Multi-task attribute-fusion model for fine-Tgrained image recognition

Li, Mengze, Kong, Ming, Kuang, Kun, Zhu, Qiang, Wu, Fei

Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, Fei Wu, "Multi-task attribute-fusion model for fine-Tgrained image recognition," Proc. SPIE 11550, Optoelectronic Imaging and Multimedia Technology VII, 115500J (10 October 2020); doi: 10.1117/12.2573311

Event: SPIE/COS Photonics Asia, 2020, Online Only

Multi-Task Attribute-Fusion Model for Fine-grained Image Recognition

Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu

Zhejiang University, Hangzhou, China

ABSTRACT

Attribute information in fine-grained image recognition often provides more accurate and rich information related to categories. How to effectively combine such knowledge to guide image classification tasks has been one of the research hotspots in computer vision in recent years. We believe that using the association relationship between attributes to fuse attribute information can obtain a more accurate representation of the image. In this paper, we propose a novel Multi-Task Attribute Fusion Model (MTAF) which makes two major improvements to the traditional multi-task learning framework: 1) *Attribute-Aware Feature Discrimination*: combine the spatial attention and the channel attention mechanism to enhance the feature map of the CNN, so that attribute can be associated to important positions and important channels of the image; 2) *Transformer-Based Feature Fusion*: introduce the Transformer model to better learn the logical association between attributes, so that the reconstructed features are able to achieve a best classification performance. We have verified our algorithm on two datasets, one is the own-collected medical dataset for thyroid benign and malignant identification, and the other is an open dataset widely used for fine-grained image recognition. Experimental results on both datasets demonstrate that the proposed method can achieve higher classification accuracy than baselines.

Keywords: Fine-grained recognition, Transformer

1. INTRODUCTION

In fine-grained image recognition, algorithms that use only image information are often insufficient to achieve satisfactory results. Therefore, researches often introduce additional information to help solve problems, such as object bounding boxes, attribute annotations, and part landmarks.[1–3] Among them, the significance of object attributes has attracted researches' attention most. Attributes usually come from a more precise and detailed description of the key information of the identified object. How to effectively integrate attribute information into deep learning models to improve model performance is difficult but important for the research of computer vision.

Multi-task learning[4] is a natural way to integrate attribute information into the deep learning model, which is an inductive transfer mechanism that can utilize the correlation between tasks to promote each other, and eventually improve model performance towards the main task. There have been many attempts to use multi-task learning frameworks to combine object attribute information and solve fine-grained image classification problems, for example, Ref. 5 and 6. Specifically, the attribute features are extracted through the same network model with one or more shared parameters. During the joint learning of multiple attributes, the correlation between the features can be sensed, and the classification results of multiple attributes can be obtained. However, this method has two major difficulties:

Further author information: (Send correspondence to Kun Kuang)
Mengze Li E-mail: 3150104805@zju.edu.cn
Ming Kong E-mail: zjukongming@zju.edu.cn
Kun Kuang E-mail: kunkuang@zju.edu.cn
Qiang Zhu E-mail: zhuq@zju.edu.cn
Fei Wu E-mail: wufei@zju.edu.cn

Optoelectronic Imaging and Multimedia Technology VII, edited by Qionghai Dai
Tsutomu Shimura, Zhenrong Zheng, Proc. of SPIE Vol. 11550, 115500J
© 2020 SPIE · CCC code: 0277-786X/20/$21 · doi: 10.1117/12.2573311

1. The expression of attributes in the image varies and depends on the specific position of the target object in the image. This makes the attribute-related information usually fuzzy, difficult to describe, or even invisible, and the difference between the attributes is subtle. How to effectively extract the attribute information is still very challenging.

2. The different attributes of the target object are related, even implying a strictly logical reasoning relationship. But there has been no effective mechanism by which the attributes can guide the main task for best performance in the multi-task learning framework.

