Learning from Multi-channel Videos: Methods and Benchmarks

Linjie Li
The “V” in Video

Single-channel Video: taking video frames only to represent a video

Video Frames = Video Input

Language Input → Video-Language (VidL) Model

Vision
The “V” in Video

Single-channel Video: taking video frames only to represent a video

Example: leveraging expert vision features to generate captions for single-channel videos
The “V” in Video

Single-channel Video: taking video frames only to represent a video

SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, CVPR 2022
Videos are Multi-channel in Nature

Multi-channel Video: Visual Frames + Subtitle + Audio

- Video Frames
- ASR/Subtitles
- Audio

Video
Videos are Multi-channel in Nature

Examples from Multi-modal Transformer for Video Retrieval, ECCV 2020
Videos are Multi-channel in Nature

Multi-channel Video: Visual Frames + Subtitle + Audio

How to encode the information from all channels?
Modeling Multi-Channel Videos with Expert Features: MMT

Multi-modal Transformer for Video Retrieval, ECCV 2020
Modeling Multi-Channel Videos with Expert Features: MMT

7 Expert Features
- OCR
  - Pre-trained scene text detector -> pre-trained text recognition model trained on Synth90K -> word2vec embeddings
- Speech
  - Speech transcripts extracted using ASR API, embedded with Word2Vec
- Face
  - Pre-trained Face detector -> pre-trained ResNet50 on VGGFace2 for face classification
- Audio
  - Pretrained CNN models for audio recognition on YT8M
- Scene
  - Pre-trained 2D CNN on Place365 for Scene Classification
- Appearance
  - Pre-trained 2D CNN on ImageNet for Image Classification
- Motion
  - Pre-trained 3D CNN on Kinetics for Action Recognition
Modeling Multi-Channel Videos with Expert Features: MMT

7 Expert Features
- OCR
- Speech
- Face
- Audio
- Scene
- Appearance
- Motion

Text queries in MSRVTT are not specifically designed to describe multi-channel information in videos. Speech and audio expert features are helpful for retrieval performance.
Modeling Multi-Channel Videos with Expert Features: MMT

7 Expert Features
- OCR
- Speech
- Face
- Audio
- Scene
- Appearance
- Motion

Not all visual features are helpful for downstream retrieval performance.

Fig. 4: MSRVTT performance (mean rank; lower is better) after training from scratch, when using only one expert (left), when using all experts but one (middle), when gradually adding experts by greedy search (right).
Modeling Multi-Channel Videos with Expert Features

- **Visual Expert Features**
  - Motion: Pre-trained 3D CNN on Kinetics for Action Recognition
  - Appearance: Pre-trained 2D CNN on ImageNet for Image Classification

- **Audio Expert Features**
  - Pre-trained audio recognition models (e.g., on YT8M or AudioSet)

- **Downstream Applications**
  - Video-Text Retrieval (and Video Understanding)

---

Pre-extracted features from expert models, which are often well-supervised.
Modeling Multi-Channel Videos with Expert Features

- Visual Expert Features
  - Motion: Pre-trained 3D CNN on Kinetics for Action Recognition
  - Appearance: Pre-trained 2D CNN on ImageNet for Image Classification

- Audio Expert Features
  - Pre-trained audio recognition models (e.g., on YT8M or AudioSet)

- Downstream Applications
  - Video-Text Retrieval and Audio-Video Retrieval

Random initialized audio encoder, learned end-to-end through self-supervised multi-modal pre-training

AVLnet: Learning Audio-Visual Language Representations from Instructional Videos, Interspeech 2021

Everything at Once – Multi-modal Fusion Transformer for Video Retrieval, CVPR 2022
(Single-channel) Video+Language Tasks

Data Collection

Task: Write a text description about the video

Examples from SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, CVPR 2022
Multi-channel Video+Language Tasks

TVQA: Localized, Compositional Video Question Answering, EMNLP 2018
Multi-channel Video+Language Tasks

**TVR: Video-Subtitle Moment Retrieval**

Video Corpus

- **Video 1**
  - Bailey: I don't care if he's sleeping, just wake him up.
  - Alex: There were two donors, Izzie. Our heart flatlined.

- **Video 2**
  - Izzie: Well, for what it's worth, I take issue with ...
  - Meredith: This is what I'm saying...

- **Video 3**
  - Castle: I'm so sorry for everything.
  - Mia: Come on, I did some pretty extraordinary things yesterday.

