Searchable Abstracts Document

SIAM International Conference on Data Mining

(SDM24)

April 18–20, 2024

The Westin Houston, Memorial City, Houston, Texas, U.S.

This document was current as of April 5, 2024 Abstracts appear as submitted.

3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 U.S.
Telephone: 800-447-7426 (U.S. & Canada) +1-215-382-9800 (Worldwide)
meetings@siam.org

IP1

The Critical Role of Cyber Infrastructure in City Innovation and Beyond

Cities, human's greatest inventions, offer many opportunities for science and technology innovations. The increasingly available big data paints a bright future for our cities. However, significant hurdles stand in our way, primarily due to the absence of the right cyber infrastructure to easily share, analyze, and act on data. In this talk, I will share my firsthand experiences working with city practitioners. I have seen how difficult it is to make real progress without the appropriate digital tools and systems in place. This shortfall prevents us from addressing some of the most pressing challenges cities face. I will discuss how to create better cyber infrastructure for our cities. Moreover, I will argue that such infrastructure is vital not only for city innovations but also across

Zhenhui (Jessie) Li
Pennsylvania State University
jessielzh@gmail.com

IP2

Explainable Machine Learning for Robust Data Analysis

The past decade has been a very exciting time for machine learning (ML) research. Significant research effort has focused on improving predictive performance of Deep Neural Networks (DNN) by proposing increasingly complex architectures which have surpassed even human-level performance. Even though these methods demonstrate incredible potential in saving valuable man-hours and minimizing inadvertent human mistakes, their adoption has been met with rightful skepticism and extreme circumspection in critical applications like medical diagnosis, credit risk analysis, etc. The most paramount of these challenges is the lack of rationale behind DNN predictions - making them notoriously a black-box in nature. In extreme cases, this can create a lack of alignment between the designer's intended behavior and the model's actual performance. In this talk, I will discuss our recent research on explainable deep learning, in particular, I will discuss the concept learning models and show how the concept-based learning models and example-based learning models can be designed for explainable deep tabular learning. I will also discuss the potential of using Knowledge Graphs to enhance the explainability and performance of Large Language Models (LLMs)

Aidong Zhang
University of Virginia
aidong@virginia.edu

IP3

Well-being, AI, and You: Developing AI-based Technology to Enhance our Well-being

Many applications claim to enhance our well-being, whether directly by aiding meditation and exercise, or indirectly, by guiding us to our destinations, assessing our sleep quality, or helping us manage our daily tasks. However, the truth is that the potential of technology to improve our well-being often eludes us, and this is happening at the dawn of an era where AI is supposed to usher in a new generation of personalized assistants. Presently, we find ourselves more distracted than ever, devoting excessive time to pondering life's minutiae, and struggling to fully embrace the present moment. Part of the reason that our well-being is not benefiting fully from technology is the fact that each of these apps focuses on a specific aspect of well-being, lacking coordination with other apps. This situation is reminiscent of the early days of computer programming when each program interacted directly with the computer's hardware. Drawing from this analogy, this talk will begin by describing a set of mechanisms that can facilitate better cooperation between well-being applications, effectively proposing an operating system for well-being. This operating system comprises a data repository, referred to as a personal timeline, which captures your past experiences and future aspirations. It also includes mechanisms for utilizing your personal data to provide improved recommendations and life plans, and, lastly, a module to assist in nurturing and navigating crucial relationships in your life. We will then delve into the technical challenges involved in building the components of the operating system. In particular, we will focus on the creation of your life experiences timeline from the digital data you create on a daily basis. In this context, we will identify opportunities for language models to be a core component on which we build systems for querying personal timelines and for supporting other components of the operating system. In particular, the challenge of answering questions about your timeline raises important challenges in the intersection of large language models, data mining and structured data.

Alon Halevy
Amazon Web Services
halevy@amazon.com

IP4

AI for Biodiversity: Combatting Extinction Together

We are in the middle of the 6th extinction, losing the planet's biodiversity at an unprecedented rate and scale. In many cases, we do not even have the basic numbers of what species we are losing and how many. New data collection technology, such as GPS, high-definition cameras, UAVs, genotyping, and crowdsourcing, are generating data about the living planet that are orders of magnitude richer than any previously collected. AI can turn these data into high resolution information source about living organisms, enabling scientific inquiry, conservation, and policy decisions. The talk will present a vision and examples of trustworthy AI for biodiversity, discussing opportunities and challenges.

Tanya Y. Berger-Wolf
The Ohio State University
berger-wolf.1@osu.edu

CP1

Local Differential Privacy in Graph Neural Networks: a Reconstruction Approach

Graph Neural Networks have achieved tremendous success in modeling complex graph data in a variety of applications. However, there are limited studies investigating privacy protection in GNNs. In this work, we propose a learning framework that can provide local node privacy for users, while incurring low utility loss. We focus on a decentralized notion of Differential Privacy, namely Local Differential Privacy, and apply randomization mechanisms to perturb both feature and label data at the node level before they are collected by a server for model training. Specifically, we investigate the application of randomiza-

tion mechanisms in high-dimensional feature settings and propose an LDP protocol with strict privacy guarantees. Based on frequency estimation in statistical analysis of randomized data, we develop reconstruction methods to approximate features and labels from perturbed data. We also formulate this learning framework to utilize frequency estimates of graph clusters to supervise the training procedure at a sub-graph level. Extensive experiments on real-world and semi-synthetic datasets demonstrate the validity of our proposed model.

Karuna Bhaila, Wen Huang
University of Arkansas
kbhaila@uark.edu, wenhuang@uark.edu

Yongkai Wu
Clemson University
yongkaw@clemson.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

CP1

GENIUS: Subteam Replacement with Clustering-Based Graph Neural Networks

In this talk, we present GENIUS, an innovative graph neural network framework that revolutionizes subteam replacement in team social networks. We delve into the building blocks of GENIUS, highlighting its key features: (1) capturing team social network knowledge for subteam replacement by deploying team-level attention GNNs (TAGs) and self-supervised positive team contrasting training scheme, (2) generating unsupervised team social network member clusters to prune candidates for fast computation, and (3) incorporating a subteam recommender that selects new subteams of flexible sizes. We further address the critical limitations of existing graph kernel-based methods and provide detailed theoretical analysis on how GENIUS overcomes them. We demonstrate the efficacy of the proposed method in terms of (1) effectiveness: being able to select better subteam members that significantly increase the similarity between the new and original teams, based on both quantitative and qualitative assessments, and (2) efficiency: achieving more than 600× speed-up in average running time.

Chuxuan Hu, Qinghai Zhou, Hanghang Tong
University of Illinois at Urbana-Champaign
chuxuan3@illinois.edu, qinghai2@illinois.edu, htong@illinois.edu

CP1

Laplacian Score Benefit Adaptive Filter Selection for Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as powerful tools for representation learning on structured data. The graph convolutional filter (GCF) for aggregating neighbor information is shown to be the key factor that leads to GNNs' success. Various GCFs are designed but ***how to select the proper filter*** that can best benefit the data and the task remains an open problem. In this paper, we introduce the **Ada**ptive **F**ilter **S**election (AdaFS) framework that addresses two critical issues: (1) defining a criterion to establish a strong base filter set; and (2) adaptively selecting filters for a specific task, even when labeled data is limited, by employing Laplacian score regularization. We further connect this multiple GCF learning process and the well-developed multiple kernel learning problem to provide a solid rationale for filter selection. With experiments on 9 datasets, AdaFS gets the best average performance.

Yewen Wang
amazon
wyw10804@gmail.com

Shichang Zhang, Junghoo Cho
UCLA
shichang@cs.ucla.edu, cho@cs.ucla.edu

Yizhou Sun
University of California, Los Angeles
yzsun@cs.ucla.edu

CP1

Self-Similar Graph Neural Network for Hierarchical Graph Learning

Many real-world networks, such as graph-structured molecules or social networks, exhibit latent hierarchical structures at many different resolutions. Existing hierarchical graph neural networks (GNNs) mainly focus on modifying graph global pooling regions into partitioned clusters, while keeping the convolutional layers unchanged. However, these approaches may suffer from a loss of expressive power in learned representations due to the uncontrolled growth of the neighborhood, leading to a failure in capturing true hierarchies. Furthermore, many real-world hierarchical graphs possess an underlying fractal structure, which is crucial to unraveling the formation mechanism of networks. Unfortunately, existing hierarchical GNNs often overlook this important aspect of graph hierarchy. To tackle these challenges, this paper proposes a generic framework for hierarchical network representation learning. We propose the Self-Similar Graph Neural Network (SS-GNN), which leverages localized representations by excluding redundant nodes and edges. At each resolution of the coarsened map, SS-GNN extracts both intra- and inter-cluster embeddings to preserve the discriminative power of the model with a theoretical guarantee. To exploit the graph fractal structure, we introduce a novel module for measuring self-similarity between resolutions and a characterized objective function for automatic adjustment of model parameters.

Zheng Zhang, Liang Zhao
Emory University
zheng.zhang@emory.edu, liang.zhao@emory.edu

CP2

An Exemplars-Based Approach for Explainable Clustering: Complexity and Efficient Approximation Algorithms

Explainable AI (XAI) is an important area but remains relatively understudied for clustering. We propose an explainable-by-design clustering approach that not only finds clusters but also exemplars to explain each cluster. The use of exemplars for understanding is supported by the exemplar-based school of concept definition in psychology. We show that finding a small set of exemplars to explain even a single cluster is computationally intractable; hence, the overall problem is challenging. We develop an approximation algorithm that provides provable performance guarantees with respect to clustering quality as well as the number of exemplars used. This basic algorithm explains

all the instances in every cluster whilst another approximation algorithm uses a bounded number of exemplars to allow simpler explanations and provably covers a large fraction of all the instances. Experimental results show that our work is useful in domains involving difficult to understand deep embeddings of images and text.

