NOVEL DEEP LEARNING BASED COMPRESSIVE SENSING FRAMEWORK FOR RECONSTRUCTION-FREE RECOGNITION

RONAK GUPTA

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2021

©Indian Institute of Technology Delhi (IITD), New Delhi, 2021

NOVEL DEEP LEARNING BASED COMPRESSIVE SENSING FRAMEWORK FOR RECONSTRUCTION-FREE RECOGNITION

by

RONAK GUPTA

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2021

To my family

Certificate

This is to certify that the thesis titled **Novel Deep Learning based Compressive Sensing Framework for Reconstruction-free Recognition** being submitted by **Mr. Ronak Gupta** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of **Doctor of Philosophy** is a record of bona-fide research work carried out by his under my guidance and supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. 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.

Professor Santanu Chaudhury
Department of Electrical Engineering
Indian Institute of Technology Delhi
Delhi - 110016, India.

Professor Brejesh Lall
Department of Electrical Engineering
Indian Institute of Technology Delhi
Delhi - 110016, India.

Professor Narendra Ahuja
Information Technology Research Academy
Media Lab Asia
Delhi - 110030, India.

Acknowledgements

I would like to express my sincere gratitude to my advisors **Prof. Santanu Chaudhury**, **Prof. Brejesh Lall** and **Prof. Narendra Ahuja**. This thesis would not have been possible without their guidance, critical review and encouragement. Their intelligent ideas and comprehensive understanding helped me to establish the overall direction of the research, got me past many significant challenges and always motivated me to strive for excellence.

I thank the members of my thesis committee: Prof. S.D. Joshi, Prof. Prem Kalra and Prof. Sumantra Dutta Roy for their insightful suggestions which encouraged me to widen my research from various perspectives. I sincerely admire the contribution of my lab mates Ms. Swati Bhugra, Mr. Manoj Sharma, Mr. Vinay Kaushik, Dr. Prerana Mukherjee, Mr. Ajeet Singh and Dr. Siddharth Srivastava for extending their unstinting support and cooperating attitude during the course of this study. My thanks to Ushaji, Manrikaji, Dilipji and Amitji for all the timely logistic support.

I acknowledge the significant contribution of my parents. Their patience and support will always be my inspiration.

Ronak Gupta

Abstract

In computer vision, there are several types of inference tasks such as image classification, pose estimation, object localization, object segmentation, action classification, gesture classification, anomaly detection, etc. These inference tasks are not meant for human consumption; they help us to automate the visual understanding of the machine and decision-making process. Few tasks such as image classification, action classification in videos, anomaly detection in videos etc. require a method to capture spatial, appearance, temporal or spatio-temporal information. The raw data collected for these tasks have a lot of spatial, temporal and perceptual redundancies that have a minimal role while inferring. In this thesis, we present compressive data sensing techniques for the mentioned inference tasks to explore the spatial redundancy, particularly while avoiding additional decompression/reconstruction. Hence our inference techniques are also called reconstruction-free inference or redundancy-free inference.

The benefit of applying compressive sensing to inference tasks is that the computational load decreases, as the number of computations to be performed over compressed representation reduces compared to that for raw images/videos, for a given deep neural network architecture. Hence, relatively fast inference speed can be obtained over the same computational resource with compressive sensing. It can implicitly provide privacy to the sensitive information present in the raw data while performing inference with deep neural networks. The performance of deep neural networks with and without compressive sensing is high for inference tasks because the discriminative information is preserved in compressed representation.

Data sensing techniques such as random sensing, data-driven sparse compressive sensing and data-driven sensing with convolutional autoencoder are presented in this thesis. The

strength of these data sensing techniques is demonstrated for inference tasks such as action recognition in videos, image classification and fall detection in videos (anomalous activity recognition) by employing deep learning architectures. In random sensing, we perform block-based compressive sensing with fixed sensing matrices, which preserves spatial neighbourhood information in compressed representation at the block level. So convolution operation over the compressed representation has been explored. Also, 3D convolution kernels can work efficiently for learning spatio-temporal classifiers over the stack of compressed video frames. We present the results of random sensing applied for fall detection in videos. In data-driven sparse compressive sensing, again the compressed representation is learned at the block level. Hence applying 2D convolutions to the compressed representation of images, can capture spatial and appearance information. However, the predictions for video classification over the compressed video representation obtained from data-driven sparse sensing matrix, are not that accurate. In data-driven sensing with convolutional autoencoder, there are pixel unshuffling and shuffling operations involved which preserve spatial arrangement of video frames across the channels. Since the convolutional encoder does not shuffle the temporal arrangement, 3D convolutional kernels can be applied over the stack of the compressed representation of frames to learn a spatio-temporal classifier.

