Seminar on Current Works in Computer Vision

CLEVRER: Collision Events for Video Representation and Reasoning
Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion
Motivation

- Human intelligence: Ability to reason about temporal and causal events
- Most video reasoning benchmarks focus on pattern recognition
• Focus on causal structure instead of pattern recognition

• Explore temporal and causal structures using videos of objects with simple appearance

• CLEVRER is introduced
• Diagnostic video dataset for temporal and causal reasoning

I. Descriptive
Q: What shape is the object that collides with the cyan cylinder?  A: cylinder
Q: How many metal objects are moving when the video ends?  A: 3

II. Explanatory
Q: Which of the following is responsible for the gray cylinder’s colliding with the cube?
   a) The presence of the sphere  A: b)
   b) The collision between the gray cylinder and the cyan cylinder

III. Predictive
Q: Which event will happen next
   a) The cube collides with the red object
   b) The cyan cylinder collides with the red object  A: a)

IV. Counterfactual
Q: Without the gray object, which event will not happen?
   a) The cyan cylinder collides with the sphere
   b) The red object and the sphere collide  A: a), b)
Overview

• Introduction
• Related Work
• Structure of CLEVRER
• Baseline Evaluation on CLEVRER
• Neuro-Symbolic Dynamic Reasoning
• Conclusion
Datasets

- CLEVRER includes all of the components that other datasets offer

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Video</th>
<th>Diagnostic Annotations</th>
<th>Temporal Relation</th>
<th>Explanation</th>
<th>Prediction</th>
<th>Counterfactual</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>CLEVR</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>COG</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>VCR</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>GQA</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>TGIF-QA</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>MovieQA</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>MarioQA</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>TVQA</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Social-IQ</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>CLEVRER (ours)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
• Diagnostic dataset
  ➢ Visual reasoning abilities

• Requires complex reasoning to solve
  ➢ Synthetic images
  ➢ Automatically generated questions
Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
CLEVR vs CLEVRER

• CLEVR operates on images, CLEVRER on videos

• Objects have similar compositional intrinsic attributes

• Same software and parameters used for rendering
  ➢ Image generation
  ➢ Video generation

• Different focus of questions
  ➢ CLEVR: Static properties of objects
  ➢ CLEVRER: Temporal and causal aspects of objects and events in videos
• Can be positioned in the context of 3 research directions
  ➢ Video understanding
  ➢ Visual question answering
  ➢ Physical and causal reasoning
Video Understanding

- Large scale video datasets => joint video and language understanding
  - Video captioning
  - Localizing video segments from natural language queries
  - Video question answering

- MovieQA, TGIF-QA, TVQA
  - Real-world videos and human-generated questions

- Social-IQ
  - Causal relations in human social interactions

- COG, MarioQA
  - Simulated environments used to generate data and controllable reasoning tasks
CLEVRER

- Causal relations in object dynamics and physical interactions
- Introduces tasks: Description, explanation, prediction, counterfactuals
- Emphasize compositionality in visual and logic context
• VQA
  ➢ Based on cloud-sourced real images and human-generated questions

• CLEVR
• VCR
  ➢ Explanations and hypothesis judgements based on common sense

• GQA
  ➢ Synthetic compositional questions to real images
• Related to research on learning scene dynamics for physical and causal reasoning.
  - Directly from images
  - OR from symbolic representation of the environment

• In CLEVRER: How a learned dynamics model contributes to causal reasoning
Overview

- Introduction
- Related Work
- **Structure of CLEVRER**
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion
The CLEVRER Dataset

I. Descriptive
Q: What shape is the object that collides with the cyan cylinder? A: cylinder
Q: How many metal objects are moving when the video ends? A: 3

II. Explanatory
Q: Which of the following is responsible for the gray cylinder’s colliding with the cube? A: b)
   a) The presence of the sphere
   b) The collision between the gray cylinder and the cyan cylinder

III. Predictive
Q: Which event will happen next
   a) The cube collides with the red object
   b) The cyan cylinder collides with the red object A: a)

