Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling – Supplementary File

Jie Lei*[1], Linjie Li*[2], Luowei Zhou[2], Zhe Gan[2], Tamara L. Berg[1], Mohit Bansal[1], Jingjing Liu[2]
[1]UNC Chapel Hill [2]Microsoft Dynamics 365 AI Research
{jielei, tlberg, mbansal}@cs.unc.edu
{lindesy.li, luowei.zhou, zhe.gan, jingjl}@microsoft.com

Method	feature	test-dev	test-std
BUTD [1]	R	65.32	65.67
grid-feat [6]	G	66.47	-
ViLBERT [9]	R	70.55	70.92
VL-BERT [10]	R	71.16	-
Pixel-BERT [4]	G	71.35	71.42
LXMERT [11]	R	72.42	72.54
UNITER [1]	R	72.70	72.91
Oscar [8]	R	73.16	73.44
ClipBERT 1×1	G	69.08	69.43

Table 1: Comparison with state-of-the-art methods on VQA. *G* stands for grid features, *R* stands for region features.

1. Additional Experiments

Visual Question Answering. As ClipBERT is designed based on 2D CNN, and is pre-trained on image-text corpus, it is also directly applicable to image-text downstream tasks, such as image-based question answering. We show ClipBERT's performance on VQA 2.0 dataset [3] in Table 1. The model is finetuned from the image-text pre-trained weights on 8GPUs for 13 epochs, with batch size 32 and learning rate 5e-5. ClipBERT shows a reasonable performance compared to the strong pre-training baselines. Note that ClipBERT uses grid features [6, 4] instead of the commonly used region features, which is much more computation efficient, *e.g.*, extracting grid features is around 80× faster than extracting region features according to the computation time reported in [6].

2. Downstream Task Adaptation

Our ClipBERT is quite generic, once trained, it can be easily adopted and transferred for various downstream tasks. In particular, in this work, we focus on text-to-video retrieval and video question answering.

Text-to-video Retrieval. We use a two-layer MLP with the last layer `[CLS]` token hidden state for a two way (*i.e.*, matched or not matched) classification for retrieval. We use LogSumExp loss for training. Denote the two-way classification logit output for clip τ_i from the video associated with the j-th example as $\boldsymbol{g}_{\tau_i}^{(j)} \in \mathbb{R}^2$, where $i = 1, \ldots, N_{train}$ for training ($i = 1, \ldots, N_{test}$ for inference; see Section 3 of the main paper). The LogSumExp prediction $\boldsymbol{p}^{(j)} \in \mathbb{R}^2$ is defined as:

$$\boldsymbol{p}^{(j)} = \frac{\sum_{i=1}^{N_{train}} e^{\boldsymbol{g}_{\tau_i}^{(j)}}}{\mathrm{sum}(\sum_{i=1}^{N_{train}} e^{\boldsymbol{g}_{\tau_i}^{(j)}})}. \tag{1}$$

We then use a negative log likelihood loss for training:

$$L = -\frac{1}{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{D}|} \log \boldsymbol{p}^{(j)}[y_j], \tag{2}$$

where $\mathcal{D}$ is the dataset, y_j is the index of the ground-truth answer for the j-th example.

We conduct experiments on three popular text-to-video retrieval datasets, MSRVTT [13], DiDeMo [2], and ActivityNet Captions [7]. Table 2 shows the training details for models on each of the datasets.

Dataset	#Epochs	Bsz ×Grad-Accu ×#GPUs	LR
MSRVTT	20	16×1×8	5e-5
DiDeMo	20	8×4×8	5e-5
ActivityNet Captions	80	16×2×8	5e-5

Table 2: Training details for text-to-video retrieval tasks. Bsz is short for batch size. Grad-Accu stands for gradient accumulation steps. LR means initial learning rate.

Video Question Answering. Similar to text-to-video retrieval task, we take the last layer `[CLS]` token hidden state through a two-layer MLP for classification. We use LogSumExp to aggregate prediction from multiple clips to calculate loss. The formulation of LogSumExp loss is simlar to Equation 1 except that the dimension of $\boldsymbol{g}_{\tau_i}$ equals to the number of answer candidates.

We conduct experiments on three video QA datasets, TGIF-QA [5], MSRVTT-QA [12], and MSRVTT MC Test [14]. For TGIF-QA, we evaluate three sub-tasks,

* Equal contribution.

Dataset	#Epochs	Bsz×Grad-Accu ×#GPUs	LR
TGIF-QA Action	55	32×1×8	1e-4
TGIF-QA Transition	15	32×1×8	1e-4
TGIF-QA FrameQA	15	32×1×8	1e-4
MSRVTT-QA	10	16×1×4	5e-5

Table 3: Training details for video question answering tasks. Bsz is short for batch size. Grad-Accu stands for gradient accumulation steps. LR means initial learning rate.

i.e., Action, Transition, and FrameQA. We train a separate model for each of the evaluated TGIF-QA tasks. For MSRVTT MC Test, as it uses the same training set as the MSRVTT retrieval task, we directly use the trained retrieval model to rank the five candidate answers. Table 2 shows the training details for models on TGIF-QA tasks and MSRVTT-QA.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 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 *CVPR*, 2017.

[4] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[5] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[6] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020.

[7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[8] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[10] 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 *ICLR*, 2020.

[11] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.

[12] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, 2017.

[13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[14] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018.