

Unified Pretraining Framework for Document Understanding

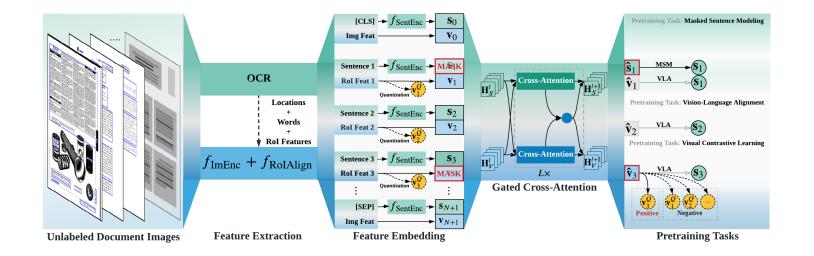
Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, Tong Sun

> Adobe Research, Adobe Document Cloud





Problem and Contribution



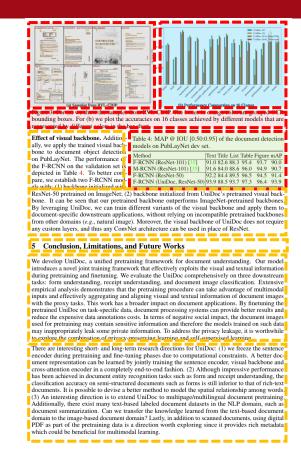
- 1. We introduce *UniDoc*, a powerful pretraining *framework* for document understanding. *UniDoc* is capable of learning contextual *textual and visual* information and *cross-modal correlations* within a *single framework*, which leads to better performance.
- 2. We present Masked Sentence Modeling for language modeling, Visual Contrastive Learning for vision modeling, and Vision-Language Alignment for pretraining.
- 3. Extensive experiments and analysis provide useful insights on the effectiveness of the pretraining tasks and show outstanding performance on various downstream tasks.

Motivation

1. Documents are composed of semantic regions

UniDoc: Unified Pretraining Framework for Document Understanding				
ſ	Juxiang Gu ¹ , Jason Kuen ¹ , Vlad I. Morariu ¹ , Handong Zhao ¹ , Nikolose Barnpalio ² , Kajiv Jain ¹ , Ani Nenkova ¹ , Tong Sun ¹ ¹ Adobe Research. ² Adobe Document Cloud { jigu, Juen, morariu, hazhao, harmpal 1, raji jain, nenkova, tsun)@adobe.com			
	<pre>{jigu,kuen,morariu,naznao,oarmpaii,rajijain,nenkova,tsun/wadobe.com</pre>			
	Abstract Document intelligence automates the extraction of information from documents and supports many business applications. Recent self-supervised learning methods on large-scale unlabeled document datasets have opened up promising directions towards reducing annotation efforts by training models with self-supervised ob- jectives. However, most of the existing document pretraining methods are still language-dominated. We present UnliDoc, a new unified pretraining framework for document understanding. UniDoc is designed to support most document under- standing tasks, extending the Transformer to take multimodal embeddings as input. Each input element is composed of words and visual features from a semantic re- gion of the input document image. An important feature of UniDice is that it learns a generic representation to model sentences, learn similarities, and align modalities. Extensive empirical analysis demonstrates that the pretraining procedure learns better joint representations and leads to improvements in downstream tasks.			
	introduction			
under docu usual pape conte chall almo semi- word of B for d	mean timeling encer is no not research arise that metudies extensingues for information extraction at studing. Unlike planitext documents in natural language processing (NLP)[1, 2, 3], a physics nent can be composed of multiple elements: tables, figures, charts, <i>etc.</i> In addition, a document y includes rich visual information, and can be one of various types of documents (scientifi , form, resume, <i>etc.</i>), with various combinations of multiple elements and layouts. Complet real layout, noisy data, form and style variations make automatic document understanding ver- enging. For example, to understand text-rich documents such as letters, a system meeds to force structured documents such as forms requires the system to analyze spatially distributed sho- p, paying particular attention to the spatial arrangement of the words. Following the success IT [4] on NLP tasks, there has been growing interest in developing pretraining method cument understanding [5, 6, 7, 8]. Pretrained models have achieved state-of-the-art (SoTA mance across diverse document understanding [8]. [0].			
howe comp docu hiera form the si trans	training databases help pretraining anoders with the current pretraining stugs. (1) documents are ver, we observe three major problems with the current pretraining setup; (1) documents are sole of semantic regions. Most of the recent document pretraining works follow BERT and spl ments into words. However, unlike the sequence-to-sequence learning in NLP, documents hav chical structure (words form sentences, sentences form a semantic region, and semantic region a document). Also, the importance of words and sentences are highly context-dependent, i.e. we word or sentence may have different importance in a different context. Moreover, currer former-based document pretraining models suffer from input length constraints. Also, input hecemes are bised on the documents or multi-page documents. (2) documents are			

