Deep Representation Learning for Keypoint localization

Shaoli Huang

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2017

To my family
Mingjiang Liang and *Jingyi Huang*

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Shaoli Huang

Acknowledgements

First and foremost, I would like to thank my supervisor **Prof. Dacheng Tao**, who not only guide me to the field of computer vision but also provide me advice on life and careers.

I would also like to thank my parents, my brother and my sisters for giving me love and support. I am very thankful to my dear wife Mingjiang Liang, who has been with me these years. She takes care of the family and allows me spending more time on the research study. I am also thankful for the unwavering love and general happiness that she has brought into my life. Along with her, I want to thank my daughter, Jingyi Huang. She has been a pure joy and has made my life much more fun. I am also thankful to my mother-in-law Fengying Lei, who takes care of my family when I was writing the thesis.

I also would like to give special thanks to Mingming Gong for numerous discussions that have played a significant role in bringing clarity to my ideas. I also would like to thank Dr. Jun Li and Dr. Zhe Xu who spend much time on having a discussion with me.

Finally, I would like like to thank my colleagues and friends I met in Sydney: Shirui Pan, Ruxin Wang, Tongliang Liu, Chang Xu, Haishuai Wang, Huan Fu and so many others.

Abstract

Keypoint localization aims to locate points of interest from the input image. This technique has become an important tool for many computer vision tasks such as fine-grained visual categorization, object detection, and pose estimation. Tremendous effort, therefore, has been devoted to improving the performance of keypoint localization. However, most of the proposed methods supervise keypoint detectors using a confidence map generated from ground-truth keypoint locations. Furthermore, the maximum achievable localization accuracy differs from keypoint to keypoint, because it is determined by the underlying keypoint structures. Thus the keypoint detector often fails to detect ambiguous keypoints if trained with strict supervision, that is, permitting only a small localization error. Training with looser supervision could help detect the ambiguous keypoints, but this comes at a cost to localization accuracy for those keypoints with distinctive appearances. In this thesis, we propose hierarchically supervised nets (HSNs), a method that imposes hierarchical supervision within deep convolutional neural networks (CNNs) for keypoint localization. To achieve this, we firstly propose a fully convolutional Inception network with several branches of varying depths to obtain hierarchical feature representations. Then, we build a coarse part detector on top of each branch of features and a fine part detector which takes features from all the branches as the input.

Collecting image data with keypoint annotations is harder than with image labels. One may collect images from Flickr or Google images by searching keywords and then perform refinement processes to build a classification dataset, while keypoint annotation requires human to click the rough location of the keypoint for each image. To address the

problem of insufficient part annotations, we propose a part detection framework that combines deep representation learning and domain adaptation within the same training process. We adopt one of the coarse detector from HSNs as the baseline and perform a quantitative evaluation on CUB200-2011 and BirdSnap dataset. Interestingly, our method trained on only 10 species images achieves 61.4% PCK accuracy on the testing set of 190 unseen species.

Finally, we explore the application of keypoint localization in the task of fine-grained visual categorization. We propose a new part-based model that consists of a localization module to detect object parts (where pathway) and a classification module to classify fine-grained categories at the subordinate level (what pathway). Experimental results reveal that our method with keypoint localization achieves the state-of-the-art performance on Caltech-UCSD Birds-200-2011 dataset.

Contents

Contents i

List of Figures v

List of Tables ix

1 Introduction 1
- 1.1 Objectives and Motivation . . . 1
- 1.2 Problems and Challenges . . . 3
  - 1.2.1 Keypoints Localization . . . 3
  - 1.2.2 Human Pose Estimation . . . 5
  - 1.2.3 Bird Part Localization . . . 8
- 1.3 Convolutional Neural Network . . . 9
- 1.4 Fine-grained Visual Categorization . . . 10
- 1.5 Contributions and Thesis Outline . . . 11
  - 1.5.1 Contributions . . . 11
  - 1.5.2 Outline . . . 12

2 Hierarchically Supervisided Nets for Keypoint Localization 14
- 2.1 Introduction . . . 16
- 2.2 Related Works . . . 18
  - 2.2.1 Bird part detection . . . 18
  - 2.2.2 Human pose estimation . . . 19
- 2.3 Hierarchically Supervised Nets . . . 20
  - 2.3.1 Network Architecture . . . 20

