Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Graph Neural Network (GNN) in Image and Video Understanding Using Deep Learning for Computer Vision Applications

Pradhyumna P, Shreya G P, Mohana
Electronics and Telecommunication Engineering, RV College of Engineering®
Bengaluru, Karnataka, India.

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532631

Abstract- **Graph neural networks (GNNs) is an information - processing system that uses message passing among graph nodes. In recent years, GNN variants including graph attention network (GAT), graph convolutional network (GCN), and graph recurrent network (GRN) have shown revolutionary performance in computer vision applications using deep learning and artificial intelligence. These neural network model extensions, collect information in the form of graphs. GNN may be divided into three groups based on the challenges it solves: link prediction, node classification, graph classification. Machines can differentiate and recognise objects in image and video using standard CNNs. Extensive amount of research work needs to be done before robots can have same visual intuition as humans. GNN architectures, on the other hand, may be used to solve various image categorization and video challenges. The number of GNN applications in computer vision not limited, continues to expand. Human-object interaction, actin understanding, image categorization from a few shots and many more. In this paper use of GNN in image and video understanding, design aspects, architecture, applications and implementation challenges towards computer vision is described. GNN is a strong tool for analysing graph data and is still a relatively active area that needs further researches attention to solve many computer vision applications.**

Keywords— Graph Neural Networks (GNNs), Convolutional Neural Network (CNN), Gated Adversarial Transformer (GAT), Atomic Visual Action(AVA), Human-Object Interactions (HOI), Graph Parsing Neural Network (GPNN).

I. INTRODUCTION

Graphs are one of the most useful data structures for various applications areas such as learning cell fingerprints, exploring traffic networks, prescribing companions to interpersonal interaction and modeling body systems. Non-Euclidean graphs need to be addressed by these activities which contain details of interpersonal relationships which can be mishandled by using deep traditional learning models. Nodes in graphs regularly contain valuable data that is ignored in profoundly moderated unregulated learning techniques. GNNs are proposed to consolidate the graph structure and feature details to pursue better presentations on graphs with integration and feature distribution. Because of its persuading execution and high interpretation, GNN has as of late become broadly utilized. GNNs can model the relationship of nodes of the graph and generate a numerical representation of it. GNNs are particularly essential since there is so much factual information that can be expressed also as graphs. Social networks, chemical compounds, maps, and transportation systems are just a few examples. Due to advancements in DL, in variety of image comprehension tasks, including image categorization, object recognition, as well as semantic segmentation. Deep learning algorithms' capacity to learn at many levels of abstraction from data accounts for this resounding success. Different levels of abstraction are required for different activities. This generalized image classification challenge helps determine which object classes are included in a given image (usually from a collection of predefined classifications). Image classification methodology based on GNN models is evolving, since GNN, which derive their motivation from CNN, are used in this domain. If given a large training dataset of labelled class, many of these models, including GNN, provide interesting results. Recent video understanding datasets such as AVA and Charades, on the other hand, have lagged behind in comparison. Understanding the interactions between actors, objects, and other context in a scene is one of the many reasons why video comprehension is very important. Furthermore, because these interactions aren't always visible in a single frame, reasoning over large time intervals is required. Because video has an additional temporal axis, it has a much higher dimensional signal than single images, and we believe that learning these unlabeled interactions directly from current datasets with huge convolutional networks is not possible. Effective video comprehension necessitates long-term reasoning about the links between objects, actors and their surroundings. Because graph-structured data is ubiquitous, it can be employed in a wide range of scenarios.

II. LITERATURE SURVEY

Anurag Arnab [1] describes how to make a proposal when supervision is not available, they suggest a message-passing GNN that can use explicit representations of objects and clearly represents these spatio-temporal interconnections. This strategy is demonstrated on two different challenges in video spatio-temporal action identification that require relational thinking. It also demonstrates how this method may more successfully model relationships between significant things in images, numerically and qualitatively. Natural video events are usually outcome of spatio-temporal engagements among actors and objects, and they frequently involve a large number of object types. Yubo Zhang [2] proposes a study that

