Superpixel-based Domain-Knowledge Infusion in Computer Vision

Gunjan Chhablani[1] and Abheesht Sharma[1] and Harshit Pandey[3] and Tirtharaj Dash[1,2]

1- Dept. of CS & IS, BITS Pilani, Goa Campus, Goa, India
2- APPCAIR, BITS Pilani, Goa Campus, Goa, India
3- Dept. of CS, Savitribai Phule Pune University, Pune, India
(First 3 authors: Equal contribution; Random order)

Abstract. Superpixels are higher-order perceptual groups of pixels in an image, often carrying much more information than raw pixels. There is an inherent relational structure to the relationship among different superpixels of an image. This relational information can convey some form of domain information about the image, e.g. relationship between superpixels representing two eyes in a cat image. Our interest in this paper is to construct computer vision models, specifically those based on Deep Neural Networks (DNNs) to incorporate these superpixels information. We propose a methodology to construct a hybrid model that leverages (a) Convolutional Neural Network (CNN) to deal with spatial information in an image, and (b) Graph Neural Network (GNN) to deal with relational superpixel information in the image. The proposed deep model is learned using a generic hybrid loss function that we call a 'hybrid' loss. We evaluate the predictive performance of our proposed hybrid vision model on four popular image classification datasets: MNIST, FMNIST, CIFAR-10 and CIFAR-100. Moreover, we evaluate our method on three real-world classification tasks: COVID-19 X-Ray Detection, LFW Face Recognition, and SOCOFing Fingerprint Identification. The results demonstrate that the relational superpixel information provided via a GNN could improve the performance of standard CNN-based vision systems.

1 Introduction

Deep learning, especially in the task of image classification and recognition has taken centerstage, mostly due to the introduction of ILSVRC challenge[1]. There have been significant architectural innovations for convolutional neural networks (CNNs). In the last few years, the core approach of basic convolution has been adopted to graph-structured data via the introduction of graph neural networks (GNNs)

A graph is a representation of binary relations and GNNs can deal with these relational data. In the case of tasks involving images, binary relations can be easily seen at the level of 'image superpixels'. Superpixels are a higher-order perceptual groups of pixels in an image, often conveying much more information than low-level raw pixels and sharing some common characteristics such as intensity levels.

[1] https://image-net.org/challenges/LSVRC/

One can observe that the superpixels share a binary relation with their adjacent neighbours, resulting in a graph structure. Our present work deals with treating the relational information conveyed by the superpixels as domain-information about an image and providing this information to aid in image classification tasks. It has been observed that incorporating domain knowledge can enhance the performance of deep learning systems [1]. In this work, we leverage spatial information from the image and infuse domain knowledge in the form of binary relations procured from the superpixel graph representation of the image. CNN filters tend to learn parameters based on pixel-level information. We hypothesise that fusing such superpixel-level information can provide a higher-level understanding of the image and aid a CNN in classification tasks. Specifically, we treat the graph resulting from the image superpixels as an input to a GNN and learn a CNN-based vision system together with the GNN. The coupled hybrid CNN+GNN system is expected to be knowledge-rich both at the level of raw pixels and at the level of superpixels.

The major contributions of our paper are as follows: (1) Treating superpixel representation of an image as a graph and allowing a GNN to extract higher-level domain information about an image; (2) We construct an image classification model by coupling the GNN with a CNN-based baseline; (3) We conduct a series of empirical evaluations of our coupled CNN+GNN hybrid system on four different popular image classification benchmarks and three case studies.

The rest of the paper is organised as follows. In Sec. 2 we provide details of our methodology including an introduction to loss suitable for learning the hybrid model. Section 3 provides details of our experiments. Section 4 provides brief description of related work. Section 5 concludes the paper.

2 Superpixels Integrated Vision Model

2.1 Superpixel Graph Construction

Simple Linear Iterative Clustering (SLIC) [2] is a simple and easy-to-use algorithm requiring an input parameter: an approximate number of superpixels along with the input image, and outputs a segmented image. We use SLIC to construct the segmented version of the input image: each segmented patch in the image is now a superpixel. We then treat the centroid of every superpixel as a node in the graph. We link these nodes by building a radius graph [3]. For every pair of nodes, we form an edge between them if and only if the Euclidean distance between them is less than a pre-ordained radius, r. Mathematically, a radius graph is defined as $G = (V, E)$ such that $V = \{x_i : x_i \text{ is a superpixel}\}$, and $E = \{\{x_i, x_j\} : x_{i,j} \in V, x_i \neq x_j, ||x_i -_j || \leq r\}$.

2.2 The Coupled CNN+GNN Model

Our main motive in this work is to learn jointly from the image and from its superpixel representation. This results in a complementary combination of a vision model (CNN) and a relational model (GNN). The CNN takes care of

feature extraction (low-level features such as edges in initial layers to complex objects at the latter layers), while the graph-based model takes the superpixel-based radius graph and extracts relational information about the image that can act as domain-knowledge. A block diagram in Fig. 1 demonstrates this hybrid combination.

