MERLOT RESERVE

Presenter: Chiawei Tang
Department of Computer Science
Virginia Tech
02/20/2023

https://reurl.cc/Dmv8dR
Outline
what we will lean in this presentation

• MERLOT
• VATT
• MERLOT RESERVE
MERLOT
Multimodal Event Representation Learning Over Time
Visual Commonsense Reasoning

- what might happen next
- what are people’s intentions
Visual Commonsense Reasoning

- what might happen next
- what are people’s intentions

Vision
Language
Audio
Self-Supervised Learning

Supervised

\[ x \overset{\text{Model}}{\rightarrow} y \overset{\text{Loss}}{\rightarrow} \hat{y} \]

Self-supervised

\[ x \overset{\text{Model}}{\rightarrow} x' \overset{\text{Loss}}{\rightarrow} x'' \]
Self-Supervised Learning

Self-supervised

Model

\[ y \rightarrow \text{Loss} \rightarrow x'' \]

\[ x' \]

\[ x \]

\[ z_i \leftarrow \text{Maximize agreement} \rightarrow z_j \]

\[ g(\cdot) \]

\[ h_i \leftrightarrow \text{Representation} \leftrightarrow h_j \]

\[ f(\cdot) \]

\[ \tilde{x}_i \rightarrow t \sim \mathcal{T} \rightarrow \tilde{x}_j \]

\[ \tilde{x}_i \rightarrow t' \sim \mathcal{T} \rightarrow \tilde{x}_j \]

SimCLR: Chen et al, 2020
Self-Supervised Learning

Self-supervised

Model

\[ y \xrightarrow{\text{Loss}} x'' \]

\[ x' \]

\[ x \]

MoCo He et al. 2020

https://medium.com/geekculture/understanding-contrastive-learning-and-moco-efe491e4eed9
Self-Supervised Learning

- Masked token prediction
- Next sentence prediction

Model for Task 1

Model for Task 2

Model for Task 3

Model for Zero-shot Task

细调

零样本
Self-Supervised Learning

How to empower Vision Transformers with large-scale, unlabeled data?

https://www.youtube.com/watch?v=rgXxAFIBido
Self-Supervised Learning

How to empower Vision Transformers with large-scale, unlabeled data?

Self-supervision from the multimodal videos (video frames, audio, and text)

https://www.youtube.com/watch?v=rgXxAFIBido
A model that learns commonsense representations of multimodal events by self-supervised pretraining over 6M unlabelled YouTube videos
Dataset

HowTo100M

To learn about a broad range of objects, actions, and scenes

YouTube: Popular Topics

With ENGLISH ASR
Not too long (> 20 minutes)
Visually “ungrounded”
(video games commentaries)
Unlikely to contain objects

https://www.di.ens.fr/willow/research/howto100m/
https://web.eecs.umich.edu/~fouhey/2017/VLOG/index.html
The Office

32 Byte Pair Encode (BPE) tokens each

Segment 1 - $s_1$

Segment 2 - $s_2$

Segment 3 - $s_3$
Segment 3 - $s_3$

- $I_1$
- $I_{t-n}$
- $I_t$

$w_1$  At 8 a.m. today, someone poisons the coffee.
$w_2$  Do not drink the coffee.
$w_3$  No~~~
...

- an image frame $I_t$, extracted from the middle timestep of the segment
- the words $w_t$ spoken during the segment, with a total length of $L$ tokens.
At 8 a.m. today, someone poisons the coffee.
Different image size for videos

MERLOT

Image Encoder

Vision Transformer (ViT)

Transformer Encoder

Patch + Position Embedding
* Extra learnable [class] embedding

Linear Projection of Flattened Patches

Trainable

1. Split an image into patches
2. Flatten the patches
3. Produce lower-dimensional linear embeddings from the flattened patches
4. Add positional embeddings
5. Feed the sequence as an input to a standard transformer encoder
6. Pretrain the model with image labels
7. Finetune on the downstream dataset for image classification

https://theaisummer.com/vision-transformer/
MERLOT

- Contrastive frame-transcript matching
- (Attention) Masked Language Modeling
- Temporal Reordering
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular saw.

For my morning routine, I …
We’re making a greenhouse. It’s thin plastic, so it’ll be easy to cut. So I’ll cut it with a circular saw.

