Paper Presentation

Speaker

Shaon Sutradhar

PhD Candidate

University of A Coruña

Spain

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

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, Jingjing Liu

How do intelligent agents understand visual and textual clues in real world videos?

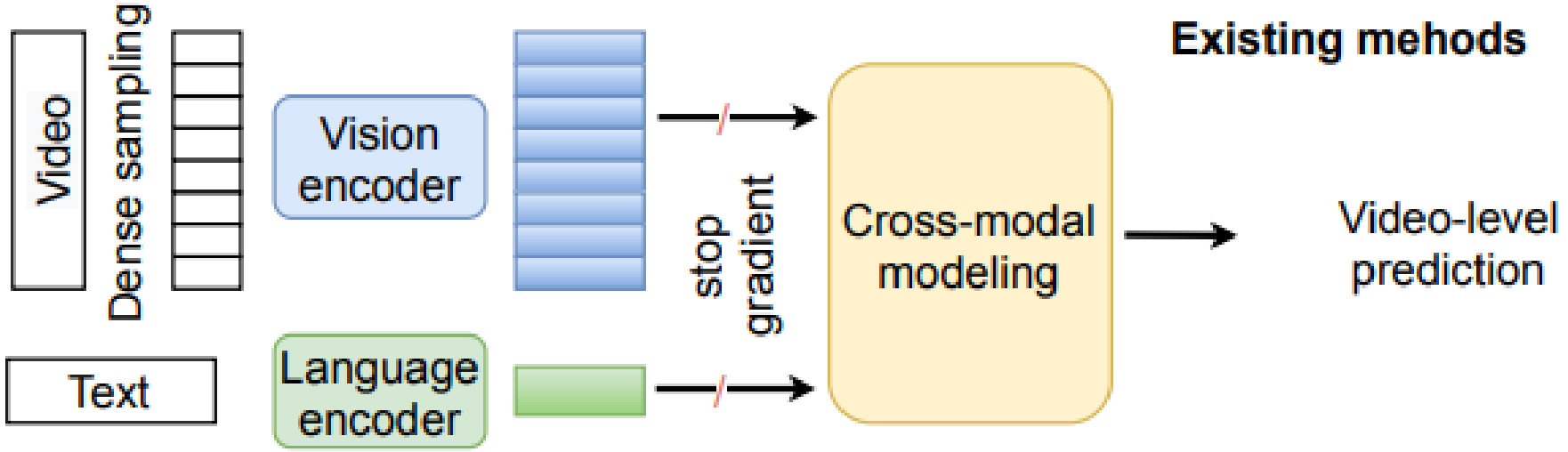
Video and Language Models

- Traditional Cross-modal approaches
  - Text-to-video retrival
  - Video captioning
  - Video question answering
  - Video moment retrival