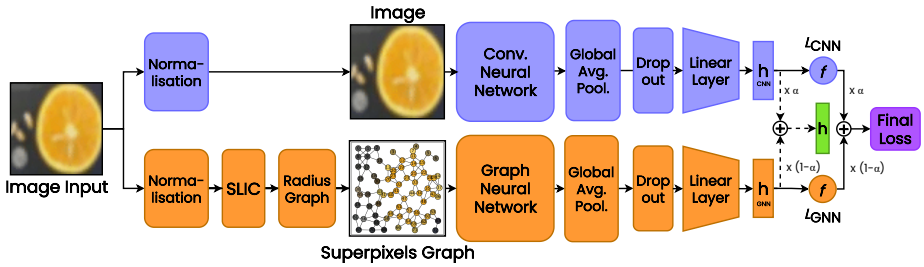

Fig. 1: Block diagram of the proposed CNN+GNN model

Regardless of the detailed architectural specifics, the goal is to construct a representation of the input (image) that consists of both spatial information (extracted by the CNN) and the domain information (extracted by the GNN). This representation can then be used to construct a feed-forward fully-connected neural network that outputs a class label for the input image. In order to train the coupled model in an end-to-end fashion and utilise the information from the combined representation adequately, we propose a simple hybrid loss as follows:

$$\mathcal{L}_{CG} = \alpha \mathcal{L}_C + (1 - \alpha)\mathcal{L}_G$$

Here $\mathcal{L}_C$ and $\mathcal{L}_G$ denote the cross-entropy loss for CNN and GNN respectively. The tunable parameter α determines the relative importance of the two models during training, i.e., a value $\alpha = 0.5$ would mean that both the raw spatial information and the domain information are equally important.

During prediction, we construct a hybrid logits from the logits computed by the CNN and GNN for an input image:

$$\hat{y} = \arg\max_j (\alpha h_C + (1 - \alpha) h_G)$$

where h_C and h_G are the logits from CNN and GNN respectively.

3 Empirical Evaluation

3.1 Aims

Our aim in this study is to integrate superpixel-level domain knowledge into vision systems, specifically those based on CNNs. Our empirical experiments attempt to answer the following questions: Can GNNs construct rich relational representations from superpixels? [RQ1]; Can GNN-constructed knowledge improve CNN-based vision systems? [RQ2]

3.2 Materials

We test our hypothesis on four popular benchmarks for image classification: MNIST, FMNIST, CIFAR-10 and CIFAR-100. Each of these datasets are supplied with official train-test splits. Additionally, we perform three case-studies: COVID-19 Chest X-Ray Detection[2], LFW Facial Recognition[3] and SOCOFing Fingerprint Identification[4] datasets. For the LFW dataset, we discard classes with less than 20 faces. For all the domain-specific datasets, we do a stratified train-validation-test split in 80:10:10. We use the method described in Sec. 2.1 to construct superpixel-based radius graph for each image in these datasets. For convenience, we refer this as a procedure *RadiusGraph* that takes two inputs: a set of images and the number of superpixels (n), and returns a set of radius graph, each corresponding to an image in the input image-set.

We use the PyTorch library for the implementation of CNN and PyTorch Geometric for GNN models. We conduct all our experiments in the OVHCloud[5] using "AI Training" platform with following configurations: 32GB NVIDIA V100S GPU, 45GB main memory, 14 Intel Xeon 2.90GHz processors and a NVIDIA-DGX station with 32GB Tesla V100 GPU, 256GB main memory, 40 Intel Xeon 2.20GHz processors.

3.3 Method

Our method is straightforward. For each dataset D, We use the $RadiusGraph(D, n)$ procedure to get graphs D' for all the dataset images, where n is the approximate number of superpixels. Then, we split the dataset into train (D_{tr}, D'_{tr}) and test (D_{te}, D'_{te}) splits, such that D'_{te} are made of the images in D_{te}. Then, we train CNN model on D_{tr}, and the coupled CNN+GNN model on D'_{tr}, and compare their performances on respective test sets.
The following details pertaining to the above steps are relevant: (1) We use $n = 75$ for MNIST and FMNIST, $n = 100$ for CIFAR, LFW and SOCOFing and $n = 200$ for COVID to construct superpixel graphs using *RadiusGraph*; the number of radial neighbours is set to 5 for MNIST, FMNIST and CIFAR datasets, 27 for COVID, 10 for LFW, and 15 for SOCOfing; This was decided using graph visualisation. (2) The node feature-vector for the radius graphs are: the normalised pixel value and the location of centroid in the image, resulting in 3-length feature-vector for MNIST and FMNIST, and 5-length feature-vector for CIFAR, COVID and LFW; (3) We use 90:10 split on D_{tr} (correspondingly, D'_{tr}) to obtain a validation set useful for hyperparameter tuning; (4) We use the AdamW optimiser [4] for training all our models; (5) The hyperparameters are obtained by a sweep across grids such as batch-size: $\{128, 256\}$, learning rate: {1e-3,1e-4,1e-5}, weight-decay parameter: $\{0, 0.001\}$ for MNIST, FMNIST, CIFAR datasets. For other datasets, we use the best parameters obtained during

