FACEBOOK

VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding

Facebook AI

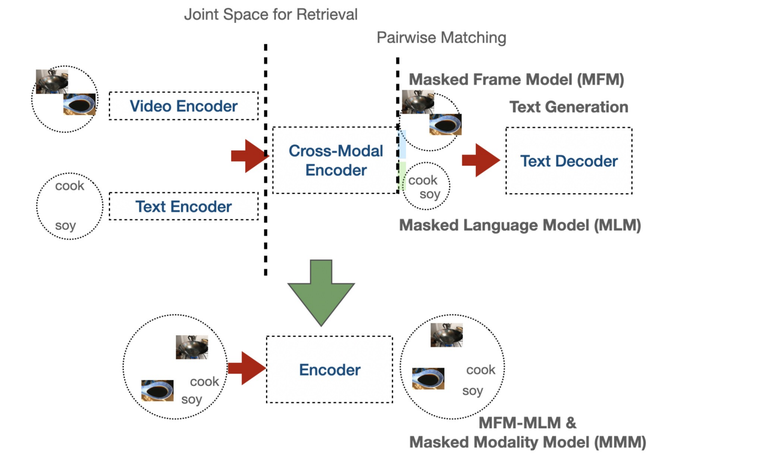
Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora , Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze and Luke Zettlemoyer

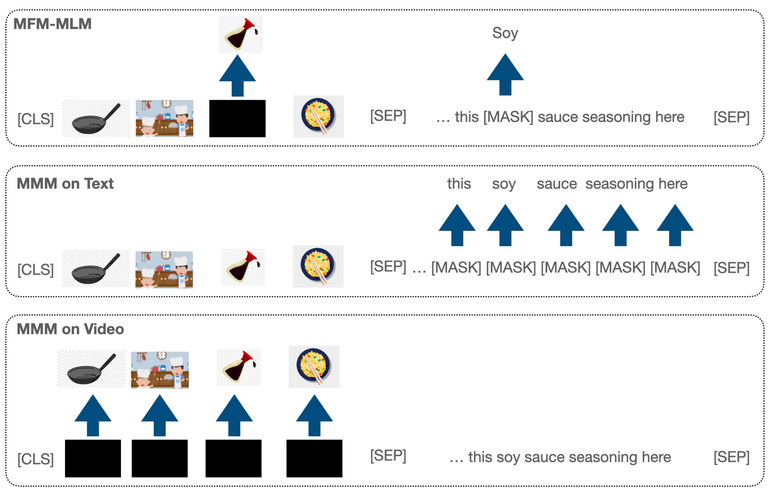
Abstract

We present a simplified, task-agnostic multi-modal pre-training approach that can accept either video or text input, or both for a variety of end tasks.

We focus on video-text understanding: Video with Caption transcribed from ASR. Existing works adopt either task-specific (retrieval) or multi-task pre-training.
Advantage of Task-agnostic:
- Modality-agnostic: MM fusion.
- One encoder, one loss. Smaller model.

We introduce new pretraining masking schemes that better mix across modalities (e.g. by forcing masks for text to predict the closest video embeddings) while also maintaining separability (e.g. unimodal predictions are sometimes required, without using all the input).





- Existing work adopts masked language model (MLM) for text and masked frame model (MFM) for video frames. One modality tends to perform self-attention on its own modality. We propose masked modality modal (MMM): learn fusion of video and text.

Model Comparison

Model	Paradigm	#params.	#loss	#unimodal/cross en/decoder	Joint Retrieval	Generation
MMT(Gabeur et al., 2020)	task-specific alignment	127.3M	1	2/0/0	yes	no
ActBERT(Zhu and Yang, 2020)	weakly supervised/MTL	n/a (3 typed attentions)	4	0/1(modal-typed attn.)/0	no(pair)	extra decoder
VideoAsMT(Korbar et al., 2020)	weakly supervised/MTL	286M(base)/801M(large)	1	1/1/1	no (gen.)	yes
HERO(Li et al., 2020b)	SSL(w/ sup. video feat.)/MTL	159M	5	1(query)/2/0	no(pair)	extra decoder
UniVL(Luo et al., 2020)	SSL/MTL	260M	5	2/1/1	yes	yes
VLM	SSL/Task-agnostic	**110M**	**1**	0/**1**/0(shared w/ encoder)	yes	yes

