Learning from Imperfect Supervision in Visual Pattern Classification and Localization

by Fan Ma

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Yi Yang

University of Technology Sydney

Faculty of Engineering and Information Technology

August 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Fan Ma declare that this thesis, is submitted in fulfilment of the requirements for the award of *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.

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

Signature: Production Note:
Signature removed prior to publication.

Date: 24th August 2022

Abstract

Machine learning algorithms have achieved tremendous success on various computer vision tasks in past decades. Large-scale well-annotated data, such as ImageNet and ActivityNet, are necessary for learning a valuable model. However, high-quality training samples are often insufficient in practice, and it is labor-intensive and time-consuming to produce intense supervision for different learning tasks. Designing algorithms with imperfect training data thus becomes significant in the current data explosion era.

In this dissertation, imperfect supervision is categorized into three classes: 1) Limited supervision where only a small portion of training samples are annotated; 2) Noisy supervision where some labels of training samples are corrupted; 3) Weak supervision where the labels of training data are imprecise to provide expected outputs. Several models are developed to learn from the supervision of different data types. A self-paced co-training algorithm is proposed to improve the model performance when limited training samples are available. I have also proved that our algorithm can achieve a better model with diverse classifiers. Moreover, a self-reweighting mechanism based on online learned class centroids is introduced to prevent the model from deteriorating by noisy supervision. Experiments are conducted on several image recognition datasets demonstrating the superiority of our designed algorithms under both limited and noisy supervision. Furthermore, two practical applications of temporal localization are studied when weak supervision is available. The first task is the temporal action localization, where only a single frame is annotated for each action instance. The goal is to produce precise temporal boundaries for action instances. An efficient frame expanding algorithm has been introduced to improve the temporal action localization performance. The other task uses query language to temporally localize moments in videos where only language-video pairs are available in the training data. The connections between the video clips and concepts in query sentences are formed by decoupling the core concepts in the query sentence.

This thesis demonstrates that our well-designed algorithms yield excellent results when only imperfect data are available in various vision tasks, ranging from image classification, object detection, and temporal localization in videos.

Dissertation directed by Professor Yi Yang
Australian Artificial Intelligence Institute, School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney

ACKNOWLEDGMENTS

First and foremost, I want to acknowledge the help and guidance of my principal supervisor, Prof. Yi Yang. Yi is a brilliant person not only in research fields but also in social communication. In the past four years, I have learned a lot from him. He encourages and helps me to build industry connections to stay at the forefront of several computer vision tasks. Moreover, he always points out my problem quickly and instructs me to make life-long plans. “Think before doing” is the most vital spirit I have learned from him. His calm and rational advice has been essential to my survival of several years as a graduate student and to my development as a researcher.

I would also like to thank my co-supervisor Prof. Deyu Meng, who was also my master tutor. I got to know Yi because of his recommendation. He always encourages me to stay focused on the research and to do some exciting work. His passion and optimism for research and life also inspire me to enjoy the process of doing academic research. Many thanks to my co-authors, Linchao Zhu, Zheng Shou, Xin Yu, Zhicheng Yan, Yu Wu. They have provided many insightful ideas and contributed a lot to the paper writing. I also want to thank my colleagues, Xiaohan Wang, Ruijie Quan, and Zhun Zhong, who are also my roommates and friends. We have had a wonderful time in the past four years.

I also want to thank my group members and colleagues, Xiaojun Chang, Jiaxu Miao, Hehe Fan, Pingbo Pan, Yanbin Liu, Zongxin Yang, Qingji Guan, Yawei Luo, Qianyu Feng, Yunqiu Xu, Guang Li, Youjiang Xu, Xuanmeng Zhang, Yuhang Ding, Guangrui Li, Yang He, Ping Liu, Yutian Lin, Peike Li, Xuanyi Dong, Tianqi Tang, Bingwen Hu, Minfeng Zhu, Fengda Zhu, Hu Zhang, Zhedong Zheng, Liang Zheng, Xiaolin Zhang, Aming Wu, and many others. I was very fortunate to collaborate with or discuss with them. These discussions inspired and motivated many of my research works.

