Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation (2021)

- Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, Steven C.H. Hoi

Presentation Author
Mathieu Ravaut

Nanyang Technological University

Table of content

Introduction

- Vision and Language Pre-training (VLP) aims to learn representations from image-text pairs (in an unsupervised manner).
- The goal is to improve downstream vision and language tasks : visual question answering (VQA), natural language for vision reasoning (NLVR), etc.

Introduction

Current challenges in Vision and Language :

- Image and text (token embeddings) features live in very **different spaces**. How to get a unique multimodal encoding ?
- In vision, the **object detector** is typically computationally intensive, and requires **annotations** (bounding boxes) which are expensive to collect, especially for a large-scale pre-training purpose.
- Datasets are collected from the web and are **noisy** : mismatch between the image content and the text description.

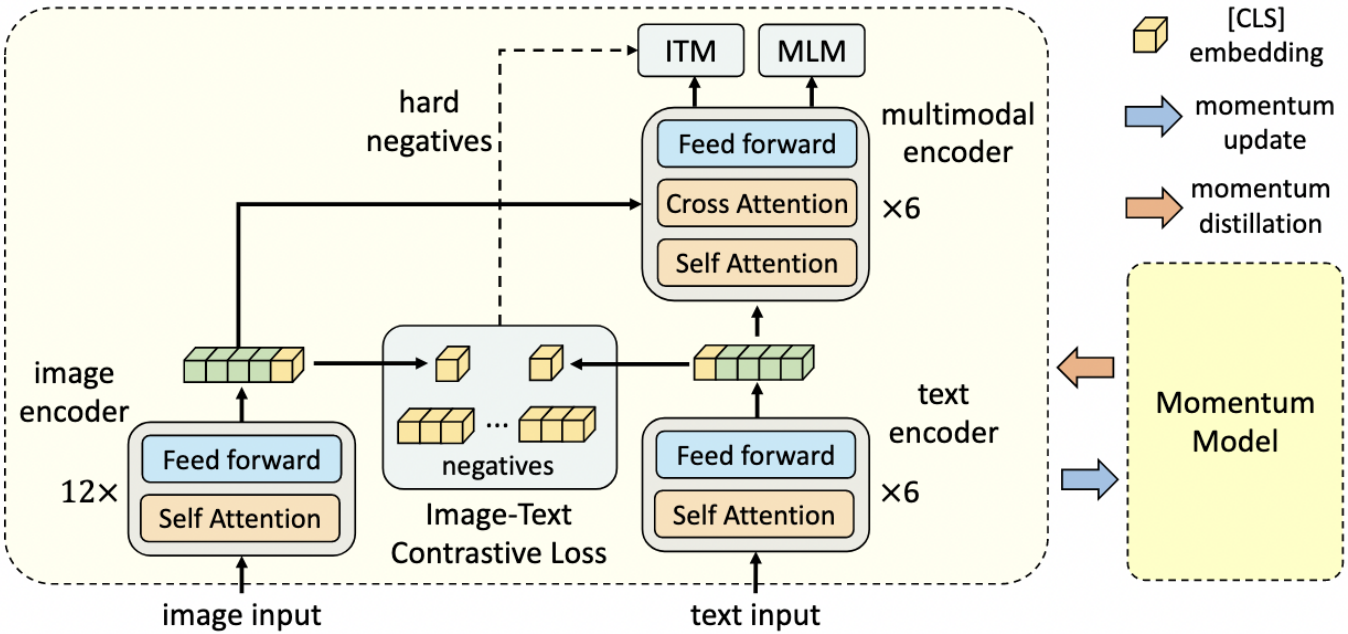
Approach

ALBEF : **A**lign **BE**Fore **F**use

- Encode image and text independently.
- Use a detector-free image encoder.
- Fuse image and text features with a multimodal encoder using cross-modal attention.
- Multiple training objectives :
  1. Image-text contrastive loss (ITC) on unimodal representations.
  2. Masked Language Modeling (MLM) with both image and text context.
  3. Image-Text Matching (ITM)
- Noisy supervision with Momentum Distillation.