Ian Davidson
University of California, Davis
davidson@cs.ucdavis.edu

MICHAEL Livanos
UC Davis
mjlivanos@ucdavis.edu

Antoine Gourru
University of Lyon 2
antoine.gourru@gmail.com

Peter Walker
Naval Medical Research Center
peter.b.walker.mil@mail.mil

JULIEN Velcin
University Lyon 2
julien.velcin@univ-lyon2.fr

S.S. Ravi
University of Virginia
sravi0@gmail.com

CP2

Pupae: Intuitive and Actionable Explanations for Time Series Anomalies

In recent years there has been significant progress in time series anomaly detection. However, after detecting an (perhaps tentative) anomaly, can we explain it? Such explanations would be useful to triage anomalies. For example, in an oil refinery, should we respond to an anomaly by dispatching a hydraulic engineer, or an intern to replace the battery on a sensor? There have been some parallel efforts to explain anomalies, however many proposed techniques produce explanations that are indirect, and often seem more complex than the anomaly they seek to explain. Our review of the literature/checklists/user-manuals used by frontline practitioners in various domains reveals an interesting near-universal commonality. Most practitioners discuss, explain and report anomalies in the following format: The anomaly would be like normal data A, if not for the corruption B. The reader will appreciate that is a type of counterfactual explanation. In this work we introduce a domain agnostic counterfactual explanation technique to produce explanations for time series anomalies. As we will show, our method can produce both visual and text-based explanations that are objectively correct, intuitive and in many circumstances, directly actionable.

Audrey E. Der
University of California, Riverside
ader003@ucr.edu

Chin-Chia Michael Yeh, Yan Zheng, Junpeng Wang, Zhongfang Zhuang, Liang Wang, Wei Zhang
Visa Research
miyeh@visa.com, yazheng@visa.com, junpenwa@visa.com, zzhuang@visa.com, liawang@visa.com, wzhan@visa.com

Eamonn Keogh
University of California, Riverside
eamonn@cs.ucr.edu

CP2

MISS: Multiclass Interpretable Scoring Systems

We present a novel, machine-learning approach for constructing Multiclass Interpretable Scoring Systems (MISS) - a fully data-driven methodology for generating single, sparse, and user-friendly scoring systems for multiclass classification problems. Scoring systems are commonly utilized as decision support models in healthcare, criminal justice, and other domains where interpretability of predictions and ease of use are crucial. Prior methods for data-driven scoring, such as SLIM (Supersparse Linear Integer Model), were limited to binary classification tasks and extensions to multiclass domains were primarily accomplished via one-versus-all-type techniques. The scores produced by our method can be easily transformed into class probabilities via the softmax function. We demonstrate techniques for dimensionality reduction and heuristics that enhance the training efficiency and decrease the optimality gap, a measure that can certify the optimality of the model. Our approach has been extensively evaluated on datasets from various domains, and the results indicate that it is competitive with other machine learning models in terms of classification performance metrics and provides well-calibrated class probabilities.

Michal K. Grzeszczyk
Sano Centre for Computational Medicine
m.grzeszczyk@sanoscience.org

Tomasz Trzcinski
Warsaw University of Technology
tomasz.trzcinski@pw.edu.pl

Arkadiusz Sitek
Massachusetts General Hospital
Harvard Medical School
asitek@mgh.harvard.edu

CP2

XGExplainer: Robust Evaluation-based Explanation for Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful tool for machine learning on graph datasets. Although GNNs can achieve high accuracy on several tasks, the explainability of the predictions remains a challenge. Existing works in GNN explainability aim to extract the key features contributing to the prediction made by a pre-trained model. For instance, perturbation-based methods focus on evaluating the potential explanatory subgraphs using the pre-trained model itself as an evaluator to determine whether the subgraphs capture the informative features. However, we show that this approach can fail to recognize informative subgraphs that become out-of-distribution relative to the training data. To address this limitation, we propose XGExplainer, a method designed to enhance the robustness of perturbation-based explainers. It achieves this by training a specialized GNN model, i.e., a robust evaluator model that aims at estimating the true graph distribution from randomized subgraphs of the input graph. Our method is geared towards enhancing the generalizability of existing explainability techniques by decoupling the pre-trained model from the evaluator, whose primary role is to gauge the informativeness of potential explanatory subgraphs. Our experiments show that XGExplainer con-

sistently improves the performance of local and global explainer techniques and outperforms state-of-the-art methods on all datasets for node and graph classification tasks.

Ryoji Kubo, Djellel Difallah
New York University Abu Dhabi
ryojikubo@nyu.edu, djellel@nyu.edu

CP2

STES: A Spatiotemporal Explanation Supervision Framework

Explanation supervision is a technique that guides a deep learning model to have correct attention during training and thus improve both the interpretability and predictability of the model. However, the exploration of explanation supervision methods for spatiotemporal prediction has been limited. In this paper, we propose a framework for explanation-supervised spatiotemporal forecasting which aims to explicitly incorporate human-annotated spatiotemporal explanations as supervision signals, achieved by introducing a unique objective that integrates human explanations for general spatiotemporal predictive models. Specifically, to extend the explanation supervision technique to spatiotemporal prediction, our framework addresses several inherent challenges associated with spatiotemporal data. Firstly, it tackles the difficulty of identifying and correcting the spatiotemporal reasoning process. Secondly, it addresses the challenge of handling the absence of human explanation annotation through interpolation techniques. Lastly, it handles the varying influence of different time points. To evaluate the effectiveness of our approach, we conducted extensive experiments on two real-world spatiotemporal datasets. The results demonstrate the superiority of our methods in improving the interpretability of explanations and the performance of the backbone deep neural network models, surpassing existing state-of-the-art explanation supervision methods.

Dazhou Yu
Emory University
dyu62@emory.edu

Binbin Chen
University of Pennsylvania
chenbb@seas.upenn.edu

Yun Li
Emory University
yli230@emory.edu

Suman Dhakal
George Mason University
sdhakal2@gmu.edu

Yifei Zhang, Zhenke Liu, Minxing Zhang
Emory University
yzh3443@emory.edu, zliu365@emory.edu,
mzha329@emory.edu

Jie Zhang
George Mason University
jzhang7@gmu.edu

Liang Zhao
Emory University
lzhao41@emory.edu

CP3

Differences Between Hard and Noisy-Labeled Samples: An Empirical Study

Extracting noisy or incorrectly labeled samples from a labeled dataset with hard/difficult samples is an important, and yet under-explored topic. Current methods focus in general either on noisy labels, or on hard samples, but not jointly on both. When the two types of data are both present, these methods often fail to distinguish them, which results in a decline in the overall performance of the model. We propose a systematic empirical study that provides insights into the similarities and more importantly the differences between hard and noisy samples. The method consists of designing synthetic datasets customized with different hardness and noisiness levels for different samples. These controlled experiments pave the way for the evaluation and development of methods that distinguish between hard and noisy samples. We evaluate how various data-partitioning methods are able to remove noisy samples while retaining hard samples. Our study highlights the advantages of using a metric in data partitioning that we propose and call *static centroid distance*. The resulting data-partitioning method outperforms others: It leads to a high test accuracy on models trained on the filtered datasets, as shown both for datasets with synthetic label noise and for datasets with *real-world* label noise. It also significantly outperforms other methods when employed within a semi-supervised learning framework.

Mahsa Forouzesh
EPFL
mahsa.forouzesh93@gmail.com

Patrick Thiran
Ecole Polytechnique Fédérale de Lausanne (EPFL)
patrick.thiran@epfl.ch

CP3

Yolo-Ocr: End-to-End Compound Figure Separation and Label Recognition of Images in Scientific Publications

Scientific publications, especially biomedical publications, contain a large number of compound figures, which are composed of multiple graphs, plots, and drawings. With the growing interest in data mining, scientific image understanding, and retrieval, compound figure separation and label recognition have become vital steps for various downstream tasks. However, existing studies are difficult to apply to increasingly complex scenarios, and they usually treat these two tasks separately. In this work, we propose a new model called YOLO-OCR to do compound figure separation and label recognition simultaneously. the YOLO-OCR realizes object detection, text detection, and text recognition altogether in a unified end-to-end trainable network. Benefiting from shared convolution features, the model has fewer computation costs and higher performance. To reduce annotation costing, we train the model on a synthesized compound figure dataset and then finetune the model in actual compound figure datasets based on an active learning strategy. The results show that the proposed method achieves a new state-of-the-art performance on the ImageCLEF 2016 dataset and our dataset. In addition, we developed an online system based on the proposed model to help researchers conveniently separate compound figures. The project is publicly available at

https://www.chatfigures.com/figure-separation

Shuo Meng, Xinshuo Liang, Shuai zhang, Leqi lei, Hanbai wu, Saira iqbal, Jinlian hu
City University of Hong Kong
shuomeng2-c@my.cityu.edu.hk, xinsliang2-c@my.cityu.edu.hk, zhangshuai.tz@my.cityu.edu.hk, leqilei2-c@my.cityu.edu.hk, hanbaiwu3-c@my.cityu.edu.hk, sairiqbal2-c@my.cityu.edu.hk, jinliahu@cityu.edu.hk

CP3

Multi-Polytope Machine for Classification

In numerous machine learning applications, there is a preference for classifiers characterized by a polyhedral description, as they are intended for utilization within optimization frameworks or for interpretability purposes. Here, we present a structured classifier designed to cater to downstream decision-making tasks. The classification method is achieved through the process of partitioning the feature domain into clusters and encompassing each cluster within a polytope. We employ a combined approach that integrates semi-supervised k-means with SVM. This unified optimization framework enables the simultaneous generation of multiple polytopes. The central concept involves using a k-means-based clustering method for the clustering step, followed by the utilization of SVM to construct hyperplanes between each pair of clusters. Notably, the clustering process for each class considers classification loss as well as information from other classes when allocating sample points to clusters. We propose an algorithm to solve the integer program. Our numerical experiments demonstrate the competitiveness of the proposed method across a wide spectrum of datasets, exhibiting its efficacy in comparison to existing hyperplane-based classifiers and nonlinear classifiers.

