

Deep Learning Based Fine-Grained Species Identification

by Qiyu Liao

Thesis submitted in fulfilment of the requirements for
the degree of

C02029 Doctor of Philosophy

under the supervision of Min Xu, Dadong Wang

University of Technology Sydney
Faculty of Engineering and Information Technology

March, 2021

Certificate of Original Authorship Template

Graduate research students are required to make a declaration of original authorship when they submit the thesis for examination and in the final bound copies. Please note, the Research Training Program (RTP) statement is for all students. The Certificate of Original Authorship must be placed within the thesis, immediately after the thesis title page.

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Qiyu Liao declare that this thesis, is submitted in fulfilment of the requirements for the award C02029 Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

**If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).*

**If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).*

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 18/11/2021

Collaborative doctoral research degree statement

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Indigenous Cultural and Intellectual Property (ICIP) statement

This thesis includes Indigenous Cultural and Intellectual Property (ICIP) belonging to UTS and CSIRO, custodians or traditional owners. Where I have used ICIP, I have followed the relevant protocols and consulted with appropriate Indigenous people/communities about its inclusion in my thesis. ICIP rights are Indigenous heritage and will always remain with these groups. To use, adapt or reference the ICIP contained in this work, you will need to consult with the relevant Indigenous groups and follow cultural protocols.

Abstract

Fine-Grained Visual Categorization (FGVC) is a challenging research topic in computer vision. It deals with the classification of visual data at a subordinate level. This thesis investigates four categories of FGVC methods based on deep learning, including general convolutional neural networks, object part localization methods, approaches using CNN ensemble or higher-order feature encoding, and methods utilizing recurrent visual attention. Overall performance comparison has been conducted to analyse their advantages and disadvantages.

We proposed a new regression-based part detection structure and a novel part-based model, which increased the classification accuracy of PS-CNN from 76.4% to 82.4% on the CUB-200-2011 benchmark dataset.

Inspired by the second-order pooling, we proposed a highly interpretable method with a compressed structure to significantly reduce the computation complexity while improving the fine-grained categorization accuracy. The proposed model provides a supervised selection of the most discriminative second-order channels. With the proposed method, the computation and the feature dimension are linearly reduced to 4% of the original bilinear pooling. By applying matrix normalization and a Fisher-Recurrent-Attention structure, we achieved the best result among the VGG-16 based FGVC models.

Following the conception of attention crop and attention drop in the Fisher-Recurrent-Attention model, we proposed a forcing module to constrain the network to extract more diverse features for FGVC. The forcing module focuses more on confusion regions which are essential for the fine-grained classification. Experimental results show that the proposed forcing module can improve the attention and prediction of the network when an input image is panned or zoomed, and the double prediction performs better than the single prediction.

The existing FGVC methods often come with enormous amounts of computation and require large memory space. This makes these models inadequate for mobile applications. We proposed a Category Attention Transferring Convolutional Neural Network (CAT-CNN) to transfer the attention knowledge from a large-scale FGVC network to a small but efficient network to improve its presentation capability. Using the proposed model, we improved the classification accuracy of the efficient networks by up to 5.7% on the CUB-2011-200 dataset without increasing computation time or memory cost, which makes FGVC feasible on mobile devices.

We also conducted abundant studies to investigate the relationship between attention and classification accuracy of our proposed deep learning models, visualized and analysed the attentional activations of these models. We hope that our findings may inspire further research efforts to advance the FGVC for a wide range of real-world applications.

keywords: Image classification, Fine-grained classification, Deep learning, Species identification.

