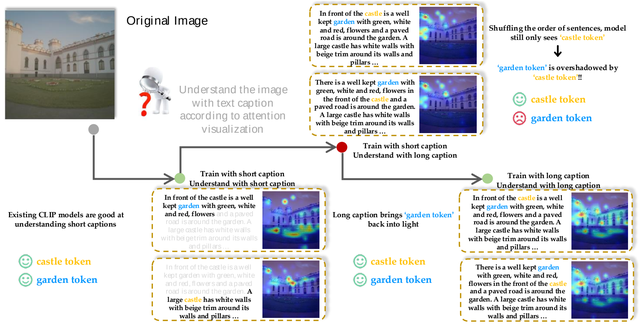
LoTLIP: Improving Language-Image Pre-training for Long Text Understanding

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, Zheng-Jun Zha

Project Page

Overview

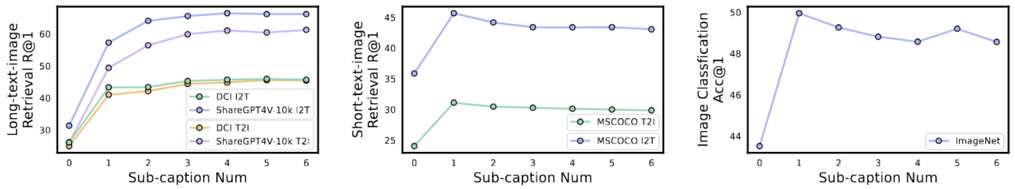
- Understanding long text beyond the reach of most language-image pre-training (LIP) model.
- We empirically confirm that the key reason is that training images are usually paired with **short captions**, leaving certain tokens easily overshadowed by salient tokens.



Contributions:

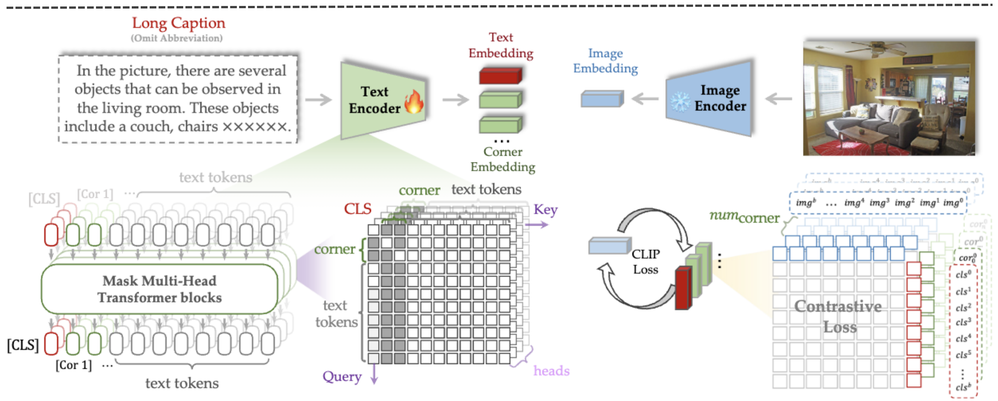
- Using 3 MLLMs to Re-caption **100M** images with long texts;
- **Enhance long text understanding** of LIP with the re-captioned long texts;
- Collected long text-image pairs from 3 dataset for constructing **long-text-image retrieval tasks**.

Method



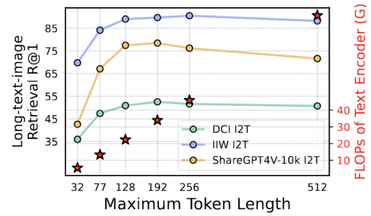
(a) long-text-image retrieval (b) short-text-image retrieval (c) image classification

As the texts getting longer, the performance in long-text-image retrieval tasks **improves**. However, there is a performance **degradation** in short-text-image retrieval task and ImageNet classification task.



To **enhance** both **long and short text understanding**, we add extra text tokens for text encoders, termed corner tokens, to aggregate diverse text features with help of designed attention mechanism.

More Analysis



- 77 token limitation is **insufficient** for a model training with long texts;
- the FLOPs of the text encoder **increases** with the text token number limitation.

