Multi-stage Semantic Attention with Transformer for Multi-label Image Classification

Qi Du[1, *], Ying Ma[1] and Jianmin Li[1]

[1]*College of Computer and Information Engineering, Xiamen University of Technology, No.600 Ligong Road, Jimei District, Xiamen, 361024, Fujian Province, China*
duqi13228882809@163.com, maying@xmut.edu.cn, lijm@xmut.edu.cn
[*]*Corresponding author.*

Abstract
Multi-label image classification is a fundamental classification task, which seeks to assign numerous possible labels to an image. Many deep convolutional neural network (CNN)-based approaches to discovering the semantics of labels and learning the semantic representation of images by modeling label correlation have been proposed in recent years. However, some small and similar objects cannot be predicted accurately due to the limitation of convolutional kernel representation capability. As a result, in order to solve this problem, this paper introduces twins-transformer. Since different stages of image representation of this model capture different levels or scales of features and have different discriminative capacities, we design a multi-stage semantic attention with transformer (MAST) framework to learn the semantic representation of images using its own multi-stage mechanism, while employing a three-layer standard transformer decoder as an effective component for feature fusion. Experiments conducted on the VOC 2007 dataset show that MSAT achieves better experimental results and improves the performance of multi-label image classification tasks to some extent.

Keywords: *Multi-label Image Classification, Transformer, Semantic Attention.*

1 INTRODUCTION

Multi-label image classification (MLIC) is a significant computer vision task that aims to assign multiple labels to an image based on its content. MLIC is more general and practical than traditional single-label image classification since any image in the physical world is likely to contain semantic information of many categories. For example, a news document can belong to multiple categories. Therefore, it plays a crucial role in a wide range of applications such as human attribute recognition [19], medical image recognition [11], and recommender systems [15] [27]. However, the task faces some unique challenges due to the rich semantic information and complex dependencies of images and their labels.

Early multi-label classification algorithms [7] [22] [29] recognized each object in isolation and naively divided this problem into multiple binary classification tasks. Moving into the deep convolutional neural network (CNN) [16] phase, great progress has been made in image classification, and the accuracy of existing multi-label image recognition methods based on CNN and their variants [13] [14] [21] has been improved. However, because these methods ignore the complicated topology between objects in an image, its performance is intrinsically constrained, preventing further improvements in the accuracy of multi-label image classification. Moreover, in CNN, convolution is computed in one local region at a time with inductive bias, and local information is easily ignored after sufficient convolution time, leading to unsatisfactory performance of CNN-based models in multi-label classification.

Transformers [8] [23] were originally designed to mine long-term dependencies between word embeddings in the field of natural language processing (NLP) and have recently been found to be able to be used for various computer vision tasks as well. ViT [9] enables transformer to replace the long-term dominance of convolutional neural networks in the field of computer vision. One of the advantages of transformer is its multi-head attention mechanism, which can extract features from different parts of an object class or different views, thus better recognizing objects with occlusions and viewpoint changes, gradually improving the performance

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1193-1199, 2023.
https://doi.org/10.2991/978-94-6463-040-4_178

of various computer vision tasks. In this paper, we use twins-transformer [6] for feature extraction of images, and it is worth noting that twins-transformer comes with a multi-stage mechanism that captures features at different levels or scales with different discriminative capabilities. Therefore, better performance can be obtained by making good use of the multi-layer features of twins-transformer. In this paper, we make full use of this feature and apply its own multi-stage mechanism to fuse image features with label semantic features. In the fusion, image features and label semantic features are first fused at a shallow level by matrix multiplication, and then fused at a deep level in the transformer decoder; meanwhile, we further propose a semantic attention module to obtain a better feature representation.

We conducted experiments on the VOC 2007 benchmark multi-label image dataset. The approach outperforms existing computational methods, according to the results of the experiments. The performance of image classification is improved compared with existing image classification methods. In conclusion, the contributions of this paper are summarized as follows.

