GAPNET: GENERIC-ATTRIBUTE-POSE NETWORK FOR FINE-GRAINED VISUAL CATEGORIZATION USING MULTI-ATTRIBUTE ATTENTION MODULE

Minjeong Ju[⋆] *Hobin Ryu*[⋆] *Sangkeun Moon*[†] *Chang D. Yoo*[⋆]

[⋆] Korea Advanced Institute of Science and Technology (KAIST)
[†] Korea Electric Power Corporation (KEPCO)

ABSTRACT

This paper proposes a multi-task learning framework for fine-grained visual categorization (FGVC) referred to as Generic-Attribute-Pose Network (GAPNet) that is capable of attending discriminating parts depending on the pose and part-attribute of an object using multi-attribute attention. FGVC is a challenging task that involves categorical data with small inter-class variation and large intra-class variation. Multi-Attribute Attention Module (MAAM) guides the GAPNet to focus on multiple parts of the image feature by emphasizing appropriate feature channels given both pose and part-attribute features. Experiments on Caltech-UCSD Birds and NABirds datasets demonstrate that GAPNet is competitive with other state-of-the-art methods, and ablation study on GAPNet conditioned on pose and part-attribute feature shows that GAPNet performs best when conditioned on both pose and part-attribute features.

Index Terms— Fine-Grained Visual Categorization, Multi-Task Learning, Attention Mechanism

1. INTRODUCTION

Fine-grained visual categorization (FGVC) is a task of distinguishing different species or object classes with *small inter-class variation* and *large intra-class variation*. First, different species/object classes in an image are distinguished by observing discriminating parts (e.g., the bird's beak) which is often localized to small areas in the species. This fact compels the need for an attention mechanism that can focus on multiple discriminating spatial regions. Second, an appearance of a species may look very different depending on the pose. This suggests that in order for the classifier to possess the capability to distinguish among different species or objects belonging to different classes, it must either be robust to pose or bear the capability of accurately predicting the class depending on the pose which must be predicted beforehand. The Grad-CAM++ [1] results reveal classifier to identify different bird species places attention on the head of the bird for classification in order to attain robustness against pose. However, it still suffers from large variations in pose and has difficulty deciphering discriminating parts of the object which can lead to accurate prediction.

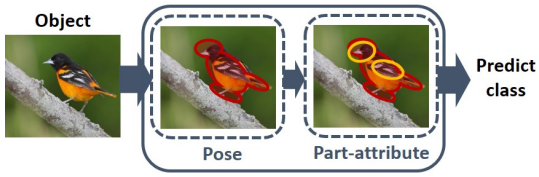

Fig. 1. The discriminative parts that determine the fine-grained species are located conditioned on the pose of a bird.

In order to overcome the challenges mentioned above, this paper proposes a Generic-Attribute-Pose Network (GAPNet) that is capable of discriminating various parts of the object under different pose using the Multi-Attribute Attention Module (MAAM) for accurately predicting the class. MAAM is capable of determining channels that are closest to query channels by taking a maximum of the cross-inner dot-product between the key-value channels and the query channels. Given pose together with part-attribute features as query, MAAM emphasizes channels of the image feature map that are best-matched with the part-attribute features. With the assistance of MAAM, multiple discriminating parts of the image feature is attended conditioned on the pose, as illustrated in Fig.1. Quantitative and qualitative evaluations for GAPNet on Caltech-UCSD Birds (CUB-Birds) [2] and NABirds [3] datasets demonstrates that GAPNet is competitive with the other state-of-the-art methods.

2. RELATED WORKS

2.1. Fine-Grained Visual Categorization (FGVC)

There are two prior approaches to FGVC which are relevant to the proposed method: a part-based localization and an end-to-end feature encoding. Part-based localization approach tries to locate various semantic parts of the objects. In [4], a fully convolutional network to locate multiple object parts and a two-stream classification network that encodes object-level and part-level cues are simultaneously proposed. In [5], an