argues that action detection is a difficult problem to solve since the models which must be trained are massive, and obtaining tagged data is costly. To overcome this constraint, they recommend incorporating domain knowledge into the model's structure to make optimization easier. Santiago Castro [3] presents the majority of research work on language-assisted video understanding has centered on two tasks: firstly, usage of multiple choice questions for video question answering, here models perform well because candidate solutions are easily available. Second, to capture a video which uses an open-ended assessment framework. They suggest fill in the blanks as just a video comprehension assessment framework that corrects previous assessment problems but more closely matches real-life circumstances in which many possibilities are unavailable. Using this associated text and video, the model must predict a concealed noun phrase inside this video description, which assesses the system's knowledge of the film. The dataset is built from the VATEX dataset, with blurred captions generated by stamping noun phrases in the English captioning in VATEX. To construct an instance, we select the first English caption that usually contains one noun phrase as recognized by spaCy1, then blank these kinds of nouns at random. As a result, we start only with the VATEX v1.1 training set, a randomized subset of size 1000 from the test dataset, respectively, to construct our training, validating, and testing data. To acquire additional right answers for each space in the verification and test sets, we used a crowd annotating technique. The key aim for gathering such additional annotations, as previously said, is to account for the diversity of words and to have several alternatives for each space. K. Sasabuchi [4] presents a learning-from-observation framework for extracting precise action sequences from a video of a human demonstration split and understood with vocal instructions. Splitting is based on minimum local points in hand velocity, which link human daily movements with object-centered facial contact transitions required for robot motion generation. They first established that hand velocity motion splitting is a reliable signal for partitioning daily tasks. Secondly, they generated a new motion description dataset with the goal of better understanding everyday human actions. Also provides the information of attention-based models, researchers developed Gated Adversarial Transformer (GAT). To increase the model's performance even further, they applied adversarial training methodologies. A regularization term was added to the loss function, giving the model adversarial robustness to both attention mappings and the final output space. Matthew Hutchinson [5] describes deep learning research in computer vision is a natural extension of video understanding. The application of artificial neural network (ANN) machine learning (ML) approaches has tremendously benefitted the field of image interpretation. They grouped deep learning model building blocks and state-of-the-art model families, and specified standard metrics for assessing models. They also listed datasets suitable as benchmarks and pre-training sources, discussed data preparation stages and techniques, and organized deep learning model building blocks and state-of-the-art model families. Ishan Dave [6] proposes a temporal contrastive learning framework outperforms state of the art outcomes in a variety of downstream video comprehension tasks, including action recognition, limited-label action classification, and action classification and nearest-neighbor video retrieval on several videos and datasets. On three different datasets, this paper gives comprehensive experimental evidence and achieves best-in-class results in a range of downstream video interpretation tasks. Beyond instance discrimination, the success of our methodology demonstrates the advantages of contrastive learning.

A. Gupta [7] proposes a simple idea which offers world features, a basic concept in which each feature at each layer has its own spatial transformation, and the feature map is only altered when needed. Results show that a network created with these World Features may be utilised to mimic eye movements including saccades, fixation, and smooth pursuit on pre-recorded video in a batch environment. It finds that numerous eye movements are achievable, allowing for a wide range of augmentations without sacrificing relative feature position. They utilised these concepts to supply the model with a transformation for each video in the experiments, but learnt transformations might also be employed. Yubo Zhang [8] proposes a notion that action detection is a difficult problem to solve based on the fact that the models are large to be trained and obtaining labelled data is costly. To overcome this issue, they suggest incorporating domain expertise into the model's structure to make optimization easier. This model surpasses existing best practices, proving the method's utility in modelling spatial correlation and reasoning about linkages. Most significantly, the performance of this model highlights the need of incorporating relational and temporal knowledge into design methods for detection of action. H. Huang [9] describes a new dynamic hidden graph module in videos for modelling complicated object-object interactions, with two instantiations: a visual graph for capturing appearance/motion changes among objects and a location graph to capture relative spatiotemporal position changes among objects. The suggested graph module may explicitly capture interactions among objects in streaming video contexts by evaluating object relations at the same time both in time domain and frequency domain, which distinguishes our work from prior techniques.

Y. Chen [10] presents a novel way of explanation that aggregates a collection of features globally across the coordinate space before moving to an interacting area whereby relationship thinking may be calculated successfully. J. Zhou [11] describes many learning activities including working with graph data, which offers a wealth of relational information between parts. The usage of a model that can learn from graph inputs is required for simulating the physical systems, establishing fingerprints, predicting proteins interface, and diagnosing illnesses. They did a thorough examination of graph neural networks. They divide GNN models into variants based on compute modules, graph kinds, and training kinds. They also describe a number of general frameworks and present a number of theoretical studies. F. Scarselli [12] proposed GNN model, for processing data it extends existing neural network approaches shown in this graph domain. The mentioned GNN model can handle most common graph topologies directly, including acyclic, cyclical, guided, and unguided graphs. The paradigm revolves upon information diffusion and relaxation processes. The generic framework, prior generative method for organized data processing, and approaches based on vague walk mod are all incorporated into the approach. K. Simonyan [13] proposes the 2 stream ConvNet design that includes both spatial and

temporal networks. First they show that despite minimal training samples, the results of a ConvNet taught on multi frame intensive optical flow might be great. Finally, they demonstrate how multitask learning may be utilised to improve the quantity of data collected from two separate action classification datasets.

