Modular Graph Transformer Networks for Multi-Label Image Classification

Hoang D. Nguyen [1], Xuan-Son Vu [2], Duc-Trong Le [3]

[1] School of Computing Science, University of Glasgow, Singapore
[2] Department of Computing Science, Umeå University, Sweden
[3] University of Engineering and Technology, Vietnam National University, Vietnam
harry.nguyen@glasgow.ac.uk, sonvx@cs.umu.se, trongld@vnu.edu.vn

Abstract

With the recent advances in graph neural networks, there is a rising number of studies on graph-based multi-label classification with the consideration of object dependencies within visual data. Nevertheless, graph representations can become indistinguishable due to the complex nature of label relationships. We propose a multi-label image classification framework based on graph transformer networks to fully exploit inter-label interactions. The paper presents a modular learning scheme to enhance the classification performance by segregating the computational graph into multiple sub-graphs based on modularity. Our approach, named Modular Graph Transformer Networks (MGTN), is capable of employing multiple backbones for better information propagation over different sub-graphs guided by graph transformers and convolutions. We validate our framework on MS-COCO and Fashion550K datasets to demonstrate improvements for multi-label image classification. The source code is available at https://github.com/ReML-AI/MGTN.

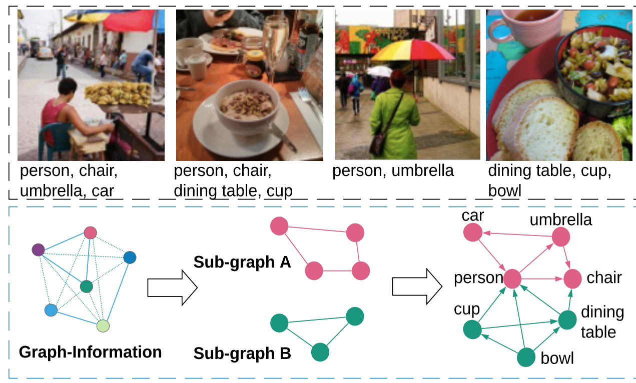

Figure 1: We segregate the graph into sub-graphs to learn inter-connected dependencies over the object labels to better model the multi-label image recognition task. In this figure, *"person, chair, umbrella, car"* is in one sub-graph, *"dining table, cup, bowl"* is in another sub-graph.

Introduction

Real-world images generally embodies rich and diverse semantic information with multiple objects or actions; therefore, multi-label classification has attracted a large number of recent studies in the artificial intelligence (AI) community (Wang et al. 2020a; Yeh et al. 2017; Zhu et al. 2017). Recognising object labels in images has many applications, ranging from social tag recommendation (Nam et al. 2019; Vu et al. 2020) and fashion trend analysis (Inoue et al. 2017) to functional genomics (Bi and Kwok 2011). The core challenge in multi-label learning is to understand and model object dependencies to exploit attributive knowledge. One of the early approaches developed by Wang et al. (2016) combined convolutional neural networks (CNN) with recurrent neural networks (RNN) to learn the semantic relevance and dependency of multiple labels in order to boost the classification performance. Nevertheless, this approach is prone to the high computational cost and the sub-optimal reciprocity between visual and semantic information. In reality, objects are inter-connected which reflect as the network nature of object label dependencies. Kipf and Welling (2017) proposed semi-supervised learning on network data using graph convolutional network (GCN) unveiled spectral graph convolutions for classification tasks. The graph-based approach was adopted with images by Chen et al. (2019b) to demonstrate the state-of-the-art performance for multi-label image recognition. Furthermore, Li et al. (2019) and Wang et al. (2020b) proposed several topological and architectural changes to enhance the learning capabilities with minor performance improvements.

This paper introduces Modular Graph Transformer Networks (MGTN) for multi-label image recognition by integrating semantic and topological label information in a harmonious way. Multi-label classification is decomposed into the segregated learning of multiple sub-graphs based on the modularity of object dependencies, leading to better performance in visual representation learning. In Figure 1, objects such as *bowl, cup, dining table, chair, umbrella, car* and *person* may co-occur in the physical world; nevertheless, the object labels appear to be clustered into sub-networks in the data. Therefore, multi-label learning entails better designated visual representation understanding as well as to reduce overfitting of the popular labels. In this example, we segregate the network into sub-graphs: G_1(*"bowl, cup, dining table"*) and G_2(*"person, chair, umbrella, car"*) to model the multilearning recognition task. The information propagation through sub-networks is guided by graph neural networks with the use of multiple modular backbones.

Compared with existing multi-label classification studies, our proposed MGTN establishes a new state-of-the-art with a number of the following contributions:

- We propose end-to-end graph transformer networks for the multi-label classification task. In this work, object label dependencies are transformed with graph transformer networks to actively distribute gradient information among multiple sub-networks of labels for distinguishable representation learning of visual data.
- The study investigates several strategies for integrating semantic and network properties of object labels, including label embeddings and Eigenvector-based enhancement, to better support the multi-label classification task.
- We evaluate our method with comprehensive experiments on benchmarking datasets, including Microsoft COCO (MS-COCO) and Fashion550K. The experiment results show significant mAP improvements of 9.7% on MS-COCO and 6.4% on Fashion550K compared to the baselines. MGTN outperforms the most recent state-of-the-art models by the increment of 0.4% and 3.7% in mAP on MS-COCO and Fashion550K, respectively.

