Video-and-Language Pre-training

Luowei Zhou

06/20/2021
Outline

• Data as fuel – The rise of pre-training data
• Method Overview and Taxonomy
• Reconstructive Methods and Contrastive Methods
• Video-Language-\emph{Audio} – The new favorite?
• From image to video and back
• Downstream Tasks and Results
• Video-And-Language Understanding Evaluation (VALUE) benchmark
• Conclusion
Pre-training isn’t new

• In fact, it is rather pervasive!

Figure credits:
Pre-training isn’t new

- This has inspired a series of work at the intersection of image and language, thanks to the availability of large high-quality curated datasets (e.g., COCO, Conceptual Captions).

<table>
<thead>
<tr>
<th></th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Split COCO Captions VG</td>
<td>Conceptual Captions SBU</td>
</tr>
<tr>
<td></td>
<td>Captions Dense Captions</td>
<td>Captions</td>
</tr>
<tr>
<td>train</td>
<td>533K (106K)</td>
<td>3.0M (3.0M)</td>
</tr>
<tr>
<td>val</td>
<td>25K (5K)</td>
<td>14K (14K)</td>
</tr>
<tr>
<td></td>
<td>5.06M (101K)</td>
<td>990K (990K)</td>
</tr>
<tr>
<td></td>
<td>106K (2.1K)</td>
<td>10K (10K)</td>
</tr>
</tbody>
</table>

Table 1: Statistics on the datasets used for pre-training. Each cell shows #image-text pairs (#images)

- But not so much in the video domain.

Table credit: Chen et al., UNITER: UNiversal Image-TExt Representation Learning, ECCV 2020.
Pre-training isn’t new

• The reasons why the video-and-language field has been lagging behind is mainly due to:

  • The challenge in harvesting large-scale data;

  • The challenge in annotating those data.
Evolution of video-language datasets

As a comparison, 500hr worth of videos are uploaded to YouTube per minute!
Video credit: COIN dataset
The Era of Pre-training

• “Free” annotations become accessible (i.e., subtitles or ASR transcripts)

Figure credit: Making Scallion Pancake Beef Rolls: https://www.youtube.com/watch?v=vTmgLKtx49Y
Video-and-Language Pre-training

• Paired video clips and subtitles

“Keep rolling tight and squeeze the air out to its side and you can kind of pull a little bit.”

• The resulted datasets are magnitudes bigger!

Figure credit: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html
Pre-training Data

• The major video-and-language dataset for pre-training:

\textit{HowTo100M Dataset}

[Miech et al., ICCV 2019]

• 1.22M instructional videos from YouTube
• Each video is 6 minutes long on average
• Over 100 million pairs of video clips and associated narrations
Pre-training Data

• Emerging public video-and-language datasets for pre-training:

**TV Dataset**
[Lei et al., EMNLP 2018]
- 22K video clips from 6 popular TV shows
- Each video clip is 60-90 seconds long
- Dialogue (“character: subtitle”) is provided

**Auto-captions on GIF Dataset**
[Pan et al., arXiv 2020]
- 163K GIFs automatically crawled from web
- Each GIF is a few seconds long
- Cover a variety of categories

Figure credits: from the original papers
Method Overview
Reconstructive Methods

• BERT-inspired; usually adopt the early fusion architecture.

• Usually leverage pre-trained unimodal feature/backbone (e.g., BERT, I3D)

• Image counterparts: ViLBERT/VLP/UNITER/OSCAR

Figure credit: Sun et al., VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019.
Background (BERT)

• BERT – Bidirectional Encoder Representations from Transformers

• Training Objectives
  • Masked Language Modeling (MLM)
  • Next Sentence Prediction (NSP)

---

Figure credits: https://www.kdnuggets.com/2018/12/bert-sota-nlp-model-explained.html
https://amitness.com/2020/02/albert-visual-summary/
VideoBRET

• Pre-training: 312K cooking videos from YouTube
• Video feature: Kinetics-pretrained S3D; then tokenize into 21K clusters using hierarchical K-means. Multi-Modal Encoder: BERT-large.
• Objectives: Masked Language Modeling (MLM), Masked Frame Modeling (MFM), Video-Text Matching (VTM)
VideoBRET

