Mind Your Neighbours: Image Annotation with Metadata Neighbourhood Graph Co-Attention Networks

Junjie Zhang[1,2] Qi Wu[3,*] Jian Zhang[1] Chunhua Shen[3] Jianfeng Lu[2]
[1]University of Technology Sydney, Australia [2]Nanjing University of Science and Technology, China
[3]The University of Adelaide, Australia
{junjie.zhang@student., jian.zhang@}uts.edu.au lujf@njust.edu.cn
{qi.wu01, chunhua.shen}@adelaide.edu.au

Abstract

As the visual reflections of our daily lives, images are frequently shared on the social network, which generates the abundant 'metadata' that records user interactions with images. Due to the diverse contents and complex styles, some images can be challenging to recognise when neglecting the context. Images with the similar metadata, such as 'relevant topics and textual descriptions', 'common friends of users' and 'nearby locations', form a neighbourhood for each image, which can be used to assist the annotation. In this paper, we propose a Metadata Neighbourhood Graph Co-Attention Network (MangoNet) to model the correlations between each target image and its neighbours. To accurately capture the visual clues from the neighbourhood, a co-attention mechanism is introduced to embed the target image and its neighbours as graph nodes, while the graph edges capture the node pair correlations. By reasoning on the neighbourhood graph, we obtain the graph representation to help annotate the target image. Experimental results on three benchmark datasets indicate that our proposed model achieves the best performance compared to the state-of-the-art methods.

1. Introduction

With the rise of the social network, people like to capture vivid moments and share them on the internet. These images are generated and spread at an explosive pace, which yields the urgent need for an efficient annotation method to assist users to understand and retrieve images. Significant progresses have been made on the image annotation by uncovering relationships between image pixel contents and labels, such as the classification [3, 4], clustering [8, 32], and graph inference [16, 27]. Most recently, deep neural networks [9, 24] have demonstrated advanced abilities of the image feature learning, which inspire various network-based models [29, 31]. These models treat the individual image as an independent object and focus on solving the annotation without the context information. However, due to the diverse contents and complex styles, some images are still difficult to annotate on their own.

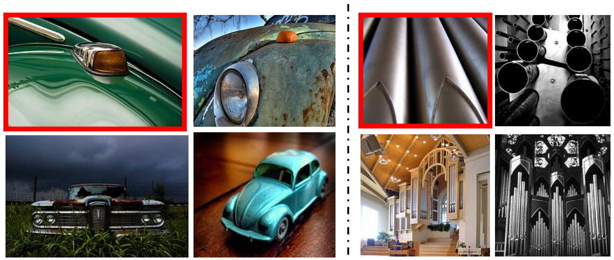

Figure 1. For the target images with the red boxes, they are hard to recognise on their own. However, in the context of the neighbours with the similar metadata, such as 'vehicle, vintage, Beetle' and 'church, instrument, art', it is more clear that the target images are a car and a pipe organ. Based on this motivation, we propose a neighbourhood graph as *'neighbourhood watch'* to assist the image annotation.

Social networks like the Flickr, Instagram and Facebook record the user interactions with images as the vast amount of the metadata, which is presented in various forms. The most common metadata includes the collections, *i.e.* image groups created by users; textual descriptions, *i.e.* tags and captions; as well as user profiles, *i.e.* user-names, locations and friends. As a means to communicate with other users, these metadata can be as informative as visual pixels [12, 22] to understand images. See Fig. 1 as an example, images are hard to be recognised and annotated without seeing its metadata related images. There are several lines of works [1, 25] conducted by utilising the metadata to assist the annotation, where different types of metadata are

[*]Corresponding Author

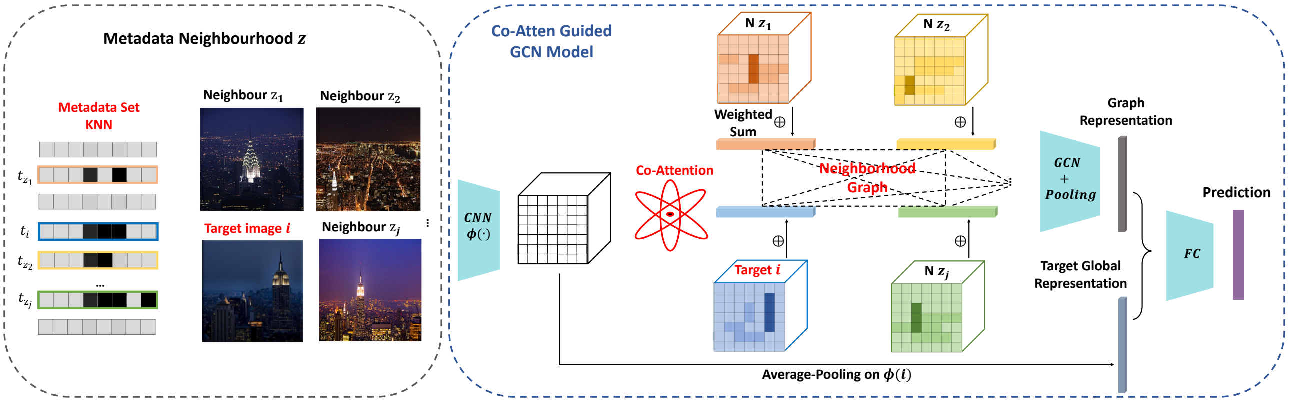

Figure 2. The framework of the proposed model. The neighbourhood z of the target image i is decided by measuring their metadata similarities. Then we establish the neighbourhood graph using image representations as nodes and correlations as edges. To accurately harvest visual clues from its neighbours, we introduce a co-attention mechanism to guide the Graph Convolutional Network (GCN) and obtain the graph representation, which is then concatenated with the target global feature to generate the label confidence.

studied to be embedded into the annotation framework. In [12], authors propose to non-parametrically use the metadata to generate image neighbours and train an annotation model based on the visual features from the target image and its neighbours. However, the features from image neighbours are independently embedded from each other, and only global features are considered.

