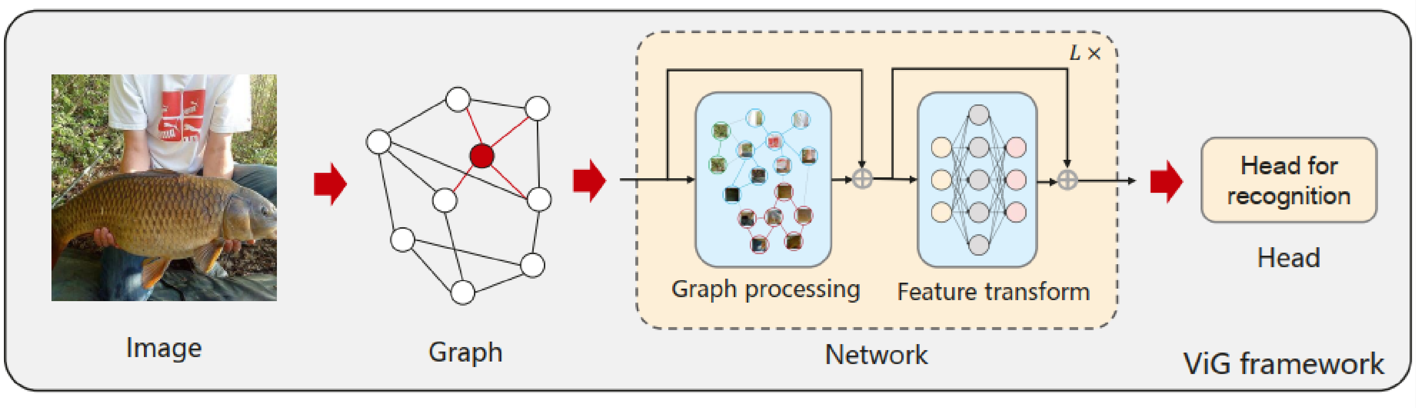
Vision GNN-Powered Object Detection

Sagar Prakash Barad

Harnessing Vision GNNs as Backbone Feature Extractors for RetinaNet and Mask R-CNNs

