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

10/8/2020

Parag Singla
Associate Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Chetan Arora
Associate Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgments

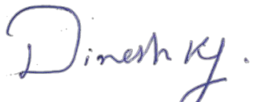
"If you can meet with Triumph and Disaster
And treat those two impostors just the same
Yours is the Earth and everything that's in it,
And - which is more - you'll be a Man, my son!"

— *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 तरंगदैर्ध्य 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 विज़न की समस्या का निर्माण करता है और फिर अधिकतम पोस्टीरियर प्रायिकता के साथ लेबलिंग कॉन्फ़िगरेशन 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 पोटेंशिअल 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 कॉन्फ़िगरेशन 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 एनोटेट 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 एनोटेट 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 वज़न 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

List of Figures

List of Tables