Video-Text Pre-training

Kevin Lin, Linjie Li, Chung-Ching Lin
6/19/2022
Pretrain-and-finetune paradigm

- Pre-train on **large** amounts of datasets is very helpful for performance improvement on target tasks with **small** datasets.
Evolution of Video-Text Pre-training

Representative Video-Text Models until CVPR 2021

Many more methods have been proposed since then ...
Comprehensive Benchmarks

**VALUE** (Li et al. '21)

**Comprehensive Image-Text Models**

**ClipsBERT** (Lei et al. '21)
**Clip4Clip** (Luo et al. '21)
**ClipsCaption** (Tang et al. '21)
**Flamingo** (Alayrac et al. '22)

**Transfer Image-Text Models**

**GIT** (Wang et al. '22)
**Clip4Clip** (Wang et al. '21)
**Clip4Caption** (Akbari et al. '21)

**Modeling Multi-channel Videos**

**HERO** (Li et al. '20)
**HiT** (Buch et al. '22)
**MMT** (Gabeur et al. '20)
**MV-GPT** (Seo et al. '22)

**Probing Analysis**

**Contrast Sets** (Park et al. '22)
**ATP** (Luo et al. '21)
**HD-VILA** (Xue et al. '22)

**Enhanced Pre-training Data**

**Frozen** (Bain et al. '21)
**MERLOT** (Zeller et al. '21)
**HD-VILA** (Xue et al. '22)

**Advanced Pre-training Tasks**

**Support-Set** (Patrick et al. '20)
**VIOLET** (Fu et al. '21)
**OA-Trans** (Wang et al. '22)
**ALPRO** (Li et al. '22)
**BridgeFormer** (Ge et al. '21)

**Applications to Video Understanding**

**VideoCLIP** (Xu et al. '20)
**TACo** (Yang et al. '21)
**ActionCLIP** (Wang et al. '21)
**EfficientPrompt** (Ju et al. '21)
**Bridge-Prompt** (Li et al. '22)
**P3IV** (Zhao et al. '22)

**More Languages**

**Uni-Percyever** (Zhu et al. '22)
**MMP** (Huang et al. '21)
**VICTOR** (Lei et al. '21)
**SkillNet** (Dai et al. '22)

**Unified Modeling**

**LAVENDER** (Li et al. '22)

**Pioneering work in Video-Text Pre-training**

**VideoBERT** (Sun et al. '19)
**UniVL** (Luo et al. '20)
**HTM** (Miech et al. '19)
**MIL-NCE** (Miech et al. '20)
Agenda

Overview of Video-Text Pre-training

Learning from Multi-channel Videos: Methods and Benchmarks

Advanced Topics in Video-Text Pretraining
Overview of Video-Text Pre-training

Kevin Lin
**Pioneering work in Video-Text Pre-training**

- **VideoBERT** (Sun et al. '19)  
- **UniVL** (Luo et al. '20)  
- **HTM** (Miech et al. '19)  
- **MIL-NCE** (Miech et al. '20)

**Enhanced Pre-training**

- **Frozen** (Bain et al. '21)  
- **GIT** (Wang et al. '22)  
- **MERLOT** (Zeller et al. '21)  
- **HD-VILA** (Xue et al. '22)

**Advanced Pre-training Tasks**

- **Support-Set** (Patrick et al. '20)  
- **OA-Trans** (Wang et al. '22)  
- **ALPRO** (Li et al. '22)  
- **BridgeFormer** (Ge et al. '22)

**Comprehensive Benchmarks**

- **VALUE** (Li et al. '21)  
- **Merlot** (Zeller et al. '21)  
- **Merlot Reserve** (Zeller et al. '22)  
- **Noise Estimation** (Amrani et al. '20)

**Transfer Image-Text Models**

- **ClipBERT** (Lei et al. '21)  
- **Clip4Clip** (Luo et al. '21)  
- **Clip4Caption** (Tang et al. '21)  
- **Flamingo** (Alayrac et al. '22)

