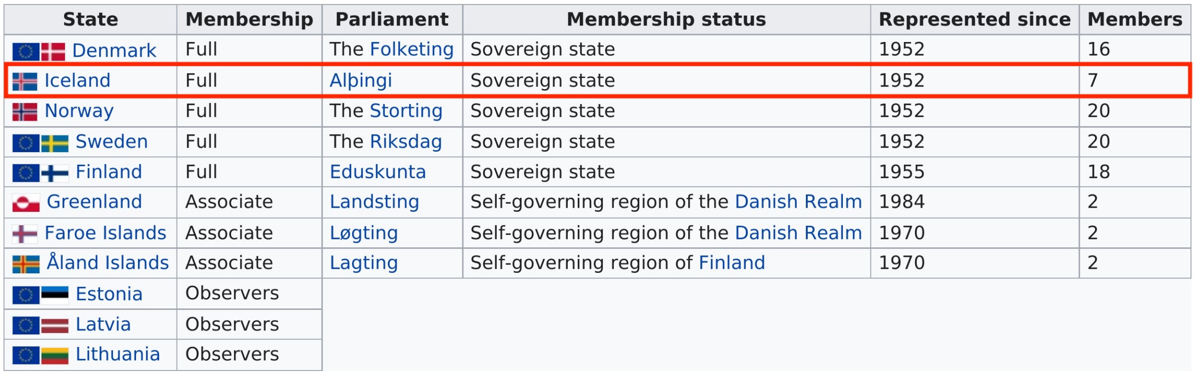
FM radio stations						
Frequency	Call sign	Name	Format	Owner	Target city/market	City of license
89.7 FM	KUSD	South Dakota Public Broadcasting	NPR	SD Board of Directors for Educational Telecommunications	Yankton/Vermillion	Vermillion
93.1 FM	KKYA	KK93	Country	Riverfront Broadcasting LLC	Yankton/Vermillion	Yankton
94.3 FM	KDAM	The Dam	Mainstream Rock	Riverfront Broadcasting LLC	Yankton/Vermillion	Hartington
104.1 FM	WNAX-FM	The Wolf 104.1	Country	Saga Communications	Yankton/Vermillion	Yankton
106.3 FM	KVHT	Classic Hits 106.3	Classic Hits	Cullhane Communications, Inc.	Yankton/Vermillion	Vermillion

Figure 5: **DocFormerv2 predictions on WikiTableQuestions**: For the question `"which of these stations do not have a 'k' in their call sign?"`, DocFormerv2 correctly predicts `"WNAX-FM"` (4th row, 2nd column). This particular example requires reasoning over spatial, visual and language features. Image file in test: `csv_200csv_18_0.jpeg`. The **Red** box showing the answer for locating the answer.



District	Location	Communities served
Agape Christian Academy	Burton Township, Ohio and Troy Township, Ohio	Accepts applications prior to the start of each school year
Hawken School	Gates Mills, Ohio	College preparatory day school: online application, site visit and testing
Hershey Montessori Farm School	Huntsburg Township, Ohio	parent-owned, and chartered by Ohio Department of Education: application deadline January each year
Notre Dame-Cathedral Latin	Munson Township, Ohio	Roman Catholic Diocese of Cleveland: open to 8th grade students who have attended a Catholic elementary school and others who have not
Solon/Bainbridge Montessori School of Languages	Bainbridge Township, Ohio	nonsectarian Montessori School: quarterly enrollment periods
Saint Anselm School	Chester Township, Ohio	Roman Catholic Diocese of Cleveland K - 8th grade; preschool
Saint Helen's School	Newbury, Ohio	Roman Catholic Diocese of Cleveland K - 8th grade; parishioners and non-parishioners
Saint Mary's School	Chardon, Ohio	Roman Catholic Diocese of Cleveland preschool - 8th grade; parishioners and non-parishioners

Figure 6: **DocFormerv2 predictions on WikiTableQuestions**: For the question `"how many of the schools serve the roman catholic diocese of cleveland?"`, DocFormerv2 correctly predicts `"4"`. This particular example requires arithmetic counting and reasoning over multiple rows as the model needs to count to generate the final answer. We point to the reader that no arithmetic specific pre-training like Tapas (Herzig et al. 2020) was done for DocFormerv2 . Image file in test: `csv_200csv_8_0.jpeg`



State	Membership	Parliament	Membership status	Represented since	Members
Denmark	Full	The Folketing	Sovereign state	1952	16
Iceland	Full	Alþingi	Sovereign state	1952	7
Norway	Full	The Storting	Sovereign state	1952	20
Sweden	Full	The Riksdag	Sovereign state	1952	20
Finland	Full	Eduskunta	Sovereign state	1955	18
Greenland	Associate	Landsting	Self-governing region of the Danish Realm	1984	2
Faroe Islands	Associate	Løgting	Self-governing region of the Danish Realm	1970	2
Åland Islands	Associate	Lagting	Self-governing region of Finland	1970	2
Estonia	Observers				
Latvia	Observers				
Lithuania	Observers				

Figure 7: **DocFormerv2 predictions on WikiTableQuestions**: For the question `"which state has a full membership and also has a membership status under sovereign state with only 7 members?"`, DocFormerv2 correctly predicts `"Iceland"`. We hypothesize that our Token-to-Line pre-training helps in such scenarios as our approach has efficiently learned to make sense of local information (here information in a row). Image file in test: `csv_201csv_20_0.jpeg`

we look at more visualizations by DocFormerv2 on the testset. One important aspect of this VDU we would like to mention is the OCR is not in human reading-order.

