Research statement

Srihari Vemuru

Abstract: My research interests span computer vision and Artificial General Intelligence. My current research mainly focuses on computer vision. More specifically, I am working on deep learning methods for object detection, object tracking, and video feature analysis. The applications range from autonomous vehicles to personality classification of psychiatric patients. I plan on targeting these broad topics in my future research:

1. Using attention concepts of NLP in detecting, modeling, and tracking objects in visible and infrared videos to understand their behaviors and interactions.

2. Using state-of-the-art computer vision technologies to detect and understand visual cues exhibited by people under different contexts such as driving, manufacturing, etc.

3. Expanding CNNs to non-Euclidean data sets, such as computer graphics objects or computer-aided design (CAD) parts.

Research projects

Understanding Social Behavior in Dyadic and Small Group Interactions

Currently, I am working on building a behavior analysis model for 2021 Understanding Social Behavior in Dyadic and Small Group Interactions Challenge at ICCV.

The task is to automatically recognize personality of single individuals during a dyadic interaction, from two individual views. Personality of a person is categorized into the Big-Five personality traits. These are the five broad personality traits described by the *Big five* theory: openness, conscientiousness, extroversion, agreeableness, and neuroticism (OCEAN). Context information is expected to be exploited to solve the problem. Audio-visual data associated with this track, as well as the self-reported Big-Five personality labels are available. Utterance level transcriptions are provided so that verbal communication can also be exploited. The proposed baseline presents a transformer-based context-aware model to regress self-reported personality traits of a target person. Feature extraction is performed by a R(2+1)D network for the visual chunks and VGGish for audio. The visual features from the R(2+1)D's 3rd residual block are concatenated to spatiotemporal encodings (STE). The VGGish's audio features and handcrafted metadata features are incorporated to visual context/query features and the result transformed to the set of Query, Keys, and Values as input to a transformer network. The latter consists of layers equipped with Local and Extended Context Transformer units. Such units implement multiheaded attention and provide their updated queries, which are combined and fed to the next layer. Finally, the output of the final transformer layer is fed to a fully-connected (FC) layer to regress per-chunk OCEAN scores.

One of the important traits of a person to determine their personality is their eyes. More specifically, the blink characteristics. I am planning on using dlib's pre-trained face detector based on a modification to the standard Histogram of Oriented Gradients + Linear SVM method for object detection. Using Blink Retrieval Algorithm proposed by [1], I'll be able to retrieve all the *blink*'s of a person in a video, where *blink* is a tuple consisting of the amplitude, velocity and the frequency of the blink.

Handling Complex Queries Using Query Trees

Contemporary search engines fail to handle even slightly complex queries. Search engines process queries by identifying keywords and searching against them in knowledge bases or indexed web pages. The results are, therefore, dependent on the keywords and how well the search engine handles them. In this project, I, along with a team member, proposed a three-step approach called parsing, tree generation, and querying (PTGQ) for effective searching of larger and more expressive queries of potentially unbounded complexity. PTGQ parses a complex query and constructs a query tree where each node represents a simple query. It then processes the complex query by recursively querying a back-end search engine, going over the corresponding query tree in postorder. Using PTGQ makes sure that the search engine always handles a simpler query containing very few keywords. Results demonstrated that PTGQ can handle queries of much higher complexity than standalone search engines.

PTGQ is a three-step approach for processing complex queries. PTGQ breaks the search query into smaller pieces at relevant positions, orders these into its corresponding query tree, and processes it recursively with a search engine. The three steps involved in PTGQ are:

1. Dependency Parsing: This generates a dependency tree from the search query that describes the structure of sentences and the relationships among the words present as a tree.

2. Query Tree Construction: This iterates over the dependency tree to identify keywords using syntax analyzers, generates corresponding simple queries, and constructs a query tree with them. Each node in the query tree corresponds to a simple query. The query tree obtained at the end of this step is answerable in postorder.

3. Progressive Querying: This recurses using a search engine over the query tree in postorder to answer each node present in the tree. Each node in the query tree waits for answers from its children, adds them to its corresponding simple query, and searches this newly formed query in a search engine. The result of the root node is the answer to the search query.