The structure of the paper is as follows. Firstly, we review the recent studies for multi-label classification in related work section. Secondly, the approach section describes our proposed MGTN framework with multiple optimisation strategies in great details. In our experiments, the new state-of-the-art results are demonstrated. Lastly, we conclude our paper with findings and contributions in the final section.

Related Work

Modelling visual data with their associated labels have drawn great research interest in machine learning and computer vision communities. Multi-tag appears to be a typical property of Internet media; thus, multi-label classification is a fundamental task with many real-world applications (Chen et al. 2019a; Ge, Yang, and Yu 2018; Yeh et al. 2017). Early approaches were derived from single-label multi-class classification, which decomposed the multi-label classification tasks into multiple sub-problems for learning. Tsoumakas and Katakis (2007) synthesised the multi-label nature of datasets and suggested the use of multiple binary classifiers. Their approach, however, completely ignored the interrelationships among various labels in visual data. Gong et al. (2013) investigated a number of multi-label loss functions for training convolutional neural networks, which catered for the deviation between multiple predicted labels and the ground truth. Nevertheless, label co-occurrence dependencies were analysed as essential in multi-label classification problems (Xue et al. 2011). Wang et al. (2016) proposed a unified framework to model the label dependencies explicitly. In their experiments, visual features were adapted based on the previous prediction outcomes by encoding attention models in an integral CNN-RNN framework. As a result, their probabilistic approach gained a performance boost on recognising smaller objects after learning the dominant ones; however, its training was not without high computational costs and scalability issues.

To exploit label dependencies, many existing works proposed semi-supervised learning using graph representations for multi-label classification. Kipf and Welling (2017) encoded graph structures using neural networks, or Graph Convolutional Networks (GCN), to learn representations for efficient information propagation on multiple labels. Chen et al. (2019b) adopted this spectral graph convolution approach to capture object label correlations for recognising multiple objects in images. Prior knowledge such as semantic label embeddings and data-driven adjacency matrix were employed to learn inter-dependent object classifiers. Instead of using the correlation matrix, Li et al. (2019) constructed it via a plug-and-play label graph module, which takes label embeddings as input. The module was further enhanced with a L1-norm regularisation of the inferred matrix and the identity matrix to avoid the over-smoothing problem on nodes' features. Likewise, Wang et al. (2020b) proposed a novel label graph superimposing framework. The framework firstly transforms the statistical graph (i.e., correlation matrix) into a superimposing label graph by integrating with a knowledge graph (e.g., ConceptNet of Speer, Chin, and Havasi (2017)). The superimposed graph was fed into a multi-layer graph convolution layer to learn the label correlation representation, which was later injected with CNN features to generate label predictions. These mentioned works are considered as our baselines in the experiment section.

Approach

Multi-label image classification entails learning of visual and topological information of inter-correlated objects. Although visual representations play a major role, there are sub-optimal learning outcomes for objects with fewer observations and limited visual details in the dataset. The semantic and topological structures of objects and their labels, therefore, furnish ancillary knowledge to surpass these limitations. Integrating structural properties into deep neural networks strengthens the learning capability of image recognition. In this work, we develop modular graph transformer networks to enhance the information propagation and representation learning for multi-labelled visual data.

This paper proposes a framework with the use of graph transformer and convolution layers to classify visual data with multiple backbones in a modular way, as shown in Figure 2. It is semi-supervised multi-label learning, which provides the control of information propagation with interconnected label information. We employ the concept of divide and conquer in model development, where each backbone is responsible for learning a representation of a set of objects. Such representation yields ample and detailed signals on different sets of object details. The combination of multiple bare backbone units lead to better performance than a single complex backbone. We validate our proposed approach on multiple public datasets, including MS-COCO and Fashion550K, to illustrate the effectiveness in comparison with existing state-of-the-art algorithms.

