**570 Project Ideas and Progress**

**Ideas**

* Use MiT-b2 encoder backbone for local/global
* Slot Generator
  + Compress channels of embeddings from target image through 1x1 conv layer
  + Apply pixel wise softmax operation to create masks
  + Apply global weighted average pooling between generated masks and target image feature embedded in local extractor

**Progress**

* Created a datasetloader for downloading datasets from the internet (DAVIS-16, DUTS, FBMS)

**To Do**

* Siamese Encoder from F2Net
* Local Feature Extractor
* Global Feature Extractor

References

* Guided Slot Attention for VOS: <https://arxiv.org/pdf/2303.08314>
* F2Net: <https://arxiv.org/pdf/2012.02534>
* Slot Attention: <https://arxiv.org/pdf/2006.15055>

Paper Topics

2-3 pages literature review

2-3 pages implementation and proposal

Title: Investigating Real Time Computer Vision Techniques for Various Optimizations

Abstract

Computer vision is a rapidly evolving field within artificial intelligence that focuses on enabling machines to interpret and understand visual data from their surroundings. By mimicking human visual perception, computer vision systems aim to extract meaningful information from images or video, such as object recognition, tracking, and scene understanding. Recent advancements in deep learning and neural networks have significantly improved the accuracy and efficiency of these systems, allowing for applications across various domains, including autonomous vehicles, medical imaging, and surveillance. Despite these successes, there are many optimizations to be made in the realm of object detection, segmentation, and tracking, as well as gaze following. This paper explores the current state of the art methodology in multi-person gaze following, unsupervised video object segmentation, and low occupancy flow fields.

Background/Overview

The application of real-time computer vision models has many speed and accuracy requirements in order to be effective in modern day usage. In addition, there exists an abundance of applications for accurate real-time outputs from a computer vision system. One of which is multi-person gaze detection and following. Gaze Following is defined as the prediction of the pixel-wise 2D location where a person in the image is looking. Prior efforts in this direction have focused primarily on CNN-based architectures to perform the task. In the paper discussed, a novel transformer-based architecture for 2D gaze prediction is introduced.

Additionally, another basic ability of machine learning architecture is video object segmentation (VOS). This entails identifying and segmenting desired objects in a video sequence. There are two forms of basic VOS: semi-supervised and unsupervised. Semi-supervised VOS entails providing an initial segmentation mask in the first frame and its objective is to track the masked object throughout the video sequence. The second and more difficult methodology is unsupervised segmentation which aims to mask and track the most prominent object in a scene. In both of these approaches, however, one limitation to high accuracy real time segmentation is the existence of complex background and foreground elements which make tracking extremely difficult. In the paper discussed, a slot attention mechanism is used to distinctly separate the foreground and background in an unsupervised environment to produce state of the art object segmentation.

Lastly, one growing need for accurate and high-speed computer vision models is for self-driving vehicles. Existing works either perform object detection followed by trajectory forecasting of the detected objects, or use an object free method and predict dense occupancy and flow grids for the whole scene. Trajectory forecasting poses a safety concern as the number of detections needs to be kept low for efficiency reasons, sacrificing object recall, and producing flow grids for the whole scene is computationally expensive due to the high-dimensionality of the output grid, and suffers from the limited receptive field inherent to fully convolutional networks. In the paper discussed an architecture to take the successes and faults of both methods is introduced by avoiding unnecessary computation and overcomes limited receptive fields through a global attention mechanism.

Problem Definition

The problem addressed in this paper is the need for more computationally efficient and reliable computer vision models for various applications. As the use of artificial intelligence for automated vision related tasks increases, scene analysis, object tracking, and object segmentation must improve to ensure proper functionality and data collection. The three external research papers discussed in this paper introduce several architecture changes to already existing computer vision problems resulting in state of the art accuracy and/or speed. In specific, this project seeks to address these novel architecture changes:

1. **ViT transformers combined with an attention framework to computationally improve multi-person gaze following**
2. **Guided background and foreground slot attention along with k-means to improve video object segmentation**
3. **Deformable convolutions and cross attention to predict occupancy probability and reverse flow for self-driving vehicles**

These papers seek to address these problems with computer vision:

1. **How to improve background and foreground separation for object masking**
2. **How to improve the computational efficiency of gaze token processing for scenes requiring gaze following**
3. **How to improve computation speed while also widening the receptive field for occupancy area prediction**

Summary and critique of selected papers

**Sharingan: A Transformer Architecture for Multi-Person Gaze Following:** Sharingan was created as a remedy to multi-person gaze following. Previously, many gaze following architectures computed the gaze of one person at a time, requiring recalculation for every person in a frame which decreased computational efficiency for complex scenes with many people. This novel architecture allows a transformer to use scene tokens and person specific gaze tokens to interact within an attention framework to predict gaze heatmaps. Since multiple tokens can be made for multiple people in a scene, this improves training efficiency and also provides improvements in inference. Additionally, the proposed decoder for the architecture is able to output two different variants for gaze detection. One is a heatmap variant which follows the traditional task of predicting a heatmap where the maximum of the map indicates the 2D gaze point. The second is a 2D point regression which predicts the gaze point by progressively regressing those the coordinates of the gaze point using a convolution-based fusion model.