R. Girdhar [14] introduces a method for recognizing and localizing living beings in video footage, use the Actions Converter model. They employ a transformer architecture for collecting characteristics from the spatiotemporal environment around the individual whose behaviors are being classified. They demonstrated that even the Actions Converter network can acquire spatiotemporal information from several other human Behaviour and items inside a film clip and use it to recognize and localize human activities. Siyuan Qi [15] describes the challenge of identifying and distinguishing human-object interactions (HOI) in photos and videos in this work. A Graph Parsing Neural Network (GPNN) is introduced, an end-to-end differentiable architecture that incorporates structural information. They test this model on various data sets such as V-COCO, HICO-DET and CAD-120, which on photos and videos, are all HOI recognition standards. This technique outperforms current methods, indicating that GPNN was adaptable to large datasets. And can be used in both spatial and temporal scenarios. Boncelet [16] describes testing of the proposed method's performance, where two picture understanding tasks were chosen: Emotion recognition at the group level and incident identification both task is extremely meaningful, and synthesizing multiple cues necessitates the interplay of numerous deep models. Understanding an image includes not just recognising the items in it, but also grasping their fundamental relationships and interconnections. GNNs may take advantage of such linkages during the feature learning and forecasting phases by spreading nodal messages through the network and aggregating the outputs. Danfei Xu [17] suggests the use of scene graphs, a graphical framework for a picture that is visually anchored, to formally model the objects and their interactions. They also propose a revolutionary end-to-end paradigm for creating structured scene representations from an input image. They developed a novel end to end model that solves the challenge of automatically constructing a visibly anchored virtual environment from an image by continuous passing of messages between the primal and dual sub-graphs along the topological structure of a scene graph. Yubo Zhang [18] describes Action detection as an example of a difficult problem: the models that must be trained are enormous, yet labelled data is difficult to get by. To overcome this constraint, they recommend incorporating domain knowledge into the model's structure to make optimization easier. The suggested methodology outperforms the by 5.5 percent mAP in the I3D base and 4.8 percent mAP on AVA dataset.

III. GRAPH NEURAL NETWORK MODEL AND ARCHITECTURE FOR IMAGE AND VIDEO UNDRSTANDING

Image categorization, a classic computer vision problem, where convolutional neural networks (CNN) being the most prominent one. GNNs, which get their motivation from CNN, have been used in this arena as well. Main goal is to improve zero-shot and few-shot learning task models performance. Zero shot learning (ZSL) is the process of training a model to recognize classes it has never seen before. ZSL image categorization is to control structural information. As a result, GNN appears to be highly tempting in this regard. The information needed to lead the ZSL work may be found in knowledge graphs. The type of information that each technique represents in the graph varies by knowledge. Graphs of such kind may well be built on commonalities between both the photos themselves or those of the objects recovered using object recognition in the photos. Semantic information from embeddings of the image class labels may also be included in the graphs. GNNs may then be used to enhance the ZSL picture classification-recognition process by applying them to this structured data. The process of creating a label for a video based on its frames is video classification, strong video level classification not only delivers correct frame labels, and also best represents the entire movie based on the characteristics and annotation of the individual frames. The process of creating a label for a video based on its frames is video classification. A strong video level classification not only delivers correct frame labels, and also best represents the entire movie based on the characteristics and annotation of the individual frames. Paper [1], propose a message passing graph GNN to spatio-temporal interactions and for object representations it uses explicit object if monitoring is available else implicit object shall be used. Their approach broadens earlier structured models for video comprehension, allowing us to investigate how varied graph representation and structure choices impact the model's performance. This shows how to apply a strategy to two separate tasks in videos that require related reasoning – on AVA and UCF101-24 it uses an action detection model of spatio temporal relation, on the recently released Action Genome dataset it uses video scene graph categorization on dataset. It also demonstrates that this strategy may more successfully model relationships between significant things in the picture, both numerically and qualitatively.

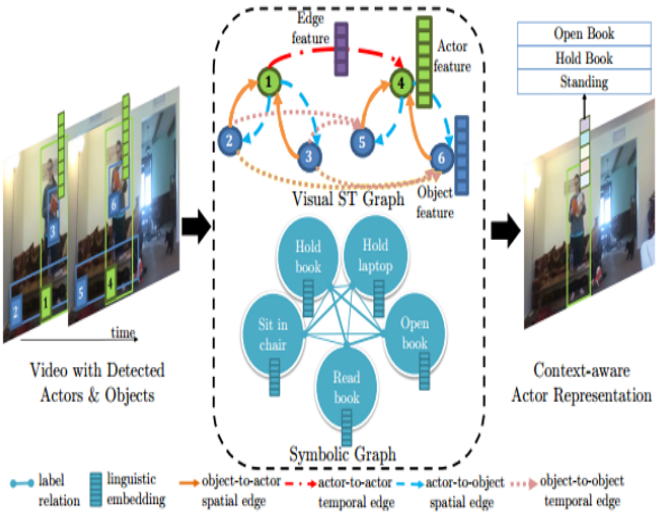

Fig.1. Graph-structured data representation [19]