**TVR: Video-Subtitle Moment Retrieval**

- Castle passes the flowers to Mia and Mia takes them. *(video-only)*
- Castle apologizes to the woman while handing her flowers. *(video-text)*

**TVC: Multimodal Video Captioning**

- Ted: Just not on a boat.
- Captain: Fair enough.

**Captions**
- The Captain says its ok if Ted will not be on the ship. *(text-only)*
- The Captain agrees and points at Ted with a glass in his hand. *(video-text)*

**Query:** Alex is on the phone with Izzie and he is updating her on the heart situation.

TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, ECCV 2020
VLP for Multi-Channel Video+Language: HERO

• Architecture
  • **Cross-Modal Transformer** for local temporal alignments between frames and subtitle sentences
  • **Temporal Transformer** for global temporal context
VLP for Multi-Channel Video+Language: HERO

• Architecture
  • **Cross-Modal Transformer** for local temporal alignments between frames and subtitle sentences
  • **Temporal Transformer** for global temporal context

• Pre-training Tasks
  • **Video Subtitle Matching** to learn both global and local alignment between the sampled query and video clips
  • **Frame Order Modeling** to reconstruct the orders of shuffled frames

HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, EMNLP 2020
VLP for Multi-Channel Video+Language: HERO
VLP for Multi-Channel Video+Language: HERO

• Generalization to Single-channel Videos

<table>
<thead>
<tr>
<th>Method \ Task</th>
<th>DiDeMo</th>
<th>DiDeMo w/ ASR</th>
<th>MSR-VTT</th>
<th>MSR-VTT w/ ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
<td>R@100</td>
<td>R@1</td>
</tr>
<tr>
<td>SOTA Baseline</td>
<td>1.59</td>
<td>6.71</td>
<td>25.44</td>
<td>14.90</td>
</tr>
<tr>
<td>HERO</td>
<td>2.14</td>
<td>11.43</td>
<td>36.09</td>
<td>16.80</td>
</tr>
</tbody>
</table>

When augmenting single-channel videos with ASR inputs, the performance is improved.
VLP for Multi-Channel Video+Language: MV-GPT

- Multimodal Video Generative Pre-training with Bi-directional Utterance Generation
  - **Forward Generation**: input frames + present utterance -> future utterance
  - **Backward Generation**: input frames + future utterance -> present utterance
- A pure Transformer-based architecture, end2end trained
VLP for Multi-Channel Video+Language: MERLOT Reserve

- A Pure Transformer-based Architecture
  - Image Encoder: ViT
  - Audio Encoder: Audio Spectrum Transformer
  - Text/Joint Encoder: Transformer

- Pre-training on YT-Temporal 1B data
  - Contrastive Masked Span Matching on text and audio modalities
  - Contrastive matching between transcripts and frames
  - Trained in an end2end manner from scratch
VLP for Multi-Channel Video+Language: MERLOT Reserve

Video-Text Pre-training can help Visual Commonsense Reasoning, an image-text task

Audio pre-training helps, even for the audio-less VCR
Unified Architecture for Multi-Channel Video Encoding: VATT

A plain Transformer architecture for all modalities.

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, NeurIPS 2021
Benchmarking VidL Models
Benchmarking VidL Models

Video Input

Language Input

Video-Text Pre-trained Model

Task

Capability

Data
Benchmarking VidL Models

Video Input

Language Input

Video-Text
Pre-trained Model

Data

Task

Capability
# Benchmarking VidL Models

**Capability 1:** A general VidL system should do well on diverse tasks/domains/datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval Tasks</th>
<th>QA Tasks</th>
<th>Captioning Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>HowTo100M [Miech et al.]</td>
<td>MSRVTT, YouCook2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HERO [Li et al.]</td>
<td>MSRVTT, DiDeMo, TVR, How2R</td>
<td>TVQA, How2QA</td>
<td>TVC</td>
</tr>
<tr>
<td>ActBERT [Zhu and Yang]</td>
<td>MSRVTT, YouCook2</td>
<td>MSRVTT, LMSDC</td>
<td>YouCook2</td>
</tr>
<tr>
<td>ClipBERT [Lei et al.]</td>
<td>MSRVTT, ActivityNet, DiDeMo</td>
<td>TGIF, MSRVTT</td>
<td>-</td>
</tr>
<tr>
<td>VQA-T [Yang et al.]</td>
<td>-</td>
<td>MSRVTT, MSVD, How2QA, ...</td>
<td>-</td>
</tr>
<tr>
<td>Frozen in Time [Bain et al.]</td>
<td>MSRVTT, MSVD, DiDeMo, LSMDC</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VIOLET [Fu et al.]</td>
<td>MSRVTT, MSVD, YouCook2, LSMDC</td>
<td>TGIF, MSRVTT, MSVD, LSMDC</td>
<td>-</td>
</tr>
<tr>
<td>MV-GPT [Seo et al.]</td>
<td>MSRVTT</td>
<td>MSRVTT, ActivityNet</td>
<td>YouCook2, MSRVTT, ...</td>
</tr>
<tr>
<td>MERLOT [Zeller et al.]</td>
<td>-</td>
<td>TGIF, MSRVTT, LSMDC, TVQA,...</td>
<td>-</td>
</tr>
<tr>
<td>MERLOT RESERVE [Zeller et al.]</td>
<td>-</td>
<td>TVQA, MSRVTT, MSVD, ...</td>
<td>-</td>
</tr>
</tbody>
</table>
**Benchmarking VidL Models**

