Supplementary Material for ACMMM 2022 Conference Paper2506

Anonymous Author(s)

1 DATASET STATISTICS

We construct our pretraining dataset with four public datasets including Conceptual Captions 3M[9], SBU Captions[8], COCO[7], and Visual Genome[4]. These data contain 4M images in total. Detailed dataset statistics are available in Table 1.

Dataset	Images	Texts
Conceptual Caption 3M [9]	2.97M	2.97M
SBU Caption[8]	859K	859K
COCO[7]	113K	567K
Visual Genome [4]	108K	5.41M

Table 1: Statistics of datasets for pretraining.

2 EXPERIMENTS ON IMAGE-TEXT CLASSIFICATION

To evaluate the effectiveness and generalization of our proposed method, we finetune our model on two typical image-text classification tasks: Visual Question Answering(VQA) and Neutral language for Visual Reasoning(NLVR)[10].

Visual Question Answering requires the model to predict the answer according to the given image and question. We evaluate our model on VQAv2[1] dataset. In practice, it is common to view this task as an image-text classification task with 3129 answer classes, so are we. As we finetuned our model on the training set and validation set following previous work, then we submit the results referenced on the test set to the evaluation server [1] to get the final score.

Neural Language for Visual Reasoning is a binary classification task. Given a triplet of two images and one question, a common way is to reformulate the triplet input into two pairs, each pair consisting of a different image but the same question text. both of the pairs will be fed into our model and get the representations, the classification head takes the concatenation of the representations of the two pairs and outputs the classification results.

Note that, during the pre-training process, the number of VCM decoder layers is set as 6, and we finetune our pre-trained model on VAQv2 and NLVR2 dataset for 10 epochs using a batchsize of 64, respectively. The learning rate of the unimodal encoders is 10^{-6}, the learning rate of other parts of the model is five times of the unimodal encoders'.

Table 2 presents our results on the above two tasks. The results of methods which using region-based visual features are listed in the upper half of the table and the results of patch-feature-based methods are listed on the bottom half of the table. Compared with the previous methods, our model can always achieve good performance with the absolute improvement of 1 point of VQA score. As for the visual reasoning task, our model also achieve the best performance compared with other patch-feature-based models. All the results have demonstrated the effectiveness and generalization of our model.

	Time		VQAv2		NLVR2	
Models	ViLT's	Ours	test-dev	test-std	dev	test-P
UNITER$_B$[2]	900ms	-	72.70	72.91	77.18	77.85
UNITER$_L$[2]	-	-	73.82	74.02	79.12	79.98
UNIMO$_L$[6]	-	-	75.06	75.27	-	-
VinVL$_B$[12]	650ms	-	75.95	76.12	82.05	<u>83.08</u>
VinVL$_L$[12]	-	-	<u>76.52</u>	<u>76.60</u>	**82.67**	**83.98**
ViLT[3]	15ms	28ms	71.26	-	75.70	76.13
VisualParsing[11]	-	-	74.00	74.17	77.61	78.05
ALBEF-4M[5]	-	52ms	74.54	74.70	80.24	80.50
Ours	-	53ms	**77.67**	**77.79**	<u>82.33</u>	<u>83.08</u>

Table 2: Comparison with existing VLP methods on VQAv2, NLVR2. The best scores are in bold, and the second-best scores are <u>underlined</u>. We also report the VQA inference time mersured by ViLT and in our hardware environment setting