Figure 1 shows the graph-structured data to understand the activity of video. The visual ST graph with unique edge types, actor-to-actor temporal, object-to-actor spatial etc., and unique node types is a heterogeneous graph with varied semantics and dimensions. First, it describes objects and actors visual spatio-temporal interactions. Second, co-occurrences, for example, are a common connection between labels. These signals can be represented visually and

symbolically in a mixed spatial temporal and symbolic attributed graphs. This hybrid graph to conduct supervised learning on recognized semantic elements, such as objects and actors, to create perspective models that may be used to tackle subsequent video processing tasks [19].

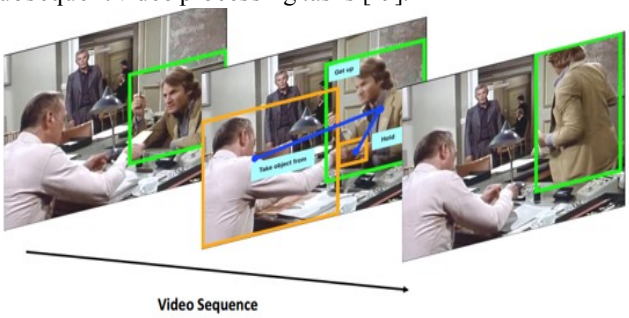

Fig.2. Action detection in video sequence [2]

Figure 2 shows the action detection in video sequence (AVA dataset).Person rising from their seat and collecting a letter from some other person seated beside a table. Information is genuinely necessary for recognising and localising this activity out of the 2359296 pixels inside the 36 frames of this snap, the actor's motion, location and interactions with other actors and the text are all important indications. The rest of the video's data, such as the wall color or the light on the table, is extraneous and should be ignored. Action region detection on such intuitive insights, it's vital to collect both deep temporal features and spatial interactions across actors and objects when detecting actions [2].

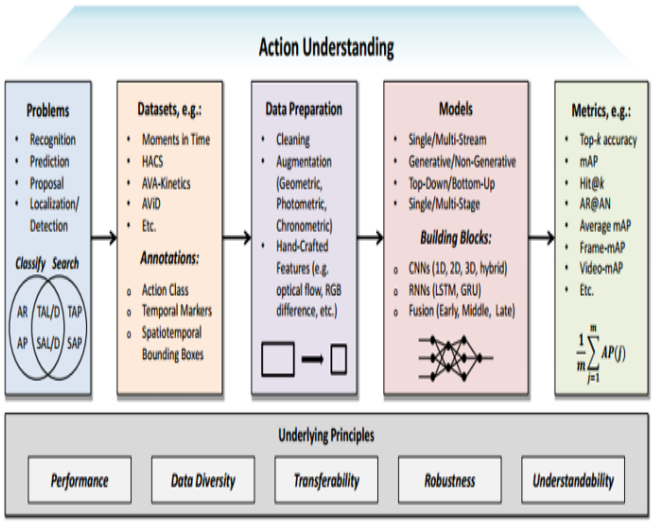

Fig.3. Action understanding [5]

Action issues, video action data, data processing approaches, deep learning models, and assessment measures all fall under the umbrella of action understanding as shown in figure 3 [5]. The ideas of computing performance, data variety, traceability, model resilience, and readability underpin these processes in computer vision and deep learning. Summary of action phases (dataset selection, problem formation, model construction, dataset preparation, and metrics basis evaluation) as well as core assumptions (data diversity, computational performance, robustness, transferability and understandability).

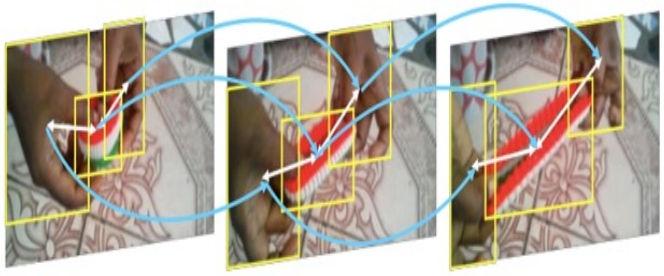

Fig.4. Object – to - object interaction [9]

Figure 4 shows object to object interaction, mainly two relations should be considered for recognising such interactions: first, interactions between various images into a single frame. Second, transitions of such interactions between different items and then the same item across successive frames [9]. The former is referred to as a spatial relationship, whereas the latter is referred to as a temporal relationship. Both are necessary for recognising multi-object operations. An efficient strategic recognition model will accurately and concurrently capture both relationships.