In this paper, we propose a new Multi-Task Attribute Fusion Model (MTAF) to solve the above two difficulties in the traditional multi-task learning process. Firstly, we use the attention mechanism to extract and separate global features and multiple attribute features. Secondly, we construct an attribute fusion model based on the Transformer technique. Compared with the existing fine-grained classification models based on attributes, we emphasize the construction of the correlation between attributes and have better feature fusion and reconstruction capabilities.

In order to validate our model, we apply MTAF to the classification problem between benign and malignant thyroid. In addition to the classification accuracy exceeding the state-of-the-art algorithm, we also find that our MTAF produces results which humans understand, and greatly improves the interpretability of the model. In addition, we apply the algorithm to a widely publicized fine-grained classification dataset, and the classification results show that our model achieves excellent performance as well.

Our contributions include the following three areas:

1. We propose a novel model MTAF to solve the problem of fine-grained classification with attribute annotations, which can use attribute information to improve classification accuracy.

2. We construct an attribute fusion module based on the Transformer structure. It can fuse and reconstruct attribute features and the global feature.

3. We conduct an evaluation of the performance of a real-world medical image classification dataset and a well-known fine-grained classification dataset. The extensive experiments show that our model achieves good performance.

2. RELATED WORK

There are extensive studies[7] for using the detailed information carried by object attributes to improve the accuracy of the image recognition task. Based on the convolution neural network, a multi-task model[5] is used to introduce attribute information for guiding the model to focus on attribute-related features, and increase the generalization ability of the learned model. In Ref. 6, the multi-task model predicts the categories of each attribute, and uses weighted stitching as a supplement to image features, and introduces attribute information into the model feature representation. In Ref. 8, the attribute label is used to guide the model to pay attention to the spatial local information of the image, and it emphasizes the attribute-related image regions with discriminative ability. Earlier works lack effective techniques to accurately match attributes to image information. They also lack a rigorous mechanism to effectively mine the association between attributes, seamlessly integrating into the global feature extraction process, and eventually leading to global optimization of classifiers.

Inspired by human thinking, the introduction of attention mechanism can more effectively relate object attributes to the learning process, and further improve model performance. Ref. 9 uses the spatial attention mechanism to strengthen the model's attention to local areas containing key information, and effectively improve the model's ability to distinguish nuances. In Ref. 10, channel attention mechanism is designed into the feature extraction part of the convolution neural network, so that the generated feature map focuses on the channel containing important information, and improves the model's attention to important types of features. Ref. 11 combines the spatial attention mechanism and the channel attention mechanism to augment the convolution neural network, so that the feature extraction can focus on both important positions and important types of features. The above work strengthens feature extraction through the use of attributes via the attention mechanism,

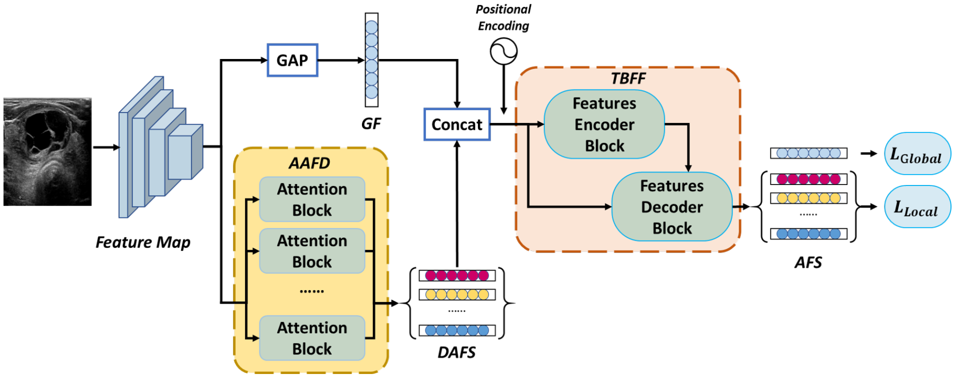

Figure 1. Overall Architecture of Multi-Task Attribute Fusion Model (MTAF). For the input image, the discriminative attribute feature sequence (DAFS) is generated with the Attribute-Aware Feature Discrimination (AAFD) module. Combining with the global feature (GF) and the positional encoding, Transformer-based Feature Fusion (TBFF) module fuse and refine the correlation between each element of the feature sequence, and the augmented feature sequence (AFS) generated for the final classification tasks.

but this enhancement is based on the independent relationship between the attributes and lacks a mechanism to effectively use the association between attributes to globally optimize the classification performance.