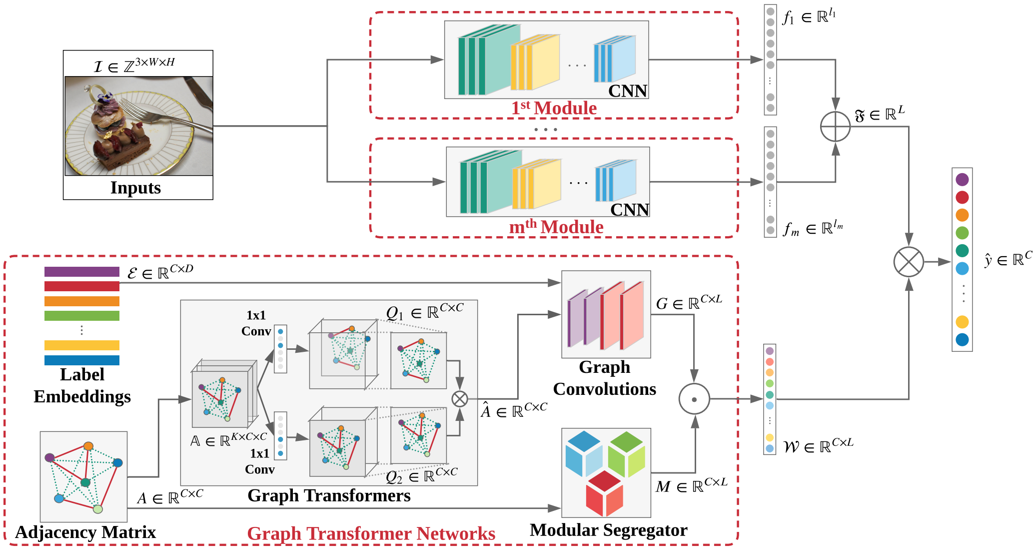

Figure 2: Modular Graph Transformer Network (MGTN) supports multi-label learning over multiple modules of CNNs for recognising object labels in images. The framework has configurable building blocks to integrate semantic information $\mathcal{E}$ and topological information A into visual representation learning. MGTN enables information propagation over multiple sub-graphs guided by graph transformer networks with a modularity-based segregator for highly effective learning of visual data.

Preliminaries

Real-world objects typically are associated with one or more labels, which appear to be in correlated patterns within the data observations. We denote the dataset $\mathcal{D}$ which consists of images and their corresponding labels. $\mathcal{I}$ is defined as an input tensor with the dimension of $W \times H$ and 3-channel RGB. The objective is to assign multiple labels out of C object classes to a single input. In this work, we use the multi-label classification loss for the optimisation task without explicit regularisation:

$$\mathcal{L} = -\frac{1}{C}\sum_{c=1}^{C} y_c \log(\sigma(\hat{y}_c)) + (1 - y_c)\log(1 - \sigma(\hat{y}_c)) \quad (1)$$

where $\sigma(\cdot)$ is known as the sigmoid function.

The inter-dependencies of object labels can be integrated as knowledge to guide information propagation. Moreover, we define the knowledge graph G based on the topological structure of inter-connected object labels discovered in the data sets. Specifically, $G = (V, E, A)$, where V denotes the object labels, and E is the set of edges with the adjacency matrix A. We aim to attune every aspect of the graph to provide compelling results by unfolding the graph transformer and convolutional networks.

Multi-label classification is performed based on the dyadic architecture: CNNs for learning the image-level representation f and graph neural networks for discovering the classifier mapping W. It allows the use of label-level word embedding $\mathcal{E}$ and topological information A to support visual recognition via stacked graph convolution layers. As a result, the final predicted scores are computed as $\hat{y} = Wf$.

Modular Graph Transformer Networks

The integration of the topological information helps to reduce the uncertainty in multi-label learning (Chen et al. 2019b; Wang et al. 2020b). However, the existing approaches tend to strictly favour node pairs with strong relationships, thereby leading to the low diversity in predicting label combinations. Inspired by Graph Transformer Networks (Yun et al. 2019), we propose a more flexible way to leverage label correlations in the matrix for the multi-label classification task on visual data.

Referring to Chen et al. (2019b), we compute the probability matrix P as $P_{ij} = \varrho * A_{ij}/\mathcal{d}_i$, where $\mathcal{d}_i = \sum_k A_{i,k}$ is the degree matrix and ϱ is 0.25. With the objective of removing weak connection edges, i.e., noisy signals, previous works apply a single cut-off threshold in the normalisation of the adjacency matrix. It may cause indistinguishable representations due to the elimination of values below the threshold. Hence, we propose the use of multiple K real-value thresholds denoted as $\mathcal{T} = [t_1, t_2, ..., t_K]$, in which $t_i \in [0, 1]$ and $t_i < t_j \; \forall i < j$.

The adjacency tensor $\mathbb{A} \in \mathbb{R}^{K\times C\times C}$ consists of $\{\mathbb{A}_k \in \mathbb{R}^{C\times C}\}, k = \{1, ..., K\}$. We set $\mathbb{A}_1$ as the identity matrix I, and for all $k \geq 2$, we have:

$$\mathbb{A}_{kij} = \begin{cases} 1 & \text{if } P_{ij} \in [t_{k-1}, t_k), i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Similar to Yun et al. (2019), the two softly chosen adjacency matrices $Q_1, Q_2 \in \mathbb{R}^{C\times C}$ are inferred via two 1×1 convolutions as follows:

$$Q_1 = \phi(\mathbb{A}, \text{softmax}(W^1_\phi)) \text{ and } Q_2 = \phi(\mathbb{A}, \text{softmax}(W^2_\phi)) \quad (3)$$

where ϕ is the convolution layer, and $W^1_\phi, W^2_\phi \in \mathbb{R}^{1\times 1\times K}$ are parameters to be learned. The final transformed adjacency matrix $\hat{A} \in \mathbb{R}^{C\times C}$ is by $\hat{A} = \eta(Q_1 Q_2 + I)$, where $\eta(A) = \mathcal{d}^{\frac{-1}{2}} A \mathcal{d}^{\frac{-1}{2}}$ is the matrix normalisation method as Kipf and Welling (2017).

