Unsupervised Vision-and-Language Pre-training via Retrieval-based Multi-Granular Alignment

Mingyang Zhou[1*] Licheng Yu[3*] Amanpreet Singh[3] Mengjiao Wang[3] Zhou Yu[2] Ning Zhang[3]
[1]Uiversity of California, Davis [2] Columbia University
[3]Meta AI
minzhou@ucdavis.edu, zy2461@columbia.edu, {lichengyu, asg, mengjiaow, ningzhang}@fb.com

A. Details of Motivation Study

As introduced in Section 1 , we try to answer two questions: (i) whether presenting a joint image-text data from non-parallel sources would improve the learned joint embedding space than alternatively presenting uni-modal data during pre-training. (ii) If we fed joint image-text data to the model, how does its existing latent alignment affect the cross-modal representation learning.

We conduct the unsupervised vision and language pre-training on Conceptual Captions (CC) by shuffling the image-text pairs. For pre-training objectives, we apply standard MLM + MRM. All other pre-training setup is the same as introduced in Section 4.3 . We first compare the round-robin and joint MLM + MRM pre-training, whose results are shown in Table 3. We then evaluate how the alignment degree of the pre-training dataset affects the model performance, where the degree is controlled by the ratio of originally aligned image-text data in Conceptual Captions. Table 4 shows the detailed results of each downstream task. Their Meta-Ave scores are also plotted in Fig. 1 . From these results, we obtained two important messages: (i) joint image-and-text input is more optimal for UVLP than alternatively presenting uni-modal data from unparallel image and text corpus. (ii) The more the latent semantic alignment exists in the image-text data the better the pre-trained model performs.

We further explore the realistic unsupervised V+L pre-training, where the images and texts are from two different sources. Specifically, we sample the images from Conceptual Captions and the texts from Book Corpus respectively. Table 1 shows that the pre-trained model on our weakly aligned CC image and BC sentence corpus far outperforms that on random pairs, indicating it also holds that better latent image-text alignment leads to better pre-trained model's performance under realistic setting.

*The two authors contribute equally.

	VQA2 Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+ Devs	Meta-Ave
random	70.3	51.2	75.3	76.5	68.3
proposed	**71.2**	**67.1**	**77.1**	**79.7**	**73.8**

Table 1. Pre-training on realistic CC + BC data

B. Effectiveness of Weighted ITM

We compared the performance of pre-training our model with or without weighted ITM. The models are pre-trained on CC images and texts. As shown in Table 2, weighted ITM are consistently better than treating all the retrieved pairs with the same weight.

	VQA2 Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+ Devs	Meta-Ave
w/o w_{ITM}	71.9	72.6	77.0	79.7	75.3
w_{ITM}	**72.1**	**73.4**	**77.3**	**80.3**	**75.8**

Table 2. Ablation Study on weighted ITM

C. Downstream Task Details

We describe the details of fine-tuning on the four different downstream tasks: Visual Question Answering (VQA2), Natural Language for Visual Reasoning (NLVR2), Visual Entailment (VE), and Referring Expression (RefCOCO+). We mainly follow the setup of UNITER [1] for each downstream task with minor adjustments.

VQA2 Given a question about an image, the task is to predict the answer to the question. Following [6], we take 3,129 most frequent answers as answer candidates. We use both training and validation sets from VQA 2.0 for fine-tuning. Following UNITER, we also leverage data from Visual Genome [2] to augment the best performance on the test-dev split. We fine-tune the model with a binary cross-entropy loss with a peak learning rate of 6×10^{-5} for 20 epochs. The training batch size is set as 480.

NLVR2 NLVR2 is a task for visual reasoning. The objective is to determine whether a natural language statement

Pre-training	VQA2 Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+ Dev	RefCOCO+ TestA	RefCOCO+ TestB	Meta-Ave
Round-Robin MLM+MRM	70.4	51.1	74.8	73.3	78.3	**67.4**	67.4
Joint MLM+MRM	**70.6**	**52.4**	**74.9**	**74.5**	**79.4**	66.8	**68.1**

Table 3. Detailed evaluation results on four V+L downstream tasks with two different data feeding strategy for UVLP: (1) joint image-text data (joint MLM+MRM); (2) alternative uni-modal data (round-robin MLM+MRM).

Paired Ratio	VQA2 Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+ Dev	RefCOCO+ TestA	RefCOCO+ TestB	Meta-Ave
0%	70.6	52.4	74.9	74.5	79.4	66.8	68.1
20%	71.1	70.0	76.4	76.3	80.3	67.5	73.5
40%	71.4	71.6	77.2	77.9	82.4	68.8	74.5
60%	71.9	74.5	77.8	79.9	84.4	69.9	76.0
80%	72.2	75.7	78.4	80.9	85.7	71.8	76.8
100%	72.5	75.9	78.7	82.1	86.6	75.0	77.3

Table 4. Detailed evaluation results on four V+L downstream tasks with 6 sets of image and text corpus of different latent cross-modal alignment degree. The alignment degree is controlled by changing the ratio of original aligned image-text data from 0% to 100%.