Transformer is an encoder-decoder architecture sequence-to-sequence model based on attention mechanism,[12] which can effectively learn the interdependence between sequence elements. After being widely verified in the field of NLP, its ability to express complex dependencies and its ability to extract comprehensive features have also extended it to the field of computer vision. In Ref. 7, the Transformer-style architecture is used to analyze the video sequence to identify and locate human-specific actions, and good results and interpretability are obtained. In Ref. 13, Transformer is introduced to decode the image feature sequences extracted by the convolution neural network, which emphasizes the internal interaction of the image features and complex reasoning, and achieves very good results in the problem of image annotation. These studies further prove Transformer's superior cross-media feature fusion ability to capture better internal correlation of attribute features. In this paper, for the first time, we propose to use Transformer technology to mine, fuse and reconstruct image features in an end-to-end multi-task learning framework to optimize the main task in a global way. We will elaborate on the algorithm flow in the next chapter.

3. METHOD

The traditional flow for a multi-task learning framework that introduces attribute information is the following: first obtains a feature map through a convolution neural network, and then uses the last layer of the network to extract the Global Feature and Attribute Features, respectively, finally combines the two types of features to form a feature sequence for classification. Compared to the existing framework, two characteristic modules are specially designed to enhance the learning process, elaborated in Fig. 1:

1. *Attribute-Aware Feature Discrimination (AAFD)*: The input of this module is a feature map from a convolution neural network. Attention Block is specially designed to use spatial attention and channel attention for obtaining a more distinguished feature representation, namely Discriminative Attribute Features (DAF).
2. *Transformed-Based Feature Fusion (TBFF)*: The input of this module is to connect the Global Feature separated from the Feature Map with the DAF obtained through AAFD. We use Transformer technology to deeply mine, fuse, and reconstruct the combined features, and produce the global optimal Augmented Feature Sequence (AFS).

The main goal of attribute-based multi-task learning is to use attribute information to guide the learning process and ultimately achieve the optimal performance of the main task. Our two characteristic modules, AAFD

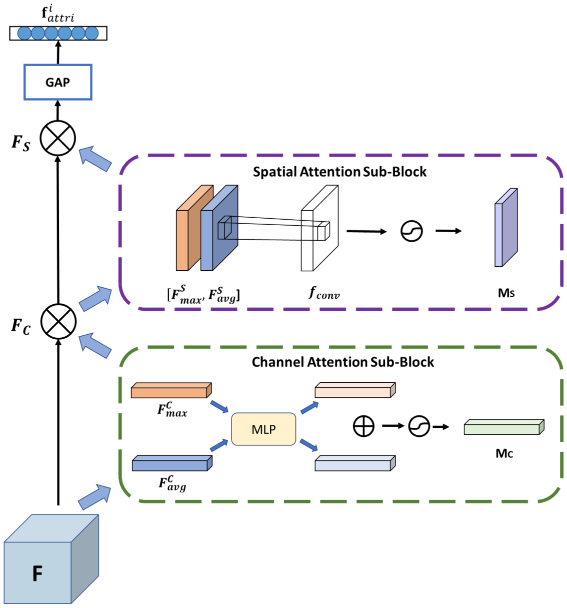

Figure 2. Structure of Attention Block in AAFD. The feature maps are sequentially processed by the channel and spatial attention sub-block and performed global average pooling to get the attribute feature.

and TBFF, play different roles at different stages of multi-task learning: AAFD focuses on the feature extraction phase for relating attribute information to the image (visually perceptible information), and is oriented to focus on the correct spatial location and channel category; and TBFF focuses on the feature classification stage, the relationship between multiple attributes (implicit knowledge of logical reasoning) is used to guide the fusion and reconstruction of features. It is worth pointing out that these two innovated modules integrate seamlessly in the complete multi-task learning process, guide each other, and unify to the global goal of optimizing the performance of the main task. This is also the core innovation and value of our work.

