Supplementary Material: Adaptive Fine-tuning for Vision and Language Pre-trained Models

Shentong Mo[†,*], **Jingfei Xia**[†], **Ihor Markevych**
Carnegie Mellon University
Pittsburgh, PA 15213
shentonm,jingfeix,imarkevy@andrew.cmu.edu

1 More Experimental Results

Visual Question Answering (VQA). In the VQA task, we follow the experimental protocol in BUTD [1], aim to answer a question at the perceptual level given a natural image by choosing the correct answer from a shared set composed of 3,129 answers. Specifically, we conduct experiments on the VQA v2.0 dataset [3] based on the images from COCO [5] dataset. We split the dataset into train set (83k images and 444k questions), validation set (41k images and 214k questions), and test set (81k images and 448k questions). We report the results in Table 1. Compared to previous methods, our AFVL achieves better performance in terms of accuracy on both test-dev and test-std datasets. This infers that the pair-wise contrastive loss between visual and linguistic representations is beneficial for learning alignments between the whole sentence and each image.

Table 1: Comparison results on the VQA and VCR datasets.

Model	VQA		VCR					
	test-dev	test-std	Q $\rightarrow A$		QA $\rightarrow R$		Q $\rightarrow AR$	
			val	test	val	test	val	test
LXMERT [9]	72.42	72.54	-	-	-	-	-	-
ViLBERT [6]	70.55	70.92	72.40	73.30	74.50	74.60	54.00	54.80
VisualBERT [4]	70.08	71.00	70.80	71.60	73.20	73.20	52.20	52.40
VL-BERT [8]	71.72	72.18	73.80	-	74.40	-	55.20	-
UNITER [2]	72.27	72.46	-	75.00	-	77.20	-	58.20
CVLP [7]	72.77	72.90	-	-	-	-	-	-
AFVL (ours)	**72.83**	**73.05**	**75.33**	**75.65**	**76.52**	**77.87**	**58.65**	**59.47**

Visual Commonsense Reasoning (VCR). In the VCR task, we need to select the right answer to the given question and provide the rationale explanation for a higher-level cognitive and commonsense understanding of the given image. In the experiments, we use an image and a list of categorized ROIs from the VCR dataset [10] to pick the correct one from 4 candidate answers and 4 candidate rationales, respectively. The task (Q $\rightarrow$ AR) can be split into two sub-tasks: question answering (Q$\rightarrow$A) and answer justification (QA$\rightarrow$R). We also split the VCR dataset into training (213k questions and 80k images), validation (27k questions and 10k images), and test (25k questions and 10k images) sets. The results are reported in Table 1. Our AFVL achieves competitive performance, although we do not use the larger Conceptual Captions dataset in VL-BERT$_{\text{large}}$. This implies that the pair-wise contrastive loss proposed at the pre-training stage is beneficial to eliminate the semantic confusion between vision and language. VL-BERT$_{\text{large}}$ also validates the importance of pre-training on a massive-scale dataset to improve the model's capacity.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Table 2: Ablation study on pair-wise contrastive pre-training, adaptive fine-tuning and batch size. MLM, NSP, PwCL, and AF denote Masked Language Modeling, Next Sentence Prediction, Pair-wise Contrastive Loss and Adaptive Fine-tuning.

MLM	NSP	PwCL	AF	batch size	test-dev (↑)	test-std (↑)	APS (↑)
✓	✓			64	70.11±0.12	71.03±0.15	0.43±0.08
✓		✓		64	70.82±0.13	71.56±0.15	0.52±0.06
	✓	✓		64	70.06±0.12	70.95±0.13	0.41±0.06
✓	✓	✓		64	71.73±0.15	72.08±0.17	0.68±0.04
✓	✓	✓	✓	64	71.68±0.14	72.02±0.16	0.67±0.05
✓	✓	✓	✓	128	72.18±0.13	72.32±0.16	0.72±0.04
✓	✓	✓	✓	256	72.42±0.11	72.67±0.13	0.76±0.03
✓	✓	✓	✓	512	72.83±0.05	**73.05±0.07**	**0.83±0.02**
✓	✓	✓	✓	1024	**72.88±0.08**	72.96±0.06	0.81±0.02

2 Ablation Studies

Effect of each module and batch size. In Table 2, we explore the effect of each pre-training task proposed in our AFVL, which consists of Masked Language Modeling, Next Sentence Prediction, Pair-wise Contrastive Loss, and Adaptive Fine-tuning. We can observe that with the incorporation of the pair-wise contrastive loss, our AFVL achieves better performance than the baseline without the pair-wise contrastive loss between linguistic and visual embeddings. This demonstrates the effectiveness of the pair-wise contrastive loss proposed in our AFVL. Our AFVL with adaptive fine-tuning achieves comparable performance while achieving fewer training parameters and saving computation resources.