Approach

Model architecture :



Approach

[1] **Image-Text Contrastive learning** (ITC) :

- Learn better quality unimodal (image, text) representations by **matching parallel image-text pairs**.
- MoCo approach : maintain 2 queues of size M, one for each encoder.
- $s(I,T) = g_v(\mathbf{v}_{cls})^T . g_w^{'}(\mathbf{w}_{cls}^{'})$, $s(T,I) = g_w(\mathbf{w}_{cls})^T . g_v^{'}(\mathbf{v}_{cls}^{'})$ where g_v, g_w are the vision, text encoders ; $g_w^{'}(\mathbf{w}_{cls}^{'})$, $g_v^{'}(\mathbf{v}_{cls}^{'})$ are the normalized CLS representations.
- Apply softmax, then for each data point (I,T), we have :

$$p_m^{i2t}(I) = \frac{e^{\frac{s(I,T_m)}{\tau}}}{\Sigma_{m=1}^{M} e^{\frac{s(I,T_m)}{\tau}}}, \; p_m^{t2i}(T) = \frac{e^{\frac{s(T,I_m)}{\tau}}}{\Sigma_{m=1}^{M} e^{\frac{s(T,I_m)}{\tau}}}$$

Approach

The final ITC loss becomes :

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

where H is the cross-entropy, $y^{i2t}(I)$ and $y^{t2i}(T)$ are the ground-truth one-hot similarities.

Approach

[2] **Masked Language Modeling** (MLM) :

- Mask out 15% of text tokens (like in BERT), and predict them with the other tokens **and the image as well**.
- The associated loss is :

$$\mathcal{L}_{mlm} = \mathbb{E}_{(I,\hat{T})\sim D} H(p^{msk}(I,\hat{T}), y^{msk})$$

where y^{msk} is the ground-truth one-hot vocabulary distribution.

Approach

[3] **Image-Text Matching** (ITM) :

- Predict whether an image-text pair is positive (matched) or negative (non-matched).
- The associated loss is :

$$\mathcal{L}_{itm} = \mathbb{E}_{(I,T)\sim D} H(p^{itm}(I,T), y^{itm})$$

where y^{itm} is the 2-D ground-truth one-hot vector.

- To get hard negatives (ITM_{hard}) : sample from the current mini batch, according to the contrastive similarity in [1].

Approach

Momentum distillation (MoD) :

- Positive pairs are only weakly correlated.
- However, the one-hot labels for ITC and MLM penalize all negative predictions regardless of correctness.
- Momentum model : exponential-moving-average of the unimodal and multimodal encoders.
- Constrain the base model to match predictions from the momentum model.

Approach

New ITC loss function :

- $s'(I,T) = g_v^{'}(\mathbf{v}_{cls}^{'})^T . g_w^{'}(\mathbf{w}_{cls}^{'})$, $s(T,I) = g_w^{'}(\mathbf{w}_{cls}^{'})^T . g_v^{'}(\mathbf{v}_{cls}^{'})$ where $g_v^{'}$, $g_w^{'}$ are the momentum model unimodal encoders.
- $q_m^{i2t}(I) = \frac{e^{\frac{s'(I,T_m)}{\tau}}}{\Sigma_{m=1}^{M} e^{\frac{s'(I,T_m)}{\tau}}}$, $q_m^{t2i}(T) = \frac{e^{\frac{s'(T,I_m)}{\tau}}}{\Sigma_{m=1}^{M} e^{\frac{s'(T,I_m)}{\tau}}}$
- New loss function :

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

Approach

Similarly, new MLM loss function :

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

Approach

The final training objective is :

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{itc} + \alpha\mathcal{L}_{itc}^{KL} + (1-\alpha)\mathcal{L}_{mlm} + \alpha\mathcal{L}_{mlm}^{KL} + \mathcal{L}_{itm}$$

Approach

ALBEF from a Mutual Information (MI) perspective :

- ALBEF maximizes a lower bound on the mutual information between different *views* of an image-text pair.
- ITC, MLM and MoD are mechanisms generating the different views.
- ITC and MLM generate views by taking partial information : modality separation (ITC), word masking (MLM).
- MoD performs data augmentation to the original views.

Approach

ITC can be re-written as :

$$\mathcal{L}_{\text{itc}} = -\frac{1}{2}\mathbb{E}_{p(I,T)}\big[\log \frac{\exp(s(I,T)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)} + \log \frac{\exp(s(T,I)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)}\big]$$

which is of the form of the NCE loss :

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E}_{p(a,b)}\left[\log \frac{\exp(s(a,b))}{\sum_{\hat{b}\in\hat{B}}\exp(s(a,\hat{b}))}\right]$$

which maximizes the MI between a and b.