3.1 Attribute-Aware Feature Discrimination

For a fine-grained classification task with N kinds of attributes, AAFD consists of N attention blocks, Each of which is used to generate the corresponding attribute feature, f^i_{Attri}. And the output of the AAFD model is the discriminative attribute feature sequence (DAFS), which is expressed as $f_{Attri} = [f^1_{Attri}, f^2_{Attri}, \ldots, f^N_{Attri}]$

Each attention block has the same structure, as shown in Fig. 2. A block consists of two sub-block, named *Channel Attention Sub-Block (CASB)* and *Spatial Attention Sub-Block (SASB)* ,respectively. As each channel of the feature map can be regarded as a feature detector, CASB is able to enhance the differences of importance between features. Based on the channel attention map, the spatial attention map can further introduce the difference of spatial information in the feature map, thus, SASB is able to find the spatial focus of the attribute features. The overall attention calculation process is as follows:

$$\begin{aligned} F_C &= M_C(F) \otimes F \\ F_S &= M_S(F) \otimes F_C \end{aligned} \tag{1}$$

For the last-layer feature maps of the convolution network $F^{(C\times H\times W)}$, we sequentially introduce a 1D channel attention map $M_C \in \mathbb{R}^{(C\times 1\times 1)}$, and a 2D spatial attention map $M_S \in \mathbb{R}^{(1\times H\times W)}$, where $\otimes$ denotes element-wise multiplication. The output of channel attention sub-block is expressed as F_C, and the output of the spatial attention sub-block is expressed as F_S. Finally, the discriminative attribute feature vector f^i_{attri} is obtained by making global average pooling on the modified feature maps F_S. The following describes the detail of both channel and spatial attention sub-blocks.

3.1.1 Channel Attention Sub-Block

We first aggregate spatial information of the feature map F, and get two spatial context descriptor: F^C_{max} and F^C_{avg}. Both descriptors are sent to a shared network, which is composed as a multi-layer perceptron (MLP) with one hidden layer, to generate the channel attention map $M_C \in \mathbb{R}^{(C\times 1\times 1)}$. The output vectors are merged by element-wise summarization. To sum up, the calculation process of Channel Attention Sub-Block is:

$$M_C(F) = \sigma(W_1(W_0(F^C_{avg})) + W_1(W_0(F^C_{max}))) \tag{2}$$

where σ denotes a sigmoid activation function, and W_0, W_1 is the weight of MLP.

3.1.2 Spatial Attention Sub-Block

After the process of Channel Attention Sub-Block, we get a channel feature map F_C. We aggregate channel information of F_C by using two pooling operations, generating two 2D maps: $F^S_{max}, F^S_{avg} \in \mathbb{R}^{\nVdash\times\mathbb{H}\times\mathbb{W}}$. Then concatenate them and convolve by a standard convolution layer, producing the 2D spatial attention map M_S. The calculation process of the Spatial Attention Sub-Block is:

$$M_S(F) = \sigma(f_{conv}([F^S_{avg}, F^S_{max}])) \tag{3}$$

where σ denotes a sigmoid activation function, and f_{conv} the convolution layer. Gotten channel feature map F_C and spatial attention map M_S, the spatial feature map F_S is generated.

3.2 Transformer-Based Feature Fusion

The global and attribute features are regarded as independent of each other. We design a feature fusion model to refuse and refine the features, introducing the correlation between the attributes. Referring to the Transformer,[12] the feature fusion module is based on an encoder-decoder architecture, where the encoder is to construct the correlation between features, and the decoder is to generate the sequence of augmented features with fused correlation.