In this paper, we address the image annotation problem by routinely checking its neighbours in a graph, which is constructed by the equipped meta information of the image. The whole framework is shown in Fig. 2. Since the metadata explicitly or implicitly indicates the connections between images, for example, semantically similar textual descriptions like tags and captions usually associated with images that have similar visual appearances [8], friends who share same interests have the high probabilities of following images with similar topics [26], and landmark photos are always taken at the fixed locations [17]. We locate the image neighbours by measuring the similarities among their metadata. Then we establish a graph network to model the correlations between the target image and its neighbours. The whole neighbourhood is represented as a graph, where each node is the corresponding image feature. The graph edges indicate the correlations between node pairs. Considering the diverse visual appearances of neighbours, different attention should be paid according to its content. Therefore, we introduce a co-attention mechanism to obtain the node representation. That is, we obtain the co-attention maps by successively switching the attention between the target image and its neighbours. Given the graph structure, we can perform reasoning on the graph and infer the representation by applying the graph convolutional operations. The graph representation is finally concatenated with the target global feature to predict the label confidence, which is intuitive since we want to capture the connections among image neighbours to assist the annotation of the target image. We name our proposed model as **Meta**data **n**eighbourhood **g**raph **co**-attention **net**work (MangoNet).

In summary, the main contributions of our method are as follows:

1) We propose to annotate images by exploring their neighbourhoods, which are allocated by the metadata similarity. A neighbourhood graph network is established to model the correlations between each image and its neighbours. The learned graph representation is used to assist the annotation of the target image.

2) To accurately capture the relevant regions in each image neighbour that are beneficial for understanding the target image, we introduce a graph co-attention mechanism to obtain the node representations in the graph.

3) We evaluate our proposed method on three benchmark datasets. Our model achieves state-of-the-art performances on all of them.

2. Related Works

Image annotation as a traditional vision task has been extensively studied for decades. Given a training set of images with manually annotated labels, early works are conducted by leveraging the pixel contents using hand-crafted features [20]. The classification-based methods [3, 4] represent each label as an independent class and train classifiers separately, while the voting-based methods [8, 32] aim at transferring labels from the training set, which is sensitive to the metric used to allocate the neighbours. In addition to modelling the semantic correlations between label pairs, the probabilistic graphical models are employed in [16, 27]. With the development of deep learning methods in recent years, several neural network-based models [6, 29, 31] are proposed to extract the advanced image features and capturing high-order label dependencies. Despite the training samples are collected from the social network, these works focus on tackling the image annotation without the context information.

Image Annotation with Metadata The most commonly used metadata for the image annotation is a set of user provided tags, where a multi-modal representation is learned for the image feature and associated tags. In [1, 7], the CCA and KCCA (Kernel Canonical Correlation Analysis) approaches are adopted to build a latent semantic space, while generative models are obtained in [25] to uncover the multi-modal association. In addition to user tags, there are investigations conducted on other types of metadata. GPS, EXIF and time-stamps are used in [13, 17] to annotate the landmark images, while in [26], friendships are contributed to the label recommendation. In [22], multi-type of textual features and network linkage information are used to construct a CRF-based inference model for the image annotation. Johnson *et al.* propose to allocate image neighbours by non-parametrically exploring the metadata in [12], and incorporate features from the target image and its neighbours to annotate. We adopt a similar setting in our model, however, different from [12], a graph-based solution is proposed to model the image neighbourhood.

Attention Mechanism Instead of using the holistic feature to represent an image, given the multi-label property of the image, the attention mechanisms are applied for the image annotation. In [10, 33], the attention mechanism is adopted to capture the correlations between the image content and associated labels. Different from these works, we not only apply the attention for the target image regions but also use a co-attention mechanism to guide the attentions between the target image and its neighbours.

Graph Neural Network The graph is an optimal representation of the structured information. In a graph, nodes are connected by edges, which indicate the pair-wise relationships between corresponding nodes. In the Graph Neural Network (GNN) model [23], the neighbourhood information is propagated through the graph and the hidden state of each node is updated by the multi-layer perceptrons (MLP), while in [18], a recurrent gating mechanism is adopted to update the graph hidden states and extended to output sequences, noted as the Gated Graph Sequence network (GGNN). In [14], a scalable approach Graph Convolutional Network (GCN) is proposed to learn on the graph-structured data via convolutional operations. In [30], the video classification is studied by modelling the frames as the spatial-time graph and applying the GCN to infer the video category, the region information and time sequences are used to establish the graph edges. In [28], the self-attention mechanism is introduced into the GCN to compute the node representation. Each node is embedded by attending over its neighbours. Different from previous works, we ground the image neighbourhood by leveraging their metadata and apply the GCN to infer on the proposed neighbourhood graph for the image annotation problem, and a novel co-attention mechanism is introduced to model the correlations between the target image and its neighbours.