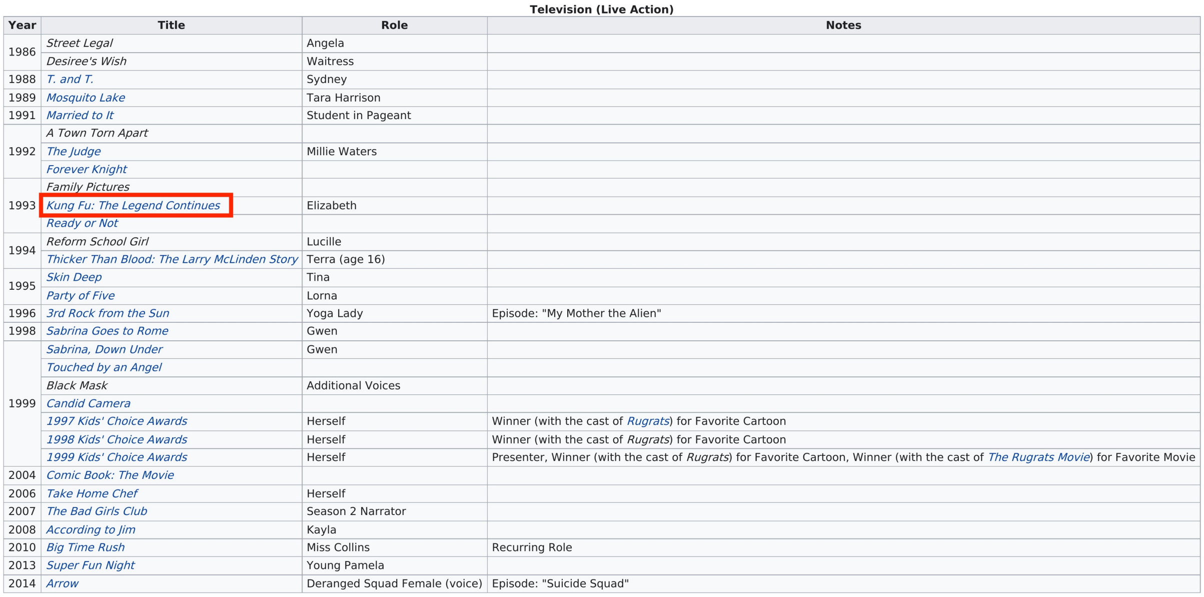
In Figure 9 and 10, we see that DocFormerv2 correctly predicts entity labels on these two samples, and the model can handle the long sequence entity labelling well. We observe some prediction errors on class “Other' and “Header', as shown in Figure 11. It is mainly due to lack of training samples for these two particular classes.

CORD Visualizations

DocFormerv2 matches the state-of-the-art performance of 97.70% F1-score on CORD (Seunghyun et al. 2019) dataset. Please see Section 4.4 in the main paper. In this sub-section we look at CORD (Seunghyun et al. 2019) visualizations by DocFormerv2 . We explicitly show hard-cases where DocFormerv2 does well, see Figures 13. In the groundtruth visualization, we show the correct entity class on the top-left of the entity boxes. Figure 12 shows some wrong predictions by DocFormerv2 . It is because class menu.etc has smaller training samples, and the order of the words in this images is different from the reading order.

Scene-Text VQA Visualizations

Figure 14 shows examples of correct and wrong predictions from DocFormerv2. As we can see, DocFormerv2



Television (Live Action)

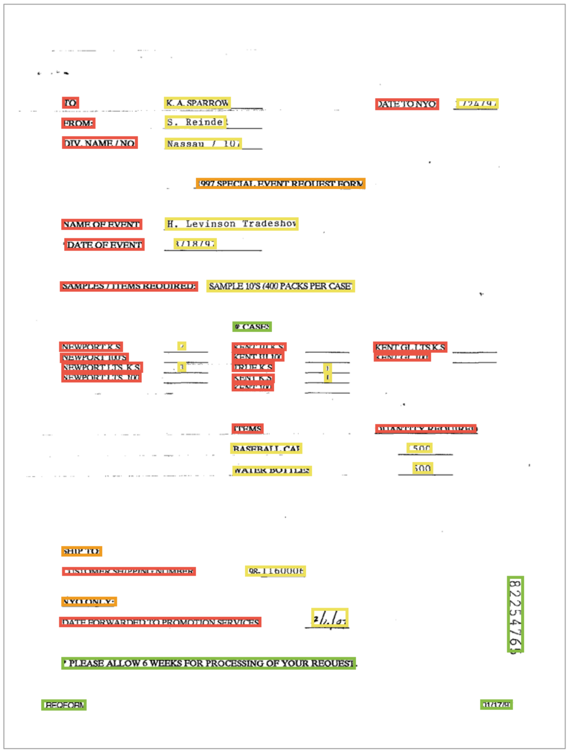
Year	Title	Role	Notes
1986	*Street Legal*	Angela	
	Desiree's Wish	Waitress	
1988	*T. and T.*	Sydney	
1989	*Mosquito Lake*	Tara Harrison	
1991	*Married to It*	Student in Pageant	
1992	*A Town Torn Apart*		
	The Judge	Millie Waters	
	Forever Knight		
1993	*Family Pictures*		
	Kung Fu: The Legend Continues	Elizabeth	
	Ready or Not		
1994	*Reform School Girl*	Lucille	
	Thicker Than Blood: The Larry McLinden Story	Terra (age 16)	
1995	*Skin Deep*	Tina	
	Party of Five	Lorna	
1996	*3rd Rock from the Sun*	Yoga Lady	Episode: "My Mother the Alien"
1998	*Sabrina Goes to Rome*	Gwen	
1999	*Sabrina, Down Under*	Gwen	
	Touched by an Angel		
	Black Mask	Additional Voices	
	Candid Camera		
	1997 Kids' Choice Awards	Herself	Winner (with the cast of *Rugrats*) for Favorite Cartoon
	1998 Kids' Choice Awards	Herself	Winner (with the cast of *Rugrats*) for Favorite Cartoon
	1999 Kids' Choice Awards	Herself	Presenter, Winner (with the cast of *Rugrats*) for Favorite Cartoon, Winner (with the cast of *The Rugrats Movie*) for Favorite Movie
2004	*Comic Book: The Movie*		
2006	*Take Home Chef*	Herself	
2007	*The Bad Girls Club*	Season 2 Narrator	
2008	*According to Jim*	Kayla	
2010	*Big Time Rush*	Miss Collins	Recurring Role
2013	*Super Fun Night*	Young Pamela	
2014	*Arrow*	Deranged Squad Female (voice)	Episode: "Suicide Squad"