Contents

1	Introduction	1
1.1	Background, Motivation and Challenges	3
1.2	Main Contributions	5
1.3	Research Objectives	7
1.4	Thesis Outline	7
2	Literature Review	8
2.1	Traditional Methods	9
2.1.1	General CNN Backbones	9
2.1.2	Destruction and Construction Learning (DCL)	14
2.1.3	Selective Sparse Sampling Net (S3N)	15
2.1.4	Anti-Perturbation Inference Net (API-Net)	15
2.1.5	Mutual-Channel Loss (MC-Loss)	16
2.1.6	Attribute Mix+	16
2.2	Part Based Methods	18
2.2.1	Part-Based R-CNN (PBR-CNN)	19
2.2.2	Pose Normalized Nets	21
2.2.3	Multi-Proposal Consensus (PL-MPC)	21
2.2.4	Deep Localisation, Alignment, and Classification (Deep LAC)	23
2.2.5	Part-Stack CNN (PS-CNN)	24
2.2.6	AttNet&AffNet	26
2.3	Model Duplication and Feature Encoding Methods	26
2.3.1	Subset Feature Learning Networks	27
2.3.2	Mixture of Deep CNN (MD-CNN)	28
2.3.3	CNN Tree	29
2.3.4	Multiple Granularity CNN	29
2.3.5	Bilinear CNN Models	30
2.4	Attention-Based Methods	32

2.4.1	FCN Attention	33
2.4.2	Diversified Visual Attention Net (DVAN)	34
2.4.3	Recurrent Attention CNN (RA-CNN)	35
2.4.4	Multi-Attention CNN (MA-CNN)	37
2.4.5	Navigator-Teacher-Scrutinizer Network (NTS-Net)	37
2.4.6	Mixture of Granularity-Specific Experts CNN (MGE-CNN)	38
2.4.7	Weakly Supervised Data Augmentation Network (WS-DAN)	39
2.4.8	Weakly Supervised Complementary Parts Models	40
2.4.9	Diversification Block (DB)	41
2.5	Fine-Grained Visual Classification Datasets	42
2.6	Performance Comparison and Analysis	43
3	Part Collaboration Convolutional Neural Network	48
3.1	PC-CNN Structure	49
3.1.1	Part Localization Network	50
3.1.2	Two Stream Feature Extraction Structure	52
3.1.3	Feature Unifying Layer and Classifier	54
3.1.4	Training Methodology	55
3.2	Experiment Result of PC-CNN	56
3.2.1	Implementation Details	56
3.2.2	Part Localization Result	57
3.2.3	Categorization Results	57
3.2.4	Discussion on the Proposed PC-CNN	60
3.3	Conclusion	61
4	Squeezed Bilinear Pooling	63
4.1	Squeezed Bilinear CNN	64
4.1.1	Fisher Feature Selector	66
4.1.2	Squeezed Bilinear CNN	69
4.1.3	Fisher Recurrent Attention Structure	70
4.2	Experimental Results of SBP	71
4.2.1	Implementation Details	71
4.2.2	Configuration and Comparison with Compact Bilinear	72
4.2.3	Experiments with Different Datasets	74
4.2.4	Validation via Visualization	76
4.3	Conclusion	77

5	Learning Enhanced Features and Inferring Twice	78
5.1	Method	79
5.1.1	Forcing Net	80
5.1.2	CAM-Based Cropping	81
5.2	Experimental Results of F-Net	82
5.2.1	Implementation Details	83
5.2.2	Quantitative Results	83
5.2.3	Ablation Study	84
5.3	Conclusion	87
6	Category Attention Teaching CNN	88
6.1	Category Attention Teaching CNN	90
6.2	Experimental Result of CAT-CNN	92
6.2.1	Dataset and Implementation Details	92
6.2.2	Configuration and comparison with the baseline	93
6.2.3	Fast-Slow CAT-CNN for General FGVC	94
6.2.4	Cross Model CAT-CNN for Efficient FGVC	96
6.2.5	Validation via Visualization	97
6.3	Conclusion	98
7	Conclusions and Future Work	99
7.1	Conclusions	99
7.2	Future Work	100
7.2.1	Extend the Squeezed Bilinear Pooling to General Structure . .	100
7.2.2	Weakly Supervised Part Discovery	103
8	Publications	104
	Bibliography	106