**Probing Analysis**

- **Contrast Sets** (Park et al. '22)  
- **Probing Analysis** (Buch et al. '22)  
- **Frozen** (Bain et al. '21)  
- **GIT** (Wang et al. '22)

**Modeling Multi-channel Videos**

- **VAT** (Akbari et al. '21)  
- **MV-GPT** (Seo et al. '22)  
- **AVLNet** (Rouditchenko et al. '21)  
- **HERO** (Li et al. '20)  
- **HiT** (Liu et al. '21)  
- **MMT** (Gabeur et al. '20)  
- **VideoCLIP** (Xu et al. '20)

**Applications to Video Understanding**

- **EfficientPrompt** (Ju et al. '21)  
- **MMV** (Alayrac et al. '20)  
- **TACo** (Yang et al. '21)  
- **ActionCLIP** (Wang et al. '21)  
- **Bridge-Prompt** (Han et al. '22)  
- **ATP** (Buch et al. '22)  
- **TAN** (Han et al. '22)  
- **CLAP** (Xu et al. '22)  
- **TACo** (Yang et al. '21)

**More Languages**

- **TAN** (Han et al. '22)  
- **MMP** (Huang et al. '21)  
- **VICTOR** (Lei et al. '21)  
- **Tencent-MSVE** (Zeng et al. '22)  
- **BridgePrompt** (Li et al. '22)  
- **BridgeFormer** (Li et al. '22)  
- **SkillNet** (Dai et al. '22)

**Unified Modeling**

- **UniPerceiver** (Zhu et al. '22)
- **LAVENDER** (Li et al. '22)  
- **HiT** (Liu et al. '21)  
- **Clap** (Xu et al. '22)  
- **UniVL** (Luo et al. '20)  
- **HTM** (Miech et al. '19)  
- **MIL-NCE** (Miech et al. '20)

**VideoCLIP** (Xu et al. '20)
Outline

• Data and challenges
• Pioneer work in video-text pre-training
• Advanced pre-training tasks
• Transferring image-text model
Video-and-Language 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
Slide credit: CVPR 2021 VQA2VLN Tutorial
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.”

Figure credit: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html
Slide credit: CVPR 2021 VQA2VLN Tutorial
Evolution of Video-and-Language Datasets

Credit: CVPR 2021 VQA2VLN Tutorial
Evolution of Video-and-Language Datasets

Merlot: Multimodal neural script knowledge models, NeurIPS 2021
HowTo100M

• 136M video clips from YouTube videos
• Each clip is paired with an automatically transcribed narration
• 23K activities

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, ICCV 2019
Challenges in training data

- Noisy transcript (automatically generated with ASR tools)
- Constrained domains (instruction videos)
- Temporally misaligned
- Computing resources demanding
now I'm just kind of grilling these tomatoes in this pan I want to get the maximum flavor usually you always use tomatoes raw as it but I just want to add that little dimension of cooked a slightly charged tomatoes yum!

Grill the tomatoes in a pan and then put them on a plate.

Language styles are different!

Examples from Youcook dataset: Towards Automatic Learning of Procedures From Web Instructional Videos, AAAI 2018
Recent Data Sources

• Researchers have been working on collecting better quality (well-aligned) data for the pre-training.

Example from shutterstock. Link: Smiling Beautiful Family Four Play Catch Stock Footage Video (100% Royalty-free) 1058262028 | Shutterstock
WebVid-2M & WebVid-10M