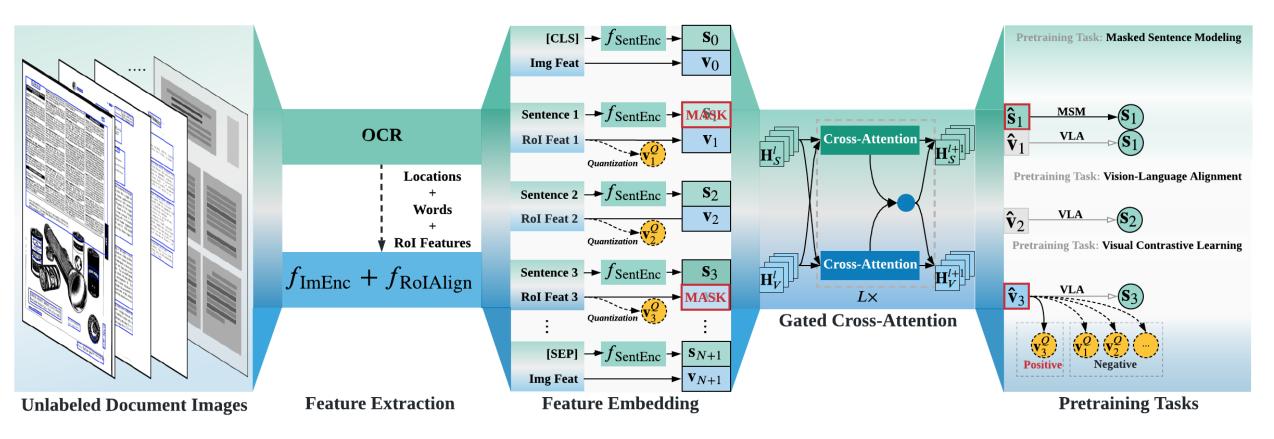
2. Documents are more than words



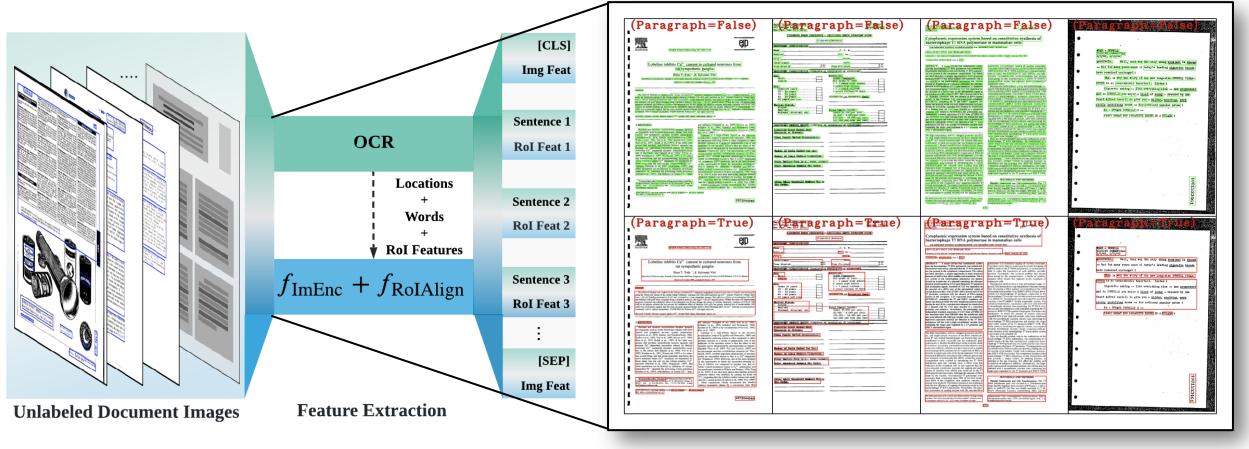
3. Documents have spatial layout

lagure 3: Fo	r (a) we sho xes. For (b)	w the samples i) we plot the ac	from RVL-CDIP. The curacies on 16 class	boxes in orange es achieved by d	color are grouped lifferent models the	OCR at are
T						
		ne. Addition- d visual back-	Table 4: MAP @ IC models on PubLay		f the document dete	ction
n PubLayN e F-RCNN picted in	let. The per l on the val Table 4. To	ect detection rformance of idation set is better com- -RCNN mod-	lethod -RCNN (ResNet-10 I-RCNN (ResNet-10 -RCNN (ResNet-50 RCNN (Jeiper B	1) [35] 91.0 82 01) [35] 91.6 84 0 92.2 84	tle List Table Figure .6 88.3 95.4 93.7 .0 88.6 96.0 94.9 .4 89.5 96.5 94.5 5 93.7 97.3 96.4	90.0 90.7 91.4
any custom l	oyom, and t	_	let architecture con-	e vool in plane		quire
ang centored 5 Conch Ve develop	usion, Lir UniDoc, a	nitations, an	nd Future Work	o vood in plane 5 document unde	of Des ^M et.	nodel
Concle Concle de develop introduces a uring pretra cysks: form inpirical ar	usion, Lir UniDoc, a novel joint t understandi alysis demo	nitations, an unified pretra raining framew inetuning. We ing, receipt un onstrates that t	nd Future Work ining framework for ork that effectively e evaluate the UniDoo derstanding, and doo he pretraining proce	s document und xploits the visual comprehensive cument image c dure can take ac	of DevNet. erstanding. Our m I and textual inform Iy on three downst lassification. Exte dvantage of multin	nodel ation ream nsive nodal
Conclu- Conclu- ve develop introduces a curing pretra exists: form inpirical arr puts and e with the pro- with the pro- retrained Ur retrained Ur reduce the ex-	united and a second	nitations, au unified pretra raining framew inetuning. We ing, receipt un onstrates that t gregating and is work has a l sk-specific data a annotations c	nd Future Work ining framework for ork that effectively e evaluate the UniDoc derstanding, and do he pretraining proce aligning visual and broader impact on d i, document processi osts. In terms of neg	document und source the source of the comprehensive current image c dure can take ac extual informati focument applica g systems can p trive social impa	erstanding. Our n and textual inform ly on three downst lassification. Exte dvantage of multin on of document in tions. By finetunin rovide better result ct, the document in	nodel ation ream nsive nodal nages g the s and nages
5 Conclu- ter develop introduces a uring pretra- tesks: form inputs and e with the pro- retrained U reduce the ep- duce the ep- duce the ep- duce the ep- tesed for pretrained a gapping and a second test of the second test of test of test of test test of test of test of test test of test of test of test of test of test of test test of test of test of test of test of test of test of test test of test of	unidation of the second state of the second st	nitations, an unified pretra raining framew inetuning. We ing, receipt un onstrates that t gregating and his work has a is work has a is work chas a is work in a contain sensiti k some private	nd Future Work ining Iramework for orok that effectively e vealuate the UniDoc derstanding, and do the pretraining procee aligning visual and 1 broader impact on d i, document processis osts. In terms of neg tve information and t	document und xploits the visual comprehensive zument image c dure can take ac extual informati cument applica g systems can p ative social impa herefore the moc ess the privacy 1	erstanding. Our n and textual inform Jy on three downst lassification. Exte Vantage of multin on of document in tions. By finetunin tovide better result ct, the document in lels trained on such	nodel ation ream nodal nages g the s and nages a data
