

IMPROVING PREDICTION OF MACHINE LEARNING MODELS USING AUXILIARY INFORMATION FOR COMPUTER VISION PROBLEMS

DINESH KHANDELWAL



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2020**

IMPROVING PREDICTION OF MACHINE LEARNING MODELS USING AUXILIARY INFORMATION FOR COMPUTER VISION PROBLEMS

by

DINESH KHANDELWAL

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2020

DEDICATED TO
My Family.

Certificate

This is to certify that the thesis titled **Improving Prediction of Machine Learning Models Using Auxiliary Information for Computer Vision Problems** being submitted by **Dinesh Khandelwal** for the award of **Doctor of Philosophy** in Department of Computer Science and Engineering is a record of bona fide work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly. In particular, work done in Chapters 2 and 4 were done jointly with undergraduate students. In each case, the part done by the collaborators appeared in their respective Bachelor's theses.



12/08/20

Parag Singla

Associate Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016



10/8/2020

Chetan Arora

Associate Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

Acknowledgments

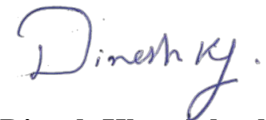
*“If you can meet with Triumph and Disaster
And treat those two impostors just the same
Yours is the Earth and everything that’s in it,
And - which is more - you’ll be a Man, my son!”*

— If by **Rudyard Kipling**

First of all, I would like to express my sincere gratitude to my advisors **Prof. Parag Singla** and **Prof. Chetan Arora** for their continuous support to my research, Their patience with me and motivation, enthusiasm for research helped in my PhD journey. I was fortunate to meet wonderful friends Ankit Anand, Shashank Sharma, Suvam Patra, Rajesh Kedia, Kunal Dahiya, Britty Baby, Vishal Sharma, Happy Mittal, Neetu Jindal, Himanshu Jain and Yashoteja Prabhu for helping me by their positive and valuable suggestions during my low phases in both research and personal life. I immensely enjoyed working with Suyash Agrawal, Anuj Mahajan and Kush Bhatia on research problems during my PhD.

I could not have finished my study without the enduring support of my family. I deeply appreciate the love and support of my wife, Shweta, who supported me in every possible way. Special thanks to my mother, father, sister, my mother-in-law, and father-in-law for their support and love. I would like to express my heartfelt gratitude to my son, Vivaan, born during the last year of my PhD. Thank you for being so patient, spending most of the time with your mother and allowing me to concentrate on my PhD.

Last but not least, I would like to express my immense gratitude towards the almighty for his blessings and grace, which helped me throughout my journey.



Dinesh Khandelwal

Abstract

In computer vision, one is interested in extracting useful information from a given image or video. However, depending upon the sensor involved, a typical camera system may only capture a subset of the available information about a scene. For example, an RGB camera does not capture the depth information. Finite wavelength response, spatial and temporal resolution, as well as occlusion and noise during the capturing process, further reduce the available information. This results in inherent ambiguity in the inference for many perception tasks in computer vision. A common way to get around the ambiguities in an inference problem is to use auxiliary information in the form of priors and/or evidence. The focus of this thesis is to propose novel techniques and algorithms for exploiting such auxiliary information to improve prediction performance in common computer vision problems.

In an MRF-MAP framework, one formulates a computer vision problem as a Markov Random Field and then computes the labeling configuration with maximum a posteriori probability. The framework captures the prior conditional dependencies between the pixel/node labels through the use of clique potential, with larger cliques allowing for wider and more complex priors. However, the inference in such higher order MRF-MAP problems is NP-hard. This has led researchers to focus on various subsets of the problem, for which efficient algorithms for optimal inference may be proposed. One such subset, which is also the focus of this thesis, is the MRF-MAP problems with submodular clique potentials. In the first part of the thesis, we give an efficient algorithm for provably optimal inference even when the clique size grows to 25. Here we make use of an important observation that only an extremely small subset of constraints defined by a clique potential are typically useful for defining the optimal MAP configuration. We exploit the observation and perform inference with a small number of such constraints in a particular iteration. As the iterations progress, we bring the constraints as per the need, ultimately converging to the optimal configuration using a fraction of memory and orders of magnitude improvement in inference time. The algorithm is called the Lazy Generic Cuts, due to its use of Generic Cuts as a baseline, and characteristic behavior of lazily bringing in the required constraints.