- Well-aligned video-text pairs from high-quality video sources

<table>
<thead>
<tr>
<th>dataset</th>
<th>domain</th>
<th>#clips</th>
<th>avg dur. (secs)</th>
<th>#sent</th>
<th>time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPII Cook [47]</td>
<td>cooking</td>
<td>44</td>
<td>600</td>
<td>6K</td>
<td>8</td>
</tr>
<tr>
<td>TACos [44]</td>
<td>cooking</td>
<td>7K</td>
<td>360</td>
<td>18K</td>
<td>15.9</td>
</tr>
<tr>
<td>DideMo [3]</td>
<td>flickr</td>
<td>27K</td>
<td>28</td>
<td>41K</td>
<td>87</td>
</tr>
<tr>
<td>MSR-VTT [65]</td>
<td>youtube</td>
<td>10K</td>
<td>15</td>
<td>200K</td>
<td>40</td>
</tr>
<tr>
<td>Charades [53]</td>
<td>home</td>
<td>10K</td>
<td>30</td>
<td>16K</td>
<td>82</td>
</tr>
<tr>
<td>LSMDC15 [46]</td>
<td>movies</td>
<td>118K</td>
<td>4.8</td>
<td>118K</td>
<td>158</td>
</tr>
<tr>
<td>YouCook II [70]</td>
<td>cooking</td>
<td>14K</td>
<td>316</td>
<td>14K</td>
<td>176</td>
</tr>
<tr>
<td>ActivityNet [24]</td>
<td>youtube</td>
<td>100K</td>
<td>180</td>
<td>100K</td>
<td>849</td>
</tr>
<tr>
<td>CMD [5]</td>
<td>movies</td>
<td>34K</td>
<td>132</td>
<td>34K</td>
<td>1.3K</td>
</tr>
<tr>
<td>WebVid-2M</td>
<td>open</td>
<td>2.5M</td>
<td>18</td>
<td>2.5M</td>
<td>13K</td>
</tr>
<tr>
<td>HT100M [37]</td>
<td>instruction</td>
<td>136M</td>
<td>4</td>
<td>136M</td>
<td>134.5K</td>
</tr>
</tbody>
</table>

Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, ICCV 2021
Scale It Up

• Collect larger scale, more diverse videos from YouTube

YT-Temporal 180M (NeurIPS21):
30% of videos are about local news & monetized contents

YT-Temporal 1B (CVPR22):
Scale it up in terms of video domains and # videos

HD-VILA-100M (CVPR22):
High-resolution videos (720p)
The more the better?

- Researchers are exploring to use a small subset of data for domain-specific pre-training
The more the better?

• VideoCC3M: Mining audio-video clips

<table>
<thead>
<tr>
<th>Pretraining Data</th>
<th>Modality</th>
<th># Caps</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finetuned</td>
<td>V</td>
<td>-</td>
<td>30.2</td>
<td>60.7</td>
<td>71.1</td>
</tr>
<tr>
<td>HowTo100M [55]</td>
<td>V</td>
<td>130M</td>
<td>33.1</td>
<td>62.3</td>
<td>72.3</td>
</tr>
<tr>
<td>VideoCC3M</td>
<td>V</td>
<td>970K</td>
<td>35.0</td>
<td>63.1</td>
<td>75.1</td>
</tr>
<tr>
<td>VideoCC3M</td>
<td>A+V</td>
<td>970K</td>
<td>35.8</td>
<td>65.1</td>
<td>76.9</td>
</tr>
<tr>
<td>Zero-shot</td>
<td>V</td>
<td>130M</td>
<td>8.6</td>
<td>16.9</td>
<td>25.8</td>
</tr>
<tr>
<td>HowTo100M [55]</td>
<td>V</td>
<td>970K</td>
<td>18.9</td>
<td>37.5</td>
<td>47.1</td>
</tr>
<tr>
<td>VideoCC3M</td>
<td>A+V</td>
<td>970K</td>
<td>19.4</td>
<td>39.5</td>
<td>50.3</td>
</tr>
</tbody>
</table>

Table 2. Effect of pretraining data on text-video retrieval for the MSR-VTT dataset. # Caps: Number of unique captions. Training on VideoCC3M provides much better performance than Howto100M, with a fraction of the dataset size (VideoCC3M has only 970K captions and 6.3M clips compared to the 130M clips in HowTo100M). The performance boost is particularly large for the zero-shot setting.
Outline

• Data and challenges
• Pioneer work in video-text pre-training
• Advanced pre-training tasks
• Transferring image-text model
Most existing approaches can be roughly classified into two categories:

**Dual Encoder**
- Textual Encoder
- Visual Encoder
- Dot product or Cosine similarity

**Fusion Encoder**
- Multimodal fusion
- Textual Encoder
- Visual Encoder
Dual Encoder

• Large-scale contrastive video-text learning

• Favorable architecture for image-text retrieval

[HTM, Miech et al., 2019], [MIL-NCE, Miech et al., 2020], [Support Set, Patrick et al., 2020], [Frozen, Bain et al., 2021], [VideoCLIP, Xu et al., 2021]

Figure credit: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019
Fusion Encoder

- Deep fusion: better model the interactions between video and text

- Strong improvements on Video QA and Video Captioning

[VideoBERT, Sun et al., 2019], [UniVL, Luo et al., 2020], [ClipBERT, Lei et al., 2021], [MERLOT, Zellers et al., 2021]

Figure credit: Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling, CVPR 2021
Overview of the representative VLP models for video-and-language

<table>
<thead>
<tr>
<th>Model</th>
<th>Multimodal Fusion</th>
<th>Vision Encoder</th>
<th>Text Encoder</th>
<th>Decoder</th>
<th>E2E</th>
<th>Pre-training Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>VideoBERT (Sun et al., 2019b)</td>
<td>3D CNN, Emb.</td>
<td>3D CNN</td>
<td>Xformer</td>
<td></td>
<td></td>
<td>MLM+VTM+MVM</td>
</tr>
<tr>
<td>ActBERT (Zhu and Yang, 2020)</td>
<td>OD</td>
<td>Emb.</td>
<td>X</td>
<td></td>
<td></td>
<td>MLM+VTM+MVM</td>
</tr>
<tr>
<td>CBT (Sun et al., 2019a)</td>
<td>3D CNN, Xformer</td>
<td>3D CNN</td>
<td>X</td>
<td></td>
<td></td>
<td>VTC</td>
</tr>
<tr>
<td>HERO (Li et al., 2020b)</td>
<td>2D+3D CNN, Emb.</td>
<td>Emb.</td>
<td>X</td>
<td></td>
<td></td>
<td>MLM+VTM+FOM</td>
</tr>
<tr>
<td>UniVL (Luo et al., 2020)</td>
<td>2D+3D CNN, Xformer</td>
<td>Emb.</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC+MLM+VTM</td>
</tr>
<tr>
<td>ClipBERT (Lei et al., 2021b)</td>
<td>2D CNN, Emb.</td>
<td>Emb.</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM+VTM+CG</td>
</tr>
<tr>
<td>VLM (Xu et al., 2021a)</td>
<td>2D CNN, Emb.</td>
<td>Emb.</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM-MF+MMM</td>
</tr>
<tr>
<td>DeCEMBERT (Tang et al., 2021b)</td>
<td>2D CNN, Emb.</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM+VTM+CA</td>
</tr>
<tr>
<td>TACo (Yang et al., 2021b)</td>
<td>2D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTM+VTC</td>
</tr>
<tr>
<td>VQA-T (Yang et al., 2021a)</td>
<td>3D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTM+VTC</td>
</tr>
<tr>
<td>VICTOR (Lei et al., 2021a)</td>
<td>2D CNN, Emb.</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM+VTC+MFM</td>
</tr>
<tr>
<td>MERLOT (Zellers et al., 2021)</td>
<td>2D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM+VTC+FOM</td>
</tr>
<tr>
<td>MV-GPT (Seo et al., 2022)</td>
<td>3D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>MLM+CG</td>
</tr>
<tr>
<td>HTM (Miech et al., 2019)</td>
<td>3D CNN, Word2Vec</td>
<td>Word2Vec</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC</td>
</tr>
<tr>
<td>MIL-NCE (Miech et al., 2020)</td>
<td>3D CNN, Word2Vec</td>
<td>Word2Vec</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC</td>
</tr>
<tr>
<td>Support Set (Patrick et al., 2020)</td>
<td>2D+3D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC+CG</td>
</tr>
<tr>
<td>Frozen (Bain et al., 2021)</td>
<td>3D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC</td>
</tr>
<tr>
<td>VideoCLIP (Xu et al., 2021b)</td>
<td>3D CNN, Xformer</td>
<td>Xformer</td>
<td>✓</td>
<td></td>
<td></td>
<td>VTC</td>
</tr>
</tbody>
</table>
Video-Text Contrastive Learning (VTC)