Lastly, I would like to thank my family and my girlfriend Yue Yang for their selfless support and love.

Fan Ma

Sydney, Australia, 2022

List of Publications

Conference Papers:

1. **Fan Ma**, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, Zheng Shou. "SF-Net: Single-Frame Supervision for Temporal Action Localization", European Conference on Computer Vision (ECCV 2020)

2. **Fan Ma**, Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, Zhicheng Yan. "Unified Transformer Tracker for Object Tracking", IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)

Journal Papers :

3. **Fan Ma**, Deyu Meng, Xuanyi Dong, Yi Yang, "Self-Paced Multi-view Co-training", Journal of Machine Learning Research (JMLR)

4. **Fan Ma**, Yu Wu, Xin Yu, Yi Yang, "Learning With Noisy Labels via Self-Reweighting From Class Centroids", IEEE Transactions on Neural Networks and Learning Systems (TNNLS)

5. **Fan Ma**, Linchao Zhu, Yi Yang, "Weakly Supervised Moment Localization with Decoupled Consistent Concept Prediction", International Journal of Computer Vision (IJCV)

Table of Contents

List of Publications v

List of Figures xi

List of Tables xv

1 Introduction 1

1.1 Motivation 1

1.2 Research Contribution 5

1.3 Thesis Organization 7

2 Literature Review 9

2.1 Semi-supervised Learning 9

2.1.1 Co-training 9

2.1.2 Self-paced Learning 11

2.2 Learning from Noisy Labels 11

2.3 Weakly Supervised Learning 13

2.3.1 Temporal Action Localization 13

2.3.2 Localizing Moments with Natural Language 14

3 Semi-supervised Learning with Self-paced Co-training 15

3.1 Introduction 15

3.2 Self-paced Co-training 17

3.2.1 Learning Objective 17

3.2.2 Hard Co-regularization Term 19

3.2.3 Soft Co-regularization Term 21

3.3 Optimization 22

3.3.1 Alternative Optimization Strategy 23

3.3.2 Parallel Training 25

3.4 Rationality Exploration 27
3.5 Experiments 31
3.5.1 Toy Data 33
3.5.2 Classification on Multi-view Features 35
3.5.3 Image Classification 38
3.5.4 Person Re-identification 40
3.5.5 Object Detection 45
3.6 Summary 48

4 Noisy Learning with Self-reweighting from Class Centroids 49
4.1 Introduction 49
4.2 Preliminaries 52
4.2.1 Empirical Risk Minimization 52
4.2.2 MixUp 53
4.3 Self-reweighting from Class Centroids 54
4.3.1 Framework 54
4.3.2 Sample Weight Generation 56
4.3.3 Class Centroid Update 58
4.3.4 Training 59
4.3.5 Analysis of SRCC 60
4.4 Experiments 63
4.4.1 Datasets 63
4.4.2 Noise Setting 63
4.4.3 Implementation Details 64
4.4.4 Ablation Study 65
4.4.5 Comparisons with State-of-the-art 70
4.5 Summary 73

5 Single Frame Supervision for Temporal Action Localization 75
5.1 Introduction 75
5.2 Single Frame Network 78
5.2.1 Problem Setup 78
5.2.2 Framework 79
5.2.3 Pseudo Label Mining 79
5.2.4 Training 83
5.2.5 Inference 84