Dzung Phan, Lam Nguyen, Jayant Kalagnanam, Chandra Reddy
IBM T.J. Watson Research Center
phandu@us.ibm.com, lamnguyen.mltd@ibm.com, jayant@us.ibm.com, creddy@us.ibm.com

CP3

Feature-Engineered Random Forests

In this paper, we present a method for constructing a feature-engineered random forest by transforming the features of a given data set using a set of diverse and randomized transforms. The transformed features are then used for creating splits at each node of a random forest. In particular, we use sum-product features because of their strong expressive power. This type of on-the-fly feature engineering has significant advantages over traditional random forests because it adds to the diversity of the splits. Such a diversity further helps in variance reduction; over and above the variance reduction ability offered by traditional random forests. We show the advantages of the proposed approach over traditional random forests and other well-established baselines using extensive experimental evaluation.

Saket Sathe
IBM T. J. Watson Research Center
saksathe@amazon.com

CP3

CoLafier: Collaborative Noisy Label Purifier With Local Intrinsic Dimensionality Guidance

Deep neural networks have significantly advanced various machine learning tasks, yet their performance often suffers from noisy labels in real-world data. To tackle this, we present CoLafier, an innovative approach utilizing Local Intrinsic Dimensionality (LID) to enhance learning with noisy labels. CoLafier consists of two subnets: LID-dis and LID-gen. LID-dis is a specialized classifier. Trained with our uniquely crafted scheme, LID-dis uses a sample's features and label to create an enhanced internal representation, which helps in distinguishing correct from incorrect labels in diverse noise scenarios. In contrast to LID-dis, LID-gen, functioning as a regular classifier, operates solely on the sample's features. During training, CoLafier utilizes two augmented views per instance to feed both subnets. CoLafier considers the LID scores from the two views as produced by LID-dis to assign weights in an adapted loss function for both subnets. Concurrently, LID-gen, serving as a classifier, suggests pseudo-labels. LID-dis then processes these pseudo-labels along with two views to derive LID scores. Finally, these LID scores along with the differences in predictions from the two subnets guide the label update decisions. This dual-view and dual-subnet approach enhances the overall reliability of the framework. CoLafier has shown enhanced prediction accuracy, particularly in scenarios with severe label noise, outperforming existing methods.

Dongyu Zhang
Worcester Polytechnic Institute
dzhang5@wpi.edu

Ruofan Hu
Worcester Polytechnic Institute
Data Science Program
rhu@wpi.edu

Elke Rundensteiner
Worcester Polytechnic Institute
rundenst@wpi.edu

CP4

Training Sparse Graph Neural Networks Via Pruning and Sprouting

With the emergence of large-scale graphs and deeper graph neural networks (GNNs), sparsifying GNNs including graph connections and model parameters has attracted a lot of attention. However, most existing GNN sparsification methods apply traditional neural network pruning techniques to sparsify graphs in an iterative cycle (train-then-sparsify), which not only incurs high training costs but also limits model performance. In this paper, we propose a novel Pruning and Sprouting framework for GNN (PSGNN) that not only enhances the efficiency of inference, but also boosts the performance of GNN trained on a core subgraph beyond the original graph. Based on during-training pruning, our framework gradually sparsifies the graph connections and model weights simultaneously. More specifically, PSGNN removes edges in the original graph according to the predicted label similarity between nodes from a global view. Additionally, with our graph sprouting strategy, PSGNN can generate new edges to include important yet missing topological and feature in-

formation in the original graph, while maintaining the sparsity of the graph. Extensive experiments on node classification task across different GNN architectures and graph datasets demonstrate that our proposed PSGNN method improves the performance over existing methods while saving training and inference costs.

Xueqi Ma
University of Melbourne
xueqim@student.unimelb.edu.au

Xingjun Ma
Fudan University
xingjunma@fudan.edu.cn

Sarah M. Erfani
The University of Melbourne
The University of Melbourne
sarah.erfani@unimelb.edu.au

James Bailey
The University of Melbourne
baileyj@unimelb.edu.au

CP4

Prompt Based Tri-Channel Graph Convolution Neural Network for Aspect Sentiment Triplet Extraction

Aspect Sentiment Triplet Extraction (ASTE) is an emerging task to extract a given sentence's triplets, which consist of aspects, opinions, and sentiments. Recent studies tend to address this task with a table-filling paradigm, wherein word relations are encoded in a two-dimensional table, and the process involves clarifying all the individual cells to extract triples. However, these studies ignore the deep interaction between neighbor cells, which we find quite helpful for accurate extraction. To this end, we propose a novel model for the ASTE task, called Prompt-based Tri Channel Graph Convolution Neural Network (PT-GCN), which converts the relation table into a graph to explore more comprehensive relational information. Specifically, we treat the original table cells as nodes and utilize a prompt attention score computation module to determine the edges' weights. This enables us to construct a target-aware grid-like graph to enhance the overall extraction process. After that, a triple-channel convolution module is conducted to extract precise sentiment knowledge. Extensive experiments on the benchmark datasets show that our model achieves state-of-the-art performance. The code is available at https://github.com/KunPunCN/PT-GCN.

Kun Peng
Institute of Information Engineering, Chinese Academy of Sci
Institute of Information Engineering, Chinese Academy of Sci
pengkun@iie.ac.cn

Lei Jiang
Chinese Academy of Sciences
jianglei@iie.ac.cn

Hao Peng
Beihang University, China
penghao@buaa.edu.cn

Rui Liu
Chinese Academy of Sciences
liurui3221@iie.ac.cn

Zhengtao Yu
Kunming University of Science and Technology
ztyu@hotmail.com

Jiaqian Ren
Chinese Academy of Sciences
renjiaqian@iie.ac.cn

Zhifeng Hao
College of Science, University of Shantou
zfhao@stu.edu.cn

CP4

Tensorized Hypergraph Neural Networks

Hypergraph neural networks (HGNN) have recently become attractive and received significant attention due to their excellent performance in various domains. However, most existing HGNNs rely on first-order approximations of hypergraph connectivity patterns, which ignores important high-order information. To address this issue, we propose a novel adjacency-tensor-based Tensorized Hypergraph Neural Network (THNN). THNN is a faithful hypergraph modeling framework through high-order outer product feature message passing and is a natural tensor extension of the adjacency-matrix-based graph neural networks. The proposed THNN is equivalent to a high-order polynomial regression scheme, which enables THNN with the ability to efficiently extract high-order information from uniform hypergraphs. Moreover, in consideration of the exponential complexity of directly processing high-order outer product features, we propose using a partially symmetric CP decomposition approach to reduce model complexity to a linear degree. Additionally, we propose two simple yet effective extensions of our method for non-uniform hypergraphs commonly found in real-world applications. Results from experiments on two widely used hypergraph datasets for 3-D visual object classification show the model's promising performance.

Maolin Wang, Yaoming Zhen
City University of Hong Kong
morin.w98@gmail.com, zhen8-c@my.cityu.edu.hk

Yu Pan
Harbin Institute of Technology Shenzhen
iperryuu@gmail.com

Yao Zhao, Chenyi Zhuang
AntGroup
nanxiao.zy@antgroup.com, chenyi.zcy@antgroup.com

Zenglin Xu
Harbin Institute of Technology Shenzhen
zenglin@gmail.com

Ruocheng Guo
ByteDance Research
rguo.asu@gmail.com

Xiangyu Zhao
City University of Hong Kong

xianzhao@cityu.edu.hk

CP4

Non-Euclidean Spatial Graph Neural Network

Spatial networks are networks whose graph topology is constrained by their embedded spatial space. Understanding the coupled spatial-graph properties is crucial for extracting powerful representations from spatial networks. Therefore, merely combining individual spatial and network representations cannot reveal the underlying interaction mechanism of spatial networks. Besides, existing spatial network representation learning methods can only consider networks embedded in Euclidean space, and can not well exploit the rich geometric information carried by irregular and non-uniform non-Euclidean space. In order to address this issue, in this paper we propose a novel generic framework to learn the representation of spatial networks that are embedded in non-Euclidean manifold space. Specifically, a novel message-passing-based neural network is proposed to combine graph topology and spatial geometry, where spatial geometry is extracted as messages on the edges. We theoretically guarantee that the learned representations are provably invariant to important symmetries such as rotation or translation, and simultaneously maintain sufficient ability in distinguishing different geometric structures. The strength of our proposed method is demonstrated through extensive experiments on both synthetic and real-world datasets.