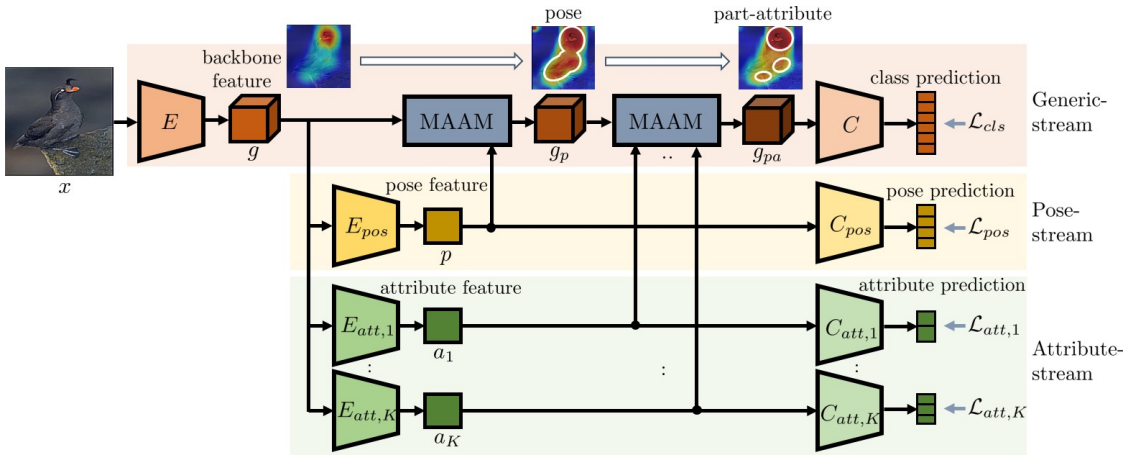

Fig. 2. The overall architecture of the proposed Generic-Attribute-Pose Network (GAPNet) for fine-grained classification.

algorithm that performs pose estimation and forms the unified object representation as the concatenation of hierarchical pose-aligned regions features, which is then fed into a classification network is proposed. [6] considers the object's part locations in weakly-supervised manners without part annotations; thus, it is hard to find precise part locations for these methods. Another line of work focuses on encoding features to contain discriminating regions of the object. [7, 8] consider spatially invariant modeling of pairwise features, based on bilinear and kernel methods. [9, 10] use multi-agents to exploit discriminative features while sharing information.

2.2. Attention Mechanism

Attention mechanism not only learns where to focus but also improves the way how regions of interests are represented which is conducive in better understanding the input image. [11] proposes a channel-wise attention mechanism using global average pooling to exploit the inter-channel relationship. [12, 13] sequentially applies channel-wise and spatial attention modules, which emphasizes the features of essential parts and dilutes the features of less important parts. FGVC methods with attention mechanism [14, 15] have been proposed to capture multiple discriminating parts of the objects.

3. METHODS

To exploit discriminating features obtained from multiple parts of the object, GAPNet incorporates MAAM to emphasizes features for inferring the pose and part-attributes.

3.1. Multi-Attribute Attention Module

Squeeze-and-Excitation (SE) module [12] has been commonly used as a channel-wise attention mechanism that learns which channel to highlight. The input feature X is "squeezed" by global average pooling and passed to a fully-connected layer so that the important channels of X are emphasized. This channel-wise attention SE module does not consider spatial information as spatial information is lost after the global average pooling. For accurate fine-grained classification, accentuating subtle discriminating regions which is localized to small regions in the image can be conducive, and for this reason, it is important to preserve the spatial information of the object.

For this reason, we propose a Multi-Attribute Attention Module (MAAM) to accentuate multiple discriminating parts of the object while preserving the spatial information of the image. MAAM takes two inputs, a key-value X and a query A, both of which are normalized using L2-norm. The key-value $X \in \mathbb{R}^{C\times H\times W}$ and the query $A \in \mathbb{R}^{K\times H\times W}$ are composed of channel feature $\{\mathbf{x}_c\}_{c=1}^{C}$ and $\{\mathbf{a}_k\}_{k=1}^{K}$, respectively. As shown in Fig.3, MAAM emphasizes each channel of the key-value X conditioned on the query A as follows:

$$
\begin{aligned}
S &= \begin{bmatrix} \mathbf{x_1}\cdot\mathbf{a_1} & \mathbf{x_1}\cdot\mathbf{a_2} & \cdots & \mathbf{x_c}\cdot\mathbf{a_k} \\ \mathbf{x_2}\cdot\mathbf{a_1} & \mathbf{x_2}\cdot\mathbf{a_2} & \cdots & \mathbf{x_c}\cdot\mathbf{a_k} \\ \vdots & & & \\ \mathbf{x_c}\cdot\mathbf{a_1} & \mathbf{x_c}\cdot\mathbf{a_2} & \cdots & \mathbf{x_c}\cdot\mathbf{a_k} \end{bmatrix} \in \mathbb{R}^{C\times K}, \\
\alpha &= \begin{bmatrix} \max(\mathbf{x_1}\cdot\mathbf{a_1}, \mathbf{x_1}\cdot\mathbf{a_2}, \cdots, \mathbf{x_c}\cdot\mathbf{a_k}) \\ \max(\mathbf{x_2}\cdot\mathbf{a_1}, \mathbf{x_2}\cdot\mathbf{a_2}, \cdots, \mathbf{x_c}\cdot\mathbf{a_k}) \\ \vdots \\ \max(\mathbf{x_c}\cdot\mathbf{a_1}, \mathbf{x_c}\cdot\mathbf{a_2}, \cdots, \mathbf{x_c}\cdot\mathbf{a_k}) \end{bmatrix} \in \mathbb{R}^{C}, \\
Z &= \alpha \odot X + X.
\end{aligned}
$$

MAAM computes cross-inner dot-product between each channel of the key-value X and that of query A to construct the similarity matrix S. By taking the maximum along the rows of S, the similarity weights $\boldsymbol{\alpha}$ is obtained. When the channel feature of the key-value and that of the query are similar, the cross-inner dot-product will be high. This process allows us align features across different streams with

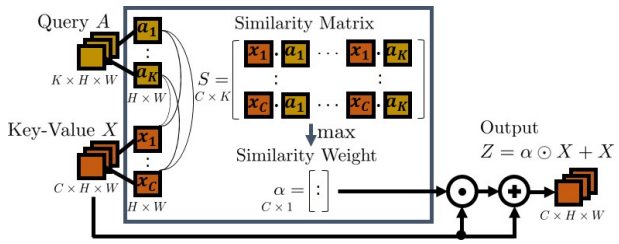

Fig. 3. Multi-Attribute Attention Module (MAAM).

the intensity determined by the dot-product score. Finally, output Z is obtained by multiplying the similarity weights $\boldsymbol{\alpha}$ with the key-value X in channel-wise followed by adding X as a residual connection, so that the output $Z \in \mathbb{R}^{C \times H \times W}$ emphasizes the query information by considering the correspondence between the query A and the key-value X. Using backbone image feature as the key-value, MAAM produces pose-conditioned and part-attribute-conditioned features using pose and part-attribute features as query.

3.2. Generic-Attribute-Pose Network (GAPNet)

GAPNet is composed of three streams, as illustrated in Fig.2. In the Generic stream, the backbone feature g is obtained as the output of the backbone feature extractor E with input image x,. In the Pose-stream, a pose feature extractor E_{pos} takes the backbone feature g and outputs a pose feature p which holds pose-specific information. In the Attribute-stream, K attribute feature extractors $E_{att,1}, \cdots, E_{att,K}$ extract K part-attribute features $a_1, a_2, \cdots, a_K$. The first MAAM in the Generic-stream emphasizes channels of the backbone feature g that have high correlation with the pose feature p to generate pose-conditioned feature g_p. Then, g_p is accentuated once again by the second MAAM in producing g_{pa} using the K part-attribute features $a_1, a_2, \cdots, a_K$ such that g_{pa} holds high concentration of detailed parts information such as shape and color. The final feature g_{pa} is passed to a fine-grained classifier C, and it predicts the category of the input image along with the prediction score. A fine-grained classification loss $\mathcal{L}_{cls}$ is evaluated by minimizing the cross-entropy. In order to obtain pose-specific and part-attribute information, the pose and attribute features are trained to predict the pose and the part-attributes, respectively, by minimizing each cross-entropy loss. First, the pose classifier C_{pos} takes the pose feature p, predicts the scores for each pose category e, and draws pose loss $\mathcal{L}_{pos}$. By clustering images with their P part location coordinates, images that contain objects with similar pose are grouped into a cluster, and the cluster's index is used as the ground-truth pose label of the images. Considering one of the P parts as a center point, relative coordinates of the other parts are calculated and represented as a vector $\mathbf{p} = (x_1, y_1, \cdots, x_{P-1}, y_{P-1})$. In order to be consistent with various image sizes, the vector $\mathbf{p}$ is normalized using L2-norm. Then, k-means clustering is conducted, and the results are used to label the pose of each image. Second, each part-attribute with binary multi-labels is predicted by K attribute classifiers $C_{att,1}, C_{att,2}, \cdots, C_{att,K}$, which are related to detail object's part information such as shape or color. For example, the 'tail color' of a bird is one of the part-attributes, which has a value of a mixture of red and yellow.