The effective use of complex higher order priors requires not only the efficient inference algorithm, but the capability to learn such clique potentials from the training data. In the second part of the thesis, we propose a novel algorithm for learning higher order submodular clique potentials. We use the standard max-margin framework within a structured SVM formulation. However, imposing additional submodularity constraints on parameters creates scalability issues, since the number of submodularity constraints increases exponentially with the clique size. It becomes difficult for existing algorithms to scale to large problems arising out of learning higher order potential in an MRF-MAP formulation. As an important contri-

bution, we propose a relaxation of submodularity constraints and give an efficient subgradient based method for the learning. The proposed algorithm, also exploits the Lazy Generic Cuts for optimal inference in the inner loop. Whereas the earlier state of the art could learn clique potentials of size only upto 9, using our algorithm one can learn the clique potentials even upto size 16.

In the last decade, fuelled by the improvement in the computation capabilities and the availability of large annotated datasets, deep neural networks have become a de-facto technique for many computer vision problems. Though our earlier work on MRF-MAP problems can be used as post-processing with many such deep neural network techniques, the lack of end-to-end training and inference of deep neural network with the MRF-MAP module, hampers the overall prediction performance. In the third and last part of the thesis, we explore the direct use of auxiliary information in the form of evidence, to improve predictions of deep neural networks. Unlike priors, which are dataset level statistics, evidence is much more targeted, and represent additional sample level information at the test time. We see the abundance of such auxiliary information around us. For example, while the primary inference task of inference may be semantic or instance level segmentation of an image, we often have image level tags or associated image captions. Although one can design an appropriate deep neural network which takes multi-modal input, the approach requires training such models from scratch, requiring a large amount of annotated data with labels for the primary task as well as auxiliary information. On the other hand, we propose a novel multi-task learning (MTL) framework for exploiting the auxiliary information. We propose to model the task of interest as the primary task in an MTL framework. The auxiliary information in the form of evidence is incorporated by modeling it as the output of a secondary task. Given a particular sample, we first perform inference on the secondary task and back-propagate the loss on the secondary task based upon the given auxiliary information. This allows the model to adapt the weights as per the given sample, which are then used to perform inference on the primary task. The advantages of the proposed framework include the possibility of using standard pre-trained models for both primary and secondary tasks, as well as the ability to train even when no jointly annotated data is available. The second advantage flows directly from the use of the MTL framework.

In summary, this thesis proposes novel techniques to exploit the auxiliary information, either in the form of prior or evidence, to improve the predictive accuracy of underlying machine learning models. In each case, we demonstrate the efficacy of our proposed techniques by experimenting on real-world data and improving the performance of the state-of-the-art.

सार

कंप्यूटर दृष्टि में, किसी को दी गई छवि या वीडियो से उपयोगी जानकारी निकालने में रुचि है। हालांकि, इसमें शामिल सेंसर के आधार पर, एक विशिष्ट कैमरा सिस्टम केवल एक दृश्य के बारे में उपलब्ध जानकारी के उपसमुच्चय को पकड़ सकता है। उदाहरण के लिए, एक आरजीबी कैमरा गहराई की जानकारी को कैप्चर नहीं करता है। परिमित प्रक्रिया के दौरान परिमित तरंगदैर्घ्य प्रतिक्रिया, स्थानिक और लौकिक संकल्प, साथ ही रोड़ा और शोर, उपलब्ध जानकारी को और कम कर देते हैं। यह कंप्यूटर दृष्टि में कई धारणा कार्यों के लिए निहित अस्पष्टता का परिणाम है। एक अनुमान समस्या में अस्पष्टताओं के आसपास जाने का एक सामान्य तरीका है प्रायर और / या इविडेंस के रूप में सहायक जानकारी का उपयोग करना। इस थीसिस का फोकस आम कंप्यूटर दृष्टि समस्याओं में भविष्यवाणी के प्रदर्शन में सुधार के लिए इस तरह की सहायक जानकारी के दोहन के लिए उपन्यास तकनीकों और एल्गोरिदम का प्रस्ताव करना है।