Zheng Zhang, Sirui Li
Emory University
zheng.zhang@emory.edu, sirui.li@emory.edu

Jingcheng Zhou
The University of Manchester
j.zhou-74@sms.ed.ac.uk

Junxiang Wang
Emory University
junxiang.wang@alumni.emory.edu

Abhinav Angirekula
Thomas Jefferson High School, Virginia, United States
abhinava2560@gmail.com

Allen Zhang, Liang Zhao
Emory University
allen.zhang@emory.edu, liang.zhao@emory.edu

CP5

On the Number of Iterations of the DBA Algorithm

The DTW Barycenter Averaging (DBA) algorithm is a widely used algorithm for estimating the mean of a given set of point sequences. In this context, the mean is defined as a point sequence that minimises the sum of dynamic time warping distances (DTW). The algorithm is similar to the k-means algorithm in the sense that it alternately repeats two steps: (1) computing an optimal assignment to the points of the current mean, and (2) computing an optimal mean under the current assignment. The popularity of DBA can be attributed to the fact that it works well in practice, despite any theoretical guarantees to be known. In our paper, we aim to initiate a theoretical study of the number of iterations that DBA performs until convergence. We assume the algorithm is given n sequences of m points in $\mathbb{R}^d$ and a parameter k that specifies the length of the mean sequence to be computed. We show that, in contrast to its fast running time in practice, the number of iterations can be exponential in k in the worst case - even if the number of input sequences is $n = 2$. We complement these findings with experiments on real-world data that suggest this worst-case behaviour is likely degenerate. To better understand the performance of the algorithm on non-degenerate input, we study DBA in the model of smoothed analysis, upper-bounding the expected number of iterations in the worst case under random perturbations of the input.

Frederik Brüning
University of Bonn
bruening@uni-bonn.de

Anne Driemel
Hausdorff Center for Mathematics
University of Bonn
driemel@cs.uni-bonn.de

Alperen Ergür
The University of Texas at San Antonio
alperen.ergur@utsa.edu

Heiko Röglin
Department of Computer Science
University of Bonn, Germany
roeglin@cs.uni-bonn.de

CP5

Robust Sparse Online Learning for Data Streams with Streaming Features

Sparse online learning has received extensive attention during the past few years. Most of existing algorithms that utilize ℓ_1-norm regularization or ℓ_1-ball projection assume that the feature space is fixed or changes by following explicit constraints. However, this assumption does not always hold in many real applications. Motivated by this observation, we propose a new online learning algorithm tailored for data streams described by open feature spaces, where new features can be occurred, and old features may be vanished over various time spans. Our algorithm named RSOL provides a strategy to adapt quickly to such feature dynamics by encouraging sparse model representation with an ℓ_1- and ℓ_2-mixed regularizer. We leverage the proximal operator of the $\ell_{1,2}$-mixed norm and show that our RSOL algorithm enjoys a closed-form solution at each iteration. A sub-linear regret bound of our proposed algorithm is guaranteed with a solid theoretical analysis. Empirical results benchmarked on nine streaming datasets validate the effectiveness of the proposed RSOL method over three state-of-the-art algorithms. **Keywords**: online learning, sparse learning, streaming feature selection, open feature spaces, $\ell_{1,2}$ mixed norm

Zhong Chen
Southern Illinois University
zhong.chen@cs.siu.edu

Yi He
Old Dominion University
yihe@cs.odu.edu

Di Wu
Southwest University
wudi.cigit@gmail.com

Huixin Zhan
Cedars-Sinai Medical Center
huixin.zhan@cshs.org

Victor Sheng
Texas Tech University
victor.sheng@ttu.edu

Kun Zhang
Xavier University of Louisiana
kzhang@xula.edu

CP5

Utility-Oriented String Mining

A string is often provided with numerical scores (*utilities*) which quantify the importance, interest, profit, or risk of the letters occurring at every position of the string. Motivated by the abundance of strings with utilities, we introduce Utility-oriented String Mining (USM), a natural generalization of the classic frequent substring mining problem. Given a string S of length n and a threshold $\mathcal{V}$, USM asks for every string R whose utility $U(R)$ is at least $\mathcal{V}$, where U is a function that maps R to a utility score based on the utilities of all letters of every occurrence of R in S. Our work makes the following contributions: 1) We identify a class of utility functions for which USM admits an $\mathcal{O}(n^2)$-time algorithm. 2) We prove that no listing algorithm solves the USM problem in subquadratic time for every utility function. 3) We propose an $\mathcal{O}(n \log n)$-time algorithm that solves USM for a class of monotone functions. 4) We design another $\mathcal{O}(n \log n)$-time algorithm for the same problem that is comparable in runtime but offers drastic space savings in practice when, in addition, a lower bound on the length of the output strings is provided as input. 5) We demonstrate experimentally using publicly available, billion-letter datasets that our algorithms are many times more efficient, in terms of runtime and/or space, compared to an Apriori-like baseline which employs advanced string processing tools.

Veronica Guerrini
University of Pisa
veronica.guerrini@unipi.it

Giulia Bernardini
University of Trieste
giulia.bernardini@units.it

Huiping Chen
University of Birmingham
h.chen.13@bham.ac.uk

Alessio Conte, Roberto Grossi
Università di Pisa
alessio.conte@unipi.it, roberto.grossi@unipi.it

Grigorios Loukides
King's College London
gloukides@acm.org

Nadia Pisanti
University of Pisa
nadia.pisanti@unipi.it

Solon P. Pissis
Centrum Wiskunde Informatica
solon.pissis@cwi.nl

CP5

EsaCL: An Efficient Continual Learning Algorithm

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Weijieying Ren
The Pennsylvania State University
wjr5337@psu.edu

CP6

Towards Tuning-Free Minimum-Volume Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) is a versatile and powerful tool for discovering latent structures in data matrices, with many variations proposed in the literature. Recently, Leplat et al. (2019) introduced a minimum-volume NMF for the identifiable recovery of rank-deficient matrices in the presence of noise. The performance of their formulation, however, requires the selection of a tuning parameter whose optimal value depends on the unknown noise level. In this work, we propose an alternative formulation of minimum-volume NMF inspired by the square-root lasso and its tuning-free properties. Our formulation also requires the selection of a tuning parameter, but its optimal value does not depend on the noise level. To fit our NMF model, we propose a majorization-minimization (MM) algorithm that comes with global convergence guarantees. We show empirically that the optimal choice of our tuning parameter is insensitive to the noise level in the data.

Duc Toan Nguyen
Texas Christian University
duc.toan.nguyen@tcu.edu

Eric Chi
Rice University
echi@rice.edu

CP6

Semi-Supervised Clustering Via Structural Entropy with Different Constraints

Semi-supervised clustering techniques have emerged as valuable tools for leveraging prior information in the form of constraints to improve the quality of clustering outcomes. Despite the proliferation of such methods, the ability to seamlessly integrate various types of constraints remains limited. While structural entropy has proven to be a powerful clustering approach with wide-ranging applications, it has lacked a variant capable of accommodating these constraints. In this work, we present Semi-supervised clustering via Structural Entropy (SSE), a novel method that can incorporate different types of constraints from diverse sources to perform both partitioning and hierarchical clustering. Specifically, we formulate a uniform view for the commonly used pairwise and label constraints for both types of clustering. Then, we design objectives that incorporate these constraints into structural entropy and develop tailored algorithms for their optimization. We evaluate SSE on nine clustering datasets and compare it with eleven semi-supervised partitioning and hierarchical clustering methods. Experimental results demonstrate the su-

periority of SSE on clustering accuracy with different types of constraints. Additionally, the functionality of SSE for biological data analysis is demonstrated by cell clustering experiments conducted on four single-cell RNAseq datasets.

Guangjie Zeng
Beihang University
zengguangjie@buaa.edu.cn

Hao Peng
Beihang University, China
penghao@buaa.edu.cn

Angsheng Li
Beihang University
angsheng@buaa.edu.cn

Zhiwei Liu
Salesforce AI Research
zhiweiliu@salesforce.com

Runze Yang
Beihang University
yangrunze@buaa.edu.cn

Chunyang Liu
Didi Chuxing
liuchunyang@didiglobal.com

Lifang He
Lehigh University
lih319@lehigh.edu

CP6

Decentralized Stochastic Compositional Gradient Descent for AUPRC Maximization

In this paper, we consider the large-scale Area Under the Precision-Recall Curve (AUPRC) maximization problem for the imbalanced data classification task. Existing optimization methods for AUPRC maximization only focus on the single-machine setting, which are not applicable to the distributed data. To address this problem, we propose a novel decentralized stochastic compositional gradient descent method for large-scale AUPRC maximization. Our theoretical analysis shows that it can achieve a better sample complexity $\mathcal{O}(1/\epsilon^4)$ than $\mathcal{O}(1/\epsilon^6)$ of existing decentralized methods, but has the same communication complexity $\mathcal{O}(1/\epsilon^4)$. To further reduce the communication cost, we developed a novel communication-efficient decentralized stochastic compositional gradient descent method, whose communication complexity is improved to $\mathcal{O}(1/\epsilon^{4-4\alpha})$ (where $\alpha \in (0, 1/4)$). To the best of our knowledge, this is the first work achieving such favorable sample and communication complexities. Finally, we conduct extensive experiments for imbalanced data classification and the empirical results confirm the superior performance of our proposed methods.