3. The MangoNet

The key characteristic of our proposed Metadata Neighbourhood Graph Co-Attention Network (MangoNet) is to represent the image neighbourhood as a graph to assist the image annotation. The neighbourhood graph is established via the metadata. A co-attention mechanism is introduced to guide the visual attention between the target image and its neighbours to obtain node representations. The whole framework is shown in Fig. 2.

In the following sections, we first describe how to locate image neighbours by measuring their metadata similarities; then we introduce the architecture of the neighbourhood graph and the node representation updating process by unifying the instance-level and co-attention mechanisms. The training and implementation details are given at last.

3.1. Neighbourhood Graph Co-Attention

3.1.1 Graph Construction

The motivation behind the neighbourhood graph is that a graph structure, where its edges indicate the correlations between image nodes, can competently represent the image neighbourhood. By reasoning on the graph, we aggregate node features as the graph representation to assist the target image annotation. Metadata, as a means of bridges between images, it connects images with each other. We first locate image neighbours by measuring their metadata similarities.

Formally, let I be a set of images, V be the set of manual labels $|V| = C$, and $D = \{(i, v)|i \in I, v \subseteq V\}$ be the image dataset, where each image is associated with a subset of labels. Let T be the vocabulary of the metadata carried by I. Given the vocabulary T, each image i is associated with a subset of $t_i \subseteq T$ metadata. We use the Jaccard metric to measure the similarity between the metadata pair, that is, given two images $i, j \in I$ with $t_i, t_j \subseteq T$, $\varphi(i, j) = |t_i \cap t_j|/|t_i \cup t_j|$, where $\varphi(i, j) \in [0, 1]$. Based on the metadata similarity, the nearest neighbour approach can be applied to locate the neighbours.

Since most of the metadata are generated based on the user behaviours [12], metadata vocabularies can be very large, and the metadata associated with each image can be imprecise at the certain level [15]. To accurately and efficiently locate the neighbours, we conduct the hierarchy

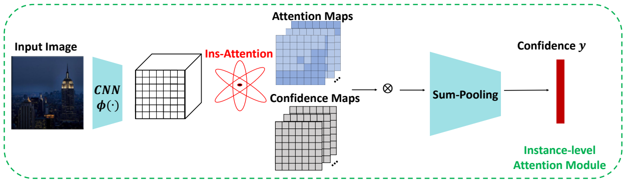

Figure 3. The instance-level attention module we used to capture the regional semantic correspondences between image content and associated labels.

search strategy or the semantic search strategy in the light of circumstances. Specifically, for each image in the large-scale dataset with large metadata vocabulary size like the NUS-WIDE [5], we perform the neighbour search based on the user tags within a sub-image set, which is consist of the images from the collections with similar topics. For the dataset with relatively clean semantic metadata such as the MS-COCO [19] with human-labelled captions, we extract metadata representations of each image as the weighted sum of the semantic representations of visual attributes. We set the search parameters as m neighbours for each image and obtain the candidate neighbourhood $z = \{z_1, ..., z_m\}$ for each image i. The image i and its neighbours z are fully connected with each other to form a neighbourhood graph.

3.1.2 Graph Co-Attention Mechanism

Node Attention Since we intend to obtain visual clues from image neighbours to assist the annotation of the target image, for each neighbour, different attention should be paid according to its content. We adopt a co-attention mechanism to generate attention maps for the target image and its neighbours, *i.e.* 'mind the neighbours'. The co-attention mechanism can be viewed as to learn image visual correlations to contribute to the node representation.

Considering we only have the image-level supervision information *i.e.* annotated ground-truth, we propose to generate the instance-level attention maps regarding individual image first, then combine them as the co-attention maps we desired, which is consistent with our motivation that visual clues from the neighbour are harvested from common semantic classes with the target image, and the target image is revisited with the clues from its neighbourhood. The framework of the instance-level attention module and the proposed co-attention module are shown in Fig. 3 and Fig. 4 respectively.

More specifically, multiple labels associated with an image are always semantically related to different image regions. By referring to [33], we adopt an instance-level attention module to capture this multi-label property for each image. That is, learning attention weights for each label (instance) respectively from the image ground-truth, noted as the InsAtten module. We employ the ResNet-101 [9] as the backbone CNN model to extract the outputs of the convolutional layer as visual features $\phi \in N \times d$, where $N = H \times W$. Given the image i with feature $\phi(i)$ and associated label subset v, the attention weights for each label are generated as follows:

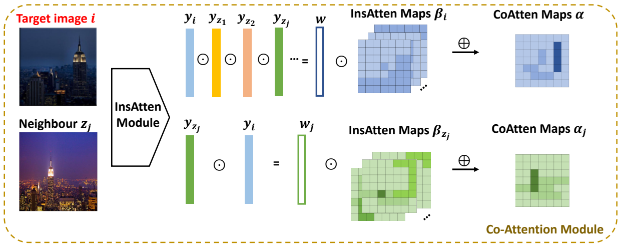

Figure 4. The co-attention attention module. The co-attention maps are the semantic overlap of instance-level attention maps between the target image i and its neighbourhood z.

