Allen AI提出MERLOT，视频理解领域新SOTA！

原创 Yimin_饭煲 夕小瑶的卖萌屋 2022-02-08 09:00

文 | Yimin_饭煲

2021年，多模态领域大概是人工智能研究者们关注者最多的一个领域了。随着各种模态数据集的增长和算力的发展，研究者们开始不断地尝试在一个模型中融合来自各个模态的信息。

而在多模态领域的研究中，和视频相关的任务被认为是最复杂的。

一方面，高质量的视频数据集比图像数据集更加困难，因此数据集的数量和质量往往受限；另一方面，视频数据集中含有文本、图像、语音等多个模态的信息，还要考虑时间线，融合起来比单纯的图像-文本数据更加复杂。

在AI领域久负盛名的Allen研究所向这一复杂的问题发起了挑战，提出了MERLOT系列工作。

第一篇 **MERLOT:Multimodal Neural Script Knowledge Models** 发表于Neurips 2021，使用了大量的视频数据进行自监督预训练，在12个视频问答任务上取得了SOTA；

而第二篇 **MERLOT Reserve:Neural Script Knowledge through Vision and Language and Sound** 则于今年年初刚刚发布，进一步深度融合了视频中的语音信息，在多个任务上又取得了明显提升。下面，就让我们一起来学习这两篇十分Solid的工作吧~

> **论文链接:**
> MERLOT:
> https://arxiv.org/pdf/2106.02636.pdf
> MERLOT Reserve:
> https://arxiv.org/pdf/2201.02639.pdf

