KnowVrDU: A Unified Knowledge-aware Prompt-Tuning Framework for Visually-rich Document Understanding

Yunqi Zhang , Yubo Chen , jingzhe zhu , Jinyu Xu , Shuai Yang ,

Zhaoliang Wu , Liang Huang , Yongfeng Huang and Shuai Chen

Department of Electronic Engineering & BNRist, Tsinghua University

Ant Group

Introduction

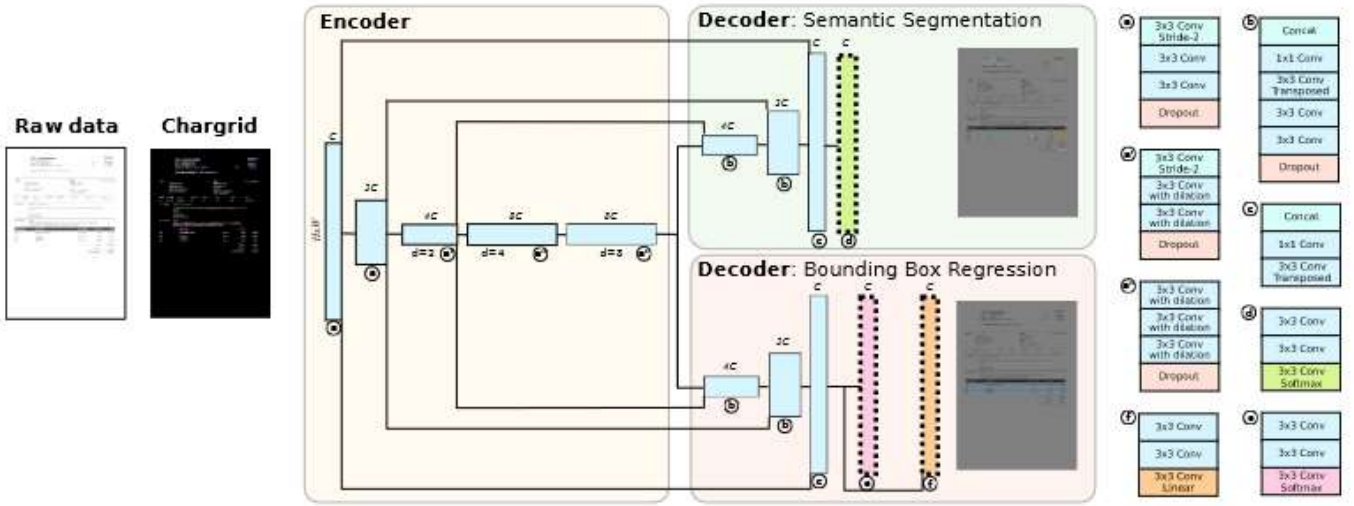
- Visually-rich Document Understanding(VrDU)
  - Automatically analyze and extract significant factual texts from image or digital-born documents
  - Subtasks: document information extraction, document visual question answering and document image classification[1]