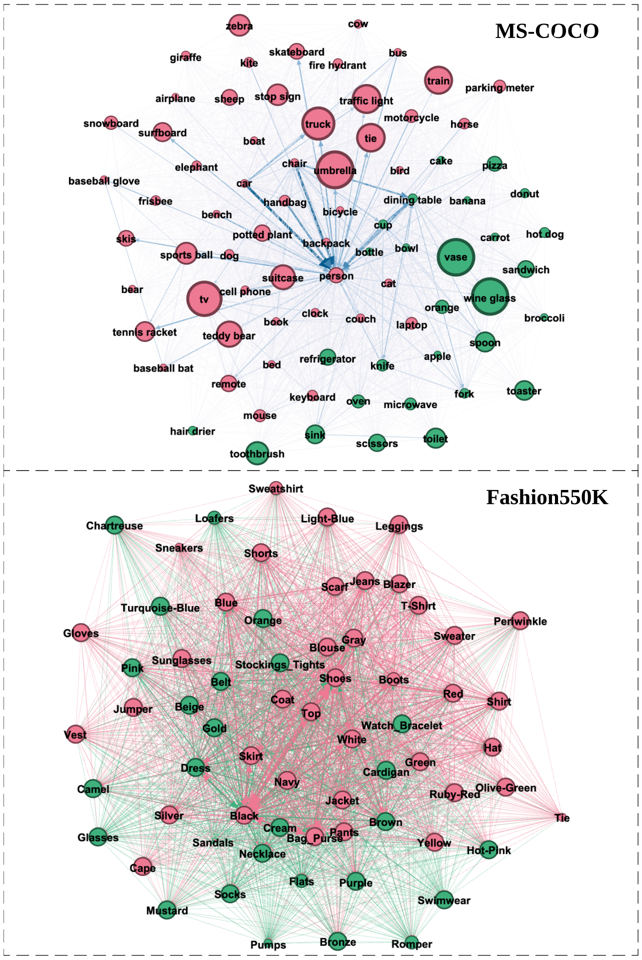

Figure 3: Network Analyses on MS-COCO and Fashion550K reveal the partitions of inter-connected object labels. Both datasets consist of two node communities highlighted in *red* and *green* using the Clauset-Newman-Moore agglomeration algorithm. The sizes of the nodes reflect the relative importance of inter-dependent object labels based on the eigenvector centrality measure.

Furthermore, we decompose the graph learning networks into multiple sub-units, i.e., modules, for recognising different highly inter-connected sets of objects. The breakdown of these partitions may reveal a-priori unknown knowledge structures of objects in visual data, thereby leading to better learning of their representations. By segregating the propagation of unfolded sub-networks, this approach aims to improve classification performance as well as to reduce over-fitting towards unpopular object labels due to their nature of appearances and co-occurrences. Therefore, V can be divided into multiple modules of vertices, i.e., $V = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_m\}$, where m is the number of modules, $\mathcal{V}_k$ is a set of objects belong to the sub-graph k. The learning of each $\mathcal{V}_k$ can be done in dynamic configurations with multiple architectural backbones.

The approach entails the discovery and analysis of highly inter-connected structures in a network; therefore, a hierarchical agglomeration algorithm (Clauset, Newman, and Moore 2004) is employed to uncover sub-graphs of the network in a unsupervised manner. It is based on the modularity $\mathfrak{Q}$ of a graph which is computed as the following.

$$\mathfrak{Q} = \frac{1}{2\mathfrak{m}} \sum_{i,j} \left[A_{i,j} - \frac{\mathscr{d}_i \mathscr{d}_j}{2\mathfrak{m}} \right] \delta(c_i, c_j) \tag{4}$$

where $\mathfrak{m} = \frac{1}{2} \sum_{i,j} A_{i,j}$, $\mathscr{d}_i = \sum_k A_{i,k}$ is the degree matrix, and $\delta(u, v)$ is 1 if $u = v$ otherwise 0.

The community detection begins with each node, or object label, in its own partition and continuously joins different partitions in order to maximise the modularity score $\mathfrak{Q}$ of the sub-graphs. As a result, m sub-graphs can be discovered and we derive the sub-graph assignment $S \in \mathbb{Z}^C$, in which $S_i = p$ with the partition number p. Figure 3 illustrates the segregation of object labels on MS-COCO and Fashion550K datasets, in which multiple labels are visually manifested in a coordinated and meaningful manner.

Our approach employs multiple CNNs with configurable backbones for the m sub-graphs. Each CNN module aims to learn the visual representations of each highly inter-connected set of object classes. The image-level representation, denoted as $f_p \in \mathbb{R}^{l_p}$, has l_p number of features which are then concentrated into a long feature $\mathfrak{F} \in \mathbb{R}^L$.

The integration of segregated learning happens with gradient distribution based on graph convolutions, in which classifier mappings are deployed to divide and conquer information propagation into the multiple CNNs. We define a control tensor $M \in \mathbb{R}^{C \times L}$ with a threshold τ as the following.

$$M_{iv} = \begin{cases} \tau & \text{if } S_i = p \text{ and } v \in f_p \\ \frac{1-\tau}{m-1} & \text{otherwise} \end{cases} \tag{5}$$

The threshold $\tau \in [0, 1]$ provides MGTN with a way to manipulate information sharing among multiple sub-graphs. The classifier mappings are computed as $\mathscr{W} = G \odot M$ before inferring the learning prediction scores as $\hat{y} = \mathscr{W}\mathfrak{F}$. MGTN leverages on the concept of decomposing multiple sub-graphs first, then linking and combining them to form a complete learning framework for multi-label classification.