सार

कंप्यूटर विज़न में, कई प्रकार के अनुमान कार्य होते हैं जैसे छवि वर्गीकरण, मुद्रा अनुमान, वस्तु स्थानीयकरण, वस्तु विभाजन, क्रिया वर्गीकरण, हावभाव वर्गीकरण, विसंगति का पता लगाना, आदि। ये अनुमान कार्य मानव उपभोग के लिए नहीं हैं; वे मशीन और निर्णय लेने की प्रक्रिया की दृश्य समझ को स्वचालित करने में हमारी मदद करते हैं। कुछ कार्यों जैसे छवि वर्गीकरण, वीडियो में क्रिया वर्गीकरण, वीडियो में विसंगति का पता लगाना आदि के लिए स्थानिक, उपस्थिति, लौकिक या स्थानिक–लौकिक जानकारी को पकड़ने के लिए एक विधि की आवश्यकता होती है। इन कार्यों के लिए एकत्र किए गए कच्चे डेटा में बहुत अधिक स्थानिक, लौकिक और अवधारणात्मक अतिरेक होते हैं जिनकी कम से कम भूमिका होती है। इस थीसिस में, हम विशेष रूप से अतिरिक्त डीकंप्रेसन/पुनर्निर्माण से बचते हुए, स्थानिक अतिरेक का पता लगाने के लिए उल्लिखित अनुमान कार्यों के लिए संपीड़ित डेटा सेंसिंग तकनीक प्रस्तुत करते हैं। इसलिए हमारी अनुमान तकनीकों को पुनर्निर्माण–मुक्त अनुमान या अतिरेक–मुक्त अनुमान भी कहा जाता है। अनुमान कार्यों के लिए कंप्रेसिव सेंसिंग को लागू करने का लाभ यह है कि कम्प्यूटेशनल लोड कम हो जाता है, क्योंकि किसी दिए गए डीप न्यूरल नेटवर्क आर्किटेक्चर के लिए, कंप्रेस्ड रिप्रेजेंटेशन पर किए जाने वाले कंप्यूटेशंस की संख्या कच्ची छवियों / वीडियो की तुलना में कम हो जाती है। इसलिए, कंप्रेसिव सेंसिंग के साथ एक ही कम्प्यूटेशनल संसाधन पर अपेक्षाकृत तेज अनुमान गति प्राप्त की जा सकती है। यह गहरे तंत्रिका नेटवर्क के साथ अनुमान लगाते समय कच्चे डेटा में मौजूद संवेदनशील जानकारी को गुप्त रूप से गोपनीयता प्रदान कर सकता है। कंप्रेसिव सेंसिंग के साथ और बिना डीप न्यूरल नेटवर्क का प्रदर्शन अनुमान कार्यों के लिए उच्च है क्योंकि विभेदक जानकारी को संपीड़ित प्रतिनिधित्व में संरक्षित किया जाता है। इस थीसिस में डेटा सेंसिंग तकनीक जैसे रैंडम सेंसिंग, डेटा–ड्रिवेन स्पैस कंप्रेसिव सेंसिंग और डेटा–ड्रिवेन सेंसिंग विद कन्वेन्शनल ऑटोएन्कोडर प्रस्तुत किए गए हैं। इन डेटा सेंसिंग तकनीकों की ताकत का प्रदर्शन वीडियो में एक्शन रिकग्निशन, इमेज क्लासिफिकेशन और वीडियो में फॉल डिटेक्शन (विसंगत गतिविधि मान्यता) जैसे गहन शिक्षण आर्किटेक्चर को नियोजित करके किया जाता है। रैंडम सेंसिंग में, हम फिक्स्ड सेंसिंग मैट्रिसेस के साथ ब्लॉक–आधारित कंप्रेसिव सेंसिंग करते हैं, जो ब्लॉक स्तर पर संपीड़ित प्रतिनिधित्व में स्थानिक पड़ोस