List of Figures

1.1	Two subclasses of gulls from the general FGVC dataset, CUB-2011-200, illustrate the major challenge in FGVC. (a) and (b) shows California gulls and Glaucous gulls, respectively. The inner-class variances, e.g., pose, scale, etc., are more visually obvious than the inter-class variances.	2
2.1	Framework of AlexNet [52].	10
2.2	Inception module of GoogLeNet [84].	10
2.3	The architecture of VGG-16 [80].	11
2.4	A residual block.	11
2.5	Model scaling in [85]. The proposed compounding scaling (e) combines the three scaling (b,c,d) into a single strategy.	12
2.6	The prediction accuracy and model size comparison with multiple models on the ImageNet dataset [85].	13
2.7	The general framework of the Destruction and Construction Learning [10].	14
2.8	The general framework of Selective Sparse Sampling Net [16].	15
2.9	Overview of the inference procedure of Anti-Perturbation Inference Net [18].	16
2.10	MC-Loss layer in a CNN structure [9].	17
2.11	The inference structure of the MCL [9]. (a) is the MCL components and (b) is the output feature from CNN with/without MCL.	17
2.12	The workflow of Mix+ model [54].	18
2.13	Rough framework of traditional part detection-based methods.	19
2.14	Workflow of the Part-based R-CNN [108].	20
2.15	Pose normalized nets [6].	21
2.16	The workflow of the multi-proposal consensus [79].	22
2.17	Framework of deep Localisation, Alignment and Classification [59].	23
2.18	Part-stack CNN [45].	24

2.19	The workflow of AttNet&AffNet [35].	26
2.20	Framework of subset feature learning networks [30].	27
2.21	Framework of MixDCNN [29].	29
2.22	Framework of CNN tree [94].	30
2.23	Framework of Multiple Granularity NN [91].	31
2.24	Framework of bilinear CNN model [62].	31
2.25	Framework of FCN attention [64].	34
2.26	The framework of diversified visual attention networks [111].	35
2.27	The DVAN attention component [111].	35
2.28	The workflow of recurrent attention CNN [22].	36
2.29	The workflow of Multi-attention CNN [112].	37
2.30	The workflow of the NTS-Net. The feature extractor extracts the deep feature map from each input image, and feeds the feature map into to Navigator network to generate the activations. Here we select the top-3 activations and crop the correspondent regions as the second layer classification navigator [103].	38
2.31	The workflow of the MGE-CNN [107].	39
2.32	The workflow of Weakly Supervised Data Augmentation Network [42].	40
2.33	The workflow of Weakly Supervised Complementary Parts Models [28].	41
2.34	The workflow model with Diversification Blocks [81].	42
2.35	Samples from the three benchmark FGVC datasets. For each dataset, three images from each of randomly selected two classes are presented to show intuitively the FGVC datasets used in our experiments. . . .	44
3.1	The network architecture of the proposed PC-CNN categorization model. With the PC-CNN, 1) the input image is resized into 224×224 for object stream, and cropped into M 113×113 parts for part streams; 2) M part streams take cropped part images as input, and extract the most representative $M \times 16 \times 3 \times 3$ features independently; 3) the object stream takes the object image as input and extract a global 2048-dimensional feature set; 4) unify object and part features and utilize pose information to achieve the final multi-view feature for the classifier with 3 fully connected layers.	50