एमआरएफ-एमएपी फ्रेमवर्क में, मार्कोव रैंडम फील्ड के रूप में कंप्यूटर विज्ञान की समस्या का निर्माण करता है और फिर अधिकतम पोस्टीरियर प्रायिकता के साथ लेबलिंग कॉन्फिगरेशन की गणना करता है। फ्रेमवर्क संभावित के उपयोग के माध्यम से पिक्सेल / नोड लेबल के बीच पूर्व सशर्त निर्भरता को कैप्चर करता है, जिसमें व्यापक और अधिक जटिल प्रायर के लिए बड़े क्लस्टर की अनुमति होती है। हालांकि, इस तरह के उच्च क्रम में एमआरएफ-एमएपी की समस्याओं का अनुमान एनपी-हार्ड है। इसने शोधकर्ताओं को समस्या के विभिन्न उपसमुच्चय पर ध्यान केंद्रित करने के लिए प्रेरित किया है, जिसके लिए इष्टतम अनुमान के लिए कुशल एल्गोरिदम प्रस्तावित किया जा सकता है। एक ऐसा उपसमुच्चय, जो इस थीसिस का फोकस भी है, सबमॉड्यूलर क्लिक पोटेंशियल के साथ एमआरएफ-एमएपी समस्याएं हैं। थीसिस के पहले भाग में, हम तब भी जब तक कि आकार 25 तक बढ़ जाता है, तब भी इष्टतम इष्टतम निष्कासन के लिए एक कुशल एल्गोरिथ्म दिया जाता है। यहाँ हम एक महत्वपूर्ण अवलोकन का उपयोग करते हैं कि केवल एक छोटी क्षमता के द्वारा परिभाषित बाधाओं का एक बहुत छोटा उपसमुच्चय आमतौर पर उपयोगी होता है। इष्टतम एमएपी विन्यास को परिभाषित करने के लिए। हम अवलोकन का शोषण करते हैं और एक विशेष पुनरावृत्ति में इस तरह की बाधाओं की एक छोटी संख्या के साथ निष्कासन करते हैं। पुनरावृत्तियों की प्रगति के रूप में, हम आवश्यकता के अनुसार बाधाओं को लाते हैं, अंततः स्मृति के एक अंश का उपयोग करके इष्टतम कॉन्फिगरेशन में परिवर्तित होते हैं और परिक्षेपण समय में परिमाण में सुधार के आदेश देते हैं। बेसलाइन के रूप में जेनेरिक कट्स के उपयोग के कारण एल्गोरिथ्म को लेज़ी जेनेरिक कट्स कहा जाता है, और आवश्यक बाधाओं में लेज़ीली के विशिष्ट व्यवहार को दर्शाता है।

जटिल उच्च क्रम के प्रायर के प्रभावी उपयोग के लिए न केवल कुशल निष्कर्ष एल्गोरिथ्म की आवश्यकता होती है, बल्कि प्रशिक्षण डेटा से इस तरह के क्लिक क्षमता को सीखने की क्षमता भी होती है। थीसिस के दूसरे भाग में, हम उच्च क्रम सबमॉड्यूलर क्लिक क्षमता सीखने के लिए एक उपन्यास एल्गोरिथ्म का प्रस्ताव करते हैं। हम एक संरचित एसवीएम निर्माण के भीतर मानक अधिकतम-मार्जिन ढांचे का उपयोग करते हैं। हालांकि, मापदंडों पर अतिरिक्त सबमॉड्यूलरिटी बाधाओं को लागू करने से स्केलेबिलिटी समस्याएं पैदा होती हैं, क्योंकि सबमॉड्यूलरिटी बाधाओं की संख्या क्लिक आकार के साथ तेजी से बढ़ जाती है। मौजूदा एल्गोरिदम के लिए एमआरएफ-एमएपी फॉर्मूलेशन में उच्च आदेश क्षमता सीखने से उत्पन्न होने वाली बड़ी समस्याओं के लिए मुश्किल हो जाता है। एक महत्वपूर्ण योगदान के रूप में, हम सबमॉड्यूलरिटी बाधाओं की छूट का प्रस्ताव करते हैं और सीखने के लिए एक कुशल सबग्रेडिएंट आधारित विधि देते हैं। प्रस्तावित एल्गोरिथ्म, आंतरिक लूप में इष्टतम हस्तक्षेप के लिए आलसी जेनेरिक कट्स का भी उपयोग करता है। जबकि कला की पूर्व स्थिति केवल 9 तक की आकार की क्षमता को सीख सकती है, हमारे एल्गोरिथ्म का उपयोग करके कोई भी आकार 16 तक की क्लिक संभावनाओं को जान सकता है।