Hongchang Gao, Yubin Duan, Yihan Zhang, Jie Wu
Temple University
hongchanggao@gmail.com, yubin.duan@temple.edu, yihan.zhang0002@temple.edu, jiewu@temple.edu

CP7

Active Learning for Graphs with Noisy Structures

Graph Neural Networks (GNNs) have seen significant success in tasks such as node classification, largely contingent upon the availability of sufficient labeled nodes. Yet, the excessive cost of labeling large-scale graphs led to a focus on active learning on graphs, which aims for effective data selection to maximize downstream model performance. Notably, most existing methods assume reliable graph topology, while real-world scenarios often present noisy graphs. Given this, designing a successful active learning framework for noisy graphs is highly needed but challenging, as selecting data for labeling and obtaining a clean graph are two tasks naturally interdependent: selecting high-quality data requires clean graph structure while cleaning noisy graph structure requires sufficient labeled data. Considering the complexity mentioned above, we propose an active learning framework, GALClean, which has been specifically designed to adopt an iterative approach for conducting both data selection and graph purification simultaneously with best information learned from the prior iteration. Importantly, we summarize GALClean as an instance of the Expectation-Maximization algorithm, which provides a theoretical understanding of its design and mechanisms. This theory naturally leads to an enhanced version, GALClean+. Extensive experiments have demonstrated the effectiveness and robustness of our proposed method across various types and levels of noisy graphs.

Hongliang Chi
Rensselaer Polytechnic Institute
chih3@rpi.edu

CP7

Treatment-Aware Hyperbolic Representation Learning for Causal Effect Estimation with Social Networks

Estimating the individual treatment effect (ITE) from observational data is a crucial research topic that holds significant value across multiple domains. How to identify hidden confounders poses a key challenge in ITE estimation. Recent studies have incorporated the structural information of social networks to tackle this challenge, achieving notable advancements. However, these methods utilize graph neural networks to learn the representation of hidden confounders in Euclidean space, disregarding two critical issues: (1) the social networks often exhibit a scale-free structure, while Euclidean embeddings suffer from high distortion when used to embed such graphs, and (2) each ego-centric network within a social network manifests a treatment-related characteristic, implying significant patterns of hidden confounders. To address these issues, we propose a novel method called Treatment-Aware Hyperbolic Representation Learning (TAHyper). Firstly, TAHyper employs the hyperbolic space to encode the social networks, thereby effectively reducing the distortion of confounder representation caused by Euclidean embeddings. Secondly, we design a treatment-aware relationship identification module that enhances the representation of hidden confounders by identifying whether an individual and her neighbors receive the same treatment. Extensive experiments on two benchmark datasets are conducted to demonstrate the superiority of our method.

Ziqiang Cui

City University of Hong Kong
ziqiang.cui@my.cityu.edu.hk

Xing Tang, Yang Qiao
Tencent
shawntang@tencent.com, sunnyqiao@tencent.com

Bowei He
City University of Hong Kong
boweihe2- c@my.cityu.edu.hk

Liang Chen, Xiuqiang He
Tencent
leocchen@tencent.com, xiuqianghe@tencent.com

Chen Ma
City University of Hong Kong
chenma@cityu.edu.hk

CP7

H^2abm: Heterogeneous Agent-Based Model on Hypergraphs to Capture Group Interactions

Heterogeneous agent-based models (HABMs) can simulate the dynamics of multiple types of entities and their interactions on contact networks. In recent years, they have gathered great interest and are widely applied in multiple fields, such as personalized recommendations, publication ranking, and epidemic modeling. Nevertheless, conventional HABMs on graphs can only capture pair-wise interactions between agents but fail to capture the more complex dynamics of group interactions (e.g., multiple people in the same location simultaneously), consequently leading to suboptimal performance. To address this, we propose using hypergraphs to capture such group interactions better and extend the current graph-based HABMs to hypergraphs. Specifically, we use MRSA (Methicillin-resistant Staphylococcus aureus, a kind of infectious disease acquired by patients during treatment at healthcare facilities) spread in the University of Virginia hospital as an example to showcase how we extend an existing graph-based HABM, Graph-HeterSIS, to a hypergraph-based HABM (H^2ABM), Hypergraph-HeterSIS. We show how the hypergraphs can capture the structural difference between contacts before and during the first wave of COVID-19 outbreak in Virginia better than graphs. Our experiments show that H^2ABM better captures the underlying group interactions and better fits and forecasts MRSA cases.

Jiaming Cui
Georgia Institute of technology
jiamingcui1997@gatech.edu

Vivek Anand
College of Computing, Georgia Institute of Technology
vivekanand@gatech.edu

Jack Heavey
Department of Computer Science, University of Virginia
jch7jm@virginia.edu

Anil Vullikanti
University of Virginia
vsakumar@virginia.edu

B. Aditya Prakash
College of Computing, Georgia Institute of Technology
adityap@cc.gatech.edu

CP7

Helper Recommendation with Seniority Control in Online Health Community

Online health communities (OHCs) provide an essential platform for patients with similar health conditions to share experiences and offer moral support. However, many time-sensitive questions from patients often remain unanswered due to the multitude of threads and the random nature of patient visits in OHCs. Traditional recommendation systems solely based on similarity for recommendations cannot be directly applied in OHCs. They tend to overlook the influence of patients' dynamically changing features (e.g., health stages), affecting their ability to provide meaningful responses to questions. To address this, we propose a novel recommender system scenario designed for OHCs, which differs from traditional recommender systems in several ways. Firstly, it's challenging to model the social support factors that form helper-seeker links in OHCs. Secondly, the impact of patients' historical activities is complex to quantify. Lastly, ensuring recommended helpers have the requisite expertise is crucial. To overcome these challenges, we develop a Monotonically regularIzed diseNTangled Variational Autoencoders (MINT) model. This model formulates interactions between seekers and helpers as a dynamic graph, using encoded historical activities as node features. We also introduce a graph-based disentangle VAE to capture patient features and a monotonic regularizer to ensure the logical pairing of seekers and helpers. Our extensive experiments show the effectiveness of our approach.

Junruo Gao
Chinatelecom Cloud Technology Co., Ltd
junruo.gao@gmail.com

Chen Ling
Emory University
chen.ling@emory.edu

Carl Yang
Department of Computer Science
Emory University
j.carlyang@emory.edu

Liang Zhao
Emory University
liang.zhao@emory.edu

CP7

Graph Summarization for Preserving Spectral Characteristics

How does the graph change if we summarize it by merging nodes? How can we summarize the graph while preserving its spectral characteristics? Graph summarization aims to present a graph in a compact summary graph form while keeping its important structural information. Existing methods primarily focus on preserving the adjacency matrix. In contrast, spectral graph theory provides a powerful tool to describe the characteristics of a graph. In this paper, we propose a novel graph summarization method that preserves the spectral characteristics, including spectral moments and heat traces. We analyze the change of the spectral characteristics after summarization and design a simple yet effective summarization method based on ag-

glomerative clustering. Our approach is extensively evaluated on real-world datasets. The experimental results show that our method excels in preserving the spectral characteristics and obtains better performance on the subsequent graph classification task.

CP8

Pattern-Based Time Series Semantic Segmentation with Gradual State Transitions

Time series semantic segmentation is the task of extracting time intervals from the time series data that share a similar meaning within the application domain in an unsupervised manner. State-of-the-art algorithms typically treat this problem as change point detection, resulting in discrete state transitions. However, in real-world applications, states often transition gradually. This leads to a novel, more challenging variation of the traditional time series segmentation task, for which we present PaTSS, a novel, domain-agnostic algorithm to uncover those gradual state transitions. PaTSS learns a distribution over the semantic segments based on an embedding space derived from mined sequential patterns. An extensive experimental evaluation on 107 benchmark time series shows that PaTSS is capable of detecting gradual state transitions, a task current methods are unable to perform.

Louis Carpentier, Wannes Meert
KU Leuven
louis.carpentier@kuleuven.be, wannes.meert@kuleuven.be

Len Feremans
University of Antwerp
UA
len.feremans@uantwerpen.be

Mathias Verbeke
KU Leuven
mathias.verbeke@kuleuven.be

CP8

Towards Entity-Aware Conditional Variational Inference for Heterogeneous Time-Series Prediction: An Application to Hydrology

Many environmental systems (e.g., hydrology basins) can be modeled as entity whose response (e.g., streamflow) depends on drivers (e.g., weather) conditioned on their characteristics (e.g., soil properties). We introduce Entity-aware Conditional Variational Inference (EA-CVI), a novel probabilistic inverse modeling approach, to deduce entity characteristics from observed driver-response data. EA-CVI infers probabilistic latent representations that can accurately predict response for diverse entities, particularly in out-of-sample few-shot settings. EA-CVI's latent embeddings encapsulate diverse entity characteristics within compact, low-dimensional representations. EA-CVI proficiently identifies dominant modes of variation in responses and offers the opportunity to infer a physical interpretation of the underlying attributes that shape these responses. EA-CVI can also generate new data samples by sampling from the learned distribution, making it useful in zero-shot scenarios. EA-CVI addresses the need for uncertainty estimation, particularly during extreme events, rendering it essential for data-driven decision-making in real-world applications. Extensive evaluations on a renowned hydrology benchmark dataset, CAMELS-GB, validate EA-CVI's abilities.