Figure 8: **DocFormerv2 predictions on WikiTableQuestions**: For the question "which movie came out in 1993 and has the role of elizabeth?", DocFormerv2 correctly predicts "Kung Fu: The Legend Continues". We hypothesize that our Token-to-Line pre-training helps in such scenarios as our approach has efficiently learned to make sense of local information (here information in a row). Image file in test: csv_201csv_23_0.jpeg

is able to understand information from different modalities to give correct answers. For example, for the fourth image in the last row, the model needs to understand the meanings of "establishment" (language), "next to" (spatial), "red" (visual), etc. to infer the correct answer. There are multiple reasons for the error cases, including OCR error, hard images/questions, ground truth mismatch, etc.. For example, for the last image in the fourth row, DocFormerv2 gives prediction g5a (vs. ground truth gsa). This is because OCR wrongly recognizes the word gsa as g5a. For the last image in the fifth row, it is even difficult for human to figure out the correct answer (we saw big discrepancies in the annotations from different annotators). For the last image in the last row, the ground truth is tongue-in-cheek warnings and DocFormerv2 gives prediction tongue-in-cheek warnings at waste land site. We can see our prediction is reasonable and just gives more information in the prediction.

Conclusion

In this supplemental, we present more experiments and ablations for our main paper DocFormerv2 (Appalaraju et al. 2023). We provide further evidence DocFormerv2 is very effective for the challenging task of Visual Document Understanding. Furthermore, we show evidence that our work generalizes to other domain like Scene-Text VQA in-spite of explicitly not being trained on natural images.



TO: K. A. SPARROW DATE TO NYO: 1/24/97

FROM: S. Reindel

DIV. NAME / NO: Nassau / 107

1997 SPECIAL EVENT REQUEST FORM

NAME OF EVENT: H. Levinson Tradeshow

DATE OF EVENT: 3/18/97

SAMPLES / ITEMS REQUIRED: SAMPLE 10'S (400 PACKS PER CASE)

CASES

NEWPORT K.S. 2 KENT III K.S. KENT GL LTS K.S.

NEWPORT 100'S KENT III 100 KENT GL 100

NEWPORT LTS K.S. 1 TRUE K.S. 1

NEWPORT LTS 100 KENT K.S. 1

KENT 100

ITEMS QUANTITY REQUIRED

BASEBALL CAP 1500

WATER BOTTLES 500

SHIP TO:

CUSTOMER SHIPPING NUMBER 198-1160006

NYO ONLY:

DATE FORWARDED TO PROMOTION SERVICES 2/1/97

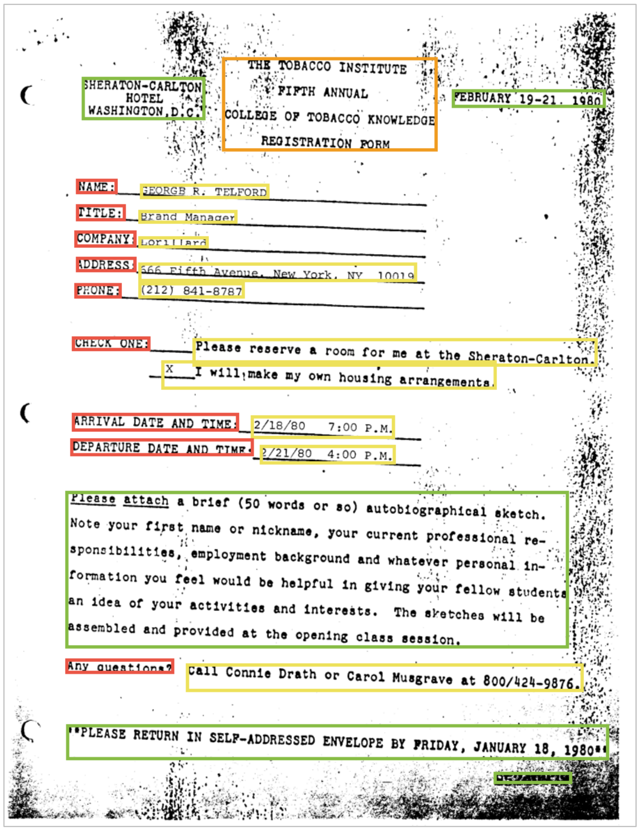
* PLEASE ALLOW 6 WEEKS FOR PROCESSING OF YOUR REQUEST.