GAPNet is trained by minimizing the loss $\mathcal{L}$, which is a linear combination of the fine-grained classification loss $\mathcal{L}_{cls}$, pose loss $\mathcal{L}_{pos}$, and attribute loss $\{\mathcal{L}_{att,i}\}_{i=1}^{K}$. The values of loss weights λ_{pos} and λ_{att} are determined by experiments.

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{att}\frac{1}{K}\sum_{i}^{K}\mathcal{L}_{att,i}.$$

4. EXPERIMENTS

4.1. Datasets

CUB-Birds. CUB-Birds [2] is a well-known fine-grained benchmark dataset for birds classification. It contains 200 bird classes with 15 part locations, 312 binary attributes, and one bounding box, and divided into $5,994$ images of the training set and $5,794$ images of the test set. The locations of the six parts of birds are selected to generate the ground-truth pose labels, which have high correlations with their pose: 'Crown,' 'Beak,' 'Belly,' 'Left Wing,' 'Right Wing,' and 'Tail.' By considering the 'Belly' as the center point, the birds are grouped into nine clusters where each cluster corresponds to a pose label, as shown in Fig.4. Clustering with six parts performed better than using all the 15 parts, and the number of clusters is set to nine, which seems more meaningful than the number between 5 and 20. Moreover, $K = 28$ multi-label attributes are used as the part-attribute labels, which are parent attributes of the 312 binary attributes.

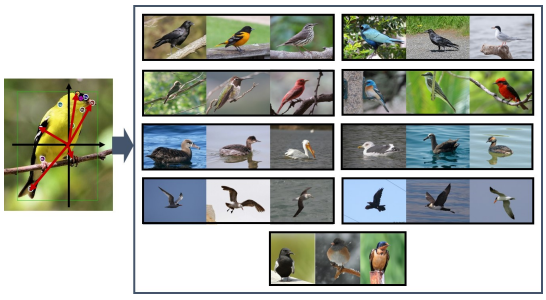

Fig. 4. The nine pose clusters on the CUB-Birds dataset.

NABirds. NABirds [3] is another fine-grained bird dataset, covering 555 bird species with $23,929$ training and $24,633$ test images. It has 11 part location annotations, which is the subset of the CUB-Birds parts, so the same six part locations as CUB-Birds were used to generate the pose labels. As it

does not have the binary attributes, the Attribute-stream in GAPNet is ignored, and only the Pose-stream is activated.

4.2. Quantitative Results

To evaluate GAPNet, extensive experiments are conducted as illustrated in Table 1-2. ResNet-50 [16] pre-trained on ImageNet dataset is employed as the backbone feature extractor E. The pose and attribute feature extractors are composed of one convolutional layer with a kernel size of 1. In Table 1, 'Patch-wise SE - $A_1A_2A_3$' denotes a variant of SE [11] that does not lose spatial information to some extent due to $A_i \times A_i$ patch-wise attention for stronger baseline that SE [11]. For both datasets, GAPNet outperforms the baseline methods including 'Patch-wise SE' and shows competitive performances with the other state-of-the-art methods [17, 18, 8, 19].

Method	Acc.(%)
ResNet-50 [16]	84.5
SE [11]	85.1
Patch-wise SE - 222/888/842	85.2/85.3/85.4
BAM [12]	84.8
CBAM [13]	84.9
MaxEnt-CNN [17]	80.4
FCAN [18]	84.3
Kernel-Pooling [8]	84.7
DT-RAM [19]	86.0
A^3M [20]	**86.2**
GAPNet w. Pose-Attribute	**86.2**

Table 1. Performance of GAPNet on CUB-Birds.