For my morning routine, I …
Contrastive Loss is one of the first training objectives that was used for contrastive learning. It takes as input a pair of samples that are either similar or dissimilar, and it brings similar samples closer and dissimilar samples far apart.
Contrastive loss is one of the first training objectives that was used for contrastive learning. It takes as input a pair of samples that are either similar or dissimilar, and it brings similar samples closer and dissimilar samples far apart.

\[
\ell_{i,j} = - \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \left[ k \neq i \right] \exp(\text{sim}(z_i, z_k)/\tau)},
\]

- similar vectors to be as close to 1 as possible, since \(-\log(1) = 0\)
- negative examples to be close to 0, since any non-zero values will reduce the value of similar vectors

\[
\text{Softmax } \sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ and } z = (z_1, \ldots, z_K) \in \mathbb{R}^K
\]
MERLOT

Contrastive frame-transcript matching

We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular saw.

For my morning routine, I …
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular [MASK].

mask = saw
We’re making a greenhouse. It’s thin plastic, so it’ll be easy to cut. So I’ll cut it with a circular [MASK].

 problems?

mask = saw
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular [MASK].
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular \textbf{[MASK]}.

people tend to ramble, and often mention key objects multiple times

\textbf{Attention Masking}

\texttt{mask} = \texttt{saw}
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular saw.

For my morning routine, I ...

So I will cut it with a circular saw.
We’re making a greenhouse.

It’s thin plastic, so it’ll be easy to cut.

So I’ll cut it with a circular saw.

For my morning routine, I ...

SpanBERT
• masking contiguous random spans, rather than random tokens
• training the span boundary representations to predict the entire content of the masked span.
MERLOT

Temporal Reordering

Replace segment-level position embeddings

\[ \text{image}_t \rightarrow \text{image}_\text{unk}_0 \]
MERLOT
Temporal Reordering

He was almost to the top.

His kids were already at the top.

Reordering Loss

$\text{concat}(h_{t_i}, h_{t_j})$ \(\rightarrow\) MLP

Cross Entropy

$t_i < t_j \text{ or } t_i > t_j$
MERLOT
Temporal Reordering

Reordering Loss

concat($h_t^i, h^i_t$) → MLP → Cross Entropy

$t_i < t_j$ or $t_i > t_j$

Temporal ordering

Joint Vision & Language

Image Encoder

CLS $I_1 \preceq I_2$ CLS

CLS $1,1$ $H,W$ CLS $1,1$ $H,W$

Image Encoder
Image Tasks

Visual Commonsense Reasoning

- what might happen next
- what are people’s intentions

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone’s order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person2] is delivering food to the table, and she might not know whose order is whose.
## Image Tasks

### Visual Commonsense Reasoning

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

### Unsupervised ordering of Visual Stories

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman (↑)</th>
<th>Pairwise acc (↑)</th>
<th>Distance (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP [89]</td>
<td>.609</td>
<td>78.7</td>
<td>.638</td>
</tr>
<tr>
<td>UNITER [22]</td>
<td>.545</td>
<td>75.2</td>
<td>.745</td>
</tr>
<tr>
<td><strong>MERLOT</strong></td>
<td><strong>.733</strong></td>
<td><strong>84.5</strong></td>
<td><strong>.498</strong></td>
</tr>
</tbody>
</table>
Video Tasks

TVQA

TVQA links depicted objects to visual concepts in questions and answers.