82254765

REQFORM 01/17/97

entity: TO::class: <question>
entity: FROM::class: <question>
entity: 1/24/97:class: <answer>
entity: 2:class: <answer>
entity: 1:class: <answer>
entity: 1:class: <answer
entity: 1:class: <answer>
entity: ITEMS:class: <question >
entity: zbulan:class: <answer>
entity: 82254765:class: <other>
entity: REQFORM:class: <other>
entity: 1500:class: <answer>
entity: 500:class: <answer>
entity: K. A. SPARROW:class: <answer>
entity: DATE TO NYO::class: <question>
entity: S. Reindel:class: <answer>
entity: Nassau/ 107:class: <answer>
entity: DIV. NAME/ NO::class: <question>
entity: 1997 SPECIAL EVENT REQUEST FORM:class: <header>
entity: NAME OF EVENT::class: <question>
entity: \"DATE OF EVENT::class: <question>
entity: 3/18/97:class: <answer>
entity: H. Levinson Tradeshow:class: <answer>
entity: SAMPLE 10'S (400 PACKS PER CASE):class: <answer>
entity: SAMPLES/ ITEMS REQUIRED::class: <question>
entity: NEWPORT K. S.:class: <question>
entity: NEWPORT 100's:class: <question>
entity: NEWPORT LTS K. S.:class: <question>
entity: NEWPORT LTS. 100:class: <question>
entity: KENT III K. S.:class: <question>
entity: # CASES:class: <header>
entity: KENT GL LTS K. S.:class: <question>
entity: GL 100:class: <question>
entity: KENT III 100:class: <question>
entity: QUANTITY REQUIRED:class: <question>
entity: BASEBALL CAP:class: <answer>
entity: WATER BOTTLES:class: <answer >
entity: SHIP TO::class: <header>
entity: CUSTOMER SHIPPING NUMBER:class: <question>
entity: 198- 1160006:class: <answer>
entity: NYO ONLY::class: <header>
entity: DATE FORWARDED TO PROMOTION SERVICES::class: <question>
entity: PLEASE ALLOW 6 WEEKS FOR PROCESSING OF YOUR REQUEST:class: <other>

Figure 9: **The results of Entity labelling on the sample 82254765 in FUNSD dataset**: Left image shows GT and the right list is the prediction made by DocFormerv2 which correctly match with GT. Best viewed in color. **Red** boxes are question entities, **Yellow** boxes are answer entities, **Orange** boxes are header entities, and **Green** boxes are other entities.



SHERATON-CARLTON
HOTEL
WASHINGTON,D.C.

THE TOBACCO INSTITUTE
FIFTH ANNUAL
COLLEGE OF TOBACCO KNOWLEDGE
REGISTRATION FORM

FEBRUARY 19-21, 1980

NAME: GEORGE R. TELFORD
TITLE: Brand Manager
COMPANY: Lorillard
ADDRESS: 666 Fifth Avenue, New York, NY 10019
PHONE: (212) 841-8787

CHECK ONE: ___ Please reserve a room for me at the Sheraton-Carlton.
X I will make my own housing arrangements.

ARRIVAL DATE AND TIME: 2/18/80 7:00 P.M.
DEPARTURE DATE AND TIME: 2/21/80 4:00 P.M.

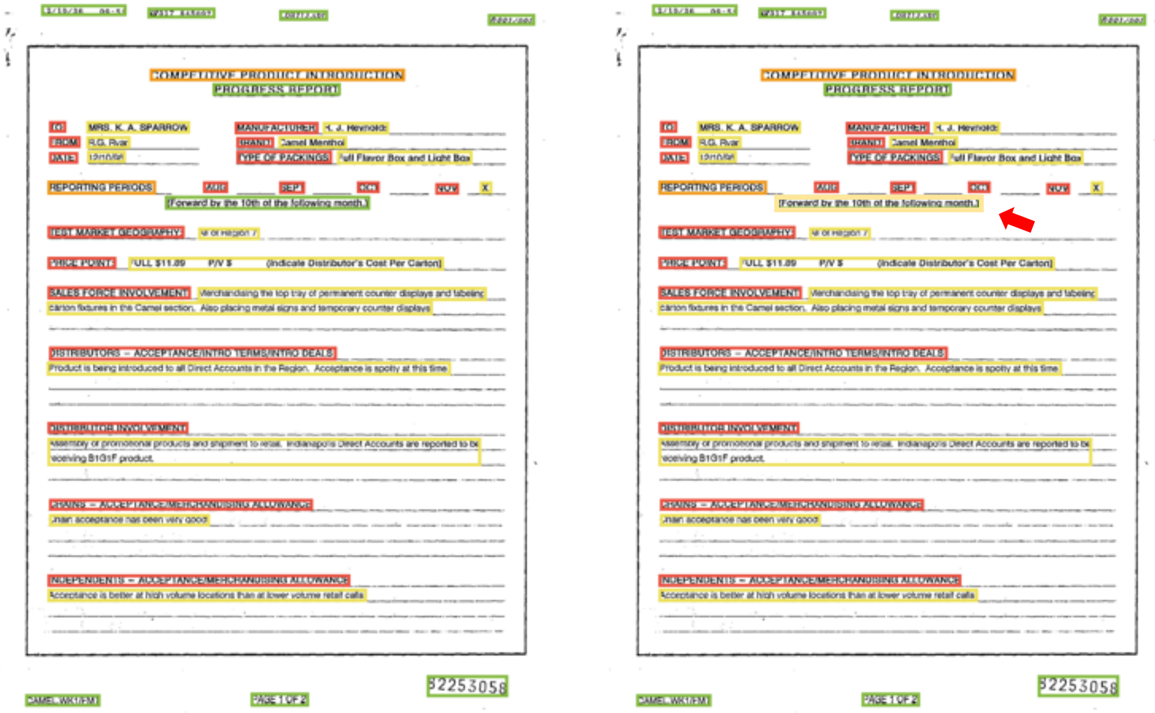
Please attach a brief (50 words or so) autobiographical sketch. Note your first name or nickname, your current professional responsibilities, employment background and whatever personal information you feel would be helpful in giving your fellow students an idea of your activities and interests. The sketches will be assembled and provided at the opening class session.