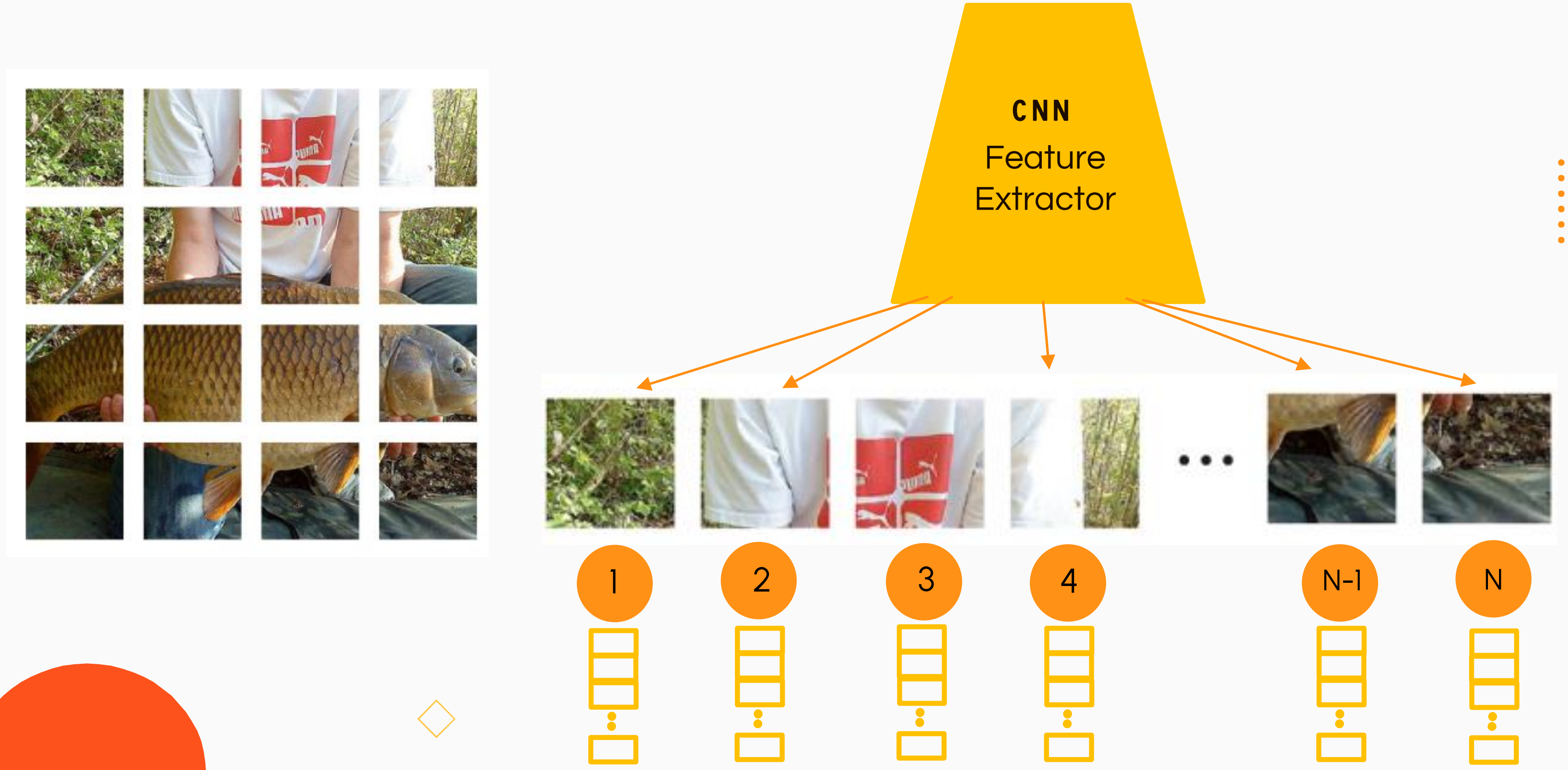


RetinaNet ▶ ~~ResNet~~ + FPN + Focal Loss
Vision GNN

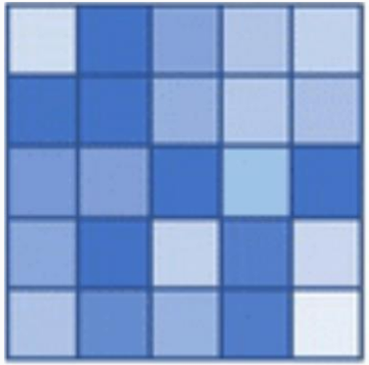
Dataset(s) – COCO Dataset, PubTables 1M Dataset, FinTab Dataset

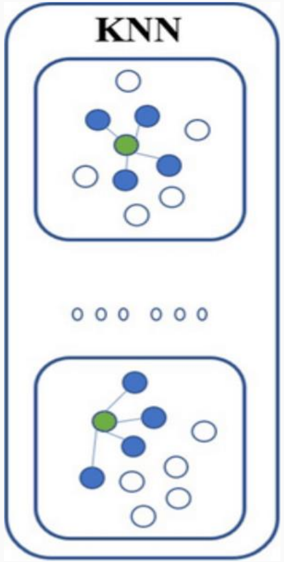
Wang, X., Liu, Z., Zhang, S., Li, B., Wang, T., & Zhang, J. (2022). Vision GNN: An Image is Worth Graph of Nodes. *Advances in Neural Information Processing Systems, 35*. Retrieved from https://arxiv.org/abs/2206.00272

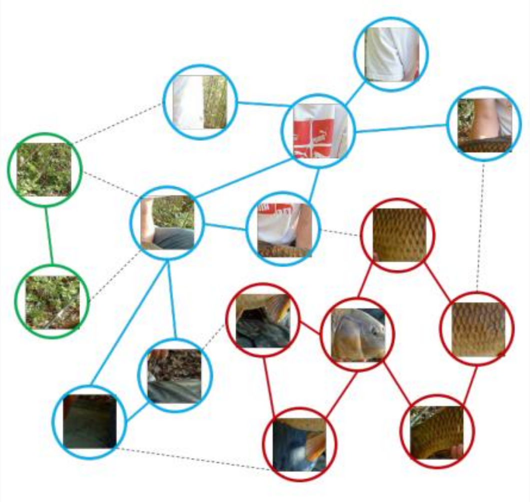
Vision GNN Works by...



