Advanced Topics in Video-Text Pre-training

Chung-Ching Lin
Comprehensive Benchmarks

- ClipBERT (Lei et al. '21)
- Clip4Clip (Luo et al. '21), Clip4Caption (Tang et al. '21)
- Transfer Image-Text Models
  - Flamingo (Alayrac et al. '22)
  - GIT (Wang et al. '22)

Modeling Multi-channel Videos

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

Probing Analysis

- ATP (Buch et al. '22)
- Contrast Sets (Park 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)
- MMV (Alayrac et al. '20)
- VideoCLIP (Yang et al. '21)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)

Applications to Video Understanding

- Bridge-Prompt (Buch et al. '22)
- P3IV (Zhao et al. '22)

Advanced Pre-training Tasks

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

More Languages

- Tencent-MVSE (Zeng et al. '21)
- SkillNet (Dai et al. '22)
- LAVENDER (Zhu et al. '22)
- Uni-Perceiver (Zeng et al. '22)

Enhanced Pre-training Data

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

Advanced Pre-training

- Frozen (Bain et al. '21)

Support-Set

- Noise Estimation (Amrani et al. '20)
- Support-Set Tasks (Patrick et al. '20)
- VIOLET (Fu et al. '21)
- OA-Trans (Wang et al. '22)
- ALPRO (Li et al. '22)
- BridgeFormer (Li et al. '22)
Agenda

• Unified modeling
  • Modality-agnostic
  • Task-agnostic
• Applications to video understanding
  • Video-text contrastive models
  • Transfer image-text models
  • Prompt tuning
• More languages
Unification

• Modalities: unify backbones across modalities

• Tasks: unify VL tasks and V-only task as text generation
Unification

- **Modalities**: unify backbones (text, image, video, audio)
  - Uni-Perceiver, All in One, SkillNet, VATT

- **Tasks**: unify VL tasks V-only task as text generation
  - LAVENDER, Socratic Models, GIT, Flamingo
Modality-specific to modality-agnostic

• A modality-agnostic Transformer encoder and lightweight modality-specific tokenizer
• Encodes different task inputs and targets from arbitrary modalities into a unified representation

Modality-agnostic model

- Encodes different task inputs and targets from arbitrary modalities into a unified representation

\[ P(x, y) \propto \exp\{\cos(f(x), f(y))/\tau\} \]
Pre-training on Multi-Modal Tasks

Pre-training dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Images</th>
<th>#Videos</th>
<th>#Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet-21k [17]</td>
<td>14.2M</td>
<td>0</td>
<td>21K</td>
</tr>
<tr>
<td>Kinetics-700 [32]</td>
<td>0</td>
<td>542K</td>
<td>700</td>
</tr>
<tr>
<td>Moments in Time [57]</td>
<td>0</td>
<td>792K</td>
<td>339</td>
</tr>
<tr>
<td>Books&amp;Wiki [93]</td>
<td>0</td>
<td>0</td>
<td>101M</td>
</tr>
<tr>
<td>PAQ [41]</td>
<td>0</td>
<td>0</td>
<td>65M</td>
</tr>
<tr>
<td>CC3M [68]</td>
<td>3.0M</td>
<td>0</td>
<td>3.0M</td>
</tr>
<tr>
<td>CC12M [9]</td>
<td>11.1M</td>
<td>0</td>
<td>11.1M</td>
</tr>
<tr>
<td>COCO Caption [12]</td>
<td>113K</td>
<td>0</td>
<td>567K</td>
</tr>
<tr>
<td>Visual Genome [36]</td>
<td>108K</td>
<td>0</td>
<td>5.41M</td>
</tr>
<tr>
<td>SBU [58]</td>
<td>830K</td>
<td>0</td>
<td>830K</td>
</tr>
<tr>
<td>YFCC [31]</td>
<td>14.8M</td>
<td>0</td>
<td>14.8M</td>
</tr>
</tbody>
</table>