5.3 Experiment . . . 85

5.3.1 Datasets . . . 85

5.3.2 Implementation Details . . . 85

5.3.3 Single Frame Annotation . . . 86

5.3.4 Evaluation Metric . . . 90

5.3.5 Ablation Study . . . 90

5.3.6 Comparisons with State-of-the-art . . . 95

5.4 Summary . . . 97

6 Weakly-supervised Moment Localization with Decoupled Consistent Concept Prediction 99

6.1 Introduction . . . 99

6.2 Decoupled Consistent Concept Prediction . . . 103

6.2.1 Problem Setup . . . 103

6.2.2 Framework . . . 105

6.2.3 Training . . . 109

6.2.4 Inference . . . 110

6.3 Experiments . . . 112

6.3.1 Datasets . . . 112

6.3.2 Implementation Details . . . 112

6.3.3 Evaluation Metric . . . 113

6.3.4 Ablation Study . . . 113

6.3.5 Comparisons with State-of-the-art . . . 120

6.4 Summary . . . 123

7 Conclusion and Future Work 125

Bibliography 127

List of Figures

Figure **Page**

3.1 Toy problems for co-training. The first column is toy data generated by different Gaussian distributions, (a) and (e) are two-Gaussian data in which each distribution corresponds to one class and (i) is four-Gaussian data in which two distributions correspond to one class. The canonical co-training results on toy data are shown in the second column. Last two columns are results of SPaCo with different γ. The blue and yellow dots denote the samples from two classes, and black triangles and stars are labeled points. 32

3.2 Visualized illustrations over the selected unlabeled examples during iterations of our method. Yellow and blue dots denote the predictions on unlabeled examples, respectively. Yellow stars are the selected pseudo-labeled examples of the first class, and blue triangles denote the pseudo-labeled examples of the second class. The first row presents the view using features along the vertical axis, and the second row represents the view using features along the horizontal axis. The third row is the fused predictions from both the first and the second views. Black triangles and stars denote the labeled points. 33

3.3 Convergence tendency of accuracy for SPaCo with hard and soft regularization terms under different λ updating strategy, and λ is adjusted by the number of samples added in each iteration. The left figure is the trend of the mean accuracy on the test set over iteration rounds for SPaCo with hard regularization term, and the right figure is the result for SPaCo with soft regularization term. 36

3.4 Typical selected pseudo-labeled samples during training, where the bounding boxes with different colors indicate the generated pseudo-boxes by our method for different classes. 46

4.1 Training samples of different noise types. Noise samples are marked with red boxes. Falsely annotated labels with symmetric noise could belong to any other classes in the training set. In asymmetric noise, noise samples are only from a certain class. 50

4.2 Framework of our self-reweighting from class centroids (SRCC). We use solid and dash lines to denote the forward and update operations. At each training step, we first extract features and calculate the class centroids for input images (the upper part of the figure). Then we randomly mix two images by linearly interpolating two original images. The weight of the mixed data is evaluated by the similarity between its feature and all class centroids. The reweighted losses are used to update the network. The class centroids and the network are iteratively updated to learn feature representations and classify images. 54

4.3 A toy experiment on synthetic data illustrates the effectiveness of our SRCC on regularizing the decision boundaries. The clean data (the first row) are generated from five Gaussian distributions with different means and standard deviations. The noisy data (the second row) are generated by randomly changing labels of examples in the clean data. The decision boundaries are displayed via Mlxtend (Raschka, 2018). 57

4.4 Transition matrices of different noise types at 40% noise rate (five classes contained). 64

4.5 Test accuracy vs. the number of epochs for SRCC and the compared methods. 67

4.6 Test accuracy vs. the number of epochs for SRCC and the compared methods trained with the fixed learning rate. 67

5.1 Different ways of annotating actions while watching a video. (a) Annotating actions in the fully-supervised way. The start and end time of each action instance are required to be annotated. (b) Annotating actions in the weakly-supervised setting. Only action classes are required to be given. (c) Annotating actions in our single frame supervision. Each action instance should have one timestamp. Note that the time is automatically generated by the annotation tool. Compared to the weakly-supervised annotation, the single frame annotation requires only a few extra pauses to annotate repeated seen actions in one video. 76

5.2 Overall training framework of our proposed SF-Net. Given single frame supervision, we employ two novel frame mining strategies to label pseudo action frames and background frames. The detailed architecture of SF-Net is shown on the right. SF-Net consists of a classification module to classify each labeled frame and the whole video, and an actionness module to predict the probability of each frame being action. 78

5.3 The inference framework of SF-Net. The classification module outputs the classification score C of each frame for identifying possible target actions in the given video. The action module produces the actionness score determining the possibility of a frame containing target actions. The actionness score together with the classification score are used to generate action segment based on the threshold. 84

