VIDEO-AND-LANGUAGE LEARNING VIA SPARSE SAMPLING: AN END-TO-END APPROACH FROM RAW VIDEO PIXELS AND TEXT TOKENS

Lâm Thanh Ngân[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

What ?

We introduce **ClipBERT**, an end-to-end framework for video-and-language learning with three key contributions:

- **Sparse Sampling:** Using only 1-2 frames per clip instead of dense video frames.
- **End-to-End Learning:** Learning directly from raw pixels without pre-extracted features.
- **Image-Text Pre-training:** Leveraging image-text data to transfer to video tasks.

Why ?

- **Expensive Pre-extraction:** Traditional methods require pre-extracted features (ResNet, 3D CNNs), which is costly and inflexible.
- **Computational Overhead:** Processing dense video frames consumes computational resources and storage.
- **Limited Video-Text Data:** Large-scale video-text datasets are scarce compared to abundant image-text data.

Overview

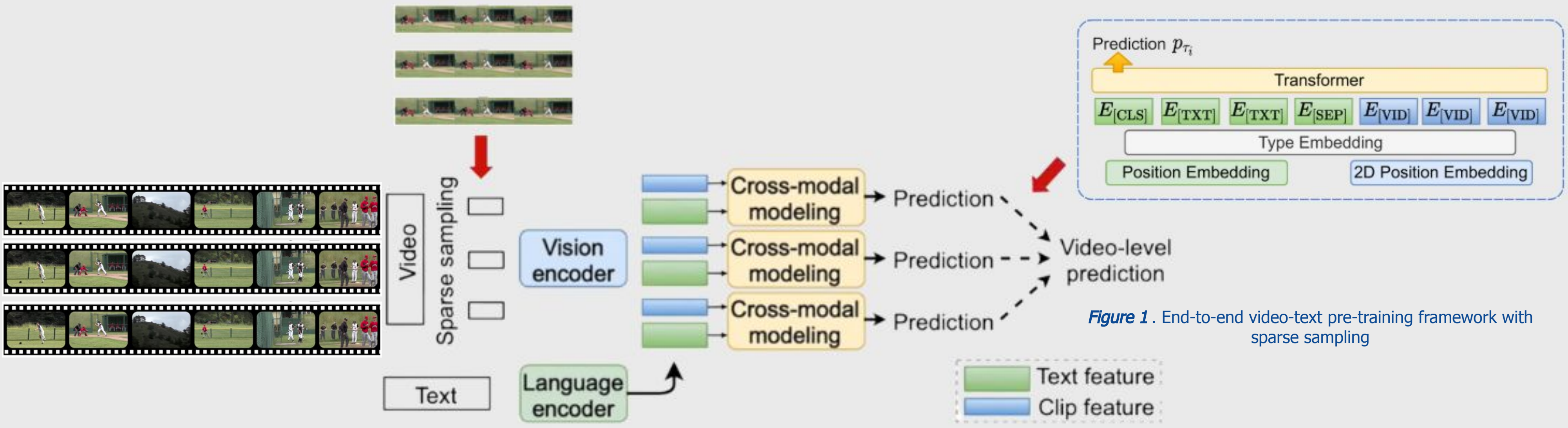

Figure 1. End-to-end video-text pre-training framework with sparse sampling

Description

1. Model Architecture

- **Visual Encoder:** ResNet-50 (2D CNN) extracts 144 grid features per frame, fused via mean-pooling.
- **Cross-modal Transformer:** 12-layer Transformer fuses visual and text features with Type Embedding and 2D Position Embedding.
- **Input Format:** [CLS] + text tokens + [SEP] + visual tokens → unified sequence for Transformer.
- **Why 2D CNN?** Faster, less memory than 3D CNNs, and proven effective on video understanding tasks.

Figure 2. Text feature extraction with word embedding and position embedding.

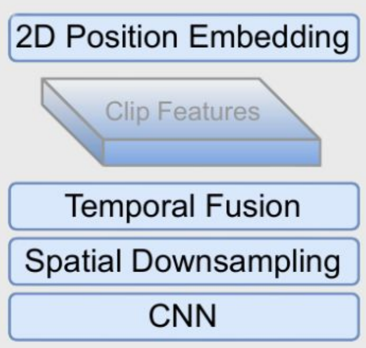

Figure 3. Visual feature extraction pipeline with CNN, spatial downsampling, and temporal fusion.L

2. Training Strategy

- **Sparse Sampling:** Randomly sample 1-4 clips per video; each clip contains only 1-2 frames, acting as data augmentation.
- **Pre-training Data:** COCO Captions + Visual Genome (5.6M image-text pairs).
- **Pre-training Objectives:** Masked Language Modeling (MLM) + Image-Text Matching (ITM).
- **Fine-tuning:** Task-specific heads for Text-to-Video Retrieval and Video QA.

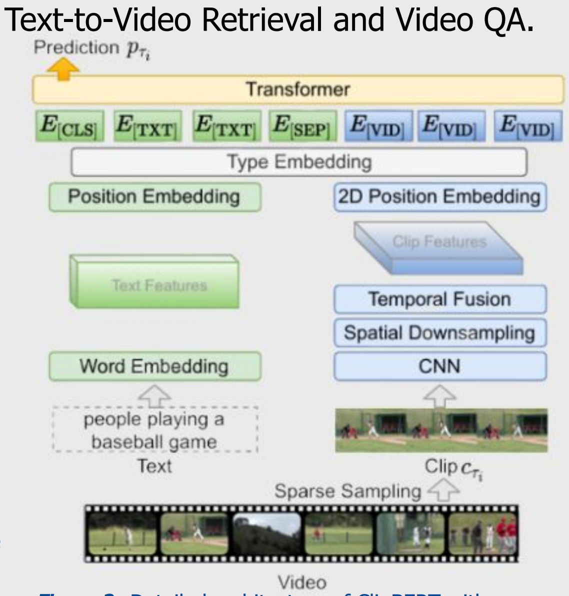

Figure 3. Detailed architecture of ClipBERT with sparse sampling and cross-modal transformer.

3. Experiments & Results

- **Sparse vs Dense:** Sparse random sampling with only 4 clips outperforms dense uniform sampling with 16 clips on both Retrieval and QA tasks.
- **MSRVTT Retrieval:** ClipBERT 8×2 achieves 22.0 R@1, surpassing HERO (16.8) and ActBERT (16.3) by a large margin.

Sampling Method	N_{train}	MSRVTT Retrieval R1	R5	R10	MdR	MSRVTT-QA Acc.
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**

Figure 4. Comparison of different sampling methods on retrieval and QA tasks.

Method	R1	R5	R10	MdR
HERO [35] ASR, PT	20.5	47.6	60.9	-
JSFusion [74]	10.2	31.2	43.2	13.0
HT [43] PT	14.9	40.2	52.8	9.0
ActBERT [80] PT	16.3	42.8	56.9	10.0
HERO [35] 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**

Figure 5. Comparison with state-of-the-art methods on text-to-video retrieval task.

NII

Lâm Thanh Ngân – University of Information Technology, VNU-HCM

TEL : +84 326098634 Email : lamthanhngan751@gmail.com