Graph Represnation





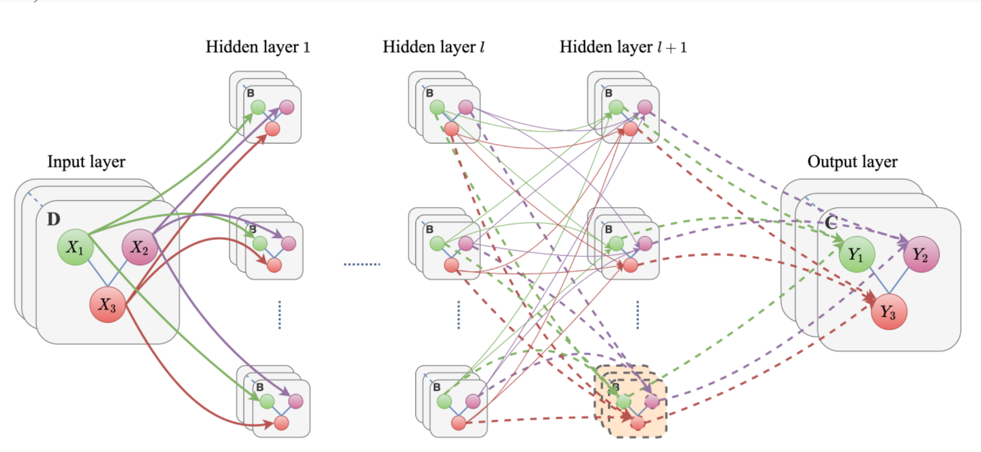


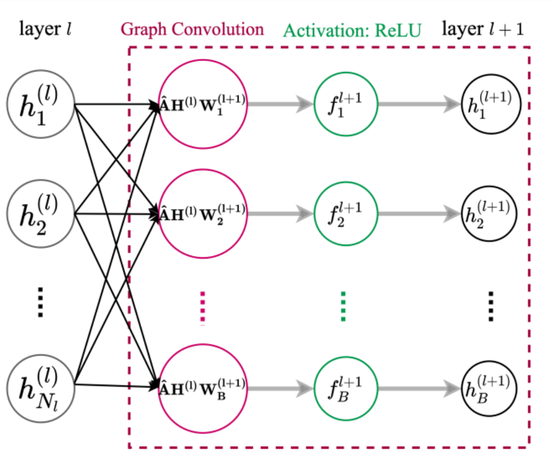
Directed Graph

Benefits

1. Generalized Data Structure
2. Flexibility for Complex Objects
3. Part-Object Relationships

Graph Convolution





Heidari, N. (2020, March 27). Progressive Graph Convolutional networks for Semi-Supervised Node Classification. arXiv.org. https://arxiv.org/abs/2003.12277

$$\mathbf{h}_{\mathcal{N}_v}^t = \text{AGGREGATE}_t\left(\{\mathbf{h}_u^{t-1}, \forall u \in \mathcal{N}_v\}\right)$$
$$\mathbf{h}_v^t = \sigma\left(\mathbf{W}^t \cdot [\mathbf{h}_v^{t-1} \| \mathbf{h}_{\mathcal{N}_v}^t]\right)$$

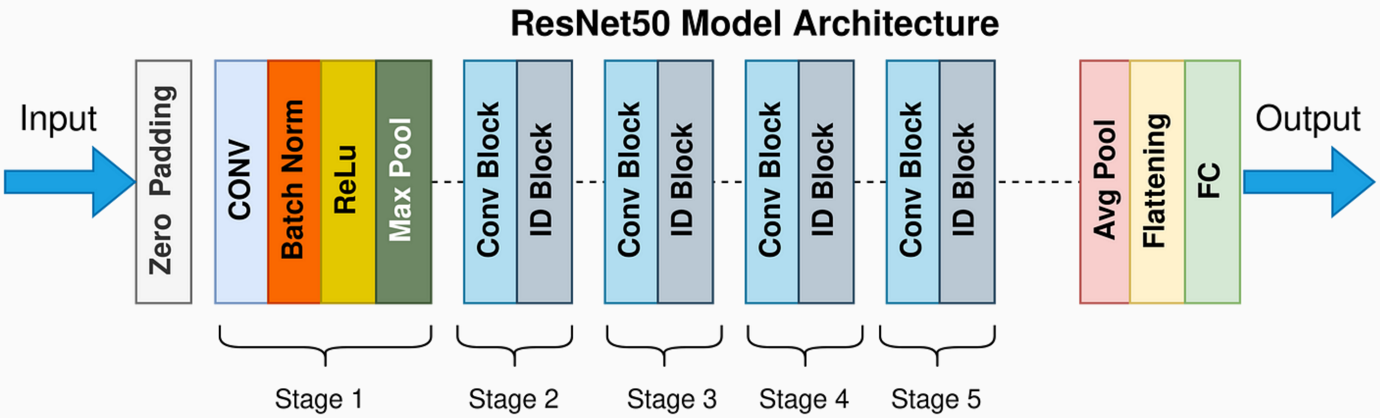
$$\mathbf{h}_{\mathcal{N}_v}^t = \max(\{\sigma\left(\mathbf{W}_{\text{pool}}\mathbf{h}_u^{t-1} + \mathbf{b}\right), \forall u \in \mathcal{N}_v\})$$

