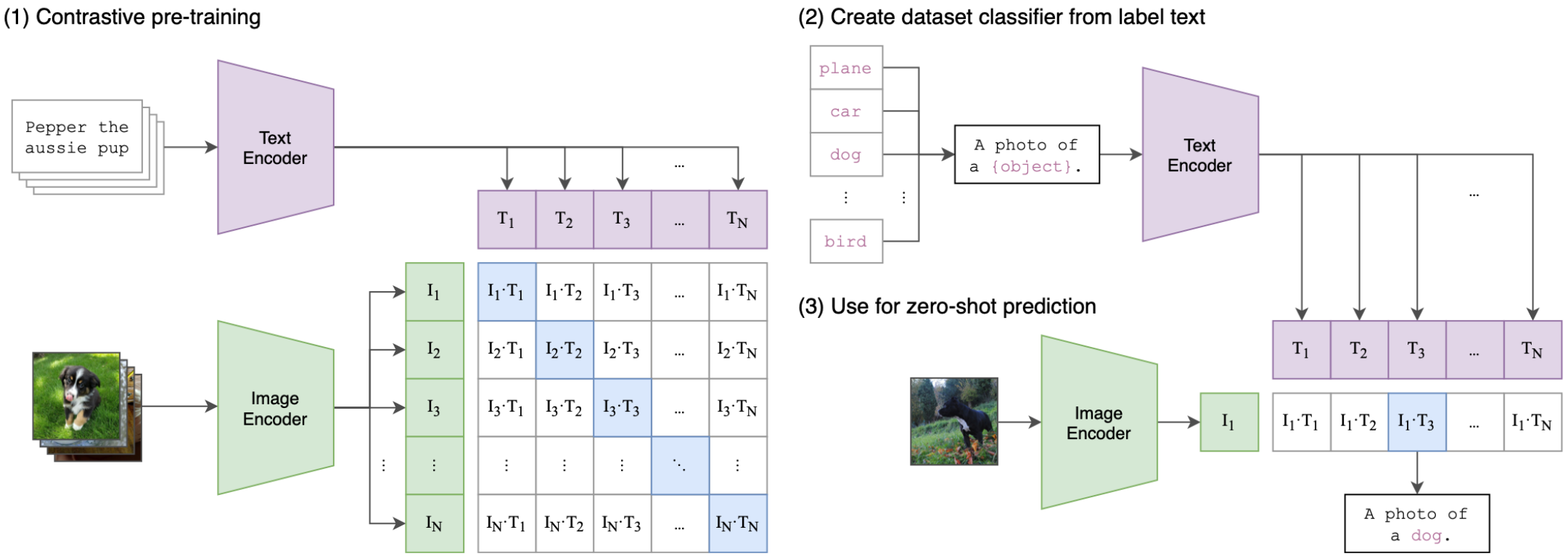
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Zhuoyi Yang