5.4 Interface for annotating a single frame. First step is to pause the video when annotators notice an action while watching the video. The second step is to select the target class for the paused frame. After annotating an action instance, the annotator can click the video to keep watching the video for the next action instance. Note that the time is automatically generated by the annotation tool. After watching a whole video, the annotator can press the generate button to save all records into a csv file. 86

5.5 Statistics of human annotated single frame on three datasets. X-axis: single frame falls in the relative portion of the whole action; Y-axis: percentage of annotated frames. We use different colors to denote annotation distribution on different datasets. 89

5.6 Qualitative Results on BEOID dataset. GT denotes the ground-truth and the action segment is marked with blue. Our proposed method detects all the action instances in the videos. 95

6.1 Illustration of video moment retrieval task with natural language in the fully-supervised and weakly-supervised settings. The start and end timestamps of the video moment for each text query is given in the fully-supervised setting. These temporal annotations are not available in the weakly-supervised setting.100

6.2 We demonstrate that localizing moments in a video with natural language can be well addressed through the localization of concepts in the video. The nouns and verbs in the query sentence are extracted to be localized in the video. The predictions from all concepts are merged to retrieve moments for the query sentence. 102

6.3 The framework of our decoupled consistent concept prediction (DCCP). The concepts in the query sentence are first extracted and then passed to the embedding layer. We introduce the concept prediction module to build semantic connections between the language and the video features. Details of the visual concept mining module are described in Section 6.2.2. The concept prediction module outputs the localization map and the visual embedding for each object and action concept, respectively. The localization map for the concept is the probability of each video clip containing that concept. 103

6.4 Illustration of our concept prediction module. The inputs include the object and action word embeddings, and the RGB and Flow video clip features. There are two types of concept pairing modules. Each module contains three pairing blocks ("PB"). "Gating" is the feature gating block to learn the visual concept embedding. $\odot$ and $\otimes$ are element-wise and matrix multiplication operations, respectively. σ denotes the sigmoid activation function. Two classification losses and one consistency loss are utilized for training the proposed model. . 106

6.5 The number of top-50 nouns and verbs. We treat nouns in the sentence as objects and verbs as actions. We use green and red to distinguish the object and action. 114

6.6 Natural language moment retrieval results on DiDeMo. Ground-truth moments are outlined with the yellow dashed line and retrieved moments are marked with the orange rectangle. 118

6.7 Natural language moment retrieval results on Charades-STA by the proposed method. Ground-truth moments are marked in yellow and retrieved moments are marked in orange. 119

LIST OF TABLES

TABLE **Page**

3.1 Reuters multilingual dataset summarization. #dim is the dimension of corresponding language, #docs, #c, #l, #u, and #t are the numbers of documents, categories, labeled samples, unlabeled samples and test samples, respectively. 35

3.2 Results for Reuters with different semi-supervised learning algorithms. Mean accuracy with deviation for all competing methods are presented. 35

3.3 Performance comparison of all competing methods on CIFAR-10 with different labeled examples (2000 and 4000). The mean error rates (%) and standard deviation are presented. The best performance is marked in bold. 38

3.4 Results of SPaCo on the test data of CIFAR-10 during model iterations with 2000 examples labeled. We use ER to denote the error rate (%) and DR to denote the difference rate between predictions from two views. Shake and WRN represent the network names of two views. We report the cross entropy (CE) loss between the predictions from two views in the last column. 40

3.5 Mean average precision (MAP) comparison of all competing methods on Market-1501 dataset with two views. The first line is the supervised learning result using only labeled data. Self iterative training and co-training results are presented in the second and third lines, respectively. The “Rep" denotes that the co-training algorithm is trained with the replacement strategy. The last two lines show the results of our proposed SPaCo model with hard and soft regularization terms. 41

3.6 Rank 1 accuracy of all competing methods on Market-1501 dataset with two views. The first line is the supervised learning result using only labeled data. Self iterative training and co-training results are presented in the second and third lines, respectively. The “Rep" denotes that the co-training algorithm is trained with the replacement strategy. The last two lines show the results of our proposed SPaCo model with hard and soft regularization terms. 42