- what might happen next
- what are people’s intentions

https://tvqa.cs.unc.edu/
## Video Tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Split</th>
<th>Vid. Length</th>
<th>ActBERT [127]</th>
<th>ClipBERT_{8x2} [67]</th>
<th>SOTA</th>
<th>MERLOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRVTT-QA</td>
<td>Test</td>
<td>Short</td>
<td>-</td>
<td>37.4</td>
<td>41.5</td>
<td>[118]</td>
</tr>
<tr>
<td>MSR-VTT-MC</td>
<td>Test</td>
<td>Short</td>
<td>88.2</td>
<td>-</td>
<td>88.2</td>
<td>[127]</td>
</tr>
<tr>
<td>TGIF-Action</td>
<td>Test</td>
<td>Short</td>
<td>-</td>
<td>82.8</td>
<td>82.8</td>
<td>[67]</td>
</tr>
<tr>
<td>TGIF-Transition</td>
<td>Test</td>
<td>Short</td>
<td>-</td>
<td>87.8</td>
<td>87.8</td>
<td>[67]</td>
</tr>
<tr>
<td>TGIF-Frame QA</td>
<td>Test</td>
<td>Short</td>
<td>-</td>
<td>60.3</td>
<td>60.3</td>
<td>[67]</td>
</tr>
<tr>
<td>LSMDC-FiB QA</td>
<td>Test</td>
<td>Short</td>
<td>48.6</td>
<td>-</td>
<td>48.6</td>
<td>[127]</td>
</tr>
<tr>
<td>LSMDC-MC</td>
<td>Test</td>
<td>Short</td>
<td>-</td>
<td>-</td>
<td>73.5</td>
<td>[121]</td>
</tr>
<tr>
<td>ActivityNetQA</td>
<td>Test</td>
<td>Long</td>
<td>-</td>
<td>-</td>
<td>38.9</td>
<td>[118]</td>
</tr>
<tr>
<td>Drama-QA</td>
<td>Val</td>
<td>Long</td>
<td>-</td>
<td>-</td>
<td>81.0</td>
<td>[56]</td>
</tr>
<tr>
<td>TVQA</td>
<td>Test</td>
<td>Long</td>
<td>-</td>
<td>-</td>
<td>76.2</td>
<td>[56]</td>
</tr>
<tr>
<td>TVQA+</td>
<td>Test</td>
<td>Long</td>
<td>-</td>
<td>-</td>
<td>76.2</td>
<td>[56]</td>
</tr>
<tr>
<td>VLEP</td>
<td>Test</td>
<td>Long</td>
<td>-</td>
<td>-</td>
<td>67.5</td>
<td>[66]</td>
</tr>
</tbody>
</table>
The old man was riding the escalator.

He was almost to the top.

His kids were already at the top.

Some police were at the top. It was a train station.

They then got on the bus.

I went to the fair with my kids last weekend.

There were a lot of people there.

They also had a barn.

We got to see a lot of animals.

We can't wait to go back later.
VATT

Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text
VATT

One backbone share different modalities

https://openreview.net/pdf?id=RzYrn625bu8
• Video, audio, and text inputs have respective feature extractors
• Each feature extractor has different architecture according to the modality.
VATT

Modality-Specific

• Video, audio, and text inputs have respective feature extractors
• Each feature extractor has different architecture according to the modality.

https://www.youtube.com/watch?v=V3gY_hyATU8
VATT

Modality-Agnostic

https://www.youtube.com/watch?v=V3gY_hyATU8
Multi-modal features need to be projected to common space for feature fusion, but different modalities have different levels of semantic granularity.

- Vision & audio: fine-grained space
- Vision + audio & text: the lower dimensional coarse-grained space.
VATT

Multimodal Contrastive Learning

\[ \mathcal{L} = \text{NCE}(z_{v,va}, z_{a,va}) + \lambda \text{MIL-NCE}(z_{v,vt}, \{z_{t,vt}\}) \]

\[ \text{NCE}(z_{v,va}, z_{a,va}) = -\log \left( \frac{\exp(z_{v,va}^T z_{a,va} / \tau)}{\exp(z_{v,va}^T z_{a,va} / \tau) + \sum_{z' \in \mathcal{N}} \exp(z'_{v,va}^T z'_{a,va} / \tau)} \right) \]

\[ \text{MIL-NCE}(z_{v,vt}, \{z_{t,vt}\}) = -\log \left( \frac{\sum_{z_{t,vt} \in \mathcal{P}} \exp(z_{v,vt}^T z_{t,vt} / \tau)}{\sum_{z_{t,vt} \in \mathcal{P}} \exp(z_{v,vt}^T z_{t,vt} / \tau) + \sum_{z' \in \mathcal{N}} \exp(z'_{v,vt}^T z'_{t,vt} / \tau)} \right) \]
VATT

Multimodal Contrastive Learning

• Vision & Text: Multiple-Instance-Learning-NCE (MIL-NCE) loss
• For multiple positive pairs of video & text, a video is matched to multiple text inputs that are temporally close to the video input.
MERLOT RESERVE

Neural Script Knowledge through Vision and Language and Sound
At 8 a.m. today, someone poisons the coffee. Do not drink the coffee. No~~~ ...

- A frame $v_t$, from the middle of the segment
- The ASR tokens $w_t$ spoken during the segment
- The audio $a_t$ of the segment.
As the text $w_t$ was automatically transcribed by a model given audio $a_t$, it is reasonable to assume that it contains strictly less information content. Thus, for each segment $s_t$, the paper provides models with exactly one of text or audio.
Contrastive Loss

AST: Audio Spectrogram Transformer

MERLOT RESERVE

Audio Encoder

Split the audio $\alpha_t$ in each segment into three equal-sized subsegments.

