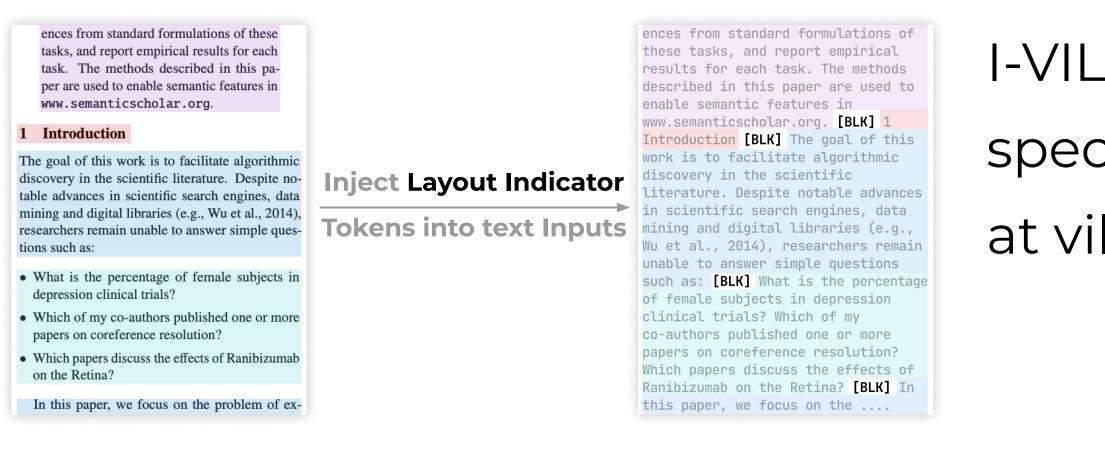


Title

We aim to extract structured data from paper PDFs. Key to the process is classifying token semantic categories. The PDF text strings are not NLP model friendly.

Section

I-VILA Model



I-VILA leads to Better Macro F1 across datasets

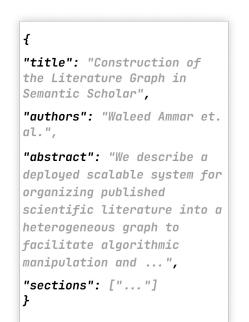
Dataset	LayoutLM	Layout using Text Lines
GROTOAP2	92.34	92.37(+0.03%)
DocBank	91.06	92.79 (+1.90%)
S2-VL	82.69	83.77 (+1.31%)

I-VILA works for different models w/o extra pre-training

Base Model	Fine-tune only	Fine-tune w using Text Lines
BERT	90.78	91.65 (+0.96%)
RoBERTa	91.64	92.04 (+0.44%)
LayoutLM	92.34	92.37 (+0.03%)

VILA: Improving Structured Content Extraction from Scientific PDFs Using VIsual LAyout Groups

