GRAM: Global Reasoning for Multi-Page VQA

Supplementary Material

Group	Parameter Name	Parameter Value	
	batch size	8	
	training steps	200K	
	warmup steps	1000	
fine-tune	fp16	True	
	training number of pages	4	
	evaluation number of pages	unlimited	
	number of image tokens	128	
	encoder learning rate	3e-5	
DooFormor [1]	decoder learning rate	3e-5	
Docronner _{concat} [1]	training text tokens per page	600	
	inference text tokens per page	400	
	encoder learning rate	3e-5	
	decoder learning rate	3e-5	
HiVT5* [4]	training text tokens per page	800	
	inference text tokens per page	8000	
	number of compression tokens per page	10	
	encoder learning rate	3e-5	
	decoder learning rate	3e-5	
	global encoder learning rate	1e-4	
GRAM	training text tokens per page	800	
	inference text tokens per page	8000	
	number of global tokens	32	
	bias adaptation constant 'c'	20	
	encoder learning rate	3e-5	
	decoder learning rate	3e-5	
GRAM _{C-Former}	global encoder learning rate	1e-4	
	C-Former learning rate	1e-4	
	training text tokens per page	800	
	inference text tokens per page	8000	
	number of global tokens	32	
	bias adaptation constant 'c'	20	
	compression length	256	

Table 1. Hyper-Parameters.

A. Parameters

We present in Tab. 1 all of the relevant hyperparameters.

B. Inference Resources Consumption

We compare three key properties of MP-DocVQA baselines and our method: inference time, memory consumption, and maximal document length. The latency and memory consumption are illustrated in Fig. 1 and Fig. 2, respectively, both as functions of the number of pages in the document. We compare the following baselines: DocFormerv2_{concat} [1], Hi-VT5 * [4], and our GRAM and GRAM_{C-Former}, utilizing the same computational resources employed in all experiments— $8 \times A100$ GPUs with 40GB of memory.

The memory consumption of DocFormerv2_{concat} [1] reaches its maximum capacity for documents with only 20 pages, while our method efficiently processes documents, spanning hundreds of pages. Moreover, the presented figures demonstrate that $\text{GRAM}_{C-Former}$ maintains a com-

parable memory footprint to the GRAM model. Nevertheless, there is potential for improvement, as HiVT5* exhibits lower memory consumption. Despite this, we achieve inference times similar to HiVT5* [4], accompanied by a noteworthy enhancement in ANLS.



Figure 1. Latency comparison. We compare the dependency between overall latency and the number of pages in input document.



Figure 2. **Memory consumption comparison**. We compare the dependency between overall memory consumption and the number of pages in input document.

C. Qualitative Results

Finally, we present a few qualitative results on the DUDE dataset in Fig. 3, showcasing the advantages of our approach over Hi-VT5* [4]. In the first three examples, we demonstrate cases where GRAM is correct and HiVT5* is wrong. The last two examples present cases where both our method and HiVT5* are incorrect.

D. Comparison with DocFormerV2_{concat}

We provide additional qualitative examples with DocFormerV2_{concat}. Examples demonstrate the effectiveness of GRAM in tackling questions that involve multiple pages in the document.

Method	ANLS by Number of Pages DUDE validation dataset					
	All	1	2-4	5-10	11-end	
GRAM	47.88	49.29	49.90	45.90	43.94	
DocFormerv2 _{concat}	44.32	46.08	47.05	42.81	38.17	
DocFormerV2 _{Longformer}	45.88	47.01	47.75	43.22	43.13	
DocFormerV2 _{AliBi}	34.73	36.55	37.00	30.99	31.25	

Table 2. **Comparison to NLP methods**. Results on DUDE validation comparing GRAM with LongFormer [2] and AliBi [3].

E. Comparison with NLP-based Approaches

We present additional experiments, comparing GRAM with two NLP-based approaches: the sparse attention-based LongFormer [2], and the bias-based AliBi [3]. Both approaches are implemented on top of DocFormerv2 for fair comparison. Results in Tab. 2 shows an advantage in our local-global approach of utilizing existing powerful models for single-page and extending them to support the multipage scenario.

References

- [1] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*, 2023. 1
- [2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020. 2
- [3] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409, 2021. 2
- [4] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. arXiv preprint arXiv:2212.05935, 2022. 1, 3

How many chapters are in the books? correct ans	swer: ""
HiVT5: "4"	
GRAM: ""	