IV. Counterfactual
Q: Without the gray object, which event will not happen?
   a) The cyan cylinder collides with the sphere
   b) The red object and the sphere collide A: a), b)
Structure of CLEVRER

- Has 2 parts
  - Videos of object motion and collisions, uses a physics engine
  - Questions that are generated about each video
- 20,000 videos and 300,000 questions
Videos

- 10,000 training, 5,000 validation and 5,000 testing videos
  - 5 second long

- Generated
  - By object motion simulating physics engine
  - By graphs engine that renders the frames
  - From simulated motion traces

- Intrinsic attributes sampled randomly for each object
  - Identical objects are prohibited => unique object with combination of the 3 attributes (Shape, Material, Color)
Randomly initialized objects + new objects with overlapping motion traces => complex causal structure

Example: A sphere collides with a cube then a cylinder.
- First collision (Sphere + cube) and the cube cause the second collision
• Each video paired with machine-generated questions
  ➢ Descriptive
  ➢ Explanatory
  ➢ Predictive
  ➢ Counterfactual

• Each question paired with functional program executable

• No questions about static object properties
Questions - Descriptive

(a) First collision

(b) Cyan cube enters

(c) Second collision

(d) Video ends

Q: What shape is the object that collides with the cyan cylinder?  
A: cylinder

Q: How many metal objects are moving when the video ends?  
A: 3
Q: Which of the following is responsible for the gray cylinder’s colliding with the cube?
   a) The presence of the sphere
   b) The collision between the gray cylinder and the cyan cylinder

A: b)
Q: Which event will happen next
a) The cube collides with the red object
b) The cyan cylinder collides with the red object

A: a)
Q: Without the gray object, which event will not happen?
   a) The cyan cylinder collides with the sphere
   b) The red object and the sphere collide

A: a), b)
• Tree-structured functional program for each question
Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion
Baseline Evaluation on CLEVRER

• Evaluate performances of different baseline models
• 3 Families of baseline models
  ➢ Language-only models
  ➢ Video question answering
  ➢ Compositional visual reasoning
• Weak baselines that depend on question input
  ➢ Assess language biases in CLEVRER

• Q-type (random)
  ➢ Uniformly sample answer from sample space
  ➢ OR randomly select choices for multiple-choice questions

• Q-type (frequent)
  ➢ Choose most frequent answer in training set for each question type
• LSTM
  - Word embedding pretrained on Google News corpus
  - To encode input question
  - And process the sequence with LSTM
• Models that rely on both video and language inputs

• CNN+MLP extracts features via CNN
  ➢ Questions encoded by taking the avg. of pretrained word embeddings
  ➢ Features sent to an MLP for answer prediction

• CNN+LSTM also extracts features via CNN
  ➢ Uses final state of an LSTM for answer prediction
• TVQA
  ➢ Multi-stream end-to-end neural model
  ➢ State of the art for video question answering

• TVQA+
  ➢ Attribute-aware object-centric features applied to TVQA

• Memory
  ➢ Incorporate heterogeneous memory with multimodal attention
Model Details - Compositional visual reasoning

• CLEVR
  ➢ Emphasizes complexity and compositionality in the logic and visual context

• IEP (V)
  ➢ IEP applies neural program execution for visual reasoning on images
  ➢ Same approach => program based video reasoning task

• Tbd-net (V)
  ➢ Spatial-temporal attention over the video feature space
Model Details - Compositional visual reasoning

• MAC (V)
  ➢ MAC modified to apply temporal attention unit across video frames
  ➢ Generate latent encoding for video

• MAC (V+)
  ➢ MAC (V) augmented
  ➢ Add segmentation masks of objects in frames + label them by their intrinsic attributes => object aware video features
• Descriptive questions
  ➢ Accuracy => compare predicted answer token with ground truth