- Borrow the idea from contrastive learning
- VTC aims to learn the correspondence between video and text

Many follow-up works propose to collect better positive and negative pairs

Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019
MIL-NCE

- Multiple Instance Learning (MIL) and Noise Contrastive Estimation (NCE)
- Try to mitigate the misalignment between video and transcript
- Consider a set of multiple positive candidate pairs

Figure 2: **Left.** Our MIL-NCE makes it possible to consider a set of multiple positive candidate pairs \(\{(x, y), (x, y^1), \ldots, (x, y^t)\}\) while the standard NCE approach would only consider the single \((x, y)\) training pair and miss the visually grounded object description *sander from pair* \((x, y^3)\) or the action description *sanding down from* \((x, y^6)\). **Right.** Given a video \(x\) and an associated set of positive narration candidates \(P\) (green triangles) that may or may not be correct, our MIL-NCE selects multiple correct positives (large blue areas) while downweighting incorrect positives (smaller blue areas) based on a discriminative ratio against negatives \(N\) (red squares). In contrast, traditional MIL considers only one positive (orange circle) while discarding the rest.
Masked Language Modeling (MLM)

- MLM is a direct adoption from NLP field
- Facilitate the multimodal fusion between video and text

Figure credit: VideoBERT: A Joint Model for Video and Language Representation Learning, ICCV 2019
Video-Text Matching (VTM)

• Given a batch of positive and negative video-text pairs, VTM aims to identify which videos and texts correspond to each other.

• Often formulate as a binary classification task

Figure credit: VideoBERT: A Joint Model for Video and Language Representation Learning, ICCV 2019
Masked Video Modeling (MVM)

• Similar to MLM, MVM is also developed to reconstruct the masked input visual tokens
• Visual features are high-dimensional and continuous
• Little-to-none effects in the pre-training

Hero: Hierarchical encoder for video+language omni-representation pre-training, EMNLP 2020
ViLT: Vision-and-language transformer without convolution or region supervision, ICML 2021
Outline

• Data and challenges
• Pioneer work in video-text pre-training
• Advanced pre-training tasks
• Transferring image-text model
Masked Visual-token Modeling

- Reconstruct the **discrete latent codes** from pre-trained DALL-E
- Promising improvements for video-and-language pre-training
Frame Order Modeling (FOM)

• During training, a percentage of the frames is randomly selected to be shuffled, and the goal is to reconstruct their original temporal order.

• Formulate FOM as a classification task and predict the timestamp.

Hero: Hierarchical encoder for video+language omni-representation pre-training, EMNLP 2020
Merlot: Multimodal neural script knowledge models, NeurIPS 2021
Object-level Supervision

- Object-level supervision can enhance cross-modality alignment

It is helpful to learn fine-grained region-entity alignment

Align and Prompt: Video-and-Language Pre-training with Entity Prompts, CVPR 2022
Object-aware Video-language Pre-training for Retrieval, CVPR 2022
Actbert: Learning global-local video-text representations, CVPR 2020
Outline

• Data and challenges
• Pioneer work in video-text pre-training
• Advanced pre-training tasks
• Transferring image-text model
Transferring Image-Text Model

• In core video problems, leveraging ImageNet pre-trained weights as an initialization is usually helpful

Can we leverage well pre-trained image-text model for video-text tasks?