पिछले दशक में, संगणना क्षमताओं में सुधार और बड़े एनोटेड किए गए डेटासेट की उपलब्धता से, गहरे तंत्रिका नेटवर्क कई कंप्यूटर दृष्टि समस्याओं के लिए एक वास्तविक तकनीक बन गए हैं। यद्यपि एमआरएफ-एमएपी समस्याओं पर हमारे पहले के काम का उपयोग कई ऐसी गहरी तंत्रिका नेटवर्क तकनीकों के साथ प्रसंस्करण के बाद किया जा सकता है, अंत-टू-एंड प्रशिक्षण की कमी और एमआरएफ-एमएपी मॉड्यूल के साथ गहरे तंत्रिका नेटवर्क की असमानता, समग्र भविष्यवाणी को बाधित करती है। प्रदर्शन। थीसिस के तीसरे और आखिरी भाग में, हम गहरे तंत्रिका नेटवर्क की भविष्यवाणियों को बेहतर बनाने के लिए सबूत के रूप में सहायक जानकारी के प्रत्यक्ष उपयोग का पता लगाते हैं। प्रायर के विपरीत, जो डेटासेट स्तर के आँकड़े हैं, साक्ष्य बहुत अधिक लक्षित होते हैं, और परीक्षण के समय में अतिरिक्त नमूना स्तर की जानकारी का प्रतिनिधित्व करते हैं। हम अपने आस-पास ऐसी सहायक सूचनाओं की प्रचुरता देखते हैं। उदाहरण के लिए, जबकि अनुमान का प्राथमिक निष्कर्ष कार्य छवि का शब्दार्थ या उदाहरण स्तर विभाजन हो सकता है, हमारे पास अक्सर छवि स्तर टैग या संबद्ध छवि कैप्शन होते हैं। यद्यपि एक उपयुक्त गहरे तंत्रिका नेटवर्क को डिज़ाइन कर सकता है जो मल्टी-मोडल इनपुट लेता है, दृष्टिकोण को खरोंच से ऐसे मॉडल की आवश्यकता होती है, जिसके लिए प्राथमिक कार्य के साथ-साथ सहायक जानकारी के साथ लेबल के साथ एनोटेड डेटा की एक बड़ी मात्रा की आवश्यकता होती है। दूसरी ओर, हम सहायक जानकारी का दोहन करने के लिए एक उपन्यास मल्टी-टास्क लर्निंग (एमटीएल) ढांचे का प्रस्ताव करते हैं। हम एक एमटीएल ढांचे में प्राथमिक कार्य के रूप में ब्याज के कार्य को मॉडल करने का प्रस्ताव करते हैं। सबूत