2.3.2 Learning and Inference 25
2.4 Experiments 29
2.4.1 Bird Part Localization 31
2.4.2 Human Pose Estimation 33
2.5 Conclusion 34

3 Transferring Part Locations Across Fine-grained Categories 35
3.1 Introduction 35
3.2 Relate Works 38
3.2.1 Part Detection. 38
3.2.2 Domain Adaptation and Active Learning 39
3.3 Our Approach 39
3.3.1 Model Formulation 39
3.3.2 Optimization with Backpropagation 43
3.4 Experiments 44
3.4.1 Datasets and Setting 44
3.4.2 Results and Analysis 45
3.5 Conclusions 46

4 Fine-grained Categorization with Part Localization 48
4.1 Introduction 48
4.2 Related Work 52
4.2.1 Keypoint Localization 52
4.2.2 Fine-Grained Visual Categorization 53
4.3 Part-Stacked CNN 56
4.3.1 Localization Network 57
4.3.2 Classification network 58
4.4 Deeper Part-Stacked CNN 61
4.4.1 Localization Network 62
4.4.2 Classification network 67
4.5 Experiments 69
4.5.1 Dataset and implementation details 70
4.5.2 Localization results for PSCNN 70

4.5.3 Classification results for PSCNN 72
4.5.4 Localization Results for DPSCNN 74
4.5.5 Classification results for DPSCNN 79
4.5.6 Model interpretation 82
4.6 Conclusion 83

5 Conclusions **87**

References **89**

List of Figures

1.1 Illustrating the pose estimation problem. 6
1.2 Illustrating the challenges of human pose estimation. 7
1.3 Illustrating the bird part localizatoin problem. 8

2.1 An illustration of the predicted keypoints from our HSN architecture. The left image contains highly accurate keypoints detected by the fine detector with strict supervision, the middle image contains keypoints from coarse detectors with loose supervisions, and the right image shows the final predictions by unifying the fine and coarse detectors. 15
2.2 Network architecture of the hierarchically supervised nets. The coarse stream learns three coarse detectors using hierarchical supervisions and while the fine stream learns a fine detector via strict supervision. Then the coarse predictions and fine predictions are unified for final prediction in inference stage. 21
2.3 Different methods for obtaining multiple-scale . (a) Input multiple resolution images. (b) Using different size of convolutional filters (c) concatenation of different resolutions of feature maps. (d) concatenation of feature maps from different layers, each of which has multiple convolutional filters. 24
2.4 An illustration of 26
2.5 Bird part detection results with occlusion,viewpoint, clustered background, and pose from the test set. 28
2.6 Pose estimation results with occlusion, crowding, deformation, and low resolution from the COCO test set. 32

LIST OF FIGURES

3.1 Illustration of the research problem. The source domain contains part annotations, while parts are not annotated in the target domain. Also, the target domain contains species which do not exist in the source domain. 37

3.2 The proposed architecture consist of three components: a feature extractor (yellow), a part classifier, and a domain classifier (blue). All these components share computation in a feed-forward pass. The feature extractor outputs feature representation as the input of the other components. The part classifier is designed to find the part location, while domain classifier is added to handle the domain shift between source and target domain. Note that the backpropagation gradients that pass from domain classifier to the feature extractor are multiplied by a negative constant during the backpropagation. 40

4.1 Overview of the proposed approach. We propose to classify fine-grained categories by modeling the subtle difference from specific object parts. Beyond classification results, the proposed DPS-CNN architecture also offers human-understandable instructions on how to classify highly similar object categories explicitly. 49

4.2 Illustration of the localization network. (a). Suppose a certain layer outputs feature maps with size 3x3, and the corresponding receptive fields are shown by dashed box. In this paper, we represent the center of each receptive filed with a feature vector at the corresponding position. (b). The first column is the input image. In the second image, each black dot is a candidate point which indicates the center of a receptive field. The final stage is to determine if a candidate point is a particular part or not. 54