Do not drink the coffee.
Do not drink the coffee.

- Sound of water splashing
- Speech of person saying the timer is starting
Do not drink the coffee.
Do not drink the coffee.
MERLOT RESERVE

Do not drink the coffee.

Frame Encoder

Audio Encoder

Vision-Language-Audio Temporal Encoder

Spilling coffee

Do not drink the coffee.

MASK
Add a third of a cup of popcorn

Now turn the heat on high

Add a lid, and then

*pouring sound*

*sizzling*

*lid clinking*

... jiggle it while it pops

*jiggling, popcorn popping*
Do not drink the coffee.
Maximize similarity $a_1$ to versus all other targets

$$\log \frac{\exp(\hat{a}_1, a_i)}{\sum_i \exp(\hat{a}_1, a_i)}$$
MERLOT RESERVE

Maximize similarity $a_1$ to versus all other targets

$$\log \frac{\exp(\hat{a}_1, a_i)}{\sum_i(\hat{a}_1, a_i)}$$
At 8 a.m. today, someone poisons the coffee.

Do not drink the coffee.

Print out the memo.

Maximize similarity $t_1$ to versus all other targets

$$
\log \frac{\exp(\hat{t}_1, t_i)}{\sum_i (t_1, t_i)}
$$
Contrastive Span Training

maximize the similarity to the encodings of the text $w_t$ and audio $a_t$

$$L_{\text{mask} \to \text{text}} = \frac{1}{|W|} \sum_{w_t \in W} \left( \log \frac{\exp(\sigma \hat{w}_t \cdot w_t)}{\sum_{w \in W} \exp(\sigma \hat{w}_t \cdot w)} \right)$$
$L = L_{text} + L_{audio} + L_{frame}$
Ablations

Visual Commonsense Reasoning

- contrastive span pretraining outperforms mask LM
- improved performance when audio is used both as input and target

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

**Visual Commonsense Reasoning**

**Dataset - YT-Temporal-1B.**

![Pretraining progress](image)

The joint encoder is a 12-layer, 768-dimensional Transformer.

<table>
<thead>
<tr>
<th>Model</th>
<th>Q→A</th>
<th>QA→R</th>
<th>Q→AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERNIE-ViL-Large [124]</td>
<td>79.2</td>
<td>83.5</td>
<td>66.3</td>
</tr>
<tr>
<td>Villa-Large [39]</td>
<td>78.9</td>
<td>83.8</td>
<td>65.7</td>
</tr>
<tr>
<td>UNITER-Large [21]</td>
<td>77.3</td>
<td>80.8</td>
<td>62.8</td>
</tr>
<tr>
<td>Villa-Base [39]</td>
<td>76.4</td>
<td>79.1</td>
<td>60.6</td>
</tr>
<tr>
<td>VilBERT [81]</td>
<td>73.3</td>
<td>74.6</td>
<td>54.8</td>
</tr>
<tr>
<td>B2T2 [4]</td>
<td>72.6</td>
<td>75.7</td>
<td>55.0</td>
</tr>
<tr>
<td>VisualBERT [77]</td>
<td>71.6</td>
<td>73.2</td>
<td>52.4</td>
</tr>
<tr>
<td><strong>MERLOT [128]</strong></td>
<td>80.6</td>
<td>80.4</td>
<td>65.1</td>
</tr>
<tr>
<td><strong>RESERVE-B</strong></td>
<td>79.3</td>
<td>78.7</td>
<td>62.6</td>
</tr>
<tr>
<td><strong>RESERVE-L</strong></td>
<td><strong>84.0</strong></td>
<td><strong>84.9</strong></td>
<td><strong>72.0</strong></td>
</tr>
</tbody>
</table>
Video Tasks

**TVQA**

TVQA links depicted objects to visual concepts in questions and answers.