3 HYPERPARAMETER STUDY ON THE NUMBER OF VCM DECODER LAYERS

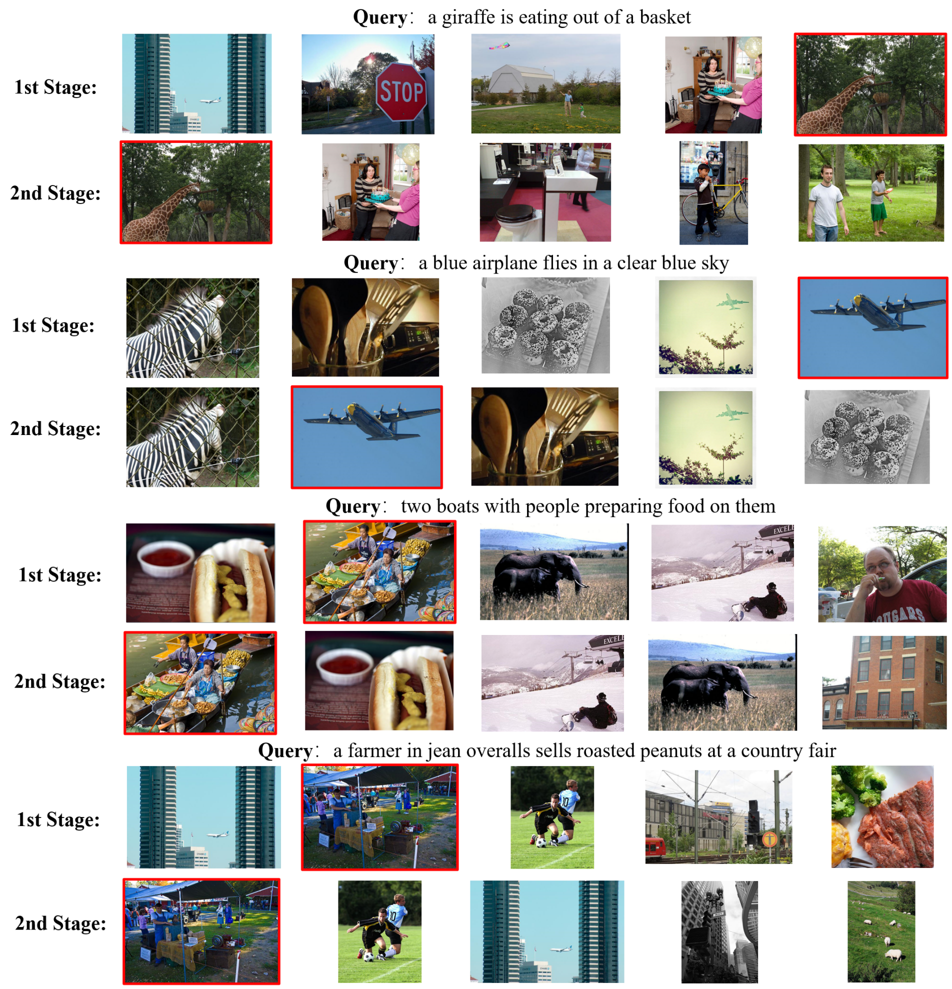
We use different numbers of the VCM decoder layers during pretraining to study the effects of the decoder layers on image-text retrieval and VQA tasks. We finetune our pretrained model on Flick30K dataset when performing the image-text retrieval task. The zero VCM decoder layer means we don't use an additional decoder. When performing VCM task, the multi-modal encoder will be used as the VCM decoder, which will take the visible embedded patches and the masked tokens as input and restore the masked parts of the images.

	VCM decoder	Image-to-Text		Text-to-Image		VQAv2
	layers	R@1	R@5	R@1	R@5	test-dev
w/o VCM	-	96.2	99.9	85.4	97.5	77.24
with VCM	0	96.1	99.9	84.7	97.0	77.00
with VCM	4	96.0	100.0	86.2	97.6	77.39
with VCM	6	96.2	**100.0**	**86.4**	**97.7**	**77.67**
with VCM	8	**96.8**	**100.0**	86.3	**97.7**	77.45

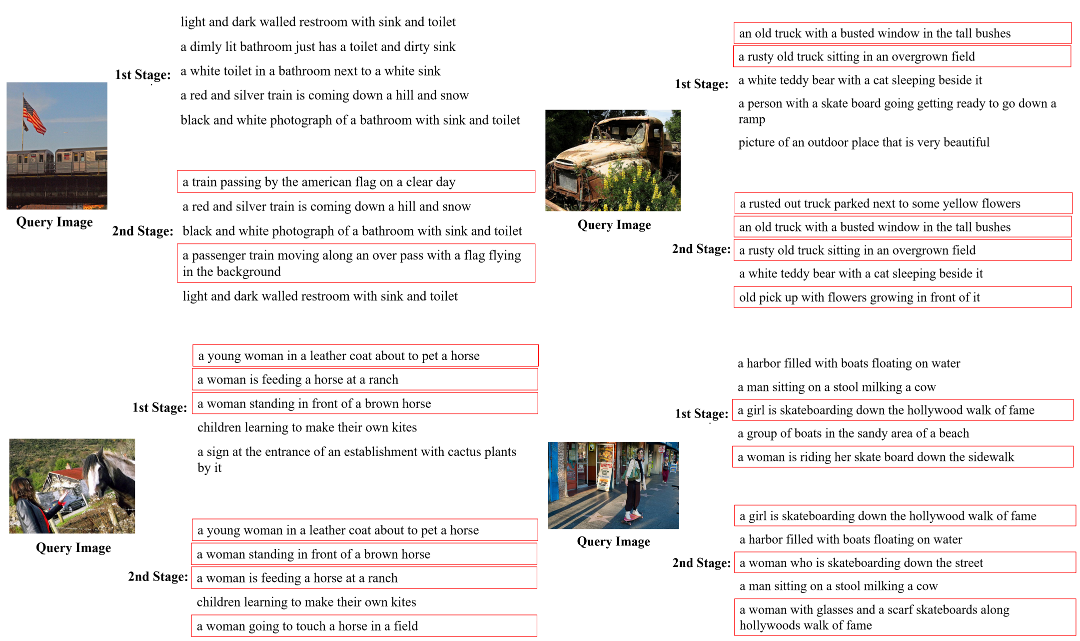
Table 3: Comparisons of pretrain models with different VCM decoder layer number on Flickr30K and VQAv2