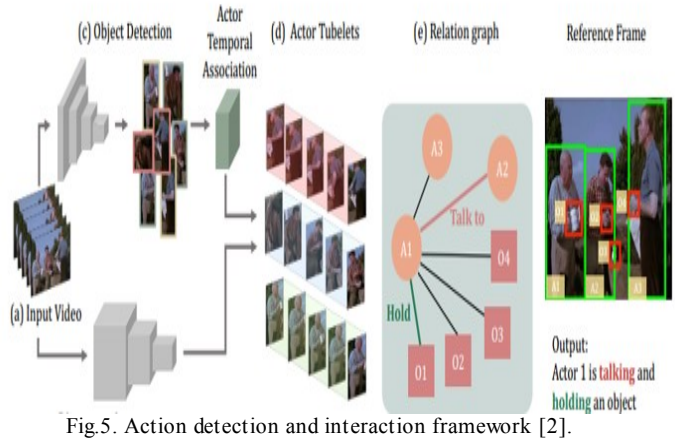

Fig.5. Action detection and interaction framework [2].

Figure 5 shows the action detection and interaction framework, it receives a video frames and runs through I3D network concurrently, each frame is subjected to object identification model, to generate person and object confidence scores. Tubelets are created by combining personal bounding boxes. Following, tubelets and object pieces (as nodes) are utilised to create an actor centric graph for each actor in the video [2].

3D convolution

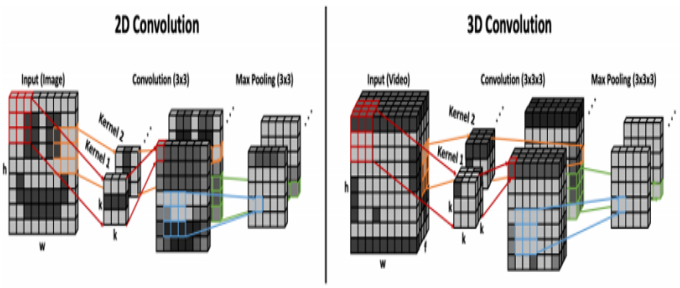

Fig.6. 2D and 3D convolution [5].

Backbone for many of the state of the art models are 1-Dimensional CNNs (C1D) uses 1D kernels, 2-Dimensional CNNs (C2D) uses 2D kernels, and 3-Dimensional CNNs (C3D) uses 3D kernels. C1D is generally used for

convolutions of embedded type features along the time dimension, whilst C2D and C3D are used to extract the feature vectors from single frames or layered frames. Figure 6 shows singular channel type samples of 2 Dimensional and 3 Dimensional convolutions. Accurate hyperspectral image classification has been a crucial yet difficult task. Convolution neural networks (CNNs) in two dimensions (2D) and three dimensions (3D) have been used to collect spectral or spatial info in multispectral photographs [5].

Graph is a form of structured arrangement of information that represents objects and their relationships. New research on graph analysis has aroused a lot of interest using ML because graphs have such a high expressive potential. In the graph field, GNNs are based on deep learning algorithms. GNN has lately gained popularity as a graph monitoring system due to its superior performance.

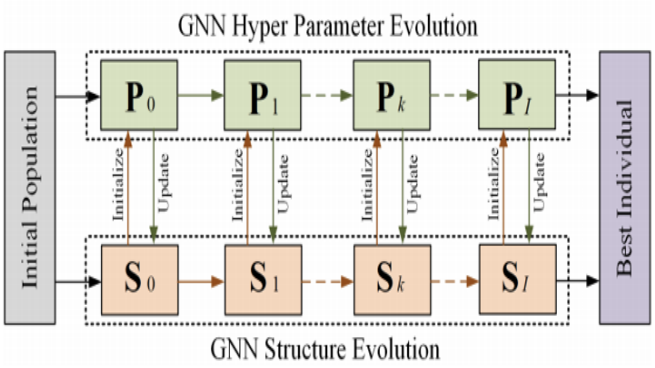

Fig.7. Genetic model for GNN architecture [11].