Rahul Ghosh, Arvind Renganathan
University of Minnesota
ghosh128@umn.edu, renga016@umn.edu

Wallace Mcaliley
USGS
wmcaliley@usgs.gov

Michael Steinbach
University of Minnesota
stei0062@umn.edu

Christopher Duffy
Penn State University
cxd111@psu.edu

Vipin Kumar
University of Minnesota
kumar001@umn.edu

CP8

EBV: Electronic Bee-Veterinarian for Principled Mining and Forecasting of Honeybee Time Series

Honeybees are vital for pollination and food production. Among many factors, extreme temperature (e.g., due to climate change) is particularly dangerous for bee health. Anticipating such extremities would allow beekeepers to take early preventive action. Thus, given sensor (temperature) time series data from beehives, how can we find patterns and do forecasting? Forecasting is crucial as it helps spot unexpected behavior and thus issue warnings to the beekeepers. In that case, what are the right models for forecasting? ARIMA, RNNs, or something else? We propose the EBV (Electronic Bee-Veterinarian) method, which has the following desirable properties:(i) principled: it is based on a) diffusion equations from physics and b) control theory for feedback-loop controllers; (ii) effective: it works well on multiple, real-world time sequences, (iii) explainable: it needs only a handful of parameters (e.g., bee strength) that beekeepers can easily understand and trust, and (iv) scalable: it performs linearly in time. We applied our method to multiple real-world time sequences and found that it yields accurate forecasting (up to 49% improvement in RMSE compared to baselines) and segmentation. Specifically, discontinuities detected by EBV mostly coincide with domain expert's opinions, showcasing our approach's potential and practical feasibility. Moreover, EBV is scalable and fast, taking about 20 minutes on a stock laptop for reconstructing two months of sensor data.

Mst Shamima Hossain
University of California, Riverside
mhoss037@ucr.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Boris Baer, Hyoseung Kim, Vassilis Tsotras
University of California, Riverside
boris.baer@ucr.edu, hyoseung@ucr.edu, vtso-

tras@ucr.edu

CP8

Early Multiple Temporal Patterns Based Event Prediction in Heterogeneous Multivariate Temporal Data

Predicting an event of interest based on heterogeneous multivariate temporal data is challenging but desirable as it allows the utilization of all types of temporal variables. In various domains, symbolic time intervals (STIs) can be used to represent real-life events that vary in duration, such as the period a traffic light remains green, or the time a patient undergoes treatment or is on medication. Further, heterogeneous multivariate temporal data may be composed of STIs along with event-driven or continuous temporal variables, such as traffic collisions or blood test values. Temporal abstraction can be used to uniformly represent heterogeneous multivariate temporal variables with STIs, from which frequent time intervals related patterns (TIRPs) can be discovered. We extend earlier work on continuous completion prediction of a single TIRP that ends with an event of interest, introducing a continuous prediction method based on *multiple* different instances of multiple TIRPs that end with the event of interest, for which we propose and evaluate several weighted aggregation functions. The proposed method overall performed better on real-life, medical, and non-medical datasets, than the use of a single TIRP, and in comparison to the baseline models (XGBoost, ResNet, LSTM-FCN, and ROCKET).

Nevo Itzhak
Ben Gurion University
nevoit@post.bgu.ac.il

Szymon Jaroszewicz
Institute of Computer Science, Polish Academy of Sciences
szymon.jaroszewicz@ipipan.waw.pl

Robert Moskovitch
Ben Gurion University
robertmo@bgu.ac.il

CP8

Time-Transformer: Integrating Local and Global Features for Better Time Series Generation

Generating time series data is a promising approach to address data deficiency problems. However, it is also challenging due to the complex temporal properties of time series data, including local correlations as well as global dependencies. Most existing generative models have failed to effectively learn both the local and global properties of time series data. To address this open problem, we propose a novel time series generative model named 'Time-Transformer AAE', which consists of an adversarial autoencoder (AAE) and a newly designed architecture named 'Time-Transformer' within the decoder. The Time-Transformer first simultaneously learns local and global features in a layer-wise parallel design, combining the abilities of Temporal Convolutional Networks and Transformer in extracting local features and global dependencies respectively. Second, a bidirectional cross attention is proposed to provide complementary guidance across the two branches and achieve proper fusion between local and global features. Experimental results demonstrate that our model can outperform existing state-of-the-art models in 5 out of 6 datasets, specifically on those with data containing both global and local properties. Furthermore, we highlight our model's ability to handle this kind of data via an artificial dataset. Finally, we show how our model performs when applied to a real-world problem: data augmentation to support learning with small datasets and imbalanced datasets.

Yuansan Liu, Sudanthi Wijewickrema
University of Melbourne
yuansanl@student.unimelb.edu.au,
sudanthi.wijewickrema@unimelb.edu.au

Ang Li
Aviation University of Air Force
angli.cs@outlook.com

Christofer Bester
University of Melbourne
christofer.bester@unimelb.edu.au

Stephen O'Leary
University of Melbourne, Melbourne
sjoleary@unimelb.edu.au

James Bailey
The University of Melbourne
baileyj@unimelb.edu.au

CP8

Message Propagation Through Time: An Algorithm for Sequence Dependency Retention in Time Series Modeling

Time series modeling, a crucial area in science, often encounters challenges when training Machine Learning (ML) models like Recurrent Neural Networks (RNNs) using the conventional mini-batch training strategy that assumes independent and identically distributed (IID) samples and initializes RNNs with zero hidden states. The IID assumption ignores temporal dependencies among samples, resulting in poor performance. This paper proposes the Message Propagation Through Time (MPTT) algorithm to effectively incorporate long temporal dependencies while preserving faster training times relative to the stateful algorithms. MPTT utilizes two memory modules to asynchronously manage initial hidden states for RNNs, fostering seamless information exchange between samples and allowing diverse mini-batches throughout epochs. MPTT further implements three policies to filter outdated and preserve essential information in the hidden states to generate informative initial hidden states for RNNs, facilitating robust training. Experimental results demonstrate that MPTT outperforms seven strategies on four climate datasets with varying levels of temporal dependencies.

Shaoming Xu, Ankush Khandelwal, Arvind Renganathan, Vipin Kumar
University of Minnesota
xu000114@umn.edu, khand035@umn.edu, renga016@umn.edu, kumar001@umn.edu

CP9

Dual-Stage Flows-Based Generative Modeling for Traceable Urban Planning

Urban planning, which aims to design feasible land-use configurations for target areas, has become increasingly essential due to the high-speed urbanization process in the

modern era. However, the traditional urban planning conducted by human designers can be a complex and onerous task. Thanks to the advancement of deep learning algorithms, researchers have started to develop automated planning techniques. While these models have exhibited promising results, they still grapple with a couple of unresolved limitations: 1) Ignoring the relationship between urban functional zones and configurations and failing to capture the relationship among different functional zones. 2) Less interpretable and stable generation process. To overcome these limitations, we propose a novel generative framework based on normalizing flows, namely Dual-stage Urban Flows (DSUF) framework. Specifically, the first stage is to utilize zone-level urban planning flows to generate urban functional zones based on given surrounding contexts and human guidance. Then we employ an Information Fusion Module to capture the relationship among functional zones and fuse the information of different aspects. The second stage is to use configuration-level urban planning flows to obtain land-use configurations derived from fused information. We design several experiments to indicate that our framework can outperform for the urban planning task.

Xuanming Hu
Arizona State University
solomonhxm@gmail.com

Wei Fan
University of Oxford
weifan.oxford@gmail.com

Dongjie Wang
University of Central Florida
dongjie.wang@ucf.edu

Pengyang Wang
University of Macau
pywang@um.edu.mo

Yong Li
Tsinghua University
liyong07@tsinghua.edu.cn

Yanjie Fu
Arizona State University
yanjie.fu@asu.edu

CP9

A Novel Hybrid Graph Learning Method for Inbound Parcel Volume Forecasting in Logistics System

Inbound parcel volume forecasting problem (IPVFP) plays an important role in the logistics system as it can facilitate various downstream applications. Despite the fact that a number of time series forecasting techniques have been developed, existing approaches fail to explicitly consider intrinsic characteristics of the logistics system, *e.g.*, parcel transport patterns, operation patterns, and their spatial-temporal dependencies. To this end, we propose a novel hybrid inbound parcel volume forecasting model to analyze the logistic spatial-temporal graph that is constructed based on logistics data and the physical location of logistics stations. The graph includes engineered features such as the transition matrix and modified Dynamic Time Warping (DTW) distance matrix, which accurately depicts the parcel transfer patterns within the system. In addition, it incorporates a dedicated attention mechanism that introduces a novel bit-embedding representation method for integer tokens, enabling to capture of dynamic correlations among different timestamps. Finally, a collaborative module comprising dilated convolution layers and Gated Recurrent Units (GRU) is integrated to capture long-term dependencies. Extensive experiments on real-world data evaluate the effectiveness of the proposed graph and model, demonstrating its superiority over 16 other advanced baseline models. We release our code and data at https://github.com/YelsAlyssa/IPVFP.

