ComputerVisionLab
Seoul National University

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Multi-Modal Representation Learning with Text-Driven Soft Masks

Jaeyoo Park
Seoul National University

Bohyung Han
Seoul National University

TUE-AM-266

Motivation

- Existing **vision-language pretraining** frameworks **rely only on image-caption pairs** with no fine-grained annotations.

- The model focuses on **the most discriminative regions**, and thus **lacks a comprehensive understanding of various attributes** observed in images despite detailed descriptions in captions.

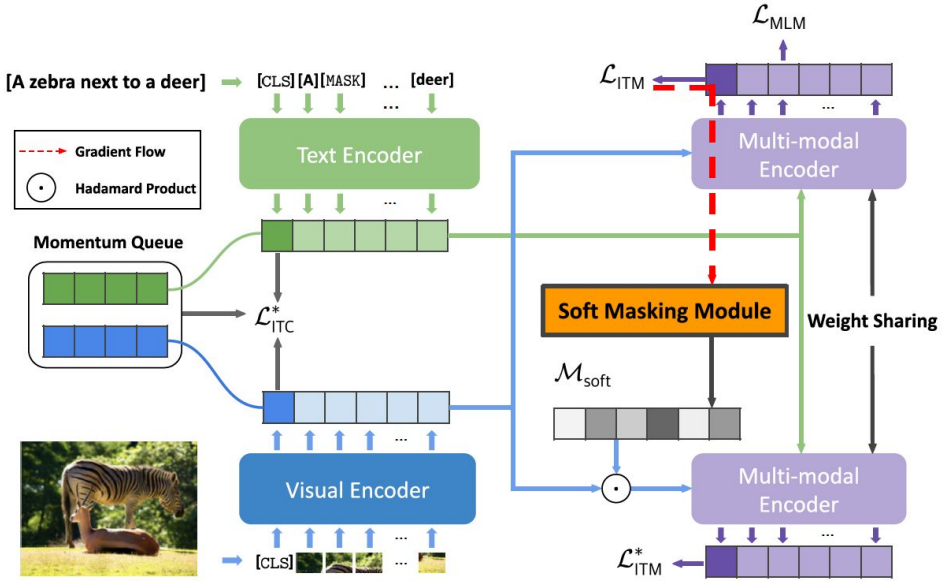
Contribution

- Propose a self-supervised visual-linguistic representation learning framework by introducing a new **operation**, **loss**, and **data augmentation** strategy to utilize **diversely augmented image-caption pairs**.

1. **Soft masking** technique on visual features based on **word-conditional Grad-CAM**
2. **Focal version of the ITC loss** to focus more on **hard but diverse** examples
3. **Multi-modal data augmentation strategies** for constructing more **diversified samples** by applying text maskings and rendering distortions on images

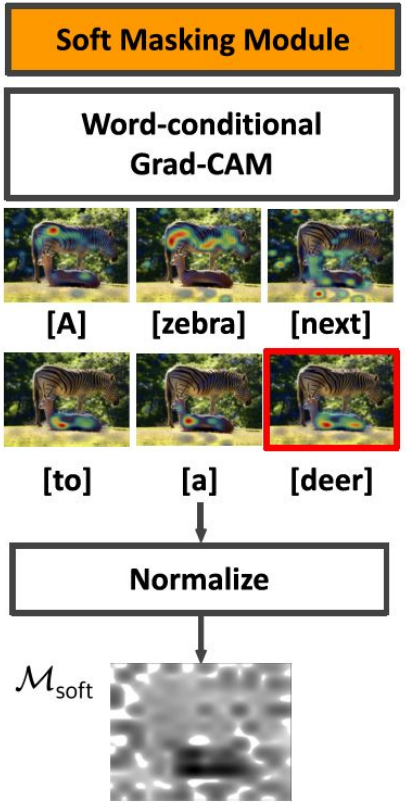
Preliminaries

- Pretraining Objectives
  - Image-Text Contrastive Loss (**ITC**)
  - Masked Language Modelling (**MLM**)
  - Image-Text Matching (**ITM**)
- Architecture
  - Visual Encoder : ViT-B/16
  - Text Encoder : First 6 layers of BERT
  - Multi-modal Encoder : Last 6 layers of BERT + Cross Attention



