MERLOT

MULTIMODAL NEURAL SCRIPT
KNOWLEDGE MODELS

Presenters: Alex Chandler and Smriti Singh

03-08-2023

Motivation

- Our capacity for common-sense reasoning is shaped by our understanding and experiences of causes and effects over time.
- Look at the image. What do we understand by looking at it?
- Machines struggle with this kind of script knowledge, and the ability to develop temporal reasoning by looking at such images.
- How can we teach a model to understand temporal information from only an image at inference time?

Challenges

- What makes this such a difficult problem to solve?
- We can’t teach machines to learn this kind of knowledge by enumerating every single fact, inference and counterfactual about a number of these images.
- Typically, the state of the art models for tasks like VCR are trained on image-caption pairs. What does this say about the capabilities of such models for temporal reasoning?

MERLOT: Introduction

- To address these challenges. the authors introduce MERLOT: Multimodal Event Representation Learning Over Time
- This is a model that learns common sense representations of multimodal events by self-supervised pre-training over 6M unlabelled YouTube videos
- MERLOT is trained to:
    - match individual video frames with contextualized representations of the associated transcripts
    - contextualize these frame level representations over time by unmasking distant word level corruptions and reordering scrambled video frames.

Related Work

- Joint Representation of written texts and images:
  - VisualBERT for image captioning
  - Problem: Static Images and heavily reliant on human annotation

- Learning from Videos with ASR transcripts:
  - In the past, web videos with ASR have been used to build weakly-supervised object detectors, action detectors/classifiers, video captioners and video reference resolvers.
  - ClipBERT model learns vision-language representations from image captions
  - More recent work has looked to learn transferable multimodal representations from uncurated sets of “How to Videos” for video understanding tasks
  - Problem: How can we design an appropriate objective for learning video-level representations?
  - Prior approaches like mostly rely on ResNet pretrained on ImageNet

- Temporal ordering and forecasting
  - “What happens next?” tasks
  - Problem: Prior methods mostly use only Visual modality

Dataset: YT-Temporal-180M

- YT-Temporal-180M is a dataset for learning multimodal script knowledge, derived from 6 million public YouTube Videos. The videos span many domains, datasets and topics.
- Diverse corpus: Encourages the model to learn about a board range of objects, actions and scenes
- Building the dataset: The authors began with 27M unique Video IDs and removed:
  - Videos without an English ASR track
  - Videos > 20 mins
  - Videos that belong to visually ungrounded categories
  - Videos that have thumbnails unlikely to contain objects
- Punctuation is added to ASR by applying a seq2seq model trained on news articles.
- The authors perform denoising on the transcripts by using the GROVER LLM, a large language model exclusively pre-trained on written text from news articles.

Dataset: contd...

- Each video V in this dataset could contain 1000s of frames. V is represented as a sequence of consecutive *video segments* {st}
- Each segment st consists of:
  - An image frame It, extracted from the middle timestep of the segment
  - The words wt spoken during the segment, with a total length of L tokens (L =32)
- To split the videos into segments, BPE is performed on each video transcript and the tokens are aligned with YouTube's word level timestamps.
- The dataset consists of 180M such segments

Some useful information about the validity of the dataset:

- Authors focus on gathering data while protecting privacy - they focus on public videos, not personal ones
- Performed a validity check on 100 randomly sampled videos, did topic modelling on 7k randomly sampled documents