Concatenating the global feature f_{Global} and the attribute feature sequence f_{Attri}, a feature sequence of length N+1 can be described as $f_S = [f^0_S, f^1_S, \ldots, f^N_S]$, where $f^0_S = f_{Global}$, $f^i_S = f^i_{Attri}$, $i = 1, \ldots, N$. Since the elements of the feature sequence are not order-independent, we also combine a unique N-D positional encoding to the corresponding feature vector in the sequence. For the global feature, the position encoding is a N-D zero vector, and for the attribute feature, the position encoding is a one-hot vector corresponding to the attribute label. A new feature vector sequence $f'_S \in \mathbb{R}^{N\times(d_{f_S}+N)}$ is obtained as an input of the feature fusion module.

The structure of the Transformer-based feature fusion module is shown in Fig. 3. Both encoder and decoder are the stacked-layer architecture that consists of two kinds of sub-layer: *Multi-Head Attention* and *Fead Forward Network(FFN)*. In the following, we introduce the structure of the two sub-layers.

3.2.1 Multi-Head Attention (MHA)

The input of attention mechanism consists a set of queries $Q \in \mathbb{R}^{m\times d}$, a set of keys $K \in \mathbb{R}^{n\times d}$ and values $V \in \mathbb{R}^{n\times d}$. The calculation of the attention mechanism can be expressed as follows:

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

Instead of performing single attention for the queries, a multi-head attention mechanism is able to attend to diverse information from different representation subspaces, and enrich the expression correlation. The calculation process of the multi-task attention mechanism is:

$$\begin{gathered} MHA(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \\ where\ head_i = A(QW^Q_i, KW^K_i, VW^V_i) \end{gathered} \tag{5}$$

where $W^Q_i, W^K_i, W^V_i \in \mathbb{R}^d \times d_h$ are the projection matrices of the i-th head, and $W^O \in \mathbb{R}^{hd_h\times d}$ is the output projection matrices that aggregate information from all the heads.

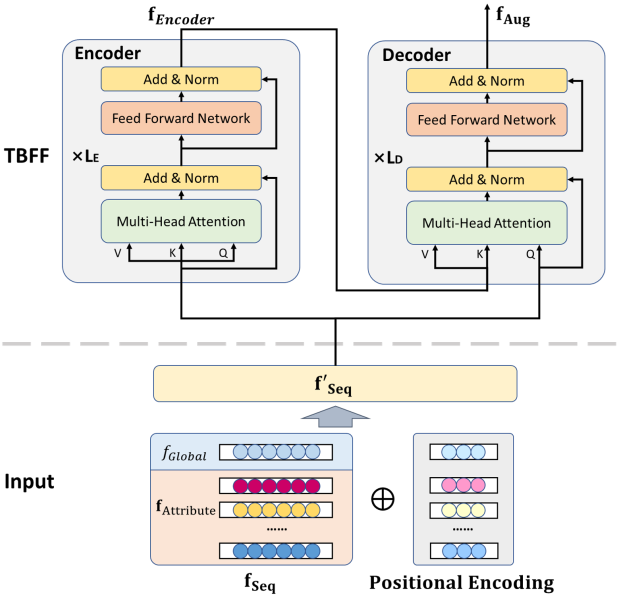

Figure 3. The framework of Transformer based Feature Fusion module. The feature sequence integrated by global feature, attribute feature sequence and position encoding are sent into a transformer-liked encoder-decoder model to generate an augmented feature sequence representation.

3.2.2 Feed Forward Network (FNN)

FFN takes input from MHA, and further transforms it using two linear layers with the activation function ReLU, which is expressed as follows:

$$FFN(x) = FC_1(ReLU(FC_0(x))) \tag{6}$$

As mentioned above, the input of the feature fusion module is f'_S. For the encoder, Q, K, V are represented as f'_S, and the output is represented as $f_{Encoder}$. For the decoder, Q is represented as f'_S while K and V are represented as $f_{Encoder}$. The output of the decoder is the augmented feature sequence f_{Aug} that takes the relevance between global and attribute features into account.