3.2	The percentage of the images in CUB-2011-200 dataset that contains each part of a bird. 99.64% of the images in the dataset contain beak, while the back only appears in 76.89% of the images. Due to occlusion, right wing, left wing, right eye, and left eye only show in around half of the images.	51
3.3	The localization network architecture. The model consists of: 1) image feature extraction network; 2) 2 fully connected layer part location prediction network; 3) 2 fully connected layer object part detection network; 4) unify part detection and predicted location output to generate the final localization.	53
3.4	Localization samples: predicted part locations and ground-truth part locations are shown in red and blue dots, respectively, the green lines show their correspondences. (a) Correct localizations on different species and environments, and (b) Some representative samples mis-predicted by the proposed model.	58
3.5	Comparison among the output heatmaps of independent and shared parameter CNNs. (1) the maximal activation maps from the outputs of part streams; (2) the maximal activation map from the outputs of object stream; (3) the heatmaps of different parts cropped from (2).	61
4.1	The proposed network architecture with three SBP based models: 1) the Squeezed Bilinear Pooling with Element-wise Normalization (SBP-EN) for fast computation, 2) the Squeezed Bilinear Pooling with Matrix Normalization (SBP-MN) by inserting the matrix square root function before the squeezing layer, and 3) the Fisher Recurrent Attention Squeezed Bilinear Pooling (FRA-SBP).	65
4.2	Classification error rate on the Cub dataset. Comparisons are made on the proposed SBP without matrix normalization and Compact Bilinear Pooling (CBP) with Tensor Sketch. Horizontal lines are the baseline performances of Fully Bilinear Pooling (FBP). ft and woft stands for with and without global fine-tuning of CNN, respectively.	73

4.3	Visualization of Squeezed Bilinear Pooling and its activation across CNN channels. 1) shows the normalized Fisher scores for the inner products of the eight most activated channels. 2) crops the activated regions of the channels and abstracts the semantemes. 3) marks overlapped regions of activations across channels to verify the effectiveness of the Fisher score measurement.	76
5.1	Overview of the proposed F-Net which consists of the feature extracting module and the forcing module. The feature extracting module is convolutional layers that extract features. The forcing module contains the original branch and the forcing branch.	80
5.2	Overview of CAM-based cropping module.	81
5.3	Visualisation of our method. The one to the left of the dotted line is where the first prediction was wrong and the second prediction was correct, and the final prediction was correct. The examples shown to the right of the dotted line are those with the first, the second, and final predictions are all correct. Each of these examples from left to right includes the original image, top-1 class activation map of the original image, prediction of the original image, cropped image, top-1 class activation map of the cropped image, prediction of the cropped image, the summation of the prediction from the original image, and the prediction from the cropped image.	85
6.1	The proposed network architecture with three training phases: 1) pre-training the category attention teacher using the class labels, 2) training the category attention teaching assistant with the supervision from the class attention of the phase one and the class labels, and 3) cross-model category attention teaching for efficient networks using the supervision from the class attention of the phase two and the class labels.	91
6.2	Visualization of the accuracy of Category Attention Teaching model with its teaching rate r_t . From left to right, the three graphs stand for ShuffleNet-V2-Large-1.0, MobileNet-V3-Large-1.0 and EfficientNet-b0 taught by a pretrained EfficientNet-b7 model. The blue lines represent the prediction accuracies of CAT-CNNs with different teaching rates. The orange lines demonstrate the baselines when we only use the corresponding student model without CAT structure.	93

6.3	Visualization of the category activation of the CAT-ShuffleNet and comparison with the teacher steam (EfficientNet-b7) and the original ShuffleNetV2-Large-1.0. The column labeled with "Image" is the original input image. The "Efficient-b7" column shows the output of the maximum activated category of the category attention maps. The "ShuffleNet" shows the maximum category activations of the ShuffleNetV2-Large-1.0, which is pretrained with the one-hot label only. The column labeled with "CAT-ShuffleNet" shows the maximum category activations of the proposed CAT-ShuffleNet model, which is trained with Category Attention Teaching from a pretrained EfficientNet-b7 model.	95
7.1	The CUB-200-2011 prediction accuracy of SBP on four different backbones: (a). MobileNet-V3; (b). ShuffleNet-V3; (c). EfficientNet-b0; (d). EfficientNet-b7. The red lines indecate the accuracies using different featrue dimension. The blue lines of dashes are the baseline of the original structures without SBP, and the blue stars is the feature dimension of the orginal structures.	102