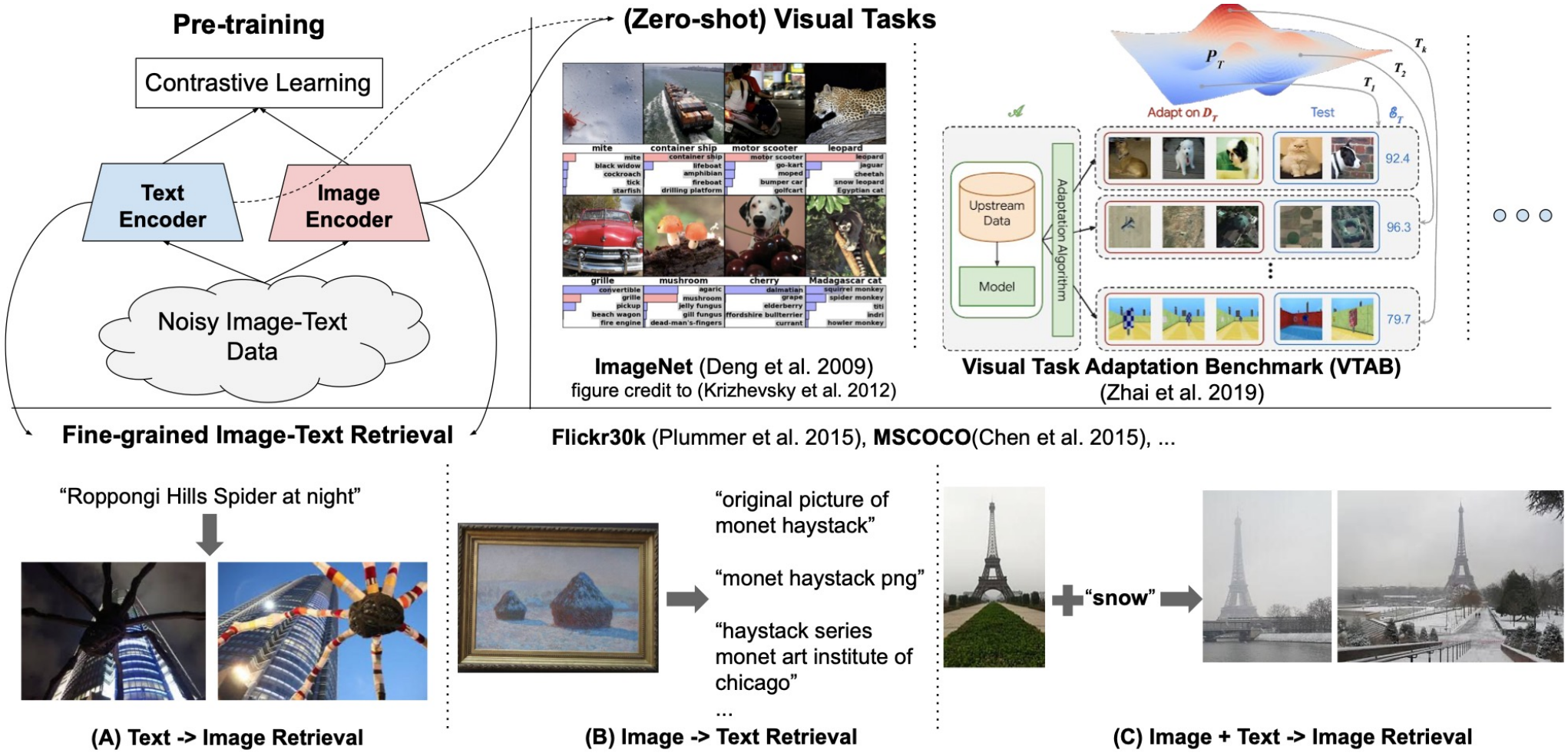
Jun. 21st, 2023

Motivations



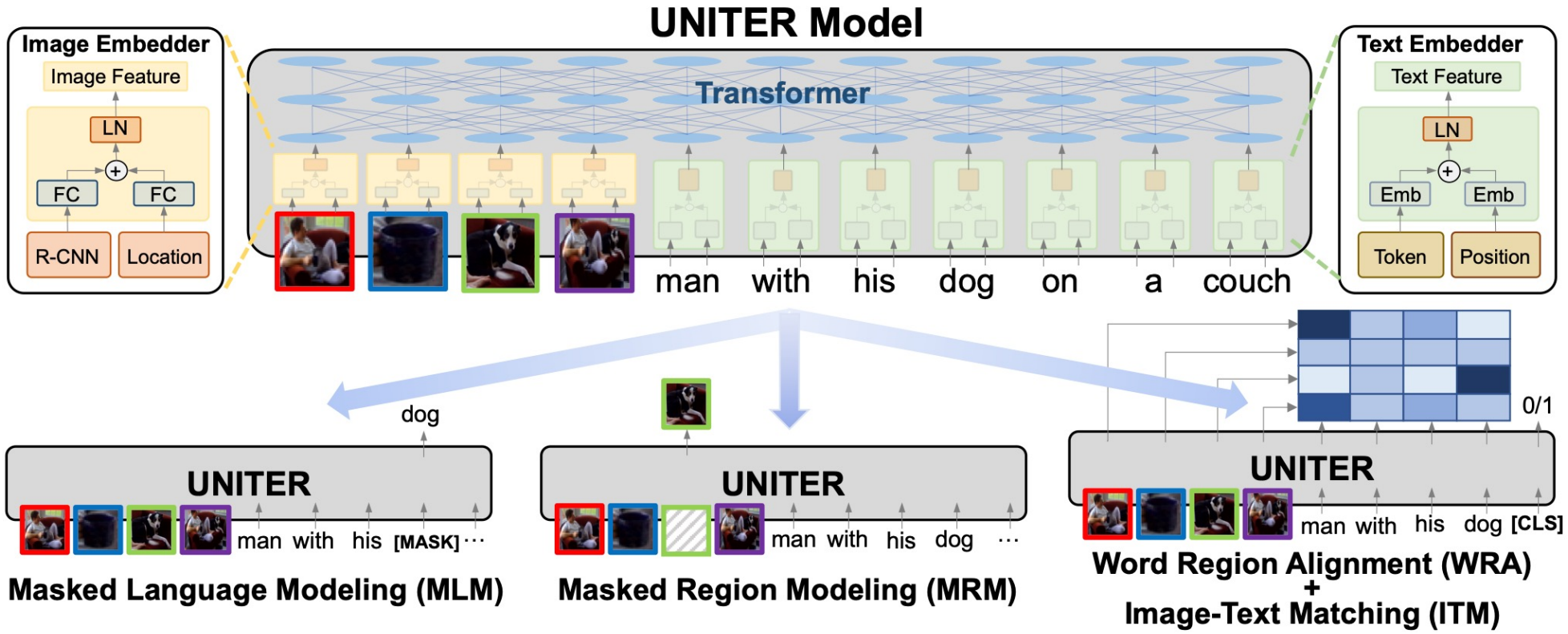
Radford, A., J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

Motivations



Jia, C., Y. Yang, Y. Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918, 2021.