Text-Driven Soft Feature Masking

- Observation
  - Global-level Image-text matching without fine-grained annotation
    - lacks in local information
  - **Word-conditional Grad-CAM** effectively captures the **discriminative regions** (**objects**, in general)
    - Even for **stop-words**
- Goal
  - Augment hard positive examples by partly masking informative regions
  - To make the model focus on less attended regions



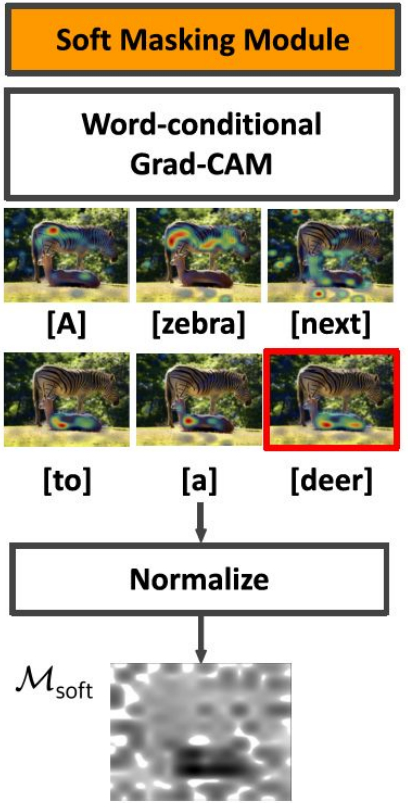
Text-Driven Soft Feature Masking

- Procedure
  - Compute **Grad-CAM** of the initial **ITM score** with respect to the **cross-attention map** of the image and text embeddings given by the multi-modal encoder

$$A_{\text{GCAM}}^{(i)} = \frac{1}{K}\sum_{k=1}^{K}\text{ReLU}\left(\frac{\partial q_{\text{ITM}}^{+(i)}}{\partial A_k^{(i)}} \odot A_k^{(i)}\right)$$

  - Randomly sample a single word index
    - To boost stochasticity

$$\mathcal{M}_{\text{soft}}^{(i)} = \mathbb{1} - \hat{A}_{\text{GCAM}}^{(i)}[i_w] \rightarrow \hat{\boldsymbol{V}}_{\text{emb}}^{(i)} = \mathcal{M}_{\text{soft}}^{(i)} \odot \boldsymbol{V}_{\text{emb}}^{(i)}$$



Focal Image-Text Contrastive Learning

- Boost the regularization effect from hard examples
- Large-scale image-caption corpora
  - Composed of **multiple datasets** with **large domain gaps**
  - The model **easily distinguishes** a lot of **samples from different datasets**
- Focal Loss
  - Alleviates the overfitting to easy examples while handling the class imbalance issue