#Corner Token	Attention Mask	Long-text-image Retrieval						Short-text-image Retrieval		Classification
		DCI		IIW		ShareGPT4v-10k		MSCOCO		ImageNet
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	Acc
0	-	47.96	44.92	84.97	81.70	73.66	66.73	43.52	30.06	48.87
1	✓	49.57	46.55	84.97	82.68	74.91	68.41	45.68	31.51	49.62
2	✓	49.46	47.82	84.97	83.33	76.49	69.72	46.56	31.59	50.34
3	✓	**49.58**	47.70	**87.09**	**84.31**	**76.51**	**70.20**	46.48	31.60	50.36
4	✓	**49.58**	**48.30**	86.76	84.15	76.25	70.14	**47.70**	**31.88**	**50.59**
2	-	48.61	47.17	86.11	81.86	76.14	69.31	47.70	31.34	49.88
2	✓	49.46	47.82	84.97	83.33	76.49	69.72	46.56	31.59	50.34

- Corner tokens **enhance** both of the short text understanding and long text understanding ability.

Long Text-Image Dataset

For Training:

Dataset	#Images	#Texts	#Sub-captions per Text	#Tokens per Text
Short-text-image Pairs Dataset				
CC3M [27]	3,018,175	3,018,175	1.01	11.29
CC12M [27]	10,445,969	10,445,969	1.00	17.48
YFCC15M [29]	14,772,456	14,772,456	1.23	13.61
LAION47M [26]	49,677,119	49,677,119	1.28	18.99
COYO24M [2]	24,658,004	24,658,004	1.21	17.07
Long-text-image Pairs Dataset				
LoTLIP	102,571,723	307,715,169	6.16	136.14

- The **100M** images are from the five short-text-image pairs dataset;
- We employ **InstructBLIP, LLaVA, and ShareGPT4V** to generate long texts based on the collected image;
- The averaged length of texts reaches **136 tokens**.

For Evaluation:

Dataset	#Images	#Texts	#Sub-captions per Text	#Tokens per Text
Long-text-image Retrieval Dataset				
DCI [31]	7,805	7,805	10.81	172.73
IIW [12]	612	612	10.16	239.73
ShareGPT4v-1k [6]	1,000	1,000	8.15	173.24
ShareGPT4v-10k [6]	10,000	10,000	8.24	173.66
Short-text-image Retrieval Dataset				
MSCOCO [18]	5,000	25,000	1.0	11.77
Flickr30k [35]	1,000	5,000	1.0	14.03

- In short-text-image-retrieval tasks, the textual inputs contain fewer than **15 tokens** on average;
- We collected long-text-image pairs from **DCI, IIW, and ShareGPT4V** datasets to construct long-text-image retrieval tasks

Main Results

Data Scale		Model	DCI		IIW		ShareGPT4v-1k		ShareGPT4v-10k		Avg.
Short	Long		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	
3M	-	FILIP [34]	10.85	11.36	31.54	29.08	26.50	26.80	8.94	8.64	19.21
3M	-	LaCLIP [11]	14.84	14.71	41.01	38.89	40.90	37.10	15.81	14.84	27.26
3M	-	SigLIP [38]	11.66	13.11	29.25	29.58	27.30	25.10	9.92	9.30	19.40
3M	-	LiT [37]	27.14	24.13	65.20	58.50	63.60	56.80	32.73	27.01	44.38
3M		LoTLIP	**49.46**	**47.82**	**84.97**	**83.33**	**93.20**	**90.00**	**76.49**	**69.72**	**74.37**
400M	-	CLIP [24]	45.45	43.01	88.24	87.58	84.50	79.80	60.22	56.16	68.12
100M	-	LiT [37]	41.78	40.90	88.07	82.68	86.00	80.00	61.41	50.69	66.44
700M	-	ALIGN [14]	56.54	57.41	92.65	90.68	86.30	85.30	65.13	62.73	74.59
12B	-	SigLIP [38]	57.78	56.22	91.99	91.01	85.80	83.40	83.40	63.08	76.59
400M	1M	Long-CLIP* [39]	51.68	57.28	89.61	**93.20**	94.70	93.40	79.24	77.06	79.52
100M		LoTLIP	**62.10**	**61.06**	**93.95**	92.48	**96.50**	**95.50**	**86.84**	**81.40**	**83.72**

Visualization