One Encoder, One Loss

- One BERT transformer consumes both video (via quantization) and text tokens:

$$\boldsymbol{h} = \mathrm{BERT}(\boldsymbol{x}_{\mathrm{masked}}),$$

- One Loss on video/text tokens (a larger MM vocab. contains both video tokens in the same batch and text tokens in vocab.):

$$\frac{\exp(x_s^T e_s)}{\exp(x_s^T e_s) + \sum_{j \in V' \cup D_{\setminus s}} \exp(x_j^T e_s)}.$$

Fine-tuning Result

Method	Accuracy
Joint Retrieval	
JSFusion(Yu et al., 2018)	83.4
Pairwise Matching	
ActBERT(Zhu and Yang, 2020)	85.7
VLM	91.64

Table 6: Video question answering (multiple-ch evaluated on MSR-VTT.

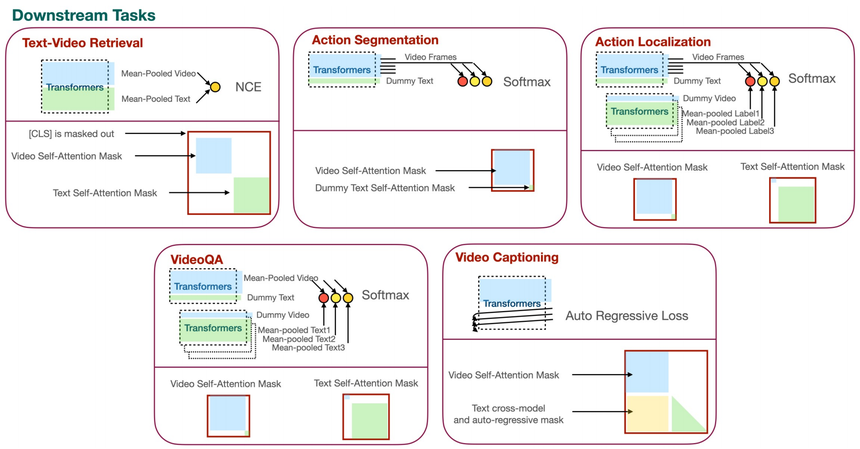
Methods	Average Recall
Joint Alignment	
Alayrac (Alayrac et al., 2016)	13.3
Zhukov (Zhukov et al., 2019)	22.4
Supervised (Zhukov et al., 2019)	31.6
HowTo100M (Miech et al., 2019)	33.6
MIL-NCE (Miech et al., 2020)	40.5
UniVL (Luo et al., 2020)	42.0
Pairwise Matching	
ActBERT (Zhu and Yang, 2020)	41.4
VLM (task-agnostic, zero-shot)	28.5
VLM (supervised on 540 videos)	46.5

Table 5: Action step localization results on CrossTask.

Method	Frame Accuracy
NN-Viterbi (Richard et al., 2018)	21.17
VGG (Simonyan and Zisserman, 2014)	25.79
TCFPN-ISBA (Ding and Xu, 2018)	34.30
CBT (Sun et al., 2019a)	53.90
MIL-NCE (Miech et al., 2020)	61.00
ActBERT (Zhu and Yang, 2020)	56.95
VLM	68.39

Table 4: Action segmentation on COIN dataset.

Fine-tuning Setup



Ablation Study

VLM	R@1	R@5	R@10	Median R
w/ MMM 50%	27.05	56.88	69.38	4.0
w/ MMM 0%	15.12	39.47	52.81	9.0
w/ MMM 30%	25.30	54.80	68.96	4.0
w/ MMM 70%	25.17	54.98	69.11	4.0
w/ min. 16 text tokens	25.84	54.43	68.29	5.0
w/ $\mathcal{L}_{\text{MFM-MLM}}$	26.93	55.92	69.86	4.0

Table 8: Ablation study of VLM for text-based video retrieval on Youcook2.

VLM	B-3	B-4	M	R-L	CIDEr
w/ MMM 50%	17.78	12.27	18.22	41.51	1.3869
w/ MMM 0%	15.47	10.54	16.49	38.83	1.2163
w/ MMM 30%	16.57	11.30	17.55	40.76	1.3215
w/ MMM 70%	16.94	11.68	17.67	41.24	1.3739
w/ min. 16 text tokens	17.25	12.00	17.67	40.62	1.3076
w/ $\mathcal{L}_{\text{MFM-MLM}}$	16.66	11.53	17.34	40.36	1.3224

Table 9: Ablation study of VLM for video captioning on Youcook2 dataset.