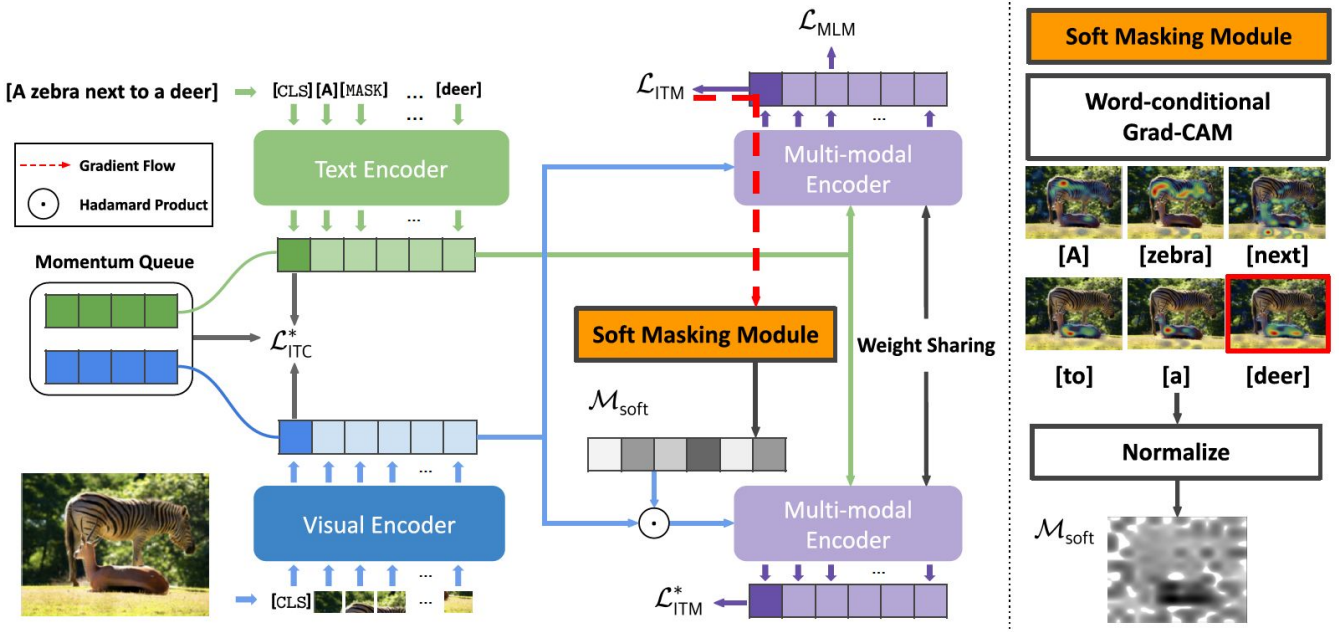
$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log p_{v2t}^{(i)} + \log p_{t2v}^{(i)}\right]$$

$$\mathcal{L}_{\text{ITC}}^{*} = -\frac{1}{2B}\sum_{i=1}^{B}\left[(1-p_{v2t}^{(i)})^{\gamma}\log p_{v2t}^{(i)} + (1-p_{t2v}^{(i)})^{\gamma}\log p_{t2v}^{(i)}\right]$$

Multi-Modal Data Augmentation

- Strong Augmentation on Image input
  - Color distortion
  - RandAugment
  - Gaussian blur
  - Random crop, resize
- Combine ITM + MLM
  - Reduce computation cost
  - Baselines (ALBEF, TCL, etc.)
    - Forward **twice**
      - Image-Text matching with **full text**
      - Mask language modeling
  - Ours
    - Forward **once**
      - Image-Text matching with **masked text** + Mask language modeling

Training Objective



$$\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{ITC}}^{*} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITM}}^{*}.$$

Experiments - Pretraining

- **Image-Text Pair Matching** of captioning datasets
- Datasets
  - COCO Captions
  - Visual Genome Dense Captions
  - SBU Captions
  - Conceptual Captions

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Experiments - Downstream Tasks

- Downstream Tasks
  - Image-Text Retrieval
    - Fine-tune
    - Zero-shot*
    - Ablation Study
  - Visual Question Answering*
  - Visual Reasoning*
  - Visual Entailment*

- * : Please refer to the paper

Experiments - Image-Text Retrieval

- Retrieve the most relevant caption from candidate images, or vice versa
  - Image-to-Text Retrieval (**TR**) / Text-to-Image Retrieval (**IR**)
  - Metric : **Recall**
- Datasets : MSCOCO Caption / Flicker30K