Figure 7. Shows the genetic model of GNN architecture. GNN structures initial population is initialized to (S0) first, with every individual being a multilayer Graph neural network where every layer is made up of components chooses randomly, such as the activation function, hidden embedding size and aggregator. With respect to (S0) the GNN parameters population (P0) is then initialized and it sets parameters accordingly which evolves as the best suit (e.g., learning rate and dropout rate). Following that, to optimize the graph neural network structures using the optimal parameter setting from P0, architecture is made from S0 to S1 [11]. After the first round of alternate development between structure and parameter, it creates a GNN architecture with ideal design and optimum parameter settings produced from SI and PI. Six encoded states of GNN architecture i.e., Hidden Dimension, Attention Head, Attention Function, Activation Function, Aggregation Function, and Skip Connection.

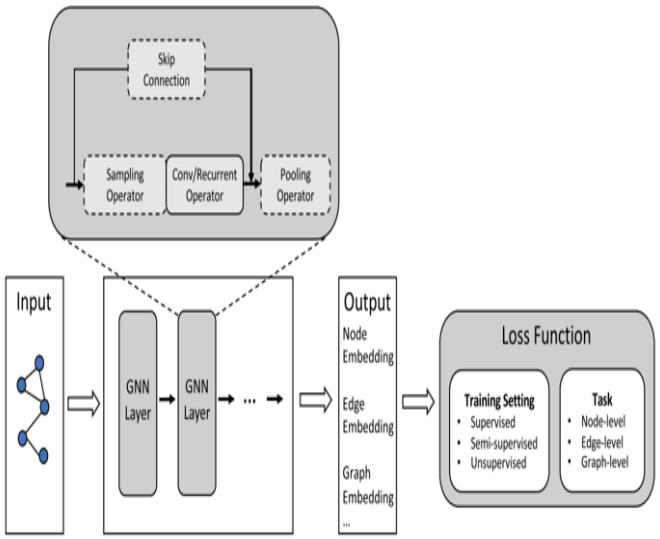

Fig.8. Design pipeline of GNN model [11].

Figure 8 shows the design pipeline for GNN model. Propagation Module is one of the most often utilized computational modules. Data is propagated across nodes using propagation module, which allows data aggregation to record both topology and feature info. Convolution and recurrent operators are typically employed in propagation modules which gather the information from neighbours, skip connection operation is typically employed to acquire detailed info from previous node representations. Sampling module frequently require to carry out graph inclination. Sampling and propagation modules are frequently combined. To extract the information, pooling modules are used from nodes when high-level subgraphs of graphs are needed [11]. A GNN model is usually developed by mixing these computing components. The recurrent operator, convolutional operator, skip connection and sampling module are used to spread info in individual layer, and to retrieve high level information pooling modules are added, as shown in figure 8. To get better representations, these layers are generally stacked. The architecture used here can simplify GNN models and outliers, such as NDCN, which mixes GNNs and ordinary differential equation systems.

Transformers in video understanding -Action Transformers, in which 3D CNN characteristics are pooled and delivered to identity exploit the Spatio-temporal data. By combining the patches generated frame at several time-steps, may use transformer to classify video. GAT-GAT models, significance of a frame based on the local and worldwide circumstances using an intra attention gate. This allows the network to comprehend the video at multiple levels of granularity.

IV. USAGE OF GNN IN VARIOUS DOMAINS

Graph neural network practical applications include traffic control [31] [32], human behavior detection [24] [25], ,adversarial attack prevention, recommender system, program verification, logical reasoning, molecular structure study, and social influence prediction Most GNN architectures can be classed as structural and non-structural depending information they process. Following are some intriguing applications from each categories: - In the graph-like structure of nano-scale molecules, the nodes are ions and edges are bonds connecting them according to GNN. In both cases, to learn about existing

molecular structures and to discover unique chemical structures GNNs can be used. This would have a great effect on the development of computer assisted medication. CNN are the most popular machine learning techniques which have remarkable answers to image categorization, a classic computer vision problem [20] [21] [22] [23]. GNNs can then be used to enhance the ZSL picture classification recognition task by applying them to this structured data. Text, like images, does not have obvious relationships.

V. APPLICATIONS OF GNN IN COMPUTER VISION

Some of the applications of GNN in computer vision applications are [26] [27] [28] [29] [30] [33] [34],

- Object Localization
- Human-Object Interactions
- Question Answering
- Object Detection
- Features Learning
- Image Classification
- Relationships in a Photo
- Visual Question Answering
- Action Recognition
- Point Clouds
- 3D Classification and Segmentation
- RGBD Semantic Segmentation
- Situation Recognition
- Social Relationship Understanding
- Zero-Shot Action Recognition

Graph neural networks have the ability to immediately analyze input graphs, thus incorporating its connectivity into the product criteria. Most popular techniques to graph theory are based on a beginning stage that translates each graph over to a smaller data type, such as a vector or a series of reals. Interactions between humans and objects- GPNN repeatedly modifies eigenvector matrices and node labeling within such a passing messages inference framework. The V-COCO, HICO-DET and CAD-120 datasets are used for testing on 3 HOI identification benchmarks on images and videos. GPNN is adaptable to large datasets and can be used in both spatiotemporal scenarios. Visual QA is a graph-based way to address visual questions. Object detection- proposes spatial-temporal Graph Convolution Network (ST-GCN), a new model of dynamic skeleton that overcomes the constraints of earlier methods by understanding both spatial and temporal variation from necessary data. Model based situation recognition is a GNN that allows to record joint interdependence between tasks effectively using neural network models formed on a network.