Architecture Overview

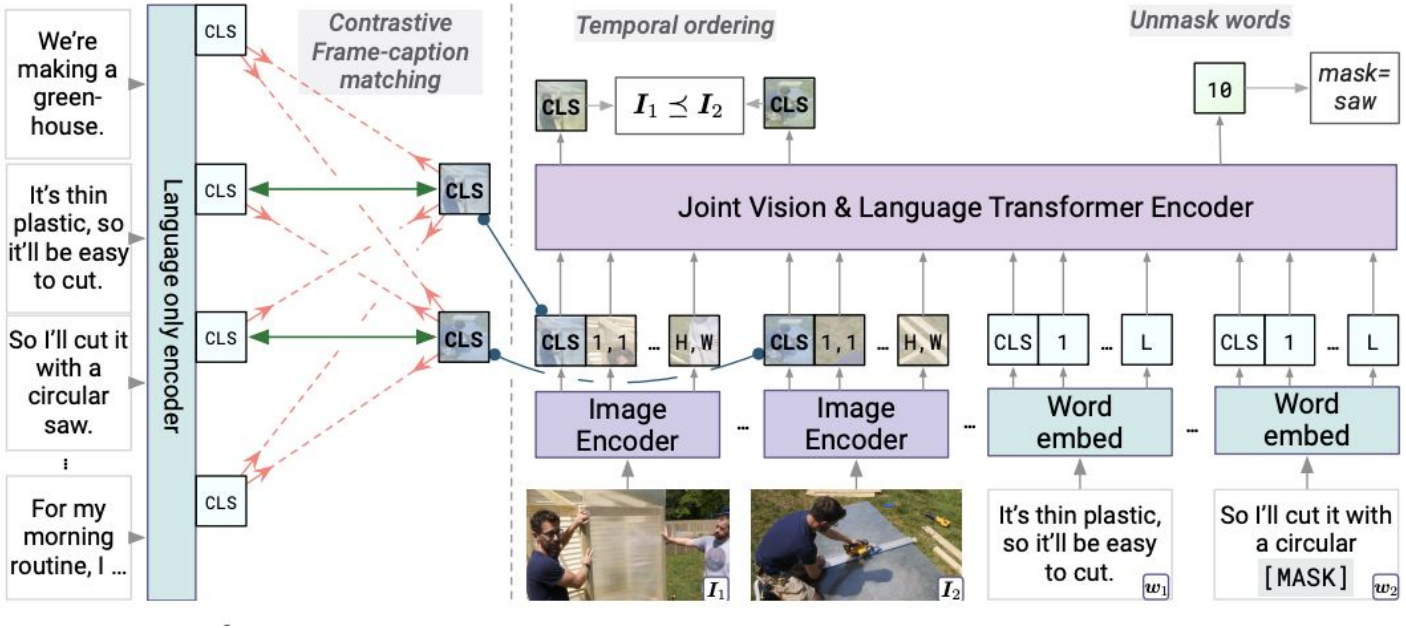
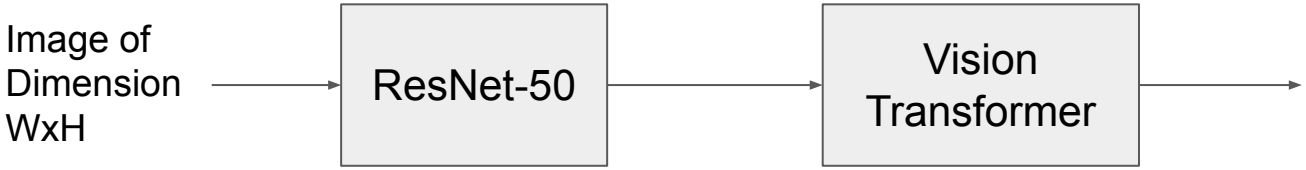
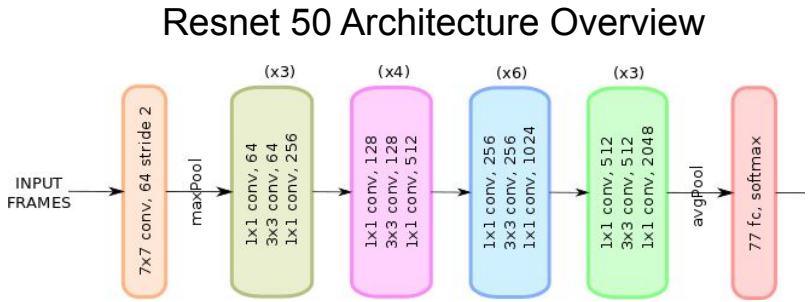
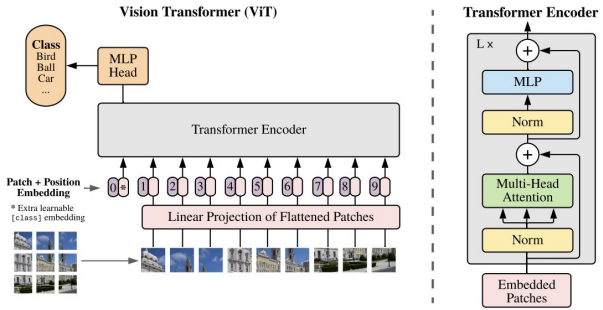


Image Encoder



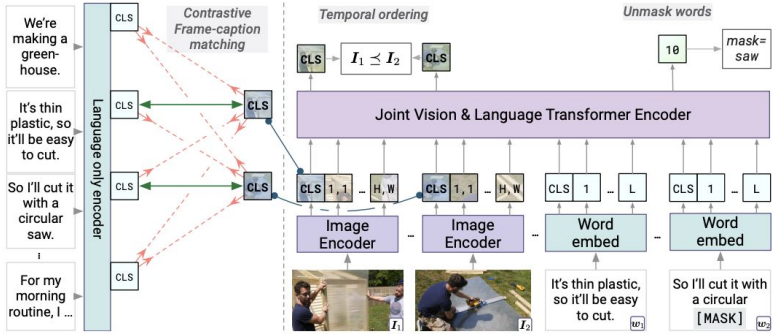
1. W/32×H/32 feature map
2. CLS hidden state one for pooling a global representation of the image
3. CLS hidden state for pretraining