Lisha Ye, Jianfeng Zhou, Zhe Yin, Kunpeng Han, Haoyuan Hu
cainiao network
yelisha.yls@cainiao.com, zhoujianfeng.zjf@cainiao.com, dimon.yz@cainiao.com, kunpeng.hkp@cainiao.com, haoyuan.huhy@cainiao.com

Dongjin Song
University of Connecticut
dongjin.song@uconn.edu

CP9

Automated Fusion of Multimodal Electronic Health Records for Better Medical Predictions

The widespread adoption of Electronic Health Record (EHR) systems in healthcare institutes has generated vast amounts of medical data, offering significant opportunities for improving healthcare services through deep learning techniques. However, the complex and diverse modalities and feature structures in real-world EHR data pose great challenges for deep learning model design. To address the multi-modality challenge in EHR data, current approaches primarily rely on hand-crafted model architectures based on intuition and empirical experiences, leading to sub-optimal model architectures and limited performance. Therefore, to automate the process of model design for mining EHR data, we propose a novel neural architecture search (NAS) framework named AutoFM, which can automatically search for the optimal model architectures for encoding diverse input modalities and fusion strategies. We conduct thorough experiments on real-world multi-modal EHR data and prediction tasks, and the results demonstrate that our framework not only achieves significant performance improvement over existing state-of-the-art methods but also discovers meaningful network architectures effectively.

Fenglong Ma, Suhan Cui, Jiaqi Wang, Yuan Zhong
Pennsylvania State University
fenglong@psu.edu, suhan@psu.edu, jqwang@psu.edu, yfz5556@psu.edu

Han Liu
Dalian University of Technology
liu.han.dut@gmail.com

Ting Wang
Stony Brook University
twang@cs.stonybrook.edu

CP9

3D Molecular Geometry Analysis with 2D Graph

Ground-state 3D geometries of molecules are essential for many molecular analysis tasks. Modern quantum mechanical methods can compute accurate 3D geometries but are computationally prohibitive. Currently, an efficient alter-

native to computing ground-state 3D molecular geometries from 2D graphs is lacking. Here, we propose a novel deep learning framework to predict 3D geometries from molecular graphs. To this end, we develop an equilibrium message passing neural network (EMPNN) to better capture ground-state geometries from molecular graphs. To provide a testbed for 3D molecular geometry analysis, we develop a benchmark that includes a dataset with precise ground-state geometries of approximately 4 million molecules. Experimental results show that EMPNN can efficiently predict more accurate ground-state 3D geometries than RDKit and other deep learning methods. Results also show that the proposed framework outperforms self-supervised learning methods on property prediction tasks.

Zhao Xu, Yaochen Xie, Youzhi Luo, Xuan Zhang
Texas A&M University
zhaoxu@tamu.edu, ethanycx@tamu.edu, yzluo@tamu.edu, xuan.zhang@tamu.edu

Xinyi Xu
Xidian University
xyxu.xd@gmail.com

Meng Liu, Kaleb Dickerson
Texas A&M University
mengliu@tamu.edu, kaleb.dickerson2001@tamu.edu

Cheng Deng
Xidian University
chdengxd@gmail.com

Maho Nakata
RIKEN
maho@riken.jp

Shuiwang Ji
Texas A&M University
sji@tamu.edu

CP9

Graph-Based Student Knowledge Profile for Online Intelligent Education

Student knowledge profile is the basis for adaptive learning applications in online learning resulting from modeling the student mastery of knowledge concepts. In recent years, typical works based on knowledge tracing (KT) expect to profile students and have achieved significant success for the next performance prediction. However, in practical online learning scenarios, current methods tend to suffer from the following challenges: 1) Prediction inconsistency: The accuracy of the next performance prediction is inconsistent with the accuracy of student knowledge profile prediction. 2) Cold start of knowledge: In online learning scenarios, it is often necessary to profile some knowledge concepts without learning records in advance. In this paper, we propose a novel Graph-based Student Knowledge Profile Model (GSKPM), along with a new end-to-end training objective, to tackle these challenges. We first define a new training objective to ensure the model is capable of inferring consistent student knowledge profiles. Then in this model, a two-stage hyper-aggregation process is employed to make full use of the topological relations between knowledge concepts and knowledge domains to provide information during profiling, especially for cold start knowledge concepts.

Jinze Wu
iFLYTEK Co., Ltd
hxwjz@mail.ustc.edu.cn

Haotian Zhang
University of science and technology of China
sosweetzhang@mail.ustc.edu.cn

Zhenya Huang
University of Science and Technology of China
huangzhy@ustc.edu.cn

Liang Ding
iFLYTEK Co., Ltd
liangding3@iflytek.com

Qi Liu
University of Science and Technology of China
qiliuql@ustc.edu.cn

Jing Sha
iFLYTEK Co., Ltd
jingsha@iflytek.com

Enhong Chen
University of Science and Technology of China
cheneh@ustc.edu.cn

Shijin Wang
iFLYTEK Co., Ltd
sjwang3@iflytek.com

CP10

Foundation Models for Spatiotemporal Tasks in the Physical World

Foundation models like ChatGPT are poised to transform society by providing general intelligence for question-answering in healthcare, education, and law. They are also expected to make dramatic impacts in the way of AI solving spatiotemporal tasks in the physical world, such as intelligent robots in manufacturing or construction. However, one major handicap is that foundation models do not understand the physical knowledge of the world, leading to unexpected model behaviors and significant safety risks. This paper discusses emerging opportunities and unique challenges of foundation models solving spatiotemporal tasks in the physical world. The emerging opportunity is to integrate foundation models with physical engine plugins (e.g., robots, digital map engines, or Earth system models). We also identify several new research directions to enhance safety for the integrated systems by knowledge-guided spatiotemporal in-context-learning, formal-logic-based verification, and safety alignment, as well as new benchmarking datasets and evaluation metrics.

Zhe Jiang, Yu Wang, Zelin Xu
University of Florida
zhe.jiang@ufl.edu, yuwang1@ufl.edu, zelin.xu@ufl.edu

CP10

Data Silences: How to Unsilence the Uncertainties in Data Science

When we wrangle the data in data science, we design the data to make it fit-for-analysis. Wrangling involves the re-

moval or reduction of uncertainties, such as outliers, missing values, mal-distributions, and the details of feature engineering. Many of the steps of data wrangling go unrecorded or poorly recorded, in terms of both what was done and also the rationale for why it was done. In this way, we impose multiple types of data silences on the data, and often on the sources (people) who are "behind" the data. In this paper, we articulate how we may perform multiple types of silencing. We challenge comfortable conceptions of the nature of data, and we call on the data-science community to devise and adopt methodologies to unsilence data.

Michael Muller
IBM Research
michael_muller@us.ibm.com

CP10

Blue Sky: Multilingual, Multimodal Domain Independent Deception Detection

Deception, a pervasive aspect of communication, has undergone a significant transformation in the digital age. With the globalization of online interactions, individuals are communicating in multiple languages, mixing languages on social media. A variety of data is now available in many languages, while the techniques for detecting deception are similar across the board. Recent studies have shown the possibility of the existence of universal linguistic cues to deception across domains within the English language; however, the existence of such cues in other languages remains unknown. Furthermore, the practical task of deception detection in low-resource languages is not a well-studied problem due to the lack of labeled data. Another dimension of deception is multimodality. For example, in fake news or disinformation, there may be a picture with an altered caption. This paper calls for a comprehensive investigation into the complexities of deceptive language across linguistic boundaries and modalities, and raises the possibility of use of multilingual transformer models and labeled data in a variety of languages to universally address the task of deception detection.

Rakesh Verma
University of Houston, U.S.
rmverma@cs.uh.edu

Dainis Boumber, Fati Qachfar
University of Houston
dainis.bomber@gmail.com, f.qachfar@gmail.com

CP11

An Image Dataset for Benchmarking Recommender Systems with Raw Pixels

The advent of large language models has inspired active and promising research focused on developing text content-based recommendation models. Meanwhile, although image features are also key signals in recommender systems, there is currently a lack of research on recommendation models that are primarily based on raw image pixels. The lack of large-scale datasets containing raw images in visually driven recommendation scenarios has been a significant barrier to the development of this research direction. To address this challenge, we introduce PixelRec, a comprehensive dataset of cover images collected from a video streaming platform. With approximately 200 million user image interactions, 30 million users, and 400,000 high-resolution short video cover images, PixelRec facilitates the development, benchmarking, and analysis of various image pixel based recommendation models. Leveraging this dataset, we establish an accessible pipeline to implement a series of vision-based recommendation models, providing extensive benchmark results for them. Our contributions include the PixelRec dataset, baseline algorithms, operational pipeline, exploratory findings, and the PixelRec benchmark. We believe PixelRec will significantly advance research on recommendation models based on image content and foster fruitful collaboration between the fields of recommender systems and computer vision. The dataset, code, and documents are available at https://github.com/westlake-repl/PixelRec.

Yu Cheng
Westlake University
chengyu@westlake.edu.cn

Yunzhu Pan
University of Electronic Science and Technology of China
yunzhupan@outlook.com

Jiaqi Zhang, Yongxin Ni
Westlake University
zhangjiaqi@westlake.edu.cn, niyongxin@westlake.edu.cn

Aixin Sun
Nanyang Technological University
axsun@ntu.edu.sg

Fajie Yuan
Westlake University
yuanfajie@westlake.edu.cn

CP11

User Migration Across Multiple Social Media Platforms

After Twitter's ownership change and policy shifts, many users reconsidered their go-to social media outlets and platforms like Mastodon, Bluesky, and Threads became attractive alternatives in the battle for users. Based on the data from over 14,000 users who migrated to these platforms within the first eight weeks after the launch of Threads, our study examines: (1) distinguishing attributes of Twitter users who migrated, compared to non-migrants; (2) temporal migration patterns and associated challenges for sustainable migration faced by each platform; and (3) how these new platforms are perceived in relation to Twitter. Our research proceeds in three stages. First, we examine migration from a broad perspective, not just one-to-one migration. Second, we leverage behavioral analysis to pinpoint the distinct migration pattern of each platform. Last, we employ a Large Language Model (LLM) to discern stances towards each platform and correlate them with the platform usage. This in-depth analysis illuminates migration patterns amid competition across social media platforms.