- Using the multi-stage mechanism that comes with Twins-transformer to learn the fusion representation of image features and label semantic features, and using transformer-decoder for deep fusion;
- To obtain a better feature representation, a semantic attention module (SAM) is designed to fuse the contexts between images, which further helps to fuse the global contexts.

2 RELATED WORK

2.1 Transformer in Vision Task

Transformers [23] were originally proposed for modeling long-term dependencies in sequence learning problems and have been widely used for natural language processing tasks. Recently, transformer-based models have also been developed for many vision tasks and have shown great potential. [2] trained a sequential transformer called iGPT to automatically regress predicted pixels. [1] designed an end-to-end object detection framework called DETR with

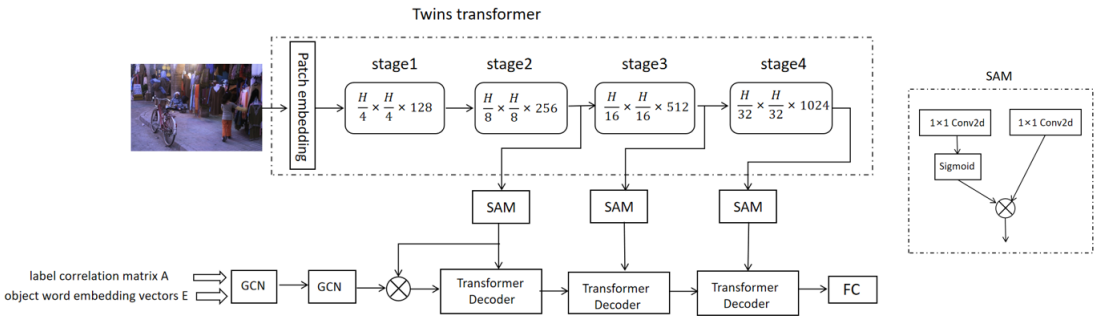

Figure 1: The overall architecture of our proposed Multi-stage Semantic Attention with Transformer model.

transformer. [9] proposed visual transformers (ViT), in which they segment an image into multiple patches and feed them into a stacked transformer architecture for classification. More advances in transformers for applications in computer vision can be found in [12].

Our approach also uses the transformer, but we utilize the twins-transformer's own multi-stage mechanism for multi-label image classification, which is fundamentally different from the improved approach in most existing work, and use the classic transformer decoder self-attention mechanism for image features and label semantic features perform deep fusion.

2.2 Multi-label Image Classification

In computer vision, multi-label classification has been a challenging and fundamental problem. By training a set of classifiers for each label, early approaches to multi-label image classification naively separated this task into numerous independent binary classification tasks. However, the approach ignored the topology between objects and the semantic dependencies between multiple classes, which are particularly important for multi-label classification. Therefore, combining such semantics has been an important research direction for multi-label classification, and various approaches have been proposed. Specifically, some previous work [24] explicitly captured class relevance through cnn-based models and recurrent neural networks (RNN), however, they are usually difficult to optimize their parameters. There are also some approaches based on probabilistic graphical models [18], which label dependencies in the covariance of probability distributions, but they tend to have a high computational complexity during statistical inference. Since labeled dependencies can be represented as graphs, some approaches apply graph convolutional networks (GCNs) [31] on these graphs to learn labeled representations, as in [5]. In addition, others have suggested the use of attention mechanisms to capture the correlations between labels. To capture the semantic and spatial relationships of these various labels, [32] proposed a spatial regularization network. To capture

label correlations, [26] introduced a spatial transformer layer and a long short-term memory (LSTM) unit. A graph-based framework [17] was also proposed to characterize the relationships between labels through knowledge graphs in order to generate more accurate image representations.