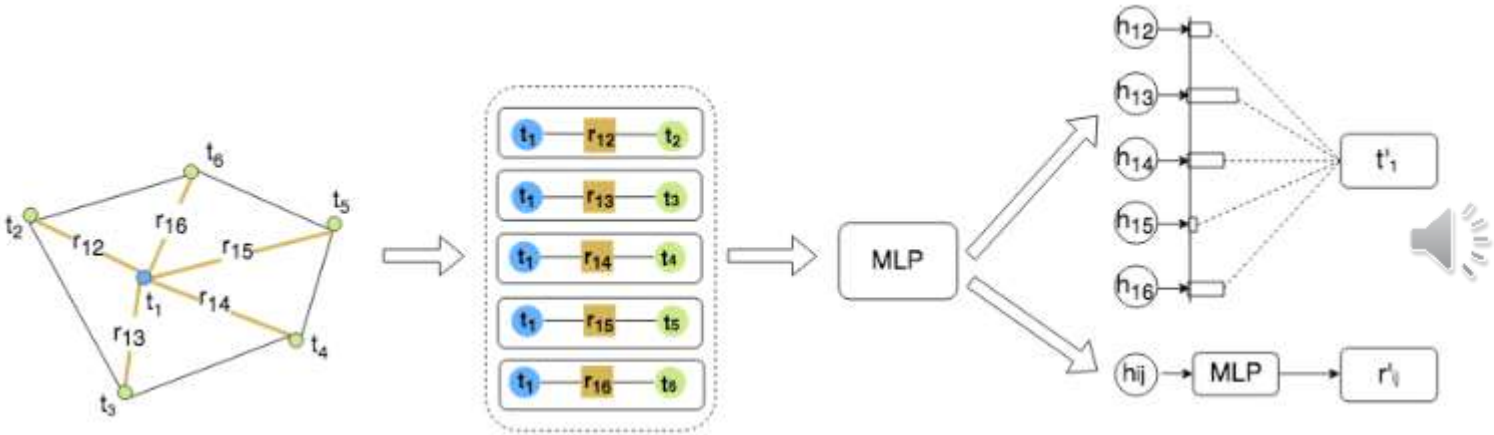
Introduction

- Critical multi-modal task
  - Cross-modal alignment learning
- Recent work
  - Learning the multi-modal interactions between text, image and layout modalities.
  - Neural network-based approaches: designing suitable network architectures
  - Pre-train and fine-tune approaches : designing objectives used at both the pre-train and fine-tune stages

Related Work

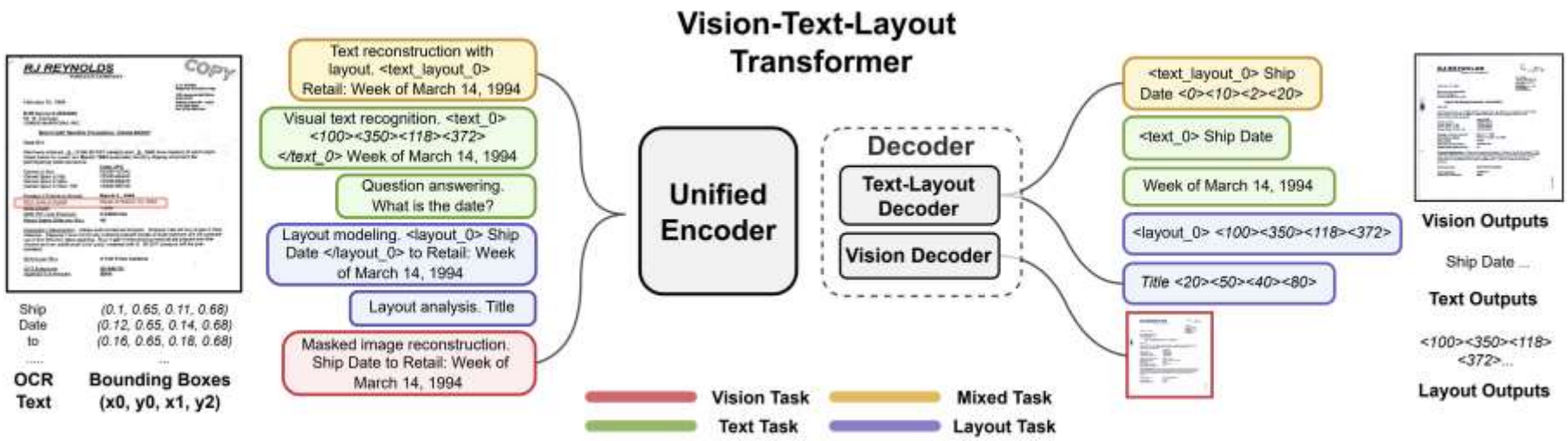
- Neural network-based approaches
  - CharGrid[2], GCN[3]