- Adding more data generally gives better results

ActBERT

• Pre-training: HowTo100M
• Video feature: object region feature from Faster RCNN; Kinetics-pretrained R(2+1)D.
• Multi-Modal Encoder: BERT-base.
• Training objectives
  • MLM, VTM
  • Masked Object (Noun) Classification
  • Masked Action (Verb) Classification

HERO (Hierarchical Encoder for Omni-representation learning)

- Objectives: MLM, MFM; New: Video-Subtitle Matching (VSM), Frame Order Modeling (FOM)

Li et al., HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. ECCV 2020.
DECEMBERT (Dense Captions and Entropy Minimization)

- Dense captions input (from a VG pre-trained dense captioning model)
- Attention Entropy Minimization (deal with the misalignment issue between video clip and subtitle through sharp attention).
Contrastive Methods

• Contrastive learning-inspired
• Usually adopt the late fusion architecture:

  ![Diagram](GIF credit: https://docs.google.com/presentation/d/1ccddJFD_j3p3h0TCqSV9ajSi2y1yOfh0-lJoK29ircs/edit#slide=id.g8c1b8d6efd_0_17)

  • Usually trained from scratch to learn a general feature representation
  • Image counterpart: CLIP
Background (Contrastive Learning)

• Given a data point $x$, contrastive methods aim to learn an encoder $f$ such that:

$$S(f(x), f(x^+)) \gg S(f(x), f(x^-)),$$

• where $x^+$ is a data point similar to $x$, referred to as a positive sample, $x^-$ is dissimilar to $x$, referred to as a negative sample.

• The score function $S$ could simply be vector inner product (or cosine similarity).

$$S(f(x), f(x^+)) = f(x)^T f(x^+)$$

• Most of the work until now is on how to define positive & negative samples.

Most of content from this section is borrowed from https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html
Background (Contrastive Learning)

• Based on the objective function, contrastive methods fall into three categories.

• Logistic Loss (e.g., the VTM/NSP objective)
  • Regress $S(f(x), f(x^+))$ to 1 and $S(f(x), f(x^-))$ to 0
Background (Contrastive Learning)

• Based on the objective function, contrastive methods fall into three categories.

  • Logistic Loss (e.g., the VTM/NSP objective)
    • Regress $S(f(x), f(x^+))$ to 1 and $S(f(x), f(x^-))$ to 0

  • Margin Loss (e.g., see later in COOT)
    • Minimize the total hinge loss:

      $\max(S(f(x), f(x^-)) - S(f(x), f(x^+)) + \Delta, 0)$
Background (Contrastive Learning)

- Noise-Contrastive Estimation (NCE) Loss
  - Use all other samples from the minibatch as negative samples
  - Cross entropy loss on an N-way Softmax classifier

\[
-\log \frac{\exp(S(f(x), f(x^+)))}{\exp(S(f(x), f(x^+))) + \sum_j \exp(S(f(x), f(x_j^-)))}
\]
CBT: Contrastive Bidirectional Transformer

**CBT: Contrastive Bidirectional Transformer**

- VL-NCE is simple, any paired clip and subtitle are considered a positive pair and the rest of the clips/subtitles in the minibatch are negatives.
- For Video NCE:

  ![Diagram](image)

  CBT is a shallow (2-layer) Transformer

- A similar objective is used in HERO (MFM with NCE).

---

MIL-NCE

• It uses VL-NCE, with a twist on multiple instance learning (MIL) to address the misalignment issue between video clip and subtitle.
COOT (Cooperative hierarchical Transformer)

- Margin loss on clip-level and video-level alignment
COOT (Cooperative hierarchical Transformer)

• Cross-modality cycle-consistency loss

Figure 3: Cross-Modality Cycle-Consistency. Starting from a sentence $s_i$, we find its nearest neighbor in the clip sequence and again its neighbor in the sentence sequence. Deviations from the start index are penalized as alignment error.
MERLOT (Multimodal Event Representation Learning Over Time)

• Objectives: i) MLM (mask visual tokens only), ii) VL-NCE (on frames), and iii) temporal reordering (similar to FOM in HERO).