3.3 Multi-Task Optimization

For the augmented feature sequence from TBFF module $f_{Aug} = [f^0_{Aug}, f^1_{Aug}, \ldots, f^N_{Aug}]$, f^0_{Aug} represents the global feature and f^i_{Aug}, i=1,...,N represents a attribute feature. The global feature f^0_{Aug} is classified by a fully connected layer. Let the category number to be K, the output is represented as $z = [z_1, z_2, \ldots, z_K] \in \mathbb{R}^K$. The predict probability for each category k is calculated as $p(k|x) = \frac{exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)}$. The cross entropy loss of the category classification is formulated as below:

$$L_{Global}(z, y_0) = -\sum_{k=1}^{K} \log(p(k))q(k) \tag{7}$$

Let y be the ground-truth category label, so that $q(y) = 1$ and $q(k) = 0$ for all $k \neq y$. In this case, minimizing the loss is equivalent to maximizing the probability of being assigned to the ground-truth.

We also use another N fully connected layers to classify the attributes. For attribute i in K_i classes, the output is described as $z_i = [{z_i}^1, {z_i}^2, \ldots, {z_i}^{K_i}] \in \mathbb{R}^{K_i}$. Therefore, the predict probability for each category k_i is calculated as $p(k_i|x) = \frac{exp(z_k)}{\sum_{i=1}^{K_i} \exp(z_i)}$. The attribute classification can be defined as the average of all the attribute losses, so the attribute classification loss is formulated as:

$$L_{Attribute}(z, y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K_i} \log(p(k))q(k) \tag{8}$$

In summary, the loss function of the entire model can be expressed as:

$$L = L_{Global} + \alpha L_{Attribute} \tag{9}$$

where the hyperparameter α is able to adjust the contribution proportion of category and attributes classification.

4. EXPERIMENT

In order to verify the classification performance of MTAF, we designed two related experiments: the first is an own-collected real-world dataset for the benign and malignant identification of thyroid nodules; and the second is the widely used fine-grained image classification problem (public dataset CUB-200-2011[14]). For both datasets, we transformed the image size to 224 × 224 pixels as the network input. All experimental basic frameworks are pre-trained on the ImageNet dataset[15] and fine-tuned on the target dataset. The network was trained for 150 rounds and the batch size was set to 16. We set the layers of encoder $L_E = 6$ and the layers of decoder $L_D = 1$, weighting coefficient $\alpha = 2$ of the loss function. The optimization algorithms are all set to stochastic gradient descent. Noted that since our proposed approach involves attribute annotation, we compare our model with state-of-the-art algorithms that is integrate attribute annotation information.

4.1 Evaluation on Thyroid Nodule Classification

The diagnosis of thyroid nodules is one of the hot topics in the current medical imaging field. In the process of diagnosing benign and malignant thyroid nodules, the ultrasound doctor will focus on observing several core indicators such as calcification, composition, margin, and shape in the nodule ultrasound image as the key basis for diagnosis. By collaborating with the medical team, we collected an image collection with 2,285 ultrasound images from 1,790 patients, and the images were labeled by medical professionals as follows. For each nodule, 1) its benign and malignant were obtained by pathological sectioning conclusion (main category); 2) their approximate positions on the ultrasound image by point set; 3) labeling the four key nodule attributes of calcification, composition, margin, and shape. This is the first known dataset with core diagnostic indicators as thyroid ultrasound image attributes, and it also becomes the data basis for verifying our MTAF algorithm.

On the thyroid nodule classification task, we first compared three currently published algorithms for thyroid nodule classification. Among them, Ref. 16 and 17 simply applied VGG-16, GoogLeNet, and inception-v3 models to classify thyroid nodule images, and 18 used the weighted summation of two network classification probabilities as the classification results. Considering that the key innovation of our MTAF model is to introduce attribute information into the traditional multi-task learning framework, with two custom modules AAFD and TBFF to separate, fuse and reconstruct attribute attributes, we added two baseline models built by ourselves for deep comparison. With Ref. 19, baseline-1 uses the same network architecture as MTAF and contains the attention mechanism module AAFD, while excluding the attribute fusion module TBFF; baseline-2 adds a simplified version of the feature fusion module following Ref. 6. The comparative experiments are mainly investigating the influence of the Transformer Fusion module in the entire learning process.