主要方法

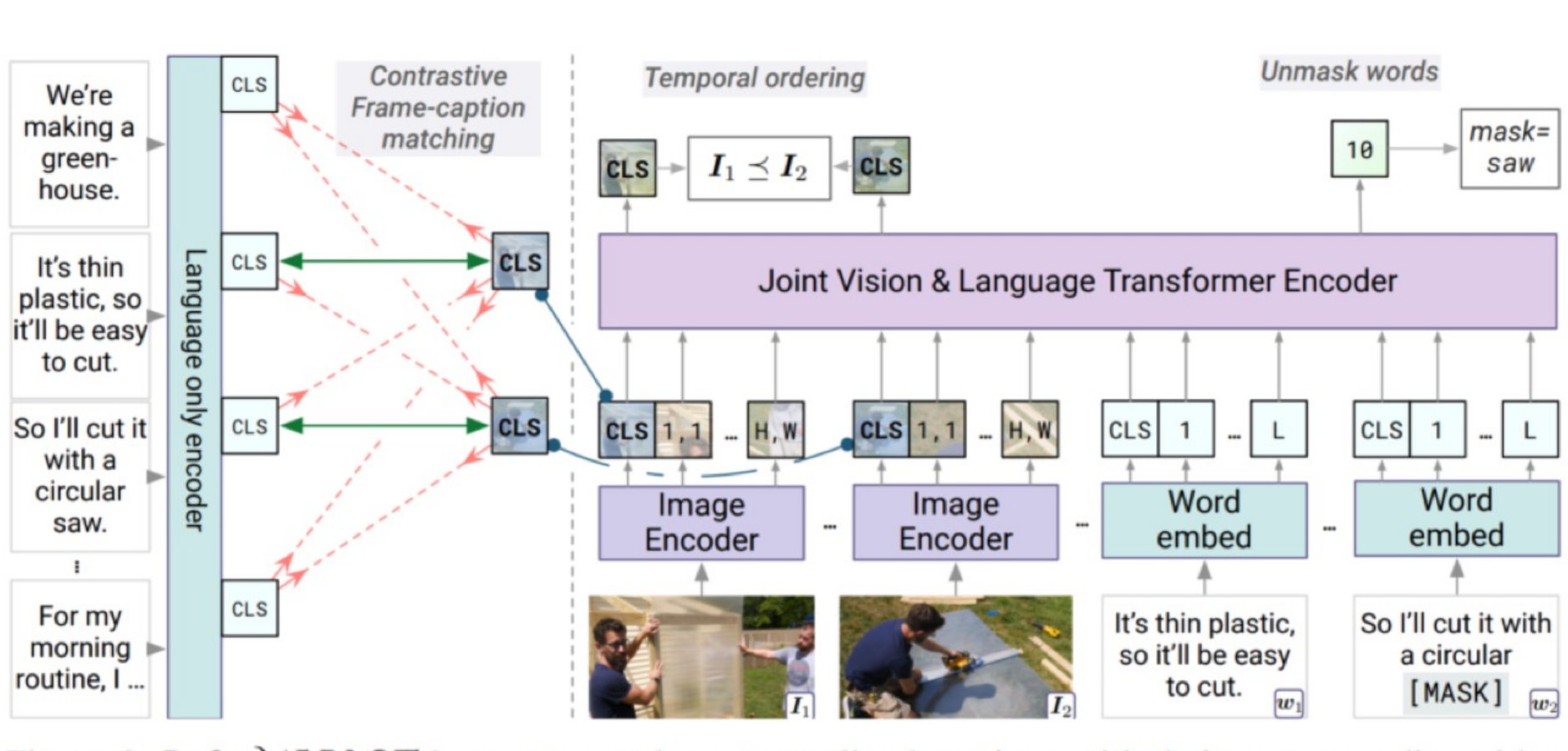

Figure 2: **Left**: MERLOT learns to match contextualized captions with their corresponding video frames. **Right**: the same image encoding is provided, along with (masked) word embeddings, into a joint vision-language Transformer model; it then unmasks ground words (like 'saw' in this example) and puts scrambled video frames into the correct order.

MERLOT这一工作使用了视觉编码器、语言编码器和联合编码器。作者们设计了三个任务来进行优化。

第一个任务是Contrastive Frame-caption matching(标题-帧匹配)，作者们使用视觉编码器编码图片得到的[CLS]表示和文本编码器编码句子得到的[CLS]表示进行对比学习，使得图像编码器具备较好的表示学习性能。

第二个任务是Masked Language Modeling，对模型的语言建模能力进行优化。

第三个任务是Temporal Reordering，在40%的情况下，随机选择一个整数i，从视频输入中的所有帧中随机选取i帧并进行打乱，将位置编码（e.g. $[image_t]$)替换为随机且独特的位置编码(e.g.$[image_{unk_0}]$)。这些随机的位置编码和原有的位置编码分别进行学习，可以让模型学到恢复被扰乱的帧顺序的能力。

这个任务的损失函数是针对一对视频帧(t_i, t_j)拼接隐状态(h_{t_i}, h_{t_j})，使用两层MLP分类器进行二分类(t_i和t_j的前后关系)。