a sectore of the s	yers, end t unDoc, a novel joint ti nining and fi understandi alaysis deme ffectively ag ty tasks. The niDoc on ta typensive dat raining may priately lease to be an eresting sho	nitations, an unified pretra raining framew inetuning. We gregating and gregating and sk-specific data a anotations c contain sensiti k some private Go of CWC)	nd Future Work ming framework for ork that effectively e valuate the UniDoo derstanding, and do he pretraining proce aligning visual and the order impact on do the output of the output down of the output set. In terms of nega- tion of the output set of the output of the end of the output of the	document und sources to the sisual comprehensive dure can take aa extual informati cument applica g systems can p trive social impa herefore the more ress the privacy 1 in the supernois s for UniDoc: (1	of DesNet erstanding. Our n and textual inform ly on three downst lassification. Exte Vantage of multin tions. By finetunin rovide better result els trained on such eakage, it is worth 20 aarting	nodel aation ream nsive nodal nages g the s and nages a data while tence
a sector of the sector of t	yers, and t asion, Lir UnDoc, a novel joint t understandi alysis demu fectively ag ty tasks. Th niDoc on ta: pensive dat raining may priately leal teresting sho ng pretrain sentation ca on encoder	unified pretra raining framew inetuning. We ing, receipt un onstrates that t gregating and sis work has a l sk-specific data a annotations c contain sensiti k some private contain sensiti k some private contain sensiti net and long-tee ing and fine-tu n be learned b	nd Future Workk ming transwork for ork that effectively e evaluate the UniDood derstanding, and dood derstanding proce- aligning visual and i , document processis ors. In terms of negr ve information and the solution of the solution of the solution of the solution of the m research direction m research direction of pointly training the point of fashior	s used in places document und comprehensive comprehensive unent image c dure can take as extual informati scument applica g systems can p trive social impa herefore the moc ess the privacy l for UniDoc: (1) of or UniDoc: (2) of or UniDoc; (2) of though (2) (2) Although (2)	of DesNet, erstanding. Our n and textual inform ly on three downst lassification. Exte Vantage of multin on of document in tions. By finetunin rovide better result to better result test trained on such eachage, it is worth erozen the sen eakage, it is worth erozen the sen onstraints. A better er, visual backbon mpressive perform	nodel ation ream nsive nodal nages g the s and nages a data while tence c doc c e and