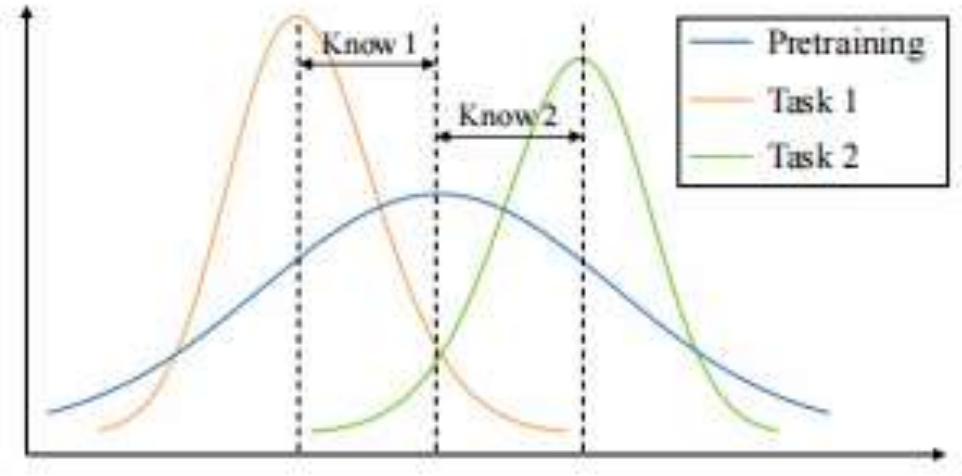
Related Work

- Pre-train and fine-tune approaches
  - Layoutlmv3[4], UDOP[5]



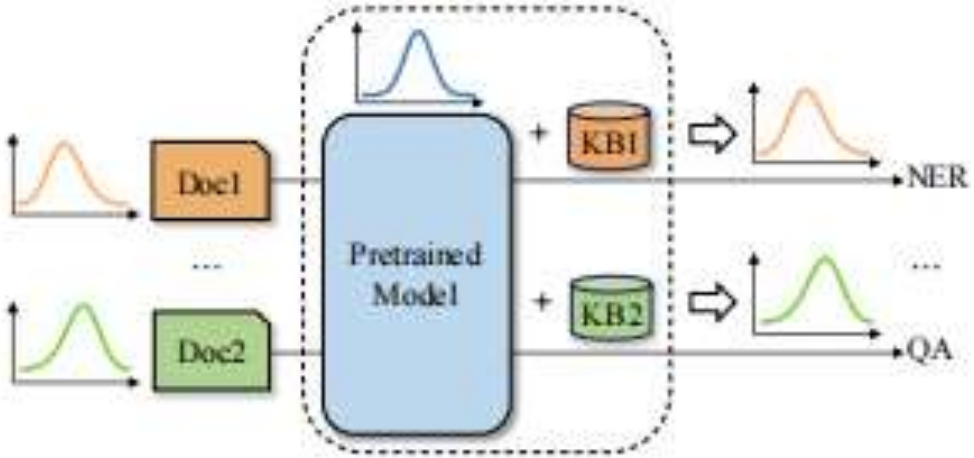
Problems

- Existing methods [6] relied on huge demands of designing dedicated architectures and annotating task-specific samples to fine-tune pretrained models.
- Significant knowledge distribution gaps between the pre-training task and VrDU tasks.
- Limited annotations are insufficient and ineffective in capturing refined task expertise [7].



