



Overview

- We release the OCR annotations for 26M document pages.
- Text** annotations are extracted with commercial Amazon Textract OCR.
- Industry documents are very varied: ranging from 1900 to 2019 and different type of documents: letters, memorandums, reports, forms, emails, etc.
- More importantly, it will allow researchers to fairly compare different pre-training strategies using the same OCR annotations.

Document intelligence datasets

Dataset	# Docs.	# Pages	Documents source	Document description	OCR-Text	OCR-BB	Layout	Document type
IIT-CDIP	6.5M	35.5M	UCSF-LTD	Industry documents	Unknown	✗	✗	✓
RVL-CDIP	-	400K	UCSF-LTD	Industry documents	✗	✗	✗	✓
PublayNet	-	364K	PubMedCentral	Journals and articles	✗	✗	✓	✗
DocBank	-	500K	arXiv	Journals and articles	✗	✗	✓	✗
DocVQA	6K	12K	UCSF-IDL	Industry documents	Microsoft OCR	✓	✗	✓
OCR-IDL	4.6M	26M	UCSF-IDL	Industry documents	Amazon Textract	✓	✗	✓

IDL Success Cases

LayoutLM [1] Pre-training Data	FUNSD Precision	FUNSD Recall	FUNSD F1
IDL-500K	0.6217	0.705	0.6607
IDL-1M	0.6545	0.737	0.6933
IDL-2M	0.6938	0.759	0.7249
IDL-11M	0.7597	0.8155	0.7866

LaTr [2] Pre-training Data	TextVQA Accuracy
✗	50.37
ST-VQA Datasets + OCR-CC	54.22
IDL-1M	55.12
IDL-11M	56.28
IDL-64M	58.03
ST-VQA Datasets + OCR-CC + IDL-64M	58.51

OCR-IDL Statistics

