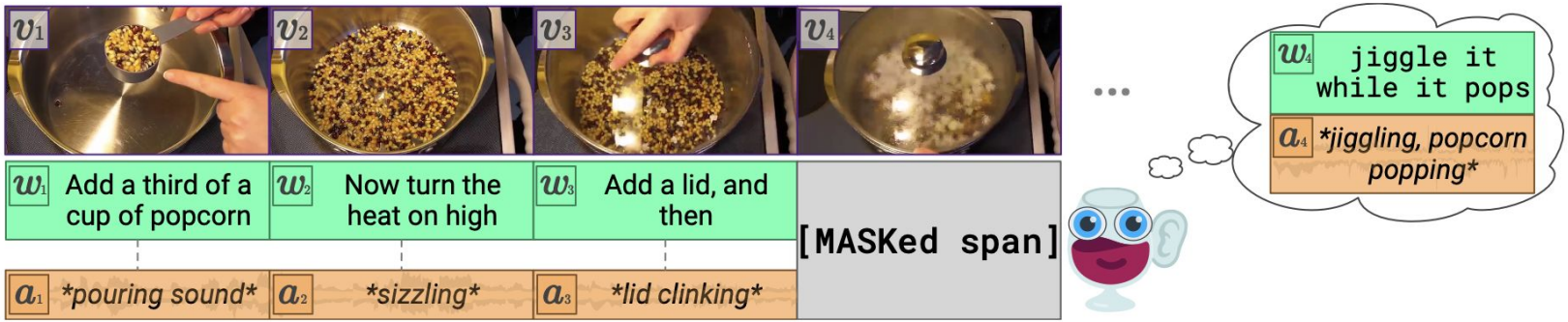
MERLOT RESERVE:
Neural Script Knowledge through Vision and Language and Sound

Rowan Zellers Jiasen Lu Ximing Lu Youngjae Yu Yanpeng Zhao
Mohammadreza Salehi Aditya Kusupati Jack Hessel Ali Farhadi Yejin Choi

Paul G. Allen School of Computer Science & Engineering, University of Washington
Allen Institute for Artificial Intelligence University of Edinburgh

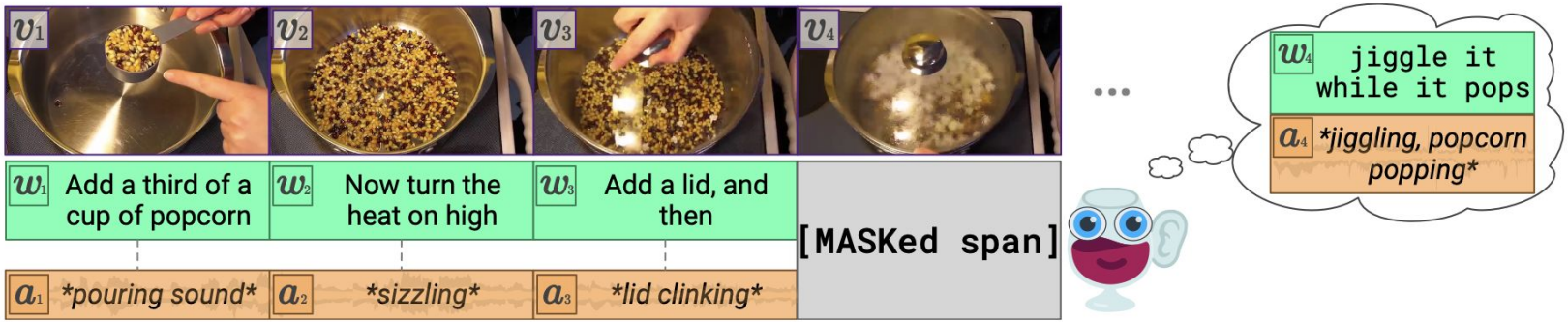
Presented by
Shoubin Yu, Bilen Ghirmai
11/07/2022

Motivation



- The world around us is **dynamic**.
- We experience and learn from it using all of our senses, reasoning over them temporally through ***multimodal script knowledge***.
- Can we build machines that likewise learn vision, language, and sound together?
- Can this paradigm enable learning neural script knowledge, that transfers to language-and-vision tasks, even those without sound?