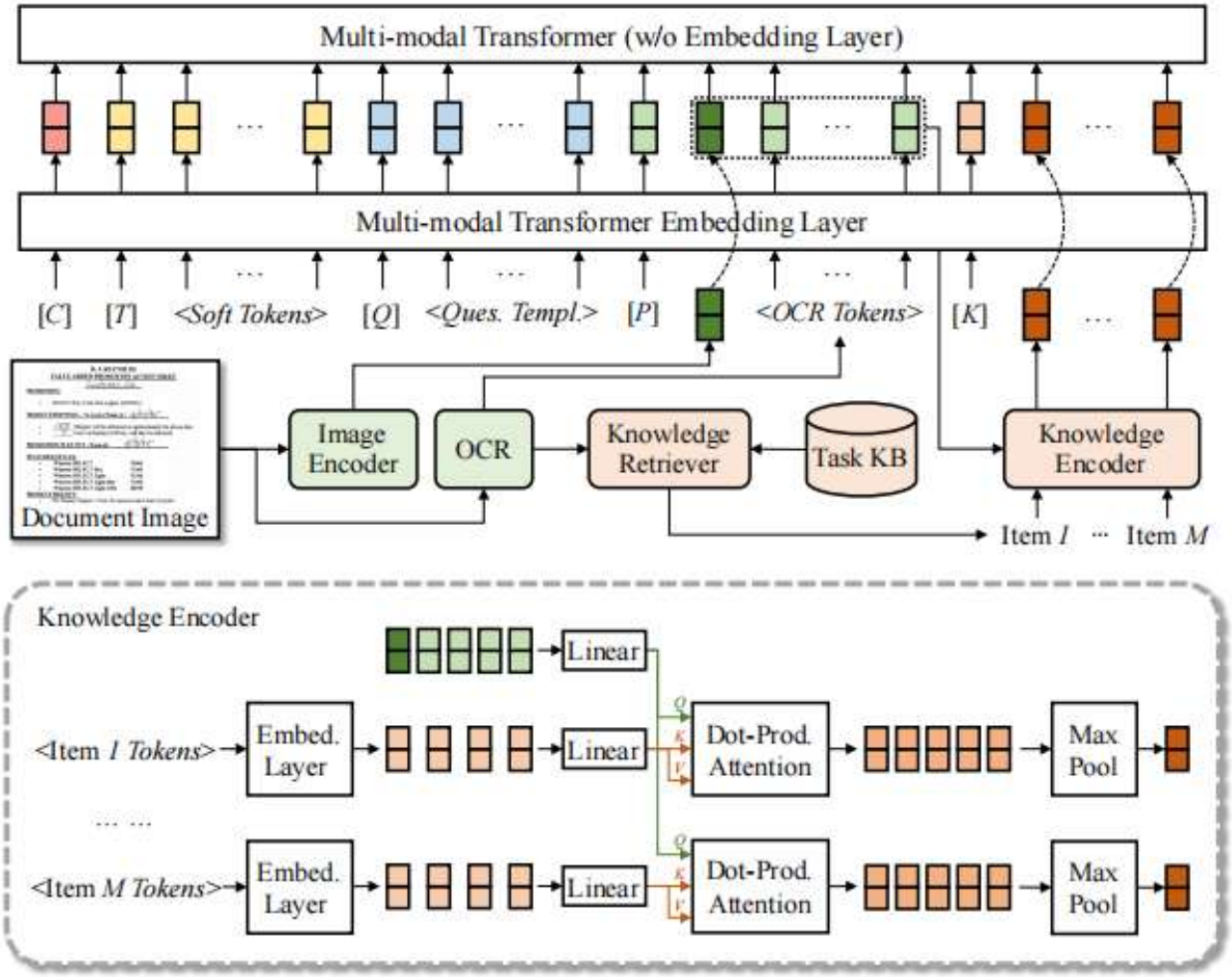
Motivation

- Recent research of natural language processing uniformly model various tasks, including information extraction and document classification, with Question Answering (QA) format.
- The lack of task-specialized knowledge caused by limited labeled data in few-shot scenarios can be compensated with large-scale knowledge bases.