के रूप में सहायक जानकारी को एक माध्यमिक कार्य के आउटपुट के रूप में मॉडलिंग द्वारा शामिल किया गया है। एक विशेष नमूने को देखते हुए, हम पहले माध्यमिक कार्य पर निष्कर्ष निकालते हैं और दी गई सहायक जानकारी के आधार पर द्वितीयक कार्य पर होने वाले नुकसान का पुनः प्रचार करते हैं। यह मॉडल को दिए गए नमूने के अनुसार वजन को अनुकूलित करने की अनुमति देता है, जो तब प्राथमिक कार्य पर अनुमान लगाने के लिए उपयोग किया जाता है। प्रस्तावित ढांचे के फायदों में प्राथमिक और माध्यमिक दोनों कार्यों के लिए मानक पूर्व-प्रशिक्षित मॉडल का उपयोग करने की संभावना शामिल है, साथ ही साथ संयुक्त रूप से एनोटेट डेटा उपलब्ध होने पर भी प्रशिक्षित करने की क्षमता शामिल है। दूसरा लाभ एमटीएल ढांचे के उपयोग से सीधे प्रवाहित होता है।

सारांश में, इस थीसिस ने अंतर्निहित मशीन लर्निंग मॉडल की पूर्वसूचक सटीकता में सुधार करने के लिए, या तो पूर्व या सबूत के रूप में, सहायक जानकारी का दोहन करने के लिए उपन्यास तकनीकों का प्रस्ताव रखा। प्रत्येक मामले में, हम वास्तविक दुनिया के आंकड़ों पर प्रयोग करके और अत्याधुनिक के प्रदर्शन में सुधार करके अपनी प्रस्तावित तकनीकों की प्रभावकारिता प्रदर्शित करते हैं।

Contents

Certificate	ii
Acknowledgments	iii
Abstract	iv
1 Introduction	1
1.1 Ambiguities Due to Loss of Scene Information	3
1.2 Dealing With Ambiguities in Computer Vision Tasks	7
1.2.1 Prior Knowledge	7
1.2.2 Evidence	9
1.2.3 Differentiation from Other Similar Learning Paradigms	10
1.3 Thesis Contributions	11
1.3.1 Inference with High Order Potentials in Markov Random Fields	11
1.3.2 Learning High Order Potentials for Markov Random Fields	13
1.3.3 Exploiting Test Time Evidence to Improve Predictions of Deep Neural Networks	13
1.3.4 Summary of Thesis	14
1.3.5 Outline of Thesis	14
2 Inference with Higher Order Potentials in Markov Random Fields	17
2.1 Related Work	19
2.2 Background	20
2.2.1 Markov Random Fields (MRFs)	20
2.2.2 Submodular Function	23
2.2.3 Generic Cuts (GC)	25
2.3 Lazy Generic Cuts (LGC)	28
2.3.1 Preliminaries	28
2.3.2 LGC Algorithm	30

2.3.3	Convergence and Correctness	34
2.4	Experiments and Results	35
2.4.1	Experiment Setup	36
2.4.2	Scalability with Clique Size	39
2.4.3	Scaling with Image Size	41
2.4.4	Effect of Algorithm Parameters	41
2.4.5	Comparison on Non-submodular Clique Potential	45
2.4.6	Comparison using Learnt Potentials	46
2.4.7	Visual Quality	46
2.5	Conclusion	47
3	Learning Higher Order Potentials For Markov Random Fields	49
3.1	Related Work	51
3.1.1	Max-margin Framework	51
3.1.2	Approximate Submodularity	52
3.2	Background	53
3.2.1	Submodularity	53
3.2.2	Structured SVMs	54
3.2.3	Subgradient based Optimization	55
3.3	Our Approach	56
3.3.1	Learning with Submodularity Constraints	56
3.3.2	Switching to Soft Submodularity Constraints	58
3.3.3	Convergence Guarantees	60
3.4	Experiments	61
3.4.1	Image Denoising	61
3.4.2	Object Detection	65
3.5	Conclusion	67
4	Exploiting Test Time Evidence to Improve Predictions of Deep Neural Networks	69
4.1	Related Work	72
4.2	Framework for Back-propagating Evidence	73
4.2.1	Background on MTL	73
4.2.2	Our Approach - Prediction	75
4.2.3	Interpreting as a Graphical Model	76
4.3	Experiments on Semantic Segmentation	77
4.3.1	State-of-the-art	77
4.3.2	Our Implementation	78

