HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training

Linjie Li*, Yen-Chun Chen*, Yu Cheng, Zhe Gan, Licheng Yu, Jingjing Liu

Microsoft Dynamics 365 AI Research

* Equal Contribution
Vision + Language Pre-training

Vision: Image
Language: Textual Descriptions

UNITER [Chen et al. 2019]

ViLBERT [Lu et al. NeurIPS 2019]

LXMERT [Tan and Bansal, EMNLP 2019]
Video + Language Pre-training

Video: Sequence of image frames
Language: Subtitles/Narrations

00:00:02 --> 00:00:04
That’s why you won’t go out with her again?
00:00:34 --> 00:00:36
- Thank God you’re here. Listen to this.
- What?
00:00:68 --> 00:00:68
(Joey:) Joey doesn’t share food!
Video + Language Pre-training

- Limitations of existing methods
  - Video + Text inputs are directly concatenated, losing the temporal alignment
  - Pre-training tasks directly borrowed from Image + Text pre-training
  - Pre-training datasets limited to narrated instructional videos from Howto100M [Miech et al. ICCV 2019]

- **HERO** (Hierarchical EncodeR for Omni-representation learning)
  - New model architecture:
    - Local temporal alignments between frames and subtitles are captured by a Cross-modal Transformer
    - Global temporal context are modeled by a Temporal Transformer
  - New Pre-training tasks: Video-Subtitle Matching and Frame Order Modeling
  - Diverse Pre-training Datasets: Howto100M and TV dataset [Lei at al. ACL 2018]
    - We further collect two downstream datasets based on Howto100M
HERO: Hierarchical EncodeR for Omni-representation learning

00:00:02 --> 00:00:04
That's why you won't go out with her again?

00:00:34 --> 00:00:36
- Thank God you're here. Listen to this.
- What?

00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!
HERO: Hierarchical EncodeR for Omni-representation learning

• Temporally align subtitle sentences with frames

• Frame features: 2D ResNet Features [He et al. CVPR 2016] and 3D SlowFast Features [Feichtenhofer et al. ICCV 2019]

• Subtitle sentences are tokenized and each word are embedded following RoBERTa [Liu et al. 2019]
**HERO:** Hierarchical EncodeR for Omni-representation learning

00:00:02 -- 00:00:04
That's why you won't go out with her again?

00:00:34 -- 00:00:36
- Thank God you're here. Listen to this.
- What?

00:00:66 -- 00:00:68
(Joey:) Joey doesn't share food!
Pre-training **HERO**

- Pre-training Tasks
  - Masked Language Modeling (MLM)
  - Masked Frame Modeling (MFM)
  - *Video-Subtitle Matching (VSM)*
  - *Frame Order Modeling (FOM)*
**Masked Language Modeling (MLM)**

Word Tokens of Subtitle $S_i$: $w_{s_i} = \{w_{s_i}^j\}_{j=1}^L$

Visual Frames aligned with $S_i$: $v_{s_i} = \{v_{s_i}^j\}_{j=1}^K$

Masking Indices: $m \in \mathbb{N}^M$

Loss Function of MLM: $L_{\text{MLM}}(\theta) = -\mathbb{E}_D \log P_{\theta}(w_{s_i}^m|w_{s_i}^m \setminus v_{s_i})$
All subtitle sentences: \( s = \{ s_i \}_{i=1}^{N_s} \)

Visual Frames: \( v = \{ v_i \}_{i=1}^{N_v} \)

Masking Indices: \( m \in \mathbb{N}^M \)

Loss Function of MFM: \( \mathcal{L}_{MFM}(\theta) = \mathbb{E}_D f_\theta(v_m|v_{\backslash m}, s) \)

(1) Masked Frame Feature Regression (MFFR)

\[ f_\theta(v_m|v_{\backslash m}, s) = \sum_{i=1}^{M} \| h_\theta(v^{(i)}_m) - r(v^{(i)}_m) \|_2 \]

(2) Masked Frame with Noise Contrastive Estimation (M-NCE)

\[ f_\theta(v_m|v_{\backslash m}, s) = \sum_{i=1}^{M} \log \text{NCE}(g_\theta(v^{(i)}_m)|g_\theta(v_{\text{neg}})) \]
Video Subtitle Matching (VSM)

VSM Flow

Collect Frames

Frame Features

VSM Frame Features

Word Features

Cosine Similarity

Cross-Modal Transformer

Temporal Transformer

Query Encoder (Subtitle as Query)

Cosine

Other video clips

Temporal Transformer

Shared

Global Alignment

Other video clips

Local Alignment

Frame Features

Word Embed.