• Multiple choice questions
  ➢ Per-option accuracy => Overall correctness on single options (Across all questions)
  ➢ Per-question accuracy => Correctness of the question as a whole (All choices selected correctly)
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Descriptive</th>
<th>Explanatory</th>
<th>Predictive</th>
<th>Counterfactual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>per opt.</td>
<td>per ques.</td>
<td>per opt.</td>
</tr>
<tr>
<td>Q-type (random)</td>
<td>29.2</td>
<td>50.1</td>
<td>8.1</td>
<td>50.7</td>
</tr>
<tr>
<td>Q-type (frequent)</td>
<td>33.0</td>
<td>50.2</td>
<td>16.5</td>
<td>50.0</td>
</tr>
<tr>
<td>LSTM</td>
<td>34.7</td>
<td>59.7</td>
<td>13.6</td>
<td>50.6</td>
</tr>
<tr>
<td>CNN+MLP</td>
<td>48.4</td>
<td>54.9</td>
<td>18.3</td>
<td>50.5</td>
</tr>
<tr>
<td>CNN+LSTM</td>
<td>51.8</td>
<td>62.0</td>
<td>17.5</td>
<td>57.9</td>
</tr>
<tr>
<td>TVQA+</td>
<td>72.0</td>
<td>63.3</td>
<td>23.7</td>
<td>70.3</td>
</tr>
<tr>
<td>Memory</td>
<td>54.7</td>
<td>53.7</td>
<td>13.9</td>
<td>50.0</td>
</tr>
<tr>
<td>IEP (V)</td>
<td>52.8</td>
<td>52.6</td>
<td>14.5</td>
<td>50.0</td>
</tr>
<tr>
<td>TbD-net (V)</td>
<td>79.5</td>
<td>61.6</td>
<td>3.8</td>
<td>50.3</td>
</tr>
<tr>
<td>MAC (V)</td>
<td>85.6</td>
<td>59.5</td>
<td>12.5</td>
<td>51.0</td>
</tr>
<tr>
<td>MAC (V+)</td>
<td>86.4</td>
<td>70.5</td>
<td>22.3</td>
<td>59.7</td>
</tr>
</tbody>
</table>

Hakan Sarp Aydemir
To do well, the model needs to
- Accurately recognize objects and events in the video
- Understand the compositional logic behind questions

Strong perception and logic operations on both visual and language input needed
• Results on descriptive questions
  ➢ Power of models that combine visuals and language

• Causal reasoning tasks (explanatory, predictive, counterfactual) need more than perception
• Object-centric representations essential for causal reasoning tasks
  ➢ Improvement of performance of MAC (V+) over MAC (V)
  ➢ Performance of TVQA+ on predictive questions

• Baseline models don’t have needed components
  ➢ Explicitly model dynamics of objects
  ➢ Explicitly model causal relations between events
  ➢ Struggle on tasks with unobserved scenes and counterfactual tasks

• Promising direction for causal tasks is combining
  ➢ Object-centric representations
  ➢ Dynamics modeling
Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion
• Unify
  - Object-centric representations
  - Dynamics modeling

• Challenges
  - Disjoint model components should operate on common set of representations of Video, question, dynamics, causal relations
  - Representation aware of compositional relations between objects and events

• Framework that joins these components via symbolic representation: NS-DR
**Model**

**I. Video Frame Parser**

- Video

**II. Dynamics Predictor**

- Mask R-CNN

**III. Question Parser**

- LSTM Encoder
  - LSTM
  - LSTM
  - LSTM

**IV. Program Executor**

- IV. Program Executor

**Question**

*What shape is the second object to collide with the gray object?*

**Answer**

*Cube*
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Descriptive</th>
<th>Explanatory</th>
<th>Predictive</th>
<th>Counterfactual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>per opt.</td>
<td>per ques.</td>
<td>per opt.</td>
</tr>
<tr>
<td>NS-DR</td>
<td>88.1</td>
<td>87.6</td>
<td>79.6</td>
<td>82.9</td>
</tr>
<tr>
<td>NS-DR (NE)</td>
<td>85.8</td>
<td>85.9</td>
<td>74.3</td>
<td>75.4</td>
</tr>
</tbody>
</table>