$$x = q(\phi(i), \eta), \quad x \in \mathcal{R}^{H \times W \times C} \tag{1}$$

$$\beta = \text{softmax}(x), \tag{2}$$

where the attention is estimated by $q(\cdot)$ with the parameters η as three convolutional layers with 512 kernels of 1×1, 512 kernels of 3×3 and C kernels of 1×1. And β is the normalised attention weights with $\beta \in \mathcal{R}^{H \times W \times C}$, of which the third dimension stands for the size of the whole label set V. Each attention map $\beta^k \in \mathcal{R}^{H \times W \times 1}$, where $k \in [1, C]$, is used to weighted sum the image feature for k_{th} label as:

$$\tilde{i^k} = \sum_l \beta^k \odot \phi(i), \tag{3}$$

where $l \in [1, N]$ indexes the spatial position. The weighted image feature $\tilde{i^k}$ represents the regions related to the k_{th} label. Then the weighted feature can be fed to the FC layer to generate the confidence for each label. To efficiently learn the attention weights, by referring to [33], we reformulate the FC layer as applying the label-specific linear classifier at every spatial location of the image feature $\phi(i)$ and then aggregating the label confidences based on β. That is, we forward $\phi(i)$ to a convolutional layer with C kernels of 1×1 to generate confidence maps $E \in \mathcal{R}^{H \times W \times C}$, then E and β are element-wise multiplied and sum-pooled to obtain the confidence vector $y \in \mathcal{R}^C$, which can be trained with the image ground-truth v.

Since the co-attention mechnisam is introduced to capture the correlations between the target image and its neighbourhood, we formulate the operation of the co-attention mechanism as the weighted sum of the instance-level attention maps. That is, the target image i and its neighbours $z(|z| = m)$ firstly pass through the backbone and InsAtten module to obtain the attention maps β_i and β_z, and confidence vectors y_i and y_z, where $y \in R^C, \beta \in R^{H \times W \times C}$.

Then the co-attention for the neighbour z_j can be computed as:

$$w_j = y_i \odot y_{z_j} \tag{4}$$

$$\alpha_j = softmax(\sum_{k=1}^{C} w_j^k \odot \beta_{z_j}^k) \tag{5}$$

$$\hat{z_j} = \sum_l \alpha_j \odot \phi(z_j) \tag{6}$$

where $\alpha_j \in R^{H \times W \times 1}$, and $\odot$ stands for the element-wise multiply with broadcasting, w_j represents the semantic overlaps between the target and its neighbour z_j. The weighted sum $\hat{z_j}$ is the weighted feature for j_{th} neighbour. We omit $\phi(\cdot)$ in the weighted feature for the expression simplicity. Similarly, for the target image i, the co-attention is computed as:

$$w = y_i \odot \{y_{z_1} ... \odot y_{z_j} ... \odot y_{z_m}\} \tag{7}$$

$$\alpha = softmax(\sum_{k=1}^{C} w^k \odot \beta_i^k) \tag{8}$$

$$\hat{i} = \sum_l \alpha \odot \phi(i) \tag{9}$$

where $\hat{i}$ is the weighted target image feature. The weighted image (and its neighbours) features will be used as node representations in the proposed neighbourhood graph.

Graph Reasoning with Attented Features After we obtain the attended representation for image i and its m neighbours as $\hat{z} = \{\hat{z_0}, \hat{z_1}, ..., \hat{z_m}\}$ [1]. The correlation between every two images can be represented as:

$$s(\hat{z_k}, \hat{z_l}) = \psi(\hat{z_k})^T \psi(\hat{z_l}), \quad \forall k, l \in [0, m] \tag{10}$$

where the $\psi(\cdot)$ is modelled as FC layer with the hidden state size 512. We apply the softmax function on each row of the correlation matrix, which normalises the sum of all the edge values connected to each node to be one. The normalised matrix S is taken as the adjacency matrix for the proposed graph.

We adopt the graph convolutional network (GCN) [14] to reason on the graph. Based on the definition of neighbourhood relations, the GCN can compute the response of each node and pass messages inside the graph. The outputs of the GCN are updated node features, which will be aggregated for further use. More specifically, one layer of the graph convolution is represented as:

$$Z' = S\hat{Z}W \tag{11}$$

where S is the introduced adjacency matrix with $(m+1) \times (m+1)$ dimensions, $\hat{Z}$ is the features of image nodes with $(m+1) \times d$ dimensions, W is the learnable weight matrix with $d \times d$ dimensions. The output Z' of one graph convolutional layer is $(m+1) \times d$ dimensions, which is followed by a ReLU activation. For the fast convergence, we use a residual unit to update the node features *i.e.* $Z' = Z' + \hat{Z}$. The updated node features are fed to an average pooling layer to obtain a $1 \times d$ representation. Moreover, we also perform an average pooling on the target image feature to obtain a $1 \times d$ global representation. Two representations are concatenated together and sent to the Fully-Connected (FC) layer to predict the final confidence vector $y_{neb} \in \mathcal{R}^C$, where C is the number of classes. See Fig. 2 for the illustration.

3.2. Implementation Details

We employ the pre-trained ResNet-101 [9] as the initial backbone CNN model $\phi(\cdot)$ to extract the convolutional features. We adopt the stage-wise training strategy: first we finetune the pretrained backbone model on each dataset, then we train the InsAtten module by referring to [33], and finally we train the CoAtten-GCN module. In practice, the size of the instance-level attention maps are initially trained with the size of 14×14 and then max-pooled as 7×7. The cross-entropy loss function and the stochastic gradient descent [2] are used for the optimisation. We train the models with the batch size 64 and the learning rate of 0.001 from the start and vary the sizes of the image neighbourhood as $m = 3/7/15$. By referring to [12], the nearest neighbour search for the training and test metadata are performed separately.