Eigenvector-based Embedding Transformation

This paper further examines a fine tuning strategy based on the connections among object labels in the network, where not all connections are equal (Bonacich 1987). We employ the concept of eigenvector-based transformation (EV) to enhance the computation of graph convolution networks. It aims to model and strengthen the learning of the complex relationships of object labels with their importance rankings. In Figure 3, the sizes of the nodes reflect the importance of the object labels, which may enhance crucial signals on their inter-relational dependencies. Instead of regularisation of the loss function, we propose the transformation of label embeddings as pre-convolutional graph processing to adjust for their relative importance in learning.

We define the eigenvector centrality $\mathcal{C}_i$ of the label i by:

$$\mathcal{C}_i = \frac{1}{\lambda} \sum_k a_{k,i} \, \mathcal{C}_k \tag{6}$$

where $\lambda \neq 0$ is the largest eigenvalue. As the result of Perron-Frobenius theorem, the eigenvector centrality $\mathcal{C}_i$ can be found as unique and positive if the graph is connected. We implement the calculation using the power iteration method, in which the $\mathcal{C}^{(k)} = \mathcal{C}^{(k-1)}A$ is repeatedly computed for $k \geq 1$. The solution is then normalised with the signed component of maximal magnitude $m(x)$ as $\mathcal{C}^{(k)} = \mathcal{C}^{(k)}/m(\mathcal{C}^{(k)})$. The calculation is stopped after 100 iterations or reaching an error tolerance of $N * 10^{-6}$. The importance matrix, then, is blended into label embeddings to support the multi-label learning process. In this work, the transformed $\mathcal{E}$ ($\mathcal{E} = E \cdot \mathcal{C}^T$) information is then convolved with the use of multiple stacks of GCN units. The graph traversal over multiple layers allows MGTN to learn an optimal embedding-to-classifiers mapping $\mathcal{W}$ with the aggregated length of visual features from multiple CNNs.

Language Embeddings

Label information plays an important role in reinforcing the learning capabilities of GCN. Based on the pre-trained language models that one uses to extract label-level word embeddings, the label information may impact differently to the initial point of the model in the optimisation space. Here we explore the use of two types of pre-trained embeddings including (1) static embeddings (e.g., FastText) and (2) contextual embeddings (e.g., BERT). Regarding contextual embeddings, instead of using word-level pre-trained embeddings, one can use character, sub-words, or byte-pair-encoding (Sennrich, Haddow, and Birch 2016) based pre-trained language models to capture contextual information. To name a few, BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019) are helping many language related tasks to achieve new state-of-the-art results. FastText and GloVe were tested by Chen et al. (2019b), Char2Vec and BERT were experimented by Vu et al. (2020). Here we investigate more into these following language models:

- **Char2Vec** (Kim et al. 2015) is a deep language model that learns at character-level inputs. Similar to Vu et al. (2020), the Char2Vec model was trained on English Wikipedia corpus with embedding dimension of 300.
- **BERT** (Devlin et al. 2018) learns contextual relations between words (or sub-words) in a text using by Transformer (Vaswani et al. 2017). Here, we use BERT_Base (12 layers) to get the label embeddings. For a given label, we average all vectors of its subwords from the last layer provided by Akbik, Blythe, and Vollgraf (2018), hereafter BERT$_{avg_last}$.
- **RoBERTa** (Liu et al. 2019) is a new improved language model based on BERT with improved training methodology, such as they removed next sentence prediction task from BERT's model and replaced by dynamic masking. Since RoBERTa is a better version of BERT, therefore, we seek to test how extracted information from different pre-trained layers works on a downstream task among Transformer's variants. We use an average vector of 12 layers in RoBERTa$_{base}$'s pre-trained model provided by Akbik, Blythe, and Vollgraf (2018) to extract label embeddings for the task, hereafter RoBERTa$_{avg_12}$.

Experiments

This section describes our implementation details and benchmarking metrics. Experiments are exhaustively conducted, and we report the relevant empirical results on two public datasets: MS-COCO and Fashion550K.

Experimental Procedure

With the objective of providing a fair comparison to the current state-of-the-art models (e.g., ML.GCN (Chen et al. 2019b), A-GCN (Li et al. 2019), and KSSNet (Wang et al. 2020b)); we select MS-COCO (Lin et al. 2014) and Fashion550K (Inoue et al. 2017) datasets for evaluation.

- **MS-COCO** (Lin et al. 2014) is the most popular multi-label image dataset. It has several main features: object segmentation, recognition in context, five captions per image among others. In total, it contains 2.5M labelled object instances in 328K images, in which 82,783 training, 40,504 validation, and 40,775 test images.
- **Fashion550K** (Inoue et al. 2017) is a multi-label fashion dataset. It contains 66 unique weakly-annotated tags with 407,772 images. These images are called as noisy-labelled data since it was created with minimal human supervision. Moreover, a *clean* collection was manually verified to improve the task with cleaning neural networks in their noisy+clean dataset. This *clean* set has 3K, 300, 2K images for training, validation, and testing respectively.

