Towards Cross-modal Backward-compatible Representation Learning for Vision-Language Models

Supplementary Material

Table 5. Cross-modal retrieval results on *nocaps*, using BLIP-Base and BLIP-Large.

Method	Text Query/Image Gallery			Image Query/Text Gallery		
	Case	R@10	R@50	Case	R@10	R@50
Original						
-	w/v	69.56	91.86	v/w	85.44	97.89
Cross-modal Backward Compatible Training						
Base	$\bar{w}_{new}/v_{old}$	67.61	91.56	$\bar{v}_{new}/w_{old}$	81.42	97.38
XBT		**69.90**	**92.52**		**85.83**	**98.04**
Base	$\bar{w}_{new}/\bar{v}_{new}$	69.02	90.57	$\bar{v}_{new}/\bar{w}_{new}$	79.29	95.98
XBT		**73.17**	**93.27**		**89.16**	**98.88**

Different VLP models. To further explore the capacity of the VLP model architecture's generalization of XBT, we evaluate it using BLIP [21] Base and Large models. We employ checkpoints of `Salesforce/blip-itm-base-coco` and `Salesforce/blip-itm-large-coco` from the Huggingface library. We apply XBT on old and new BLIP models in the same fashion with our CLIP applications, and the results appear in Table 5. From the results, we confirm that XBT provides cross-modal backward-compatibility to the BLIP models too.

5.1. Dataset Examples, More Visualization

In Figure 5, we use a t-SNE map to examine the actual distribution of embeddings in the VLP space. It's evident that the image and text embeddings are distinct. Furthermore, the intra-distribution within both image and text embeddings is similar, suggesting that they are supposed to *mirror* each other.

5.2. Further Analysis & Discussion

Table 6. Computational analysis on baselines. We evaluate with B32 as *old* and L14 as *new* model.

Method	Training Time (h)	Trainable Parameters (M)	Memory Load (GB)	Number of Samples (M)
Text-only Pretraining	1.55	6.82	0.61	67
Full-tune	5.71	434.45	1.13	4
LoRA-only	5.42	8.34	1.13	4
Base	5.66	8.35	1.13	4
XBT	2.84	8.36	0.84	4

Computational Analysis. In Table 6, we calculate the required training cost for each baseline. Despite XBT handling a larger number of training samples, the total training time (Text-only pretraining + XBT) is less than that of the other methods. Furthermore, since XBT does not utilize the old VLP model during training, it significantly reduces the memory load.

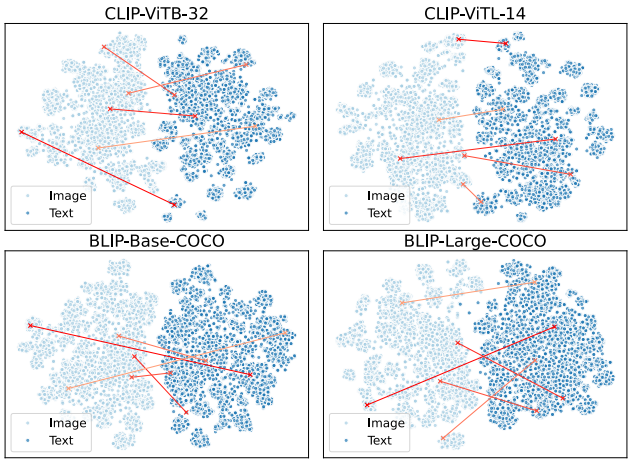

Figure 5. A tSNE visualization of 5,000 paired image-text embeddings from COCO [23] dataset, using two different CLIP models [28], and two different BLIP models [21]. Five pairs are marked as examples. The distinct distributions of image and text samples in each VLP space are observed.

Table 7. Zero-shot classification results on ImageNet [31], ImageNet-R [10], and ImageNet-Sketch [39]. $\bar{w}$ and $\bar{v}$ are used to compute scores, and accuracy (%) is metric.

Method	ImageNet	ImageNet-R	ImageNet-Sketch
CLIP-ViTB-32	55.23	40.66	35.53
CLIP-ViTL-14	66.63	62.30	52.52
XBT trained by 4M	55.44	59.21	45.67
XBT trained by 8M	55.91	61.27	47.02
XBT trained by 16M	57.99	63.53	48.69

Research question: Zero-shot Classification. As we incorporate VLP models, an intriguing research question emerges: How do VLP models, fine-tuned with XBT, perform as zero-shot classifiers? To investigate this, we conduct a zero-shot classification using the text prompt *'a photo of class name'*. As demonstrated in Table 7, XBT outperforms the old VLP in classification performance, though it falls short of the new VLP. Interestingly, we observe that as the number of supervised training samples increases, so does the classification performance. This suggests the potential for XBT-tuned models to function as zero-shot classifiers given sufficient training samples. This opens up a new research direction towards not only achieving backward compatibility, but also comparable performance to zero-shot classifiers.