• It combines reconstructive objective and contrastive objective.

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.
Generative Methods

• Video captioning inspired; usually adopt the encoder-decoder architecture

• Leverage video-to-text generation for video representation learning

• Image counterpart: VirTex
UniVL (Unified Video and Language)

- Objectives: VL-NCE, MLM, MFM, VTM; New: language reconstruction
SSB (Support-Set Bottlenecks)

- VL-NCE loss pushes away even semantically related captions.
- This paper introduces cross-captioning, which alleviates this by learning to reconstruct a sample’s text representation as a weighted combination of a support-set.

SSB (Support-Set Bottlenecks)

• A support-set contains every sample in the minibatch other than the positive sample.
CUPID (Adaptive **Curation of Pre-training Data**)

- Close the source-target domain gap
CUPID (Adaptive Curation of Pre-training Data)

- The paradigm is generic and has been applied to various models including MIL-NCE, HERO, CLIP, VLP.

<table>
<thead>
<tr>
<th>Downstream Task</th>
<th>Domain</th>
<th>Method</th>
<th>Feature Extractor</th>
<th>Our Curated Pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVR and TVQA</td>
<td>Out-of-domain</td>
<td>HERO</td>
<td>ImageNet/Kinetics pre-trained</td>
<td>200k from HowTo100M and TV</td>
</tr>
<tr>
<td>HMDB51</td>
<td>Near-domain</td>
<td>CUPID-CLIP</td>
<td>WIT [47] pre-trained backbone</td>
<td>15k from HowTo100M</td>
</tr>
<tr>
<td>YouCook2 captioning</td>
<td>In-domain</td>
<td>CUPID-VLP</td>
<td>HowTo100M pre-trained</td>
<td>15k from HowTo100M</td>
</tr>
<tr>
<td>YouCook2 retrieval</td>
<td>In-domain</td>
<td>MIL-NCE</td>
<td>HowTo100M pre-trained backbone</td>
<td>15k from HowTo100M</td>
</tr>
</tbody>
</table>

Table 4. A summary of downstream tasks, domain genres, methods, and pre-training settings.

Other Modalities (Video-Language-Audio)

Multi-Modal Versatile Network (MMV)

Video-Audio-Text Transformer (VATT)

Multimodal Clustering Networks (MCN)

Other Modalities

- Multimodal Transformer, MMT (ECCV 2020): mixture of seven types/experts of video features, including audio, appearance, motion, speech, scene, face, OCR for overlaid text, for video representation.

- Video-Audio: XDC/GDT/STiCA/AVID etc.
Image-Video Connector

• Can visual representation learned from video pre-training be useful for image tasks?
  • Yes. MMV (NeurIPS 2020) and VATT have results on ImageNet classification. MERLOT have results on VCR (a VQA dataset).

• Joint video-image encoder:

Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval

Max Bain\textsuperscript{1} Arsha Nagrani\textsuperscript{1}† Gül Varol\textsuperscript{1,2} Andrew Zisserman\textsuperscript{1}

\textsuperscript{1}Visual Geometry Group, University of Oxford
\textsuperscript{2}LIGM, École des Ponts, Univ Gustave Eiffel, CNRS

\{maxbain, arsha, gul, az\}@robots.ox.ac.uk
Image-Video Connector

• On the other hand, can image pre-training benefit video tasks?
  • Yes. See CLIP (OpenAI) and ClipBERT (CVPR 2021 Best Paper Nominee).

• ClipBERT

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective</th>
<th>Reconstructive</th>
<th>Contrastive</th>
<th>Generative</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>MFM</td>
<td>FOM</td>
<td>VTM</td>
</tr>
<tr>
<td>VideoBERT (ICCV 2019)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>ActBERT (CVPR 2020)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HERO (EMNLP 2020)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>DECEMBERT (NAACL’21)</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>CBT (arXiv 2019)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIL-NCE (CVPR 2020)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>COOT (NeurIPS 2020)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERLOT (arXiv 2021)</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>UniVL (arXiv 2020)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>SSB (ICLR 2021)</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>CUPID (arXiv 2021)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>ClipBERT (CVPR 2021)</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Downstream Tasks and Datasets