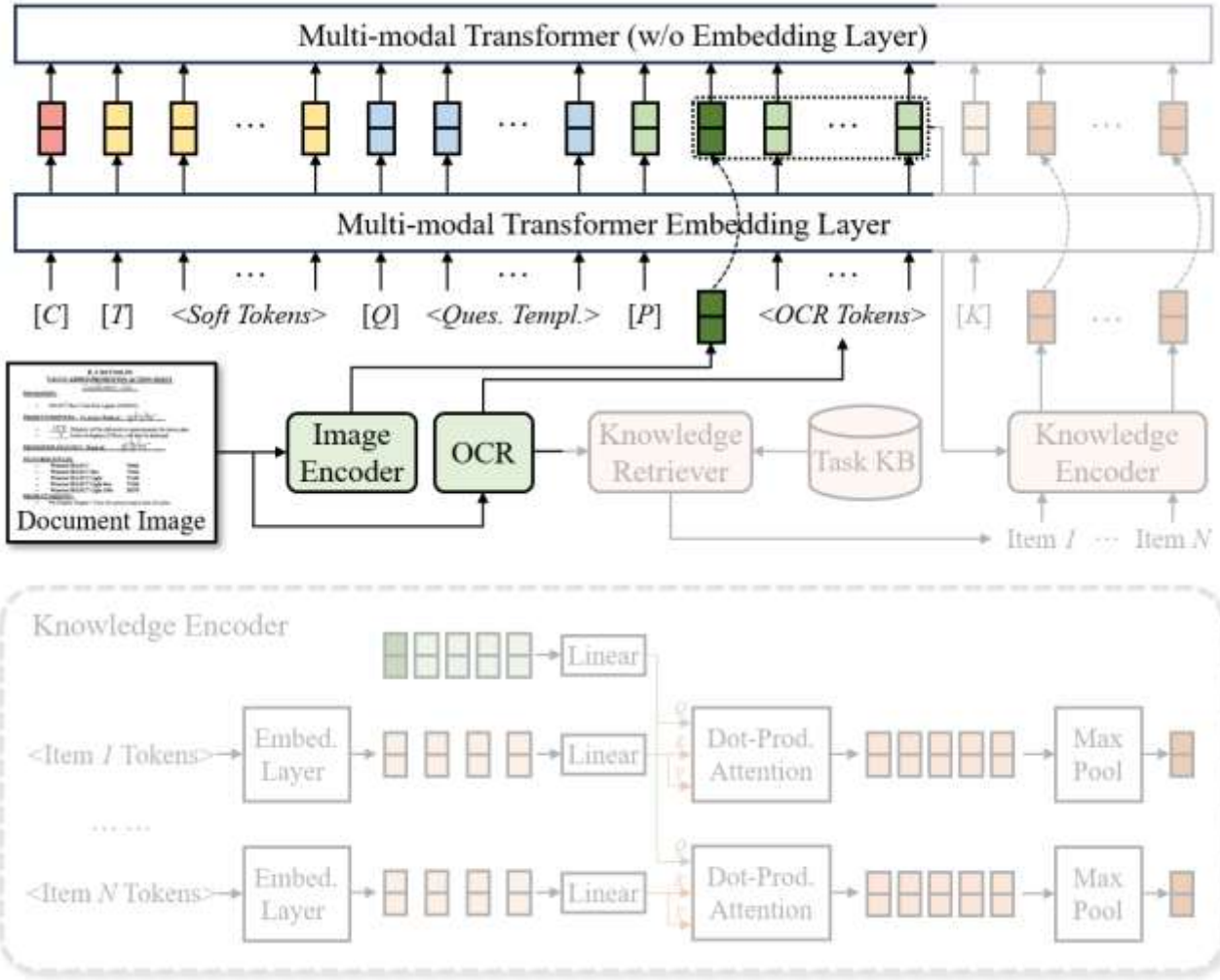
Our Approach

- Overall Framework



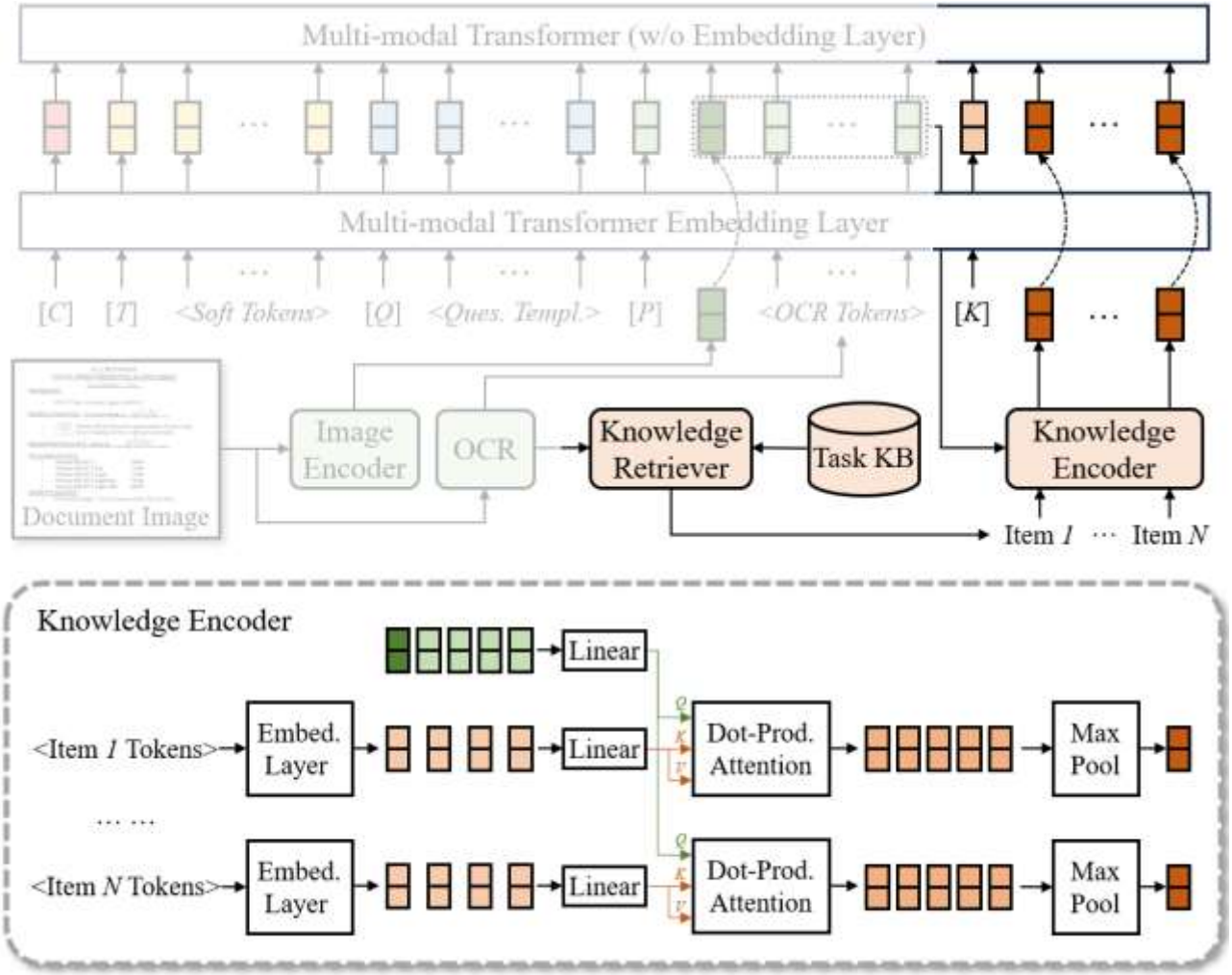
Our Approach

- Uniform Prompts Generation



Our Approach

- External Knowledge Injection



Experiments

- Datasets
  - FUNSD: 199 documents
  - CORD: 1000 documents
  - RVL-CDIP: 400000 documents
  - DocVQA:12767 documents

Dataset	Field	Train	Vaild	Test
FUNSD	4	149	-	50
CORD	30	800	100	100
RVL-CDIP	16	320k	40k	40k
DocVQA	-	10,104	1,286	1,287

- Evaluation
  - Document Information Extraction: Precision, Recall, F_1
  - Document Image Classification: Accuracy
  - Document Question Answering: ANLS