Any questions? Call Connie Drath or Carol Musgrave at 800/424-9876.

PLEASE RETURN IN SELF-ADDRESSED ENVELOPE BY FRIDAY, JANUARY 18, 1980

entity: NAME::class: <question>
entity: TITLE::class: <question>
entity: COMPANY::class: <question>
entity: ADDRESS::class: <question>
entity: PHONE::class: <question
entity: Lorillard:class: <answer>
entity: SHERATON- CARLTON HOTEL WASHINGTON, D. C.:class: <other>
entity: THE TOBACCO INSTITUTE FIFTH ANNUAL COLLEGE TOBACCO KNOWLEDGE REGISTRATION FORM:class: <header>
entity: FEBRUARY 19-21 1980:class: <other>
entity: GEORGE R. TELFORD:class: <answer>
entity: Brand Manager:class: <answer>
entity: 666 Fifth Avenue, New York, NY 10019:class: <answer>
entity: (212) 841- 8787:class: <answer>
entity: CHECK ONE::class: <question>
"entity": "Please reserve a room for me at the Sheraton- Carlton",entity: Please reserve a room for me at the Sheraton- Carlton:class: <answer>
entity: X I will my own housing arrangements.:class: <answer> entity: ARRIVAL DATE AND TIME::class: <question>
entity: 2/18/80 7:00 P. M.:class: <answer>
entity: DEPARTURE DATE AND TIME::class: <question> entity: 2/21/80 4:00 P. M.:class: <answer>
entity: Please attach a brief (50 words or so) autobiographical sketch. Note your first name or nickname, your current professional responsibilities, employment background and whatever personal in-formation you feel would be helpful in giving your fellow students an idea of your activities and interests. The sketches will be assembled and provided at the opening class session.:class:"<other>"
entity: Any questions?:class: <question>
entity: Call Connie Drath or Carol Musgrave at 800/424- 9876.:class: <answer>
entity: **PLEASE RETURN IN SELF- ADDRESSED ENVELOPE BY FRIDAY, JANUARY 18, 1980**:class: <other>

Figure 10: **The results of Entity labelling on the sample 85240939 in FUNSD dataset**: Left image shows GT and the right list is the predictions made by DocFormerv2 which perfectly matches with GT. Best viewed in color. **Red** boxes are question entities, **Yellow** boxes are answer entities, **Orange** boxes are header entities, and **Green** boxes are other entities.



COMPETITIVE PRODUCT INTRODUCTION
PROGRESS REPORT

TO: MRS. K. A. SPARROW
FROM: R.G. Rvar
DATE: 12/10/98

MANUFACTURER: R. J. Reynolds
BRAND: Camel Menthol
TYPE OF PACKINGS: Full Flavor Box and Light Box

REPORTING PERIODS: AUG ____ SEPT ____ OCT ____ NOV X
(Forward by the 10th of the following month.)

TEST MARKET GEOGRAPHY: All of Region 7

PRICE POINT: FULL $11.89 P/V $ (Indicate Distributor's Cost Per Carton)

SALES FORCE INVOLVEMENT: Merchandising the top tray of permanent counter displays and labeling carton fixtures in the Camel section. Also placing metal signs and temporary counter displays

DISTRIBUTORS – ACCEPTANCE/INTRO TERMS/INTRO DEALS
Product is being introduced to all Direct Accounts in the Region. Acceptance is spotty at this time

DISTRIBUTOR INVOLVEMENT
Assembly of promotional products and shipment to retail. Indianapolis Direct Accounts are reported to be receiving B1G1F product.

CHAINS – ACCEPTANCE/MERCHANDISING ALLOWANCE
Chain acceptance has been very good

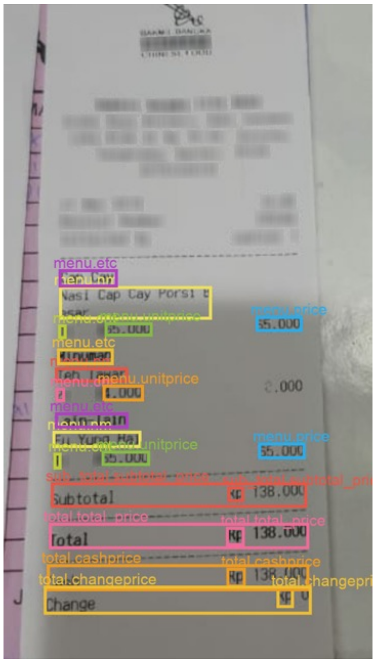
INDEPENDENTS – ACCEPTANCE/MERCHANDISING ALLOWANCE
Acceptance is better at high volume locations than at lower volume retail calls

CAMEL WK1/FM1 PAGE 1 OF 2 82253058

(a) GT Entity labelling

(b) Entity labelling predicted by DocFormerv2

Figure 11: **The failure case of entity labelling on FUNSD dataset by DocFormerv2 .** Column(a) shows the ground-truth entity labels, and column (b) is the entities labelling prediction by DocFormerv2 . **Red** boxes are question entities, Yellow boxes are answer entities, **Orange** boxes are header entities, and **Green** boxes are other entities. DocFormerv2 wrongly labels "(Forward by the 10th of the following month.)" as class "Answer", but its groundtruth class is "Other".