For our thyroid dataset, due to the limitation of the size, we use a 10-fold cross test to ensure the reliability of the results. We use DenseNet-201[20] as the backbone of the algorithm. To follow the medical application convention, we use the accuracy and $\kappa - value$ to evaluate the performance of the classifier. The experimental results are shown in Tab. 1. In terms of accuracy and kappa value, the MTAF model exceeds all the comparative algorithms, including both of our customized baseline experiments. Experimental results have demonstrated the contribution of the Transformer fusion module to improve classification performance.

Table 1. Comparison with state-of-the-arts on thyroid nodule classification dataset.

Methods	Backbone	Accuracy	κ-value
[Ko et al., 2019][16]	VGG-16	81.1%	0.618
[Chi et al., 2017][17]	GoogLeNet Inception-v3	82.2%	0.643
[Li et al., 2019][18]	DarkNet 19+ResNet 50	81.5%	0.632
Baseline-1	DenseNet-201	82.4%	0.646
Baseline-2	DenseNet-201	82.7%	0.651
MTAF	DenseNet-201	**84.3%**	**0.678**

In addition to the main task (benign and malignant identification) comparative experiments, we also performed attribute classification experiments on the thyroid dataset. The experimental results are shown in Fig. 4. The MTAF model exceeds baseline-1 and baseline-2 in the accuracy of all the attribute classifications, which proves that MTAF uses Transformer technology for attribute feature fusion not only to help improve the accuracy of the main task classification but also to obtain a more reasonable attribute feature expression to improve the precision of attribute classification.

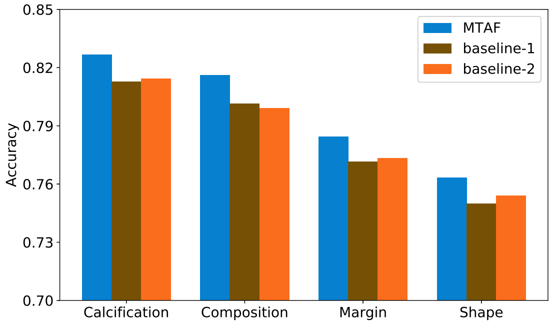

Figure 4. Comparison of the attribute classification results between MTAF and two baselines on thyroid nodule classification problem.

4.2 Evaluation on CUB-200-2011

CUB-200-2011 is a public benchmark dataset widely used on fine-grained image recognition, specifically for bird classification. It contains 11,788 pictures of 200 bird species. Each image is labeled with 15 local areas and 312 binary classification attributes, along with a bounding box. In this experiment, we transform the attribute information into 28 multi-class attributes. We will compare the performance of our algorithm with comparative experiments by accuracy.

On the CUB-200-2011 dataset, we compared three advanced fine-grained classification models[5,6,8] based on attribute features. Similar to the thyroid nodule experiment, we introduced the baseline experiment including the AAFD module while excluding the TBFF module, and the purpose is to investigate the role and the impact of the Transformer-based fusion mechanism. Noted that the backbone of the algorithms compared in this dataset are all set into ResNet-50.[21]

The experimental results are shown in Tab. 2. Our model has the highest accuracy, which is 1.3% higher than the best baseline. From the above experimental results, it can be seen that the fusion of attribute correlation information by MTAF helps to improve the performance of fine-grained classification problems, which confirms the rationality of model design.

Table 2. Comparison with state-of-the-arts on CUB-200-2011 dataset.

Methods	Accuracy
APR[6]	84.1%
Image+Part+Attribute[8]	85.5%
A3M[5]	86.2%
Baseline-1	84.8%
MTAF	**87.5%**

4.3 Interpretable Visualization

Fig. 5 shows the results of our visualization of two examples of the CUB-200-2011 dataset. We use the Grad-CAM algorithm[22] to visualize the areas of interest of the MTAF model during the calculation of attribute categories. We selected the bird's beak shape, wing color, body size, and main body color as the demonstration. The visualization results show that the area of interest of the model when extracting attribute features is concentrated in the area where the corresponding attribute is located, which further illustrates that our model can accurately describe the attribute features.