Motivation



Multimodal **E**vent **R**epresentation **L**earning **O**ver **T**ime,

with **Re**-entrant Sup**ERV**ision of **E**vents

Related Works

- Many language-and-vision tasks benefit from *early fusion* of the modalities (VisualBERT Family). Cross-modal interactions are learned in part through a ***masked language modeling*** (mask LM) objective.

- Those methods lacks audio. And it is limited to representing (and learning from) video frames paired with subtitles.

- Some recent works (MMV) consider audio input, and adpot independent encoders can be combined through *late fusion*.

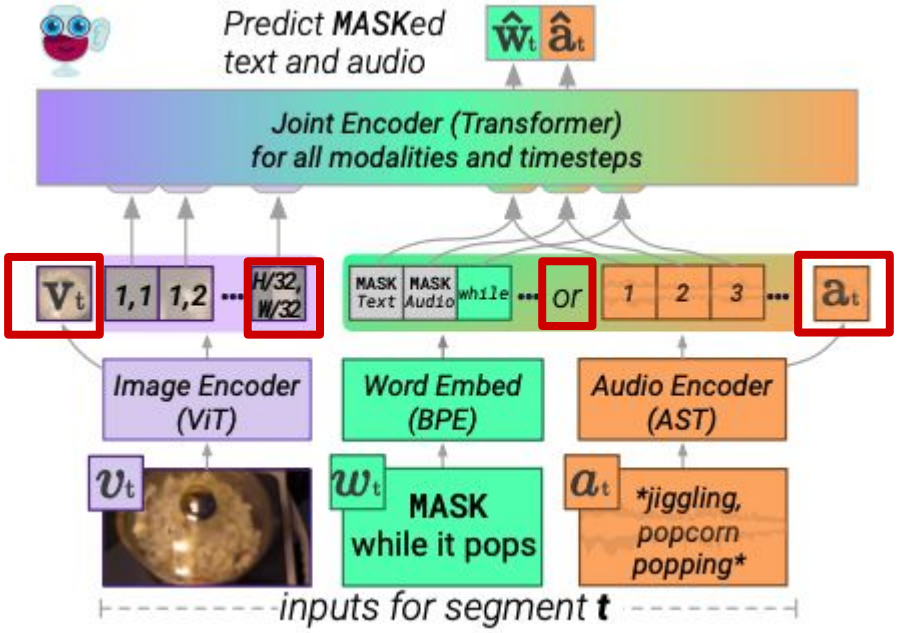
- Some recent works (VATT) consider audio input and adopt *early fusion* stage, but suffering in undesigned loss fuction and noisy data.

Proposed Method — Overview

- Video is dense, so video is spilted into non-overlapping segments in time.

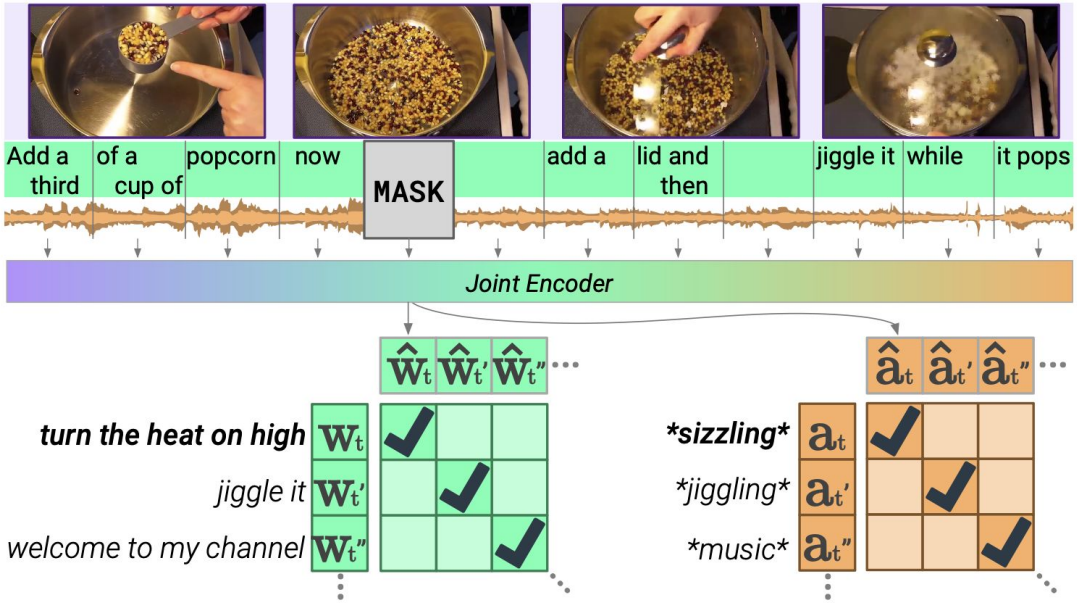
- Each segment contains:
  - v_t : the center frame w_t spoken during the segment
  - w_t : the ASR (Automatic Speech Regnition) token spoken during the segements
  - a_t : the audio of the segments