(a) GT Entity extraction

Question: What are entities of class: <*menu.nm*> ?
Answers: [Nasi Cap Cay Porsi B esar; **Cap Cay**;
Teh Tawar: Minuman; Fu Yung Hai; **Lain-lain**]
GTs: [Nasi Cap Cay Porsi B esar; Teh Tawar; Fu Yung Hai]

Question: What are entities of class: <*menu.cnt*>*?*
Answers: [1; 2; 1]

Question: What are entities of class: <*menu.unitprice*>*?*
Answers: [65.000; 4.000; 65.000]

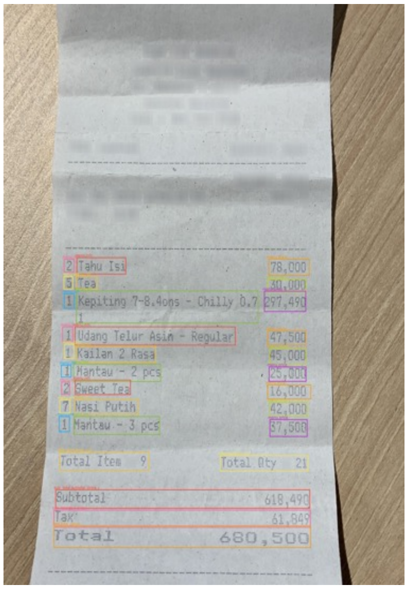
Question: What are entities of class: <*menu. price*>*?*
Answers: [65.000**;** 65.000]

Question: What are entities of class: <*menu. etc*>*?*
Answers : [**Rp: Cash 138.000**]
GTs : [Cap Cay: Minuman: Lain-lain]

Question: What are entities of class: < *sub_total.subtotal_price* >*?*
Answers: [Rp: Subtotal 138.000]
Question: What are entities of class: <*total.total_price*>*?*
Answers: [Rp: Total 138.000]
Question: What are entities of class: <*total.cash_price*>*?*
Answers: [Rp: Cash 138.000]

(b) Entity extraction by DocFormerv2

Figure 12: **DocFormerv2 Partially correct predictions on CORD dataset.** Column(a) shows the Ground-truth entities with the corresponding labels, and column (b) is the entities prediction results by DocFormerv2 . The entities highlighted in Red are wrongly extracted by DocFormerv2 (The corresponding GT entities are highlighted in green, and blacks entities are correctly extracted). `Cap Cay` and `Lain-lain` are wrongly recognized as `menu.name`. `Rp: Cash 138,000` is wrongly recognized as `menu.etc`. It may be because class `menu.etc` has smaller training samples. Best if viewed digitally and in color.



Question: What are entities of class: <*menu.nm*> ?
Answers: [Tahu Isi; Tea; Kepiting 7-8.4ons - Chilly 0.7 1;
Udang Telur Asin – Regular;
Kailan 2 Rasa: Mantau - 2 pcs;
Sweet Tea; Nasi Putih; Mantau - 3 pcs]

Question: What are entities of class: <*menu.cnt*>*?*
Answers: [2; 5; 1; 1; 1; 1; 2; 7; 1]

Question: What are entities of class: <*menu.price*>*?*
Answers: [78,000; 30,000; 297,490; 47,500;
45,000; 25,000; 16,000; 42,000; 37,500]

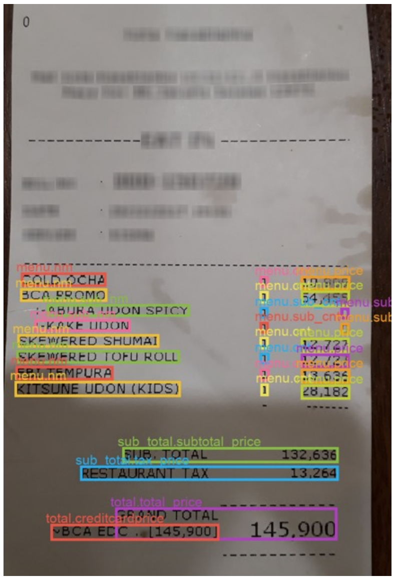
Question: What are entities of class: < *sub_total.subtotal_price* >*?*
Answers: [Subtotal 618,490]

Question: What are entities of class: <*sub_total.tax_price*>*?*
Answers: [Tax 61,849]

Question: What are entities of class: <*total.total_price*>*?*
Answers: [Total 680,500]

Question: What are entities of class: <*total.menutype_cnt*>*?*
Answers: [Total Item 9]

Question: What are entities of class: <*total.menuqty_cnt*>*?*
Answers: [Total Qty 21]



Question: What are entities of class: <*total.creditcardprice*>*?*
Answers: [BCA EDC [145,900]]

Question: What are entities of class: <*menu.nm*> ?
Answers: [COLD OCHA; BCA PROMO; SKEWERED SHUMAI;
SKEWERED TOFU ROLL; EBI TEMPURA; KITSUNE UDON (KIDS)]

