An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling


Tsu-Jui Fu[1]


Linjie Li[2]


Zhe Gan[3]


Kevin Lin[2]


William Wang[1]


Lijuan Wang[2]


Zicheng Liu[2]

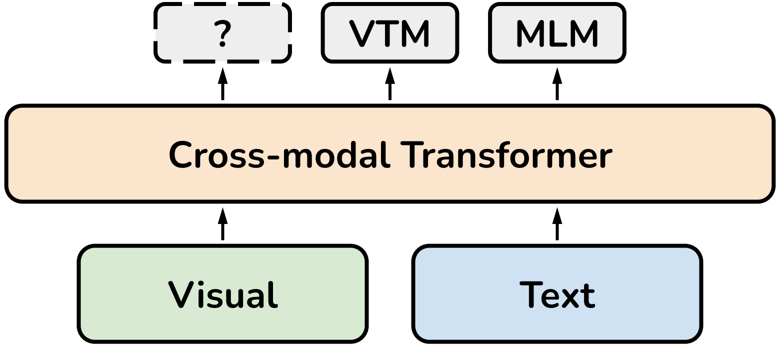
[1]UC Santa Barbara, [2]Microsoft, [3]Apple

https://tsujuifu.github.io

Large-scale Text-Visual Pre-training

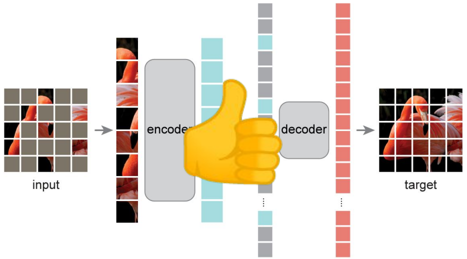
- Masked Language Modeling (**MLM**): recover missing word tokens
- Visual-Text Matching (**VTM**): alignment between visual and textual inputs

- How to enhance the visual modality ?

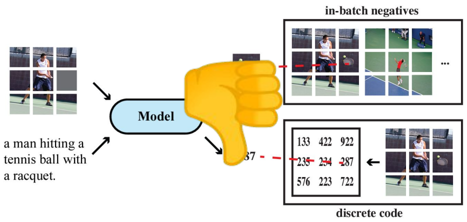


Mask Visual Modeling (MVM)

- MVM achieves promising results for self-supervised visual pre-training
  - MAE, BEiT, VideoMAE, ...
- In contrast, MVM even hurts performance on text-image pre-training
- How can we design effective MVM for **text-video pre-training** ?



Visual Pre-training



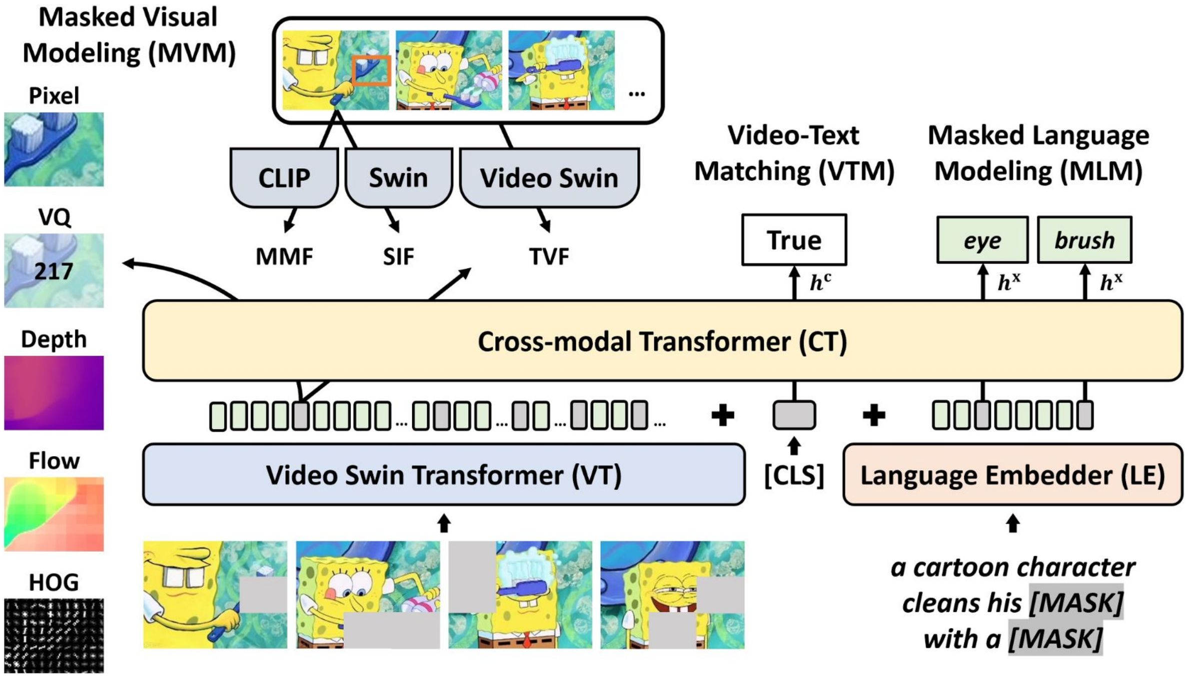
Text-Image Pre-training

Text-Video Pre-training

[CVPR'22] *Masked Autoencoders Are Scalable Vision Learners*
[CVPR'22] *An Empirical Study of Training End-to-End Vision-and-Language Transformers*

