iPerceive

Applying Common-Sense Reasoning to Multi-Modal Dense Video Captioning and Video Question Answering
“I can see nothing,” says Watson.

“On the contrary, Watson, you can see everything. You fail, however, to reason from what you see.”

Sherlock Holmes, The Adventure of the Blue Carbuncle
Why iPerceive?

• **Background**
  
  • Today’s computer vision systems are good at telling us the “what” (e.g., classification, segmentation) and “where” (e.g., detection, localization, tracking)
  
  • Common-sense reasoning, which leads to the interesting question of “why”, is a thinking gap in today’s pattern-learning-based systems

• **Common-sense reasoning**
  
  • Fundamentally different from what prior work in the domain of video understanding uses: the conventional likelihood, \( P(Y \mid X) \)

• **Need for common-sense reasoning**
  
  • Failing to factor in causality leads to the unfortunate conclusion that the co-existence of objects \( X \) and \( Y \) might be attributed to spurious observational bias [2]
  
  • For e.g., if a keyboard and mouse are often observed on a table, the model learns to develop an “association” between the two
  
  • The underlying common-sense that the keyboard and mouse are parts of a computer would not be inferred
  
  • In fact, the duo would be wrongly associated as being part of a table
  
  • In the event that a keyboard and mouse are observed outside of a tabular setting, the model can commit a cognitive error

---


Common-Sense Generation for Videos

• Problem statement
  • Determining causal context is challenging
  • How do you figure out the cause → effect relationship between objects?

• Solution
  • Use the proxy task of predicting the contextual objects of an event as the training objective
  • This process enables self-supervised common-sense learning

• Process
  • Building upon the approach in [3], we carry out the following deliberate “borrow-put” experiment for an event identified using the proposal module:
    1. “borrow” non-local context, say an object Z from another event
    2. “put” Z in the context of object X and object Y
    3. test if object X still causes the existence of object Y given Z
  • This experiment helps determine if the chance of Z is dependent on X or Y
  • Therefore, by using \( P(Y \mid do(X)) \) as the learning objective instead of \( P(Y \mid X) \), the observational bias from the “apparent” context could be alleviated

• Inference
  • Each RoI X is fed into two sibling branches:
    • Self-predictor: predicts the class of the “center” object \( x \in X \)
    • Context-predictor: predicts the “center” object’s context labels \( y \in Y \), using “do” calculus

\[
P(Y \mid do(X)) = \sum_z P(Y \mid X, z) P(z)
\]
where,
• \( X \) is the “center” object
• \( Y \) is a “context” object
• \( z \in Z \) is an object from another event, a “confounding” agent that adds spurious observational bias around objects \( X \) and \( Y \)

\[
P(Y \mid do(X)) = E_z(\text{Softmax}(f(x, z)))
\]
where,
• \( f() \) calculates the logits for N categories and \( E_z \) requires expensive sampling of \( z \) over the set of confounder objects \( Z \)

iPerceive DVC

- iPerceive DVC accepts inputs from multiple modalities and consists of three components: event proposal, common-sense reasoning and captioning.
  - **Event proposal**
    - A bi-directional LSTM accumulates visual cues from past and future context over time and predicts the endpoints of each event in the video.
  - **Common-sense reasoning module**
    - We generate common-sense vectors from the temporal events that the event proposal module localizes.
  - **Captioning module**
    - Features from all modalities are sent to the corresponding encoder-decoder Transformers.
    - Upon fusing the processed features we finally output the next word in the caption using the distribution over the vocabulary.
iPerceive VideoQA

- **Feature fusion**
  - Encode features using a convolutional encoder
  - Generate common-sense vectors from the input video sequence
  - Extract dense captions using iPerceive DVC
  - Features from all modalities (video, dense captions, QA and subtitles) are then fed to dual-layer attention: word/object and frame-level

- **Frame selection**
  - Upon fusing the attended features, we calculate frame-relevance scores
Common-Sense Reasoning Loss

- The common-sense module in iPerceive DVC and VideoQA can be split into two components:
  - **Self predictor loss**
    - For a “center” object \( x \in X \) in the video frame at time \( t \), the self-predictor loss \( L_{self} \) can be defined using negative log likelihood
  - **Context predictor loss**
    - For a “context” object \( y_i \in Y \) in the video frame at time \( t \), the context-predictor loss \( L_{cxt} \) can be defined for a pair of RoI feature vectors \( x \) and \( y_i \) using negative log likelihood
  - The overall multi-task loss for each RoI \( X \) is the self + context predictor loss

\[
L_{self}(p, x^c, t) = -\log(p|x^c^c|)
\]

where,
- \( p \) is the probability distribution output of the self-predictor over \( N \) categories for \( X \);
- \( x^c \) is the ground-truth class of RoI \( X \)

\[
L_{cxt}(p, y^c_i, t) = -\log(p_i|y^c_i|)
\]

where,
- \( y^c_i \) is the ground-truth label for \( y_i \);
- \( p_i \) is calculated using \( p_i = P(Y_i|do(X)) \) and \( p = (p_1, ..., p_N) \) is the probability over \( N \) categories

\[
L_{cat} = L_{self} + \frac{1}{K} \sum_i L_{cxt}
\]