Motivations



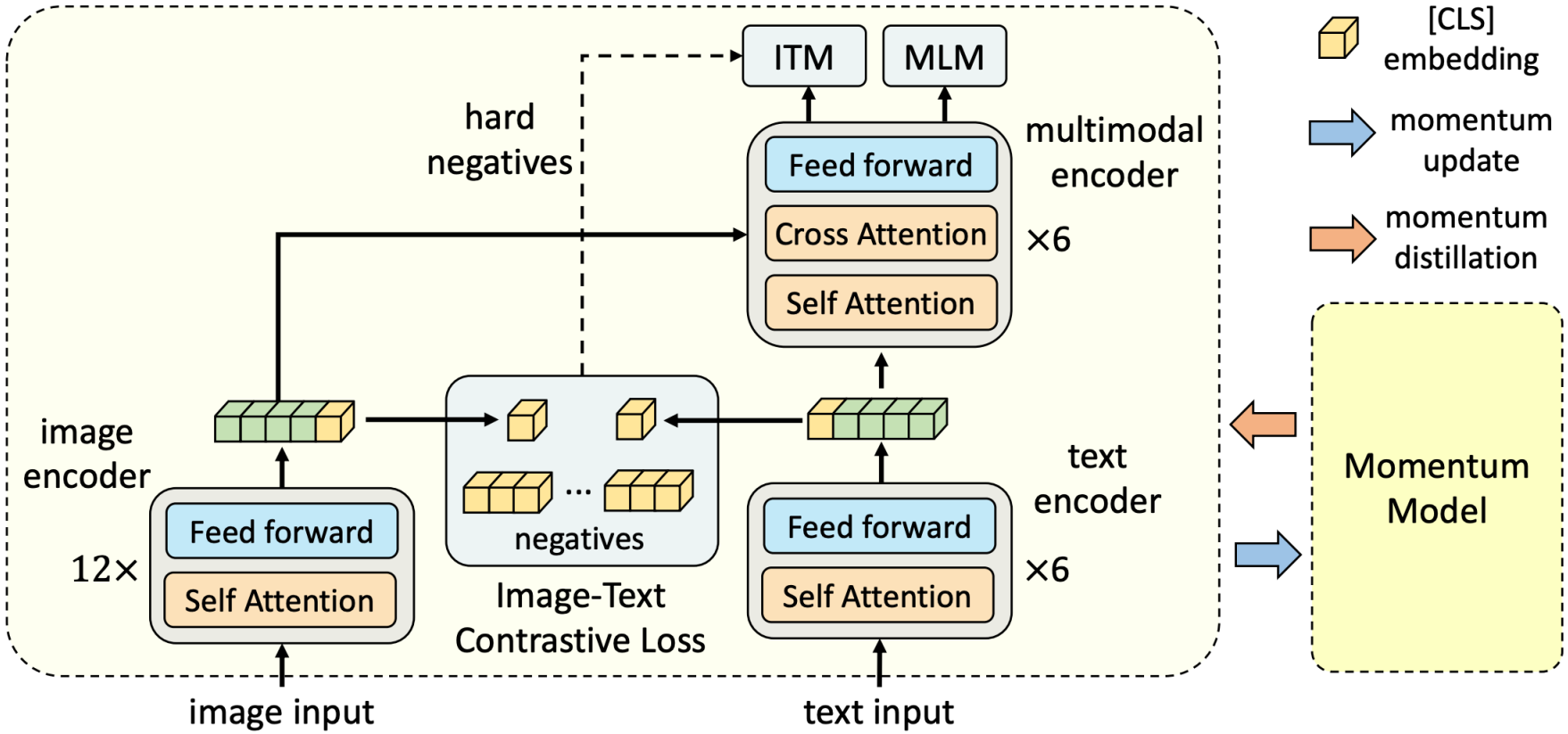
Chen, Y., L. Li, L. Yu, et al. UNITER: universal image-text representation learning. In ECCV, 2020.

Motivations

Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (pp. 12888-12900). PMLR.

