CHAPTER 10

COMPUTER SCIENCE

Doctoral Theses

01. DUA (Arti)
Locating Covert Data in Multiple IPv6 Header Fields.
Supervisor: Sr. Prof. Punam Bedi
Th 27427

Abstract

The growing usage of Internet has led to tremendous increase in cyber-attacks like stegomalware, online frauds, security breaches, scams, intrusions etc. Stegomalware are the modern attacks that use information-hiding techniques to hide and transfer malware. They exploit innocent-looking cover medium like images, videos, audios, network traffic flows to hide and carry malicious data. The use of network traffic flows to hide secret information is termed Network Steganography. Implementation of network steganographic methods create Network Covert Channels (NCCs). The NCCs are classified broadly into two classes: Storage-based NCCs, and Timing-based NCCs. Generally, storage-based NCCs offer higher capacity than timing-based NCCs. Thus, they can be a preferred choice of attackers to transfer more malware per network packet. Storage-based NCCs can operate at different layers of TCP/IP model. This thesis focuses on analysing and locating storage-based NCCs over the Internet layer protocols viz. ARP, IPv4, and IPv6. To understand the working of NCCs over these protocols, four NCCs (two over ARP, one over IPv4 and one over IPv6) are proposed. Further, these and other existing NCCs developed using the above-mentioned protocols are analysed for their feasibility over the Internet. Out of these NCCs, the ARP-based NCCs work over a LAN and hence have a limitation of short-distanced covert communication. IPv4 and IPv6-based NCCs have no such limitation and can be good candidates for covert communications. Moreover, to the best of our knowledge, work in literature discuss only detection of storage-based covert channels and no work identifies the location of hidden data present in one or a combination of header fields of a network packet. Identifying the location of covert data is important for interpreting or decoding hidden information. Thus, with this aim, this thesis proposes systems to locate the storage area of covert data in header fields of IPv4 and IPv6 protocol.

Contents

02. DWIVEDI (Kountay)

An Explainable AI-driven Approach to Non-Small Cell Lung Cancer Biomarkers Discovery.

Supervisors: Prof. Naveen Kumar and Dr. Ankit Rajpal

Th 27428

Abstract

Non-small Cell Lung Cancer (NSCLC), which has a 5-year survival rate of 17.8%, is considered the most deadly lung cancer. Its molecular variability characterizes its major subtypes, lung adenocarcinoma, and lung squamous cell carcinoma. This thesis provides eXplainable AI (XAI) driven deep learning approaches for NSCLC biomarker identification using genomic, transcriptomic, and epigenomic data. The copy number variation (CNV) data provides a measure of genomic instability. We present a modification of the standard L1-regularized gradient descent algorithm to identify NSCLC biomarkers. An XAI-based feature identification reveals 20 NSCLC-relevant biomarkers. RNA-Seq gene expression and DNA methylation data reveal transcriptomic and epigenomic variations. Using these datasets, we present an XAI-driven deep learning framework to find transcriptomic and epigenomic biomarkers. The proposed framework comprises an autoencoder to reduce high-dimensional input feature space, a feed-forward neural network for NSCLC subtype classification, and an XAI-based biomarker discovery module that discovers 52 transcriptomics and 7 epigenomics biomarkers. The discovered biomarkers are assessed for NSCLC-subtype classification and targeted therapeutic efficacy. Using the discovered genomic, transcriptomic, and epigenomic biomarkers, we achieved a 10-fold cross-validation classification accuracy of 84.95%, 95.74%, and 91.53%, respectively. Using the Drug-Gene Interaction Database, out of the discovered biomarkers, 12 genomic, 14 transcriptomic, and 4 epigenomic were found to be druggable. Using the KM Plotter tool, several discovered biomarkers were found useful in predicting NSCLC survival. While the majority of the discovered biomarkers confirm the NSCLC literature, we also discovered 14 new biomarkers that could be evaluated for lung cancer-targeted therapy. While using XAI-driven deep-learning approaches, we could exploit various omic data to discover several biomarkers useful in NSCLC subtype classification, the work presented in this thesis may be extended using multimodal data, for example, combining images with the omic data.

Contents

03. GARG (Manika)

Mitigating Academic Dishonesty in Online Assessment using Machine Learning.