We present the finetuned results in Table 3. If we don't apply an additional VCM decoder, the performance of the model on downstream tasks will drop, especially on text-to-image retrieval tasks. As we use more layers, the model becomes better for image-text retrieval tasks. Due to the limited GPU memory, we cannot use more decoder layers than eight layers. As for VQA task, the model achieves the best performance when we use a six-layer VCM decoder, more decoder layers are not helpful for the model.

[1] https://eval.ai/web/challenges/challenge-page/830/overview



(a) text-to-image retrieval



(b) image-to-text retrieval

Figure 1: Visual comparisons of image-text retrieval examples between each stage on Flickr30K dataset, we provide the top-5 results of each stage in our inference process. The results in red boxes are the ground truth.

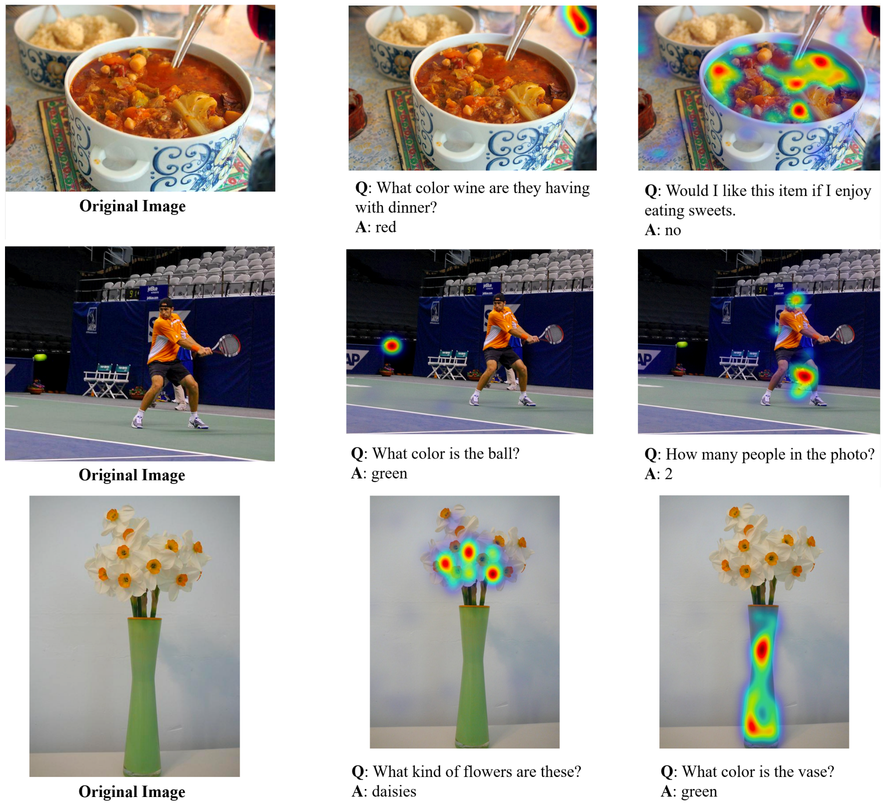


Figure 2: Grad-CAM heatmaps computed on the cross-attention maps in the last 3rd layer of the multi-model encoder for VQA model.

4 VISUALIZATIONS

We provide more image-text retrieval examples which are coming from Flick30K dataset for visual comparison in Figure 1. We also provide the Grad-CAM heatmaps of the VQA model in Figure 2, the heatmaps are computed on the cross-attention maps in the last 3rd layer of the multi-model encoder. As we can see, our model clearly understands the input question and is able to focus on the part of the image that is associated with the answer.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2425–2433.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings,*. Springer, 104–120.

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. PMLR, 5583–5594.

[4] 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. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.

[5] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*.

[6] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 2592–2607.

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Springer, 740–755.

[8] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc.

[9] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2556–2565.

[10] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 6418–6428.

[11] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing Inter-modality: Visual Parsing with Self-Attention for Vision-Language Pre-training. In *NeurIPS*.

[12] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5579–5588.