Hierarchical Self-Supervised Representation Learning for Movie Understanding

AWS AI Labs

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, Davide Modolo
Movie understanding

• Movies are:
  • a lot more complex than short YouTube videos (e.g., Kinetics)
  • a lot fewer (MovieClips: 3k vs Kinetics 650k)
  • classic video action models are not enough (e.g., SlowFast, I3D, etc.)
  • require reasoning at many levels

  From simple low-level actions ➔ Hugging

  To high-level semantic narratives ➔ The actors are sad because the boat is sinking and they don’t know if they will survive

[From MovieNet dataset]
Recent advances on movie understanding

• VidSitu’s hierarchical movie model [1]
  • Low-level video backbone encoder
  • Higher-level transformer contextualizer
  • Fully-supervised learning

• Challenges:
  • Extremely difficult to annotate large-scale movie datasets

• Our solution:
  • Self-Supervised pre-training

This paper

• Sequentially pretrain a low-level backbone and a high-level contextualizer in a self-supervised manner

• Benefits:
  • Each level can specialize better
  • Reduces data needs → Each level can be trained on its own dataset
  • SOTA performance
This paper: Low-level backbone

- Extracts low-level appearance and motion cues for people, objects and scenes from raw pixels
- High capacity
- Trained on a large amount of YouTube videos (e.g., Kinetics)
- Trained using contrastive learning objective
This paper: High-Level Contextualizer

- Learn to contextualize the neighboring low-level visual tokens
- Low capacity
- Trained on a small amount of movie data with stronger semantic and temporal structures (e.g., VidSitu, LVU)
- Trained using mask prediction objective
Results on VidSitu: Semantic Role Prediction

• Goal: predict various semantic role labels for each verb. For example, the agent ("person") and the patient ("ball") of the verb ("throw"), as well as other attributes like the scene where the verb is happening ("a gym")

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-training</th>
<th>Contextualizer</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>VidSitu paper [1]</td>
<td>Fully-Supervised (K400)</td>
<td>None</td>
<td>47.06</td>
</tr>
<tr>
<td>This paper</td>
<td>Self-Supervised (K400)</td>
<td>Self-Supervised (VS)</td>
<td>60.34 (+13.28)</td>
</tr>
<tr>
<td>This paper</td>
<td>Self-Supervised (K400)</td>
<td>Self-Supervised (LVU)</td>
<td>61.18 (+14.12)</td>
</tr>
</tbody>
</table>

Results on VidSitu: Event Relation Prediction

- Goal: 4-way classification problem between four relation types: “A is enabled by B”, “A is a reaction to B”, “A causes B”, and “A is unrelated to B”

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-training</th>
<th>Contextualizer</th>
<th>Top1-Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>VidSitu paper</td>
<td>Fully-Supervised (K400)</td>
<td>None</td>
<td>39.91</td>
</tr>
<tr>
<td>This paper</td>
<td>Self-Supervised (K400)</td>
<td>Self-Supervised (VS)</td>
<td>41.62 (+1.71)</td>
</tr>
</tbody>
</table>

Thank you!