- The audio a_t in each segments is then divided into three equal-size subsegments.



Proposed Method — Contrastive Span Training

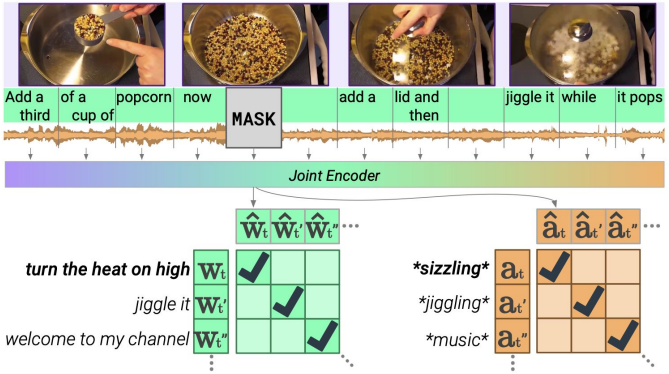
- The model is given a sequence of video segments.
- For each one, the model input includes a video frame and three subsegments that are each either text or audio.
- 25% of those text and audio subsegments are replaced with special [MASK] token.
- In this case, the model predict representation at a higher-level semantic unit than inividial tokens.



Proposed Method — Contrastive Span Training

$$\mathcal{L}_{\text{mask}\to\text{text}} = \frac{1}{|\mathcal{W}|}\sum_{\mathbf{w}_t\in\mathcal{W}}\left(\log\frac{\exp(\sigma\hat{\mathbf{w}}_t\cdot\mathbf{w}_t)}{\sum_{\mathbf{w}\in\mathcal{W}}\exp(\sigma\hat{\mathbf{w}}_t\cdot\mathbf{w})}\right). \quad (1)$$

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{frame}}. \quad (2)$$



- The model minimize the cross entorp between the masked prediction and representation.
- In addition to the masked audio/text objective, the model is also trained to match video frames with a contextualized encoding of the transcript. Thus, the symmetric frame-based loss is also applied to the training.
- The approach also enables the model to learn from both audio and text, while **discouraging** *memorization* of raw perceptual input or tokens.

Proposed Method — Avoid Shortcut Learning

- Early on, authors observed that, given input from from the same modality, and training a model to predict a perceptual modality (like audio/visual), led to ***shortcut learning.***

- ***Shortcut learning***: a low loss but poor representation.

- Authors hypothesize that this setup encourages models to learn imperceptible features, like the *exact mode of microphone* or the *chromatic aberration of the camera lens*.

- The model proposed in this paper avoid this, while they still use audio as a target, by simultaneously training on two kinds of masked videos, to do the comparison.

Experiment — Model & Pretraining Setting

- **Model Variants:**
  1. RESERVE-B, with a hidden size of 768, a 12-layer ViT-B/16 image encoder, and a 12-layer joint encoder. → 200M
  2. RESERVE-L, with a hidden size of 1024, a 24-layer ViT-L/16 image encoder, and a 24-layer joint encoder. → 644M

- **Pretraining Setting:**
  1. **Devices:** TPUv3-512 accelerators. 5 days for RESERVE-B, 16 days for RESERVE-L
  2. **Dataset:** a new dataset training dataset **YT-Temporal-1B** with 20 million English-subtitled YouTube videos and 1 billion frames.