CRIMINAL LAW, PROPERTY CRIMES	Execution The second	Viscours a series We are applied as the foreign entropy project of these flag paleses officers are started (back) to series interested with an effective posterior content of page (or posterior content of the file A to Mills the entropy term bace in a content of the file	 secting descent or dependent or water property descenting that the section of the long sector for descent section or property is section of the long sector for definition to section or developed in the section of the long sector of the long	sourceathalogue de los cistos facto batela e sourceanse atentes registos des bate May uter terro terro una constructivada de la cisto de la cisto de la cisto de la cisto de Securitario de la cisto de Securitario enclando en la cisto de la cisto de la cisto de la cisto de la cisto de Securitario de la cisto de Securitario de la cisto de Securitario de la cisto de Securitario de la cisto de la	NUMA VPL Access IN Just P Access and an exploration with the control bandwidth and a the product with ground and analysis and the children of the Access part of the Access part of the analysis of the access part of the Access part of the access of accessing of the Access part of the Access part of the Access part of the Access of the Access of the Access of the Access of the Access part of the Access o	For Applied proposed data an enclosular productivity of advanced or advanced or Apply proposed and an advanced or Applied on the Applied State of Applied State of Applied State Technologies and the Applied State of Applied State State of Applied State of Applied State Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State of Applied State Applied State of Applied State of Applied State Applied State of Applied State of Applied State Applied State of Applied State of
<u>rolos</u>	Bandward The start of	VIETE WALL-10, back high starts in All The annumery set of the an	A De Yook and A De Yook A	Les Altres controls en la de la del la de NET MILIONE DE La del	 percent percent sectors. Marcine and the sector sector sector sectors and the sector sector sectors. Marcine and the sector sector sector sectors and the sector sector sectors. Marcine and the sector sector sector sector sectors and the sector sectors. Marcine and the sector sector sector sectors and the sector sectors. Marcine and the sector sector sectors. Marcine and the sector sectors. 	Andream State (State State Sta
	Internet and the second s	The analy, by Secury of System B. Stand up to the site of the System States (strength to change the System State (strength strength strength States) (strength states (SST) We cannot step to Many (strength).	Name (most) is a solutioning effective, two out-charge (inclusion) or "controlling" of adult property further fault of flatters charge (most) of control proves the person the decision	 within your provide of any non-banks within second and provide of 	Recently instruction of load one in a model <u>sectory top</u> (1991) Recently one instruction	No man since a finite of a filler stream for a filling is 1998. For animal damage is properly the advanced is because a stream is 1999. A stream of the stre
PRESENTED IN) DPS - Law Enforcement Academy Sunta Fa, New Mexico	Concerning and Concer	The second secon	tere in a constant per a carte transming transmission per personal and transmission of the second se	subject privile year of the privile year	Resident excellent of the decomposition of the strength of the	(coll) control distances) and advances, it have get a lot of period period. The sector of the sec
		BELIEVE LINE CONTRACTOR The Assess of Records Mathematican a memory t	Section 2.1. Se	spring - Sign behavior strength of an URLP's should be the regard and spring - Sign begins the grant grant participation and participation for spring the spring of a spring of the spring s	 A more in plane takes share based as a solid or childrad or the plane based as a solid or childrad or the plane based as a solid or childrad or the plane plane based as a solid or childrad or the plane based as a solid or childrad or the plane based or the solid or	CONTRACTOR OF A DESCRIPTION OF A DESCRIP









Figure 3. Qualitative comparison between our approach and Hi-VT5 [4] indicates that the integration of our global-local encoder enhances reasoning capabilities, especially when inquiries require multi-page context.

Question: Explain Criminals in Love

GT answer: "Tragi-comedy play. Tells the story of Junior who is forced into criminal activity by his father"

GRAM: "It is a tragi-comedy which tells the story of junior and his passionate love for gail, whose dreams for a simple life go awry when junior's father, a chronic offender, coerces him into criminal activity. One by one, more of junior's friends attempt to helphim, butare forcedinto the gang led by his aunt wineva"

DocFormer concat: ""



Question: How many paragraphs are 1 page 1 GT answer: "3" GRAM: "3" DocFormer concat: "6"







Question: Which subject use in this document? GT answer: "" GRAM: "Location Map" DocFormer concat: ""



Question: How many pages with different colors are there in the document? GT answer: "3" GRAM: "3" DocFormer concat: "2"

	The second seco	
	PROFESSION OF STREET	

(c)

(d)

Figure 4. Comparisons between DocFormerV2_{concat} and GRAM.

(b) Question: Does the signature belong to the same person listed at the top of the form? GT answer: "Yes" GRAM: "Yes"

256 tokens GRAM: "BRIAN FRESKO" BRIAN JOHN ELLIOTT FRES FUTNING BOOK JOHN BLIOTT FREEK 195 MONTAGUE STREET, 14TH 16 VOITABLE STREET, Baines Repeared _____TETRICE HE TRACE Pere _____9962484 CALINDARY SEA FOR SA CUSPRE CONTRACTOR AND A A VIEW DATA OFSE REEL 🛙 🚥 Smalshat, a ship will receive and revie ation for 0 72-305.5. direct soli [whstat Brian Freskos Span Bies Feeder e and Admin Records: You m, fill it out and fax it to 720 80022 or can also call 7 nyytitlelen, il totadi jaihood aradia xeexx Il out the ele (e)