UNIT: Unifying Image and Text Recognition in One Vision Encoder

Yi Zhu[1], Yanpeng Zhou[1], Chunwei Wang[1], Yang Cao[2], Jianhua Han[1], Lu Hou[1], Hang Xu[1].

[1]Huawei Noah's Ark Lab, [2]Hong Kong University of Science and Technology

Project page: https://github.com/yeezhu/UNIT

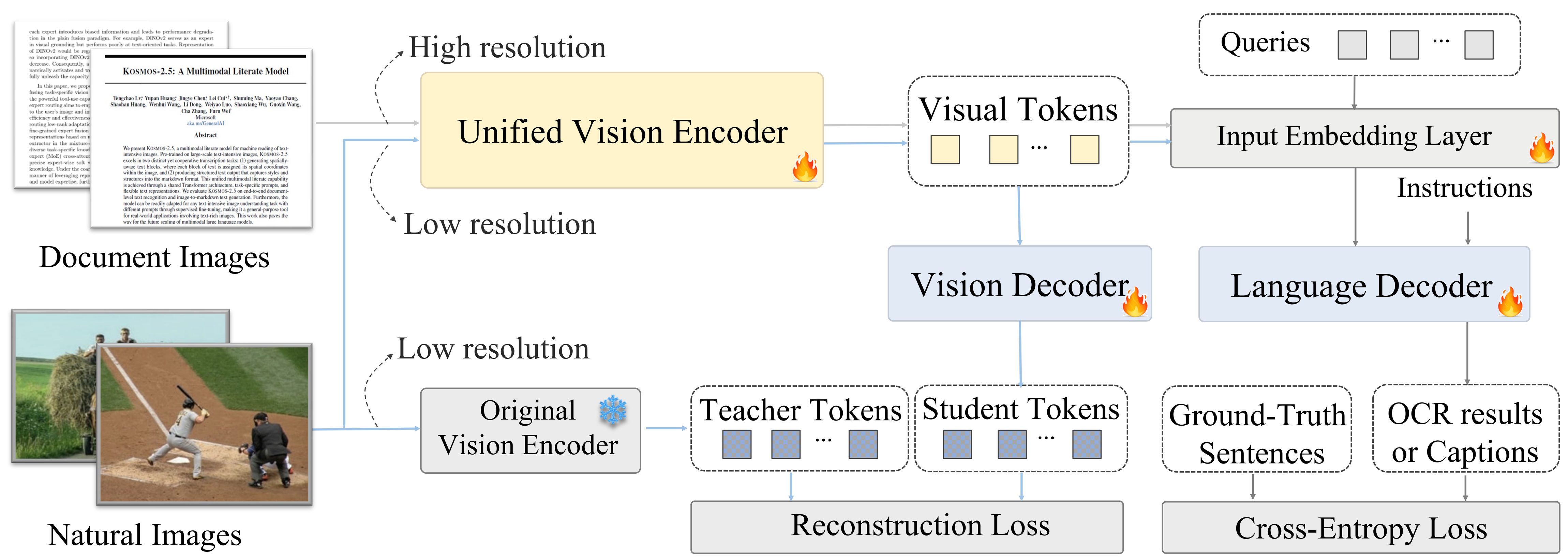
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