Diverse Targets of MVM

- Explore various MVM targets for end-to-end VidL learning
  - **Low-level**: Pixel, HOG
  - **Semantic-level**: Depth, Flow, SIF, TVF
  - **Multi-modal**: VQ, MMF



MVM on Text-Video (WebVid-2.5M)

- **Not all MVMs** are helpful for VidL
- Only **Pixel** and **SIF** bring consistent improvement on both downstream tasks
- **SIF** gains significant advance, especially on T2V

Pre-train	MVM	TGIF-Frame	DiDeMo-Retrieval			
		Accuracy	R1	R5	R10	AveR
VTM+MLM	None	68.1	28.7	57.0	69.7	51.8
+MVM	Pixel	68.3 (+0.2)	29.2 (+0.5)	58.6 (+1.6)	70.1 (+0.4)	52.6 (+0.8)
	HOG	67.3 (-0.8)	26.6 (-2.1)	54.9 (-2.1)	68.1 (-1.6)	49.8 (-2.0)
	Depth	68.0 (-0.1)	27.3 (-1.4)	55.0 (-2.0)	68.3 (-1.4)	50.2 (-1.6)
	Flow	67.6 (-0.5)	30.3 (+1.6)	58.0 (+1.0)	70.3 (+0.6)	52.9 (+1.1)
	SIF	**68.8 (+0.7)**	**35.4 (+6.7)**	**62.4 (+5.4)**	**74.9 (+5.2)**	**57.6 (+5.8)**
	TVF	68.0 (-0.1)	32.8 (+4.1)	60.5 (+3.5)	73.0 (+3.3)	55.4 (+3.6)
	VQ	68.4 (+0.3)	28.1 (-0.6)	56.6 (-0.4)	69.4 (-0.3)	51.3 (-0.5)
	MMF	67.7 (-0.4)	29.8 (+1.1)	57.8 (+0.8)	68.5 (-1.2)	52.1 (+0.3)

Combination of MVM targets on Text-Video

- Joint of different MVMs is **not encouraging**
- Explicit Pixel **conflicts with** high-level SIF
- SIF+TVF cannot bring more improvement (T2V ↓)

MVM	TGIF-Frame	DiDeMo-Retrieval			
	Accuracy	R1	R5	R10	AveR
None	68.1	28.7	57.0	69.7	51.8
Pixel	68.3 (+0.2)	29.2 (+0.5)	58.6 (+1.6)	70.1 (+0.4)	52.6 (+0.8)
Flow	67.6 (-0.5)	30.3 (+1.6)	58.0 (+1.0)	70.3 (+0.6)	52.9 (+1.1)
SIF	68.8 (+0.7)	**35.4 (+6.7)**	62.4 (+5.4)	**74.9 (+5.2)**	**57.6 (+5.8)**
TVF	68.0 (-0.1)	32.8 (+4.1)	60.5 (+3.5)	73.0 (+3.3)	55.4 (+3.6)
SIF+Pixel	68.8 (+0.7)	31.8 (+3.1)	60.4 (+3.4)	73.0 (+3.3)	55.1 (+3.3)
SIF+Flow	68.7 (+0.6)	34.4 (+5.7)	61.5 (+4.5)	72.8 (+3.1)	56.3 (+4.5)
SIF+TVF	**69.2 (+1.1)**	33.8 (+5.1)	**63.0 (+6.0)**	74.4 (+4.7)	57.1 (+5.3)

MVM on Text-Image (CC3M)

- **Challenging to learn** without visual implications from neighbor frames
- **Fit in static image**, which hurts video temporal
- MVM cannot work well on text-image data for VidL

Pre-train	MVM	TGIF-Frame	DiDeMo-Retrieval			
		Accuracy	R1	R5	R10	AveR
VTM+MLM	None	**69.8**	**36.4**	64.3	74.7	**58.4**
+MVM	Pixel	69.7 (-0.1)	35.8 (-0.6)	**64.4 (+0.1)**	74.9 (+0.2)	58.4
	HOG	69.8	34.9 (-1.5)	64.4 (+0.1)	75.1 (+0.4)	58.1 (-0.3)
	Depth	69.6 (-0.2)	32.3 (-4.1)	63.8 (-0.5)	74.2 (-0.5)	56.9 (-1.5)
	SIF	69.7 (-0.1)	31.6 (-4.8)	60.5 (-3.8)	72.5 (-2.2)	54.9 (-3.5)
	VQ	69.8	34.4 (-2.0)	62.6 (-1.7)	75.1 (+0.4)	57.4 (-1.0)
	MMF	69.8	33.6 (-2.8)	62.9 (-1.4)	**75.6 (+0.9)**	57.4 (-1.0)