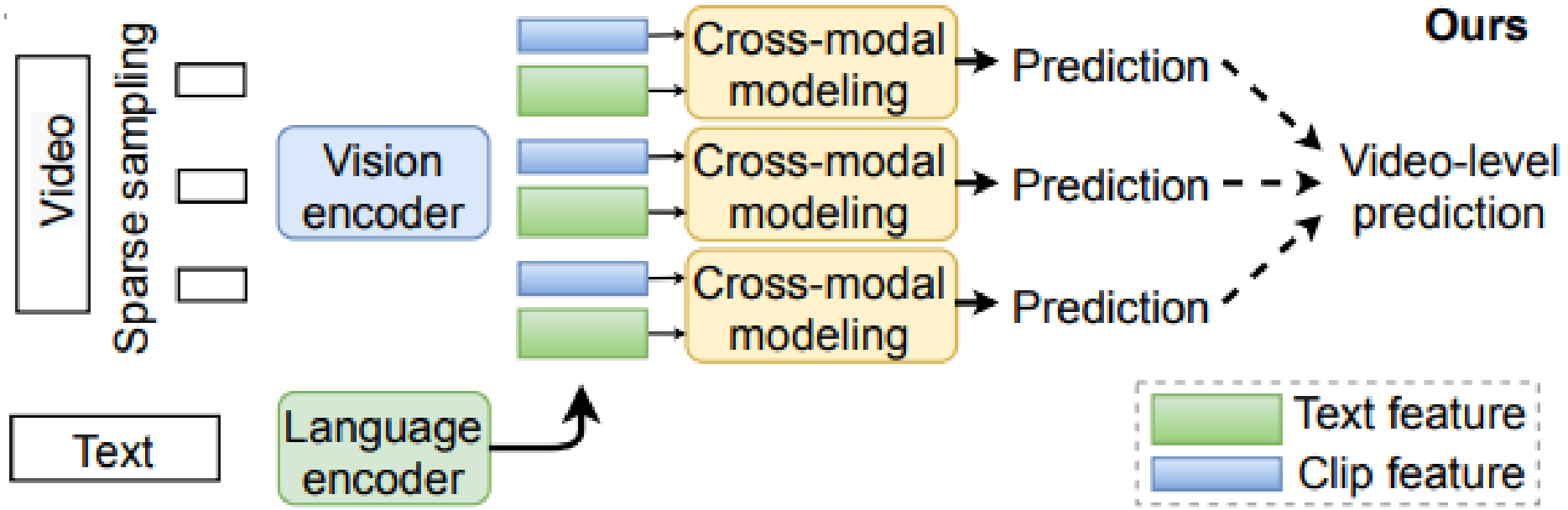
- General Mechanism
  - Step 1: extractdense video features using pre-trained vision models
  - Step 2: extract text features form pre-trained
  - Step 3: complex fusion mechanism for wrangling these static features

How do intelligent agents understand visual and textual clues in real world videos?

Drawbacks of Traditional Cross-modal approaches



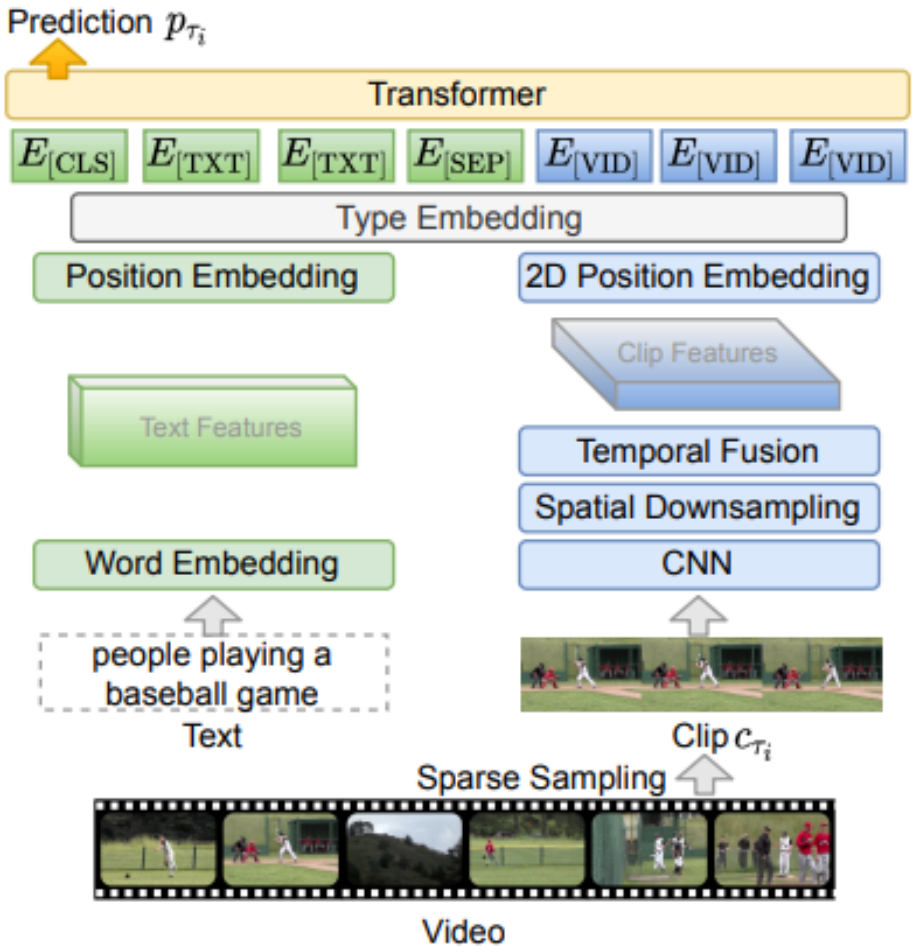
CLIPBERT for Video-and-Language Learning via Sparse Sampling - *Abstract*