Experiments

- Performance Evaluation on FUNSD and CORD

N	Method	FUNSD			CORD		
		Precision	Recall	F_1	Precision	Recall	F_1
2	RoBERTa	21.64±1.64	33.43±4.24	26.68±1.76	34.96±6.73	45.70±7.17	39.59±7.03
	LASER	30.40±4.89	35.20±7.20	32.36±5.14	-	-	-
	LayoutLMv3$_{BASE}$	44.29±6.14	58.96±7.20	50.43±6.03	47.21±6.25	58.99±4.94	52.41±5.85
	†LAGER$_{BASE}$	49.82±6.06	59.55±8.91	54.09±6.54	48.68±5.72	60.19±4.23	53.79±5.24
	KnowVrDU$_{BASE}$	50.37±5.85	59.61±6.26	54.42±6.13	49.36±5.96	59.84±3.28	54.29±4.75
	KnowVrDU$_{LARGE}$	**54.28±4.77**	**64.09±6.34**	**58.67±5.82**	**52.27±5.31**	**66.15±4.53**	**57.73±4.97**
4	RoBERTa	27.53±2.92	42.83±2.68	33.48±2.83	45.89±7.84	55.04±8.69	50.05±8.25
	LASER	44.91±2.42	50.25±3.26	47.36±2.18	-	-	-
	LayoutLMv3	65.32±3.89	77.97±2.26	71.06±3.04	54.18±5.01	64.92±3.76	59.04±4.53
	†LAGER$_{BASE}$	67.86±3.30	78.73±2.57	72.86±2.69	56.28±4.24	66.47±3.29	60.94±3.86
	KnowVrDU$_{BASE}$	69.20±3.03	78.51±2.64	73.67±2.31	56.69±4.13	68.16±4.05	62.24±5.01
	KnowVrDU$_{LARGE}$	**72.87±4.49**	**79.61±3.11**	**76.15±3.47**	**58.58±3.87**	**72.31±3.84**	**64.79±3.40**
6	RoBERTa	33.75±2.19	47.20±2.54	39.32±2.06	52.88±4.84	61.41±4.86	56.82±4.82
	LASER	48.64±2.14	53.54±2.10	50.96±1.95	-	-	-
	LayoutLMv3	71.19±3.75	80.83±1.09	75.68±2.58	60.91±3.51	69.16±2.76	64.76±3.16
	†LAGER$_{BASE}$	72.71±3.41	81.53±1.98	76.84±2.58	61.80±5.14	70.00±3.75	65.63±4.53
	KnowVrDU$_{BASE}$	72.23±3.26	83.25±1.16	77.31±2.36	63.27±2.76	72.90±4.68	67.59±3.38
	KnowVrDU$_{LARGE}$	**73.49±2.31**	**85.02±2.40**	**78.60±1.77**	**66.43±3.64**	**73.41±2.57**	**69.62±3.05**
8	RoBERTa	37.30±3.55	49.52±4.89	42.52±4.89	57.38±1.86	65.32±1.54	61.08±1.57
	LASER	-	-	-	-	-	-
	LayoutLMv3	74.31±2.19	81.75±2.60	77.85±2.29	64.49±3.24	72.21±2.17	68.12±2.77
	†LAGER$_{BASE}$	76.27±1.44	83.41±1.73	79.66±1.14	64.89±4.38	72.22±3.19	68.35±3.84
	KnowVrDU$_{BASE}$	78.14±1.68	84.65±1.29	81.44±1.42	66.17±2.83	72.99±3.71	69.23±3.16
	KnowVrDU$_{LARGE}$	**81.52±1.78**	**87.23±1.36**	**84.04±1.30**	**70.19±3.49**	**75.41±4.44**	**72.51±3.63**

Experiments

- Performance on Document Image Classification and Document Question Answering tasks

Ratio	Model	RVL-CDIP	DocVQA
1%	LayoutLMv3	29.70	16.84
	KnowVrDU$_{BASE}$	38.53	20.91
	KnowVrDU$_{LARGE}$	**45.31**	**25.42**
5%	LayoutLMv3	52.47	28.03
	KnowVrDU$_{BASE}$	56.90	31.78
	KnowVrDU$_{LARGE}$	**61.27**	**35.23**
10%	LayoutLMv3	71.14	42.11
	KnowVrDU$_{BASE}$	75.70	45.52
	KnowVrDU$_{LARGE}$	**78.22**	**49.67**
20%	LayoutLMv3	80.25	67.46
	KnowVrDU$_{BASE}$	84.03	69.81
	KnowVrDU$_{LARGE}$	**85.89**	**72.43**