[2] https://github.com/tawsifur/COVID-19-Chest-X-ray-Detection
[3] http://vis-www.cs.umass.edu/lfw/
[4] https://www.kaggle.com/ruizgara/socofing
[5] https://www.ovhcloud.com/en-ie/

MNIST/FMNIST/CIFAR grid-search; (6) CNN structure is built with 3-blocks; each block consists of a convolutional layer, a batch-norm layer, ReLU activation and a max-pool layer. The number of channels in each convolution layer is 32, 64 and 64, respectively; (7) The coupled CNN+GNN structure consists of the same CNN backbone as mentioned above; and a GNN structure consisting of three graph-attention layers with 32, 64 and 64 channels, respectively, and the number of heads is 1; (8) The value of α is set to 0.75 for all our experiments; (9) We use accuracy as the metric to measure the predictive performance of the models on D_{te}.

3.4 Results

In all our experiments, the primary intention is to show whether relational information provided via a superpixel graph is able to aid in improving the performance of a CNN-based model. One should note that the primary backbone of the CNN in the standalone model (CNN) and that in the coupled model remains the same to draw any meaningful conclusion. The principal results of our experiments are reported in Table 1. The results demonstrate that the higher-level domain-knowledge extracted by a GNN from the superpixel graph is able to substantially boost the predictive performance of CNNs in a hybrid-learning setting in some cases.

Dataset	CNN	CNN+GNN
MNIST	**99.30**	99.21
FMNIST	**91.65**	91.50
CIFAR-10	**77.80**	76.81
CIFAR-100	42.88	**46.79**

Dataset	CNN	CNN+GNN
Chest X-ray	89.09	**91.01**
Face	60.83	**66.12**
Fingerprint	65.68	**93.58**

Table 1: Predictive performance of CNN and CNN+GNN on generic datasets (L), domain-specific datasets (R)

4 Related Work

Various graph-based learning approaches for superpixel image classification can be distinguished based on how they perform neighbourhood aggregation. In MoNet [5], weighted aggregation is performed by learning a scale factor based on geometric distances. They also test their model on the MNIST Superpixels dataset. In Graph Attention Networks (GATs) [6], self-attention weights are learned. SplineCNN [7] uses B-spline bases for aggregation, whereas SGCN [8] is a variant of MoNet and uses a different distance metric. Preliminary work on enhancing Semantic Segmentation Vision-based systems using superpixel representations has been done. Mi *et al.* [9] propose a Superpixel-enhanced Region Module (SRM), which they train jointly with a Deep Neural Forest. The SRM alleviates the noise by leveraging the pixel-superpixel associations. In DrsNet [10], coarse superpixel masking and fine superpixel masking are applied to the CNN features of the input image, particularly for rare classes and background areas.

5 Conclusion

In our work, we demonstrate that superpixels graphs represent domain knowledge for images and that infusing this knowledge in a standard vision-based training process shows significant gains in accuracy. We are able to conclude that a GNN can deal with relational information conveyed by a superpixel graph and can construct high-level relations that could boost the predictive performance when coupled with a CNN model. A straightforward extension is to employ a pre-trained CNN model for domain-specific tasks and observe how the superpixel information impact the predictive performance. Our code repository is available at: https://github.com/abheesht17/super-pixels.

Acknowledgments

We sincerely thank Jean-Louis Quéguiner, Head of Artificial Intelligence, Data and Quantum Computing at OVHCloud for providing us computing resources.

References

[1] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. Incorporating domain knowledge into deep neural networks. *arXiv preprint arXiv:2103.00180*, 2021.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[3] Jon L Bentley, Donald F Stanat, and E Hollins Williams Jr. The complexity of finding fixed-radius near neighbors. *Information processing letters*, 6(6):209–212, 1977.

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[5] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017.

[6] Pedro HC Avelar, Anderson R Tavares, Thiago LT da Silveira, Clíudio R Jung, and Luís C Lamb. Superpixel image classification with graph attention networks. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 203–209. IEEE, 2020.

[7] Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2018.

[8] Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. Spatial graph convolutional networks. In *International Conference on Neural Information Processing*, pages 668–675. Springer, 2020.

[9] Li Mi and Zhenzhong Chen. Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:140–152, 2020.

[10] Liangjiang Yu and Guoliang Fan. Drsnet: Dual-resolution semantic segmentation with rare class-oriented superpixel prior. *Multimedia Tools and Applications*, 80(2):1687–1706, 2021.