Supplementary Materials for Unifying Vision-and-Language Tasks via Text Generation

Jaemin Cho [1] **Jie Lei** **Hao Tan** **Mohit Bansal**
UNC Chapel Hill
{jmincho,jielei,haotan,mbansal}@cs.unc.edu

In the supplementary materials, we include the detailed comparisons of our models with recent vision-and-language transformer baselines (Sec. A) and the implementation details (Sec. B).

A. Comparison with Baselines

In Table 1, we compare the baseline vision-and-language transformers with our VL-T5 and VL-BART in detail, including their pretraining datasets, architecture, etc.

B. Implementation Details

In Table 2 and Table 3, we show the detailed statistics of our pretraining and downstream datasets and tasks. In Table 4, we show the hyperparameters that we used in our pretraining and downstream task experiments. We provide the links to download pretraining and downstream datasets.

B.1. Pretraining Data

Overall, our pretraining dataset contains 9.18M image-text pairs on 180K distinct images. We carefully split our pretraining data to avoid any intersection between our training data and the validation/test sets of the downstream tasks (e.g., COCO Captioning, RefCOCOg). In this process, around 10K images are excluded from the training sets of COCO[1] and Visual Genome[2]. We use COCO *Karpathy val split* (Karpathy & Fei-Fei, 2015) with 5,000 images as our validation set to monitor pretraining performance.

B.2. Downstream Tasks

VQA[3], COCO caption For both VQA and COCO captioning tasks, we follow *Karpathy split* (Karpathy & Fei-Fei, 2015), which re-splits train2014 and val2014 COCO images (Lin et al., 2014) into 113,287 / 5,000 / 5,000 images for train / validation / test.

GQA[4] Following LXMERT (Tan & Bansal, 2019), we use GQA-balanced version. We use train and val splits for training and use test-dev split for validation. Train / val / test-dev splits consist of 943,000 / 132,062 / 12,578 questions, respectively.

NLVR[2] [5] Train / val / test-P splits consist of 86,373 / 6982 / 6967 sentences, respectively. We train our model on train split and use val split for validation.

VCR[6] Train / val / test splits consist of 212,923 / 26,534 / 25,263 questions, respectively. We train our model on train split and use val split for validation.

RefCOCOg[7] We use *umd* split, which consists of train / val / test sets with 42,226 / 2,573 / 5,023 sentences, respectively. Following UNITER (Chen et al., 2020) and MAttNet (Yu et al., 2018), we use ground truth COCO boxes for training, and use the detected boxes from an off-the-shelf Mask R-CNN [8] as candidates during inference.

Multi30K En-De[9] The train / val / test2016 / test2017 / test2018 splits consist of 29,000 / 1,014 / 1,000 / 1,000 / 1,017 English-German sentence pairs, respectively.

[1]UNC Chapel Hill. Correspondence to: Jaemin Cho <jmincho@cs.unc.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[1]https://cocodataset.org/#download
[2]http://visualgenome.org/api/v0/api_home.html
[3]https://visualqa.org/download.html
[4]https://cs.stanford.edu/people/dorarad/gqa/download.html
[5]http://lil.nlp.cornell.edu/nlvr/
[6]https://visualcommonsense.com/download/
[7]https://github.com/lichengunc/refer
[8]https://github.com/lichengunc/MAttNet#pre-computed-detectionsmasks
[9]https://github.com/multi30k/dataset

Table 1. Summary of baseline vision-and-language transformers. [a]Since not all models report exact parameter numbers, we provide rough estimates compared to BERT$_{Base}$ (86M; noted as P), where word embedding parameters are excluded. [b]LXMERT and XGPT are not initialized from pretrained language models. LXMERT authors found pretraining from scratch was more effective than initialization from BERT$_{Base}$ in their experiments. XGPT uses text pretraining on Conceptual captions and COCO captions with Masked LM (Devlin et al., 2019) and Masked Seq2Seq (Song et al., 2019) objectives before V&L pretraining. [c]LXMERT (text+visual+cross-modal) and ViLBERT (cross-modal) use dual-stream encoders. ViLBERT uses 768/1024-dim hidden states for text/visual streams respectively. XGPT uses AoA module (Huang et al., 2019) as visual encoder. Rest of the models use single-stream encoders. [d]For generation tasks, Unified VLP and Oscar use causal mask and reuse encoder as decoder similar to UniLM. [e]XGPT also uses shared parameters for encoder and decoder, but its decoder is right-shifted and predicts next tokens. [f]Unified VLP is initialized from UniLM, which is initialized from BERT$_{Large}$. [g]Oscar uses object tags as additional text inputs.