Method	Backbone	Acc.(%)
ResNet-50 [16]	ResNet-50	79.2
MaxEnt-CNN [17]	ResNet-50	69.2
MaxEnt-CNN [17]	DenseNet-161	83.0
GAPNet w. Pose Only	ResNet-50	**83.8**

Table 2. Performance of GAPNet on NABirds.

4.3. Ablation Study

Several experiments on CUB-Birds were conducted to observe the performance according to (1) the loss weights, (2) the order of two MAAMs for emphasizing the backbone feature, and (3) weakly-supervised pose label, as shown in Table 3. When using only one between the Pose- and Attribute-stream with various loss weights λ_{pos} and λ_{att}, the best accuracies were 85.9% and 85.5%, respectively. Applying MAAMs with the order of pose and attribute performs best, while the MAMMs with the reversed order also outperforms the GAP with only one MAAM. By clustering intermediate image features extracted from the pretrained ResNet-50 followed by PCA, pose labels are generated without part location annotations. GAPNet is trained in a weakly supervised manner with these part labels, and surprisingly shows comparable performance with the supervised one.

Method	λ_{pos}	λ_{att}	Acc.(%)
GAPNet w. Pose Only	0.2		85.6
GAPNet w. Pose Only	0.5		85.9
GAPNet w. Pose Only	0.7		85.6
GAPNet w. Pose Only	1.0		85.0
GAPNet w. Pose(weakly) Only	0.5		85.7
GAPNet w. Attribute Only		0.2	85.5
GAPNet w. Attribute Only		0.6	85.2
GAPNet w. Attribute Only		1.0	85.3
GAPNet w. Pose-Attribute	0.5	0.2	**86.2**
GAPNet w. Attribute-Pose	0.5	0.2	86.0

Table 3. Ablation study of GAPNet on CUB-Birds.

4.4. Qualitative Results

To demonstrate the effect of the two kinds of MAAMs, Grad-CAM++ [1] results that visualize where the model saw to predict the correct class are shown in Fig.5. While the ResNet-50 focuses only on a specific part of birds, GAPNet attends to overall parts after the first MAAM (g_p) and focuses on the multiple discriminative parts after the second MAAM (g_{pa}). The attention map is gradually calibrated as more information conditionally modulates the backbone feature (g).

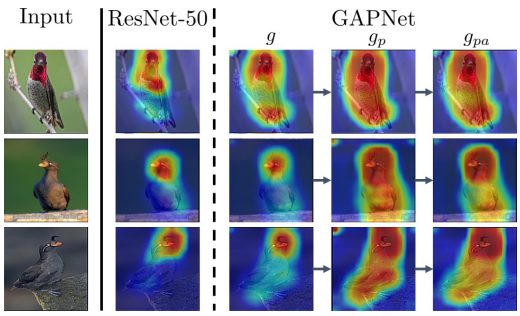

Fig. 5. The Grad-CAM++ results of GAPNet on CUB-Birds.

5. CONCLUSION

GAPNet based on MAAM is proposed for FGVC where there is small inter-class variation and large intra-class variation. MAAM guides GAPNet to highlight the most discriminating parts by modulating the backbone feature conditioned on pose and part-attribute. Extensive experiments demonstrate that the GAPNet is effective and explainable.

6. REFERENCES

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[2] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[3] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 595–604.

[4] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang, "Part-stacked cnn for fine-grained visual categorization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1173–1182.

[5] Pei Guo and Ryan Farrell, "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)*. IEEE, 2019, pp. 1876–1885.

[6] Weifeng Ge, Xiangru Lin, and Yizhou Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3034–3043.

[7] Shu Kong and Charless Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 365–374.

[8] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie, "Kernel pooling for convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2921–2930.

[9] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 420–435.

[10] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8331–8340.

[11] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[14] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[15] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 805–821.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik, "Maximum-entropy fine grained classification," in *Advances in Neural Information Processing Systems*, 2018, pp. 637–647.

[18] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin, "Fully convolutional attention networks for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, 2016.

[19] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu, "Dynamic computational time for visual attention," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1199–1209.

[20] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu, "Attribute-aware attention model for fine-grained representation learning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 2040–2048.