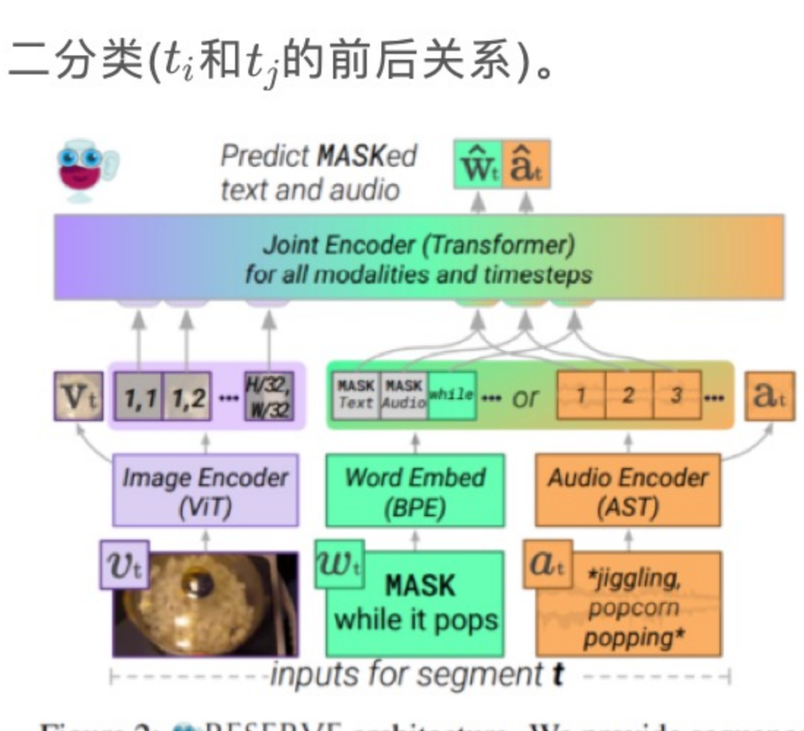

Figure 2: RESERVE architecture. We provide sequence-level representations of video frames, and *either* words or audio, to a joint encoder. The joint encoder contextualizes over modalities and timesteps, to predict what is behind MASK for audio $\hat{a}_t$ and text $\hat{w}_t$. We supervise these predictions with independently encoded targets: a_t from the audio encoder, and w_t from a separate text encoder (not shown).

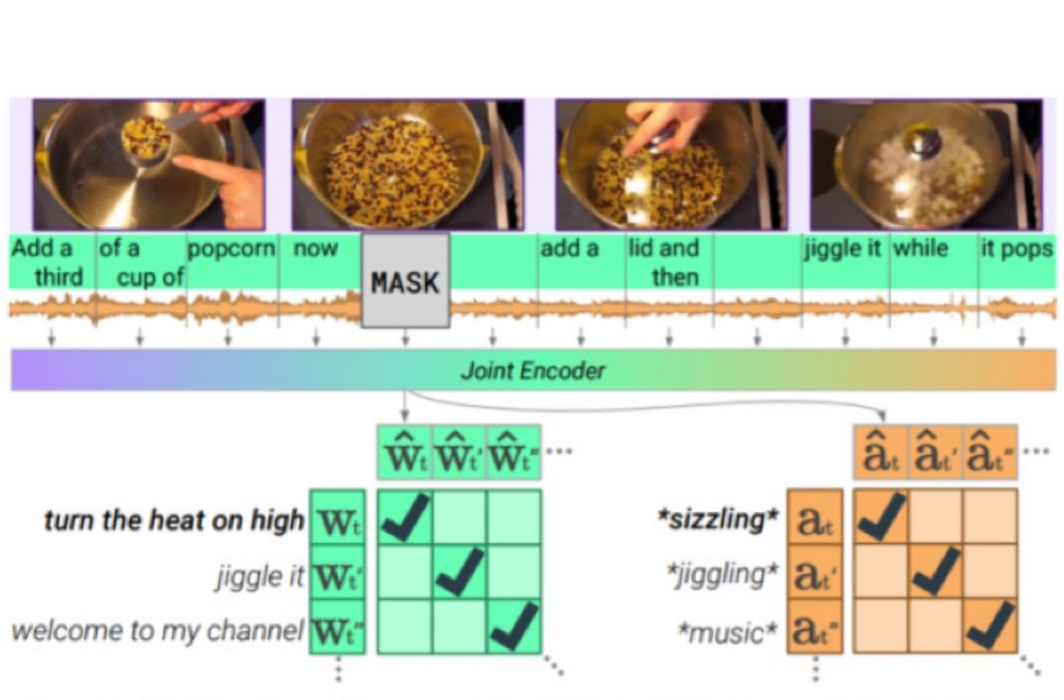

Figure 3: Contrastive span training. Given a video with all modalities temporally aligned, we MASK out a region of text and audio. The model must maximize its similarity *only to* an independent encoding of the text w_t and audio a_t.