Conclusion of the second se	yers and t usion, Lir UmDoc, a novel joint L alysis dem adjysis de	mitations, an united pretra raning framew inctuning. We up, receipt un onstrates that gregating and its work has a sk-specific data a anotations can contain sensiti s come private contain sensiti s contain sensiti s c	nd Future Workk mining framework for ork that effectively e evaluate the UniDoo aligning visual and broader impact on du aligning visual and broader impact on du aligning visual and broader impact on a document processis ots. In terms of nege information. To addu vi information. To addu vi information. To addu vi information. To addu vi information. The addu vi information. The addu vi information and the information and t	s und implements document und sploits the visual comprehensive cument image c dure can take as extual information g systems can g g systems can g s for UniDoc; (1) omputational cc sentence encode s for UniDoc; (1) omputational cc s sentence encode s of thiologia f a s form and re- s forms is still in el the spatial rel	of DesNet erstanding. Our n and textual inform by on three downst lassification. Exte vantage of multin on of document in toroide better result els trained on such eakage, it is worth els trained on such eakage, it is worth eakage, it is worth cell and the such eakage is a such	nodel aation nsive aodal aages g the s and aages a data while tence c doc c e and aance g, the
5 Concle Ve develop introduces a cring pretra i puts and e i puts and e with the pro- ved for pretra ved for	yrm and t uniDoc, a novel joint L understandia alysis demo fectively ag yr tasks. Th niDoc on ta reining may priately leal or ensuing the eresting sho ng pretraini- sentation ca on encoder ieved in dox accuracy on It is possibl sting direce t, there exist	mitations, ar united pretra- ratining framework inctuning. We ingregating and its work has a gregating and its work has a sumotations ex- contain sensiti s some private and some private work of the sense of the sense of the sense of the sense of the sense of the sense of the sense of the sense ing and fine-tunity in a completel cument entity in a completel cument entity in a seni-structure to devise a b tion is to extend the nit set sense the sense of the sense sense of the sense sense of the sense sense of the sense the sense of the sense sense of the sense of the sense sense of the sense of the sense of the sense sense of the sense of the sense of the sense sense of the sense of the sense sense of the sense	nd Future Workk iming transwork foor ork that effectively e evaluate the UniDoor derstanding, and doo he pretraining proce- aligning visual and i do the pretraining process or solar the solar solar to a solar the solar solar or solar the solar solar m research direction m research direction m pipeling solar sola	document under document under ploits the visual comprehensive comprehensive comprehensive comprehensive as extual informati- comprehensive set state privacy the visual compational of set the privacy the visual set the vi	of DesNet, erstanding. Our n and textual inform Jy on three downst lassification. Extu Vantage of multin on of document in toos. By finetunin rovide better result clt be document in lels trained on suct clt actual backbon onstraints. A better er, visual backbon morstaints. A better er, visual backbon morstaints. A better er, visual backbon periorito that of rici ationship among w document pretra e NLP domain, su	node aation ream nsive noda nage: a data while tence r doc e and aato aato aato aato aato aato aato aat