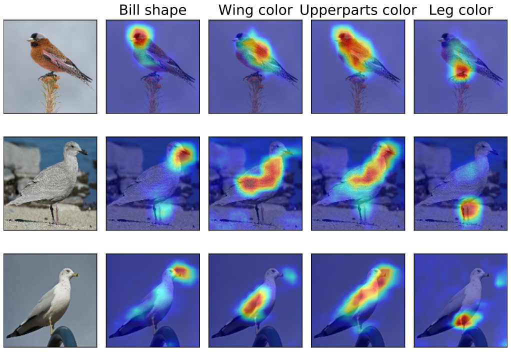

Figure 5. Examples of the Grad-CAM results performed by MTAF on CUB-200-2011 Dataset.

5. CONCLUSION

In this paper, we propose a multi-task attribute fusion model, MTAF, to solve the problem of fine-grained image recognition. We use the attention mechanism to separate the attribute features from the feature maps and introduce the Transformer architecture into the multi-task learning framework to fuse the global feature and attribute features to build the correlation between various attribute features. The experimental results show that our algorithm can effectively improve the model performance on multi-classification problems and obtain better attribute feature representations.

REFERENCES

[1] Guo, P. and Farrell, R., "Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition," in [*2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*], 1876–1885, IEEE (2019).

[2] Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., and Metaxas, D., "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1143–1152 (2016).

[3] Zhang, N., Donahue, J., Girshick, R., and Darrell, T., "Part-based r-cnns for fine-grained category detection," in [*European conference on computer vision*], 834–849, Springer (2014).

[4] Caruana, R., "Multitask learning," *Machine learning* **28**(1), 41–75 (1997).

[5] Han, K., Guo, J., Zhang, C., and Zhu, M., "Attribute-aware attention model for fine-grained representation learning," in [*Proceedings of the 26th ACM international conference on Multimedia*], 2040–2048 (2018).

[6] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., and Yang, Y., "Improving person re-identification by attribute and identity learning," *Pattern Recognition* **95**, 151–161 (2019).

[7] Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A., "Video action transformer network," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 244–253 (2019).

[8] Liu, X., Wang, J., Wen, S., Ding, E., and Lin, Y., "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in [*Thirty-First AAAI Conference on Artificial Intelligence*], (2017).

[9] Fu, J., Zheng, H., and Mei, T., "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4438–4446 (2017).

[10] Hu, J., Shen, L., and Sun, G., "Squeeze-and-excitation networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7132–7141 (2018).

[11] Woo, S., Park, J., Lee, J.-Y., and So Kweon, I., "Cbam: Convolutional block attention module," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], 3–19 (2018).

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," in [*Advances in neural information processing systems*], 5998–6008 (2017).

[13] Yu, J., Li, J., Yu, Z., and Huang, Q., "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology* (2019).

[14] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S., "The caltech-ucsd birds-200-2011 dataset," (2011).

[15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision* **115**(3), 211–252 (2015).

[16] Ko, S. Y., Lee, J. H., Yoon, J. H., Na, H., Hong, E., Han, K., Jung, I., Kim, E.-K., Moon, H. J., Park, V. Y., et al., "Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound," *Head & neck* **41**(4), 885–891 (2019).

[17] Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., and Eramian, M., "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *Journal of digital imaging* **30**(4), 477–486 (2017).

[18] Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., et al., "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet Oncology* **20**(2), 193–201 (2019).

[19] Buda, M., Wildman-Tobriner, B., Hoang, J. K., Thayer, D., Tessler, F. N., Middleton, W. D., and Mazurowski, M. A., "Management of thyroid nodules seen on us images: deep learning may match performance of radiologists," *Radiology* **292**(3), 695–701 (2019).

[20] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., "Densely connected convolutional networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4700–4708 (2017).

[21] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[22] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).