Experiment — Modality Ablation

	Configuration *for one epoch of pretraining*	VCR Q→A val (%)
V+T	Mask LM [29, 106, 128]	67.2
V+T	VirTex-style [27]	67.8
V+T	Contrastive Span	**69.7**
V+T+A	Audio as target	70.4
V+T+A	Audio as input and target	**70.7**
V+T+A	Audio as input and target, w/o strict localization	70.6
	RESERVE-B	**71.9**

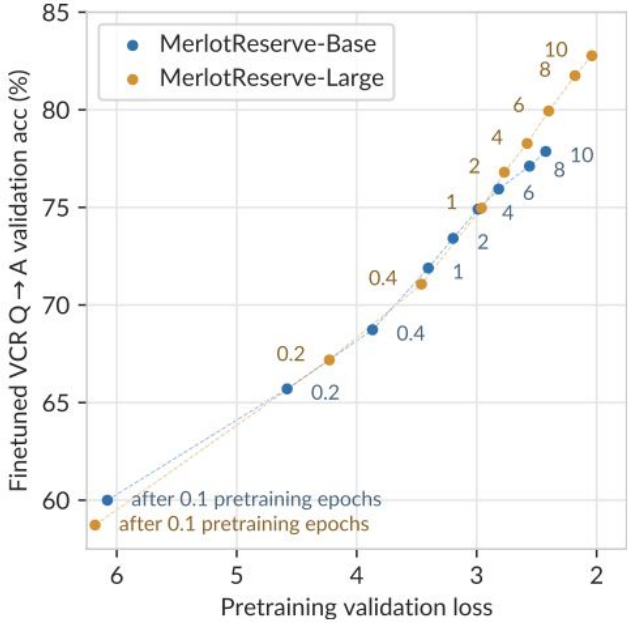
→ Contrastive Span helps ***V***ision + ***T***ext

→ ***A***udio Pretraining helps

- **Mask ML:** independently predict masked tokens.
- **VirTex:** mask text subsegments and extract their hidden states.

- **Audio as target:** only video frames and ASR text as input, the model produce both masked audio and text
- **Audio as input and target:** video, text, audio are given, the model only produce the masked text
- **Sans strict localization:** the adjacent masked regions are also taken into counts

Experiment — Pretraining Progress



- Pretraining RESERVE-B for 9 more epochs boosted performance by 5% and L by 8%.
- Large model shows similar performance with the base one under small size pretraining.

Experiment — VL Tasks

	Model	VCR test (acc; %) Q→A	QA→R	Q→AR
Caption/ObjDet-based	ERNIE-ViL-Large [124]	79.2	83.5	66.3
	Villa-Large [39]	78.9	83.8	65.7
	UNITER-Large [21]	77.3	80.8	62.8
	Villa-Base [39]	76.4	79.1	60.6
	VilBERT [81]	73.3	74.6	54.8
	B2T2 [4]	72.6	75.7	55.0
	VisualBERT [77]	71.6	73.2	52.4
Video-based	MERLOT [128]	80.6	80.4	65.1
	RESERVE-B	79.3	78.7	62.6
	RESERVE-L	**84.0**	**84.9**	**72.0**

	Model	TVQA (acc; %) Val	Test
	Human [75]	–	89.4
Subtitles	MERLOT [128]	78.7	78.4
	MMFT-BERT [109]	73.5	72.8
	Kim et al [68]	76.2	76.1
	RESERVE-B	82.5	–
	RESERVE-L	**85.9**	**85.6**
Audio	RESERVE-B	81.3	–
	RESERVE-L	**85.6**	**84.8**
Both	RESERVE-B	83.1	82.7
	RESERVE-L	**86.5**	**86.1**

Experiment — Video Classification

- The vision only REVERSE-B outperforms MTV-Large (with a larger parameter size).
- It demonstrates better representation quality.
- The gain may come from the new proposed pretraining dataset.