Results for this architecture prove state of the art performance in both the GazeFollow and VideoAttentionTarget datasets. With respect to the GazeFollow dataset, the heatmap output variant achieves an average distance of 0.108 and minimum distance of 0.054 (state of the art), and the 2D regression point achieves 0.104 (state of the art) average and 0.064 minimum on single person gaze prediction and 0.106 (state of the art) average and 0.066 (state of the art) minimum on 6 person gaze prediction. Additionally, for the VideoAttentionTarget dataset Sharingan achieves state of the art performance with 2D point regression on the 6 person gaze following with a distance of 0.118.

Although achieving this with the GazeFollow dataset, this model is unable to achieve state of the art across all fields in the VideoAttentionTarget dataset. This leads to the conclusion that this model holds the best accuracy for static images but lacks applicability in situations requiring video input. This is shown by the testing results where the Average Precision (AP) and Area Under the Curve (AUC) scores for the model achieved below average results when compared to other top performing models on the VideoAttentionTarget dataset.

**Guided Slot Attention for Unsupervised Video Object Segmentation:** This paper introduces more accurate object mask creation for unsupervised VOS. In order to improve masking, background and foreground slots were initialized with query guidance to increase separation. Additionally, the proposed feature aggregation transformer uses target features from the local extractor and reference features from the global extractor then performs slot attention using the aggregated features and guided slots and adjusts the slots using KNN. Finally, a slot decoder is used to generate the final mask for the frame. By purposefully separating the foreground and background, accurate masking can be performed in scenes with complex foreground and background objects and textures.

The main contribution of this paper is the guided slot generator and slot attention framework that allows for state of the art foreground and background separation. The slot generator takes the embeddings for the target image feature and pushes it through a 1x1 convolutional layer then applies a softmax operation on the output. Afterwards this output undergoes a global weighted average pooling with the target feature embedded in the local extractor which yields the guided slot blocks that assist in the final mask generation. Utilizing the slot attention module introduced in the Object-Centric Learning with Slot Attention paper by Google Research, slot attention can extract object-centric representations that enable generalization for unseen compositions which make it ideal for unsupervised learning.

The quantitative performance evaluation of this model utilized the DAVIS-16 and FBMS datasets. For the DAVIS -16 dataset three values were compared with other top performing models: JM​ (Jaccard Index, or Region Similarity - "IM"), FM​ (Contour-based F-measure - "FM"), GM​ (Generalized Metric - "GM"). Using the MiT-b2 encoder model this architecture achieves state of the art performance across all quantitative metrics as well as the FM metric on the FBMS dataset.

Although this model achieves state of the art performance under a specific backbone, this is at the cost of frames per second (FPS). The metrics achieved yielded 4.1 FPS, sacrificing accuracy for speed which is not feasible in real time applications. Under the two tests yielding 41.5 fps and 38.2 fps for the ResNet101 and MiT-b2 encoding backbones respectively, the performance was still above average but did not beat the state of the art models in the performance metrics. Conversely, despite not beating other models in metrics in the higher fps testing, the trade in speed causes it to have the best accuracy to fps ratio out of all other high fps models. Another criticism of the paper is that the usage of KNN filtering doesn’t guarantee convergence and doesn’t inherently optimize a function. There are other unsupervised learning algorithms available that may provide a better approach that ensures convergence such as expectation maximization or using principal component analysis for filtering.

**Implicit occupancy flow fields for perception and prediction in self-driving:** This paper introduces IMPLITO which utilizes an implicit representation to predict occupancy probability and flow over time from sensor data and HD maps. They leverage deformable convolutions and cross attention in order to overcome shortcomings of previous self driving methods such as the limited receptive field of fully convolutional explicit architectures. The model takes in a multi sweep Lidar and a Map Raster and puts it through an encoder then through an implicit decoder to predict both occupancy and reverse flow.

The proposed architecture takes in a Voxelized LiDAR, raster map, and a mini-batch containing spatio-temporal query points. This then predicts the occupancy and flow for the mini-batch in parallel. This is performed by using a convolutional encoder that computes scene features and an implicit encoder that outputs occupancy and flow estimates by using a residual block and cross attention architecture.

IMPLITO achieves state of the art performance across Mean Average Precision (mAP), Soft Intersection over Union (Soft-IoU), Expected Calibration Error (ECE), and End Point Error (EPE) metrics when compared with both object-based and object-free models on the AV2 and HwySim datasets. Additionally, IMPLITO outperforms MP3 and OccFlow in inference time with varying number of query points from 0-200,000.