**Table: Results on TVQA**

<table>
<thead>
<tr>
<th>Model</th>
<th>TVQA (acc; %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human [75]</td>
<td>–</td>
</tr>
<tr>
<td>MERLOT [128]</td>
<td>78.7</td>
</tr>
<tr>
<td>MMFT-BERT [109]</td>
<td>73.5</td>
</tr>
<tr>
<td>Kim et al [68]</td>
<td>76.2</td>
</tr>
<tr>
<td><strong>SUBTITLES</strong></td>
<td></td>
</tr>
<tr>
<td>RESERVE-B</td>
<td>82.5</td>
</tr>
<tr>
<td>RESERVE-L</td>
<td><strong>85.9</strong></td>
</tr>
<tr>
<td><strong>AUDIOS</strong></td>
<td></td>
</tr>
<tr>
<td>RESERVE-B</td>
<td>81.3</td>
</tr>
<tr>
<td>RESERVE-L</td>
<td><strong>85.6</strong></td>
</tr>
<tr>
<td><strong>BOTH</strong></td>
<td></td>
</tr>
<tr>
<td>RESERVE-B</td>
<td>83.1</td>
</tr>
<tr>
<td>RESERVE-L</td>
<td><strong>86.5</strong></td>
</tr>
</tbody>
</table>

https://tvqa.cs.unc.edu/
Activity Recognition

**Kinetics-600**

The **Kinetics-600** is a large-scale action recognition dataset which consists of around 480K videos from 600 action categories. The 480K videos are divided into 390K, 30K, 60K for training, validation and test sets, respectively.

Results on Kinetics-600

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>VATT-Base [2]</td>
<td>80.5</td>
<td>95.5</td>
</tr>
<tr>
<td>VATT-Large [2]</td>
<td>83.6</td>
<td>96.6</td>
</tr>
<tr>
<td>TimeSFormer-L [9]</td>
<td>82.2</td>
<td>95.6</td>
</tr>
<tr>
<td>Florence [125]</td>
<td>87.8</td>
<td>97.8</td>
</tr>
<tr>
<td>MTV-Base [122]</td>
<td>83.6</td>
<td>96.1</td>
</tr>
<tr>
<td>MTV-Large [122]</td>
<td>85.4</td>
<td>96.7</td>
</tr>
<tr>
<td>MTV-Huge [122]</td>
<td>89.6</td>
<td>98.3</td>
</tr>
<tr>
<td><strong>RESERVE-B</strong></td>
<td>88.1</td>
<td>95.8</td>
</tr>
<tr>
<td><strong>RESERVE-L</strong></td>
<td>89.4</td>
<td>96.3</td>
</tr>
<tr>
<td><strong>RESERVE-B</strong></td>
<td><strong>91.1</strong></td>
<td><strong>97.1</strong></td>
</tr>
</tbody>
</table>

https://www.deepmind.com/open-source/kinetics
Zero-Shot

- Situated Reasoning (STAR): The model is given a video, a templated question, and 4 answer choices.
- Action Anticipation in Epic Kitchens: Here, the goal is to predict future actions given a video clip.
- LSMDC: Videos with captions (with a MASK to be filled in).
- MSR-VTT QA: Open-ended video QA.
Why does audio help?

and forth every now and then *popcorn popping*

hadn't no control over but i knew that i could control how my room looked what

*exhausted laugh*

weight on the legs and get more stretch in the calves in these 45...

this time we're holding it with the right leg
Strength

- A new, extensive dataset named YT-Temporal-1B has been introduced as a competitor to HowTo100M and YT-Temporal-180M, which offers broader content coverage beyond just instructional videos.
- A new multimodal model to acquire knowledge from vast amounts of videos featuring accompanying vision, language and audio, achieving new SOTA.
- A new objective learning function that maximize the similarity of positive text and audio pair

https://openreview.net/forum?id=CRFSrgYtV7m
Weakness

• The paper discusses the model training process using millions of YouTube videos, which has sparked privacy apprehensions. Moreover, accessing the dataset may be challenging, as the only provided links to the associated YouTube videos are unlikely to ensure long-term availability of the data.

• The visual representations are not evaluated

• Since the model only utilizes a single frame per segment as input, it remains unclear how it can effectively represent dynamic scenes.
  • we would argue that in the context of learning multimodal script knowledge, giving the model a single frame might actually be a strength. The reason being is that our goal is for MERLOT to infer what’s going on in the world, temporally, through (partial) observations of both vision and language.

https://openreview.net/forum?id=CRFSrgYtV7m
Future Works

• An end-to-end model with feature extractions
• Multiple frames per segment
• Contrastive loss to align image and text/audio pair

https://openreview.net/forum?id=CRFSrgYtV7m
Q&A