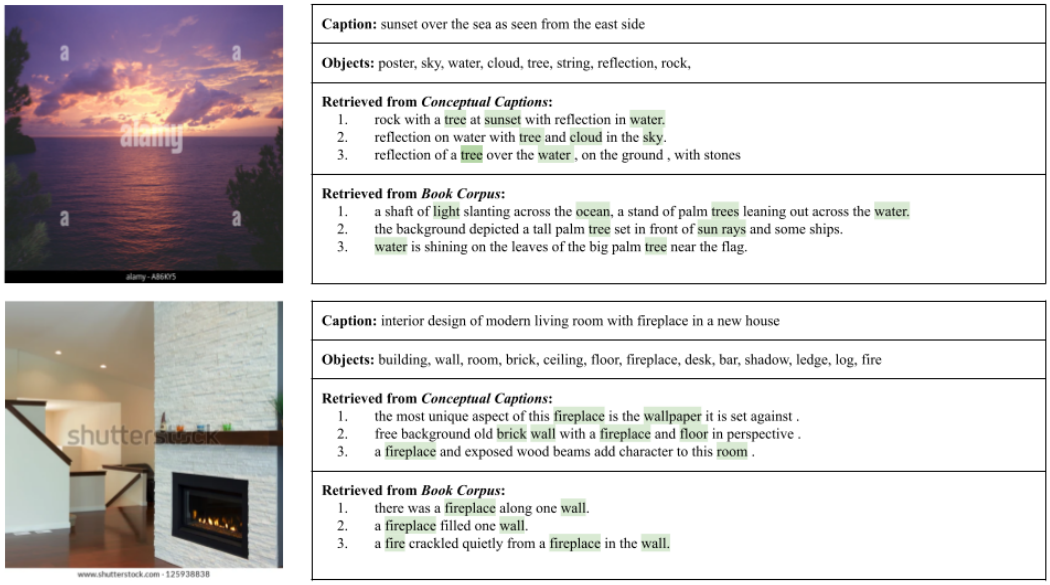

Figure 1. Examples of retrieved text from both CC and BC. The covered grounded noun phrases in retrieved sentences are highlighted in green bar for positive examples.

is true or not given a pair of input images. We follow UNITER's setup treating each data point as two text-image pairs with repeated text. The two [CLS] outputs from the model are then concatenated as the joint embedding for the example. We apply a multi-layer perceptron (MLP) classifier on top of this joint embedding for the final classification. Unlike [3] that conducts additional "pre-training" on NLVR2 datasets, we only fine-tune the model with the task-specific objective to maintain the same setting as all other downstream tasks. We train the model for 8 epochs with a batch size of 60 and a peak learning rate of 3×10^{-5}.

VE Visual Entailment is a task built on Flickr30k Images [4], where the goal is to determine the logical relationship between a natural language statement and an image. Similar to the Natural Language Inference problem in NLP, this task is formatted as a 3-way classification problem to predict if the language statement entails, contradicts, or is undetermined with respect to the given image. An MLP transformer classifier is applied to the output of the [CLS] token to make the final prediction. The model is fine-tuned using cross-entropy loss. We set the batch size as 480 and the peak learning rate as 8×10^{-5}. The model is fine-tuned for 4 epochs for this downstream task.

RefCOCO+ The referring expression task involves locating

an image region given a natural language phrase. We use RefCOCO+ [5] as the evaluation dataset. Bounding box proposals from VinVL object detectors are used for fine-tuning. A proposal box is considered correct if it has an IoU with a gold box larger than 0.5. We add an MLP layer on top of the region outputs from the Transformer to compute the alignment score between the language phrase and each proposed region. We fine-tune our model for 20 epochs with a peak-learning rate of 2×10^{-4}.

D. Additional Visualization

We present additional examples of retrieved text from both CC and BookCorpus. Specifically, we demonstrate more positive examples in Fig 1 that covers the appropriate grounded noun phrases. We also share some negative examples in Fig 1. As analyzed in the limitation section, the current language embedding model weighs all the object tags equally to generate the joint embedding representation. This can lead to mistakenly focused object tags when retrieving the text. In row 1 of Fig 2, texts retrieved cover less important noun phrases such as "finger" and "hair" instead of the more important noun phrase "baby". Row 2 of Fig 2 demonstrate mistakenly retrieved texts due to the limitation of the pre-defined object categories in the object detector. In this example, the students in the image are detected as "person" or "man", which leads to the failure of retrieving any valid text.