3.7 MAP and rank-1 accuracy of all competing methods on Market-1501 dataset with triple-views. The first line is the supervised learning result using only labeled data. SelfTrain result is presented in the second line. Phard and Psoft indicate that parallel training strategy is employed compared to serial training strategy. Fhard and Fsoft denote that the model does not update labels of unlabeled examples during iterations. Last six lines show the results of SPaCo method with hard and soft regularization term under different training strategies. 43

3.8 Performance comparison in average precision (AP) of all competing methods on the PASCAL VOC 2007 test set. The five compared methods make use of full image-level labels for training. Our method (the last four rows) requires only approximately four strong annotated images per class. Results on each class are shown in one column. We use Fast RCNN with VGG16 and RFCN with ResNet50 and ResNet101 as our base detectors to get 3-view features. . 45

4.1 Classification accuracy on synthetic noisy data. Instances are sampled from the same distribution for five seeds. 65

4.2 Classification errors on CIFAR10 and CIFAR100 in different noise rates. Mean and standard deviation are reported. 66

4.3 Classification errors on CIFAR10 for different hyper-parameter values α. . . 68

4.4 Effects of reweighting strategies on CIFAR10 and CIFAR100. 69

4.5 Overall test accuracy of models with different centroids update strategies on CIFAR10. 70

4.6 Comparisons with different state-of-the-art methods on CIFAR10 and CIFAR100. Mean and standard deviation of Top-1 Accuracy are reported. The relative degradation between the noise and clean cases is also reported in the parentheses. The best and second best results are marked in red and blue respectively. 71

4.7 Comparisons with different state-of-the-art methods in terms of test accuracy on Tiny-ImageNet. 71

4.8 Comparisons with state-of-the-art methods in terms of test accuracy on Clothing1M. 72

5.1 Single frame annotation differences between different annotators on three datasets. We show the number of action segments annotated by Annotator 1, Annotator 2, Annotator 3, and Annotator 4. In the last column, we report the total number of the ground-truth action segments for each dataset. 88

5.2 Comparisons between different methods for simulating single frame supervision on THUMOS14. “Annotation” means that the model uses human annotated frame for training. “TS” denotes that the single frame is sampled from action instances using a uniform distribution, while “TS in GT” is using a Gaussian distribution near the mid timestamp of each activity. The AVG for segment localization is the average mAP from IoU 0.1 to 0.7. 90

5.3 Segment localization mAP results at different IoU thresholds on three datasets. Weak denotes that only video-level labels are used for training. All action frames are used in the full supervision approach. SF uses extra single frame supervision with frame level classification loss. SFB means that pseudo background frames are added into the training, while the SFBA adopts the actionness module, and the SFBAE indicates the action frame mining strategy added in the model. For models trained on single frame annotations, we report mean and standard deviation results of five runs. AVG is the average mAP from IoU 0.1 to 0.7. 92

5.4 The background η analysis on THUMOS14. AVG is the average mAP at IoU 0.1 to 0.7. 93

5.5 The loss coefficients analysis on THUMOS14. AVG is the average mAP at IoU 0.1 to 0.7. 93

5.6 Classification accuracy and class-agnostic localization AP on THUMOS14. . . 94

5.7 Segment localization results on THUMOS14 dataset. The mAP values at different IoU thresholds are reported, and the column AVG indicates the average mAP at IoU thresholds from 0.1 to 0.5. * denotes the single frame labels are uniform sampled from the ground-truth annotations (see 5.3.5.1 for more details). # denotes single frame labels are manually annotated by human annotators. 96

5.8 Segment localization results on ActivityNet1.2 validation set. The AVG indicates the average mAP from IoU 0.5 to 0.95. 97

6.1 Comparisons on DiDeMo validation set in terms of vocabulary volume. Highest score is marked in bold. 114

6.2 Comparisons on DiDeMo validation set in terms of loss functions. 115

6.3 Comparisons on DiDeMo validation set in terms of video features. 115
6.4 Comparisons on DiDeMo in terms of pairing score functions. 116
6.5 Results with feature gating strategies on DiDeMo. 117
6.6 Comparisons on DiDeMo with different methods. 120
6.7 Comparisons on ActivityNet Captions with different methods. 121
6.8 Comparisons on Charades-STA with different methods. 122