Implementation. Our proposed MGTN framework is developed using PyTorch (version 1.3.1). The segregation of learning with any number of sub-networks of object labels is fully implemented. We utilise the NetworkX library to investigate the community structure using the Clauset-Newman-Moore greedy modularity maximisation in multiple runs. Based on our network analyses shown in Figure 3, both MS-COCO and Fashion550K are consistently segregated into two sub-graphs for multi-label learning. We employ dual ResNeXt-50 32x4d backbones (Xie et al. 2017) for visual feature extraction with a semi-weakly supervised pre-trained model on ImageNet (Yalniz et al. 2019). The concentration of visual presentations amounts to a tensor $\mathfrak{F}$ of 2×2048 features.

We configure our model with two GCN layers and the output dimensionality of 2048 and 4096 to match our dual backbones. We employ the threshold τ is 0.999 in the Eq(5) to manipulate the information sharing in our gradient distribution. For the graph transformer layer, without otherwise stated, we set $\mathcal{T} = [0.2, 0.4, 1.0]$ for MS-COCO and $\mathcal{T} = [0.1, 0.3, 1.0]$ for Fashion550K. The negative slope of 0.2, which is similar to Chen et al. (2019b), is set for image representation learning using LeakyReLU (Maas, Hannun, and Ng 2013) as the non-linear activation function.

For label embeddings, we explore different language models and GloVe (Pennington, Socher, and Manning 2014) is chosen to assure the reproducibility of our results for future comparison. Our data augmentation during training process is similar to Chen et al.; Wang et al. (2019b; 2020b), in which we resize images to 512×512 and randomly crop regions of 448×448 with random horizontal flips. We adopt

Method	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN (Wang et al. 2016)	61.2	-	-	-	-	-	-
SRN (Zhu et al. 2017)	77.1	81.6	65.4	71.2	82.7	69.9	75.8
Baseline(ResNet101) (He et al. 2016)	77.3	80.2	66.7	72.8	83.9	70.8	76.8
Multi-Evidence (Ge, Yang, and Yu 2018)	–	80.4	70.2	74.9	85.2	72.5	78.4
ML-GCN (Chen et al. 2019b)	82.4	84.4	71.4	77.4	85.8	74.5	79.8
A-GCN (Li et al. 2019)	83.1	84.7	72.3	78.0	85.6	75.5	80.3
KSSNet (Wang et al. 2020b)	83.7	84.6	73.2	77.2	87.8	76.2	81.5
ML-GCN (ResNeXt50 swsl)	86.2	85.8	77.3	81.3	86.2	79.7	82.8
SGTN (Our) (Vu et al. 2020)	86.6	77.2	**82.2**	79.6	76.0	**82.6**	79.2
MGTN(Base)	86.9	**89.4**	74.5	81.3	**90.9**	76.3	83.0
MGTN(Final)	**87.0**	86.1	77.9	**81.8**	87.7	79.4	**83.4**

Table 1: Performance comparisons on MS-COCO. Our MGTN outperforms all previous approaches with large margins.

Method	mAP
Baseline(ResNet50) (Inoue et al. 2017)	58.68
StyleNet (Simo-Serra and Ishikawa 2016)	53.24
ML-GCN (Chen et al. 2019b)	60.85
A-GCN (Li et al. 2019)	61.35
MGTN(Final)	**65.10**

Table 2: Performance comparisons on Fashion550K. MGTN's models are selected based on the best pre-trained weights on the validation set. Then the final performance is reported based on the test set. For other metrics, MGTN archived 77.7, 35.16, 48.42, 81.36, 41.24, 54.74 for CP, CR, CF1, OP, OR, and OF1 accordingly.

SGD as the optimiser with the momentum is set to be 0.9. Weight decay is 10^{-4}. The initial learning rate is 0.03 and 0.01 for without and with EV-enhancement label embeddings, respectively. The learning rate decays by a factor of 10 for every 20 epochs, and the network is trained for 60 epochs in total. The experiments were run on two Nvidia Tesla V100, each card has 16GB memory.

Evaluation metrics. We evaluate mAP - mean average precision, CP - average per-class precision, CR - average per-class recall, CF1 - average per-class F1 score, OP - overall precision, OR - overall recall, and OF1 - overall F1 score for benchmarking with baseline models (Chen et al. 2019b).

Experiment Results

In this sub-section, we present our comparisons with the existing state-of-the-arts on MS-COCO and Fashion550K respectively to demonstrate the effectiveness of our proposed approach for the multi-label classification task.

Results on MS-COCO. We compare several configurations of MGTN with recent state-of-the-arts including the baseline of ResNet101 (He et al. 2016), CNN-RNN (Wang et al. 2016), SRN (Zhu et al. 2017), ML-GCN (Chen et al. 2019b), A-GCN (Li et al. 2019), KSSNet (Wang et al. 2020b), ML-GCN (ResNeXT50), and SGTN (Vu et al. 2020). In Table 1, we report our quantitative results based on the graph transformer networks MGTN (Base) and the fine-tuned model with eigenvector-based transformation MGTN (Final). We observed three insights. Firstly, MGTN outperforms different baselines that do not use GTN in their architectures. Specifically, MGTN shows significant mAP improvements of 9.4% from the ResNet101 baseline and 3% from KSSNet. In our final model, the experiment results establish new state-of-the-art with substantial improvements of 9.7%, 4.58%, and 3.3% in mAP compared to the baseline (ResNet101), ML-GCN, and KSSNet, respectively. Secondly, MGTN gets better performance in comparison to SGTN (Vu et al. 2020), which incorporated GTN in its learn-

