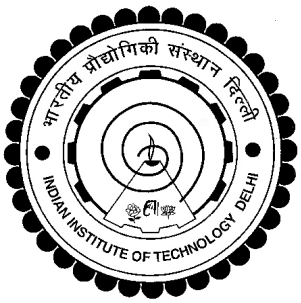
IMAGE ANNOTATION AND TAG PROPAGATION

Anurag Tripathi



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2021

IMAGE ANNOTATION AND TAG PROPAGATION

by

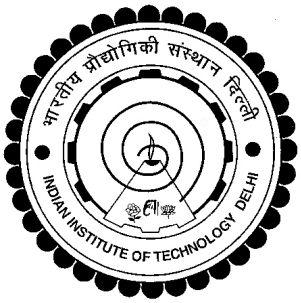
Anurag Tripathi

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

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

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2021

Certificate

This is to certify that the thesis entitled **"IMAGE ANNOTATION AND TAG PROPAGATION"** being submitted by **Mr. ANURAG TRIPATHI** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** is the record of the bonafide research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

Prof. Brejesh Lall
Department of Electrical Engineering
Indian Institute of Technology Delhi

Prof. Santanu Chaudhury
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

Acknowledgements

I express my sincere thanks to my respected supervisor, Prof. Santanu Chaudhary and Prof. Brejesh Lall for their valuable advice and guidance. I would also like to thank my friends Dr. Siddharth Srivastava, Dr. Abhinav Gupta and Dr. Prerana Mukherjee for their support at all the times. I would like to sincerely thank everyone who guided, helped and supported me in this part of my life's journey.

I would like to mention that this dissertation would not have been possible without support of my parents, my wife, my sister and kids. I would like to express my heartiest gratitude towards my wife, who stood by me patiently through all my efforts and encouraged me to pursue Ph.D.

I dedicate this thesis to my spiritual guru, Sri Dr. Rajendra Das Devacharya, for imparting strength and patience making it possible in balancing through the ups and downs of during my life and Ph.D.

Anurag Tripathi

Abstract

In the era of connected devices and internet large amount of digital data is generated which keeps on growing day by day, with images and videos contributing a majority to it. Such humongous amount of data presents two challenges. First is it's management and the second is extracting meaningful information from this data. However both of these challenges are dependent on each other. For example, if one can interpret the meaning of the data, the organization of data becomes easier and vice-versa. Moreover, text based interaction is more convenient to communicate with the visual data hence its an obvious direction to develop similar technologies for archiving and retrieving visual data. To this end, this thesis first aims towards tagging of images which aids effective organization of data. Further, we utilize these annotations in understanding images. Specifically, we work on prediction of relationship among objects within a scene. Subsequently, based on our contributions in image annotations and understanding of relationship between objects, we propose to generate novel natural scenes in a controlled manner.

Our first work is in annotation of images, where we have two contributions, first approach is probabilistic component analysis (PLCA) based framework wherein we represent the image objects as latent components. Prior work in literature assumed that tags are mutually exclusive, however our analysis shows that the tags are highly correlated. For example correlation between car and road is high as compared to car and chair. The second approach is a joint framework which performs both image classification and annotation. Here, we used the correlation between an image class and their corresponding annotation.

Once we have the annotations, next step is to investigate the semantic information within

objects of an image. We can establish the relationship between the objects by analyzing the interaction between them. This relationship between the objects or labels aids towards deeper understanding of the scene. Therefore, the third contribution of this thesis is to detect relationship using visual information i.e. classes of objects within an image. For example if subject and object in an image are *person* and *bike*, and interaction between them can be a predicate *riding on* from the visual cues. Such visual relationship can be formulated as triplets in the form of (*subject,predicate,object*). The relationship in a given image can be a verb (e.g. ride), or preposition (e.g. by), spatial phrase (e.g. in the front of), or comparative phrase (e.g. taller than). These relationships can be used for application such as image captioning etc.

Now as an inverse problem our fourth contribution is a framework to generate images containing scenes corresponding to specified objects aligned with the textual annotations/classes. The proposed technique generates scene which has a high correlation with the provided set of input objects while also maintaining the natural placement of objects within the scene. This is in contrast with earlier works where the objective is to generate a natural scene from a noise vector or conditioning the network over a variable. However, such methods have limitations in their ability to control the objects within the generated images. On the contrary, we show that by training a Generative Adversarial Network with raw image pixels as input, we can generate scenes which constitute the objects as well as generate the surrounding environment suitable for the combination of the input objects.

We have evaluated our proposed framework with the relevant dataset and metrics used in similar settings. On image annotation we have achieved state-of-the art results as compared to other method. On image generation task we have achieved state-of-the art results using the inception score as metrics.