Method	#Img	Flickr30K (1K)						MS-COCO (5K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		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 [4]	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 [10]	4M	87.9	97.5	98.8	76.3	94.2	96.8	—	—	—	—	—	—
OSCAR [24]	4M	—	—	—	—	—	—	70.0	91.1	95.5	54.0	80.8	88.5
ViLT [17]	4M	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
UNIMO [23]	4M	89.7	98.4	99.1	74.7	93.5	96.1	—	—	—	—	—	—
SOHO [13]	200K	86.5	98.1	99.3	72.5	92.7	96.1	66.4	88.2	93.8	50.6	78.0	86.7
ALBEF [21]	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
TCL [43]	4M	94.9	99.5	99.8	84.0	96.7	**98.5**	75.6	92.8	**96.7**	59.0	83.2	89.9
CODIS [9]	4M	95.1	99.4	**99.9**	83.3	96.1	97.8	75.3	92.6	96.6	58.7	82.8	89.7
VinVL [45]	6M	—	—	—	—	—	—	75.4	92.9	96.2	58.8	83.5	90.3
SoftMask++ (ours)	4M	**95.4**	**99.7**	**99.9**	**84.6**	**96.8**	**98.5**	**76.6**	**93.5**	96.6	**60.2**	**83.7**	**90.5**
ALIGN [14]	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

Experiments - Image-Text Retrieval (Zero-shot)

- Retrieve the most relevant caption from candidate images, or vice versa
  - Image-to-Text Retrieval (**TR**) / Text-to-Image Retrieval (**IR**)
  - Metric : **Recall**
- Datasets : MSCOCO Caption / Flicker30K
- **CLIP & ALIGN**
  - Trained on orders of magnitude larger datasets

Method	#Img	Flickr30K (1K)						MS-COCO (5K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		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 [4]	4M	80.7	95.7	98.0	66.2	88.4	92.9	64.1	87.7	93.3	48.8	76.7	85.8
ViLT [17]	4M	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
CLIP [33]	400M	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [14]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
ALBEF [21]	4M	90.5	98.8	99.7	76.8	93.7	96.7	68.7	89.5	94.7	50.1	76.4	84.5
TCL [43]	4M	93.0	99.1	99.6	79.6	**95.1**	97.4	71.4	90.8	95.4	53.5	79.0	87.1
CODIS [9]	4M	91.7	**99.3**	**99.8**	79.7	94.8	97.3	71.5	91.1	95.5	53.9	79.5	87.1
SoftMask++ (ours)	4M	**93.4**	**99.3**	**99.8**	**80.1**	94.9	**97.7**	**72.3**	**91.5**	**95.7**	**54.1**	**79.8**	**87.3**

Experiments - Ablation Study

- Effectiveness of each component

SoftMask	Focal ITC	MMDA	TR@1	IR@1
			73.10	56.80
√			75.74	58.24
	√		75.66	58.85
		√	75.26	58.85
√	√		76.06	59.54
√		√	75.98	59.44
	√	√	76.16	59.62
√	√	√	**76.62**	**60.15**

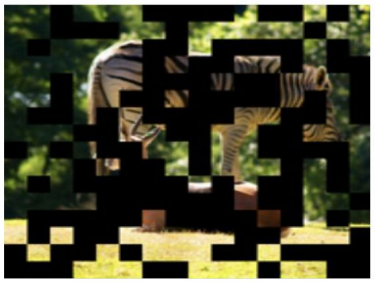
Experiments - Ablation Study

- Soft vs Random Masking

Method	TR@1	IR@1
RandMask ($p = 0.3$)	75.82	59.52
RandMask ($p = 0.5$)	76.22	59.74
SoftMask (Ours)	**76.62**	**60.15**

Text : A zebra next to a deer

(a) Our soft mask



(b) Random hard mask

- Advantages of SoftMask over RandMask
  - SoftMask can generate more diverse, and more importantly, semantically meaningful visual embeddings
  - SoftMask based on Grad-CAM does **not rely on additional hyperparameters** while RandMask often comes with a masking ratio.

