FEW-SHOT LEARNING DRIVEN BY CONTRASTIVE TECHNIQUES

ANVAYA RAI

BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY AND MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY DELHI

FEBRUARY 2024

FEW-SHOT LEARNING DRIVEN BY CONTRASTIVE TECHNIQUES

by

ANVAYA RAI

BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY AND MANAGEMENT

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

INDIAN INSTITUTE OF TECHNOLOGY DELHI

FEBRUARY 2024

Dedicated to
My Teachers
Family
&
Friends

THESIS CERTIFICATE

This is to certify that the thesis titled **Few-shot Learning driven by Contrastive Techniques**, submitted by **Anvaya Rai**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by him under my supervision. He has fulfilled the requirements for the submission of the thesis, which to the best of our knowledge has reached the required standard.

The material contained in the thesis has not been submitted in part or full to any other University or Institute for the award of any degree or diploma.

Prof. Brejesh Lall
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi -110016

Date: FEBRUARY 2024

ACKNOWLEDGEMENTS

This thesis is the outcome of a wonderful seven year working experience under the guidance of **Dr. Brejesh Lall**, to whom I wish to express my appreciation and gratitude. His vision, creativeness, and enthusiasm have not just fueled this research but also have a lasting influence in my career as a role model. I would also like to express my deep sense of gratitude towards **Dr. R. K. P. Bhatt** and **Dr. S. D. Joshi** for his support and guidance whenever required during the entire course work.

I take this opportunity to thank my friends and colleagues at IIT Delhi and C-DOT for their generous help and fruitful collaborations, especially Ms. Astha Zalani Jain and Mr. Raghwendra Prakash Singh. I would specially like to thank my friends, Mr. Dharmesh Kumar, Mr. Adarsh Tripathi and Mr. Alok Sharma, for constantly boosting my moral and motivating me in my research.

My deepest gratitude goes to my wife **Ms. Pragya**. She had given me the much needed comfort and inspiration for completing this thesis.

Finally, I would like to dedicate this work to my parents, whose constant support and encouragement have really brought me here. Lastly, I am eternally grateful to the entire universe for the love, support, and opportunities it has bestowed upon me to make a positive impact on the world.

Anvaya Rai

ABSTRACT

KEYWORDS: Hyperspherical Manifold; Extreme Learning Machines; Principal Component Analysis Extreme Learning Machine; Linear Discriminant Analysis Extreme Learning Machine; Few-shot; Contrastive Learning; Visual SLAM; Vision based navigation; Transfer Learning; Manifold Generation.

The few-shot learning concept using contrastive techniques has come into existence because the Deep neural networks (DNN) demand huge amounts of manually annotated training data. In many scenarios, this type of data acquisition is highly tedious and at times impossible. Thus, there was a need to learn from few examples, using few iterations and augmentations. In this work, we have proposed a novel contrastive learning based few-shot technique and applied it in the area of Face Recognition. We call it as ***Latent Feature Transformed Contrastive Learning*** *(LFT-CLR)*. The suggested framework can be extended to other domains as well. Additionally, we also propose a novel contrastive loss, called ***Weighted Normalized Temperature-scaled Cross Entropy Loss*** *(wNT-Xent)*, to refine a pre-trained DNN. The suggested approach has been validated using finely curated datasets based upon LFW [20].

In the proposed inference pipeline, the fine-tuned DNN is followed by PCA/ LDA based latent feature space transformation, resulting in extracting a manifold from the feature space to project the test samples. This feature space transformation helps in dimensionality reduction of the latent feature space and projection of query images onto an optimally dense subspace of the original latent feature space, leading to accurate and faster inference.

In addition to Face Recognition, we also extend this work in the domain of Visual SLAM and introduce a novel approach to accurately localize a subject in indoor environments by using the scene images captured from the subject's mobile phone camera. We present a novel deep neural network (DNN), called ***InPosNet***, that generates a concise representation of an indoor scene while being able to distinguish between their inherent symmetry. It also enables the user in real time distinction between the images of the same location but captured from different orientations, there by enabling the user to detect the orientation along with position. The novel DNN presented in the work is motivated by MobileNetv3-Small [54]. A localization accuracy of less than 1 meter from ground truth is achieved and enumerated through the experimental results. The goal is to present a vision based system that will have

the ability to be used for indoor positioning, without any need for additional infrastructure.

अमूर्त

प्रमुख शब्द: हाइपरस्फेरिकल मैनिफोल्ड; चरम शिक्षण मशीनें; प्रमुख घटक विश्लेषण चरम शिक्षण मशीन; रैखिक विभेदक विश्लेषण चरम शिक्षण मशीन; कुछ-शॉट; विरोधाभासी शिक्षा; विजुअल स्लैम; दृष्टि आधारित नेविगेशन; स्थानांतरण सीखना; अनेक गुना पीढ़ी.