Modality-Agnostic Video-Language Pre-training models

- An unified backbone architecture

(a). Conventional Pre-training Framework

(b). Ours All-in-one Transformer

(c). Flops & Performance Comparison

All in One: Exploring Unified Video-Language Pre-training, arXiv 2022
Modality-Agnostic Video-Language Pre-training models

- Token rolling operation to encode temporal representations from video clips in a non-parametric manner
Sparse multimodal Mixture-of-Expert (MoE)

• Handling multiple modalities of information with a single model
• Different parts of the parameters are specialized for processing different modalities

(a) Fully activated dense model
(b) Sparsely activated MoE

(e) SkillNet for text-to-video retrieval
(f) SkillNet for text-to-code retrieval

Sparse multimodal Mixture-of-Expert (MoE)

A puppy is playing a frisbee with a women.
Sparse multimodal Mixture-of-Expert (MoE)

- The embeddings of different modalities
  - Text
  - Sound
  - Image
  - Video
  - Code

Unification

• Modalities: architecture (across audio, code)
  • Uni-Perceiver, All in One, SkillNet

• Tasks: unify VL tasks/V-only tasks as text generation
  • LAVENDER, Socratic Models, GIT, Flamingo
Unify VidL tasks as text generation

- Unify both pre-training and downstream finetuning as Masked Language Modeling
Unify VidL tasks as text generation

• Adopt an encoder-only architecture, with a lightweight MLM head, instead of the heavy decoder in unified image-text models

Comparison to existing methods during pre-training
Unify VidL tasks as text generation

- Adopt an encoder-only architecture, with a lightweight MLM head, instead of the heavy decoder in unified image-text models

Comparison to existing methods on downstream image/video question answering task

(a) Existing VidL Methods

(b) Existing Unified Image-text Models

(c) LAVENDER

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022
Unify VidL tasks as text generation

- LAVENDER surpasses previous methods with task-specific designs on 12 video-language tasks

<table>
<thead>
<tr>
<th></th>
<th>TGIF</th>
<th>MSRVT</th>
<th>MSVD</th>
<th>Captioning</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Action</td>
<td>Transition</td>
<td>Frame</td>
<td>MSRVTT</td>
<td>LSMDC</td>
</tr>
<tr>
<td>Published</td>
<td>94.0</td>
<td>96.2</td>
<td>69.5</td>
<td>90.9</td>
<td>81.7</td>
</tr>
<tr>
<td>SOTA</td>
<td>[84]</td>
<td>[84]</td>
<td>[84]</td>
<td>[84]</td>
<td>[84]</td>
</tr>
<tr>
<td>LAVENDER</td>
<td>96.6</td>
<td>99.1</td>
<td>73.5</td>
<td>97.4</td>
<td>87.0</td>
</tr>
<tr>
<td>\Delta</td>
<td>2.6↑</td>
<td>2.9↑</td>
<td>4.0↑</td>
<td>6.5↑</td>
<td>5.3↑</td>
</tr>
</tbody>
</table>

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling, arXiv 2022
Composing Zero-Shot Multimodal Reasoning with Language

• Different pretrained models store different forms of commonsense knowledge across different domains
  • Visual-language models (VLMs) are trained on Internet-scale image captions
  • Large language models (LMs) are further trained on Internet-scale text with no images (e.g., spreadsheets, SAT questions, code)

• The diversity is symbiotic

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, arXiv 2022
Composing Zero-Shot Multimodal Reasoning with Language

• A modular framework in which multiple pretrained models may be composed zero-shot
  • via multimodal-informed prompting, to exchange information with each other and capture new multimodal capabilities, without requiring finetuning.