shannons@allenai.org | github.com/allenai/vila



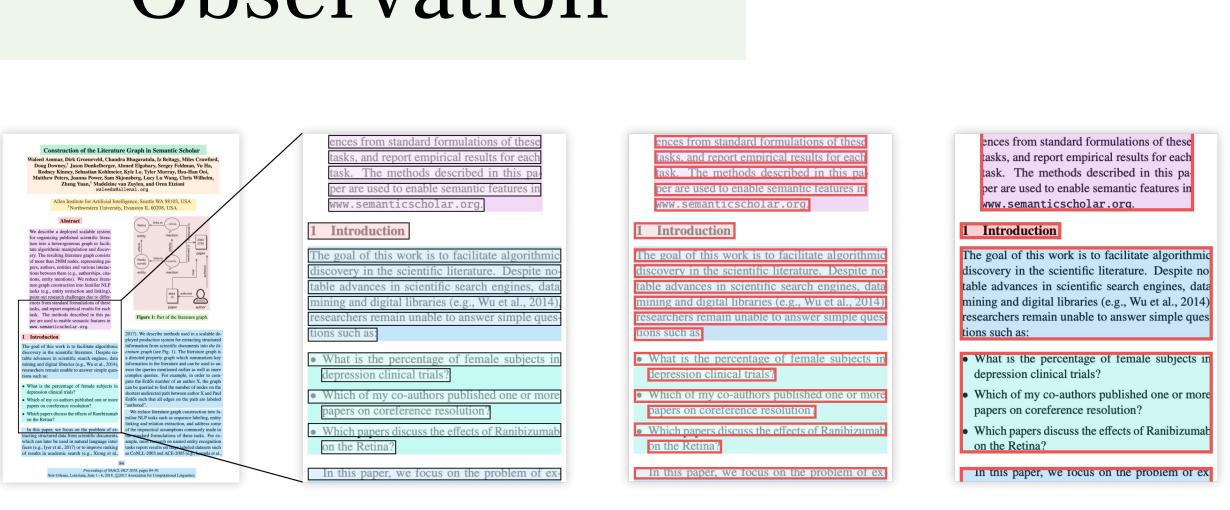
Output / Paper Metadata JSON

I-VILA injects a special token [BLK] at vila boundaries

LayoutLM + I-VILA using Text Blocks 93.38 (+1.13%) 92.00 (+1.03%) 83.44 (+0.91%)

> with I-VILA using Text Blocks 92.31 **(+1.69%)** 92.52 **(+0.96%)** 93.38 **(+1.13%)**

Observation



Tokens in the same VILA group usually have the same category -- Token Category Uniformity Assumption

Section H-VILA Model

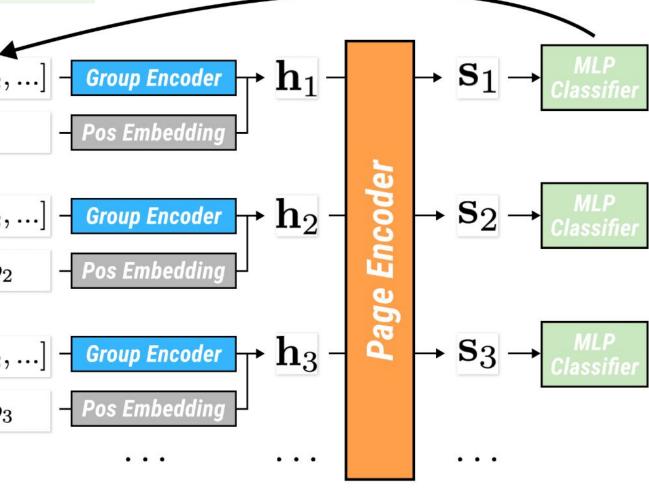
1 Introduction	$_{oxdot}$ Text $[\mathbf{t}_{1,1},\mathbf{t}_{1,2},.$
The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite no-table advances in scientific search engines, data	Box Coordinates
mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple ques- tions such as:	$-$ Text $[{f t}_{2,1},{f t}_{2,2},.$
• What is the percentage of female subjects in depression clinical trials?	${}^{ar{}}$ Box Coordinates b_2
 Which of my co-authors published one or more papers on coreference resolution? 	$\sub{\textit{Text}}[\mathbf{t}_{3,1},\mathbf{t}_{3,2},.]$
 Which papers discuss the effects of Ranibizumab on the Retina? 	Box Coordinates b_3

H-VILA is a hierarchical model that encodes VILA: It encodes the textual information in each group individually, then model the groups as a sequence. The classifier predicts the group category, which is assigned to all containing tokens as the token class.

H-VIL reduces almost 50% Inference time vs LayoutLM

Model Name	Macro F1	Inference Time (ms)
BERT	87.24	41.59 (-21%)
LayoutLM	91.06	52.56
LayoutLM + I-VILA	92.79	56.31 (+7%)
LayoutLM + H-VILA	91.27	28.07 (-47%)
LayoutLMv2	93.33	99.19 (+89%)

Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, Doug Downey



No extra pre-training is needed to achieve performance gains -- saving up to 95% computational cost.

Section

Dataset Nam Total Sample Annotatio Metho Discipline

Has VILA Groups

PDF Parsin

Vision Mode

human annotation

S2-VLUE is a new benchmark for <u>V</u>isual Layout-enhanced Scientific Document <u>Understanding Evaluation. It augments</u> existing dataset (DocBank) with visual layout groups, and forms a new dataset called S2-VL, with human annotations from 19 disciplines and different types of VILA groups sources.





Our Findings

We develop two models using VILA: I-VILA injects layout indicators and improves *accuracy*.

H-VILA is a hierarchical model encode VILA and has better **efficiency**.

S2-VLUE Benchmark

าย	GROTOAP2	DocBank+	S2-VL
es	119k	500k	1.3k
on od	Automatic	Automatic	Human
es	Life Science	Math / Physics / CS	19 Disciplines
s?			
Ŋ	Yes	No	Yes
lel	No	Yes	Yes
on	No	No	Yes