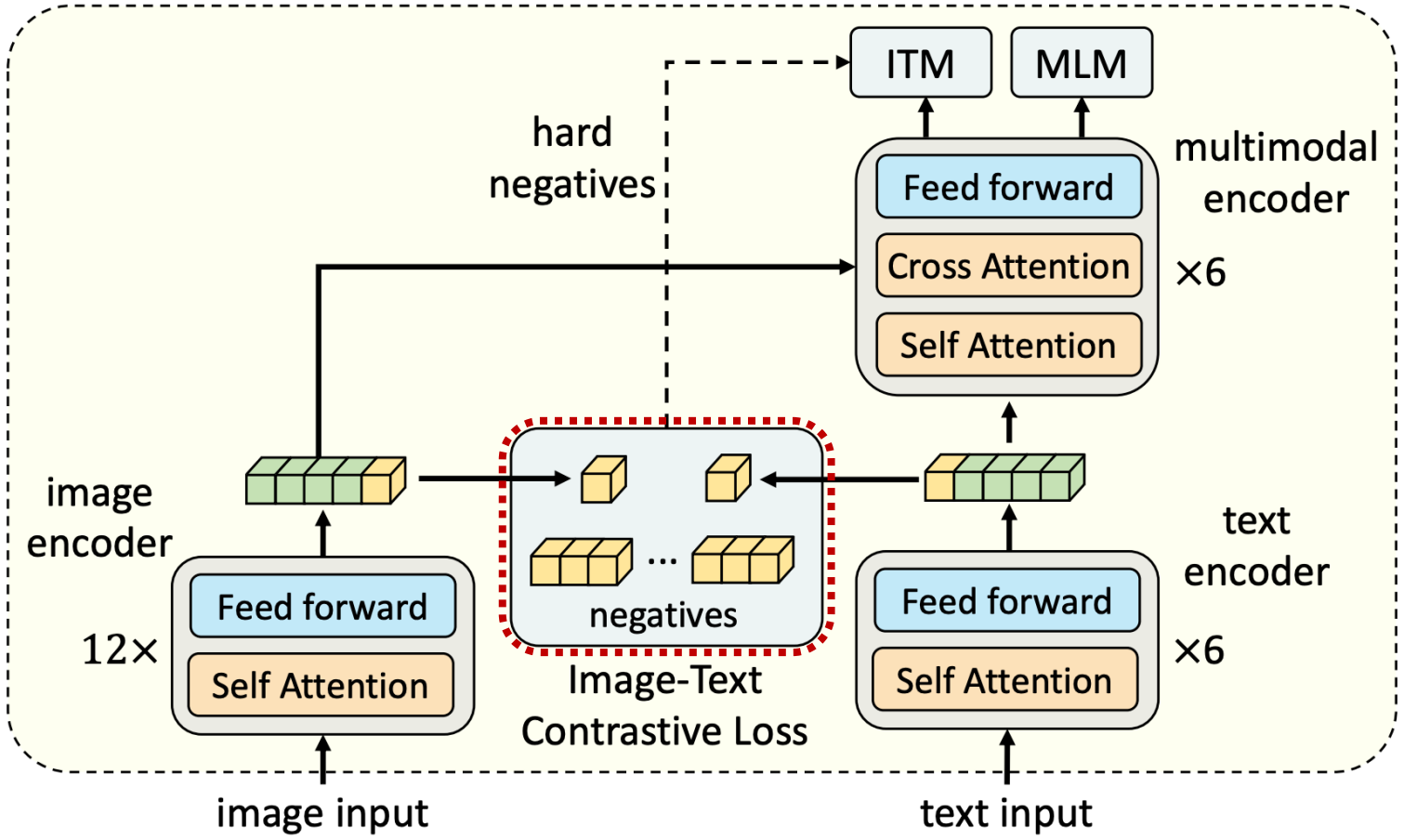
Methods

- Image-text Contrastive(ITC)
- Masked language model(MLM)
- Image-Text matching(ITM)



ALBEF

Methods



ALBEF

- Use [CLS] embedding as unimodal representation
- Encourage an image-text pair to have similar feature
- Align the unimodal features
- Learn better unimodal encoders

Methods

$$s = g_v(v_{\text{cls}})^\top g_w(w_{\text{cls}})$$

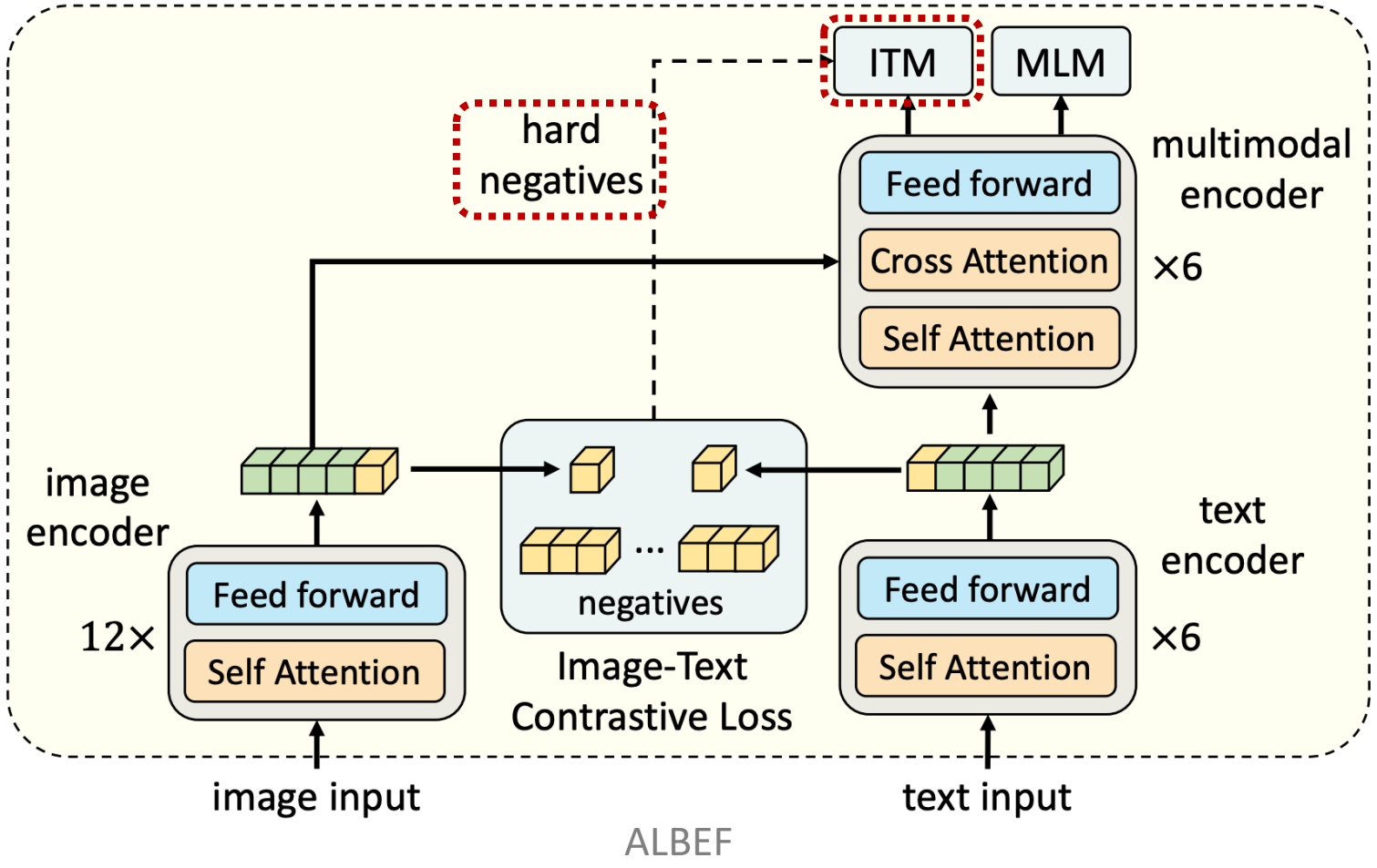
$$s(I,T) = g_v(v_{\text{cls}})^\top g'_w(w'_{\text{cls}}),\ s(I,T) = g_w(w_{\text{cls}})^\top g'_v(v'_{\text{cls}})$$

