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

Wei Wu[1] Kecheng Zheng[2,3] Shuailei Ma[4] Fan Lu[1] Yuxin Guo[5] Yifei Zhang[6]
Wei Chen[3] Qingpei Guo[2] Yujun Shen[2] Zheng-Jun Zha[1]

[1]University of Science and Technology of China, [2]Ant Group, [3]Zhejiang University, [4]Northeastern University, China,
[5]Institute of Automation, Chinese Academy of Sciences, [6]Shanghai Jiao Tong University

Project Page Code

Motivation

Why existing language-image pre-training models struggle to understand long texts?

Original Image

Understand the image with text caption according to attention visualization

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

There is a well kept garden with green, white and red, flowers in the front of the castle and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

Train with short caption
Understand with long caption

Shuffling the order of sentences, model still only sees 'castle token'

↓

'garden token' is overshadowed by 'castle token'!!

☺ castle token

☹ garden token

Train with short caption
Understand with short caption

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

Existing CLIP models are good at understanding short captions

☺ castle token

☺ garden token

Reason:

Lack of large-scale long text-image pairs for pre-training.

Pre-training Dataset

Long texts re-captioning:

- We collected 100M data from five publicly available datasets;
- InstructBLIP, LLaVA, ShareGPT4V are used for re-captioning;
- The averaged length of long captions reaches 136 tokens.

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

Data Sample

Caption from InstructBLIP: In the image, there are several objectively existing and visible objects with attributes. These include: 1. A cat: The main subject of the image is a gray and white cat, which is being fed by a person's finger. 2. The person's hand: The person's hand is visible in the image, holding the finger that is feeding the cat. 3. The cat's mouth: The cat's mouth is open as it receives the food from the person's finger. 4. The cat's teeth: The cat's teeth are visible as it opens its mouth to receive the food from the person's finger. 5. The person's fingertip: The person's fingertip is touching the cat's mouth, demonstrating the act of feeding the cat. 6. The cat's fur: The cat's fur is visible in the image, contributing to the overall visual representation of the scene.

Caption from ShareGPT4V: In the image, a person's finger is seen pointing towards a cat's nose. The cat, with its fur in shades of gray and white, is looking up at the finger with its mouth slightly open, as if curious or interested in the action. The background is a plain white wall, providing a stark contrast to the subjects in the foreground. The cat's position relative to the finger suggests it's being the center of attention, possibly for a playful interaction or a trick. The image captures a moment of connection between the person and the cat, highlighting the curiosity and engagement of the feline.

Caption from LLAVA: The image features a cat with a long whisker, sitting on a white surface and looking up at a person's hand. The person is gently touching the cat's nose with their index finger, possibly petting or teasing the cat. The cat appears to be enjoying the interaction, as it is looking up at the person's hand and possibly purring. The scene captures a moment of bonding between the cat and the person.

Motivation

Use long captions in language-image pre-training

Original Image

Understand the image with text caption according to attention visualization

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

There is a well kept garden with green, white and red, flowers in the front of the castle and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

Shuffling the order of sentences, model still only sees 'castle token'

↓

'garden token' is overshadowed by 'castle token'!!

☺ castle token

☹ garden token

Train with short caption
Understand with long caption

Train with short caption
Understand with short caption

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

Existing CLIP models are good at understanding short captions

☺ castle token

☺ garden token

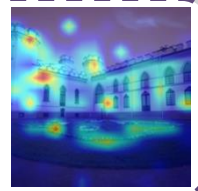
Long caption brings 'garden token' back into light

☺ castle token

☺ garden token

Train with long caption
Understand with long caption

In front of the castle is a well kept garden with green, white and red, flowers and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …



There is a well kept garden with green, white and red, flowers in the front of the castle and a paved road is around the garden. A large castle has white walls with beige trim around its walls and pillars …

The usage of long captions brings the overshadowed tokens back into the light.

Evaluation datasets

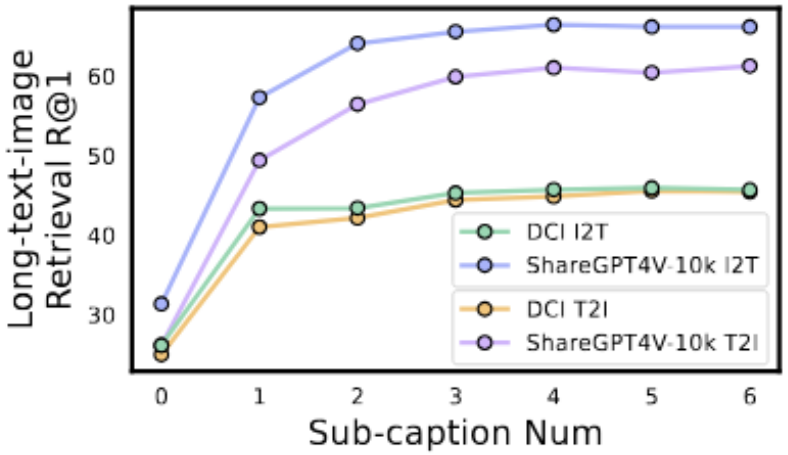
To quantitative evaluate the long text understanding ability of the pre-trained model, we construct long-text-image retrieval tasks.