VI. RESOURCES AND PLATFORMS FOR GRAPH COMPUTING IN CV APPLICATIONS

TABLE.I. STANDARD GRAPH LEARNING RESOURCES

Repository	Web Link
Network Repository	http://networkrepository.com
Graph Kernel Datasets	https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets
Relational Dataset Repository	https://relational.fit.cvut.cz/
Stanford Large Network Dataset Collection	https://snap.stanford.edu/data/
Open Graph Benchmark	https://ogb.stanford.edu

TABLE II. PLATFORMS FOR GRAPH COMPUTING

Platform	Web Link
PyTorch Geometric	https://github.com/rusty1s/pytorch_geometric
Deep Graph Library	https://github.com/dmlc/dgl
AliGraph	https://github.com/alibaba/aligraph
GraphVite	https://github.com/DeepGraphLearning/graphvite
Paddle Graph Learning	https://github.com/PaddlePaddle/PGL
Euler	https://github.com/alibaba/euler
Plato	https://github.com/tencent/plato
CogDL	https://github.com/THUDM/cogdl/
OpenNE	https://github.com/thunlp/OpenNE/tree/pytorch

TABLE III. GRAPH MODELS FOR COMPUTER VISION APPLICATIONS

Graph Model	Year	Web link
Gated Graph Neural Network (GGNN)	2015	https://github.com/yujiali/ggnn
Diffusion Convolutional Neural Network (DCNN)	2016	https://github.com/jcatw/dcnn
Graph Convolutional Network (GCN)	2017	https://github.com/tkipf/gcn
Graph Attention Network (GAT)	2017	https://github.com/PetarV-/GAT
GraphRNN	2018	https://github.com/snap-stanford/GraphRNN
Dual Graph Convolutional Network(DGCN)	2018	https://github.com/ZhuangCY/DGCN
Deep Graph Infomax (DGI)	2019	https://github.com/PetarV-/DGI

Table 1 shows the repository and web links of standard graph learning resources. Various platforms that can be used for graph computing applications is tabulated in table 2. Table 3 depicts the popular graph models can be used for computer vision applications.

VII. RESEARCH AND IMPLEMENTATION CHALLENGES

Despite the positive outcomes, previous works are continually hampered by the following two flaws:

Hyper parameters-Aside from GNN structure, a little change in hyper parameters can affect the performance of converging structural model. Currently available approaches that simply optimize structural variables with fixed hyper parameter values may result in a model that is unsatisfactory.

Scalability -The time it takes to train recurrent network contributes to the search time. Run-time computation would be required for both the controller training and the single GNN model training. Furthermore, the controller often produces and analyses potential GNN structures in a sequential fashion, which makes scaling to a vast searching space problematic.

VIII. CONCLUSION

Design and implementation of Graph Neural Network (GNN) for computer vision (CV) applications is currently an active ongoing researching topic for various application domains not only limited to CV. Still there are several unanswered concerns. Spatio-temporal graph neural network architecture has been described to explicitly simulate interactions between

actors, objects, and their environment. This approach can implicitly or explicitly characterize objects, and it generalizes the existing structured models for video comprehension. Usage of adaptive approach for better score, district task across distinct dataset proposed. On AVA, still lot of additional work needs to be done to better utilize explicit object representations.

Further, Use of GNN in image and video understanding, architecture, applications, resources, platforms, graph models and implementation challenges towards computer vision is elaborated in detail.

REFERENCES

[1] Anurag Arnab et al., "Unified Graph Structured Models for Video Understanding" *CVPR*, *arXiv:2103.15662*, 2021.

[2] Yubo Zhang et al., "A Structured Model For Action Detection" *CVPR, arXiv:1812.03544,* 2019.

[3] Santiago Castro et al.,"Fill-in-the-blank as a Challenging Video Understanding Evaluation Framework" *CVPR, arXiv:2104.04182*,2021.

[4] Saurabh Sahu et al.," Enhancing Transformer for Video Understanding Using Gated Multi-Level Attention and Temporal Adversarial Training"*CVPR, arXiv:2103.10043*, 2021.

[5] Matthew Hutchinson et al.," Video Action Understanding: A Tutorial" *CVPR, arXiv:2010.06647*, 2020.