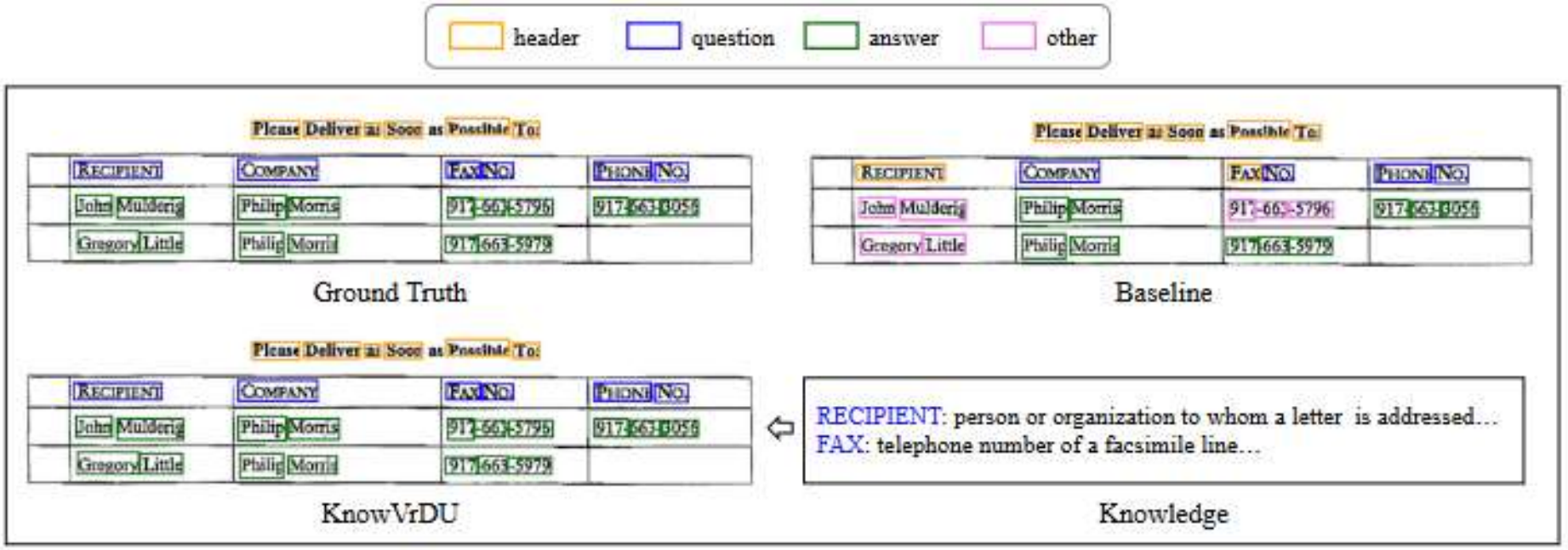
Experiments

- An ablation study of the KnowVrDU model on the CORD dataset.

Method	Prec.	Rec.	F_1
KnowVrDU$_{BASE}$	**78.14**	**84.65**	**81.44**
w/o Attention	76.41	82.27	79.23
w/o Knowledge	74.92	81.66	78.14
w/o Soft Prompts	77.10	83.49	80.16
w/o All	74.31	81.75	77.85

Experiments

- Case Study



Conclusion

- We propose a unified knowledge-aware framework for a wide applications of various downstream VrDU tasks.
- We propose to model heterogeneous VrDU structures through reformulating all tasks into extractive question answering tasks with task-specific prompts.
- We propose to reduce the knowledge gap through integrating external open-source knowledge to incorporate external knowledge bases.
- Experimental results in few-shot settings demonstrate the effectiveness of our method on a wide variety of downstream VrDU tasks.

References

[1] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei.2021. Document AI: benchmarks, models and applications. CoRR, abs/2111.08609.

[2] Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4459–4469.

[3] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019a. Graph convolution for multimodal information extraction from visually rich documents. In Proceedings of NAACL-HLT, pages 32–39.

[4] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

[5] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19254–19264.

[6] Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In ACL, pages 1331–1339.

[7] Zilong Wang and Jingbo Shang. 2022a. Towards few-shot entity recognition in document images: A label-aware sequence-to-sequence framework. In Findings of the Association for Computational Linguistics: ACL 2022, pages 4174–4186.[8] Ningning Jia, Xiang Cheng, and Sen Su. 2020. Improving knowledge graph embedding using locally and globally attentive relation paths. In European Conference on Information Retrieval, pages 17–32. Springer.

[9] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In ACL, pages 1476–1488.

Thanks for Listening!