Question: What are entities of class: <*menu.cnt*>*?*
Answers: [1; 1; 1; 1; 1; 1]

Question: What are entities of class: <*menu.price*>*?*
Answers: [10,909; 54,455; 12,727; 12,727; 13,636; 28,182]

Question: What are entities of class: < *menu.sub_nm*>*?*
Answers: [ABURA UDON SPICY: ; -KAKE UDON]

Question: What are entities of class: <*menu.sub_cnt*>*?*
Answers: [1; 1]

Question: What are entities of class: < *menu.sub_price*>*?*
Answers: [0; 0]

Question: What are entities of class: < *sub_total.subtotal_price* >*?*
Answers: [SUB, TOTAL 132,636]

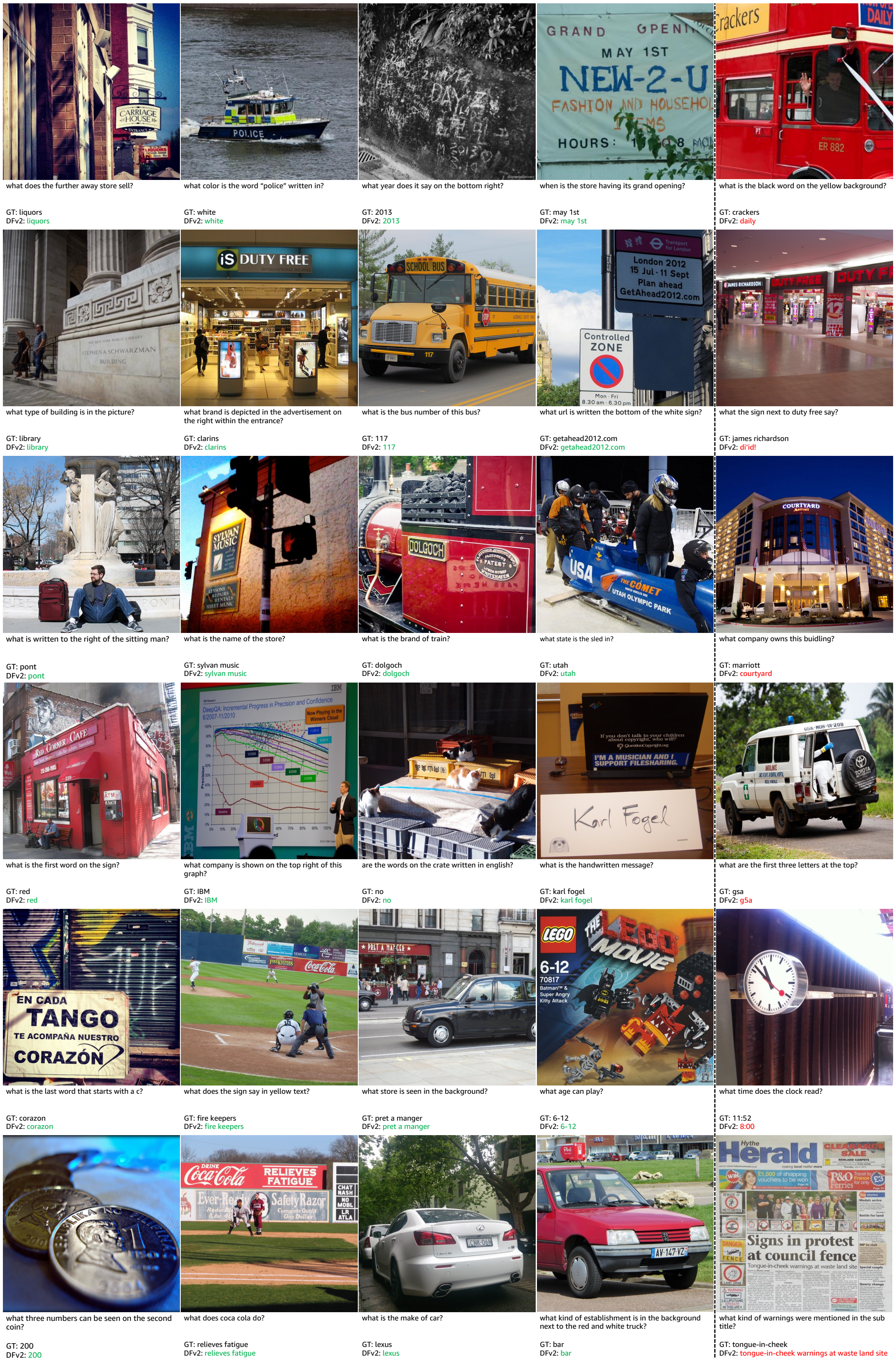
Question: What are entities of class: <*sub_total.tax_price*>*?*
Answers: [RESTAURANT TAX 13,264]

Question: What are entities of class: <*total.total_price*>*?*
Answers: [GRAND TOTAL 145,900]

(a) GT Entity extraction

(b) Entity extraction by DocFormerv2

Figure 13: **Entity extraction comparisons on CORD receipt dataset.** Column(a) shows the Groundtruth entities with the corresponding labels, and column (b) is the entities prediction results by DocFormerv2 . Note that the questions with empty values are not shown. DocFormerv2 predicted correctly all the entity regions in the image. Best if viewed digitally and in color.