Supervisor: Prof. Anita Goel

Th 27429

Abstract

The recent advancement in the field of information technology has resulted in the proliferation of online education all over the world. Much like traditional classroom education, assessments are an integral component of online education. During the online assessment, evaluation of the learning outcomes presents challenges mainly due to academic dishonesty among students. It results in unfair evaluations that raises questions credibility of online assessments. There exist several types of dishonesty in online assessments including exploiting the available Internet for finding solutions (Internet-as-a-Forbidden-Aid), illicit collaboration among students (Collusion) and third-party persons taking assessment on behalf of the genuine student (Impersonation). Several research studies have proposed solutions for addressing dishonesty in online assessments. These solutions include strategies for designing assessments that are resistant to cheating, implementing proctoring and formulating integrity policies. While these methods can be effective, their implementation is often resource-intensive and laborious, posing challenges. Other studies propose the use of Machine Learning (ML) methods for automated dishonesty detection. However, these approaches often lack clarity in selecting appropriate features and classifiers, impacting the quality of results. Furthermore, the lack of sufficient and accurate training data further leads to poorly tuned models that are prone to making errors. There is a need to develop ML models tailored to specific type of dishonesty as the test-taking patterns vary with the dishonesty type. In this thesis, we focus on MCQ-based assessments. We consider the three types of dishonesty: (1) Internet-as-a-Forbidden-Aid, (2) Collusion, and (3) Impersonation, observed in MCQ-based assessments. We developed individual ML models to detect students involved in each type of dishonesty during the assessment. The results also facilitate understanding the test-taking pattern of students and providing recommendations for cheat-proof assessment design. Finally, we proposed an Academic Dishonesty Mitigation Plan (ADMP) that provides a comprehensive approach for addressing the diverse forms of dishonesty.

Contents

04. GOLE (Pushkar)
Lightweight and Few-Shot Image-Based Plant Disease Diagnosis and Remedy Recommender System.
Supervisor: Sr. Prof. Punam bedi
Th 27430

Abstract

Early-stage plant disease diagnosis can be helpful to minimize crop yield loss and maximize the farmer’s profit and it is a big challenge in the growth of farming sector. A lot of work has been done in the literature in this area using various Deep Learning techniques but most of these techniques utilize large number of trainable weight parameters and large number of annotated leaf images for training. Therefore, the aim of this thesis is to develop a lightweight Deep Learning model which requires a smaller number of annotated leaf images for training, as annotating leaf images is laborious and time-consuming task. Hence, an attempt has been made in this thesis to first develop a hybrid Deep Learning model to identify a plant disease with a smaller number of trainable weight parameters. Next, an improved and lightweight Vision Transformer (ViT) model named "TrIncNet" is proposed in the thesis for detecting multiple type of plant diseases. Disease severity is required in order to take necessary steps for curing the identified disease as it can provide the quantitative assessment of the damage caused by the pathogen of identified disease. Hence, this thesis proposed a lightweight and few-shot framework named PDSE-Lite for plant disease diagnosis and severity estimation. The severity of identified disease is estimated by calculating the percentage of diseased pixels out of total leaf pixels in this framework. Lastly, an android mobile application named "PlantD2R2S-Lite" has been developed in this thesis for identifying plant diseases after capturing the leaf image, estimating severity of the identified disease, generating the advisory for curing the plant disease in either English or Hindi language. In order to generate this advisory, a pre-trained BERT model is fine-tuned on the text of two research papers which have detailed descriptions of apple leaf diseases and their management.

Contents

05. JAIN (Kirti)
Modeling Behavioral and Epidemic Dynamics in Social Contact Networks.
Supervisor: Prof. Vasudha Bhatnagar and Prof. Sharanjit Kaur
Th 27879

Abstract

Contagious diseases spread in the population through a contact network and their spread is a function of the complex interplay of the biology of the contagion and behavior of the population. The emotion of fear due to disease not only influences human behavior but also transmits to social contacts in the network. Therefore, modeling human behavior in conjunction with epidemic spread is a precursor to reliable estimates of epidemic size and span. In this work, we introduce a novel framework, the Individual-based Fear Model (IBFM) that associates fear-index with individuals to indicate the extent to which they follow health and hygiene protocols as a self-protective measure against disease. By considering individual-level behavior and the varying levels of fear associated with disease transmission, the framework acknowledges heterogeneous individual responses to epidemics. Since societies are organized into communities, we study the impact of the collective behavior of individuals in a community on the size and span of the epidemic. In order to improve the estimates of epidemic variables for network simulation, we propose a framework to create a wire-frame that mimics the modular contact network of the population in a geography using census data. We test the effectiveness of a demography-laced modular contact network using real-life COVID-19 data for three Indian states. The framework is a potent decision-making instrument for urban planners, demographers, and social scientists. Since network simulations of epidemic spread are computationally expensive, we launch a systemic investigation into the possibility of predicting three epidemic variables, viz. peak day, peak infections, and span of the epidemic using the Regression Chain Model trained on topological properties of networks. We find that the predictions are fairly accurate for peak day and peak cases.