सार

कनेक्टेड डिवाइस और इंटरनेट के युग में बड़ी मात्रा में डिजिटल डेटा उत्पन्न होता है जो दिन-प्रतिदिन बढ़ता रहता है, जिसमें छवियों और वीडियो का बड़ा योगदान होता है।

इतनी भारी मात्रा में डेटा दो चुनौतियां पेश करता है। पहला है इसका प्रबंधन और दूसरा है इस डेटा से सार्थक जानकारी निकालना। हालाँकि ये दोनों चुनौतियाँ एक दूसरे पर निर्भर हैं। उदाहरण के लिए, यदि कोई डेटा के अर्थ की व्याख्या कर सकता है, तो डेटा का संगठन आसान हो जाता है और इसके विपरीत। इसके अलावा, दृश्य डेटा के साथ संवाद करने के लिए पाठ आधारित बातचीत अधिक सुविधाजनक है, इसलिए दृश्य डेटा को संग्रहीत करने और पुनर्प्राप्त करने के लिए समान तकनीकों को विकसित करने की यह एक स्पष्ट दिशा है। इसके लिए, इस थीसिस का उद्देश्य सबसे पहले छवियों को टैग करना है जो डेटा के प्रभावी संगठन में सहायता करता है। इसके अलावा, हम छवियों को समझने में इन एनोटेशन का उपयोग करते हैं। विशेष रूप से, हम एक दृश्य के भीतर वस्तुओं के बीच संबंध की भविष्यवाणी पर काम करते हैं।

इसके बाद, छवि एनोटेशन और वस्तुओं के बीच संबंधों की समझ में हमारे योगदान के आधार पर, हम उपन्यास प्राकृतिक दृश्यों को नियंत्रित तरीके से उत्पन्न करने का प्रस्ताव करते हैं।

हमारा पहला काम छवियों के एनोटेशन में है, जहां हमारे दो योगदान हैं, पहला दृष्टिकोण संभाव्य घटक विश्लेषण (पीएलसीए) आधारित ढांचा है जिसमें हम छवि वस्तुओं को गुप्त घटकों के रूप में प्रस्तुत करते हैं। साहित्य में पहले के काम ने माना कि टैग परस्पर अनन्य हैं, हालांकि हमारे विश्लेषण से पता चलता है कि टैग अत्यधिक सहसंबद्ध हैं। उदाहरण के लिए कार और सड़क के बीच सहसंबंध कार और कुर्सी की तुलना में अधिक है।

दूसरा दृष्टिकोण एक संयुक्त ढांचा है जो छवि वर्गीकरण और एनोटेशन दोनों करता है।

यहां, हमने एक छवि वर्ग और उनके संबंधित एनोटेशन के बीच सहसंबंध का उपयोग किया।

एक बार हमारे पास एनोटेशन हो जाने के बाद, अगला कदम एक छवि की वस्तुओं के भीतर अर्थ संबंधी जानकारी की जांच करना है। हम वस्तुओं के बीच की बातचीत का विश्लेषण करके उनके बीच संबंध स्थापित कर सकते हैं। वस्तुओं या लेबल के बीच यह संबंध दृश्य की गहरी समझ की दिशा में सहायता करता है। इसलिए, इस थीसिस का तीसरा योगदान दृश्य जानकारी यानी छवि के भीतर वस्तुओं के वर्गों का उपयोग करके संबंधों का पता लगाना है। उदाहरण के लिए यदि किसी छवि में विषय और वस्तु {व्यक्ति} और {बाइक} हैं, और उनके बीच की बातचीत दृश्य संकेतों से एक विधेय {सवारी} हो सकती है। इस तरह के दृश्य संबंध को {विषय, विधेय, वस्तु}) के रूप में त्रिगुणों के रूप में तैयार किया जा सकता है।

किसी दी गई छवि में संबंध एक क्रिया (जैसे सवारी), या पूर्वसर्ग (जैसे द्वारा), स्थानिक वाक्यांश (जैसे सामने में), या तुलनात्मक वाक्यांश (जैसे इससे लंबा) हो सकता है। इन रिश्तों का उपयोग इमेज कैप्शनिंग आदि जैसे एप्लिकेशन के लिए किया जा सकता है।

अब एक उलटी समस्या के रूप में हमारा चौथा योगदान पाठ्य एनोटेशन/वर्गों के साथ संरेखित निर्दिष्ट वस्तुओं के अनुरूप दृश्यों वाले चित्र उत्पन्न करने के लिए एक ढांचा है। प्रस्तावित तकनीक दृश्य उत्पन्न करती है जिसमें दृश्य के भीतर वस्तुओं के प्राकृतिक स्थान को बनाए रखते हुए इनपुट वस्तुओं के प्रदान किए गए सेट के साथ उच्च सहसंबंध होता है। यह पहले के कार्यों के विपरीत है जहां उद्देश्य एक शोर वेक्टर या एक चर पर नेटवर्क कंडीशनिंग से प्राकृतिक दृश्य उत्पन्न करना है। हालांकि, ऐसी विधियों में उत्पन्न छवियों के भीतर वस्तुओं को नियंत्रित करने की उनकी क्षमता में सीमाएं हैं। इसके विपरीत, हम दिखाते हैं कि इनपुट के रूप में कच्ची छवि पिक्सेल के साथ एक जनरेटिव एडवरसैरियल नेटवर्क को प्रशिक्षित करके, हम ऐसे दृश्य उत्पन्न कर सकते हैं जो वस्तुओं का निर्माण करते हैं और साथ ही इनपुट ऑब्जेक्ट्स के संयोजन के लिए उपयुक्त परिवेश उत्पन्न करते हैं।

हमने समान सेटिंग्स में उपयोग किए गए प्रासंगिक डेटासेट और मीट्रिक के साथ हमारे प्रस्तावित ढांचे का मूल्यांकन किया है। छवि एनोटेशन पर हमने अन्य पद्धति की तुलना में अत्याधुनिक परिणाम प्राप्त किए हैं।

इमेज जनरेशन टास्क पर हमने इंसेप्शन स्कोर को मेट्रिक्स के रूप में इस्तेमाल करते हुए अत्याधुनिक परिणाम हासिल किए हैं।

Contents

List of Figures

List of Tables