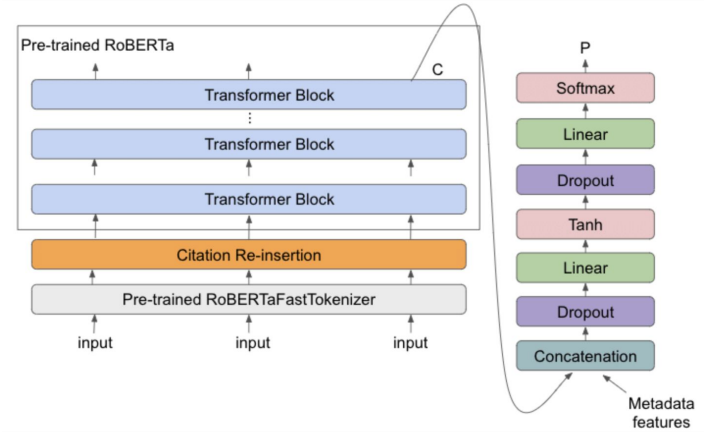




Joint Vision-Language Encoder

- Fine-tune RoBERTa base architecture
  - Note: RoBERTa is a 12-layer, 768-dimensional Transformer
  - Note: Joint Vision-Language Encoder is initialized with pretrained RoBERTa weights.
- Embed the tokens {wt} via lookup, and then add position embeddings to both language and vision components (i.e., {It}).
- Pass the independent visual and textual feature maps to our joint encoder.





Pre-training Tasks and Objectives

- Contrastive frame-transcript matching
  - Use the video transcript to compute a ‘language-only’ representation of each video segment
  - Use a contrastive loss to maximize its similarity to corresponding representations from the image encoder

Contrastive Loss Steps:

1. Represent encoded image and caption in size-768 hidden state
2. Apply L-2 Normalization
3. Compute an all-pairs dot-product between all image and text representations
4. Divide the logits by a hyperparameter
5. Apply a pairwise cross entropy loss to encourage matching captions and frames

Pre-training Tasks and Objectives Continued

Temporal Re-ordering Steps:

1. Scramble video frames replacing segment-level position embeddings with random position-embeddings
    a. Note: Random position embeddings are learned
2. Compute reordering loss:
    a. Extract hidden states from each frame at the CLS token position
    b. Concatenate hidden states for each frame and pass through two-layer MLP
    c. Optimize with cross-entropy loss

- (Attention) Masked Language Modeling
    - Randomly replace 20% with a MASK token, a random word, or the same word
    - Use cross-entropy loss function to reconstruct words
    - Introduction of **attention masking**

Experiments and Evaluation

- Transferred MERLOT to 14 downstream tasks
  - 12 of 14 tasks were video reasoning
- Tested model on three VCR settings
  - Outperforms models with heavy supervision (even models including bounding boxes)
- Compared with CLIP and UNITER
- Demonstrates SOTA on zero-shot transfer and finetuning

	Q→A	QA→R	Q→AR
ViLBERT [75]	73.3	74.6	54.8
Unicoder-VL [68]	73.4	74.4	54.9
VLBERT [69]	73.8	74.4	55.2
UNITER [22]	75.0	77.2	58.2
VILLA [36]	76.4	79.1	60.6
ERNIE-ViL [119]	77.0	80.3	62.1
MERLOT (base-sized)	**80.6**	**80.4**	**65.1**

Table 1: Results on VCR [123]. We compare against SOTA models of the same 'base' size as ours (12-layer vision-and-language Transformers). MERLOT performs best on all metrics.

	Spearman (↑)	Pairwise acc (↑)	Distance (↓)
CLIP [89]	.609	78.7	.638
UNITER [22]	.545	75.2	.745
MERLOT	**.733**	**84.5**	**.498**

Table 2: Results unscrambling SIND visual stories[50, 2]. Captions are provided in the correct order; models must arrange the images temporally. MERLOT performs best on all metrics by reasoning over the entire story, instead of independently matching images with captions.

Experiments and Evaluation Continued

- Details of 12 Downstream Tasks
  - sample a sequence of 5 to 7 still frames from each video clip
  - initialize new parameters only to map the model's pooled CLS hidden state into the output labels
  - finetune MERLOT with a softmax cross entropy loss

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.