BEVT: BERT Pretraining of Video Transformers, CVPR 2022
Video Swin Transformer, ICCV 2021
VidTr: Video Transformer Without Convolutions, ICCV 2021
Mask2Former for Video Instance Segmentation, ArXiv 2021
ClipBERT

• Pre-train with MLM + ITM on image-text pairs (COCO + VG Captions)
• Avoid the excessive cost of video-text pre-training

<table>
<thead>
<tr>
<th>Weight Initialization</th>
<th>MSRVTT Retrieval</th>
<th>MSRVTT-QA Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
</tr>
<tr>
<td>CNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>random</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>random</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>TSN, K700</td>
<td>5.7</td>
<td>22.1</td>
</tr>
<tr>
<td>ImageNet</td>
<td>7.2</td>
<td>23.3</td>
</tr>
<tr>
<td>grid-feat</td>
<td>7.4</td>
<td>21.0</td>
</tr>
<tr>
<td>image-text pre-training</td>
<td><strong>10.2</strong></td>
<td><strong>28.6</strong></td>
</tr>
</tbody>
</table>

Table 5: Impact of weight initialization strategy.

Image-text pre-training helps video-text tasks!
CLIP for X

- CLIP4Clip is post-pretrained with contrastive loss on HT100M

*Large-scale image-text pre-training also helps video-text tasks*

CLIP4Caption: CLIP for Video Caption, arXiv 2021
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, arXiv 2021
TubeDETR

- DETR style architecture for spatio-temporal video grounding
- Image-text pre-training (COCO, VG, F30K)

![Diagram of TubeDETR](image)

Figure 1. Spatio-temporal video grounding requires reasoning about space, time, and language.

*Image-text pre-training can also help advanced video-text downstream tasks*

<table>
<thead>
<tr>
<th>Pre-Training</th>
<th>Decoder Self-Attention Transfer</th>
<th>m.IoU</th>
<th>m.vIoU</th>
<th>vIoU @0.3</th>
<th>vIoU @0.5</th>
<th>m.sIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>X</td>
<td>42.8</td>
<td>23.5</td>
<td>33.2</td>
<td>20.9</td>
<td>38.5</td>
</tr>
<tr>
<td>2.</td>
<td>☑</td>
<td>43.8</td>
<td>28.6</td>
<td>39.8</td>
<td>27.3</td>
<td>46.6</td>
</tr>
<tr>
<td>3.</td>
<td>☑ Temporal</td>
<td>45.9</td>
<td><strong>30.3</strong></td>
<td><strong>42.3</strong></td>
<td><strong>29.8</strong></td>
<td><strong>47.7</strong></td>
</tr>
</tbody>
</table>

Table 2. Effect of the weight initialization for our model on the VidSTG validation set.
Flamingo with Perceiver Resampler

A flexible architecture that can take both images and videos as inputs.
Applying GIT to Video Domain

Figure 2: Network architecture of our GIT, composed of one image encoder and one text decoder. (a): The training task in both pre-training and captioning is the language modeling task to predict the associated description. (b): In VQA, the question is placed as the text prefix. (c): For video, multiple frames are sampled and encoded independently. The features are added with an extra learnable temporal embedding (initialized as 0) before concatenation.

Adaptation with sparsely-sampled frames can generate new SOTA on popular benchmark

<table>
<thead>
<tr>
<th>Video captioning</th>
<th>Video QA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSVD</td>
</tr>
<tr>
<td>Prior SOTA</td>
<td>120.6</td>
</tr>
<tr>
<td>GIT (ours)</td>
<td>180.2</td>
</tr>
<tr>
<td>Δ</td>
<td>+59.6</td>
</tr>
</tbody>
</table>
Comprehensive Benchmarks

- VALUE (Li et al. '21)
- GIT (Wang et al. '22)
- MERLOT (Zeller et al. '21)
- MERLOT RESERVE (Zeller et al. '22)
- MERLOT (Zeller et al. '21)
- HD-VILA (Xue et al. '22)
- Frozen (Bain et al. '21)

Transfer Image-Text Models

- ClipBERT (Lei et al. '21)
- Clip4Clip (Luo et al. '21)
- Clip4Caption (Tang et al. '21)
- Flamingo (Alayrac et al. '22)
- GIT (Wang et al. '22)