4. Experiments

We present the experimental results in this section and analyse the effectiveness of the proposed model. Our model is evaluated on three benchmark datasets: NUS-WIDE [5], Mirflickr [11] and MS-COCO [19]. We compare our model with several baselines and state-of-the-art methods. An ablation study is then performed to evaluate the contribution of each component of our model. We finally visualise some of the attention map examples to show their effectiveness.

4.1. Datasets

Both the NUS-WIDE [5] and Mirflickr [11] contain a large number of images collected from the Flickr website, a commonly used image-sharing social network. Each image in the dataset is manually annotated for the presence of the pre-defined label set. By referring to [12, 22], we query the metadata of images via the Flickr API. To ensure the dataset scale, we tokenise and lemmatise the most common metadata *i.e.* user tags and image collection descriptions for our experiments. We remove the duplicates in the processed metadata sets, and the image records without

[1] The weighted feature $\hat{i}$ of the target image i is noted as $\hat{z_0}$ for the unified graph representation.

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic [12]	0.527	0.667	0.679	0.393	0.475	0.768	0.450	0.567
KNN [21]	0.462	0.669	-	-	-	-	-	-
Multi-CNN [9]	0.580	0.789	0.693	0.408	0.490	0.810	0.565	0.666
CNN_Voting [12]	0.599	0.799	0.674	0.423	0.497	0.795	0.597	0.682
TagProp [8]	0.541	0.742	0.674	0.408	0.496	0.784	0.572	0.661
Link-CRF [22]	0.542	0.770	0.698	0.347	0.437	0.805	0.548	0.652
SRN [33]	0.600	0.806	0.704	0.415	0.496	0.811	0.587	0.682
NCNN [12]	0.598	0.803	0.725	0.390	0.478	0.800	0.598	0.685
MangoNet-m3	0.617	0.806	0.715	0.419	**0.507**	0.802	0.599	0.686
MangoNet-m7	0.626	**0.808**	0.720	0.415	0.505	0.804	0.599	**0.687**
MangoNet-m15	**0.628**	**0.808**	0.739	0.410	0.501	0.806	0.599	**0.687**

Table 1. Image annotation results compared with other state-of-the-art methods and our MangoNet on the NUS-WIDE, where m indicates the neighbourhood size used in our model.

valid URLs or ground-truth labels. It is worth noting that the metadata information is only used for locating the image neighbourhood, it does not affect the dataset scale or involve in the model training. We select the optimal sizes of the metadata sets based on the grid search. After the preprocess, we use $201,302$ images for the NUS-WIDE with 81 labels, $3,010$ user tags and 704 collection topics; $12,682$ images for the Mirflickr with 14 labels, 450 user tags. We then select $150,000$ and $51,302$ images for training and test respectively on the NUS-WIDE; $5,200$ and $7,482$ images for training and test respectively on the Mirflickr. For the MS-COCO [19], the descriptions of each image take the form of a set of captions. We tokenise the captions and extract most common 256 visual attributes as the metadata for this dataset. For the semantic representations of the visual attributes, we query the pre-trained word2vec. We use the official train/val split, which is $82,783$ for training and $40,504$ for test.

4.2. Evaluation Metrics

We employ several metrics to evaluate the performance of the proposed models and compared methods. By referring to previous works [12, 15, 33], we compute the average precision (AP), it ranks the retrieved results based on the relevance regarding the query. For each label, relevant images should be ranked higher than the irrelevant ones, noted as the mAP_C. To take into consideration of the label imbalance problem on the dataset splits, we also compute the overall mAP by treating each label assignment as an independent label, noted as the mAP_O. Moreover, to conduct the quantitative evaluation, we predict up to three ranked labels above the confidence threshold 0.5 for each image to compare against the ground-truth. Mean scores of per label and overall precision, recall and F1 score noted as $C_P, C_R, C_{F_1}/O_P, O_R, O_{F_1}$ are reported.

4.3. Overall Performance

We compare the proposed model with several popular and state-of-the-art annotation models, which involve utilising the image metadata or the attention mechanism. Since the different metadata and dataset splits are studied in these models, for fair comparisons, we re-implement some of them [8, 12, 22, 21, 33] by using the metadata vocabularies and dataset splits we processed, and the hand-crafted features from the original models are replaced with the average pooled 2048-d convolutional features from the backbone. For fair comparisons, the compared models share the same finetuned backbone on each dataset. We give the details of the main compared models as follows:

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic [12]	0.571	0.719	0.771	0.334	0.429	0.767	0.494	0.601
KNN [21]	0.745	0.839	-	-	-	-	-	-
Multi-CNN [9]	0.816	0.915	0.889	0.638	0.696	0.857	0.800	0.827
CNN_Voting [12]	0.825	0.916	0.902	0.630	0.685	0.860	0.798	0.828
TagProp [8]	0.818	0.856	0.776	0.625	0.657	0.864	0.594	0.704
Link-CRF [22]	0.800	0.902	0.883	0.601	0.671	0.853	0.766	0.807
SRN [33]	0.831	**0.925**	0.839	0.711	0.760	0.853	0.827	0.840
NCNN [12]	0.840	0.918	0.840	0.724	0.765	0.849	0.826	0.837
MangoNet-m3	0.849	0.924	0.870	0.713	0.769	0.865	0.822	0.843
MangoNet-m7	**0.852**	0.924	0.866	0.718	**0.772**	0.865	0.826	**0.845**
MangoNet-m15	0.851	**0.925**	0.881	0.705	0.761	0.867	0.823	0.844

Table 2. Image annotation results compared with other state-of-the-art methods and our MangoNet on the Mirflickr, where m indicates the neighbourhood size used in our model.