$$p_m^{i2t}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)},\ p_m^{t2i}(I) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)}$$

$$\mathcal{L}_{\text{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\,D}[H(p^{i2t}(I), y^{i2t}(I)) + H(p^{t2i}(T), y^{t2i}(T))]$$

- M: most recent M image-text representations from the momentum unimodal encoders. $M = 65536$
- H: cross-entropy

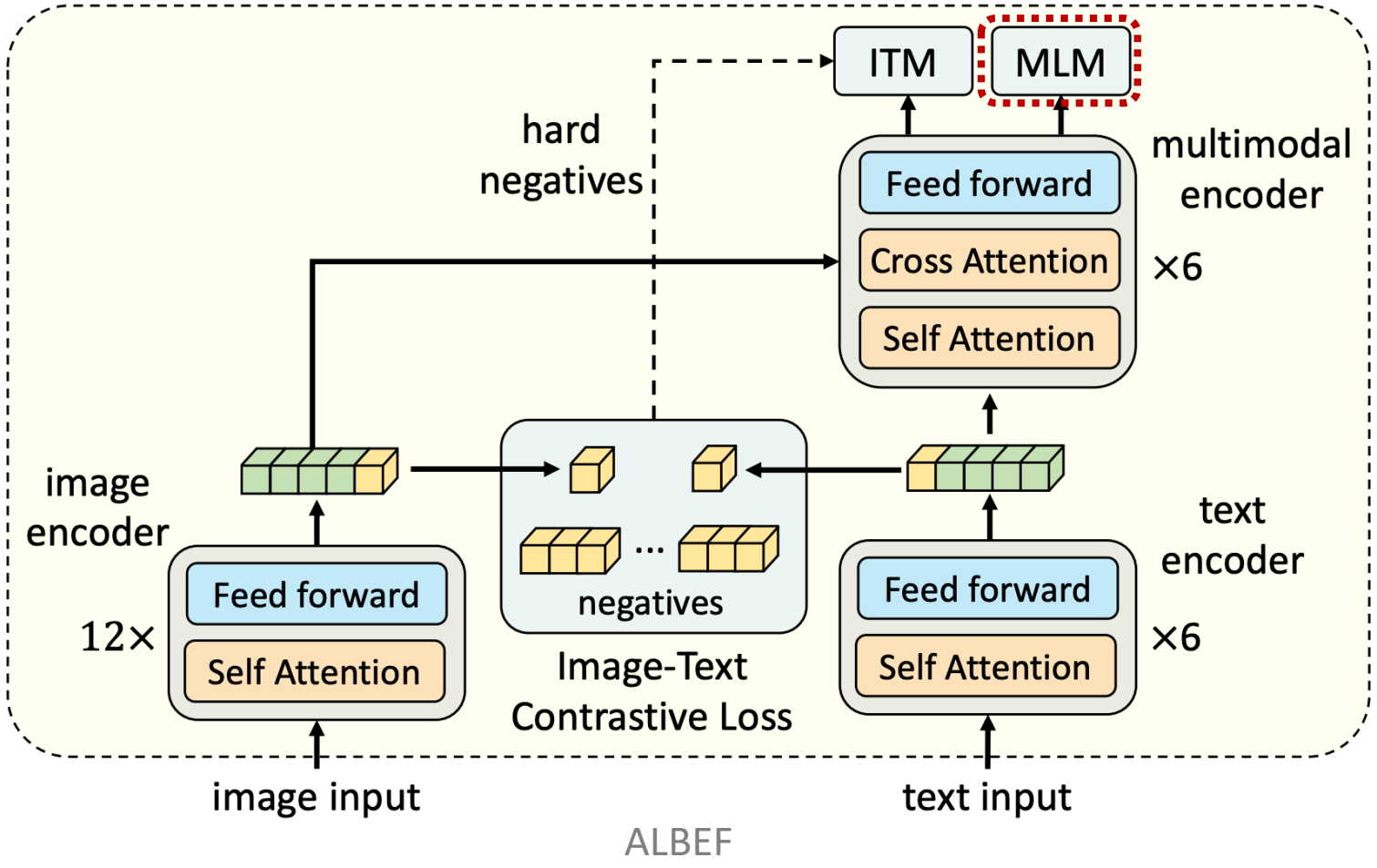
Methods



ALBEF

