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

Paper Review

Immanuel, Steve Andreas - 22110338
VLI - Sejong University

J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, J. Liu, Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021

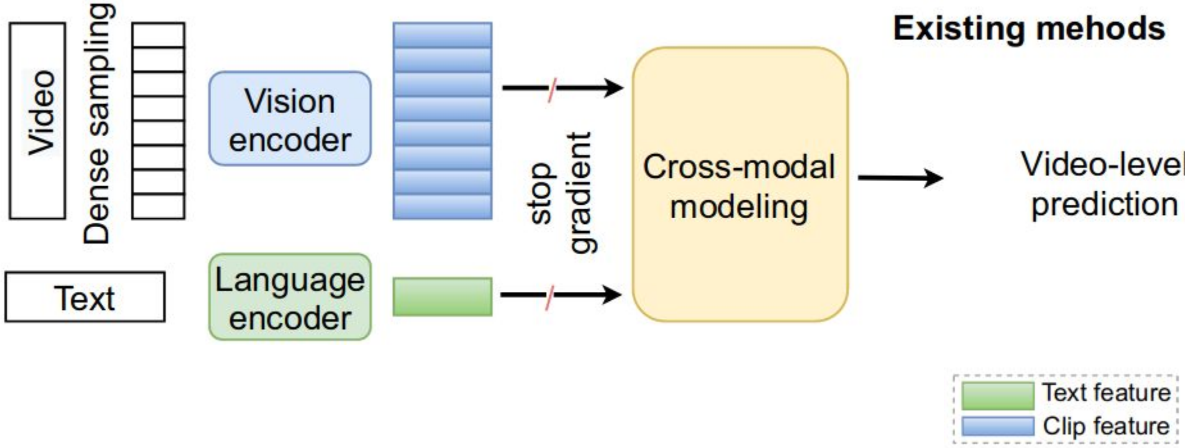
Background & Motivation

- Humans communicate with each other in a dynamic visual world using various signals, e.g. language, sign, gesture
- We want to create an intelligent agent that can **interpret** those **multimodal signals**
- Essentially, the agent has to be able to **jointly understand the visual and textual clues** that is being conveyed by those signals
- Examples of tasks to evaluate such ability are **video captioning**, **text-to-video retrieval**, and **video question answering**

Background & Motivation (Cont.)

Standard approaches consist of:

- Pre-trained vision model
- Pre-trained language model
- Multimodal fusion model

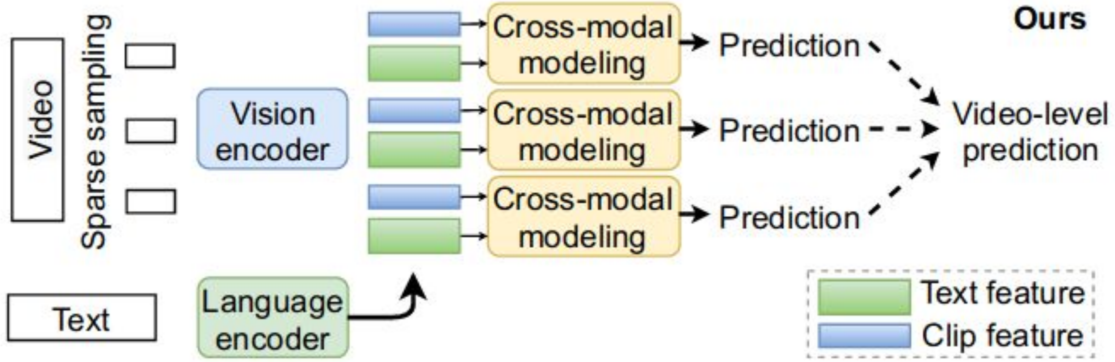


Background & Motivation (Cont.)

- Problem of standard approaches:
  - Disconnection in tasks
  - Disconnection in multimodal features
- Can be solved with **end-to-end** task-specific finetuning
- However, most approaches extract the features from **full sequence** of video frames which requires excessive demand on memory and computation

Key Ideas

- Instead of using **full sequence** of video frames, ClipBERT sparsely samples only one or a few short clips from full length video during **training**
- The hypothesis is that sparse clips already capture key visual and semantic information
- During **inference**, multiple densely-sampled clips are aggregated to obtain final video-level prediction



Key Ideas (Cont.)