like mean, sum or max function

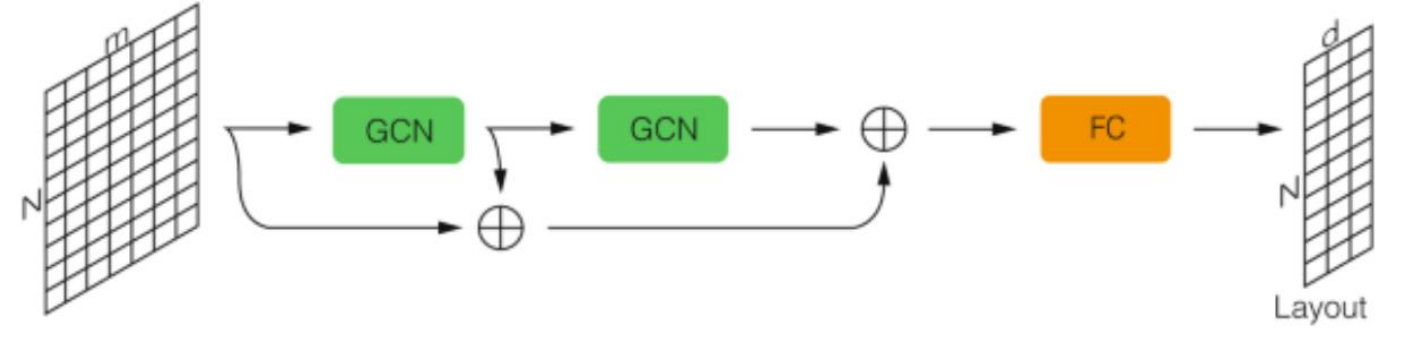
$$\mathbf{h}_v^{(k)} = \sigma(\mathbf{W}^{(k)} \cdot f_k(\mathbf{h}_v^{(k-1)}, \{\mathbf{h}_u^{(k-1)}, \forall u \in S_{\mathcal{N}(v)}\}))$$

where $\mathbf{h}_v^{(0)} = \mathbf{x}_v$, $f_k(\cdot)$ is an aggregation function, $S_{\mathcal{N}(v)}$ is a random sample of the node v's neighbors.

Comparision with CNNS



Sapireddy, S. R. (2023, July 1). ReSNEt-50: Introduction - Srinivas Rahul Sapireddy - Medium. Medium. https://srsapireddy.medium.com/resnet-50-introduction-b5435fdba66f



Both, C., Dehmamy, N., Yu, R. et al. Accelerating network layouts using graph neural networks. Nat Commun 14, 1560 (2023). https://doi.org/10.1038/s41467-023-37189-2