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic [12]	0.703	0.779	0.851	0.556	0.643	0.882	0.575	0.696
KNN [21]	0.699	0.766	-	-	-	-	-	-
Multi-CNN [9]	0.738	0.812	0.827	0.563	0.646	0.848	0.601	0.703
CNN_Voting [12]	0.750	0.818	0.816	0.558	0.635	0.839	0.582	0.687
TagProp [8]	0.720	0.814	0.815	0.570	0.641	0.832	0.607	0.703
Link-CRF [22]	0.718	0.787	0.823	0.548	0.642	0.831	0.594	0.693
SRN [33]	0.771	0.839	0.852	0.588	0.674	0.874	0.625	0.729
NCNN [12]	0.760	0.833	0.838	0.579	0.669	0.871	0.608	0.716
MangoNet-m3	0.775	0.843	0.871	0.579	0.676	0.895	0.619	0.732
MangoNet-m7	0.778	0.845	0.881	0.577	0.676	0.902	0.618	0.733
MangoNet-m15	**0.779**	**0.846**	0.876	0.584	**0.680**	0.898	0.622	**0.735**

Table 3. Image annotation results compared with other state-of-the-art methods and our MangoNet on the MS-COCO, where m indicates the neighbourhood size used in our model.

- ***Meta+Logistic*** [12]: This model is to investigate the annotation capability by only using the metadata. Each image is represented by the binary vector with the metadata vocabulary size $|T|$-dimension, and trained with the logistic loss to generate the label confidence.

- ***CNN_Voting*** [12]: This model is to investigate the contribution of the image metadata neighbourhood for the voting-based methods. Different from the KNN, which allocates the visually similar neighbourhood in the training set, this model uses the metadata neighbourhood directly generated from the test set (same as the proposed MangoNet). Then the label confidence of the test image is set to be a weighted sum of its Multi-CNN prediction and the mean of the Multi-CNN predictions of its neighbours.

- ***SRN*** [33]: This is a state-of-the-art attention-based model, which utilises the instance-level attention as the spatial and semantic regularisation to strengthen the CNN framework. We report the results on the MS-COCO from the original model since we also use the same official dataset splits. For the NUS-WIDE and Mirflickr, the dataset splits are different after the preprocess, therefore, we train and test this model using the same splits as the proposed

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.612	**0.807**	0.727	0.381	0.470	0.808	0.595	0.685
MangoNet w/o GF	0.499	0.691	0.656	0.328	0.412	0.739	0.492	0.591
InsAtten	0.579	0.801	0.700	0.383	0.468	0.822	0.562	0.668
MangoNet	**0.617**	0.806	0.715	0.419	**0.507**	0.802	0.599	**0.686**

Table 4. Ablation studies of our model on the NUS-WIDE.

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.843	0.923	0.885	0.667	0.731	0.866	0.813	0.838
MangoNet w/o GF	0.768	0.876	0.807	0.613	0.679	0.821	0.737	0.777
InsAtten	0.820	0.916	0.884	0.665	0.723	0.857	0.806	0.831
MangoNet	**0.849**	**0.924**	0.870	0.713	**0.769**	0.865	0.822	**0.843**

Table 5. Ablation studies of our model on the Mirflickr.

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.765	0.834	0.842	0.568	0.663	0.880	0.610	0.721
MangoNet w/o GF	0.692	0.774	0.811	0.504	0.599	0.854	0.547	0.667
InsAtten	0.741	0.816	0.831	0.565	0.649	0.851	0.603	0.706
MangoNet	**0.775**	**0.843**	0.871	0.579	**0.676**	0.895	0.619	**0.732**

Table 6. Ablation studies of our model on the MS-COCO.

model for these two datasets.

- ***NCNN*** [12]: This model employs the metadata neighbourhood to assist the target annotation. The image neighbours are embedded with the hidden state size 512 from image global features and then max-pooled. We use the same processed metadata neighbourhood in this model to investigate the importance of the co-attention GCN module.

Tab. 1, 2 and 3 show that our proposed models, noted as the ***MangoNet***, achieve the best performances on the overall evaluation metrics mAP_C, mAP_O and O_{F_1} on all three datasets. The ***Meta+Logistic*** [12] represents each image with a binary vector indicating the presence of the metadata and trains the classifiers with the logistic loss regarding each label. Most vision-based methods outperform this model, which proves that learning from the image visual content is crucial for the annotation. The ***Multi-CNN*** [9] is a standard multi-label annotation model trained with the logistic loss, which serves as a baseline. The ***CNN_Voting*** [12] utilises the same metadata neighbourhood we processed and averages the label confidences from the ***Multi-CNN*** on the test set neighbours. It outperforms the classic ***KNN*** [21], which indicates that the metadata can be useful for eliminating the visual ambiguous and locating the neighbours that contribute to the annotation. The superior performance against these models shows the significance of the proposed graph structure for exploring the neighbourhood feature.