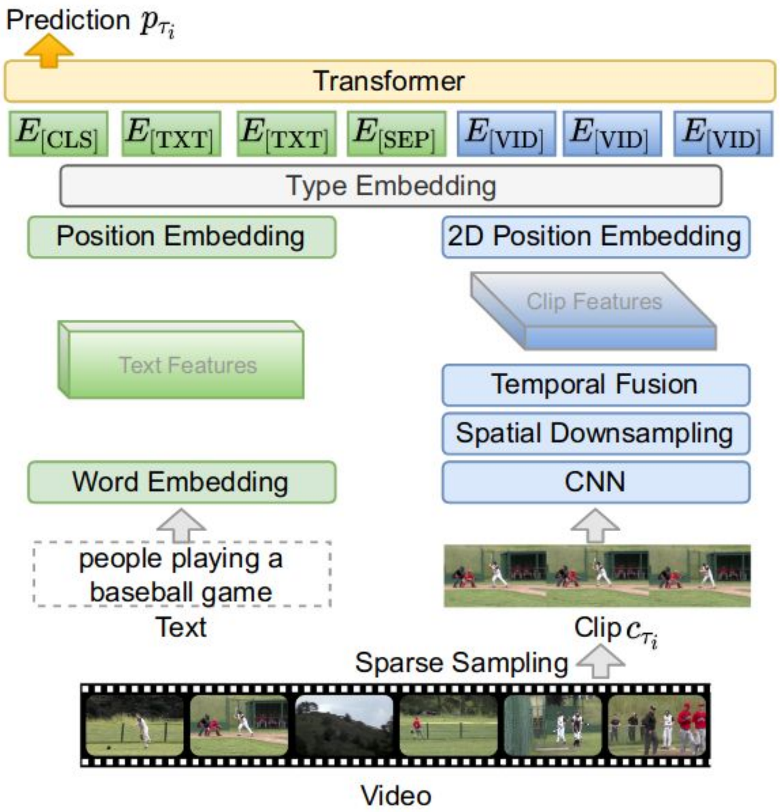
Use 2D vision feature extractor, e.g., ResNet-50 [2], instead of 3D vision feature extractor in order to:

- Study the effect of **image**-text pre-training on **video**-text tasks
- Lower memory cost and faster runtime

Main Contribution

- ClipBERT, general end-to-end learning framework for video-language tasks
- Proposed the use of end-to-end training strategy with sparsely-sampled clips, proving that "*less is more*"
- Demonstrated that **image**-text pre-training benefits **video**-text tasks

Architecture



Details:

- Text feature extractor: BERT-base [3]
- CNN: ResNet-50
- Spatial downsampling: max pooling
- Temporal fusion: mean pooling
- Prediction head: two-layer MLP

The whole model is pre-trained on image-text dataset COCO Captions [4] and Visual Genome Captions [5]

Dataset

- Text-Video Retrieval:
  - MSR-VTT [6]
  - DiDeMo [7]
  - ActivityNet Captions [8]
- Video Question Answering:
  - TGIF-QA [9]
  - MSRVTT-QA [10]
  - MSRVTT multiple-choice test [11]

Comparison with Prior Approaches

Method	R1	R5	R10	MdR
HERO [37] ASR, PT	20.5	47.6	60.9	-
JSFusion [77]	10.2	31.2	43.2	13.0
HT [46] PT	14.9	40.2	52.8	9.0
ActBERT [83] PT	16.3	42.8	56.9	10.0
HERO [37] PT	16.8	43.4	**57.7**	-
ClipBERT 4×1	**19.8**	**45.1**	57.5	**7.0**
ClipBERT 8×2	**22.0**	**46.8**	**59.9**	**6.0**

(a) MSRVTT 1K test set.

Method	R1	R5	R10	MdR
CE [41]	16.1	41.1	-	8.3
S2VT [65]	11.9	33.6	-	13.0
FSE [80]	13.9	36.0	-	11.0
ClipBERT 4×1	**19.9**	**44.5**	**56.7**	**7.0**
ClipBERT 8×2	**20.4**	**48.0**	**60.8**	**6.0**

(b) DiDeMo test set.