Contents

06. JAIN (Mansi) Nee Mansi Sood
Framework for Incremental Updates of Domain-Specific Vocabulary.
Supervisor: Dr. Harmeet Kaur
<u>Th 27880</u>

Abstract

Accessing relevant information in specific contexts requires understanding pertinent vocabulary, but conventional vocabularies often fall short in specialized fields like scientific literature, legal documents, and medical texts. Domain-specific vocabulary (DSV) enhances comprehension in these fields and is crucial for fields like Information Retrieval, Machine Learning, NLP, AI, and Natural Language Generation. Maintaining DSV relevance amidst evolving language and expanding digital content is challenging, as current methods focus on initial creation and neglect efficient updates. Our research develops a systematic framework for incremental DSV maintenance using incremental learning to integrate new terms and remove obsolete ones while preserving vocabulary integrity. We create the DSV by extracting unigrams and collocations from a labeled dataset, optimizing it using Stochastic Gradient Descent (SGD) with L1/L2 regularization, and storing the resulting sparse vocabulary and dataset footprint in a repository named DocLib. The proposed framework uses the dataset footprint from DocLib and augments it with newly received add-on datasets to update the vocabulary. Algorithms evaluate the incremental model`s stability and plasticity, demonstrating consistent performance and adaptability to new data without compromising existing knowledge. The framework optimizes processing time and memory requirements, outperforming conventional methods. Task-based evaluations for NLP tasks, like text classification, confirm the vocabulary’s applicability in discerning domain-related data. We apply this framework to the "Agricultural and Biological Sciences," "Computer Science," and "Psychology" domains tracking vocabulary evolution over 24

years, addressing significant research gaps, and ensuring continuous DSV relevance.

Contents

07. KHURANA (Alka)
Leveraging Non-negative Matrix Factorization for Extractive Text Summarization.
Supervisor: Prof. Vasudha Bhatnagar
Th 27431

Abstract

This thesis presents a design of unsupervised methods for extractive text summarization using Non-negative Matrix Factorization (NMF). NMF is a popular matrix decomposition technique, which uncovers the latent semantic space of the text and divulges inter-relation between three prime semantic units of the text. We exploit the inter-relationships among the three semantic units, viz. terms, sentences, and latent topics, in an innovative manner to extract informative sentences from the salient topics in the text. The selected sentences form the summary of the text. The main contributions of the thesis are – i. We demonstrate that use of NMF ensembles to overcome the stochasticity yields marginal performance gain and fails to justify the incurred cost. ii. We propose two novel algorithms that excavate inter-relationship among the semantic units in the text. The first approach NMF-TR is term-oriented, which leverages the information carried by the terms and quantifies the importance of a sentence as an additive function of the unique terms present in the sentence. Next approach is NMF-TP, a topic-oriented approach that quantifies the importance of a sentence as weighted contribution of latent topics. iii. The third algorithm E-Summ follows an information-theoretic approach, which exploits the information contained in topic and sentence entropies and subsequently uses Knapsack optimization algorithm to maximize the information conveyed by sentences selected for inclusion in the summary. iv. We design an unsupervised algorithm, P-Summ, that generates an extractive summary of scientific scholarly text to meet the personal knowledge needs of the user. The method delves into the latent semantic space of the document exposed by wNMF, and scores sentences in consonance with the knowledge needs of the user. We also propose a multi-granular evaluation framework, which assesses the quality of generated personal summaries at three levels of granularity.

Contents

Personalized Summarization of Scientific Scholarly Texts 7. Conclusion and Future Directions. Appendix.

08. PREETI

Modeling and Prediction of Short and Long-Horizons for Time Series Data Using Extreme Learning Machines and Deep Learning Frameworks.