3 METHOD

3.1 Architecture

In this paper, we propose a multi-stage semantic attention with transformer (MSAT) framework, which mainly consists of a semantic attention module and a feature fusion module, as shown in figure 1. First, image features are extracted using twins-transformer, while its own multi-stage mechanism is used to input the features extracted in the last three steps into the semantic attention module to obtain a better representation of high-level semantic features. Subsequently, two layers of GCN are used to obtain the label-to-label dependencies. Then, the image feature representations obtained in each stage with label semantics are fed into the transformer decoder for feature fusion. It is worth noting that there is one decoder for each stage of the framework, and the input of the latter two decoders is the output of the previous decoder and the corresponding image feature representations, which are continuously fused instead. In the final stage of our architecture, we input the final output fused features into a fully connected layer and apply the fully connected layer for prediction, where the cross-entropy loss function is used as follows.

$$L_{loss} = \sum_{i=1}^{n} y^i \, log\left(\sigma(\hat{y}^i)\right) + (1 - y^i) log\left(1 - \sigma(\hat{y}^i)\right) \quad (1)$$

Where σ() is the sigmoid function, $y^i = \{0,1\}$ indicates whether label i appears in the image, and $\hat{y}^i$ is the output of the fully connected layer.

3.2 Multi-stage Image Features

In this paper, we use twins transformer [6] as the backbone and output the features of the last three stages through a self-contained multi-stage mechanism, denoted as $F = \{f_s\}_{i=1}^{s}$, where s is the total number of stages and f_s is the image feature of the sth stage. Given an input image $x \in R^{H \times W \times 3}$, where H and W are the height and width of the image. Before the start of each stage, the image undergoes a downsampling process to reduce the resolution and adjust the number of channels in order to form a hierarchical design. The resolution of each stage is $\frac{H}{4} \times \frac{H}{4}, \frac{H}{8} \times \frac{H}{8}, \frac{H}{16} \times \frac{H}{16}, \frac{H}{32} \times \frac{H}{32}$, and the number of channels in each stage is 128, 256, 512, 1024 respectively.

3.3 Semantic Attention Module

The image features f_s of the last three stages are obtained through the previous section, and in order to make the image features that need attention more obvious, a semantic attention module is constructed in this paper. Specifically, this module is made up of two 1×1 convolutional layers and a dot product operation. The image features obtained in the latter three stages are convolved twice, and the output after one of the convolutions is convolved once with the output after the other convolution by the sigmoid activation function, and finally the output of each stage after this module $\bar{f}_s$ is obtained, as shown in equation 2.

$$\bar{f}_s = \sigma\left(W_\varphi * f_s\right)^T (W_\vartheta * f_s) \quad (2)$$

Where σ() indicates the sigmoid function; W_φ and W_ϑ denote the convolution kernel; $*$ is the convolution operation; f_s is the image feature of the sth stage.

3.4 Graph Convolutional Networks for Multi-Label Classification

The topology unique to graphs can model any two-two correlated labeling relationships, and the network model constructed in this paper adopts a similar approach to ML-GCN [5], consisting of two layers of stacked GCNs. The initialized correlation coefficient matrix A and the label embedding E, which are, from a mathematical perspective, the adjacency and diagonal matrices, respectively, used to represent the correlations of semantic labels of categories, are the inputs for the first layer of the GCN. The model output for the last GCN layer is $W \in R^{D \times C}$, where D denotes the dimensionality of the image features and C denotes the number of categories. The fundamental principle of GCN is to propagate information among the nodes in order to update the node representation. The forward propagation process of a single GCN layer can be represented as

$$H^{l+1} = \delta(H^l, A) \quad (3)$$

Where H^l represents the node of the previous layer, which is the input of the next layer H^{l+1}; $A \in R^{n \times n}$ is a correlation coefficient matrix; δ() is a nonlinear activation function, such as Sigmoid or ReLU; H^{l+1} is the new node feature of the output.

3.5 Feature Fusion Module

Previous work has mostly concentrated on capturing correlations between labels without successfully fusing image features and label embeddings, which has a negative impact on the convergence efficiency of model and prevents future improvement of multi-label image classification accuracy. To address this drawback, this paper introduces transformer decoder [23], which serves as an effective component for fusing image features and

label semantic features. Specifically, in this paper, a three-layer decoder is used, and the obtained image features $\bar{f}_s$ and the result after the dot product of the output of the last layer of GCN are used as the input of the first decoder, and similarly, in the second and third decoder layers, the output of the previous decoder and the image features of the corresponding stage are used as the input of this decoder layer to

Table 1: Comparisons of our method with previous state-of-the-art methods on PASCAL VOC 2007.