**Capability 2:** A smart VidL system should be able to leverage information from different modalities in video.

---

### Single-Channel Video vs. Multi-Channel Video

---

<table>
<thead>
<tr>
<th>Method</th>
<th>MSRVTT-QA (Single-channel)</th>
<th>TVQA (Multi-channel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HERO [Li et al.]</td>
<td>-</td>
<td>73.6</td>
</tr>
<tr>
<td>DECEMBERT [Tan et al.]</td>
<td>37.4</td>
<td>-</td>
</tr>
<tr>
<td>ClipBERT [Lei et al.]</td>
<td>37.4</td>
<td>-</td>
</tr>
<tr>
<td>SiaSamRea [Lei et al.]</td>
<td>41.6</td>
<td>-</td>
</tr>
<tr>
<td>VQA-T [Yang et al.]</td>
<td>41.5</td>
<td>-</td>
</tr>
<tr>
<td>MERLOT [Zeller et al.]</td>
<td>43.1</td>
<td>78.7</td>
</tr>
<tr>
<td>VIOLET [Fu et al.]</td>
<td>43.7</td>
<td>-</td>
</tr>
<tr>
<td>MV-GPT [Seo et al.]</td>
<td>41.7</td>
<td>-</td>
</tr>
<tr>
<td>MERLOT RESERVE [Zeller et al.]</td>
<td>-</td>
<td>86.5</td>
</tr>
</tbody>
</table>
Benchmarking VidL Models

NLP Benchmarks

GLUE
SuperGLUE
XGLUE
XTREME

(X) Cross-Lingual Transfer Evaluation of Multilingual Encoders

Publicly accessible large-scale multi-task benchmarks can facilitate advances in modeling.
VALUE Benchmark

https://value-benchmark.github.io

• A comprehensive benchmark for Video-And-Language Understanding Evaluation

Multi-channel Video
With both Video Frames and Subtitle/ASR

Diverse Video Domain
Diverse video content from YouTube, TV Episodes and Movie Clips

Various Datasets over Representative Tasks
11 datasets over 3 tasks: Retrieval, Question Answering and Captioning.

Leaderboard!
To track the advances in Video-and-Language research.
Analyzing VidL Tasks and Models

- Existing Benchmarks
- Existing Models
- Improved Benchmarks
- Better Models
Analyzing VidL Tasks

Do existing video-language tasks require temporal reasoning?

Revisiting the “Video” in Video-Language Understanding, CVPR 2022

SSV2-Template/Label Retrieval

Revealing Single Frame Bias for Video-and-Language Learning, arxiv 2022
Analyzing VidL Tasks

Shortcomings in current video retrieval benchmarks and evaluation protocols

Query: "A man doing an origami tutorial"

<table>
<thead>
<tr>
<th>Videos</th>
<th>current</th>
<th>proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;A man doing an origami tutorial&quot;</td>
<td>✓</td>
<td>1.0</td>
</tr>
<tr>
<td>&quot;A demonstration in origami&quot;</td>
<td>✗</td>
<td>1.0</td>
</tr>
<tr>
<td>&quot;a guy explains the steps of folding paper&quot;</td>
<td>✗</td>
<td>1.0</td>
</tr>
<tr>
<td>&quot;a man folding a piece of paper into a paper airplane&quot;</td>
<td>✗</td>
<td>0.8</td>
</tr>
<tr>
<td>&quot;a man drawing a star on a piece of paper&quot;</td>
<td>✗</td>
<td>0.5</td>
</tr>
<tr>
<td>&quot;two 3D character animations fighting&quot;</td>
<td>✗</td>
<td>0.0</td>
</tr>
</tbody>
</table>

• Assumptions in existing video retrieval benchmarks
  • Only a single caption is relevant to a query video and vice versa
• New evaluation protocol and training objective
  • Semantic similarity based on proxy measures with Bag of Words, Part of Speech, Synset, METEOR.