CLIPBERT for Video-and-Language Learning via Sparse Sampling - *Architecture*

- **Vision Encoder**
  - 2D ResNet-50
- **Language Encoder, Cross-modal Modeling**
  - BERT based
- **Image-Text Pre-training**
  - COCO and Visual Genome captions

- Benifitted against high cost of video-text pre-training
- Cross-modal feature learning, enabling mitual co-relation between the video and text encoders



CLIPBERT for Video-and-Language Learning via Sparse Sampling - *Experimental Settings*

- Downstream Tasks
  - Text-to-video Retrival
    - Datasets - MSRVTT , DiDeMo, ActivityNet Captions
  - Video Question Answering
    - Datasets - TGIF-QA, MSRVTT-QA, MSRVTT multiple-choice test

Average length of videos in the datasets (in Seconds)

CLIPBERT for Video-and-Language Learning via Sparse Sampling - *Experimental Results*

- Text-to-video retrieval

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.

CLIPBERT for Video-and-Language Learning via Sparse Sampling - *Experimental Results*

- Video Question Answering

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.

Application to Veriff’s identity verification

- **ID Verification Software**
  - Real-time feedback with Assisted Image Capture

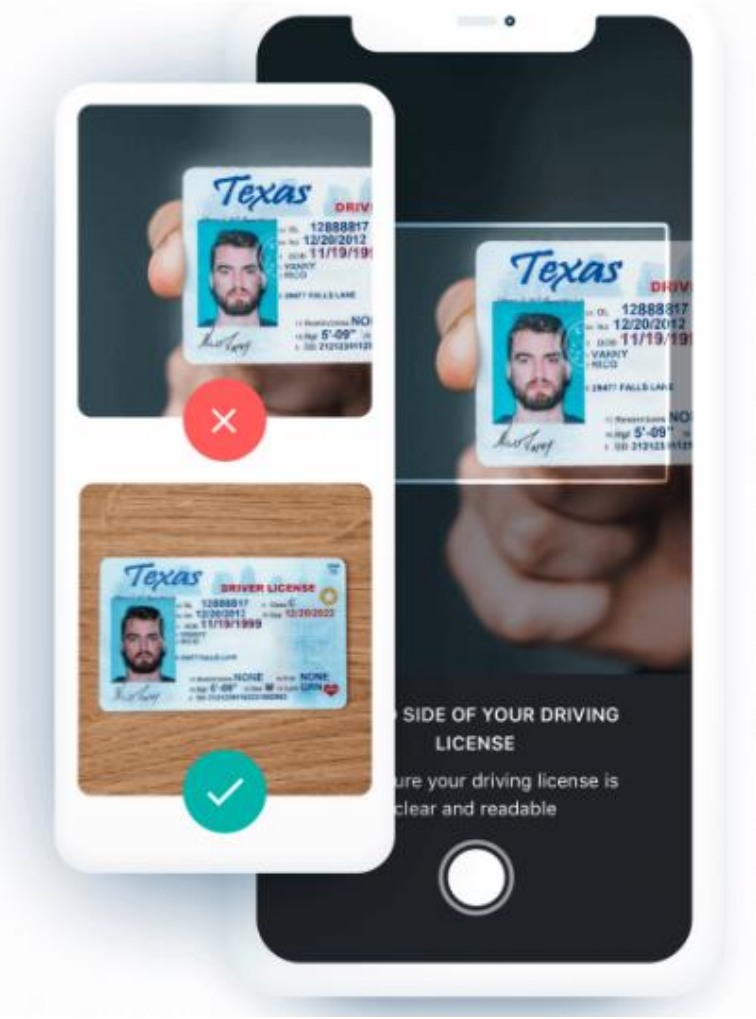


Photo taken from https://www.veriff.com/product/id-verification
Last accessed 19.10.2021

Questions

and

Thank You!