Analysing road safety using object tracking models

In this project, I analyzed the safety standards of Autonomous public transport vehicles in Montreal. I worked under Prof. Nicloas Saunier of Polytechnique University, Montreal on improving his traffic safety repository called TrafficIntelligence. TrafficIntelligence uses object detectors and trackers to get bounding boxes on vehicles in a video, and then encodes it into a state containing the position, velocity and the type of vehicle. Then, it uses various algorithms to analyze the average paths taken by the vehicles and find out if there are any possibilities of collisions over a period of time. But, TrafficIntelligence uses older object detectors and trackers. It uses Feature based tracker [4] to get the bounding boxes on the vehicles in a video. I replaced it with DeepSORT with a YOLO backbone. I trained YOLO and DeepSORT's image feature extractor with UA-DETRAC dataset. Then, I used a new object tracking metric called HOTA [2], to compare it with the baseline model. HOTA (Higher Order Tracking Accuracy) is a novel metric for evaluating multi-object tracking (MOT) performance. It is designed to overcome many of the limitations of previous metrics such as MOTA, IDF1 and Track mAP. HOTA can be thought of as a combination of three IoU scores. It divides the task of evaluating tracking into three subtasks (detection, association and localization), and calculates a score for each using an IoU (intersection over union) formulation (also known as a Jaccard Index). It then combines these three IoU scores for each subtask into the final HOTA score. In the project, while Feature based tracker gave a HOTA score of only a 35, DeepSORT gave 49. It was a significant jump in tracking accuracy.

Trying novel object detection methods on traffic images

I did a project on object detection during my internship at LightMetrics, a road safety and video analytics company. I did two tasks during my internship.

First, I experimented with a contemporary object detector with a novel heatmap based concept to detect bounding boxes in an image called CenterNet [5]. CenterNet has a feature extractor, like YOLO, but in the final layers it outputs a scaled down version of the original image, where every cell has the bounding box information if an object's center falls in it. I trained CenterNet on their traffic dataset. Although it gave the same mAP values as YOLO, CenterNet was much faster. In the second part, I used the feature extractor of YOLO, darknet53 [3], as the feature extractor for CenterNet. There was a significant increase in mAP compared to YOLO and the traditional CenterNet.

Other projects

I have done multiple projects in computer vision and NLP, the two main branches of AI/ML. One of the projects I did in computer vision was inferring the engagement levels of students in a classroom using facial recognition. For this, I used a neural net training technique called Unsupervised Domain Adaptation. Working in a team for another project, I wrote and trained a football tracking model using a traditional Kalman filter method. In NLP, for the Smart India Hackathon 2020, my team created a medical chatbot to answer questions posed by users on the Covid pandemic. The chatbot could also locate the nearest medical facility that could address the symptoms mentioned by a user.

Future trajectory

Computer vision applies to numerous real-world applications and contributes to Artificial general intelligence. I am interested in analyzing and enhancing existing deep learning models in computer vision, such as object detection, tracking, facial recognition, etc. My long-term goal is to contribute to AGI through vision research. One of my prime focuses of AGI-related applications is virtual social media. These are some of my short-term and long-term research goals:

1. Attention in computer vision: As one of the important features of the human visual system, visual attention mechanism is essential in image generation, scene classification, target detection, and tracking when applied in the field of computer vision. My goal is to use attention models in computer vision.

2. Body pose and facial expression analysis for autonomous vehicles: People produce a variety of visual cues both consciously and unconsciously. Car companies these days are keen to integrate technologies to detect such visual cues into their advanced driver assistance systems (ADAS) to enhance safety and convenience. Accurate sensing and detection of such visual cues can become critical in many applications. My goal is to understand visual cues exhibited by people under different contexts such as driving, manufacturing, etc., by combining state-of-the-art computer vision technologies.

3. CNN's on non-Euclidean data sets: Convolutional neural networks have demonstrated unprecedented success in a variety of visual cognition tasks. But, there are a large number of problems that could not benefit much from such powerful CNNs due to the non-Euclidean nature of the data set, like graphical models, such as computer graphics objects or computer-aided design (CAD) parts. Hence, there is no canonical way of representing such data in a tensor format as expected by CNNs and, thus, the standard CNNs cannot analyze such data. I aim to develop mathematical foundations and scalable algorithms to expand CNNs to a variety of different geometric domains.

Conclusion

My goal is to do solid research in computer vision and Artificial General Intelligence. I like to do research with the actual foreseeable applications. I have research experience in wide facets of AI but predominantly in computer vision. I have done numerous projects in many sub-topics of computer vision, namely object detection, tracking, body pose, and facial recognition. Even though my research interests are more aligned towards computer vision, I would like to transition to the broader AI field in the long term.

References

[1] R. Ghoddoosian, M. Galib, and V. Athitsos. A realistic dataset and baseline temporal model for early drowsiness detection, 2019.

[2] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, Oct 2020.

[3] J. Redmon and A. Farhadi. Yolov3: An incremental improvement, 2018.

[4] N. Saunier and T. Sayed. A feature-based tracking algorithm for vehicles in intersections. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 59–59, 2006.

[5] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points, 2019.