4.3.3	Methodology and Dataset	78
4.3.4	Results	79
4.3.5	Impact of Noise	82
4.4	Experiments on Instance Segmentation	83
4.4.1	State-of-the-art	83
4.4.2	Our Implementations	84
4.4.3	Methodology and Dataset	84
4.4.4	Results	85
4.4.5	Failure analysis	86
4.4.6	Dataset Analysis	87
4.5	Conclusion	91
5	Conclusion and Future Directions	93
	Bibliography	97
	List of Publications	113
	Biography	115

List of Figures

1.1	Image source: [Gonzalez and Woods, 1992]	3
1.2	Without any constraints on the output, both y_1 and y_2 can represent the given input blurred image x equally well. This shows the ill-posed nature of the image deblurring problem. Image source: [Levin, 2009].	4
1.3	There is loss of temporal information if we capture a scene by a single image. Just looking at single input image it is difficult to tell where the person is “sitting down” or “standing up”. Image source: [AlBahar, 2018]	5
1.4	This is a difficult image for object detection. But providing contextual information that the image is an upside down can help. Image source: [Edell, 2018].	6
1.5	Visible and infrared views from an airplane as it tries to land. Infrared camera clearly shows the runway (right image) and helps the pilot, whereas using a normal camera, the runway can be barely seen due to the presence of fog (left image). Image source: [mitsubishi]	6
1.6	Showing prior knowledge that natural images are generally smoother as compared to unnatural Images. Image source: [Levin, 2009]	8
1.7	Improvement in visual quality for denoising problem using cliques of different sizes.	12
2.1	A gadget for 5-clique	26
2.2	Flow chart of LGC	31

2.3	The figure displays an image patch corresponding to a 3×3 sized clique. The nodes are numbered from 1 to 9. Each node in the figure has the label a or b denoting a specific labeling configuration. We consider the 4-neighborhood i.e. up, down, left and right. For example, the neighbors of node 5 are given as 2, 4, 6 and 8. A pair of neighboring pixels defines an edge if one of them has the label a and the other has label b . The edges for the above figure are shown using solid lines. The total number of edges is $ E = 9$. Using the edge based potential, the cost of labeling is given by $\sqrt{ E } = \sqrt{9} = 3$. If $ V_a $ and $ V_b $ denote the number of pixels labeled a and b , respectively, the cost of count based potential for the above example is given by $ V_a * V_b $ i.e. $5 * 4 = 20$	37
2.4	A representative image from each of the 10 sample shape classes from the MPEG-7 Core Experiment CE-Shape-1 dataset [Latecki <i>et al.</i> , 2000]. There are 70 shape classes in the dataset with 20 images per class.	38
2.5	Comparison of visual quality using non submodular clique potentials for denoising problem. Image size is 80×80 and Clique Size is 4×4	40
2.6	Time and Memory variation for LGC with θ_{th} for image of size 80×80 and clique size 4×4	43
2.7	Time taken by LGC for various values of flow augmentations allowed in each LGC iteration. Image size is 60×60 , clique size is 4×3 and $\theta_{th} = 20$	44
2.8	Improvement in visual quality for denoising problem using cliques of different sizes. Image size is 80×80	47
3.1	Change of objective value (left) and average pixel loss on the training data (right), during different iterations.	64
3.2	The current state-of-the-art has shown learning of such clique potentials upto size 4, we seek to enhance the tractability of learning upto size 16. The figure shows the improvement in denoising output for input image shown in (a), using learned potentials of various sizes: (c) clique size = 2, (d) clique size = 4, (e) clique size = 9, and (f) clique size = 16.	64
3.3	Comparison of the denoising output using clique potentials learnt from the proposed approach vs using handcrafted count based potential of size 3×3 clique potentials. We have used structured noise by choosing random patch of size 3×3 and flip pixels of the patch with probability $p = 0.3$. Image size is 120×120	65
3.4	Comparing object detection output using DeepLabv2 Vs our approach using learnt clique potentials of size 3×3	66