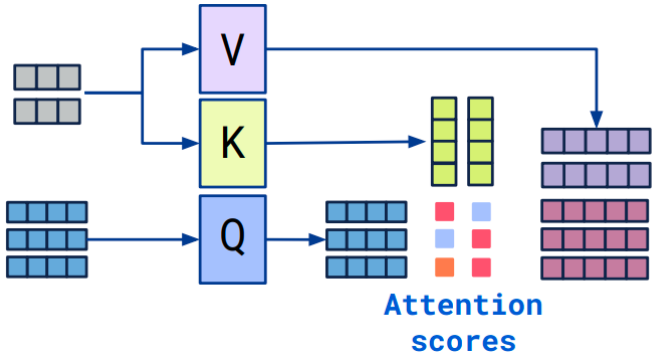
- Binary classification of positive and negatives
- Contrastive hard negative mining
- Mine more informative negatives with zero computation

Methods

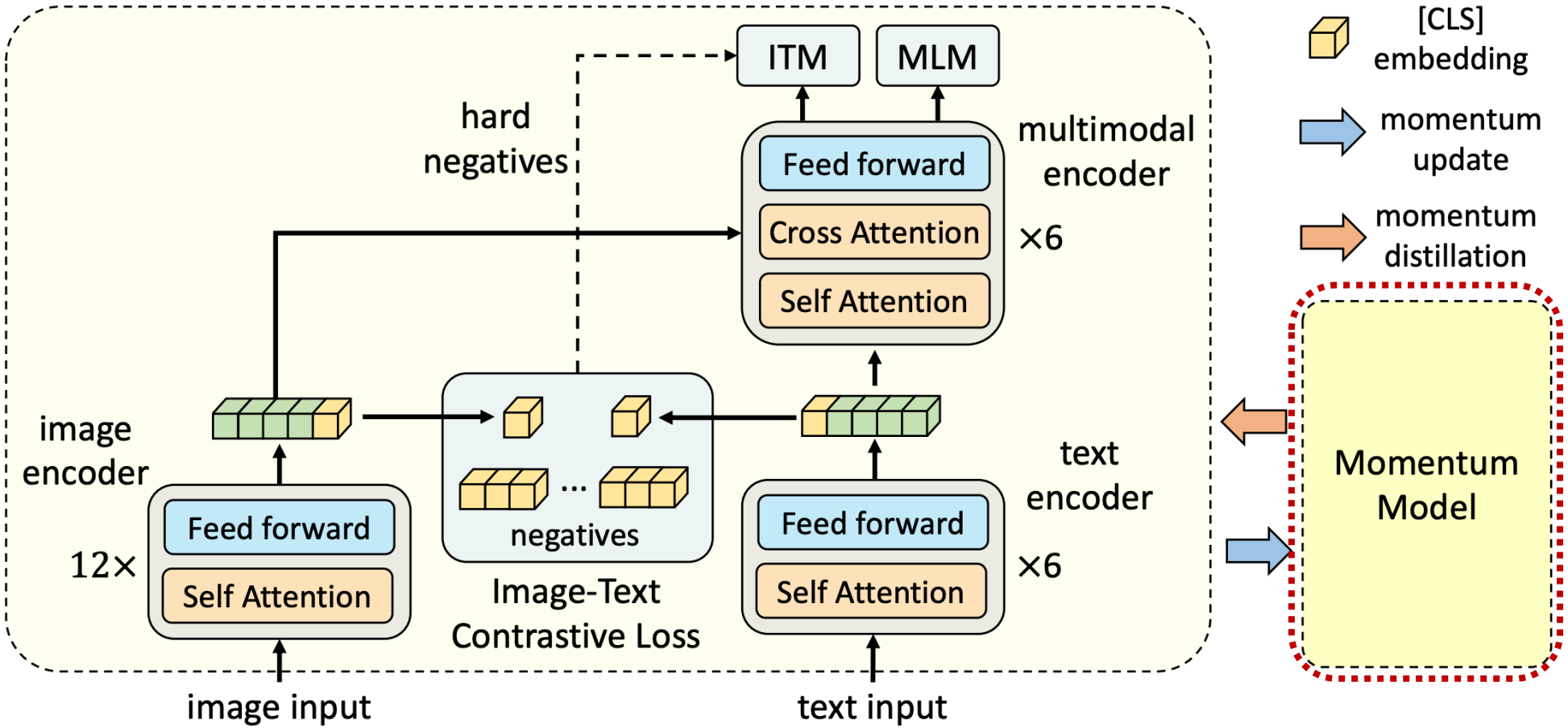


ALBEF

- Predict word based on image and contextual text



Methods



- Momentum model: a continuously-evolving teacher whose parameters are exponential moving-average of the base model's
- Generate pseudo-labels for ITC and MLM

