Supplementary Materials of Paper 5239

Anonymous Author(s)
Affiliation
Address
email

1 Broader Impact

Vision-Language Pre-training (VLP) serves for a lot of downstream vision-language tasks like visual question answering, visual reasoning, visual entailment. By extensive experiments, we show the importance of inter-modality interaction and achieve competitive performance by applying Transformer for visual embedding. Our study can benefit future researches from providing a view of designing model for VLP. By revealing the fusion mechanism of multi-modality, our work may also benefit other multi-modal tasks besides vision-language task.

At the same time, Vision-Language Pre-training may learn biased or offensive content from unsupervised image-text pairs. This may cause improper understanding of images. More work is needed to automatically filter data for pre-training.

2 Image-Text Retrieval

In this paper we focus on tasks related to visual relation understanding and inter-modal reasoning: Visual Question Answering (VQA), natural language for visual reasoning (NLVR), and fine-grained visual reasoning (Visual Entailment). We also show results on Image-Text Retrieval task. Image-text retrieval aims to retrieve the most relevant text from candidate images, or vice versa. Image-text retrieval includes two sub-tasks of image-to-text retrieval (TR) and text-to-image retrieval (IR). We follow the same practice as SOHO [4] to conduct image-text retrieval for fair comparisons. During training, we construct image-text pairs in a mini-batch by sampling aligned pairs from ground-truth annotations, and unaligned pairs from other captions within the mini-batch. To predict whether an image-text pair is aligned or not, we use the joint embedding representation of the [CLS] token from Transformers to perform binary classification. Since the binary classification objective of image-text retrieval model is consistent with the image-text matching (ITM) task in pre-training stage, we initialize the task-specific head from the pre-trained ITM head for better initialization. We adopt AdamW optimizer with a learning rate of 5e-5. The mini-batch size is set to 32. We train 10 epochs until convergence and decay the learning rate by half at 5th epoch empirically.

Experiment results on Flickr30k [12] are shown in Table 1. Our model outperforms ViLT and SOHO under all metrics on Flickr30k. The promising results of our model on image-text retrieval indicate the advantage of our fully Transformer architecture for learning cross-modal alignment.

3 Dataset Statistics

We summarizes the statistics of all our pre-training and downstream tasks in Table 2.

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 *European Conference*

Table 1: Evaluation of image-to-text retrieval (TR) and text-to-image retrieval (IR) on Flickr30K dataset. "-" indicates the detail is not reported.

Method		VSE++[2]	SCAN[8]	ViLBERT[11]	Unicoder-VL[9]	UNITER[1]	ViLT[6]	SOHO[4]	Ours
	R@1	52.9	67.4	-	86.2	85.9	83.7	86.5	**87.0**
TR	R@5	80.5	90.3	-	96.3	97.1	97.2	98.1	**98.4**
	R@10	87.2	95.8	-	99.0	98.8	98.1	99.3	**99.5**
	R@1	39.6	48.6	58.2	71.5	72.5	62.2	72.5	**73.5**
IR	R@5	70.1	77.7	84.9	90.9	92.4	87.6	92.7	**93.1**
	R@10	79.5	85.2	91.5	94.9	96.1	93.2	96.1	**96.4**

Table 2: Statistics of different tasks. Notation “*” denotes Karpathy split [5]. Notation “-” denotes not applicable.

Task	**Dataset**	**Train Split**	**Test Split**	**Metric**
Pre-training	VG [7]	train	-	-
	MSOCO [10]	train+restval*	-	-
Image-Text Retrieval	Flickr30K [12]	train	test*	Recall@1,5,10
Visual Question Answering	VQA2.0 [3]	train+val	test-dev/test-std	VQA-score [3]
Visual Reasoning	$NLVR^2$ [13]	train	dev/test-P	Top-1 Accuracy
Visual Entailment	SNLI-VE [14]	train	val/test	Top-1 Accuracy

on Computer Vision, pages 104–120. Springer, 2020.

[2] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[4] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. *arXiv preprint arXiv:2104.03135*, 2021.

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[6] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[8] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[9] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2019.

[12] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[13] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

[14] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018.