ViG Models

ViG-Ti

Flops(B): 7.1
Params(M): 1.3

ViG-S

Flops(B): 4.5
Params(M): 22.7

ViG-B

Flops(B): 17.7
Params(M): 86.8

ResNet -18

Flops(B): 1.82
Params(M6): 11.7

ResNet -50

Flops(B): 3.86
Params(M): 25.6

ResNet -101

Flops(B): 7.1
Params(M): 44.6

Model	Depth	Dimension D
ViG-Ti	12	192
ViG-S	16	320
iG-B	16	640

Results on CIFAR

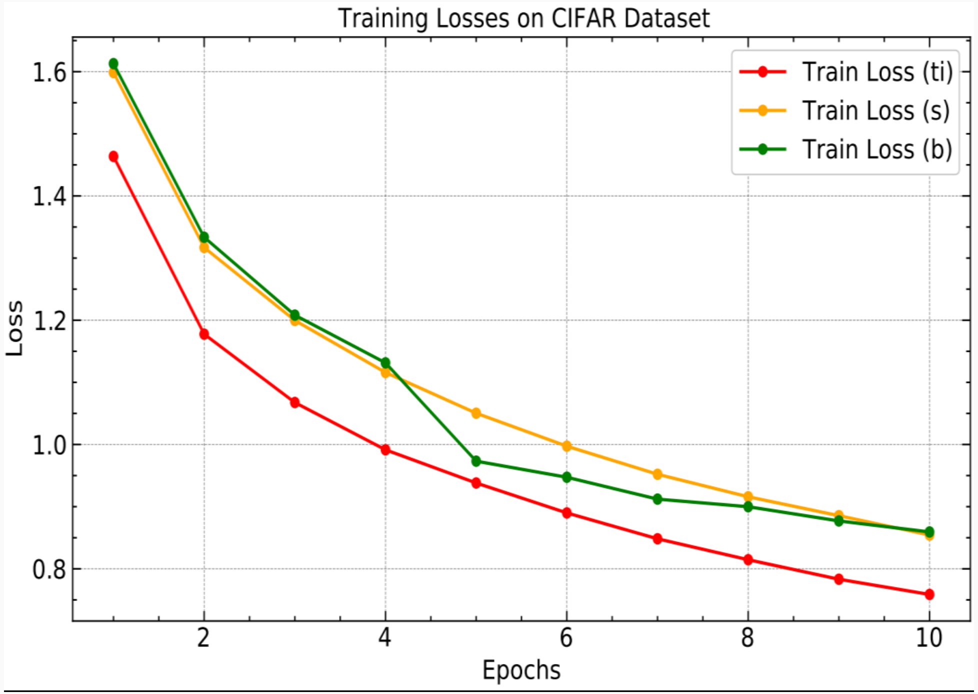


Table 3. ViG models on CIFAR Dataset

Model	Top-1	Top-5
ViG-Ti	66.1	97.67
ViG-S	65.64	97.95
ViG-B	67.71	98.78

Results on ImageNet 1k

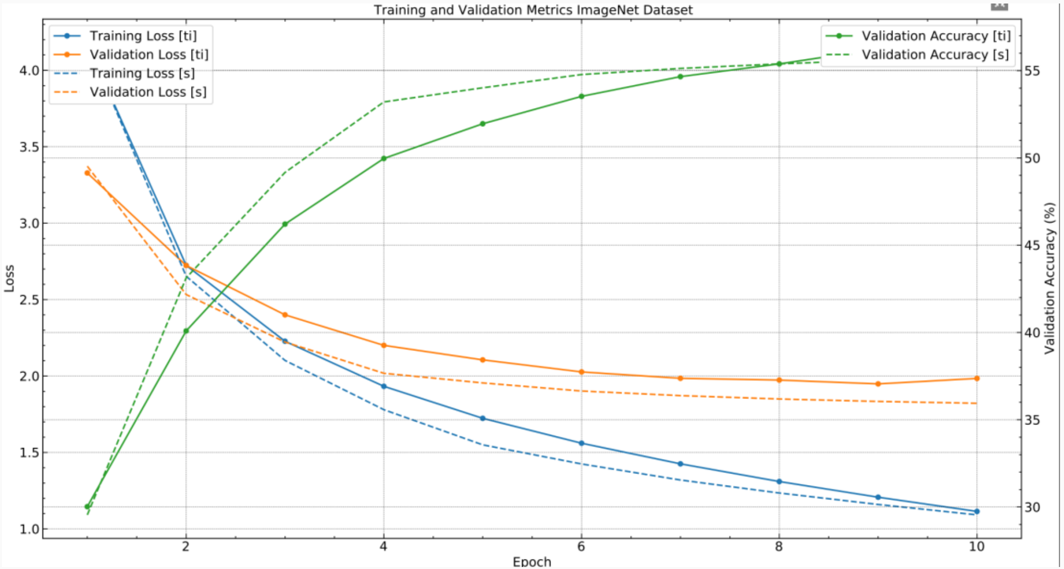


Table 4. ViG models on ImageNet Dataset

Model	Top-1	Top-5
ViG-Ti	55.60	93.32
ViG-S	55.75	93.95

Conclsuion

1. ViG models (ViG-Ti, ViG-S, ViG-B) show promise in our training, with potential for comparable or superior performance to ResNet variants (ResNet-18, ResNet-50, ResNet-101) as we increase training epochs.

2. ViG models match ResNet models in size (FLOPs and parameters) but outperform them in image classification.