Supervisors: Prof. Ram Pal Singh and Dr. Rajni Bala

Th 27432

Abstract

Time series forecasting is an important and challenging problem for various engineering applications in finance, energy, weather, and information technology. The analysis of time series data helps to understand their statistical properties and the general tendency of the data. For example, forecasting solar radiation for a given region will help identify the places for installing large-scale photovoltaic systems, designing energy-efficient buildings, and estimating energy. The application of time series forecasting can analyze the causes and conditions prevailing during past events to decide on future policies and programs. We have proposed extreme-learning and deep-learning-based frameworks to predict the short- and long-horizons values for univariate and multivariate time series (MTS). Time series are inherently noisy, non-stationary, and non-periodic. As a result, the relationship between input and output variables alters continuously. Existing methods find it difficult to learn all these changes. Univariate time series find it difficult to find memory order for the input-output phase reconstruction step. On the other hand, MTS finds it difficult to learn the dynamic interdependencies among various input variables of data. We aim to overcome the multiple limitations of traditional methods and enhance the predictive accuracy of time series forecasting. We began by proposing novel online sequential extreme learning machines with L2,1-norm regularization (L2,1 OS-ELM) for short horizon Prediction in MTS series. Next, we proposed Hybrid Kernel-based extreme learning machine (HKELM) for short horizon prediction in the MTS series. To further improve the prediction accuracy, a deep learning-based algorithm is proposed that is an autoencoder-based extreme learning machine (ELM-AE) for short-horizon prediction in a univariate time series. Long-term and long-sequence prediction for MTS data can be solved by learning the long-term dependency period in a time series. So, we propose a stacked LSTM algorithm for long-horizon forecasting in a univariate time series. Different attention-based models, like RNN variants and transformers, are actively researched to provide long-term and long-sequence predictions for a time series. The complex temporal patterns of the data prevent these models from finding a reliable period of dependency. Our proposed attention-based models, namely, Dual Stage Attention-based Transformer with Time to Vec (DST2V-Transformer) and Frequency Attention-based Transformer (FANT) for long-term and long-sequence forecasting in the MTS series. The proposed models give long-term and long-sequence predictions for time series with enhanced accuracy compared to state-of-the-art algorithms. These models can be used in practical applications to provide forecasting as a service to different companies, organizations, and individuals. This will be very beneficial to increase profit for an organization and to decrease the chances of losses.

Contents

09. TANEJA (Shashi Bhushan)
Automated Deployment in Constructive Simulation Using Hybrid AI Models.
Supervisor: Sr. Prof. Punam Bedi
Th 27433

Abstract

Constructive simulations are the applications used by the military for the training of their commanders in planning and analyses of various threats and Courses of Action. Research in the area of military modelling and simulation is towards addressing the trade-off so as to have accurate detailed outcomes while abstracted inputs with limited required human commanders in simulations exercises. In this thesis, the proposed research has addressed the automated planning by taking two specific problems of deployment of infantry unit and deployment of artillery major weapons in the mountainous terrain. The problem is generalised to the required abstractions for the research objective. The deployment of units is the important factors which influence military commanders on the concept of operations in different areas. As per the higher commander's intentions and overall plan of operations, the tactical commander appreciates the likely deployment areas on the map board followed by ground reconnaissance. The process of selection of area for deployment is based on the appreciation of a particular commander for that scope of operations. Another problem addressed here is for automated optimal deployment of the major artillery weapons in mountainous terrain. The Mountain is usually characterised by rugged terrain and variation in slopes and elevations that makes deployment a challenge. The proposed modelling approach for Automated deployments is solved for static warfare scenario and further extended to scenario to address the enemy reaction and changing location. PSO based heuristic optimisation techniques (APSO, PP-PSO) are used to generate the deployment location in the given mountainous terrain digital map. To address the dynamic enemy changing location, the hybrid approach, MOP-N is proposed which uses Monte Carlo simulation, PP-PSO and NN. Solution proposed in this research will also aid commanders in planning for the selection of ideal deployable areas in a particular terrain.

Contents

10. BHARTI (Urmil)
Framework for Designing Serverless Applications.
Supervisors: Prof. Anita Goel and Prof. S. C. Gupta
Th 27426

Abstract

Serverless computing, or Function-as-a-Service (FaaS), offers a simplified cloud computing model that is gaining popularity due to its simplicity, cost-effective billing, and inherent scalability. This model frees developers from managing server infrastructure, allowing them to concentrate on application development. Serverless applications are based on an event-driven architecture where independent functions operate and scale independently. However, serverless computing introduces significant challenges, particularly in application design due to its constrained resource environment and lack of inherent state management. Serverless functions, which are stateless and short-lived, face strict limits on CPU, memory, disk space, and execution time. These limitations can lead to abrupt terminations if a function exceeds its resource allocations. Moreover, the stateless nature of these functions complicates tasks that require state propagation between functions within workflows, presenting hurdles in applications that demand complex function compositions. The current serverless platforms do not natively support function composition. Due to this limitation, it is challenging to implement data parallel and task parallel applications where dynamic parallel function composition is required. Although serverless providers provide orchestration services for creating workflows, these services are primitive and do not allow dynamic parallel function composition to accommodate fluctuating workloads. This limitation hinders the potential of serverless applications in scenarios requiring robust, scalable parallelism, forcing some to rely on traditional server-based models to achieve functionality, thereby not fully leveraging the serverless model`s advantages. This thesis proposes a framework designed to overcome serverless computing`s limitations to address these issues. The framework assists in designing applications that can effectively utilize the scalability of serverless computing while adhering to its operational limitations. By incorporating design techniques that overcome the constraints of serverless computing and exploit its scalability, this framework ensures that serverless applications can achieve their full potential, enhancing adaptability in this new cloud computing paradigm.

Contents