MERLOT Reserve这一工作相比于MERLOT引入了新的模态：语音。

为了更好的融合来自三个模态的信息，**作者们提出了更通用，更统一的训练任务。**

对于每一个Batch的输入，只输入视频的帧和文本/语音中的一个（由于文本和语音的信息具有重复性），并且MASK文本/语音中的一部分。作者们提出了对比区域匹配(Contrastive Span Matching)这一任务，给定匹配的视频帧/文本/语音数据，以文本为例，最小化交叉熵损失函数：

$$L_{mask->text} = \frac{1}{|W|}\sum_{w_t \in W}(\log \frac{\exp(\sigma \hat{w}_t w_t)}{\sum_{w \in W}\exp(\sigma \hat{w}_t w)})$$

其中$\hat{w}_t$为[MASK]位置的隐状态表示，w_t为[MASK]掉的信息的隐状态表示，w为Batch中其他样本(负样本)的隐状态表示。同理定义了$L_{text->mask}$，定义：

$$L_{text} = L_{text->mask} + L_{mask->text}$$

同样的，可以定义L_{audio}和L_{frame}。定义总体的损失函数为：

$$L = L_{text} + L_{audio} + L_{frame}$$

作者们还使用了一些技巧来提升得到的特征表示的质量，感兴趣的小伙伴可以去原文细读~。

数据集

对于大规模的预训练工作，除开训练方法之外另一个值得关注的部分就是使用的数据集了。

MERLOT收集的数据集为YT-Temporal-180M，从600万公开的YouTube视频中抽取得到。

作者们选取的数据集比起HowTo100M和VLOG等局限于特定领域的数据集范围更大，主题更广。

后来的实验表明，**如果仅使用HowTo100M这样的数据集进行训练，会降低模型在下游任务上的性能。**

Dataset	VCR
Conceptual ∪ COCO	58.9
HowTo100M	66.3
YT-Temporal-180M	**75.2**
HowTo100M-sized YT-Temporal-180M	72.8
YTT180M, raw ASR	72.8

(d) **Diverse (video) data is important.** Applying our architecture to caption data leads to poor results. Our model performs better on HowTo100M, yet still below our (more diverse) YT-Temporal-180M, even when controlled for size. Using raw ASR (vs. denoised ASR) reduces performance.

在MERLOT Reserve这一工作中，作者们扩充了数据集，**提出了YT-Temporal-1B数据集，包含2000万Youtube视频**，进一步提升了数据集的多样性，而模型强大的性能也说明了扩充数据的有效性。

结果

作者们通过大量的实验证实了MERLOT和MERLOT Reserve的有效性。**对于MERLOT模型，作者们在12个视频问答数据集上开展了实验，大幅度刷新了SOTA。**

Tasks	Split	Vid. Length	ActBERT [127]	ClipBERT$_{8x2}$ [67]	SOTA	MERLOT
MSRVTT-QA	Test	Short	-	37.4	41.5 [118]	**43.1**
MSR-VTT-MC	Test	Short	88.2	-	88.2 [127]	**90.9**
TGIF-Action	Test	Short	-	82.8	82.8 [67]	**94.0**
TGIF-Transition	Test	Short	-	87.8	87.8 [67]	**96.2**
TGIF-Frame QA	Test	Short	-	60.3	60.3 [67]	**69.5**
LSMDC-FiB QA	Test	Short	48.6	-	48.6 [127]	**52.9**
LSMDC-MC	Test	Short	-	-	73.5 [121]	**81.7**
ActivityNetQA	Test	Long	-	-	38.9 [118]	**41.4**
Drama-QA	Val	Long	-	-	81.0 [56]	**81.4**
TVQA	Test	Long	-	-	76.2 [56]	**78.7**
TVQA+	Test	Long	-	-	76.2 [56]	**80.9**
VLEP	Test	Long	-	-	67.5 [66]	**68.4**

Table 3: Comparison with state-of-the-art methods on video reasoning tasks. MERLOT outperforms state of the art methods in **12** downstream tasks that involve short and long videos.