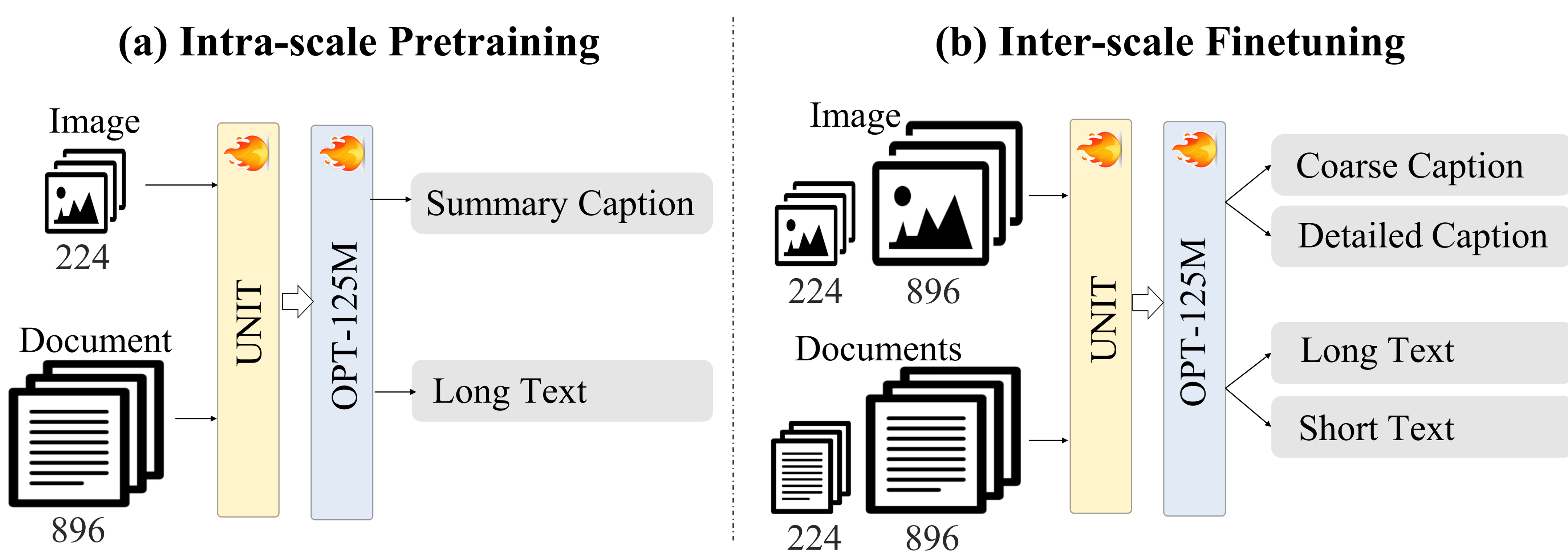
- Currently, vision encoder models like Vision Transformers (ViTs) typically excel at image recognition tasks but cannot simultaneously support text recognition like human visual recognition.
- Image recognition typically involves global feature extraction, while text recognition demands precise, localized feature extraction.
- We propose **UNIT**, a framework aimed at UNifying Image and Text recognition within a single model.

Method

- UNIT builds upon existing Vision Transformer models and integrates a lightweight **language decoder** for text prediction, alongside a small **vision decoder** to preserve the image recognition abilities of the original model.



- During **intra-scale pretraining**, UNIT learns unified representations from multi-scale inputs, where images and documents are at their commonly used resolution, to enable fundamental recognition capability.
- During **inter-scale finetuning stage**, the model introduces scale-exchanged data, featuring images and documents at resolutions different from the most commonly used ones, to enhance its scale robustness.



Experiments

Method	#Param.	ZS cls.	kNN cls.	Segm.
EfficientViT-L1 [7]	38M	71.73	79.90	33.12
SwinV2-S [29]	49M	74.70	81.12	35.57
ConvNext-B [31]	88M	75.43	81.73	38.95
MViTV2-B [27]	51M	75.92	81.39	41.39
NFNet-F3 [6]	254M	76.93	80.50	38.31
MaxViT-B [48]	119M	77.49	79.34	38.46
OpenCLIP-H/14 [43]	632M	77.19	81.10	40.04
RADIO-L/14 [44]	304M	77.25	84.03	48.70
E-RADIO-L/14 [44]	265M	77.87	83.73	45.50
RADIO-H/14 [44]	632M	78.62	84.17	49.01
UNIT (ours)	632M	**78.76**	**84.18**	**50.19**

- Our method achieves comparable results with existing vision encoders on zero-shot image classification, k-NN classification, and semantic segmentation benchmarks.
- Our method outperforms document-oriented models on OCR benchmarks.

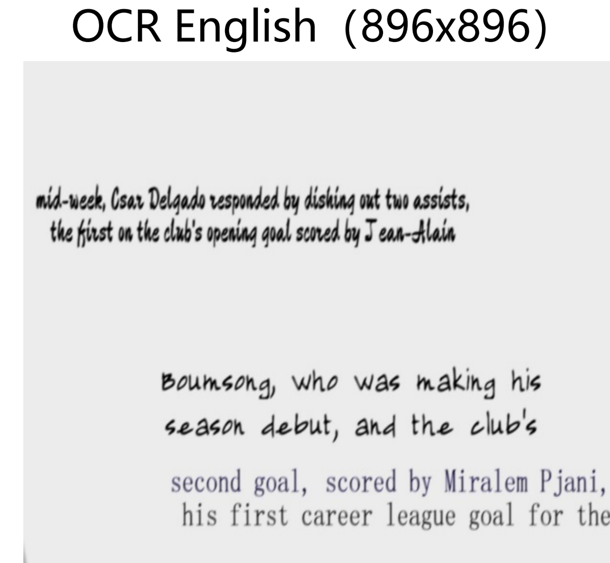
Method	Backbone	#Params	Input	FUNSD	SROIE	CORD	SYN-L-val	MD-val
Donut [23]	Swin-B	260M	1280 × 960	9.08	8.94	16.64	44.78	5.07
Nougat [5]	SAM-ViT-B	247M	896 × 672	55.35	33.64	1.57	66.76	**86.71**
Vary* [53]	SAM-ViT-B	525M	1024×1024	21.01	9.84	12.89	91.20	59.30
RADIO* [44]	ViT-H/14	632M	896×896	26.12	10.42	10.01	93.90	37.57
UNIT (ours)	ViT-H/14	632M	896×896	**67.14**	**41.48**	**58.87**	**95.33**	78.50