• Video-only tasks
  • Action Recognition: **HMDB51, UCF101**, Kinetics-600
  • Action Segmentation/Localization: COIN, CrossTask etc.
Downstream Tasks and Datasets

• Video-Language tasks
  • Video Captioning: **YouCook2, MSR-VTT, VATEX, TVC**
  • Text-to-Video Retrieval: **YouCook2, MSR-VTT, DiDeMo, ActivityNet Captions, TVR, VATEX, How2R, MSVD**
  • Video QA: **MSRVTT-QA, TGIF-QA, TVQA, How2QA**

Now, let’s place the tomatoes to the cutting board and slice the tomatoes.

**Query:** Toast the bread slices in the toaster

**Question:** What does the lady pour into pot?

**Answer:** Milk
Benchmark Results (Video-Only)

• Action Recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>Pre-training data</th>
<th>HMDB51</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised (Duan et al., ECCV 2020)</td>
<td>V</td>
<td>K400+OS</td>
<td>83.8</td>
<td>98.6</td>
</tr>
<tr>
<td>Supervised backbone (SSB, ICLR 2021)</td>
<td>V+T</td>
<td>HowTo+IG65+IM</td>
<td>81.3</td>
<td>98.0</td>
</tr>
<tr>
<td>Pure vision-based (Qian et al., CVPR 2021)</td>
<td>V</td>
<td>K600</td>
<td>70.6</td>
<td>94.4</td>
</tr>
<tr>
<td>CBT (arXiv 2019)</td>
<td>V+T</td>
<td>HowTo+ K600</td>
<td>44.5</td>
<td>79.5</td>
</tr>
<tr>
<td>MIL-NCE (CVPR 2020)</td>
<td>V+T</td>
<td>HowTo100M</td>
<td>61.0</td>
<td>91.3</td>
</tr>
<tr>
<td>MMV (NeurIPS 2020)</td>
<td>V+T+A</td>
<td>HowTo+AudioSet</td>
<td>75.0</td>
<td>95.2</td>
</tr>
</tbody>
</table>

• Multimodal pre-training has an edge over pure vision-based methods.

• Self-supervised methods are still trailing supervised counterparts.

SSB uses a backbone that is pretrained on IG65M model and another one pretrained on Imagenet. Others are from scratch.
Benchmark Results (Video-Language)

- YouCook2 captioning (video input only)

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training data</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masked Transformer (CVPR 2018)</td>
<td>None</td>
<td>3.85</td>
<td>10.68</td>
<td>37.9</td>
</tr>
<tr>
<td>VideoBERT (ICCV 2019)</td>
<td>312K videos</td>
<td>4.33</td>
<td>11.94</td>
<td>55.0</td>
</tr>
<tr>
<td>CBT (arXiv 2019)</td>
<td>HowTo+K600</td>
<td>5.12</td>
<td>12.97</td>
<td>64.0</td>
</tr>
<tr>
<td>ActBERT (CVPR 2020)</td>
<td>HowTo100M</td>
<td>5.41</td>
<td>13.30</td>
<td>65.0</td>
</tr>
<tr>
<td>CUPID (arXiv 2021)</td>
<td>HowTo100M</td>
<td>9.34</td>
<td>16.47</td>
<td>110.5</td>
</tr>
<tr>
<td>UniVL (arXiv 2020)</td>
<td>HowTo100M</td>
<td><strong>11.17</strong></td>
<td><strong>17.57</strong></td>
<td><strong>127.0</strong></td>
</tr>
</tbody>
</table>

Note: results are on micro-level metrics. For macro-level and paragraph-level metrics, see https://github.com/LuoweiZhou/YouCook2-Leaderboard#video-captioning
## Benchmark Results (Video-Language)

- YouCook2 text-to-video retrieval (video only, no audio)

