CSC494 Student-Initiated Project Proposal Natasha Sharan

Title: Improving Occlusion-Aware Object Representations in Unsupervised Video Understanding via Scene Graphs and Graph Neural Networks

Motivation: Supervised models for object tracking in videos are highly effective, but require large amounts of annotated data. In contrast, unsupervised approaches eliminate the need for labeled datasets, however they struggle with persistent object identification, especially under occlusion. Objects may be merged or misidentified once they disappear and reappear. This project investigates whether incorporating 3D spatial information can improve robustness in object tracking under occlusion.

Objective: To investigate whether representing a video as an evolving 3D scene graph, and processing it with a Graph Neural Network (GNN), can improve object tracking performance in the presence of occlusion in an unsupervised setting.

Hypothesis: By leveraging both appearance features and structural relationships across time, a GNN can learn spatial-temporal embeddings that preserve object identity across occlusions, leading to improved tracking with unsupervised methods.

Proposed Architecture:
The architecture will be refined and adjusted to best address the problem of occlusions.

1. Unsupervised Depth-Aware Object Detection Model
   - A CNN-based model which can detect keypoints from each RGB frame and form object clusters
   - Incorporates depth estimation to position objects in 3D space.
   - Outputs a set of detected objects with estimated 3D positions and visual feature vectors.

2. Scene Graph Construction
   - Each detected object becomes a node in a frame-level scene graph.
   - Edges have embeddings which represent spatial relations between objects.

3. Unsupervised Graph Neural Network (GNN)
   - Processes each scene graph throughout the video.
   - Learns spatial-temporal embeddings of detected objects.
   - Tracks object identity and position across frames, including through occlusions.
   - Can refine object representations based on context from other objects and prior frames.

Training:
The training is unsupervised, and will focus on encouraging embeddings of the same object across adjacent frames to remain similar. It will penalize unexplained abrupt shifts in predicted object position. Determining suitable loss functions for this architecture will be explored in this project as well.

Methodology:

For every frame of an input RGB video:

1. The unsupervised depth-aware detection model identifies objects in the frame.
2. A scene graph is constructed with the detected objects for the frame using 3D positions and visual features.
3. The GNN processes the scene graph and learns/updates the spatial representation of the objects over time.
4. The GNN then outputs the predicted object positions and identities in each frame.

Evaluation:

Since this is an unsupervised setting, the evaluation will involve a combination of quantitative and qualitative metrics.

- Quantitative: (On a labeled benchmark dataset such as CLEVRER),
  - Occlusion recovery rate: accuracy of re-identifying objects after occlusion.
  - 3D localization error: distance between predicted and true object positions.
- Qualitative: overlay tracked object identities across frames to visually demonstrate model performance during occlusion events.

Implementation Details:

- Language: Python
- Tools: PyTorch, Google Colab (for GPU acceleration)
- Format: Combination of .py and .ipynb

Milestones:

1. Prepare dataset for training and testing.
2. Implement object detection and scene graph construction.
3. Develop GNN architecture and self supervised training loop.
4. Train and evaluate the model.
5. Document and visualize results in final report.

Sources: (used while creating proposal)

- https://medium.com/data-science/ug-vod-the-ultimate-guide-to-video-object-detection-816a76073aef#3ddf
- https://medium.com/stanford-cs224w/scene-graph-generation-compression-and-classification-on-action-genome-dataset-9f692a1d5394
- https://viso.ai/deep-learning/object-detection/
- https://viso.ai/deep-learning/image-segmentation-using-deep-learning/