Query Encoder

Thank God you're here.
Listen to this.
What?

(00:00:34 --> 00:00:36)

(00:00:66 --> 00:00:68)

(00:00:00 --> 00:00:02)

That's why you won't go out with me, Joey.

(Joey:) Joey doesn't share food!
Video Subtitle Matching (VSM)

Start and end index of overlapping frames: $y_{st}$, $y_{ed}$

Loss function of local alignments:

$$\mathcal{L}_{local} = -\mathbb{E}_D \log(p_{st}[y_{st}]) + \log(p_{ed}[y_{ed}])$$
Video Subtitle Matching (VSM)

Positive and negative video-subtitle pairs: $(s_q, v)$, $(s_q, \hat{v})$, $(\hat{s}_q, v)$

Similarity measure: $S$

Hinge loss: $\mathcal{L}_h(S_{pos}, S_{neg}) = \max(0, \delta + S_{neg} - S_{pos})$

Loss function of global alignments:

$$\mathcal{L}_{global} = -\mathbb{E}_D[\mathcal{L}_h(S_{global}(s_q, v), S_{global}(\hat{s}_q, v)) + \mathcal{L}_h(S_{global}(s_q, v), S_{global}(s_q, \hat{v}))]$$
Frame Order Modeling (FOM)

Reorder Indices: \( \mathbf{r} = \{ r_i \}_{i=1}^R \in \mathbb{N}^R \)

Original timestamp: \( \mathbf{t} = \{ t_i \}_{i=1}^R \)

Loss Function of FOM: \( \mathcal{L}_{\text{FOM}} = -E_{\mathcal{D}} \sum_{i=1}^R \log \mathbb{P}(r_i, t_i) \)
Pre-training HERO

• Pre-training Tasks
  • Masked Language Modeling (MLM)
  • Masked Frame Modeling (MFM)
  • Video-Subtitle Matching (VSM)
  • Frame Order Modeling (FOM)

• Pre-training Datasets
  • TV Dataset
  • Howto100M Dataset
Our Pre-training Data for Video + Language

TV Dataset
- 22K video clips from 6 popular TV shows
- Each video clip is 60-90 seconds long
- Dialogue (“character name: subtitle”) is provided

Howto100M Dataset
- 1.22M instructional videos from YouTube
- Exclude videos in non-English languages and cut the rest into 60-second clips
- 660K video clips with English subtitles
Video + Language Downstream Tasks

Video: Sequence of image frames
Language: Subtitles/Narrations

Video Captioning
Caption: Joey’s dating policy: never shares food!

Text-based Video Moment Retrieval
Query: Joey’s dating policy: never shares food!

Video Question Answering
Question: Why did Joey complain about his date?
Answer: She took Joey’s fries
Downstream Task 1: Video Moment Retrieval

Video Moment Retrieval = Video Retrieval + Moment Retrieval

- **Subtask I: Video Retrieval**
  - From video corpus, retrieve the most relevant video clip described by the query

- **Subtask II: Moment Retrieval**
  - Given the query, localize the correct moment from the most relevant video clip

- **Evaluation**:
  - Average recall at K (R@K) over all queries
  - Temporal Intersection over Union (tIOU) is used to measure the performance of moment retrieval

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

**TVR** [Lei et al. 2020]
Downstream Task 1: Video Moment Retrieval
Downstream Task 1: Video Moment Retrieval

Query: Joey's dating policy: never shares food!
Downstream Task 2: Video Question Answering