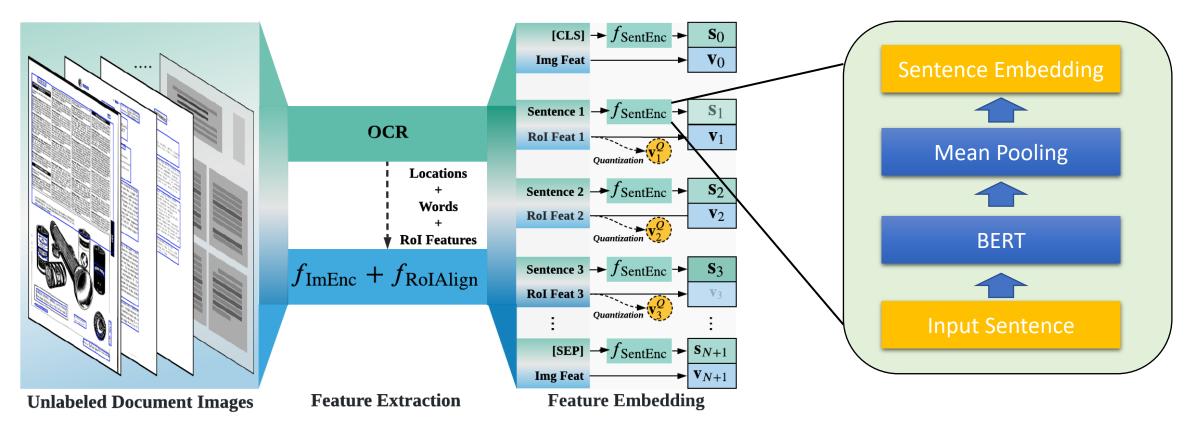
Framework



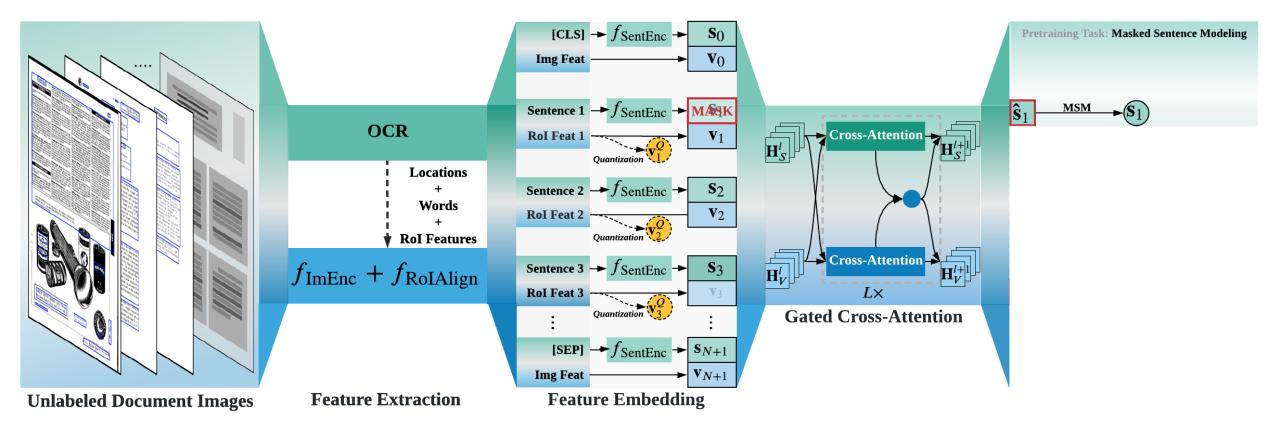
Feature Extraction



Feature Embedding

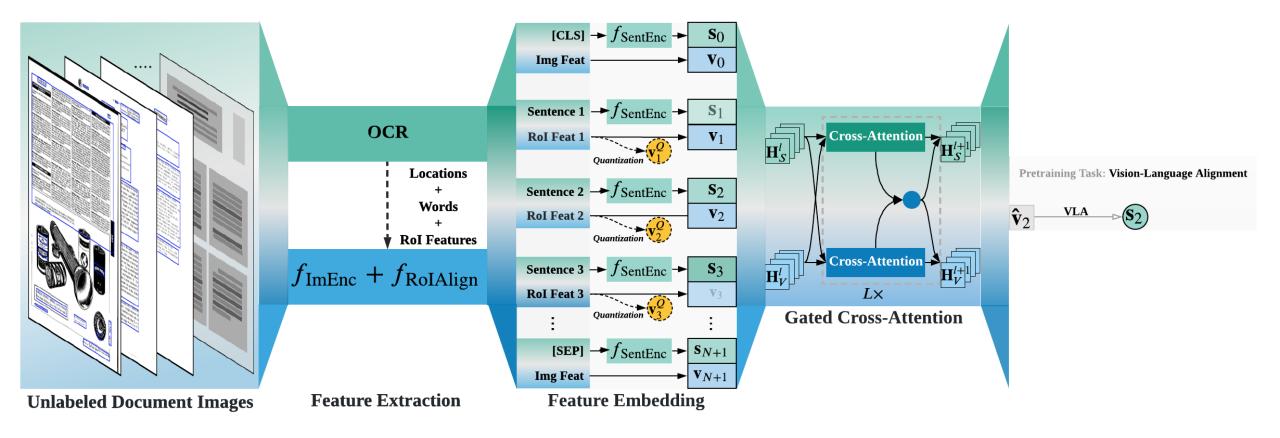


Pretraining Task: Masked Sentence Modeling



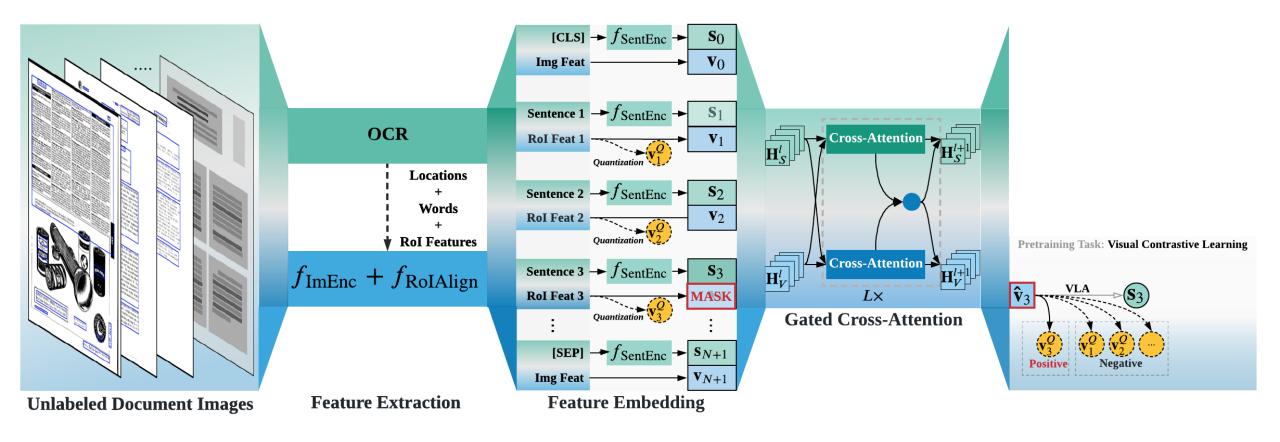
$$\mathcal{L}_{\text{MSM}}(\Theta) = \sum_{i} \text{smooth}_{L_1}(\boldsymbol{s}_i - f_{\text{UniDoc}}(\boldsymbol{s}_i | \boldsymbol{s}_{\setminus i}, \tilde{\mathbf{V}}))$$

Pretraining Task: Vision-Language Alignment



$$\mathcal{L}_{\text{VLA}}(\Theta) = \frac{1}{N \times N} ||f_{\text{Norm}}(\mathbf{S} \cdot \mathbf{S}^{\top}) - f_{\text{Norm}}(\mathbf{H}_{V}^{L} \cdot \mathbf{H}_{V}^{L\top})||_{F}^{2}$$

Pretraining Task: Visual Contrastive Learning



$$\mathcal{L}_{\text{VCL}}(\Theta) = -\sum_{\tilde{\boldsymbol{v}}_i \in \tilde{\mathbf{V}}} \left(\log \frac{\exp(\sin(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i^Q) / \kappa)}{\sum_{\boldsymbol{v}_j^Q} \exp(\sin(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_j^Q) / \kappa)} \right) + \lambda \frac{1}{CE} \sum_{c=1}^C \sum_{e=1}^E p_{c,e} \log p_{c,e}$$

