Project Proposal: **Multi-label image classification**

Team-04/TDM

Background

Multi-label Image Recognition is a fundamental task of Computer Vision, where the aim is to predict a set of objects present in an image. It has wide-ranging applications of recognition in medical diagnosis, human attributes and retail checkout. It is more challenging compared to multi-class image classification as the output space can be quite large due to the combinatorial nature of labels and implicit label dependencies have to be modelled.

Existing works on the topic include converting the multi-label problem to independent binary classifiers for each label and the use of Deep Convolutional Neural Networks (CNN's) [4, 5]. However, this method suffers from high computational complexity as it increases exponentially with the increase in number of labels and ignoring the topology structure between objects. To regularize prediction space and employ correlation between labels by converting labels into embedded label vectors, various "memory-based" methods were employed, including RNN's [6], attention-based methods [7], spatial transformers and long short-term memory (LSTM) based units [8]. To model the structural dependencies using structured learning methods, graph-based methods were introduced, which include using a Maximum Spanning Tree Algorithm in the label space [9], image-dependent conditional label structures based on a graphical Lasso framework [10] and knowledge graph modelling label relations [11]. However, all these methods are single image based and model local correlations but not global correlations.

In the given paper, a ML-GCN based method is proposed to incorporate the local as well as global correlations, which extends beyond a single image. The method uses a GCN classifier function and uses a reweighted correlation matrix to avoid overfitting and oversmoothing. This method takes care of all problems of the existing architecture problems, as the ML-GCN models label correlations as well as label dependencies and local as well as global correlations are modelled. In addition, this method provides unprecedented scalability and flexibility, which is not possible for competing approaches.

Paper Summary (150-200 words)

ML-GCN is a novel end-to-end trainable GCN-based framework for multi-label image recognition. Stacked GCNs are used where each GCN layer l takes the node representations from the previous layer (H_l) as inputs and outputs new node representations, i.e., H_{l+1} and information propagation between the nodes is controlled by the pre-defined data-driven correlation matrix A. The correlation between labels is obtained via mining of their co-occurrence patterns in the dataset - as a conditional probability which denotes the probability of occurrence of label L_j when label L_i appears:

$$P_i = \frac{M_i}{N_i}$$

Where N_i denotes the occurrence count of L_i in the training set,

$P_{ij} = P(\frac{L_j}{L_i})$ means the probability of label L_j when label L_i appears,

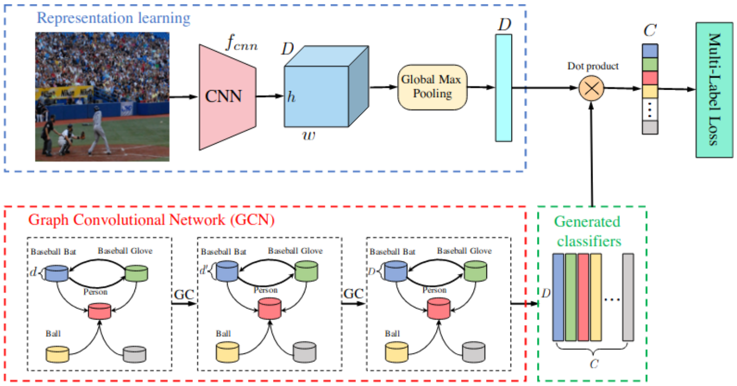
$M \in \mathbb{R}_{C \times C}$ is a matrix having the occurrence count of label pairs in the training set

To handle noisy edges and inconsistencies in the absolute number of co-occurrences, A may be thresholded to a binary correlation matrix, and each node's features can be re-weighted against those of its correlated nodes to solve the over-smoothing problem.

The generated object classifiers are applied to image representations x learned using CNN-based models like ResNet-101. The output of the stacked GCNs is the inter-dependent object classifiers matrix W, and the final predicted scores are as follows:

$$y = W.x$$

which are learned against ground truth labels using the multi-label classification loss function. The overall architecture is depicted below:



Quantitative comparison on standard datasets such as MS-COCO [cite] and VOC 2007[cite] reveals that ML-GCN outperforms SOTA methods like CNN-RNN, RNN-Attention, Order-Free RNN, ML-ZSL, SRN and Multi-Evidence in mean Average Precision by around 2%. The authors discover that ML-GCN benefits image representation learning as well, by retrieving better image features compared to vanilla ResNet.

Project Description

[Main goal]

Introducing changes in the overall model architecture to observe the change in performance. On a smaller scale, experimenting with a different feature conversion model to observe overall changes in models and using a different label encoding compared to the one present in the paper to observe performance changes in the overall architecture.

[What task/objective will you address]

· Introducing different image representation models for a possible increase in performance (e.g., DenseNet)

· Experiment different label encoding models for a possible performance boost (e.g., Word2Vec, BERT, Roberta)

- Experimenting with different architectural innovations to try and model more data-based information into the architecture (e.g., UNet, Attention layers, Graph Attention Transformers)

- Evaluate against latest benchmarks (ML-Decoder , Q2L-CvT)

- Downstream task: Extract images posted on social media, i.e., Twitter and generate alt-text for the image based on object tags.

[What data will you use? Already existing code?]

The datasets we will use are:

- MS-COCO - The objects are categorized into 80 classes with about 2.9 object labels per image.
- VOC 2007 - It contains 9,963 images from 20 object categories,

The implementation of the concept is present.

[What baseline(s) will you use? How will you evaluate your results?]

For the baseline we will be using the results given in the paper. We are also planning to include the current state of the art result in comparison to our model performance. The metrics that will be used for comparison are average per-class precision (CP), recall (CR), F1 (CF1) and the average overall precision (OP), recall (OR), F1 (OF1) for performance evaluation.

Timeframe + Work Distribution (For two months)

	Task	Work Distribution	Start and End Dates
Phase One	Understanding the implementation and ablation study for parameters	Understanding code: All three Re-weighting parameter changes: Dipanwita Threshold changes for binary correlation: Manav Network depth experiments: Tanmay	25th February – 7th March
Phase Two	Experiment with different label embeddings, image representation models	All experiments will be equally divided among us	8th March – 20th March

Phase Three	Architectural innovation and benchmarking, documentation, reports	All experiments will be equally divided among us	21st March – 15th April

[References (if any)]

1. Residual Attention: A Simple but Effective Method for Multi-Label Recognition
2. Microsoft COCO: Common Objects in Context
3. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.
5. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018.
6. Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A unified framework for multi-label image classification. In CVPR, pages 2285–2294, 2016.
7. Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In CVPR, pages 5513–5522, 2017.
8. Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In ICCV, pages 464–472, 2017
9. Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. In UAI, pages 1–10, 2014
10. Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In CVPR, pages 2977–2986, 2016.
11. Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In CVPR, pages 1576–1585, 2018.