Method	R1	R5	R10	MdR
CE [41]	18.2	47.7	-	6.0
MMT [15]	22.7	54.2	93.2	5.0
MMT [15] PT	28.7	61.4	94.5	3.3
Dense [28]	14.0	32.0	-	34.0
FSE [80]	18.2	44.8	-	7.0
HSE [80]	20.5	**49.3**	-	-
ClipBERT 4×2*	**20.9**	48.6	**62.8**	**6.0**
ClipBERT 4×2* (N_{test}=20)	**21.3**	**49.0**	**63.5**	**6.0**

(c) ActivityNet Captions val1 set.

Method	Action	Transition	FrameQA
ST-VQA [23]	60.8	67.1	49.3
Co-Memory [17]	68.2	74.3	51.5
PSAC [38]	70.4	76.9	55.7
Heterogeneous Memory [12]	73.9	77.8	53.8
HCRN [31]	75.0	81.4	55.9
QueST [25]	75.9	81.0	**59.7**
ClipBERT 1×1 (N_{test}=1)	**82.9**	**87.5**	59.4
ClipBERT 1×1	**82.8**	**87.8**	**60.3**

(a) TGIF-QA test set.

Method	Accuracy
ST-VQA [23] (by [12])	30.9
Co-Memory [17] (by [12])	32.0
AMU [74]	32.5
Heterogeneous Memory [12]	33.0
HCRN [31]	35.6
ClipBERT 4×1	**37.0**
ClipBERT 8×2	**37.4**

(b) MRSVTT-QA test set.

Method	Accuracy
SNUVL [78] (by [77])	65.4
ST-VQA [23] (by [77])	66.1
CT-SAN [79] (by [77])	66.4
MLB [27] (by [77])	76.1
JSFusion [77]	83.4
ActBERT [83] PT	85.7
ClipBERT 4×1	**87.9**
ClipBERT 8×2	**88.2**

(c) MRSVTT multiple-choice test.

Experiments Results

L	MSRVTT Retrieval				MSRVTT-QA Acc.
	R1	R5	R10	MdR	
224	6.8	24.4	35.8	20.0	**35.78**
448	**10.2**	**28.6**	**40.5**	**17.0**	**35.73**
768	**11.0**	27.8	**40.9**	**16.0**	**35.73**
1000	10.0	**28.4**	39.4	18.0	35.19

Table 1: Impact of **input image size** L.

$\mathcal{M}$	T	MSRVTT Retrieval				MSRVTT-QA Acc.
		R1	R5	R10	MdR	
-	1	10.2	28.6	40.5	17.0	35.73
Mean Pooling	2	**11.3**	31.7	44.9	14.0	**36.02**
	4	10.8	30.0	43.6	14.0	35.83
	8	10.6	**32.5**	**45.0**	**13.0**	35.69
	16	**11.6**	**33.9**	**45.8**	**13.0**	**36.05**
Conv3D	2	8.7	27.3	40.2	17.0	34.85
	16	10.1	28.9	41.7	16.0	35.03
Conv(2+1)D	2	7.3	24.1	35.6	22.0	34.13
	16	9.9	27.3	39.6	17.0	33.92

Table 2: Impact of **#frames** (T) **and temporal fusion function** ($\mathcal{M}$). We use a 1-second clip for all experiments.

Experiments Results (Cont.)

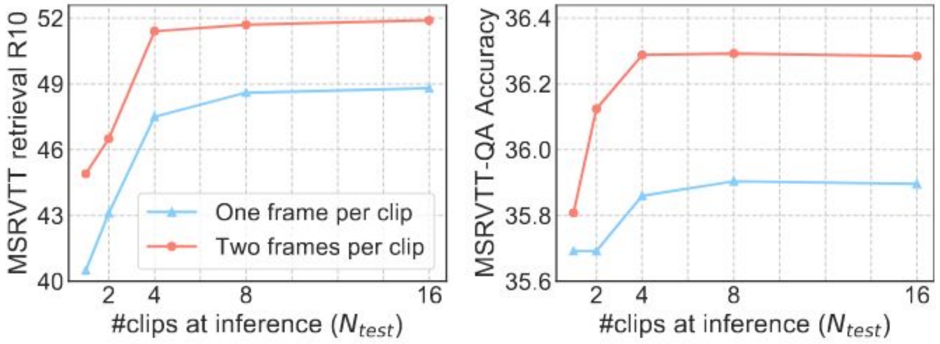


Figure 4: Impact of **#inference clips** (N_{test}).