Methods

** Contrastive Learning with Momentum**

$$s'(I,T) = g'_v(v'_{\text{cls}})^\top g'_w(w'_{\text{cls}}),\ s'(I,T) = g'_w(w'_{\text{cls}})^\top g'_v(v'_{\text{cls}})$$

$$q_m^{i2t}(I) = \frac{\exp(s'(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s'(I,T_m)/\tau)},\ q_m^{t2i}(I) = \frac{\exp(s'(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s'(T,I_m)/\tau)}$$

$$\mathcal{L}_{itc}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{itc}} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}[\text{KL}(p^{\text{i2t}}(I), q^{\text{i2t}}(I))\text{KL}(p^{\text{t2i}}(T), q^{\text{t2i}}(T))]$$

** MLM with Momentum**

$$\mathcal{L}_{\text{mlm}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{mlm}} + \alpha\mathbb{E}_{(I,\hat{T})\sim D}\text{KL}(p^{\text{msk}}(I,\hat{T}), q^{\text{msk}}(I,\hat{T}))$$

Pseudo-targets for MLM

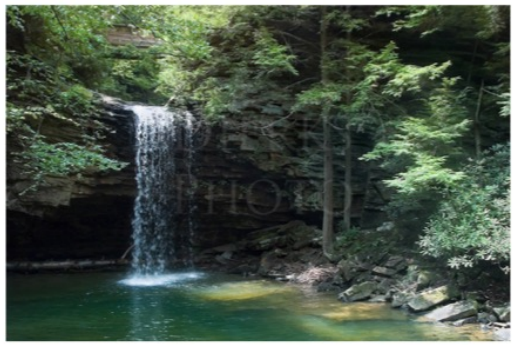
"polar bear in the [MASK]"

GT: wild

Top-5 pseudo-targets:

1. zoo
2. pool
3. water
4. pond
5. wild

"a [MASK] waterfall in the deep woods"



GT: remote

Top-5 pseudo-targets:

1. small
2. beautiful
3. little
4. secret
5. secluded

"a man [MASK] along a road in front of nature in summer"

GT: standing

Top-5 pseudo-targets:

1. walks
2. walking
3. runs
4. running
5. goes

Pseudo-targets for ITC

GT: breakdown of the car on the road
Top-5 pseudo-targets:

1. young woman get out of the car near the road
2. a woman inspects her damaged car under a tree
3. a woman looking into a car after locking her keys inside
4. young woman with a broken car calling for help
5. breakdown of the car on the road

GT: the harbor a small village
Top-5 pseudo-targets:

1. the harbour with boats and houses
2. replica of the sailing ship in the harbour
3. ships in the harbor of the town
4. the harbor a small village
5. boats lined up alongside the geographical feature category in the village

Datasets and Implement Details

** Pre-training Datasets**

- Datasets Used:
  1. Conceptual Captions [4]
  2. SBU Captions [5]
  3. COCO [41]
  4. Visual Genome [42]
  5. Conceptual 12M dataset [43] (to demonstrate scalability with larger web data)

- Statistics:
  - Unique Images: 4.0M
  - Image-Text Pairs: 5.1M
  - Total Images after including Conceptual 12M dataset: 14.1M

Implementation Details

- Model Architecture:
  1. BERTbase: 123.7M parameters
  2. ViT-B/16: 85.8M parameters

Downstream task: image-text retrieval

- Finetune the pre-trained model using the ITC and ITM loss
- Two-step inference:
    1. Compute the feature similarity score Site for all image-text pairs (fast).
    2. For each query, select the top-k candidates and compute their image-text matching score s (slow)
    3. A small k can produce a high recall.

Method	# Pre-train Images	Flickr30K (1K test set)						MSCOCO (5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA	4M	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ALIGN	1.2B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8
ALBEF	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF	14M	**95.9**	**99.8**	**100.0**	**85.6**	**97.5**	**98.9**	**77.6**	**94.3**	**97.2**	**60.7**	**84.3**	**90.5**

Table 2: Fine-tuned image-text retrieval results on Flickr30K and COCO datasets.