[6] Ishan Dave et al.,"TCLR: Temporal Contrastive Learning for Video Representation"*CVPR, arXiv:2101.07974*, 2021.

[7] Gunnar A. Sigurdsson et al.," Beyond the Camera: Neural Networks in World Coordinates"*CVPR, arXiv:2003.05614*, 2020.

[8] Yubo Zhang et al.," A Structured Model For Action Detection" *CVPR, arXiv:1812.03544,* 2018.

[9] Hao Huang et al.," Dynamic Graph Modules for Modeling Object-Object Interactions in Activity Recognition"*CVPR, arXiv:1812.05637*, 2018.

[10] Yunpeng Chen et al.," Graph-Based Global Reasoning Networks",*CVPR, arXiv:1811.12814*, 2018.

[11] Jie Zhou et al.,"Graph Neural Networks: A Review of Methods and Applications"*arXiv:1812.08434*, 2021.

[12] F. Scarselli et al.,"The Graph Neural Network Model," *IEEE Transactions on Neural Networks,vol. 20, no. 1,* pp. 61-80, 2009.

[13] Karen Simonyan et al.,"Two-Stream Convolutional Networks for Action Recognition in Videos" *Visual Geometry Group(VGG), University of Oxford*.

[14] Rohit Girdhar et al.,"Video Action Transformer Network" *CVPR, arXiv:1812.02707,* 2018.

[15] Siyuan Qi et al.," Learning Human-Object Interactions by Graph Parsing Neural Networks*" CVPR, arXiv:1808.07962*,2018.

[16] Xin Guo et al.," Graph Neural Networks for Image Understanding Based on Multiple Cues: Group Emotion Recognition and Event Recognition as Use Cases" *CVPR, arXiv:1909.12911*, 2020.

[17] Danfei Xu et al.," Scene Graph Generation by Iterative Message Passing" *CVPR, arXiv:1701.02426*,2017.

[18] Yubo Zhang et al.," A Structured Model For Action Detection" *CVPR, arXiv:1812.03544*, 2019.

[19] E. Mavroudi et al.,"Representation Learning on Visual-Symbolic Graphs for Video Understanding" *arXiv:1905.07385*, 2020.

[20] Biswas A et al.," "Survey on Edge Computing–Key Technology in Retail Industry" *Lecture Notes on Data Engineering and Communications Technologies,* 2021, vol 58. Springer, Singapore.

[21] R. J. Franklin et al.,"Anomaly Detection in Videos for Video Surveillance Applications using Neural Networks," *Fourth International Conference on Inventive Systems and Control (ICISC)*, 2020.

[22] Mohana et al.,"Performance Evaluation of Background Modeling Methods for Object Detection and Tracking," 4th *International Conference on Inventive Systems and Control (ICISC)*, 2020.

[23] A. Biswas et al.,"Classification of Objects in Video Records using Neural Network Framework," *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2018.

[24] H. Jain et al.," Weapon Detection using Artificial Intelligence and Deep Learning for Security Applications," *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020.

[25] M. R. Nehashree et al.," Simulation and Performance Analysis of Feature Extraction and Matching Algorithms for Image Processing Applications," *International Conference on Intelligent Sustainable Systems (ICISS),* 2019.

[26] R. K. Meghana et al.," Background-modelling techniques for foreground detection and Tracking using Gaussian Mixture Model," *3rd International Conference on Computing Methodologies and Communication (ICCMC),* 2019.

[27] V. P. Korakoppa et al.,"Implementation of highly efficient sorting algorithm for median filtering using FPGA Spartan 6," *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),* 2017.

[28] D. Akash et al.,"Interfacing of flash memory and DDR3 RAM memory with Kintex 7 FPGA board," *International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2017.

[29] N. Jain et al.,"Performance Analysis of Object Detection and Tracking Algorithms for Traffic Surveillance Applications using Neural Networks," *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019.

[30] C. Kumar B et al.,"YOLOv3 and YOLOv4: Multiple Object Detection for Surveillance Applications," *International Conference on Smart Systems and Inventive Technology (ICSSIT),* 2020.

[31] C. Kumar B et al.," Performance Analysis of Object Detection Algorithm for Intelligent Traffic Surveillance System," *International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020.

[32] R. J. Franklin et al.,"Traffic Signal Violation Detection using Artificial Intelligence and Deep Learning," *International Conference on Communication and Electronics Systems* (ICCES), 2020.

[33] Mohana et al.,"Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications" *International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019.*

[34] Manoharan Samuel et al., "Improved Version of Graph-Cut Algorithm for CT Images of Lung Cancer with Clinical Property Condition" *Journal of Artificial Intelligence and Capsule networks,* 2(4), 201–206.