$\mathcal{G}$	N_{train}	MSRVTT Retrieval				MSRVTT-QA Acc.
		R1	R5	R10	MdR	
-	1	12.7	34.5	48.8	11.0	36.24
Mean Pooling	2	13.3	37.1	50.6	10.0	35.94
	4	14.0	38.6	51.6	10.0	35.40
	8	13.4	36.4	49.7	11.0	35.76
	16	15.2	**39.4**	**53.1**	**9.0**	35.33
Max Pooling	2	8.5	28.7	42.2	14.0	**36.41**
	16	12.5	33.1	46.8	12.0	36.25
LogSumExp	2	**15.5**	38.4	52.6	**9.0**	**36.59**
	16	**17.4**	**41.5**	**55.5**	**8.0**	36.16

Table 3: Impact of **#training clips** (N_{train}) **and score aggregation function** ($\mathcal{G}$). All models use N_{test}=16 clips for inference.

Sampling Method	N_{train}	MSRVTT Retrieval				MSRVTT-QA Acc.
		R1	R5	R10	MdR	
Dense Uniform	16	15.5	39.6	55.0	9.0	35.88
Sparse Random	1	12.7	34.5	48.8	11.0	36.24
	2	15.5	38.4	52.6	9.0	36.59
	4	**15.7**	**41.9**	**55.3**	**8.0**	**36.67**

Table 4: Sparse random sampling *vs.* dense uniform sampling. All models use N_{test}=16 clips for inference.

Experiments Results (Cont.)

Weight Initialization		MSRVTT Retrieval				MSRVTT-QA Acc.
CNN	transformer	R1	R5	R10	MdR	
random	random	0.3	0.4	0.9	506.0	28.05
random	$\text{BERT}_{\text{BASE}}$	0.0	0.2	0.7	505.0	31.72
TSN, K700	$\text{BERT}_{\text{BASE}}$	5.7	22.1	33.1	23.0	35.40
ImageNet	$\text{BERT}_{\text{BASE}}$	7.2	23.3	35.6	21.0	35.01
grid-feat	$\text{BERT}_{\text{BASE}}$	7.4	21.0	30.7	26.0	35.27
image-text pre-training		**10.2**	**28.6**	**40.5**	**17.0**	**35.73**

Table 5: Impact of **weight initialization strategy**.

Parameters Trainable?		MSRVTT Retrieval				MSRVTT-QA Acc.
$\mathcal{F}_v$	$\mathcal{F}_l$	R1	R5	R10	MdR	
✗	✗	8.0	27.2	38.9	17.0	**35.78**
✗	✓	9.0	27.5	39.4	18.0	35.50
✓	✓	**10.2**	**28.6**	**40.5**	**17.0**	35.73

Table 6: Impact of **end-to-end training**.

Experiments Results (Cont.)

Weight Initialization		MSRVTT Retrieval				MSRVTT-QA Acc.
CNN	transformer	R1	R5	R10	MdR	
random	random	0.3	0.4	0.9	506.0	28.05
random	$\text{BERT}_{\text{BASE}}$	0.0	0.2	0.7	505.0	31.72
TSN, K700	$\text{BERT}_{\text{BASE}}$	5.7	22.1	33.1	23.0	35.40
ImageNet	$\text{BERT}_{\text{BASE}}$	7.2	23.3	35.6	21.0	35.01
grid-feat	$\text{BERT}_{\text{BASE}}$	7.4	21.0	30.7	26.0	35.27
image-text pre-training		**10.2**	**28.6**	**40.5**	**17.0**	**35.73**

Table 5: Impact of **weight initialization strategy**.

Parameters Trainable?		MSRVTT Retrieval				MSRVTT-QA Acc.
$\mathcal{F}_v$	$\mathcal{F}_l$	R1	R5	R10	MdR	
✗	✗	8.0	27.2	38.9	17.0	**35.78**
✗	✓	9.0	27.5	39.4	18.0	35.50
✓	✓	**10.2**	**28.6**	**40.5**	**17.0**	35.73

Table 6: Impact of **end-to-end training**.

Conclusion

- Sparse-sampled clips is adequate to represent the whole video
- End-to-end training helps to combat the disconnection problem
- Image-text pre-training helps the model extracts better features on video-text tasks
- Overall, they empirically proved the principle of “*less is more*”

References

[1] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, J. Liu, Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling, in CVPR, 2021.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in CVPR, 2016.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Doll ́ar, C. L. Zitnick, Microsoft COCO captions: Data collection and evaluation server (2015).

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision (2017).

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

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

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

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

[10] 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.

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