Instead of treating image neighbours equally, the ***TagProp*** [8] is a trained nearest neighbour method, where each image neighbour is re-weighted by the discriminative metric learning. Compared to this model, our proposed ***MangoNet*** represents relationships between the target image and its neighbours as a graph. By reasoning on the graph, we can aggregate the node features as the neighbourhood representation to assist the annotation, which outperforms this model by a large margin. The state-of-the-art method ***SRN*** [33] employs the instance-level attention mechanism as the spatial and semantic regularisation to boost the annotation on the individual image, while in our model, we not only consider the target regional attention but also value the visual clues harvested by the proposed graph model from the neighbourhood. The graphical solution ***Link-CRF*** [22] defines the image relations via metadata, and models them as the CRF. We have the similar motivation to use the neighbourhood defined by the metadata. However, we not only look into the neighbourhood but also propose to capture the visual clues from each neighbour by a co-attention mechanism, which achieves better results. The most related method ***NCNN*** [12] also proposes to utilise the neighbour features to assist the annotation. However, in this model, these neighbours are embedded separately from the target image, and only the holistic features are considered. Different from the ***NCNN***, our ***MangoNet*** establishes a graph structure to represent the neighbourhood and employ a co-attention mechanism to guide the message passing within the graph.

Moreover, we also investigate the influences of the different sizes of the neighbourhood, we report the results of the $m = 3/7/15$ in the tables. As we can see, in general, with the larger neighbourhood, the visual ambiguous can be further eliminated, and the proposed model achieves the better results. To indicate the significance of the proposed components on the whole label set, we show the comparisons of AP values against the NCNN (no co-attention neighbourhood graph is adapted in this model) in Fig. 5, where the x-ray stands for the NCNN value on each label, and y-ray is the corresponding value of our MangoNet. As we can see, the majority of the values are above the $y = x$, which proves the effectiveness of the proposed graph model on the whole tag set.

4.4. Ablation Study

We conduct the ablation analysis on three datasets to further investigate the individual contributions of proposed components in a tiered manner. We compare the following ablation models:

- ***NGCN***: This is a plain graph model without the proposed co-attention mechanism. To implement this model, we replace the node representations in our proposed GCN model as the holistic features, *i.e.* average-pooled convolutional features.

- ***MangoNet w/o GF***: This is the co-attention guided GCN part of the proposed model without the target global feature concatenation.

- ***InsAtten***: In this model, we train the instance-level attention module independently.

- ***MangoNet*** is the proposed full model with the neighbourhood size $m = 3$, same as all other ablation models.

The results are shown in Tab. 4, 5 and 6. As we can see, by introducing the co-attention mechanism into the graph model, our ***MangoNet*** achieves better results against the ***NGCN***, which proves the effectiveness of the proposed at-

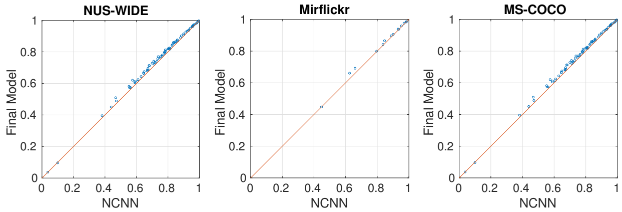

Figure 5. The AP comparisons of the proposed MangoNet against the NCNN model on the NUS-WIDE, Mirflickr and MS-COCO.

tention mechanism in capturing the correlations within the neighbourhood. Without using the global feature concatenation in the GCN module, as we can see, ***MangoNet w/o GF*** has lower performance, since the image neighbours do not guarantee to contain all the labels associated with the target image. The instance-level only attention model ***InsAtten*** performs lower than others, but based on the co-attention maps generated from the InsAtten module, our full model ***MangoNet*** achieves the best performance.

4.5. Visualisation of Attention

To verify the proposed attention mechanisms, we visualise the attention maps from the co-attention module. The brighter colour (yellow) indicates the higher attention weights. We show the examples of the co-attention maps of the given images in Fig. 6. The first column of every two rows is the target image and its co-attention map, while the rest columns are the neighbours and their corresponding attention maps. As we can see, the co-attention maps capture the correlations between the target image and its neighbours in both simple and complex scenes. For example, small subjects like 'surfboard' and 'ball' are well-captured in example 2 and 4, while in the complex scenes such as example 3, the 'people' and 'car' are also well-captured in the neighbours. The co-attention mechanism is employed to find the most relevant visual features to eliminate the recognition uncertainties, then the neighbourhood graph receives these features and communicates within the neighbourhood, which improves the ability of the annotation model to recognise the target image.

5. Conclusion

Images are connected to each other via the abundant metadata. Fully making use of these connections can assist the image annotation. In this paper, we explore the image neighbours by measuring their metadata similarities and propose a graph network to model the correlations between the target image and its neighbours. A co-attention mechanism is introduced to leverage the visual attention within the neighbourhood. Experimental results on three benchmark datasets show that the proposed model achieves the best performances against compared methods. Since the textual metadata is mainly used in our experiments, we will explore other metadata types in our future work.