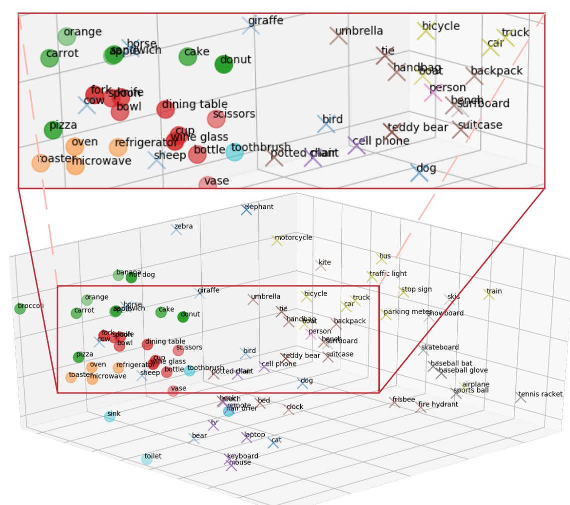

Figure 4: 3D t-SNE visualisation of MGTN's predicted results on the test set of MS-COCO. Each point is a label of MS-COCO dataset. Identical shape, either a 'circle' (○) or a 'multiplier'(×), manifests two labels belong to the same sub network based on modularity. For colourful fading variants, similar level of fading illustrates that two shapes are semantically closed and might belong to the same super-category.

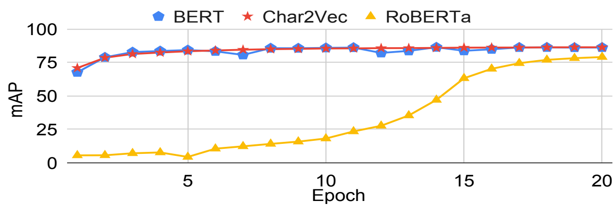

Figure 5: Learning patterns of MGTN with different label embeddings in 20 epochs. The MGTN model using RoBERTa$_{avg_12}$ label embedding shows a slow learning speed in comparison to others.

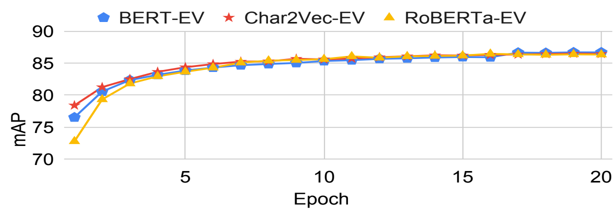

Figure 6: The EV-enhancement for label embedding helps the MGTN's model learn faster, even MGTN with the setting using the RoBERTa$_{avg_12}$ now learns faster. Note that y-axis here is ranged in $[70, 100]$ for visibility.

ing model. Lastly, we experimented with ML-GCN using ResNeXt50 (Xie et al. 2017) as a visual backbone. MGTN shows better performances for all metrics, which again confirms the effectiveness of the novel *modular graph transformer networks*.

To explore the predicted outputs on the test set of MS-COCO. We employed t-SNE (van der Maaten and Hinton 2008) to visualise the modularity of the outputs as in Figure 4. We can intuitively analyse how good MGTN understands the correlation between labels on unseen images. There are only two shapes which mean MGTN learnt the modularity information (i.e., two modularities) from the training data and predicted that information on unseen data. Similar to the motivation example in Figure 1, *'person, chair, umbrella'* are in one modularity, *'cup, bowl, dining table'* are in another modularity. Moreover, MGTN understands the correlation information between labels by resulting *'sink'* and *'toilet'* in the same colour, which means they are very closed as well.

In the same way, *'car, truck, bus, traffic light, stop sign'* stay closed to each other and have the same colour since they fall under transportation. Generally, MGTN could learn and predict both modularity information as well as understanding the label correlations to improve the muti-label classification performance.

Results on Fashion550K. We compare the final model of MGTN with state-of-the-art models on Fashion550K dataset including the baseline of ResNet50 (Inoue et al. 2017), StyleNet (Simo-Serra and Ishikawa 2016), ML-GCN (Chen et al. 2019b), and A-GCN (Li et al. 2019). The learning capabilities of our approach are asserted on the noisy dataset; because manual verification and cleaning neural networks introduced in the noisy+clean dataset may not reflect the true impact of our assessment. Also, the use of MGTN on noisy data is already shown to be superior to the fine-tuned model with clean labels in Inoue et al. (2017). The experiment results demonstrate the effectiveness of MGTN with significant improvements of 6.4%, 4.2%, and 3.7% in mAP from the baseline (ResNet50), ML-GCN, and A-GCN, respectively.

Ablation Study on Label Embeddings

We attempt answer two questions on the MS-COCO dataset: (1) do different ways of extracting label embeddings affect the final performance of MGTN? and (2) how EV-enhancement on label embeddings affect the learning?

Label Embeddings and Learning Patterns. This study tests with different language models be used to extract label representations. Figure 5 shows that different label embeddings affect to the learning speed of the downstream task significantly. Generally, this result is consistent to Chen et al. (2019b) in the sense that, after a certain number of training epochs, the model would achieve performance similarly. For RoBERTa$_{avg_12}$ label embedding, it was averaged from 12 layers, therefore, the label representation was not closed to actual meaning of word-level representation. Thus, its learning progress was almost started from scratch. This result suggests that, for label representation, it is probably better to use information from a few last layers (e.g., the last layer as in the BERT$_{avg_last}$ setting) for this task. In summary, this ablation shows that, with different ways of extracting these label embeddings, one could get "two birds with one stone" - i.e., saving compute power and getting high performances at the same time.