- The texts in short-text-image retrieval tasks contain fewer than 15 tokens on average;
- We collected three long text-image pairs datasets, where the averaged text length exceeds 170 tokens;
- The long texts from DCI and IIW are human-annotated.

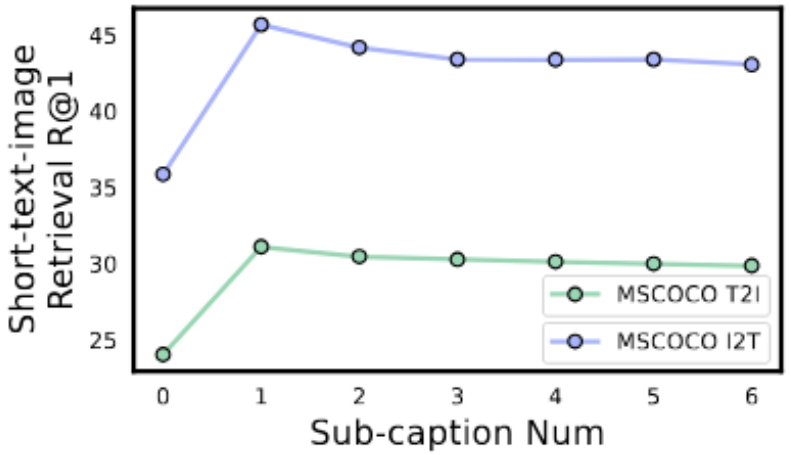
Dataset	#Images	#Texts	#Sub-captions per Text	#Tokens per Text
Long-text-image Retrieval Dataset				
DCI [31]	7,805	7,805	10.81	172.73
IIW [12]	612	612	10.16	239.73
ShareGPT4v-1k [6]	1,000	1,000	8.15	173.24
ShareGPT4v-10k [6]	10,000	10,000	8.24	173.66
Short-text-image Retrieval Dataset				
MSCOCO [18]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

Method

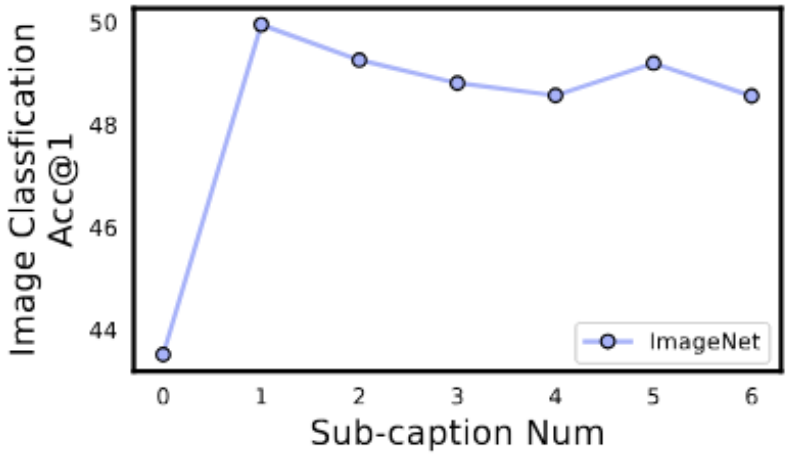
Influence of directly using long captions in pre-training