Experiments - Ablation Study

- SoftMask for MLM
  - **MLM loss for the soft-masked features** does **not contribute** to the downstream task
  - Reconstruction task is not well addressed when the signals from both modalities are unable
  - **ITM** enjoys the regularization effect since ITM is the task based on **global information**

$\mathcal{L}^*_{\text{ITM}}$	$\mathcal{L}^*_{\text{MLM}}$	TR@1	IR@1
		76.16	59.62
√		**76.62**	**60.15**
√	√	76.21	59.84

Experiments - Other Downstream Tasks

- Visual Entailment (SNLI-VE)
  - To predict whether an image semantically entails a text
    - 3-way classification (entailment/contradiction/neutral)
- NLVR^2
  - To reason about whether the text is true for the pair of images
    - Duplicate encoder weights to receive additional input image
- VQA 2.0
  - To find an answer for a question about an input image
    - Add decoder to generate answer from pre-defined candidates

	Method	#Img	SNLI-VE		NLVR^2		VQA	
			val	test	dev	test-P	dev	std
O.D.	OSCAR [24]	4M	—	—	78.1	78.4	73.2	73.4
	UNITER [4]	4M	78.6	78.3	77.2	77.9	72.7	72.9
	UNIMO [23]	4M	80.0	79.1	—	—	73.3	74.0
	VILLA [10]	4M	79.5	79.0	78.4	79.3	73.6	73.7
	VinVL [45]	6M	—	—	82.1	83.1	75.9	76.1
D.F.	ViLT [17]	4M	—	—	75.7	76.1	71.3	—
	ALBEF [21]	4M	80.1	80.3	80.2	80.5	74.5	74.7
	TCL [43]	4M	80.5	80.3	80.5	81.3	74.9	74.9
	CODIS [9]	4M	80.5	80.4	80.5	80.8	**75.0**	74.9
	SoftMask++ (ours)	4M	**80.9**	**80.6**	**80.6**	**81.6**	**75.0**	**75.1**

Experiments - Computational Complexity

- Combining ITM forward path and MLM forward path
  - Reduces overall cost & improve performance

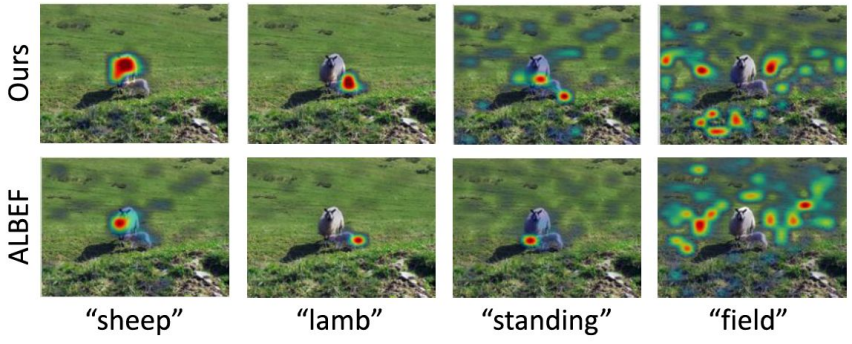
- Cost introduced by our SoftMask is negligible
  - Computing Grad-CAM & feed-forward through the multi-modal encoder

Table 6. Computation cost per a single pretraining iteration compared to ALBEF [19]. We measure computation cost on a single NVIDIA Quadro RTX GPU with 128 batch size per GPU.

Methods	Time (sec/it)	GPU Mem. (GB/GPU)
ALBEF [19]	1.94	41.5
Ours w/o SoftMask	1.91	38.7
Ours	1.98	41.4

Qualitative Results

“A dog on a beach carrying a frisbee”



“A sheep and lamb standing in the field”

- Our model provides more **accurate** and **wide** coverage of the **objects**
- Our model provides more **accurate** regions for the **action attributes**
- Failure cases
  - Our model may learn bias towards the object and scene
  - If the most discriminative parts is masked with high weights

Conclusion

- Propose a self-supervised visual-linguistic representation learning framework by introducing a new **operation**, **loss**, and **data augmentation** strategy to utilize **diversely augmented image-caption pairs**.

1. **Soft masking** technique on visual features based on **word-conditional Grad-CAM**
2. **Focal version of the ITC loss** to focus more on **hard but diverse** examples
3. **Multi-modal data augmentation strategies** for constructing more **diversified samples** by applying text maskings and rendering distortions on images