- Our method significantly outperforms the compared models on document-oriented QA tasks and demonstrates comparable performance on other QA tasks.

Method	ChartQA	DocVQA	InfoVQA	OCRBench	GQA	OKVQA	MME	MathVista
CLIP-L [43]	52.0	57.2	29.3	382	62.3	57.0	1503.6	42.7
SigLIP [65]	56.5	62.0	29.7	429	63.0	61.1	1489.4	44.2
UNIT (ours)	**61.0**	**65.5**	**31.9**	**480**	**63.9**	**61.5**	**1529.8**	**44.6**

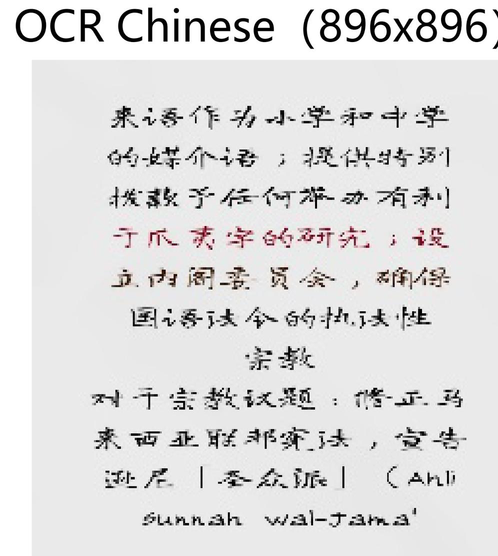
Visualization

OCR English (896x896)



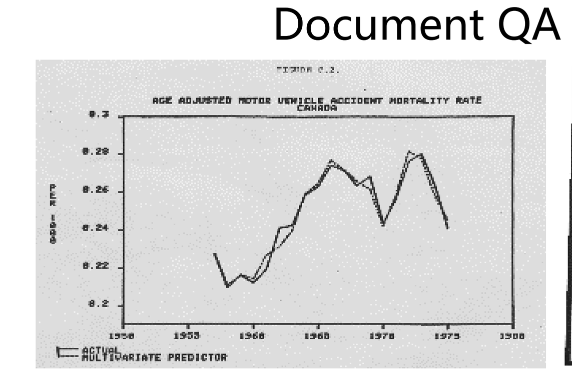
mid-week, Csar Delgado responded by disheshining out two assistants, the first on the club's opening goal scored by Jean-AlainBoumsong, who was making his season debut, and the club'ssecond goal, scored by Miralem Pjani, his first career league goal for the

OCR Chinese (896x896)

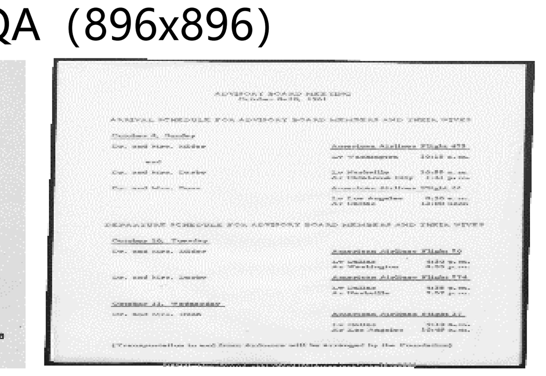


来语作为小学和中学的姨价语；提供特别按款予任何举办有利于爪字的研究；设立内阁委员会，确保国语法令的批法性宗教对于宗教议题：修正马装西亚联邦法，宣告逊尼「圣众派」（Ahi Sunnah wal-Jama'

Document QA (896x896)



Q: What is the actual value per 1000, during the year 1970?
A: 0.24



Q: When was 'advisory board meeting' scheduled?
A: October 8-10, 1961

VQA (224x224)

Q: What is the name of the flower tree? A: Cherry

Q: What fruit is typically added to the top of cereal? A: Banana