One issue is that with increasing query points, the MP3 and OccFlow model maintain constant inference time. Although IMPLITO outperforms MP3 and OccFlow and has a constant inference time at <20,000 query points. Introducing more query points causes the model to increase in inference time linearly and thus will eventually have an inferior inference time given a very large amount of query points.

Proposal for Project

**Experimental Setup:**

**Preliminary Implementation Efforts:** Initial implementation for the project is based off the architecture for *Guided Slot Attention for Unsupervised Video Object Segmentation Paper*. The MiT-b2 encoder backbone was used to process a raw rgb input image and output encoded features from the given image. The hidden sizes for the features consist of [64, 128, 320, 512]. After this the input image was put through 3 different pathways for processing. The first path was the Slot Generator, which passed the encoded input image features through a 1x1 convolution then performed the pixel-wise softmax shown below:

After this, global weighted average pooling was performed with the target image feature embedded in the local extractor using the methodology below by taking the input feature tensor and multiplying it element-wise and normalizing with the target image feature:

Secondly, the local extractor path also performs a 1x1 convolution on the target image feature then performs k-means on the features to develop sets of cluster masks. Then global weighted average pooling is performed on the embedded local feature and the cluster masks to create vectors for each cluster representing the prominence in each channel.

Lastly, the global extractor path first has the encoded features from the MiT-b2 architecture pass through a feature pyramid network similar to the one laid out in the reference paper *“Feature Pyramid Networks for Object Detection”* [7]. This consecutively dilates up the image size in order to prioritize broader features in the image. After performing this the encoded features undergo the same 1x1 convolution before having a channel-wise softmax operation performed then doing a global weighted average pooling similar to the ones in the previous paths.

**Methodology:** The objective for the semester project will be a full reimplementation of the ideas and architecture laid out in the *Guided Slot Attention for Unsupervised Video Object Segmentation* paper. This includes using an encoder backbone to extract local and global features from the images and reference images. Passing the local embedding into a slot generator to generate background and foreground slots. Using a feature aggregation transformer to combine local and global extracted features. Then using the guided slots and aggregated features inside of a slot attention module to generate refined slots which when paired with the aggregated features can be decoded into a image mask.

One change to be made to the original architecture that will be explored in this project is attempting different unsupervised learning algorithms as a means of refinement in the slot attention module instead of K-Nearest-Neighbors in order to compare the differences in accuracy between the several methods to be implemented. Expectation Maximization and Principle Component Analysis are examples of algorithms that can be used instead or in tandem with the current architecture in order to yield improvements in quantitative results.

**Feasibility and Relevance**: The main coding challenge of this implementation will be the feature aggregation transformer architecture for generating useful features from the global and local features and the slot generation and attention module. The KNN filtering, local/global extractors, and decoder are standard enough that online resources will help with architecture reimplementation.

For innovations in the computer vision related field, it is highly important that accurate and fast masking models are implemented. Areas such as autonomous vehicles, augmented reality, security surveillance, robotics, and medical imaging benefit from this project. The benefit of having this architecture use unsupervised learning is better generalization and the discovery of hidden features. This allows the proposed project model to be applicable to a wider range of situations and improves scalability for future endeavors in the computer vision field.

**Conclusion:** This project proposes an implementation of a guided slot attention mechanism for video object segmentation in an unsupervised learning context. Additionally, it aims to explore various unsupervised learning algorithms in order to integrate them into the existing slot attention module laid out in the original paper: *Guided Slot Attention for Unsupervised Video Object Segmentation.* By introducing object-centric slot attention through guided slots, background and foreground objects can be recognized and separated allowing for a more accurate masking of the most prominent object in a video object sequence. Especially for crowded and complex foreground and background scenes, performing slot attention on slots in conjunction with extracted features from the given scene can yield accurate object segmentation that is beneficial for ongoing research and technological application in object detection and tracking.

References

Lee, M., Cho, S., Lee, D., Park, C., Lee, J., & Lee, S. (2023). Guided Slot Attention for Unsupervised Video Object Segmentation. *ArXiv*. https://arxiv.org/abs/2303.08314

Tafasca, S., Gupta, A., & Odobez, J. (2023). Sharingan: A Transformer-based Architecture for Gaze Following. *ArXiv*. https://arxiv.org/abs/2310.00816

Agro, B., Sykora, Q., Casas, S., & Urtasun, R. (2023). Implicit Occupancy Flow Fields for Perception and Prediction in Self-Driving. *ArXiv*. https://arxiv.org/abs/2308.01471

Liu, D., Yu, D., Wang, C., & Zhou, P. (2020). F2Net: Learning to Focus on the Foreground for Unsupervised Video Object Segmentation. *ArXiv*. https://arxiv.org/abs/2012.02534

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-Centric Learning with Slot Attention. *ArXiv*. <https://arxiv.org/abs/2006.15055>

Qiu, S. (2018). Global Weighted Average Pooling Bridges Pixel-level Localization and Image-level Classification. *ArXiv*. https://arxiv.org/abs/1809.08264