VOC 2007	CNN-RNN	RLSD	AR	ML-GCN	SSGRL	F-GCN	DSDL	MSRN	Ours
areo	96.7	96.4	98.6	99.5	99.5	99.5	99.8	100.0	99.9
bike	83.1	92.7	97.1	98.5	97.1	98.5	98.7	98.8	98.9
bird	94.2	93.8	97.1	98.6	97.6	98.7	98.4	98.9	98.6
boat	92.8	94.1	95.5	98.1	97.8	98.2	97.9	99.1	99.2
bottle	61.2	71.2	75.6	80.8	82.6	80.9	81.9	81.6	78.9
bus	82.1	92.5	92.8	94.6	94.8	94.8	95.4	95.5	97.2
car	89.1	94.2	96.8	97.2	96.7	97.3	97.6	98.0	97.7
cat	94.2	95.7	97.3	98.2	98.1	98.3	98.3	98.2	98.2
chair	64.2	74.3	78.3	82.3	78.0	82.5	83.3	84.4	85.9
cow	83.6	90.0	92.2	95.7	97.0	95.7	95.0	96.6	97.5
table	70.0	74.2	87.6	86.4	85.6	86.6	88.6	87.5	93.2
dog	92.4	95.4	96.9	98.2	97.8	98.2	98.0	98.6	98.6
horse	91.7	96.2	96.5	98.4	98.3	98.4	97.9	98.6	99.2
motor	84.2	92.1	93.6	96.7	96.4	96.7	95.8	97.2	98.4
person	93.7	97.9	98.5	99.0	98.8	99.0	99.0	99.1	98.8
plant	59.8	66.9	81.6	84.7	84.9	84.8	86.6	87.0	86.7
sleep	93.2	93.5	93.1	96.7	96.5	96.7	95.9	97.6	97.9
sofa	75.3	73.7	83.2	84.3	79.8	84.4	86.4	86.5	86.1
train	99.7	97.5	98.5	98.9	98.4	99.0	98.6	99.4	99.6
tv	78.6	87.6	89.3	93.7	92.8	93.7	94.4	94.4	93.8
mAP	84.0	88.5	92.0	94.0	93.4	94.1	94.4	94.9	95.2

obtain the final fusion features. Since we do not need to perform autoregressive prediction, we do not use attention masks, therefore, we only use self-attention, which is defined with the same functions as those defined in the standard transformer decoder, with the following equation for the attention mechanism.

$$Atten(Q,K,V) = \omega\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Where Q, K, V are the matrices of query, key, and value respectively; d_k denotes the dimension of K, ω() denotes the softmax function. The dot product operation in equation 4 returns the similarity of each query and key value, and the output is the weighted sum of V, the greater its weight, the greater the similarity between the query and the key.

4 EXPERIMENTS

We experimented with the PASCAL VOC dataset [10] to evaluate the proposed method. We utilized the Average Precision (AP) and mean Average Precision (mAP) for the evaluation based on previous work.

4.1 Implementation Details

PyTorch was used to carry out all of the experiments. The input label features E are 300-dimensional Glove features pre-trained on the Wikipedia dataset. The backbone uses twins-transformer for feature extraction of the images, while the image features and label semantic features are fused using twins-transformer's own multi-stage mechanism to obtain image features from the last three stages of twins-transformer's module, respectively. Output dimension of f_{s2} is 256, output dimension of f_{s3} is 512, and f_{s4} has an output dimension of 1024. The

output dimensions of the two GCN layers are 512 and 1024, respectively. For training and testing, the input images are adjusted to 224x224. Our model is trained on a GeForce RTX3090-24GB GPU with a batch size of 16. We employ SGD as the optimizer, which has a momentum of 0.9 and a weight decay of 10-4. The initial learning rate is set as 0.01, which decays by a factor of 10 for every 20 epochs in total 80 epochs.