### Charts

- **Descriptive**: Test accuracy (%) vs. # Programs
- **Explanatory**: Test accuracy (%) vs. # Programs
- **Predictive**: Test accuracy (%) vs. # Programs
- **Counterfactual**: Test accuracy (%) vs. # Programs
Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion
• Systematic study of temporal and causal reasoning in videos
  ➢ Benchmark tasks introduced to better facilitate the research
  ➢ CLEVRER dataset and NS-DR model

• Capture true intelligence beyond pattern recognition
Thank You!
Backup Slides
NS-DR

- Different baseline models evaluated

- NS-DR combines components that are required for causal reasoning tasks

- Symbolic representations
Objects have 3 intrinsic attributes
- **Shapes:** Cube, Sphere, Cylinder
- **Materials:** Metal, Rubber
- **Colors:** Gray, Red, Blue, Green, Brown, Cyan, Purple, Yellow
3 Types of events
- Enter
- Exit
- Collision

Fixed number of participants for each event
- 2 for collision
- 1 for enter and exit
• Abstract representation of the video formed

• Ground truth annotations + object motion traces => model diagnostics
Videos - Video generation

• Generated from simulated motion traces
  ➢ Object’s position and time at each time step

• Each simulation lasts 7 seconds
  ➢ Motion traces first downsampled to fit 25 frames per second
  ➢ First 5 seconds sent to Blender
  ➢ Remaining 2 seconds held-out for predictive tasks
• Event can be caused by 2 ways
  ➢ By an object if the event is the first one participated by the object
  ➢ By another event if the cause event happened right before the outcome event on the same object

• Example: A sphere collides with a cube then a cylinder.
  ➢ First collision (Sphere + cube) and the cube cause the second collision

• Randomly initialized objects + new objects with overlapping motion traces => complex causal structure
• Video’s diynamical content and temporal relation

• Open-ended questions, can be answered by a single word

• Sub-types: Count, exist, color, etc.
Questions - Explanatory

• Whether object or event responsible for the event
  ➢ Event A is ancestor of Event B in the causal graph => A responsible for B
  ➢ Object B participates in Event A or any other event responsible for A => O responsible for A

• Multiple choice with at most 4 options

• Model selects all options that match
  ➢ Correct and wrong options sampled equally for balance and bias minimization
Questions - Predictive

• Predict possible occurrences of future events

• Multiple choice

• Sparse post-video events
  ➢ 2 options provided to reduce bias
Example: Which event will happen next?

a) The cube collides with the red object
b) The cyan cylinder collides with the red object

A: a)
• Outcome of the video under hypothetical conditions.
• At most 4 options
• Correct and incorrect option numbers balanced
Questions - Counterfactual

• Example: Without the gray object, which event will not happen?
  a) The cyan cylinder collides with the sphere
  b) The red object and the sphere collide

A: a), b)
Baseline Evaluation on CLEVRER

• Evaluate performances of different baseline models
  ➢ Multi-class classification for descriptive questions
  ➢ Binary classification for multiple choice questions
• Different performances => Powerful assessment to models’ strength and limitations on various domains

• Models trained until convergence
  ➢ Tuned on validation set
  ➢ Evaluated on test set
Results - Descriptive reasoning

• LSTM relies only on question input => performs poorly
  ➢ Suggests CLEVRER has very small bias on questions

• TVQA+ achieves better performance
  ➢ Video QA model
  ➢ Compositionality in question logic and visual context handling limited => doesn’t thrive

• TbD-net and MAC achieve better performance
  ➢ Designed for compositional reasoning
  ➢ Neural program execution
  ➢ Joint attention mechanism
• Poor performance of most baseline models

• Models that do well on descriptive questions (MAC (V), TbD-net (V)) perform poorly

• Performance gain on models with object-aware representations
  ➢ TVQA+ achieves high accuracy on predictive questions
  ➢ MAC (V+) better than MAC (V) for all tasks