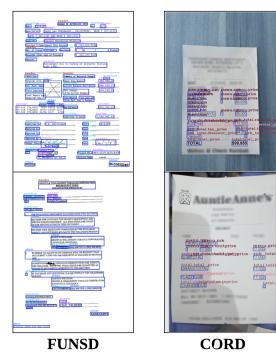
Finetuning Task: Document Classification

	scientific publicat		memo
	EVENTS: The second sec		
form	presentation	email	resume
UNIVERSITY OF RENTUCKY		2073735620	CONSCIENCTION VENT
Image: Section of the sectio	MARKING IN DESCRIPTION DESCRIPTION	<text><text><section-header><list-item><list-item><section-header><section-header></section-header></section-header></list-item></list-item></section-header></text></text>	

Samples from RVL-CDIP

				-		_	-	-	
Method	Pretraining					FUNSD	CORI	RVL-CDI	
Wielilou	Source	#Data	Scale	Max #Words	Modality	#Param.	F 1	F 1	Accuracy
BERT _{BASE} [5]	—	_	Word	512	L	110M	60.26	89.68	89.81
BERT _{LARGE} [5]	_	_	Word	512	L	340M	65.63	90.25	89.92
LayoutLM _{BASE} [5]	IIT-CDIP	11 M	Word	512	L	113M	78.66	94.72	94.42
LayoutLM _{LARGE} [5]	IIT-CDIP	11M	Word	512	L	343M	78.95	94.93	94.43
LayoutLMv2 _{BASE} [5]	IIT-CDIP	11 M	Word	512	V+L	200M	82.76	94.95	95.25
LayoutLMv2 _{LARGE} [5]	IIT-CDIP	11M	Word	512	V+L	426M	84.20	96.01	95.64
SelfDoc [6]	RVL-CDIP	320K	Region	50×512	V+L	-	83.36	-	92.81
SelfDoc+VGG-16 [6]	RVL-CDIP	320K	Region	50×512	V+L	-	_	_	93.81
TILT-Base [34]	RVL-CDIP+	1.1M	Word	512	V+L	230M	_	95.11	95.25
TILT-Large [34]	RVL-CDIP+	1.1M	Word	512	V+L	780M	_	96.33	95.52
UDoc	IIT-CDIP	1M	Region	64×512	V+L	272M	87.96	98.85	93.96
UDoc*	IIT-CDIP	1M	Region	64×512	V+L	272M	87.93	98.94	95.05 [‡]

Finetuning Task: Document Entity Recognition



Pretraining FUNSD CORD RVL-CDIP Method #Data Scale |Max #Words|Modality|#Param. **F**1 F1 Source Accuracy BERT_{BASE} [5] Word 512 60.26 89.68 110M 89.81 _ L 512 Word 340M 65.63 90.25 89.92 BERT_{LARGE} [5] L _ LayoutLM_{BASE} [5] **IIT-CDIP** 512 113M 78.66 94.72 94.42 11**M** Word L **IIT-CDIP** 512 343M 78.95 94.43 LayoutLM_{LARGE} [5] 11**M** Word 94.93 L 200M 95.25 LayoutLMv2_{BASE} [5] **IIT-CDIP** 512 V+L 82.76 11**M** Word 94.95 LayoutLMv2_{LARGE} [5] 95.64 **IIT-CDIP** 11**M** 512 V+L 426M 84.20 96.01 Word SelfDoc [6] RVL-CDIP 320K Region 50×512 V+L 92.81 83.36 _ _ SelfDoc+VGG-16 [6] **RVL-CDIP** 320K Region 50×512 V+L 93.81 _ _ TILT-Base [34] RVL-CDIP+ 512 95.25 V+L 230M 1.1M Word 95.11 _ 512 780M 95.52 TILT-Large [34] RVL-CDIP+ 1.1M Word V+L 96.33 _ 93.96 UDoc **IIT-CDIP** 1M Region 64×512 V+L 272M 87.96 98.85 UDoc* 272M IIT-CDIP 64×512 V+L 87.93 98.94 95.05[‡] 1M Region

Adobe®

At the Poster:



Additional details

- Quantitative results
 - Discussion