4.2 Experiment Results

PASCAL VOC 2007 [10] is a commonly used multi-label dataset with 9963 images representing 20 common object categories. It is divided into a training set, a validation set and a test set, where 5011 images are used for training and 4952 images are used for testing. The training set is used to train our model, while the test set is utilized to evaluate the classification performance. A variety of multi-label image classification algorithms were compared to our proposed MSAT. They are CNN-RNN [24] RLSD [28], AR [3], ML-GCN [5], SSGRL [4], F-GCN [25], DSDL [30] and MSRN [20]. Table 1 presents a comparison of our MSAT with other approaches. With an overall mAP gain of 1.2% and 1.8%, respectively, our method consistently outperforms these methods, particularly when compares to the other two current state-of-the-art methods ML-GCN and SSGRL.

5 CONCLUSIONS

In this paper, we propose a multi-stage semantic attention with transformer for Multi-label image classification (MSAT) framework. MSAT utilizes the twins-transformer's own multi-stage mechanism to fuse image features and label features. We also employ a standard transformer decoder for deeper fusion of the two features. In addition, we designed a semantic attention module to further help fuse the global context between images to obtain a better feature representation. The experimental results on the voc2007 dataset show that the MSAT model has some improvements over the classical deep multi-label image classification model, and the experiments demonstrate the effectiveness of the MSAT model.

REFERENCES

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Springer, Cham.

[2] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020, November). Generative pretraining from pixels. In *International conference on machine learning* (pp. 1691-1703). PMLR.

[3] Chen, T., Wang, Z., Li, G., & Lin, L. (2018, April). Recurrent attentional reinforcement learning for multi-label image recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[4] Chen, T., Xu, M., Hui, X., Wu, H., & Lin, L. (2019). Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 522-531).

[5] Chen, Z. M., Wei, X. S., Wang, P., & Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5177-5186).

[6] Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., ... & Shen, C. (2021). Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, *34*, 9355-9366.

[7] Clare, A., & King, R. D. (2001, September). Knowledge discovery in multi-label phenotype data. In *European conference on principles of data mining and knowledge discovery* (pp. 42-53). Springer, Berlin, Heidelberg.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[10] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303-338.

[11] Ge, Z., Mahapatra, D., Sedai, S., Garnavi, R., & Chakravorty, R. (2018). Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*.

[12] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2021). A survey on visual transformer. arXiv preprint arXiv:2012.12556

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[14] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected

convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[15] Jain, H., Prabhu, Y., & Varma, M. (2016, August). Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 935-944).

[16] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

[17] Lee, C. W., Fang, W., Yeh, C. K., & Wang, Y. C. F. (2018). Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1576-1585).

[18] Li, Q., Qiao, M., Bian, W., & Tao, D. (2016). Conditional graphical lasso for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2977-2986).

[19] Li, Y., Huang, C., Loy, C. C., & Tang, X. (2016, October). Human attribute recognition by deep hierarchical contexts. In *European conference on computer vision* (pp. 684-700). Springer, Cham.

[20] Qu, X., Che, H., Huang, J., Xu, L., & Zheng, X. (2021). Multi-layered semantic representation network for multi-label image classification. *arXiv preprint arXiv:2106.11596.*

[21] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[24] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2285-2294).

[25] Wang, Y., Xie, Y., Liu, Y., Zhou, K., & Li, X. (2020, October). Fast graph convolution network based multi-label image recognition via cross-modal fusion. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1575-1584).

[26] Wang, Z., Chen, T., Li, G., Xu, R., & Lin, L. (2017). Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision* (pp. 464-472).

[27] Yang, X., Li, Y., & Luo, J. (2015, October). Pinterest board recommendation for twitter users. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 963-966).

[28] Zhang, J., Wu, Q., Shen, C., Zhang, J., & Lu, J. (2018). Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, *20*(10), 2801-2813.

[29] Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, *26*(8), 1819-1837.

[30] Zhou, F., Huang, S., & Xing, Y. (2021, May). Deep semantic dictionary learning for multi-label image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 4, pp. 3572-3580).

[31] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57-81.

[32] Zhu, F., Li, H., Ouyang, W., Yu, N., & 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* (pp. 5513-5522).