MERLOT Reserve模型同样也有非常强大的表现，**在视频常识推理数据集上取得了不小的性能提升，超过了许多使用了其他监督信息的模型。**

	Model	VCR (test) Q→A	QA→R	Q→AR
Caption/ObjDet-based	ERNIE-ViL-Large [123]	79.2	83.5	66.3
	Villa-Large [39]	78.9	83.8	65.7
	UNITER-Large [20]	77.3	80.8	62.8
	Villa-Base [39]	76.4	79.1	60.6
	ViLBERT [80]	73.3	74.6	54.8
	B2T2 [4]	72.6	75.7	55.0
	VisualBERT [76]	71.6	73.2	52.4
Video-based	MERLOT [126]	80.6	80.4	65.1
	RESERVE-B	79.3	78.7	62.6
	RESERVE-L	**84.0**	**84.9**	**72.0**

Table 2: RESERVE gets **state-of-the-art leaderboard performance on VCR.** We compare it with the largest submitted single models, including image-caption models that utilize heavy manual supervision (e.g. object detections and captions).

由于MERLOT Reserve使用了大量的数据进行了自监督预训练，因此**在零样本学习上有着不错的性能，在STAR数据集上相比于有监督的SOTA都有着明显的提升。**

	Model	Situated Reasoning (STAR) (test acc; %) Interaction	Sequence	Prediction	Feasibility	Overall	EPIC-Kitchens (val class-mean R@5; %) Verb	Noun	Action	LSMDC (FiB test %) Acc	MSR-VTT QA (test acc %) top1	top5
	Supervised SoTA	ClipBERT [73]					AVT+ [45]			MERLOT [126]		
		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
	CLIP (VIT-B/16) [91]	39.8	40.5	35.5	36.0	38.0	16.5	12.8	2.3	2.0	3.0	11.9
	CLIP (RN50x16) [91]	39.9	41.7	36.5	**37.0**	38.7	13.4	14.5	2.1	2.3	2.3	9.7
	Just Ask (ZS)[122]										2.9	8.8
	RESERVE-B	44.4	40.1	38.1	35.0	39.4	17.9	15.6	2.7	26.1	3.7	10.8
	RESERVE-L	42.6	41.1	37.4	32.2	38.3	15.6	19.3	4.5	26.7	4.4	11.5
	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
	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**

Table 4: Zero shot results. On STAR, RESERVE obtains state-of-the-art results, outperforming finetuned video models. It performs well on EPIC-Kitchens (verb and noun forecasting), along with LSMDC, despite their long-tail distributions. On MSR-VTT QA, it outperforms past work on weakly-supervised video QA. Further, it outperforms CLIP (that cannot handle dynamic situations), and benefits from audio when given.

结语

多模态技术的发展和商业化，也许将会带来人工智能应用的新一轮爆发。以往的多模态应用面临着模型架构复杂、缺少数据、缺少算力等一系列问题，而随着Transformer结构一统天下，互联网上各模态数据的井喷式增长，计算资源越发普及，这些问题都在慢慢得到解决。

MERLOT系列工作刷新了我们认知中视频理解领域的上限，向我们展示了视频、语音、文本多模态高效融合的一种可能性。未来，让我们一起努力朝着多模态领域的"BERT"模型进发吧！

萌屋作者：Yimin_饭煲
在微软NLC组搬砖的联培博士生，爱好摄影和运动，希望卖萌屋早日开通视频业务，我来当摄影师！

作品推荐

1. 学完文本知识，我就直接看懂图片了！
2. 别再Prompt了！谷歌提出tuning新方法，强力释放GPT-3潜力
3. 大模型炼丹无从下手？谷歌、OpenAI烧了几百万刀，总结出这些方法论…