what does the further away store sell?
GT: liquors
DFv2: liquors

what color is the word "police" written in?
GT: white
DFv2: white

what year does it say on the bottom right?
GT: 2013
DFv2: 2013

when is the store having its grand opening?
GT: may 1st
DFv2: may 1st

what is the black word on the yellow background?
GT: crackers
DFv2: daily

what type of building is in the picture?
GT: library
DFv2: library

what brand is depicted in the advertisement on the right within the entrance?
GT: clarins
DFv2: clarins

what is the bus number of this bus?
GT: 117
DFv2: 117

what url is written the bottom of the white sign?
GT: getahead2012.com
DFv2: getahead2012.com

what the sign next to duty free say?
GT: james richardson
DFv2: di'id!

what is written to the right of the sitting man?
GT: pont
DFv2: pont

what is the name of the store?
GT: sylvan music
DFv2: sylvan music

what is the brand of train?
GT: dolgoch
DFv2: dolgoch

what state is the sled in?
GT: utah
DFv2: utah

what company owns this buidling?
GT: marriott
DFv2: courtyard

what is the first word on the sign?
GT: red
DFv2: red

what company is shown on the top right of this graph?
GT: IBM
DFv2: IBM

are the words on the crate written in english?
GT: no
DFv2: no

what is the handwritten message?
GT: karl fogel
DFv2: karl fogel

what are the first three letters at the top?
GT: gsa
DFv2: g5a

what is the last word that starts with a c?
GT: corazon
DFv2: corazon

what does the sign say in yellow text?
GT: fire keepers
DFv2: fire keepers

what store is seen in the background?
GT: pret a manger
DFv2: pret a manger

what age can play?
GT: 6-12
DFv2: 6-12

what time does the clock read?
GT: 11:52
DFv2: 8:00

what three numbers can be seen on the second coin?
GT: 200
DFv2: 200

what does coca cola do?
GT: relieves fatigue
DFv2: relieves fatigue

what is the make of car?
GT: lexus
DFv2: lexus

what kind of establishment is in the background next to the red and white truck?
GT: bar
DFv2: bar

what kind of warnings were mentioned in the sub title?
GT: tongue-in-cheek
DFv2: tongue-in-cheek warnings at waste land site

Figure 14: **DocFormerv2 (DFv2) predictions on Scene-Text VQA**: The first four columns show examples of correct predictions from DocFormerv2. The last column shows examples of failure cases of DocFormerv2.

References

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 993–1003.

Appalaraju, S.; Tang, P.; Dong, Q.; Sankaran, N.; Zhou, Y.; and Manmatha, R. 2023. DocFormerv2: Local Features for Document Understanding - Full Paper and Supplemental. *arXiv preprint arXiv:2306.01733*.

Biten, A. F.; Litman, R.; Xie, Y.; Appalaraju, S.; and Manmatha, R. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16548–16558.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; and Karatzas, D. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4291–4301.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; LI, S.; Zhou, X.; and Wang, W. Y. 2019. TabFact: A Large-scale Dataset for Table-based Fact Verification. *ArXiv*, abs/1909.02164.

Guillaume Jaume, J.-P. T., Hazim Kemal Ekenel. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. In *Accepted to ICDAR-OST*.

Herzig, J.; Nowak, P. K.; Müller, T.; Piccinno, F.; and Eisenschlos, J. M. 2020. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

Ho, C.-H.; Appalaraju, S.; Jasani, B.; Manmatha, R.; and Vasconcelos, N. 2022. YORO - Lightweight End to End Visual Grounding. In *ECCV Workshops*.

Hong, T.; Kim, D.; Ji, M.; Hwang, W.; Nam, D.; and Park, S. 2020. BROS: A Pre-trained Language Model for Understanding Texts in Document. *under review https://openreview.net/references/pdf?id=uCz3OR6CJT*.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv preprint arXiv:2204.08387*.

Jaume, G.; Ekenel, H. K.; and Thiran, J.-P. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2: 1–6.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; Wei, F.; and Guo, B. 2021. Swin Transformer V2: Scaling Up Capacity and Resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11999–12009.

Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1697–1706.

Mathew, M.; Karatzas, D.; Manmatha, R.; and Jawahar, C. V. 2020. DocVQA: A Dataset for VQA on Document Images. arXiv:2007.00398.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Pasupat, P.; and Liang, P. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Annual Meeting of the Association for Computational Linguistics*.

Powalski, R.; Borchmann, Ł.; Jurkiewicz, D.; Dwojak, T.; Pietruszka, M.; and Pałka, G. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, 732–747.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Seunghyun, P.; Seung, S.; Bado, L.; Junyeop, L.; Jaeheung, S.; Minjoon, S.; and Hwalsuk, L. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Tang, P.; Appalaraju, S.; Manmatha, R.; Xie, Y.; and Mahadevan, V. 2023a. Multiple-Question Multiple-Answer Text-VQA. *arXiv preprint arXiv:2311.08622*.

Tang, P.; Zhu, P.; Li, T.; Appalaraju, S.; Mahadevan, V.; and Manmatha, R. 2023b. DEED: Dynamic Early Exit on Decoder for Accelerating Encoder-Decoder Transformer Models. *arXiv preprint arXiv:2311.08623*.

Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.-Y.; and Bansal, M. 2022. Unifying Vision, Text, and Layout for Universal Document Processing. *ArXiv*, abs/2212.02623.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020a. 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*, 1192–1200.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. 2020b. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *arXiv preprint arXiv:2012.14740*.

Łukasz Borchmann; Pietruszka, M.; Stanisławek, T.; Jurkiewicz, D.; Turski, M. P.; Szyndler, K.; and Gralinski, F. 2021. DUE: End-to-End Document Understanding Benchmark. In *NeurIPS Datasets and Benchmarks*.