4.1	Three examples of Mooney images[Mooney, 1957], which are difficult to interpret even for a human observer. However, given the description: “Beautiful lady smiling in front of a screen”, the face of a woman in the images become obvious. Our proposed framework is inspired from this behavior and improves the prediction of a DNN by exploiting additional cues at the test time. Standard Mask R-CNN [He <i>et al.</i> , 2017], fails to detect a face in the above images, but the same network when plugged in our framework with the caption as auxiliary information easily detects the faces as shown in Figure 4.9.	70
4.2	The figure shows the improvement achieved by existing diverse state-of-the-art models using our framework for exploiting the evidence at prediction time. We show results for instance segmentation (1st row), and semantic segmentation (2nd row), using textual description and image level tags respectively as the auxiliary information. Both the images as well as auxiliary information are deliberately sourced from unconstrained data on the web (and not benchmark datasets) to show the easy availability of such information at the test time. . . .	71
4.3	Model Architecture for Multi-task learning (MTL)	74
4.4	Loss propagation at train and prediction time	74
4.5	Our proposed architecture for semantic segmentation using image level tags as test time auxiliary information. The architecture uses pre-trained DeepLab architecture, fine tuned with MTL strategy using image classification as auxiliary task(lower branch).	79
4.6	Sensitivity of results on number of back-propagation iterations (early stopping) and α parameter (L2-norm)	80
4.7	Comparison of visual results on semantic segmentation problem using image level tags as the test time auxiliary information. 1st and 2nd column are input and ground truth respectively. 3rd column shows baseline model trained with MTL. 4th column shows results of a naive strategy to improve the results of 3rd column by pruning predicted labels which were not present in tags. The last 4 columns show results of proposed approach in various configurations. Please see the chapter text for details on the configurations. The first row shows improvement in segmentation of ‘dining table’, second row shows correction of ‘dog’ label and third row detecting a new object ‘sofa’ by our technique using test time tags.	82
4.8	Our framework for instance segmentation using Mask R-CNN and LSTM based caption generator to use textual descriptions of an image at test time.	84

4.9	Comparison of visual results for baseline (Mask-MT) vs proposed (Mask-BP-ES) approaches. Using our strategy to exploit test time evidence, our approach is able to discover new objects (and their segmentation) which were missed earlier.	85
4.10	Our framework may lead to over-fitting on the auxiliary information if back-propagation is overused. The figure shows incorrect prediction of multiple surf-board instances when the number of back-propagations are increased from 2 to 5.	87
4.11	Input caption: “a person standing over a toilet using the restroom”. Mask-BP-ES is able to detect the person standing in the toilet. This is in contrast to Mask-MT which can only detect the toilet. Ground truth incorrectly misses the person.	87
4.12	Input caption: “a living room with a couch, coffee table, tv on a stand and a chair with ottoman”. Mask-BP-ES is able to detect clock as compared to Mask-MT. This is not annotated in the ground truth.	88
4.13	Input caption: “the large glass plant vase is installed into the wall”. Mask-BP-ES is able to detect potted plant as compared to Mask-MT. This is not annotated in the ground truth.	88
4.14	Input caption: “the little girl is busy eating her pizza at the table”. Mask-BP-ES is able to detect dining table as compared to Mask-MT. The dining table is labeled in the ground truth with incomplete segmentation.	88
4.15	Input caption: “three teddy bears sit in a sled in fake snow”. Mask-BP-ES is able to detect 4 teddy bears as compared to Mask-MT which detects 3 teddy bears. All 3 teddy bears in the front portion of image are marked as a single teddy bears in the ground truth.	89
4.16	Input caption: “a shopping cart full of food that includes bananas and milk”. Mask-BP-ES is able to detect bottle of milk as compared to Mask-MT. This is not annotated in the ground truth.	89
4.17	Input caption: “a yellow and red apple and some bananas”. Mask-BP-ES is able to separate the bananas as compared to Mask-MT. These are marked as a single banana in the ground truth.	89
4.18	Input caption: “two apples, an orange, some grapes and peanuts”. Mask-BP-ES is able to detect the an apple in the image as compared to Mask-MT, but both the apples are marked as single apple in the ground truth.	90
4.19	Input caption: ‘a bird resting outside of a boat window”. Mask-BP-ES is able to detect both bird and boat in the image as compared to Mask-MT as mentioned in the caption also. Ground truth has no annotation for boat.	90