Qualitative sampling of iPerceive DVC

Captioning results for a sample video from the ActivityNet Captions validation set show better performance owing to common-sense reasoning and end-to-end training.
Results

Comparison of iPerceive DVC with the state-of-the-art.

<table>
<thead>
<tr>
<th>Method</th>
<th>GT Proposals</th>
<th>Learned Proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B@3</td>
<td>B@4</td>
</tr>
<tr>
<td><strong>Seen full dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krishna et al. [23]</td>
<td>4.09</td>
<td>1.60</td>
</tr>
<tr>
<td>Wang et al. [49]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zhou et al. [58]</td>
<td>5.76</td>
<td>2.71</td>
</tr>
<tr>
<td>Li et al. [77]</td>
<td>4.55</td>
<td>1.62</td>
</tr>
<tr>
<td><strong>Seen part of the dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rahman et al. [38]</td>
<td>3.04</td>
<td>1.46</td>
</tr>
<tr>
<td>Iashin et al. [14]</td>
<td>4.12</td>
<td>1.81</td>
</tr>
<tr>
<td>iPerceive</td>
<td>5.23</td>
<td>2.34</td>
</tr>
<tr>
<td>Iashin et al. (all modalities)</td>
<td>5.83</td>
<td>2.86</td>
</tr>
<tr>
<td>iPerceive (all modalities)</td>
<td><strong>6.13</strong></td>
<td><strong>2.98</strong></td>
</tr>
</tbody>
</table>

Comparison of iPerceive VideoQA with the state-of-the-art.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>BRT</th>
<th>Friends</th>
<th>HMYM</th>
<th>Grey</th>
<th>House</th>
<th>Castle</th>
<th>Val (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lei et al. [24]</td>
<td>66.66</td>
<td>70.25</td>
<td>65.78</td>
<td>64.02</td>
<td>67.20</td>
<td>66.84</td>
<td>63.96</td>
<td><strong>65.85</strong></td>
</tr>
<tr>
<td>Kim et al. [19]</td>
<td>66.77</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lei et al. [18]</td>
<td>67.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kim et al. [17]</td>
<td>74.09</td>
<td>74.04</td>
<td>73.03</td>
<td>74.34</td>
<td>73.44</td>
<td>74.68</td>
<td>74.86</td>
<td>74.20</td>
</tr>
<tr>
<td>iPerceive VideoQA</td>
<td><strong>75.15</strong></td>
<td><strong>75.32</strong></td>
<td><strong>74.22</strong></td>
<td><strong>75.14</strong></td>
<td><strong>74.42</strong></td>
<td><strong>75.22</strong></td>
<td><strong>75.77</strong></td>
<td><strong>76.97</strong></td>
</tr>
</tbody>
</table>

*Algorithms were split into those which “saw” all training videos and others which trained on partially available data (since some YouTube videos which were part of the ActivityNet Captions dataset are no longer available).

Ablation studies for iPerceive DVC to assess the impact of common-sense reasoning and end-to-end training as design decisions.

<table>
<thead>
<tr>
<th>Common-Sense Reasoning</th>
<th>End-to-End Training</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td><strong>7.31</strong></td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>7.42</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td><strong>7.51</strong></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td><strong>7.87</strong></td>
</tr>
</tbody>
</table>

Ablation studies for iPerceive VideoQA to assess the impact of common-sense reasoning and iPerceive DVC as design decisions.

<table>
<thead>
<tr>
<th>Common-Sense Reasoning</th>
<th>iPerceive DVC</th>
<th>Val (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td>74.20</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>75.42</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>75.55</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td><strong>76.97</strong></td>
</tr>
</tbody>
</table>
Broader Impact

• Humans perceive their immediate world by understanding their surroundings, à la playing a video of what's around a machine and hoping it will make sense of its environment.

• We feel that video understanding in general is important to close the “gap” between man and machine.

• Our work propels the idea of causal reasoning for machines and bring us one step closer to the ultimate goal of visual-linguistic causal reasoning which is one of the distinct qualities that make us human.

• Since our work is easily portable, we hope that the promising results in our work would encourage researchers to further explore the domain of common-sense reasoning and apply it to new applications in the field of video and language understanding.
Thank you!

For more, iperceive.amanchadha.com