We also evaluate the effect of the batch size on the final performance of our AFVL pre-trained models in Table 2. As can be seen, our AFVL performs the best APS at the batch size of 512, which shows the importance of the choice of the batch size in the pair-wise contrastive loss. Adding PwCL to baselines with the same batch size increases the accuracy from 70.11 to 71.73, where the improvement (1.62) is significant in the VQA task. Increasing the batch size from 64 to 1024 can boost the accuracy from 71.68 to 72.83 (1.15), but the improvement is smaller than that of PwCL. This also conforms to the importance of larger batch size in vision-language pre-training by introducing more negative pairs. We also show the big improvements of our AFVL in VCR (3.45), NLVR2 (11.31), and RPG (10.43).

Furthermore, we compare the Averaged Pair-wise Similarity (APS) between embeddings of each text-image pair during pre-training in Table 2. Specifically, we calculate the pair-wise dot product between linguistic embeddings $\mathbf{E}'_i$ and visual embeddings $\mathbf{F}'_i$, *i.e.*, $\frac{1}{B}\sum_{i=1}^{B}(\mathbf{E}'_i \cdot \mathbf{F}'_i)$. We can observe that our AFVL with the added pair-wise contrastive loss help increase the APS between each text-image pair, which verifies the effectiveness of our pair-wise contrastive pre-training in mitigating the semantic confusion between visual and linguistic embeddings during the pre-training process.

3 Visualizations

In this appendix, we also visualize the pre-trained image and text pairs to validate the effectiveness of our AFVL in mitigating the semantic confusion between those representations. Specifically, we calculate the pair-wise similarity across pre-trained image and text pairs and report them in Figure 1. As can be seen, our AFVL can learn alignments between the whole sentence and each image during pre-training.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, and Stephen Gouldand Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 6077–6086, 2018.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

*Corresponding author,†These authors contributed equally to this work.

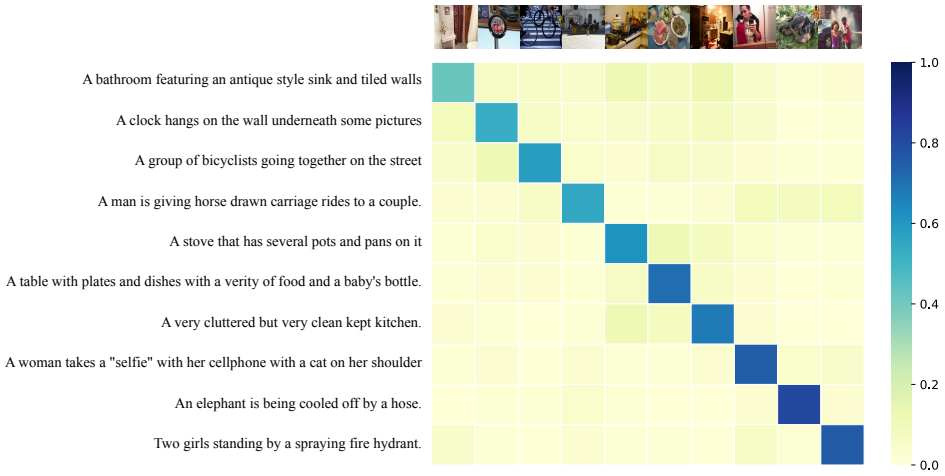

Figure 1: Heatmap visualization of cosine similarities between image and text pre-trained representations. Alignments across image and text pairs are learned during pre-training.

[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 IEEE Conference on Computer Vision and Pattern Recognition*, page 6904–6913, 2017.

[4] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[7] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. Contrastive visual-linguistic pretraining. *arXiv preprint arXiv:2007.13135*, 2020.

[8] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[9] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[10] Rowan Zellers, Yonatan Bisk, Ali Farhadi, , and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 6720–6731, 2019.