TVQA [Lei et al. EMNLP 2018]
Downstream Task 2: Video Question Answering

1. **Frame Features**
   - Collect Frames
   - Flow

2. **Cross-Modal Transformers**
   - Word Embed.
   - Shared

3. **Temporal Transformer**
   - QA-aware Video Representations

---

**Dialogue:**

**J:** Joey doesn't share food!

**Q:** Why did Joey complain about his date?

**A:** She took Joey's fries

**J:** That's why you won't go out with her again?

**Q:** Why did Joey complain about his date?

**A:** She took Joey's fries

---

**J:** Thank God you're here.
Listen to this.

**Q:** Why did Joey complain about his date?

**A:** She took Joey's fries

**J:** Joey doesn't share food!

**Q:** Why did Joey complain about his date?

**A:** She took Joey's fries
Downstream Task 2: Video Question Answering

[Diagram showing a flow of video clips with associated questions and answers, processed through Cross-Modal Transformer and Temporal Transformer.]
Downstream Task 2: Video Question Answering

Frame Features → Cross-Modal Transformer → Temporal Transformer → QA-aware Video Representations

Q: Why did Joey complain about his date?
A: She took Joey's fries

Q: Why did Joey complain about his date?
A: She took Joey's fries

Q: Why did Joey complain about his date?
A: She took Joey's fries

That's why you won't go out with her again?

Thank God you're here.
Listen to this.
What?

(Johnny:) Joey doesn't share food!
Downstream Data Collection

Text-based Video Moment Retrieval

- **Howto100M-R**
  - 67K text queries are collected for 30K 60-second video clips from Howto100M
- **Instructions:**
  - First, select a video segment
  - Then, write a caption describing the selected segment

Caption:

8 to 20 words
Downstream Data Collection

Video Question Answering

• Howto100M-QA
  • QA collected for video segments annotated from video moment retrieval
  • On average, 2 questions per video segment
  • One correct answer and three wrong answers are written by the same annotator
  • Using adversarial matching [Zeller et al. CVPR 2019] to construct harder negative answers
Ablation Study

1. Best combination: MLM + MNCE + FOM + VSM
2. QA tasks benefit from FOM
3. Retrieval tasks benefit from VSM
4. Adding more data generally give better results
Ablation Study

- Comparison with two baseline models with/without pre-training
  - F-TRM
    - A flat BERT-like encoder
    - Input is a single sequence by concatenating video frames and subtitle sentences
  - H-TRM
    - Replacing Cross-modal Transformer with RoBERTa to encode subtitle only
    - Max-pooled subtitle sentence embeddings is added to temporally aligned frame embeddings

<table>
<thead>
<tr>
<th>Pre-training</th>
<th>Model</th>
<th>TVR</th>
<th>TVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
<td>R@100</td>
</tr>
<tr>
<td>No(^8)</td>
<td>F-TRM</td>
<td>1.99</td>
<td>7.76</td>
</tr>
<tr>
<td></td>
<td>H-TRM</td>
<td>2.97</td>
<td>10.65</td>
</tr>
<tr>
<td></td>
<td>HERO</td>
<td>2.98</td>
<td>10.65</td>
</tr>
<tr>
<td>Yes</td>
<td>H-TRM</td>
<td>3.12</td>
<td>11.08</td>
</tr>
<tr>
<td></td>
<td>HERO</td>
<td>4.44</td>
<td>14.69</td>
</tr>
</tbody>
</table>

1. Without pre-training, HERO and H-TRM outperforms F-TRM
   - Inherent temporal alignment between two modalities of videos is important

2. With pre-training, HERO outperforms H-TRM
   - Cross-modal interactions between visual frames and its local textual context is critical
Comparison with SOTA Models

1. Compared to task-specific SOTA models, HERO outperforms with/without pre-training.
2. Pre-training greatly lift HERO’s performance on downstream tasks.
3. HERO achieves state-of-the-art results on all four downstream tasks.

<table>
<thead>
<tr>
<th>Method</th>
<th>TVR</th>
<th></th>
<th>TVQA</th>
<th>Howto100M-QA</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>XML (Lei et al., 2020)</td>
<td>2.70</td>
<td>8.93</td>
<td>15.34</td>
<td>2.06</td>
</tr>
<tr>
<td>STAGE (Lei et al., 2019)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HERO w/o pre-training</td>
<td>2.98</td>
<td>10.65</td>
<td>18.42</td>
<td>2.17</td>
</tr>
<tr>
<td>HERO w/ pre-training</td>
<td><strong>4.34</strong></td>
<td><strong>13.97</strong></td>
<td><strong>21.78</strong></td>
<td><strong>2.98</strong></td>
</tr>
</tbody>
</table>
Thank You