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training data</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Median R</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGLMM (CVPR 2015)</td>
<td>None</td>
<td>4.6</td>
<td>14.3</td>
<td>21.6</td>
<td>75</td>
</tr>
<tr>
<td>Miech et al. (ICCV 2019)</td>
<td>None</td>
<td>4.2</td>
<td>13.7</td>
<td>21.5</td>
<td>65</td>
</tr>
<tr>
<td>COOT (NeurIPS 2020)</td>
<td>None</td>
<td>5.9</td>
<td>16.7</td>
<td>24.8</td>
<td>50</td>
</tr>
<tr>
<td>Miech et al. (zero-shot)</td>
<td>HowTo100M</td>
<td>6.1</td>
<td>17.3</td>
<td>24.8</td>
<td>46</td>
</tr>
<tr>
<td>ActBERT (zero-shot)</td>
<td>HowTo100M</td>
<td>9.6</td>
<td>26.7</td>
<td>38.0</td>
<td>19</td>
</tr>
<tr>
<td>MIL-NCE (zero-shot)</td>
<td>HowTo100M</td>
<td>13.9</td>
<td>36.3</td>
<td>48.9</td>
<td>11</td>
</tr>
<tr>
<td>MMV (zero-shot)</td>
<td>HowTo+AudioSet</td>
<td>11.7</td>
<td>33.4</td>
<td>45.4</td>
<td>13</td>
</tr>
<tr>
<td>MCN (zero-shot)</td>
<td>HowTo100M</td>
<td>18.1</td>
<td>35.5</td>
<td>45.2</td>
<td>-</td>
</tr>
<tr>
<td>Miech et al. (ICCV 2019)</td>
<td>HowTo100M</td>
<td>8.2</td>
<td>24.5</td>
<td>35.3</td>
<td>24</td>
</tr>
<tr>
<td>COOT (NeurIPS 2020)</td>
<td>HowTo100M</td>
<td>16.7</td>
<td>40.2</td>
<td>52.3</td>
<td>9</td>
</tr>
<tr>
<td>DECEMBERT (NAACL 2021)</td>
<td>HowTo100M</td>
<td>17.0</td>
<td>43.8</td>
<td>59.8</td>
<td>9</td>
</tr>
<tr>
<td>CUPID (arXiv 2021)</td>
<td>HowTo100M</td>
<td>17.7</td>
<td>43.2</td>
<td>57.1</td>
<td>7</td>
</tr>
<tr>
<td>UniVL (arXiv 2020)</td>
<td>HowTo100M</td>
<td>28.9</td>
<td>57.6</td>
<td>70.0</td>
<td>4</td>
</tr>
</tbody>
</table>

Pre-trained models generalize well
Pre-training wins again
Benchmark Results (Video-Language)

- MSR-VTT text-to-video retrieval (video only, no audio)

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training data</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Median R</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSB, w/o pre-training (ICLR 2021)</td>
<td>None</td>
<td>27.4</td>
<td>56.3</td>
<td>67.7</td>
<td>3</td>
</tr>
<tr>
<td>Miech et al. (ICCV 2019)</td>
<td>HowTo100M</td>
<td>14.9</td>
<td>40.2</td>
<td>52.8</td>
<td>9</td>
</tr>
<tr>
<td>ActBERT (CVPR 2020)</td>
<td>HowTo100M</td>
<td>16.3</td>
<td>42.8</td>
<td>56.9</td>
<td>10</td>
</tr>
<tr>
<td>HERO (EMNLP 2020)</td>
<td>HowTo100M+TV</td>
<td>16.8</td>
<td>43.4</td>
<td>57.7</td>
<td>-</td>
</tr>
<tr>
<td>UniVL (arXiv 2020)</td>
<td>HowTo100M</td>
<td>21.2</td>
<td>49.6</td>
<td>63.1</td>
<td>6</td>
</tr>
<tr>
<td>NoiseEstimation (AAAI 2021)</td>
<td>HowTo100M</td>
<td>17.4</td>
<td>41.6</td>
<td>53.6</td>
<td>8</td>
</tr>
<tr>
<td>SSB (ICLR 2021)</td>
<td>HowTo100M</td>
<td><strong>30.1</strong></td>
<td><strong>58.5</strong></td>
<td><strong>69.3</strong></td>
<td><strong>3</strong></td>
</tr>
<tr>
<td>ClipBERT (CVPR 2021)</td>
<td>COCO and VG</td>
<td>22.0</td>
<td>46.8</td>
<td>59.9</td>
<td>6</td>
</tr>
<tr>
<td>DECEMBERT (NAACL 2021)</td>
<td>HowTo100M</td>
<td>17.5</td>
<td>44.3</td>
<td>58.6</td>
<td>9</td>
</tr>
</tbody>
</table>