Effects of EV-Enhancement on Different Label Embeddings. The goal of this ablation study is to address that, with EV-enhancement, the MGTN model could even learn faster and hence, save more computing power. More importantly, the effects are consistent across all tested label embeddings. Figure 6 shows that EV-enhancement helps MGTN with different label embeddings learn faster and achieve optimal performance in less than 20 epochs.

Conclusion

This paper presents an end-to-end framework, named Modular Graph Transformer Networks (MGTN), to solve the multi-label classification task on visual data. The framework integrates multiple CNN backbones on unfolded subnetworks that are segregated from the original one based on the graph modularity. Additionally, it also exploits topological and semantic properties among labels via the graph transformer and eigenvector-based embedding layers respectively to enhance the label correlation representation in GCN. This work unveils new opportunities to surpass the limitations of single backbone systems for better learning of network niches and reduced overfitting potentials. Extensive experiments on two benchmark datasets manifest the advantages of MGTN via significant improvements against state-of-the-art algorithms in classifying images with multiple labels.

References

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649.

Bi, W.; and Kwok, J. T. 2011. Multi-label classification on tree-and dag-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 17–24.

Bonacich, P. 1987. Power and centrality: A family of measures. *American journal of sociology* 92(5): 1170–1182.

Chen, H.; Miao, S.; Xu, D.; Hager, G. D.; Harrison, A. P.; and Com, A. P. H. 2019a. Deep Hierarchical Multi-label Classification of Chest X-ray Images. *Proceedings of Machine Learning Research* 102: 109–120. URL http://proceedings.mlr.press/v102/chen19a/chen19a.pdf.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019b. Multi-Label Image Recognition with Graph Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5177–5186. URL http://arxiv.org/abs/1904.03582.

Clauset, A.; Newman, M. E.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 70(6): 6. ISSN 1063651X. doi:10.1103/PhysRevE.70.066111.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* .

Ge, W.; Yang, S.; and Yu, Y. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1277–1286.

Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; and Ioffe, S. 2013. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Inoue, N.; Simo-Serra, E.; Yamasaki, T.; and Ishikawa, H. 2017. Multi-Label Fashion Image Classification with Minimal Human Supervision. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*.

Kim, Y.; Jernite, Y.; Sontag, D.; and Rush, A. M. 2015. Character-Aware Neural Language Models. *CoRR* abs/1508.06615. URL http://arxiv.org/abs/1508.06615.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–14.

Li, Q.; Peng, X.; Qiao, Y.; and Peng, Q. 2019. Learning Category Correlations for Multi-label Image Recognition with Graph Networks. *arXiv preprint arXiv:1909.13005* URL http://arxiv.org/abs/1909.13005.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692. URL http://arxiv.org/abs/1907.11692.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30/1, 3.

Nam, J.; Kim, Y.-B.; Mencia, E. L.; Park, S.; Sarikaya, R.; and Fürnkranz, J. 2019. Learning Context-dependent Label Permutations for Multi-label Classification. In *International Conference on Machine Learning*, 4733–4742.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. URL http://www.aclweb.org/anthology/D14-1162.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-1162.

Simo-Serra, E.; and Ishikawa, H. 2016. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 298–307.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451. AAAI Press.

Tsoumakas, G.; and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3(3): 1–13.

van der Maaten, L.; and Hinton, G. E. 2008. Visualizing Data using t-SNE. In *Journal of Machine Learning Research*, 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vu, X.-S.; Le, D.-T.; Edlund, C.; Jiang, L.; and Nguyen, H. D. 2020. Privacy-Preserving Visual Content Tagging using Graph Transformer Networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2299–2307.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2285–2294. ISSN 00214671. URL http://arxiv.org/abs/1604.04573.

Wang, L.; Liu, Y.; Qin, C.; Sun, G.; and Fu, Y. 2020a. Dual Relation Semi-Supervised Multi-Label Learning. In *AAAI*, 6227–6234.

Wang, Y.; He, D.; Li, F.; Long, X.; Zhou, Z.; Ma, J.; and Wen, S. 2020b. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34/07, 12265–12272.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xue, X.; Zhang, W.; Zhang, J.; Wu, B.; Fan, J.; and Lu, Y. 2011. Correlative multi-label multi-instance image annotation. In *2011 International Conference on Computer Vision*, 651–658. IEEE.

Yalniz, I. Z.; Jégou, H.; Chen, K.; Paluri, M.; and Mahajan, D. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* URL http://arxiv.org/abs/1905.00546.

Yeh, C.-K.; Wu, W.-C.; Ko, W.-J.; and Wang, Y.-C. F. 2017. Learning deep latent space for multi-label classification. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yun, S.; Jeong, M.; Kim, R.; Kang, J.; and Kim, H. J. 2019. Graph Transformer Networks. In *Advances in Neural Information Processing Systems*, 11960–11970.

Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; and Wang, X. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5513–5522.