Title

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

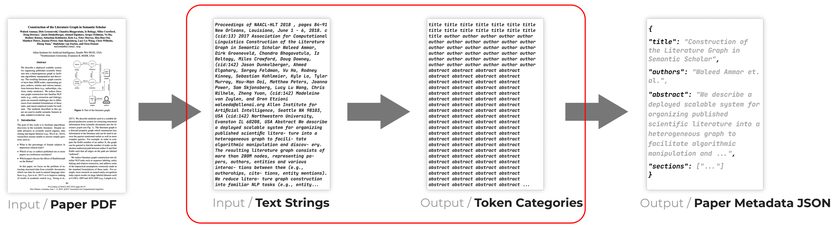
Author

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

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

Section

## The Task




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

## Observation

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

Section

## Our Findings

We develop two models using VILA:

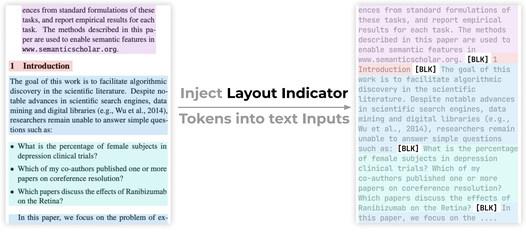
**I-VILA** injects layout indicators and improves ***accuracy***.

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

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

Section

## I-VILA Model




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

### I-VILA leads to Better Macro F1 across datasets

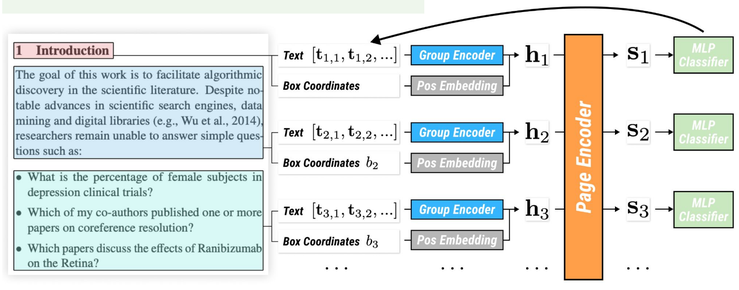
| Dataset | LayoutLM | LayoutLM + I-VILA | |
|---|---|---|---|
| | | *using Text Lines* | *using Text Blocks* |
| GROTOAP2 | 92.34 | 92.37(+0.03% ) | **93.38 (+1.13%)** |
| DocBank | 91.06 | **92.79 (+1.90%)** | 92.00 (+1.03%) |
| S2-VL | 82.69 | **83.77 (+1.31%)** | 83.44 (+0.91%) |

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

| Base Model | Fine-tune only | Fine-tune with I-VILA | |
|---|---|---|---|
| | | *using Text Lines* | *using Text Blocks* |
| BERT | 90.78 | 91.65 (+0.96%) | 92.31 **(+1.69%)** |
| RoBERTa | 91.64 | 92.04 (+0.44%) | 92.52 **(+0.96%)** |
| LayoutLM | 92.34 | 92.37 (+0.03%) | 93.38 **(+1.13%)** |

Section

## H-VILA Model




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%) |

Section

## S2-VLUE Benchmark

| Dataset Name | GROTOAP2 | DocBank+ | S2-VL |
|---|---|---|---|
| Total Samples | 119k | 500k | 1.3k |
| Annotation Method | Automatic | Automatic | Human |
| Disciplines | Life Science | Math / Physics / CS | 19 Disciplines |
| Has VILA Groups? | | | |
| PDF Parsing | Yes | No | Yes |
| Vision Model | No | Yes | Yes |
| human annotation | No | No | Yes |

S2-VLUE is a new benchmark for Visual Layout-enhanced Scientific Document Understanding Evaluation. It augments 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.

arXiv 2106.00676 | Github allenai/vila