We also demonstrate more examples on text-to-image attention between the pre-trained U-VisualBert and μ-VLA on the Conceptual Captions Validation set in Fig 3, 4, 5, 6. These examples provide additional evidence on the better local alignment captured by μ-VLA.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 1

[3] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *NACCL*, 2021. 2

[4] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[5] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 3

[6] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 1

Caption: small baby girl is given her first solid food on a spoon and is unsure about the process

Objects: baby, wire, head, blanket, bib, hair, fork, hand, face, finger, spoon, design, cheek, nose, ear, mouth, handle

Retrieved from *Conceptual Captions*:
1. baby feeding spoon handmade in yellow birch small smooth bowl and tapered handle carved with an oval ring
2. finger in the nose with removable fur collar - listing
3. close up of a baby doll s head showing the ear

Retrieved from *Book Corpus*:
1. the man's face was highlighted as young man somewhere in his thirties with glasses and a mustache.
2. a pair of thick glasses hung crookedly on the man's ears and a few tufts of brown hair jutted out from beneath a stocking cap that matched his shirt.
3. the man, with his greying hair, kept reaching up with the index finger on his right hand.

Caption: students have occupied the building of the academy

Objects: building, wall, chair, sign, line, floor, person, writing, door, man, letter

Retrieved from *Conceptual Captions*:
1. a hand - painted sign by person decorates the door .
2. person puts a letter through door .
3. a sign on a door in the basement of person

Retrieved from *Book Corpus*:
1. a man with a shaved head sat behind the reception desk.
2. the man waiting in my office wore an unbuttoned black herringbone suit and a striped red tie hanging out to one side against a light-blue oxford shirt.
3. i look up to see a pudgy middle-aged man with balding black hair standing near the entrance of the office.

Figure 2. Examples of retrieved text from both CC and BC. The mistakenly covered grounded noun phrases in retrieved sentences are highlighted in red bar for negative examples.

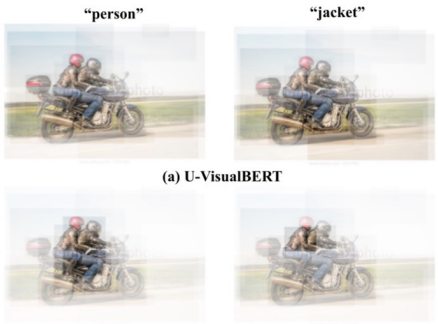

Figure 3. Text-to-image attention given the aligned pair whose caption is "person in a leather jacket riding a motorcycle on the road".

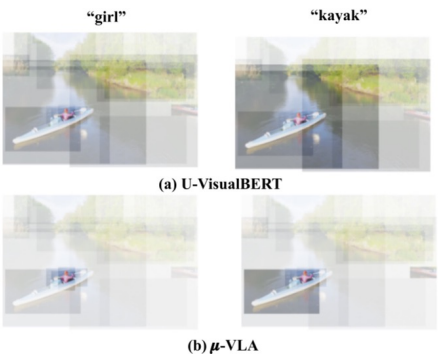

Figure 4. Text-to-image attention given the aligned pair whose caption is "girl in a blue kayak floating on the picturesque river at sunset".

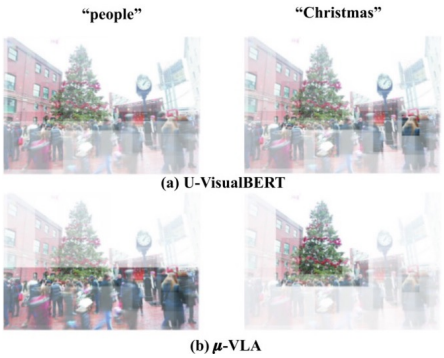

Figure 5. Text-to-image attention given the aligned pair whose caption is "people walking by the christmas tree and stage area".

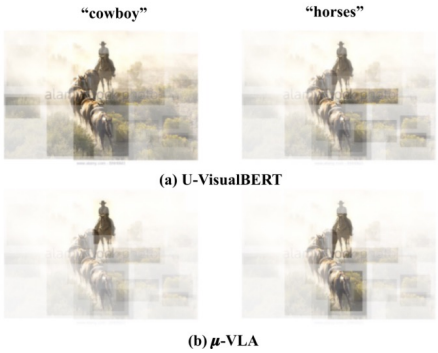

Figure 6. Text-to-image attention given the aligned pair whose caption is "single cowboy guiding a line of horses through the desert".