(a) long-text-image retrieval



(b) short-text-image retrieval



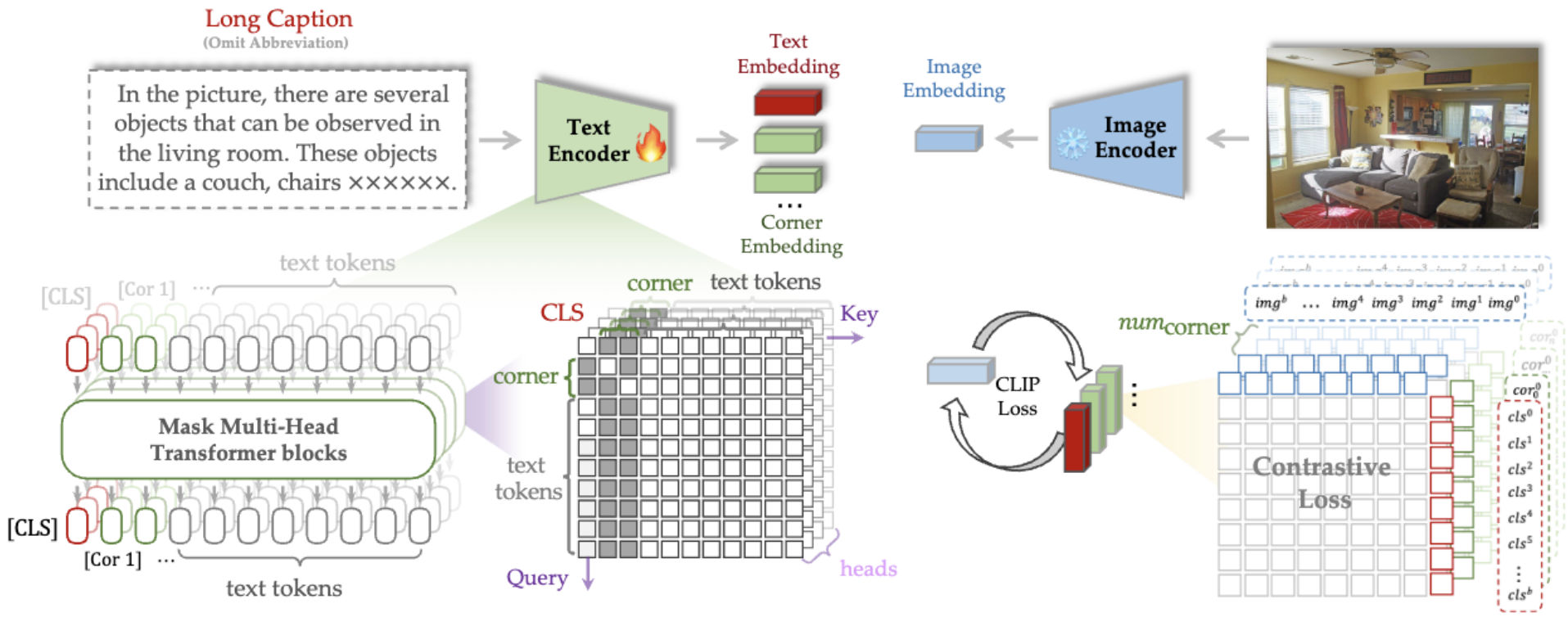
(c) image classification

- (a) The performance on long-text-image retrieval tasks is consistently improved as the texts get longer;
- (b)(c) there is a performance degradation on short-text-image retrieval and image classification tasks.

Directly learning with long texts negatively impacts the ability of understanding short texts.

Method

Regain short text understanding ability and further enhance long text understanding ability



- Integrating corner tokens ([Cor 1], [Cor 2], . · · ·) to aggregate diverse textual information;
- Designing attention mask mechanism for text encoder to promote the diversity of the aggregated feature.

Ablation Studies

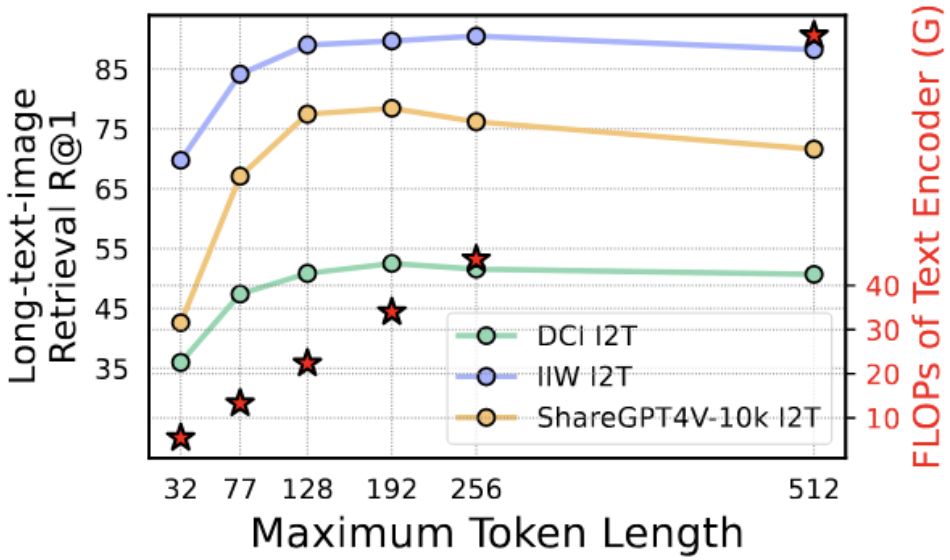
Analysis on corner tokens and attention mask

#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

- More corner tokens improve the performance on long texts and short texts related tasks;
- Use Attention mask mechanism improve image classification accuracy by 0.46%.

Ablation Studies

Influence of token number limitation



- The performance of the pre-trained model on different tasks improves when the token number limitation increases up to 192, exceeds the commonly used 77 tokens;
- The FLOPs of the text encoder, which increases with the text token number limitation.
- To balance the training efficiency and performance, we set token number limitation to 128.

Experimental Results

Comparison with SOTA methods

Long text-image Retrieval

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**

- LoTLIP trained with 100M data exceeds all state-of-the-art methods on long-text-image retrieval tasks, even though these methods pre-trained on a much larger scale of data.

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks!

Project Page

Code