-
- 4.20 Input caption: “a lot of strawberries and oranges sitting in a bowl”. Mask-BP-ES is able to correctly detect an extra orange as compared to Mask-MT. These all oranges are marked as a single orange in the ground truth. 90
- 4.21 Input caption: ‘a man holds a controller and keyboard with his hands’. Mask-BP-ES is able to detect the remote which is missed by Mask-MT. The ground truth annotation misses the remote. 91

List of Tables

2.1	Comparison of time and memory of different inference algorithms on count based potential with increasing clique size. The potential is submodular for all clique sizes. A '-' in the table means that the algorithm ran out of memory. Image size is 80×80 . For LGC $\theta_{th} = 40$	39
2.2	Comparison of time and memory for different inference algorithms on edge based potential with increasing clique size. It may be noted that the potential is submodular for clique size 2×2 only. Image size is 80×80 . For LGC $\theta_{th} = 40$	40
2.3	Comparison of time and memory of different inference algorithms on count based potential with increasing image size. Clique size is 4×4 . For LGC $\theta_{th} = 40$	41
2.4	Inference time in seconds and memory in MB with increasing clique size. Image size is 40×40 . $\theta_{th} = 20$ for LGC.	42
2.5	Inference time required by different components of LGC algorithm on 4×4 clique problem with $\theta_{th} = 40$	42
2.6	Number of LGC iterations and fraction of DFCs in active set at convergence. Clique size is 4×4	43
2.7	Comparison of minimum energy value reached by different inference algorithms on edge based potential with increasing clique size. Image size is 80×80 and $\theta_{th} = 40$ for LGC	45
2.8	Comparison of time and memory for different inference algorithms on edge based potential with increasing image size. Clique size is 4×4 . For LGC $\theta_{th} = 40$	45
2.9	Comparison of minimum energy value reached by different inference algorithms on edge based potential with increasing image size. Clique size is 4×4 . For LGC $\theta_{th} = 40$	45
2.10	Comparison of inference time and memory for different inference algorithms on learned potential in seconds with increasing clique size. Image size is 80×80 and $\theta_{th} = 40$ for LGC.	46

3.1	Average pixel loss on 10 testing images with increasing clique size. Image size is 120×120 . We have used structured noise of patch size 4×4 and with flipping probability $p = 0.7$	62
3.2	Average pixel loss on 10 testing images with varying clique size of learned potential and patch size of structured noise. Image size is 120×120 and flipping probability $p = 0.7$. We observe that as the patch size of noise increases, we need higher and higher order potentials to achieve a certain accuracy.	63
4.1	Comparison of results for semantic segmentation. The two columns shows results of various approaches with and without using multi scale (MS) strategy. 1st row shows result from baseline DeepLab, 2nd row shows the results for DeepLab trained using MTL framework along with image level tags. 3rd row shows result of 2nd row after naive pruning of tags absent in the auxiliary information. The last 4 rows show results of proposed approach in various configurations. Please see the chapter text for details on the configurations. . .	80
4.2	Object category-wise comparison of results for semantic segmentation on Pascal VOC. The first column shows results with naive pruning of labels not present in image tags. The last two columns are variations of proposed methodology. Numbers denote mIoU.	81
4.3	Performance of DL-BP-ES and DL-MT-BP-ES-Pr models as noisy tags are added in the auxiliary information. Multi-scaling was off in this experiment. . .	83
4.4	Comparison of results for instance segmentation. Mask R-CNN and Mask-MT represent baseline pre-trained network and baseline trained with MTL strategy respectively. The bottom two rows are variations of proposed framework. $AP_{0.5}$ refers to AP at IoU of 0.5. AP_L , AP_M , AP_S represent $AP_{0.5}$ values for large, medium and small objects, respectively	86
4.5	Comparison of baseline (Mask-MT) and proposed (Mask-BP-ES) approaches on at 0.5 IoU and 0.9 confidence threshold. P: Precision, R: Recall, F1: F-measure	86