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, arXiv 2022
Applications to video understanding

• Video-text contrastive models
  • MMV, VideoCLIP, TACo

• Transfer image-text models
  • ActionCLIP

• With Prompt Tuning
  • EfficientPrompt, Bridge-Prompt
Video-text contrastive models

[HTM, Miech et al., 2019], [MIL-NCE, Miech et al., 2020]

Figure credit: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019
Video-text contrastive models

**MMV:** Self-Supervised MultiModal Versatile Networks, NeurIPS 2020

**VideoCLIP:** Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021

**TACo:** Token-aware Cascade Contrastive Learning for Video-Text Alignment, ICCV 2021
Video-text contrastive models

MMV: Self-Supervised MultiModal Versatile Networks, NeurIPS 2020

\[ \mathcal{L}(x) = \lambda_{va} \text{NCE}(x_v, x_a) + \lambda_{vl} \text{MIL-NCE}(x_v, x_t) \]

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021

TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment, ICCV 2021
Transfer image-text models

• Attaching importance to the semantic information of label texts rather than simply mapping them into numbers.

ActionCLIP: A New Paradigm for Video Action Recognition, arXiv 2021
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, arXiv 2021
Prompting

- Efficiently adapt one pre-trained visual-language model to novel tasks with minimal training


Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos, CVPR 2022
Continuous prompt vectors

- Optimize a few random vectors that convert the novel tasks into the same format as the pre-training objectives
Model the semantics across adjacent actions

- Simultaneously exploits both out-of-context and contextual information from a series of ordinal actions in instructional videos
More languages

- MMP
- VICTOR
- Tencent-MVSE
Multilingual Multimodal Pre-training

MMP: Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models, NAACL 2021
Understanding Chinese Video and Language via Contrastive Multimodal Pre-Training

VICTOR: Understanding Chinese Video and Language via Contrastive Multimodal Pre-Training, arXiv 2021
A Large-Scale Benchmark Dataset for Multi-Modal Video Similarity Evaluation

Tencent-MVSE: A Large-Scale Benchmark Dataset for Multi-Modal Video Similarity Evaluation, CVPR 22
Summary

**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)

**Comprehensive Benchmarks**
- VALUE (Li et al. '21)
- GIT (Wang et al. '22)

**Modeling Multi-channel Videos**
- HERO (Li et al. '20)
- MMT (Gabeur et al. '20)
- HIT (Liu et al. '21)
- ClipBERT (Lei et al. '21)
- Clip4Caption (Tang et al. '21)

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

**Video Text Understanding**
- VATT (Akbari et al. '21)
- MV-GPT (Seo et al. '22)

**Applications to Video**
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)
- Bridge-Prompt (Han et al. '22)
- P³IV (Zhao 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)

**Enhanced Pre-training Data**
- MERLOT (Zeller et al. '21)
- MERLOT RESERVE (Zeller et al. '22)

**Noise Estimation**
- Noise Estimation (Amrani et al. '20)

**Unified Modeling**
- Uni-Perceiver (Zhu et al. '22)
- SkillNet (Dai 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)
- VideoCLIP (Xu et al. '20)
- HiT (Liu et al. '21)
- VATT (Akbari et al. '21)
- MMV (Alayrac et al. '20)
- ActionCLIP (Wang et al. '21)
- EfficientPrompt (Ju et al. '21)
- Bridge-Prompt (Han et al. '22)
- P³IV (Zhao et al. '22)

**Support-Set Task**
- Support-Set (Patrick et al. '20)
- VIOLET (Fu et al. '21)

**More Languages**
- Tencent-MVSE (Zeng et al. '21)

---

**VideoCLIP (Xu et al. '20)**
- Applications to Video Understanding
- Video Text Understanding
- Video Understanding
- Video CLIP

**VideoBERT (Sun et al. '19)**
- Video Text Pre-training
- Video Text Understanding
- Video Understanding
- Video CLIP

**UniVL (Luo et al. '20)**
- Video Text Pre-training
- Video Text Understanding
- Video Understanding
- Video CLIP

**HTM (Miech et al. '19)**
- Video Text Pre-training
- Video Text Understanding
- Video Understanding
- Video CLIP

**MIL-NCE (Miech et al. '20)**
- Video Text Pre-training
- Video Text Understanding
- Video Understanding
- Video CLIP