Ablation Studies

- SpanBERT masking after attention masking helps performance through ablation study
  - SpanBERT idea: randomly corrupt tokens in each direction of mask to mitigate existence of BPE artifacts
- Six main ablation studies given in table 4

Training setup	VCR	TVQA+
One segment (N=1)	73.8	75.2
One segment, attention masking	73.5	74.5
Four segments	74.1	73.3
🍷 Four segments, attention masking	**75.2**	**75.8**

(a) **Context helps together with attention masking.** Pretraining on more segments at once improves performance, but more context can encourage language-only representation learning. Attention masking counteracts this, giving an additional 1 point boost.

Training setup	VCR	TVQA+
No contrastive V-L loss	57.5	67.6
No temporal ordering loss	**75.5**	75.6
🍷 All losses	75.2	**75.8**

(b) **Contrastive V+L loss is crucial.** Removing it makes performance drop significantly; the temporal ordering loss is not as important for downstream finetuning.

	VCR
No boxes	74.8
🍷 Drawn-on boxes	**79.4**

(c) **Drawing on bounding boxes helps**, suggesting that our model uses it to decode the 'referring expression' information (e.g. `person1`).

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.

# epochs	VCR
5 epochs	75.2
10 epochs	75.9
20 epochs	77.0
30 epochs	78.5
🍷 40 epochs	**79.4**

(e) **Training for longer helps,** with performance increasing monotonically over training iterations.

Table 4: Ablation study on the validation set of VCR question answering ($Q \rightarrow A$) and TVQA+, in accuracy (%). We put a 🍷 next to the configurations we chose for MERLOT.

Critique: Dataset

Strengths

- The dataset spans many domains, covering a lot of topics
- The authors did not limit the dataset only to instructional videos
- An effort to protect privacy: Giving users the right to opt out of the dataset

Limitations:

- The dataset consists videos only in English
- Large scale dataset - likely contains offensive/biased content (“do not advocate for deployment”)
- Filtering criteria: Thumbnails are not always necessarily accurate
- Video descriptions are repetitive and may have incorrect captions
- Sanity checks happen on very limited portions of the dataset

Critique: Model Architecture

- Strengths
  - MERLOT Image Encoder 50X faster than Faster-RCNN
  - Initializing Joint Vision-Language Encoder with RoBERTa weights gives encoder wide range of natural language processing capabilities, and likely reduces overall training time
- Limitations
  - Lack of analysis as to why grid-based hybrid ResNet/Vision Transformer was best choice, other than efficiency reasons
    - ResNet is a relatively outdated architecture
  - Lacking details about combining ResNet and Vision Transformer
  - Removing the C5 block in ResNet-50 and adding average-pooling the final-layer region cells has unknown, but likely negative, impact on performance
  - Model doesn’t learn from audio

Critique: Pre-Training

Strengths

- Pre-training objectives ensure full stack visual reasoning from recognition subtasks
- Effective Loss function
  - Contrastive Loss combined with temporal reordering yields strong results for downstream tasks

Limitations:

- Potential loss in intra-video coherence when the treatment of all other frame-caption pairs in the batch as negative samples, even if they come from the same video
- Fairly computationally expensive pre-training process - 30 hours on a v3-1024 TPU pod.

Critique: Experiments and Evaluation

- Strengths
  - Tested on wide variety of Image and Video Reasoning Tasks
  - Ablation studies highlighted importance component of model architecture, pre-training techniques, loss function, and dataset.
  - Provided detailed examples and explanations in the appendix for various tasks
- Limitations
  - Lack of human evaluation
  - Lack of experimentation on architectural changes
    - Could have experimented with other Image Encoders and Joint Vision & Language Transformer Encoders
      - Although it is reasonable to assume that they decided not to due to energy use
    - Could have measured importance of initializing with RoBERTa weights

Conclusion

- Presents MERLOT, an end-to-end vision and language model that learns without labeled data
- Provides YT-Temporal-180M, a new filtered dataset of 6 million YouTube videos
- Demonstrates SOTA zero-shot and fine tuned performance
- Achieves SOTA performance on tasks requiring event and temporal reasoning over both static images and videos

What we would have done differently

- Expanded the dataset to consist of multilingual videos
- Use sound as an extra input
- Included human evaluation for the model task performance evaluation
- Checked for racial/sexist bias in dataset
- Experimented with transformer only architecture
  - Use shifted window approach for computing self-attention (SWIN like architecture)
- Licensed the dataset, given that the authors do not advocate for deployment

Future Work for Paper / Reading

- Suggested future work by authors:
  - “1) exploring finer-grained temporal reasoning pretraining objectives vs. frame ordering e.g., a temporal frame localization within transcripts”
  - “2) learning multilingually from non-English videos and communities on YouTube.”
- Existing Future works:
  - MERLOT Reserve: Neural Script Knowledge Through Vision and Language and Sound
  - Flamingo: a Visual Language Model for Few-Shot Learning

Thank You