4.3 The network architecture of the proposed Part-Stacked CNN model. The model consists of 1) a fully convolutional network for part landmark localization; 2) a part stream where multiple parts share the same feature extraction procedure, while being separated by a novel part crop layer given detected part locations; 3) an object stream with lower spatial resolution input images to capture bounding-box level supervision; and 4) three fully connected layers to achieve the final classification results based on a concatenated feature map containing information from all parts and the bounding box. 56
4.4 Demonstration of the localization network. The training process is denoted inside the dashed box. For inference, a Gaussian kernel is then introduced to remove noise. The results are M 2D part locations in the 27×27 *conv5* feature map. 58
4.5 Demonstration of the localization network. Training process is denoted inside the dashed box. For inference, a Gaussian kernel is then introduced to remove noise. The results are M 2D part locations in the 27×27 *conv5* feature map. 62
4.6 Network architecture of the proposed Deeper Part-Stacked CNN. The model consists of: (1) a fully convolutional network for part landmark localization; (2) a part stream where multiple parts share the same feature extraction procedure, while being separated by a novel part crop layer given detected part locations; (3) an object stream to capture global information; and (4) Feature fusion layer with input feature vectors from part stream and object stream to achieve the final feature representation. 65
4.7 Different strategies for feature fusion which are illustrated in (a) Fully connected,(b) Scale Sum, (c) Scale Max and (d) Scale Average Max respectively. 66
4.8 Typical localization results on CUB-200-2011 test set. We show 6 of the 15 detected parts here. They are: beak (red), belly (green), crown (blue), right eye (yellow), right leg (magenta), tail (cyan). Better viewed in color. 71

4.9 Typical localization results on CUB-200-2011 test set. Better viewed in color. 77
4.10 Feature maps visualization of *Inception-4a* layer. Each example image is followed by three rows of top six scoring feature maps, which are from the part stream, object stream and and baseline BN-inception network respectively. Red dash box indicates a failure case of visualization using the model learned by our approach. 78
4.11 Example of the prediction manual generated by the proposed approach. Given a test image, the system reports its predicted class label with some typical exemplar images. Part-based comparison criteria between the predicted class and its most similar classes are shown in the right part of the image. The number in brackets shows the confidence of classifying two categories by introducing a specific part. We present top three object parts for each pair of comparison. For each of the parts, three part-center-cropped patches are shown for the predicted class (upper rows) and the compared class (lower rows) respectively. 86

List of Tables

2.1 Comparison with methods that report per-part PCK(%) and average PCK(%) on CUB200-2011. The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead,Left Eye,Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing, Tail, and Throat 25

2.2 Comparison of PCP(%) and over-all PCP(%) on CUB200-2011. The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead, Eye, Leg, Wing, Nape,Tail, and Throat 25

2.3 Performance comparison between using strict supervision only and hierarchical supervision. 30

2.4 Results on COCO keypoint on test-dev and test-standard split . . 30

3.1 Part transferring results for different splits of CUB200-2011 dataset. Per-part PCKs(%) and mean PCK(%) are given.The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead,Left Eye,Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing, Tail, and Throat 44

3.2 Part transferring from CUB200-2011(Source) to BirdSnap(Target). Per-part PCKs(%) and mean PCK(%) are given. 45

4.1 *APK* for each object part in the CUB-200-2011 test set in descending order. 71

4.2 Comparison of different model architectures on localization results. "conv5" stands for the first 5 convolutional layers in CaffeNet; "conv6(256)" stands for the additional 1×1 convolutional layer with 256 output channels; "cls" denotes the classification layer with $M + 1$ output channels; "gaussian" represents a Gaussian kernel for smoothing. 72
4.3 The effect of increasing the number of object parts on the classification accuracy. 72
4.4 The effect of increasing the number of object parts on the classification accuracy. 73
4.5 Comparison with state-of-the-art methods on the CUB-200-2011 dataset. To conduct fair comparisons, for all the methods using deep features, we report their results on the standard seven-layer architecture (mostly *ALexNet* except *VGG-m* for [52]) if possible. Note that our method achieves comparable results with state-of-the-art while running in real-time. 74
4.6 Receptive field size of different layers. 76
4.7 Comparison of per-part PCK(%) and over-all APK(%) on CUB200-2011. The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead,Left Eye,Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing,Tail, and Throat 76
4.8 Localization recall of candidate points selected by *inception-4a* layer with different α values. The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead,Left Eye,Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing, Tail, and Throat 77
4.9 Localization recall of candidate points selected by *inception-4a* layer with different α values. The abbreviated part names from left to right are: Back, Beak, Belly, Breast, Crown, Forehead,Left Eye,Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing, Tail, and Throat 80
4.10 Comparison of different settings of our approach on CUB200-2011 . 80

4.11 Comparison with state-of-the-art methods on the CUB-200-2011 dataset. 81