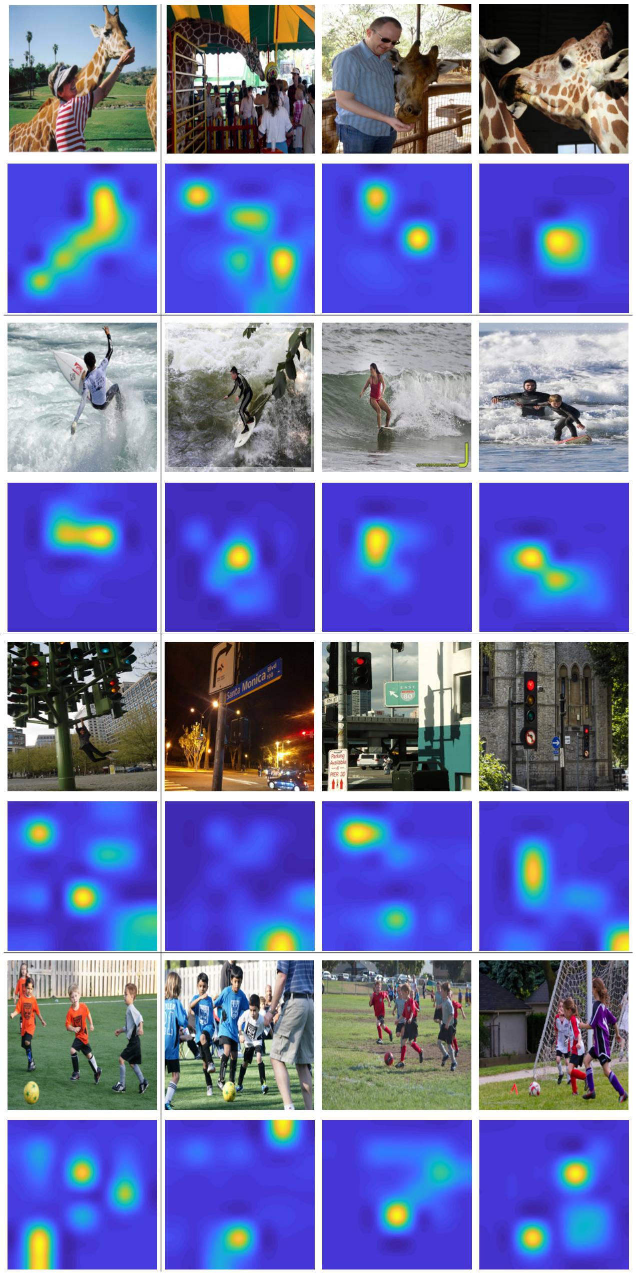

Figure 6. The visualisations of the co-attention maps of the targets and their neighbours. The first column is the target and its attention map, the rest columns are the neighbours, where the neighbourhood size $m = 3$.

6. Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] Lamberto Ballan, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo. A cross-media model for automatic image annotation. In *ICMR*, page 73. ACM, 2014.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Trans. Intelligent Syst. and Technology*, 2(3):27, 2011.

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Pro. ACM Int. Conf. Image and Video Retrieval*, page 48. ACM, 2009.

[6] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013.

[7] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[8] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316. IEEE, 2009.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Tao Mei, Yueying He, and Zhonghua Zhao. Learning social image embedding with deep multimodal attention networks. In *Proc. Thematic Workshops of ACM Multimedia 2017*, pages 460–468. ACM, 2017.

[11] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Pro. ACM Int. Conf. Multimedia Info. Retrieval*, pages 39–43. ACM, 2008.

[12] Justin Johnson, Lamberto Ballan, and Fei-Fei Li. Love thy neighbors: Image annotation by exploiting image metadata. In *ICCV*, pages 4624–4632, 2015.

[13] Dhiraj Joshi, Jiebo Luo, Jie Yu, Phoury Lei, and Andrew Gallagher. Using geotags to derive rich tag-clouds for image annotation. In *Social media modeling and computing*, pages 239–256. Springer, 2011.

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2016.

[15] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys*, 49(1):14, 2016.

[16] Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. *Proc. Uncertainty in Artificial Intell*, 2014.

[17] Yunpeng Li, David J Crandall, and Daniel P Huttenlocher. Landmark classification in large-scale image collections. In *ICCV*, pages 1957–1964. IEEE, 2009.

[18] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. *CoRR*, abs/1511.05493, 2015.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[21] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. *IJCV*, 90(1):88–105, 2010.

[22] Julian McAuley and Jure Leskovec. Image labeling on a network: using social-network metadata for image classification. In *ECCV*, pages 828–841. Springer, 2012.

[23] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[25] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NeurIPS*, pages 2222–2230, 2012.

[26] Zak Stone, Todd Zickler, and Trevor Darrell. Autotagging facebook: Social network context improves photo annotation. In *CVPR Workshops*, pages 1–8. IEEE, 2008.

[27] Mingkui Tan, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Junbin Gao, Fuyuan Hu, and Zhen Zhang. Learning graph structure for multi-label image classification via clique generation. In *CVPR*, pages 4100–4109, 2015.

[28] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[29] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A unified framework for multi-label image classification. *CVPR*, pages 2285–2294, 2016.

[30] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. *ECCV*, 2018.

[31] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *TPAMI*, 38(9):1901–1907, 2016.

[32] Aron Yu and Kristen Grauman. Predicting useful neighborhoods for lazy local learning. In *Advances in Neural Information Processing Systems*, pages 1916–1924, 2014.

[33] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. *CVPR*, 2017.