	V&L Pretraining		Hyperparameters						
	Dataset	# Imgs	Arch. type	Backbone	# Layers	# Params[a]	Hidden dim	# Regions	Position Emb
LXMERT	COCO+VG	180K	Encoder	-[b]	9+5+5[c]	2P	768	36	absolute
ViLBERT	CC	3M	Encoder	BERT$_{Base}$	12[c]	2.5P	768/1024[c]	10∼36	absolute
UNITER$_{Base}$	CC+SBU+COCO+VG	4M	Encoder	BERT$_{Base}$	12	P	768	10∼100	absolute
Unified VLP	CC	3M	Encoder[d]	UniLM[f]	12	P	768	100	absolute
Oscar$_{Base}$	CC+SBU+COCO+VG+Flickr30K	4M	Encoder[d]	BERT$_{Base}$	12	P	768	50[g]	absolute
XGPT	CC+COCO	3M	Enc-Dec[e]	-[b]	1[c]+12+12	P	768	100	absolute
VL-T5	COCO+VG	180K	Enc-Dec	T5$_{Base}$	12+12	2P	768	36	relative
VL-BART	COCO+VG	180K	Enc-Dec	BART$_{Base}$	6+6	P	768	36	absolute

Table 2. Pretraining tasks used in our vision-and-language pretraining. The images that have any intersection with evaluation set of downstream tasks (e.g., COCO caption, RefCOCOg) and the held-out validation set for pretraining are excluded.

Task	Image source	Text source	# Examples
Multimodal language modeling	COCO, VG	COCO caption, VG caption	4.9M (# captions)
Visual question answering	COCO, VG	VQA, GQA, Visual7W	2.5M (# questions)
Image-text matching	COCO	COCO caption	533K (# captions)
Visual grounding	COCO, VG	object&attribute tags	163K (# images)
Grounded captioning	COCO, VG	object&attribute tags	163K (# images)

Table 3. Statistics of the datasets used in downstream tasks. The data that are not used for training/validation (e.g., COCO test2015 images) and data for leaderboard submissions (e.g., test-dev/test-std for VQA, test for GQA) are excluded.

Datasets	Image source	# Images (train)	# Text (train)	Metric
VQA	COCO	123K (113K)	658K (605K)	VQA-score
GQA	VG	82.7K (82.3K)	1.08M (1.07M)	Accuracy
NLVR^2	Web Crawled	238K (206K)	100K (86K)	Accuracy
RefCOCOg	COCO	26K (21K)	95K (80K)	Accuracy
VCR	Movie Clips	110K (80K)	290K (212K)	Accuracy
COCO Caption	COCO	123K (113K)	616K (566K)	BLEU,CIDEr,METEOR,SPICE
Multi30K En-De	Flickr30K	31K (29K)	31K (29K)	BLEU

Table 4. Hyperparameters for pretraining and downtream tasks

Model	Task	Learning rate	Batch size	Epochs
VL-T5	Pretraining	1e-4	320	30
	VCR Pretraining	5e-5	80	20
	VQA	5e-5	320	20
	GQA	1e-5	240	20
	NLVR^2	5e-5	120	20
	RefCOCOg	5e-5	360	20
	VCR	5e-5	16	20
	COCO Caption	3e-5	320	20
	Multi30K En-De	5e-5	120	20
VL-BART	Pretraining	1e-4	600	30
	VCR Pretraining	5e-5	120	20
	VQA	5e-5	600	20
	GQA	1e-5	800	20
	NLVR^2	5e-5	400	20
	RefCOCOg	5e-5	1200	20
	VCR	5e-5	48	20
	COCO Caption	3e-5	520	20
	Multi30K En-De	5e-5	320	20

References

Chen, Y.-c., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, 2020. URL https://arxiv.org/abs/1909.11740.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, oct 2019. URL http://arxiv.org/abs/1810.04805.

Huang, L., Wang, W., Chen, J., and Wei, X. Y. Attention on attention for image captioning. In *ICCV*, pp. 4633–4642, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00473.

Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. ISBN 9781467369640. doi: 10.1109/TPAMI.2016.2598339.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. ISBN 978-3-319-10601-4. doi: 10.1007/978-3-319-10602-1_48.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML*, 2019. URL http://arxiv.org/abs/1905.02450.

Tan, H. and Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*, 2019. URL http://arxiv.org/abs/1908.07490.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. MAttNet : Modular Attention Network for Referring Expression Comprehension. In *CVPR*, 2018. URL https://arxiv.org/abs/1801.08186.