तकनीकों का उपयोग करके कुछ-शॉट सीखने की अवधारणा अस्तित्व में आई है क्योंकि डीप न्यूरल नेटवर्क (डीएनएन) बड़ी मात्रा में मैन्युअल रूप से एनोटेट प्रशिक्षण डेटा की मांग करते हैं। कई परिदृश्यों में, इस प्रकार का डेटा अधिग्रहण अत्यधिक कठिन और कभी-कभी असंभव होता है। इस प्रकार, कुछ पुनरावृत्तियों और संवर्द्धनों का उपयोग करके कुछ उदाहरणों से सीखने की आवश्यकता थी। इस कार्य में, हमने एक नवीन विरोधाभासी शिक्षण आधारित कुछ-शॉट तकनीक का प्रस्ताव रखा है और इसे चेहरा पहचान के क्षेत्र में लागू किया है। इसे हम **वेटेड सिमसीएलआर** कहते हैं। सुझाए गए ढांचे को अन्य डोमेन तक भी बढ़ाया जा सकता है। इसके अतिरिक्त, हम पूर्व-प्रशिक्षित डीएनएन को परिष्कृत करने के लिए एक नवीन विपरीत हानि का भी प्रस्ताव करते हैं, जिसे **भारित सामान्यीकृत तापमान-स्केल्ड क्रॉस एन्ट्रॉपी लॉस** कहा जाता है। सुझाए गए दृष्टिकोण को एलएफडब्ल्यू पर आधारित बारीक क्यूरेटेड डेटासेट का उपयोग करके मान्य किया गया है।

सुझाई गई अनुमान पाइपलाइन में, फाइन-ट्यून किए गए डीएनएन के बाद पीसीए/एलडीए आधारित अव्यक्त फीचर स्पेस परिवर्तन होता है। यह फ़ीचर स्पेस परिवर्तन अव्यक्त फ़ीचर स्पेस की आयामीता में कमी और मूल अव्यक्त फ़ीचर स्पेस के इष्टतम घने उप-स्थान पर क्वेरी छवियों के प्रक्षेपण में मदद करता है, जिससे सटीक और तेज़ अनुमान लगाया जा सकता है।

फेस रिकग्निशन के अलावा, हम विज़ुअल स्लैम के क्षेत्र में भी इस काम का विस्तार करते हैं और विषय के मोबाइल फोन कैमरे से कैप्चर की गई दृश्य छवियों का उपयोग करके इनडोर वातावरण में किसी विषय को सटीक रूप से स्थानीयकृत करने के लिए एक नया दृष्टिकोण पेश करते हैं। हम **इनपोसनेट** नामक एक नया डीप न्यूरल नेटवर्क डीएनएन प्रस्तुत करते हैं, जो उनकी अंतर्निहित समरूपता के बीच अंतर करने में सक्षम होने के साथ-साथ एक इनडोर दृश्य का संक्षिप्त प्रतिनिधित्व उत्पन्न करता है। यह उपयोगकर्ता को एक ही स्थान की लेकिन अलग-अलग ओरिएंटेशन से कैप्चर की गई छवियों के बीच वास्तविक समय में अंतर करने में सक्षम बनाता है, जिससे उपयोगकर्ता स्थिति के साथ-साथ ओरिएंटेशन का पता लगाने में सक्षम होता है। कार्य में प्रस्तुत उपन्यास डीएनएन मोबाइलनेट बनाम 3-छोटा से प्रेरित है। प्रायोगिक परिणामों के माध्यम से जमीनी सच्चाई से 1 मीटर से कम की स्थानीयकरण सटीकता प्राप्त की जाती है और उसकी गणना की जाती है। लक्ष्य एक दृष्टि आधारित प्रणाली प्रस्तुत करना है जिसमें अतिरिक्त बुनियादी ढांचे या बाहरी हार्डवेयर की आवश्यकता के बिना, इनडोर पोजिशनिंग के लिए उपयोग करने की क्षमता होगी।

ABBREVIATIONS

ELM	Extreme Learning Machines
PCA-ELM	Principal Component Analysis Extreme Learning Machine
LDA-ELM	Linear Discriminant Analysis Extreme Learning Machine
FSL	Few-shot Learning
CL	Contrastive Learning
SLAM	Simultaneous Localisation and Motion
NT-Xent	Normalized Temperature-scaled Cross Entropy Loss
MoCo	Momentum Contrast
SimCLR	Simple framework for Contrastive Learning
SimSiam	Simple Siamese Representation
NNCLR	Nearest Neighbour Contrastive Learning
NNSiam	Nearest Neighbour Siamese Representation
BYOL	Bootstrap Your Own Latent
SLFN	Single Hidden Layer Feed forward Network
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
LARS	Layer-wise Adaptive Rate Scaling
DWT	Discrete Wavelet Transform
SOTA	State of the Art
LFT-CLR	Latent Feature Transformed Contrastive Learning
wNT-Xent	Weighted NT-Xent Loss
InPosNet	Indoor Positioning Network

NOTATION

S	Dimension of Feature Transformed Latent Space
D	Dimension of Original Latent Feature Space
J	Number of Novel Dataset classes
U	Feature Transformed Latent Space
x	Image sample from Novel Dataset
z	Image sample representation in Projected Space
τ	Normalising Temprature for Loss Function
θ	Online Network Parameter
ζ	Target Network Parameter
Φ	Activation Function

Contents

List of Tables

List of Figures