Downstream task: zero-shot image-text retrieval

ALBEF outperforms CLIP and ALIGN which are trained on orders of magnitude larger datasets

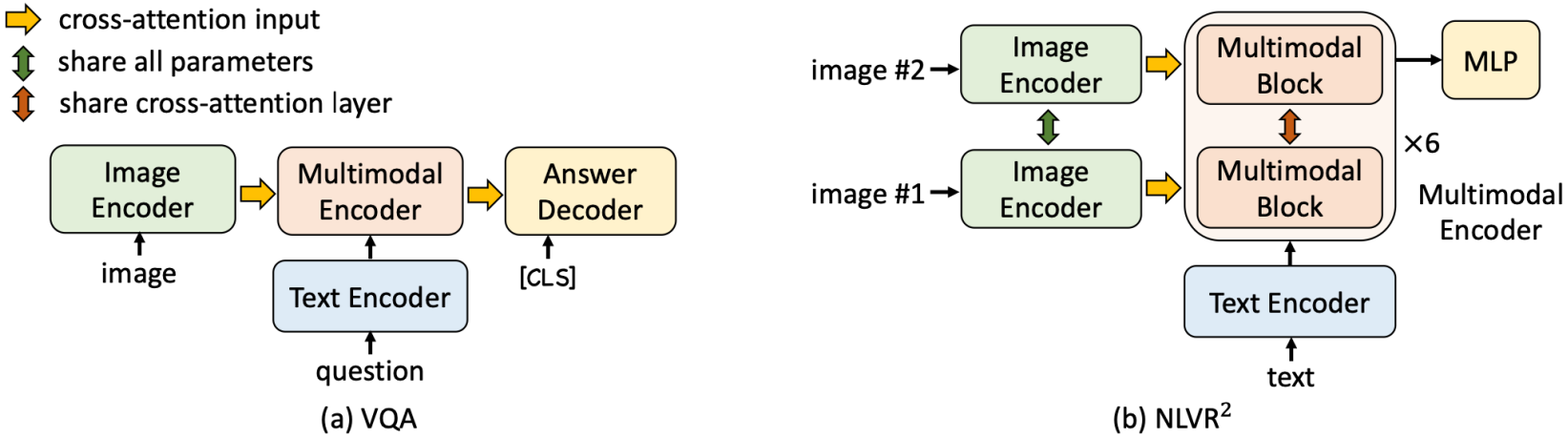
Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

Table 3: Zero-shot image-text retrieval results on Flickr30K.

Downstream task: Vision Language tasks

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

Table 4: Comparison with state-of-the-art methods on downstream vision-language tasks.



Downstream task: Vision Grounding

Method	Val	TestA	TestB
ARN [57]	32.78	34.35	32.13
CCL [58]	34.29	36.91	33.56
$\text{ALBEF}_{\text{itc}}$	51.58	60.09	40.19
$\text{ALBEF}_{\text{itm}}$	**58.46**	**65.89**	**46.25**

Table 5: Weakly-supervised visual grounding on RefCOCO+ [56] dataset.

Figure 4: Grad-CAM visualization on the cross-attention maps in the 3rd layer of the multimodal encoder.

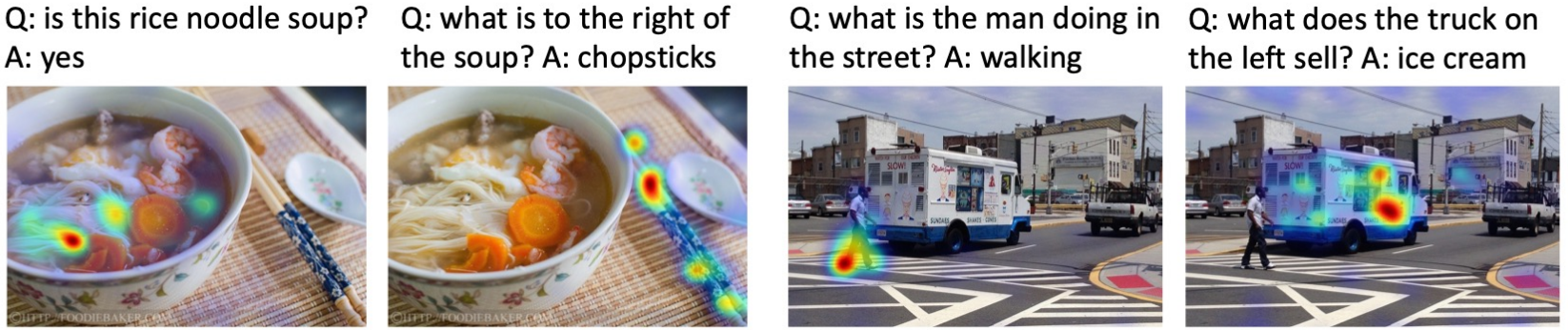


Figure 5: Grad-CAM visualizations on the cross-attention maps of the multimodal encoder for the VQA model.

Figure 6: Grad-CAM visualizations on the cross-attention maps corresponding to individual words.

QUIZ 1

In ALBEF, why does it make sense to align before merging?

Answer: In traditional methods, these two steps are often intertwined, which can lead to competition between alignment and fusion. To solve this problem, ALBEF proposes a new strategy, which is to align first before merging. In this way, models can be aligned in low-dimensional space and then fused, thus avoiding the competition between alignment and fusion.

QUIZ 2

Question: Key ideas of ALBEF (List at list 3Points to earn credits)