On Semantic Similarity in Video Retrieval, CVPR 2021
Analyzing VidL Models

When do existing video-language models fail?

Verb Manipulation
A: A girl feeding a brown horse.
B: A girl *rides* a brown horse.
C: Football team playing football on a field.
D: The man is drinking beer.
E: Two men playing a video game.

Entity Manipulation
A: Over her shoulders, HARRY glances at RON, who lowers his gaze for a moment.
B: Over her shoulders RON glances at HARRY, who lowers his gaze for a moment.
C: He tries to shake him off.
D: HARRY studies it.
E: RON whispers to HARRY.
Pioneering work in Video-Text Pre-training

- **VideoBERT** (Sun et al. '19)
- **UniVL** (Luo et al. '20)
- **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)
- **VATT** (Akbari et al. '21)
- **MV-GPT** (Seo et al. '22)
- **AVLNet** (Rouditchenko et al. '21)
- **ATP** (Buch et al. '22)
- **Frozen** (Bain et al. '21)
- **MERLOT** (Zeller et al. '21)
- **HD-VILA** (Xue et al. '22)
- **Noise Estimation** (Amrani et al. '20)
- **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)
- **MMV** (Alayrac et al. '20)
- **VideoCLIP** (Xu et al. '20)
- **TACo** (Yang et al. '21)
- **ActionCLIP** (Wang et al. '21)
- **EfficientPrompt** (Ju et al. '21)
- **Bridge-Prompt** (Zhao et al. '22)
- **TAN** (Han et al. '22)
- **CLAP** (Xu et al. '22)
- **Universe** (Zhou et al. '21)
- **Socratic Models** (Zeng et al. '22)
- **UniPerceiver** (Zhu et al. '22)
- **LAVENDER** (Li et al. '22)
- **Uniform Modeling** (Dai et al. '22)
- **More Languages** (Zheng et al. '21)
- **Tencent-MSVE** (Zeng et al. '21)
- **BridgePrompt** (Ju et al. '21)
- **All-in-One** (Wang et al. '22)

---

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

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

**Advanced Pre-training Tasks**
- **OA-Trans** (Wang et al. '22)
- **ALPRO** (Li et al. '22)

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

**Transfer Image-Text Models**
- **Flamingo** (Alayrac et al. '22)

**Modeling Multi-channel Videos**
- **Clip4Clip** (Luo et al. '21)
- **Clip4Caption** (Tang et al. '21)
- **VATT** (Akbari et al. '21)
- **MV-GPT** (Seo et al. '22)
- **AVLNet** (Rouditchenko et al. '21)

**Applications to Video Understanding**
- **ATP** (Buch et al. '22)
- **Frozen** (Bain et al. '21)
- **MERLOT** (Zeller et al. '21)
- **HD-VILA** (Xue et al. '22)
- **Noise Estimation** (Amrani et al. '20)
- **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)
- **MMV** (Alayrac et al. '20)
- **VideoCLIP** (Xu et al. '20)
- **TACo** (Yang et al. '21)
- **ActionCLIP** (Wang et al. '21)
- **EfficientPrompt** (Ju et al. '21)
- **Bridge-Prompt** (Zhao et al. '22)
- **TAN** (Han et al. '22)
- **CLAP** (Xu et al. '22)
- **Universe** (Zhou et al. '21)
- **Socratic Models** (Zeng et al. '22)
- **UniPerceiver** (Zhu et al. '22)
- **LAVENDER** (Li et al. '22)
- **Uniform Modeling** (Dai et al. '22)
- **More Languages** (Zheng et al. '21)
- **Tencent-MSVE** (Zeng et al. '21)
- **BridgePrompt** (Ju et al. '21)
- **All-in-One** (Wang et al. '22)
Looking forward

• How to effectively **leverage advances in unimodal models**
  • E2E pre-training > pre-extracted expert features
  • E2E pre-training (from scratch) is usually time-consuming, extremely data-hungry, very computationally expensive

• Benchmarks on truly **multi-channel video understanding**
  • Existing VidL datasets do not test on audio understanding in video