	Model	Kinetics-600 (%) Top-1	Top-5
Vision Only	VATT-Base[2]	80.5	95.5
	VATT-Large [2]	83.6	96.6
	TimeSFormer-L [9]	82.2	95.6
	Florence [125]	87.8	97.8
	MTV-Base [122]	83.6	96.1
	MTV-Large [122]	85.4	96.7
	MTV-Huge [122]	89.6	98.3
	RESERVE-B	88.1	95.8
	RESERVE-L	89.4	96.3
+Audio	RESERVE-B	89.7	96.6
	RESERVE-L	**91.1**	**97.1**

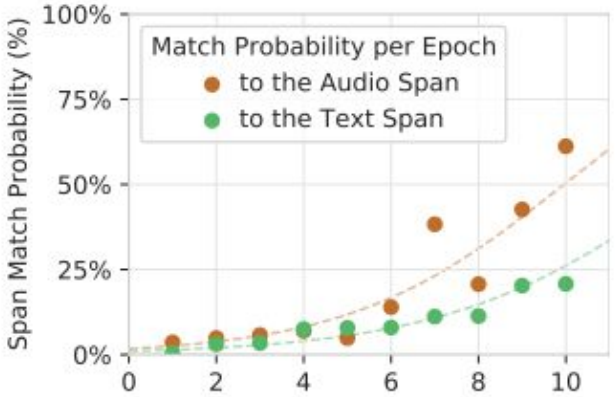
Experiment — Zero-shot Setting

	Model	Situated Reasoning (STAR) (test acc; %)					EPIC-Kitchens (val class-mean R@5; %)			LSMDC (FiB test %)	MSR-VTT QA (test acc %)	
		Interaction	Sequence	Prediction	Feasibility	Overall	Verb	Noun	Action	Acc	top1	top5
	Supervised SoTA	ClipBERT [74]					AVT+ [46]			MERLOT [128]		
		39.8	43.6	32.3	31.4	36.7	28.2	32.0	15.9	52.9	43.1	
zero-shot	Random	25.0	25.0	25.0	25.0	25.0	6.2	2.3	0.1	0.1	0.1	0.5
zero-shot	CLIP (VIT-B/16) [92]	39.8	40.5	35.5	36.0	38.0	16.5	12.8	2.3	2.0	3.0	11.9
zero-shot	CLIP (RN50x16) [92]	39.9	41.7	36.5	**37.0**	38.7	13.4	14.5	2.1	2.3	2.3	9.7
zero-shot	Just Ask (ZS)[123]										2.9	8.8
zero-shot	RESERVE-B	44.4	40.1	38.1	35.0	39.4	17.9	15.6	2.7	26.1	3.7	10.8
zero-shot	RESERVE-L	42.6	41.1	37.4	32.2	38.3	15.6	19.3	4.5	26.7	4.4	11.5
zero-shot	RESERVE-B (+audio)	**44.8**	**42.4**	**38.8**	36.2	**40.5**	20.9	17.5	3.7	29.1	4.0	12.0
zero-shot	RESERVE-L (+audio)	43.9	42.6	37.6	33.6	39.4	**23.2**	**23.7**	**4.8**	**31.0**	**5.8**	**13.6**

- The perverse results on STAR may indicate that the REVERSE learns a biased script knowledge from big data.

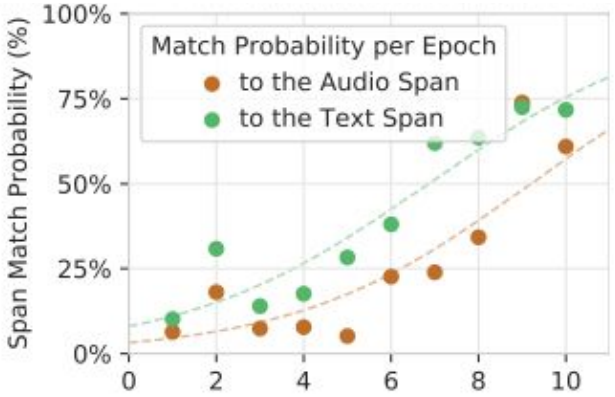
Experiment — Qualitative Analysis

- The plot shows RESERVE-B's probability of correctly identifying the correct audio or text span, as it progress through 10 epochs of pretraining.
- Audio provides orthogonal supervision to text.



Experiment — Qualitative Analysis

- The plot shows RESERVE-B's probability of correctly identifying the correct audio or text span, as it progress through 10 epochs of pretraining.
- Audio provides orthogonal supervision to text.



Conclusion

- This paper introduces RESERVE, which learns jointly through sound, language and vision, guided through a new pretraining objective (contrastive span training).
- This work extends the multimodal representation learning systems to be more semantic-specific.
- A new large-scale video pretraining dataset is also proposed.

Thanks for your attention!
Q & A