की जानकारी को संरक्षित करता है। इसलिए संकुचित प्रतिनिधित्व पर कनवल्शन ऑपरेशन का पता लगाया गया है। इसके अलावा, 3 कनवल्शन कर्नेल संपीड़ित वीडियो फ्रेम के स्टैक पर अनुपात–अस्थायी क्लासिफायर सीखने के लिए कुशलता से काम कर सकते हैं। हम वीडियो में फॉल डिटेक्शन के लिए लागू रैंडम सेंसिंग के परिणाम प्रस्तुत करते हैं। डेटा–संचालित विरल कंप्रेसिव सेंसिंग में, फिर से ब्लॉक स्तर पर कंप्रेस्ड रिप्रेजेंटेशन सीखा जाता है। इसलिए छवियों के संकुचित प्रतिनिधित्व के लिए 2 संकल्पों को लागू करना, स्थानिक और उपस्थिति जानकारी को कैप्चर कर सकता है। हालांकि, डेटा–संचालित विरल सेंसिंग मैट्रिक्स से प्राप्त संपीड़ित वीडियो प्रतिनिधित्व पर वीडियो वर्गीकरण के लिए भविष्यवाणियां सटीक नहीं हैं। कन्वेन्शनल ऑटोएन्कोडर के साथ डेटा–संचालित सेंसिंग में, पिक्सेल अनशफलिंग और शफलिंग ऑपरेशन शामिल होते हैं जो चैनलों में वीडियो फ्रेम की स्थानिक व्यवस्था को संरक्षित करते हैं। चूँकि कनवल्शनल एनकोडर टेम्पोरल अरेंजमेंट में फेरबदल नहीं करता है, स्पैटो–टेम्पोरल क्लासिफायरियर सीखने के लिए फ्रेम के कंप्रेस्ड रिप्रेजेंटेशन के स्टैक पर 3 कन्वेन्शनल कर्नेल लागू किया जा सकता है।

Contents

Certificate iii

Acknowledgements v

Abstract vii

List of Figures xiii

List of Tables xv

1 Introduction 1

1.1 Scope and Objectives 2

1.2 Context 5

1.3 Major Contributions of the Thesis 7

1.4 Outline of Thesis 8

2 Prior Works 11

2.1 Literature review 11

2.1.1 Deep learning architectures for Action recognition 11

2.1.2 Anomalous activity recognition 17

2.1.3 Compressive sensing for computational imaging 20

3 Fall detection with compressive sensing 25

3.1 Introduction 25

3.2 Proposed framework for fall detection . . . 27
3.3 Experimental Results . . . 29
3.3.1 Fall and Action Datasets . . . 29
3.3.2 Implementation Details . . . 31
3.4 Conclusion . . . 32

4 Data Adaptive Compressed Sensing using deep neural network for Image recognition 33
4.1 Introduction . . . 33
4.2 Proposed framework . . . 35
4.2.1 Data Adaptive CS sampling sub-network . . . 35
4.2.2 Image recognition sub-network . . . 37
4.3 Experimental Results . . . 37
4.4 Conclusion . . . 40

5 Data-driven sensing for action recognition 41
5.1 Introduction . . . 41
5.2 Proposed framework . . . 43
5.2.1 Data-driven sensing through Convolutional Autoencoder (CAE) . . . 43
5.2.2 Learning Action recognition in compressed domain . . . 46
5.3 Experimental Results . . . 46
5.3.1 Implementation details . . . 48
5.4 Discussion on design of privacy protocol for anomaly detection . . . 49
5.5 Conclusion . . . 50

6 Conclusions and Future work 51
6.1 Conclusions . . . 51
6.1.1 Conclusion of this research . . . 51
6.2 Future Work . . . 53

Bibliography 55

Publications 65

Biography 66

List of Figures

1.1 The three objectives and their position in the source complexity and task integration matrix . . . 3

1.2 Protocol for our works . . . 7

3.1 Compression technique . . . 28

3.2 Fall detection architecture . . . 28

3.3 Fall example from multiple cameras [5] . . . 30

4.1 Proposed framework for Data-Adaptive CS based image recognition . . . 35

4.2 Proposed sensing network . . . 36

5.1 Proposed framework for data-driven sensing and reconstruction free action recognition . . . 43

5.2 Data driven Sensing framework for learning undersampled measurements (dimensions shown for undersampling ratio of 4x4) . . . 45

5.3 Inception module in our CAE architecture . . . 45

5.4 Action recognition architecture . . . 46

5.5 Privacy Protocol . . . 50

6.1 CS based computer vision platform . . . 53

List of Tables

3.1 Accuracy on test split of Kinetics dataset with different deep learning architectures 30

3.2 Accuracy on test split of Kinetics-10* with our 3D ConvNet architecture 30

3.3 Performance of various techniques over Multi-camera fall dataset and UR fall dataset 31

4.1 Performance of different CS based image classification frameworks on MNIST dataset 39

4.2 Performance of different CS based image classification frameworks on CIFAR-10 dataset 39

4.3 Performance of different CS based image classification frameworks on F-MNIST dataset 39

5.1 Performances of different Action recognition framework in compressed domain 47

5.2 Avg. Accuracy on UCF101 splits 48

5.3 Performance with respect to number of Inception modules in Proposed data-driven sensing encoder 48