MVM on Text-Image & Text-Video

- Not trivial to find superior MVM combination
- **Video (SIF) + Image (None)** is our default setting

Pre-train	MVM		TGIF-Frame	DiDeMo-Retrieval			
	WebVid	CC3M	Accuracy	R1	R5	R10	AveR
VTM+MLM	None		69.7	36.7	66.5	76.6	59.9
+MVM	SIF	None	71.1 (+1.4)	38.8 (+2.1)	**69.6 (+3.1)**	**80.0 (+3.4)**	**62.8 (+2.9)**
	SIF	Pixel	**71.3 (+1.6)**	**39.7 (+3.0)**	69.3 (+2.8)	78.4 (+1.8)	62.5 (+2.6)

SIF Extractor *vs.* Downstream

- Classification accuracy is crucial but **not positively correlated**
- **Similar inductive biases** is another key
- Trade-off between **informative and feasible** learning

SIF		IN-1K	TGIF-Frame	DiDeMo-Retrieval			
Model	**Train**	**Accuracy**	**Accuracy**	**R1**	**R5**	**R10**	**AveR**
	None		68.1	28.7	57.0	69.7	51.8
Res-50	IN-1K	76.1	67.3 (-0.8)	29.1 (+0.4)	58.1 (+1.1)	69.3 (-0.4)	52.2 (+0.4)
Swin-T	IN-1K	81.2	**68.9 (+0.8)**	33.8 (+5.1)	**63.6 (+6.6)**	74.2 (+4.5)	57.2 (+5.4)
DeiT	IN-1K	83.4	68.4 (+0.3)	31.4 (+2.7)	59.4 (+2.4)	72.2 (+2.5)	54.3 (+2.5)
Swin-B	IN-1K	83.5	68.3 (+0.2)	34.9 (+6.2)	63.4 (+6.4)	73.9 (+4.2)	57.4 (+5.6)
Swin-B	IN-22K	85.2	68.8 (+0.7)	**35.4 (+6.7)**	62.4 (+5.4)	**74.9 (+5.2)**	**57.6 (+5.8)**
Swin-L	IN-22K	**86.3**	68.2 (+0.1)	33.2 (+4.5)	62.4 (+5.4)	72.6 (+2.9)	56.1 (+4.3)

Comparison with SOTA

- Video Question Answering (VideoQA)

Method	#Pre-train	TGIF			MSRVTT		LSMDC		MSVD
		Act.	Trans.	Frame	MC	QA	MC	FiB	QA
ClipBERT	0.2M	82.8	87.8	60.3	88.2	37.4	-	-	-
ALRPO	5M	-	-	-	-	42.1	-	-	46.3
JustAsk	69M	-	-	-	-	41.5	-	-	46.3
MERLOT	180M	94.0	96.2	69.5	90.9	43.1	81.7	52.9	-
VIOLET	186M	92.5	95.7	68.9	91.9	43.9	82.8	53.7	47.9
All-in-One	283M	95.5	94.7	66.3	92.3	46.8	84.4	-	48.3
VIOLETv2	5M	**94.8**	**99.0**	**72.8**	**97.6**	**44.5**	**84.4**	**56.9**	**54.7**

Comparison with SOTA

- Text-to-Video Retrieval (T2V)

Method	#Pre-train	MARVTT			DiDeMo			LSMDC		
		R1	R5	R10	R1	R5	R10	R1	R5	R10
ClipBERT	0.2M	22.0	46.8	59.9	20.4	48.0	60.8	-	-	-
Frozen	5M	31.0	59.5	70.5	31.0	59.8	72.8	15.0	30.8	39.8
ALPRO	5M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
B-Former	5M	**37.6**	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5
All-in-One	138M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
VIOLET	186M	34.5	63.0	73.4	32.6	62.8	74.7	16.1	36.6	41.2
Clip4Clip	400M	42.1	71.9	81.4	43.4	70.2	80.6	21.6	41.8	49.8
VIOLETv2	5M	37.2	**64.8**	**75.8**	**47.9**	**76.5**	**84.1**	**24.0**	**43.5**	**54.1**

Summary

- Explore **various MVM targets** for VidL learning
  - Low-level: **Pixel**, HOG
  - Semantic-level: Depth, Flow, **SIF**, TVF
  - Multi-modal: VQ, MMF

- Best setting should be **Text-Video (SIF) + Text-Image (None)**
  - Not trivial to find superior combination of MVM

- Features extractor is also crucial
  - Classification accuracy is **not always positively correlated**
  - **Similar inductive biases** is the key
  - Trade-off between **informative and feasible** learning