Limited gain possibly due to domain discrepancy
# Benchmark Results (Video-Language)

## Video QA

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training data</th>
<th>MSRVTT-QA</th>
<th>TVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAGE (ACL 2020)</td>
<td>None</td>
<td>-</td>
<td>70.23</td>
</tr>
<tr>
<td>HCRN (CVPR 2020)</td>
<td>None</td>
<td>27.4</td>
<td>-</td>
</tr>
<tr>
<td>HERO (EMNLP 2020)</td>
<td>HowTo100M+TV</td>
<td>-</td>
<td>73.61</td>
</tr>
<tr>
<td>NoiseEstimation (AAAI)</td>
<td>HowTo100M</td>
<td>35.1</td>
<td>-</td>
</tr>
<tr>
<td>DECEMBERT (NAACL 2021)</td>
<td>HowTo100M</td>
<td>37.4</td>
<td>-</td>
</tr>
<tr>
<td>ClipBERT (CVPR 2021)</td>
<td>COCO+VG</td>
<td>37.4</td>
<td>-</td>
</tr>
<tr>
<td>CoMVT (CVPR 2021)</td>
<td>HowTo100M</td>
<td>39.5</td>
<td>-</td>
</tr>
<tr>
<td>VQA-T (arXiv 2021)</td>
<td>HowToVQA69M</td>
<td>41.5</td>
<td>-</td>
</tr>
<tr>
<td>MERLOT (arXiv 2021)</td>
<td>YT-Temporal-180M</td>
<td>43.1</td>
<td>78.7</td>
</tr>
</tbody>
</table>

YT-Temporal-180M is larger than HowTo100M and contains diverse topics; this allows it to go beyond literal descriptions and capture more commonsense knowledge that could benefit QA.

Seo et al., Look Before you Speak: Visually Contextualized Utterances. CVPR 2021.
Video-And-Language Understanding Evaluation (VALUE)

VALUE competition will be held in conjunction with CLVL workshop at ICCV 2021!

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Video Source</th>
<th>More info</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retrieval Tasks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TVR</td>
<td>TV episodes</td>
<td></td>
<td>Average(R@1, 5, 10) with tIoU &gt;= 0.7</td>
</tr>
<tr>
<td>How2R</td>
<td>YouTube (HowTo100M)</td>
<td></td>
<td>Average(R@1, 5, 10) with tIoU &gt;= 0.7</td>
</tr>
<tr>
<td>YC2R</td>
<td>YouTube</td>
<td></td>
<td>Average(R@1, 5, 10)</td>
</tr>
<tr>
<td>VATEX EN-R</td>
<td>YouTube</td>
<td></td>
<td>Average(R@1, 5, 10)</td>
</tr>
<tr>
<td><strong>QA Tasks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TVQA</td>
<td>TV episodes</td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>How2QA</td>
<td>YouTube (HowTo100M)</td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>VIOLIN</td>
<td>TV episodes, Movie clips</td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>VLEP</td>
<td>TV episodes, YouTube</td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td><strong>Captioning Tasks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TVC</td>
<td>TV episodes</td>
<td></td>
<td>CIDEr-D</td>
</tr>
<tr>
<td>YC2C</td>
<td>YouTube</td>
<td></td>
<td>CIDEr-D</td>
</tr>
<tr>
<td>VATEX-EN-C</td>
<td>YouTube</td>
<td></td>
<td>CIDEr-D</td>
</tr>
</tbody>
</table>
Conclusion

• Video-and-Language Pre-training is a nascent field with great potential.

• Limitations
  • The use of different modalities (video, audio), pretraining datasets (HowTo100M, Kinetics-600), architectures (S3D, SlowFast), pre-training (supervised, self-supervised) makes it difficult to have fair comparisons.
  • More unified benchmarks need to be proposed. VALUE is a good start.

• Future Directions
  • Further scale up the data and its domain diversity
  • Multimodal and multilingual
Thank you!
Any questions?