Experiments

Implementation and training details :

- Text encoder : BERT-base (124M parameters).
- Image encoder : ViT-B/16 (86M parameters).
- 30 epochs, batch size 512, 8 NVIDIA A100 GPUs.
- Cosine schedule, with warmup from $lr = 1e-5$ to $lr = 1e-4$ for 1,000 iterations.
- In the contrastive learning : $m = 0.995$, $M = 65,536$ (maximum M size introduced in the MoCo paper).

Experiments

Pre-training data is built from several datasets :

- Conceptual Captions (web).
- SBU Captions (web).
- Conceptual 12M (web)
- COCO (in-domain).
- Visual Genome (in-domain).

Leading to a total of 14.1M images, 5.1M image-text pairs.

Experiments

To fine-tune on each downstream task :

- Use the task's loss.
- **Also use the KL-divergence between the model's predictions and the momentum model's predictions.** α is set to 0.4

Results

Image-to-text retrieval :

- **Tasks** : Retrieve the corresponding text/image from a given text/image.
- **Datasets** : Flickr30K and COCO.
- Zero-shot retrieval on Flickr30K using the model fine-tuned on COCO.
- Fine-tuning with the ITC and ITM losses.
- Since there are multiple ground truth texts for each image, relax the ground-truth label of ITC to have several positives.

Results

SOTA in zero-shot image-text retrieval on Flickr30K when pre-trained with just 4M images, further gains when pre-trained with 14M images.

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.

Results

SOTA in image-text retrieval on Flickr30K and COCO when pre-trained with 14M images.

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**

Results

Visual question answering (VQA) :

- **Task** : Predict an answer given an image and a question.
- **Datasets** : VQA.
- Add a 6-layer autoregressive answer decoder, initialized with the multimodal encoder weights.
- Only generate from the 3,192 answer candidates (for fair comparison).

Results

Natural language for visual reasoning (NLVR) :

- **Task** : Predict whether a text describes a pair of images.
- **Datasets** : NLVR.
- Extend the multimodal encoder to enable reasoning over 2 images.
- Append a MLP classifier on the multimodal encoder.
- Additional pre-training task : given a pair of images and a text, predict whether the text describes the first image, the second, or none of them.

Results

Visual entailment :

- **Task** : Predict whether the relationship between an image and a text is entailment, neutral, or contradictory.
- **Dataset** : SNLI-VE.
- Add a 3-way MLP classifier on top of the multimodal encoder.

Results

SOTA on VQA, NLVR and SNLI-VE :

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 [51]	-	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**

Results

Visual grounding :

- **Task** : Localize the region in an image related to the text description.
- **Dataset** : RefCOCO+.

Results

SOTA on visual grounding :

Method	Val	TestA	TestB
ARN [54]	32.78	34.35	32.13
CCL [55]	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**

Results

Ablation study :

#Pre-train Images	Training tasks	TR (flickr test)	IR (flickr test)	SNLI-VE (test)	NLVR^2 (test-P)	VQA (test-dev)
4M	MLM + ITM	93.96	88.55	77.06	77.51	71.40
	ITC + MLM + ITM	96.55	91.69	79.15	79.88	73.29
	ITC + MLM + ITM_{hard}	97.01	92.16	79.77	80.35	73.81
	ITC_{MoD} + MLM + ITM_{hard}	97.33	92.43	79.99	80.34	74.06
	Full (ITC_{MoD} + MLM_{MoD} + ITM_{hard})	97.47	92.58	80.12	80.44	74.42
	ALBEF (Full + $\text{MoD}_{\text{Downstream}}$)	97.83	92.65	80.30	80.50	74.54
14M	ALBEF	98.70	94.07	80.91	83.14	75.84

Conclusion

Takeaways :

- Contrastive learning is elegantly leveraged :
  - To align image-text representations (ITC).
  - To sample negatives for the Image-Text matching (ITM) pre-training obejctive.
  - Use the momentum model from MoCo to relax one-hot labels from ITC and MLM (momentum distillation).
- Very efficient pre-training objective, leading to new SOTA on a wide diversity of vision-text tasks (image-text retrieval, visual question answering, natural language for visual reasoning, visual grounding) and datasets (Flickr30K, COCO, VQA, NLVR, SNLI-VE, RefCOCO+).