Enhanced Pre-training

- MERLOT (Zeller et al. '21)
- HD-VILA (Xue et al. '22)

Advanced Pre-training Tasks

- Support-Set (Patrick et al. '20)
- VIOLET (Fu et al. '21)
- OA-Trans (Wang et al. '22)
- ALPRO (Li et al. '22)
- BridgeFormer (Ge et al. '22)

Probing Analysis

- Contrast Sets (Park et al. '22)
- ATP (Buch et al. '22)

Applications to Video Understanding

- VideoCLIP (Xu et al. '20)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)

More Languages

- Tencent-MSVE (Zeng et al. '21)

Unified Modeling

- UniPerceiver (Zhu et al. '22)
- SkillNet (Dai et al. '22)
- LAVENDER (Li et al. '22)

VideoCLIP (Xu et al. '20)

VideoBERT (Sun et al. '19)

UniVL (Luo et al. '20)

HTM (Miech et al. '19)

MIL-NCE (Miech et al. '20)

HERO (Li et al. '20)

MMT (Li et al. '21)

HiT (Liu et al. '21)

VATT (Akbari et al. '21)

MV-GPT (Seo et al. '22)

AVLNet (Rouditchenko et al. '21)

ATP (Buch et al. '22)

TAN (Han et al. '22)

CLAP (Xu et al. '22)

Bridge-Prompt (Li et al. '22)

P3IV (Zhao et al. '22)

More Languages

- Tencent-MSVE (Zeng et al. '21)

Unified Modeling

- UniPerceiver (Zhu et al. '22)
- SkillNet (Dai et al. '22)
- LAVENDER (Li et al. '22)

Advanced Pre-training Tasks

- Support-Set (Patrick et al. '20)
- VIOLET (Fu et al. '21)
- OA-Trans (Wang et al. '22)
- ALPRO (Li et al. '22)
- BridgeFormer (Ge et al. '22)

Probing Analysis

- Contrast Sets (Park et al. '22)
- ATP (Buch et al. '22)

Applications to Video Understanding

- VideoCLIP (Xu et al. '20)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)

More Languages

- Tencent-MSVE (Zeng et al. '21)

Unified Modeling

- UniPerceiver (Zhu et al. '22)
- SkillNet (Dai et al. '22)
- LAVENDER (Li et al. '22)

Advanced Pre-training Tasks

- Support-Set (Patrick et al. '20)
- VIOLET (Fu et al. '21)
- OA-Trans (Wang et al. '22)
- ALPRO (Li et al. '22)
- BridgeFormer (Ge et al. '22)

Probing Analysis

- Contrast Sets (Park et al. '22)
- ATP (Buch et al. '22)

Applications to Video Understanding

- VideoCLIP (Xu et al. '20)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)

More Languages

- Tencent-MSVE (Zeng et al. '21)

Unified Modeling

- UniPerceiver (Zhu et al. '22)
- SkillNet (Dai et al. '22)
- LAVENDER (Li et al. '22)

Advanced Pre-training Tasks

- Support-Set (Patrick et al. '20)
- VIOLET (Fu et al. '21)
- OA-Trans (Wang et al. '22)
- ALPRO (Li et al. '22)
- BridgeFormer (Ge et al. '22)

Probing Analysis

- Contrast Sets (Park et al. '22)
- ATP (Buch et al. '22)

Applications to Video Understanding

- VideoCLIP (Xu et al. '20)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)

More Languages

- Tencent-MSVE (Zeng et al. '21)

Unified Modeling

- UniPerceiver (Zhu et al. '22)
- SkillNet (Dai et al. '22)
- LAVENDER (Li et al. '22)
Looking forward

• How to effectively transfer image-text model to video-text tasks?
  • Many recent methods only use a naïve frame concatenation

• Temporal modeling has not been well-explored
  • Most existing studies focus mainly on spatial modeling