Ujun Jeong, Ayushi Nirmal, Kritshekhar Jha, Susan Xu Tang, H. Russell Bernard, Huan Liu
Arizona State University
ujeong1@asu.edu, anirmal1@asu.edu, kjha9@asu.edu, susan.tang@asu.edu, asuruss@asu.edu, huanliu@asu.edu

CP11

Disinformation Detection: An Evolving Challenge

in the Age of LLMs

The advent of generative Large Language Models (LLMs) such as ChatGPT has catalyzed transformative advancements across multiple domains. However, alongside these advancements, they have also introduced potential threats. One critical concern is the misuse of LLMs by disinformation spreaders, leveraging these models to generate highly persuasive yet misleading content that challenges the disinformation detection system. This work aims to address this issue by answering three research questions: (1) To what extent can the current disinformation detection technique reliably detect LLM-generated disinformation? (2) If traditional techniques prove less effective, can LLMs themself be exploited to serve as a robust defense against advanced disinformation? and, (3) Should both these strategies falter, what novel approaches can be proposed to counter this burgeoning threat effectively? A holistic exploration for the formation and detection of disinformation is conducted to foster this line of research.

Bohan Jiang
Arizona State University
School of Computing and Augmented Intelligence
bjiang14@asu.edu

Zhen Tan, Ayushi Nirmal, Huan Liu
Arizona State University
ztan36@asu.edu, anirmal1@asu.edu, huanliu@asu.edu

CP11

Label Distribution Learning-Enhanced Dual-Knn for Text Classification

Many text classification methods usually introduce external information (e.g., label descriptions and knowledge bases) to improve the classification performance. Compared to external information, some internal information generated by the model itself during training, like text embeddings and predicted label probability distributions, are exploited poorly when predicting the outcomes of some texts. In this paper, we focus on leveraging this internal information, proposing a dual k nearest neighbor (DkNN) framework with two kNN modules, to retrieve several neighbors from the training set and augment the distribution of labels. For the kNN module, it is easily confused and may cause incorrect predictions when retrieving some nearest neighbors from noisy datasets (datasets with labeling errors) or similar datasets (datasets with similar labels). To address this issue, we also introduce a label distribution learning module that can learn label similarity, and generate a better label distribution to help models distinguish texts more effectively. This module eases model overfitting and improves final classification performance, hence enhancing the quality of the retrieved neighbors by kNN modules during inference. Extensive experiments on datasets demonstrate that our method improves the classification performance of pre-trained and non-pre-trained models effectively.

Bo Yuan, Yulin Chen
Zhejiang University
354151531@qq.com, sylvia_cyl@qq.com

Zhen Tan
Arizona State University
ztan36@asu.edu

Jinyan Wang
Zhejiang University
3180103768@zju.edu.cn

Huan Liu
Arizona State University
huanliu@asu.edu

Yin Zhang
Zhejiang University
yinzh@zju.edu.cn

CP11

Refining Pre-Trained Language Models for Domain Adaptation with Entity-Aware Discriminative and Contrastive Learning

With the rapid advancement of pre-trained language models (PLMs), the adaptation of these models to specialized domains has emerged as an essential area of research. However, PLMs encounter substantial challenges when deployed in highly specialized fields, such as the Chinese military equipment domain. The intricate nature of the domain entities, characterized by multi-class character combinations and the dynamic variability of names, has exposed a significant deficiency in the ability of PLMs to accurately recognize and comprehend them. In response to this challenge, we introduce a novel method to augment PLMs with domain-specific entity knowledge through discriminative and contrastive learning objectives. The discriminative objective guides the PLM in discerning domain entities by distinguishing between the subwords of common words and those of specific entities, while the contrastive objective strategically infuses domain entity semantics from a pre-established intermediate entity embedding vocabulary. By leveraging this methodology, we trained a unified military equipment-oriented PLM. To assess the model's performance, we construct a set of domain test datasets and conduct a comprehensive evaluation. The experimental results demonstrate that our model significantly surpasses baseline models in all evaluated metrics, thereby underscoring the effectiveness of the proposed method.

Jian Yang
Xidian University
ferryjian@gmail.com

Xinyu Hu
Sun Yat-sen University
husense@foxmail.com

Yulong Shen
Xidian University
ylshen@mail.xidian.edu.cn

Gang Xiao
National Key Laboratory of Complex System Simulation
searchware@qq.com

CP12

Test-Time Training for Spatial-Temporal Forecasting

Despite the recent success of deep neural networks in spatialtemporal forecasting, existing methods suffer from distribution shifts between the training and test data, failing to address the non-stationary and abrupt changes at test time. To solve this problem, we propose a novel test-time training framework for spatial-temporal forecasting. Instead of employing a fixed trained model, we adapt the

trained model with only one or a mini-batch of test examples to address the test data shifts. The unique spatial structure with hundreds of geographical locations offers an effective batch size to explore the test-time distribution and avoid overfitting. To implement test-time training on spatial-temporal data, we devise a bidirectional cycle-consistent architecture consisting of a forward and a backward cyclic network. Each network has a shared encoder and two direction-aware decoders. At the test time, two self-supervised auxiliary tasks (forward?backward and backward?forward reconstruction) are proposed to adapt the trained model without accessing the target labels. Besides, the bi-cyclic structure of our model can also improve the forecasting task at training time, and ensure consistency between the training and test time. Comprehensive experiments are performed on various spatial-temporal forecasting datasets, demonstrating the effectiveness of the test-time training framework and the bidirectional-cyclic structure.

Changlu Chen
University of Technology Sydney
changlu.chen@student.uts.edu.au

Yanbin Liu
Australian National University
csyanbin@gmail.com

Ling Chen, Chengqi Zhang
University of Technology, Sydney
ling.chen@uts.edu.au, chengqi.zhang@uts.edu.au

CP12

Geospatial Topological Relation Extraction from Text with Knowledge Augmentation

Geospatial topological relation extraction (GeoTopoRE) aims to extract topological relations between named geospatial entities (i.e., geo-entities) in text. It is a domain-specific relation extraction (RE) task essential in geospatial knowledge graph construction and spatial reasoning. Unlike general-purpose RE, which primarily depends on semantic and syntactic cues, GeoTopoRE requires integrating geometric knowledge about geo-entities. This is essential for accurately capturing or inferring the complex geospatial relations among entities. GeoTopoRE is not studied systematically and lacks dedicated datasets for evaluation, posing significant challenges to developing and assessing effective models. This study presents two major contributions: (i) the introduction of a high-quality, human-labeled dataset WikiTopo for GeoTopoRE, and (ii) a novel framework GeoWISE designed to adapt existing RE models to GeoTopoRE, with integrated semantic and external geospatial domain knowledge. We leverage coarse-to-fine-grained natural language inference (NLI) to align externally sourced knowledge with the semantic text context, enhanced by geospatial expertise. This integrated knowledge is then conveyed to language models as geospatial cues, enabling a nuanced understanding of topological relations. Empirical results demonstrate the efficacy of our framework in few-shot settings, showing significant and consistent improvements in the GeoTopoRE task for diverse state-of-the-art RE models.

Wei Hu, Bowen Jin, Minhao Jiang, Sizhe Zhou, Zhaonan Wang
University of Illinois Urbana-Champaign
weih9@illinois.edu, bowenj4@illinois.edu, minhaoj2@illinois.edu, sizhez@illinois.edu, znwang@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

Shaowen Wang
University of Illinois at Urbana-Champaign
shaowen@illinois.edu

CP12

Only Attending What Matter within Trajectories — Memory-Efficient Trajectory Attention

Human-generated Spatial-Temporal Data (HSTD), represented as trajectory sequences, has undergone a data revolution, thanks to advances in mobile sensing, data mining, and AI. Previous studies have revealed the effectiveness of employing attention mechanisms to analyze massive HSTD. However, traditional attention models face challenges when managing lengthy and noisy trajectories as their computation comes with large memory overheads. Furthermore, attention scores within HSTD trajectories are sparse, and clustered with varying lengths. To address these challenges, we introduce an innovative strategy named Memory-efficient Trajectory Attention (MeTA). We leverage complicated spatial-temporal features and design an innovative feature-based trajectory partition technique to shrink trajectory length. Additionally, we present a learnable dynamic sorting mechanism, with which attention is only computed between sub-trajectories that have prominent correlations. Empirical validations using real-world HSTD demonstrate that our approach not only yields competitive results but also significantly lowers memory usage compared with state-of-the-art methods. Our approach presents innovative solutions for memory-efficient trajectory attention, offering valuable insights for handling HSTD efficiently.

Mingzhi Hu
Worcester Polytechnic Institute
Worcester Polytechnic Institute
mhu3@wpi.edu

Xin Zhang
San Diego State University
xzhang19@sdsu.edu

Yanhua Li
Worcester Polytechnic Institute
Worcester, MA
yli15@wpi.edu

Yiqun Xie
University of Maryland
xie@umd.edu

Xiaowei Jia
U. of Pittsburgh
xiaowei@pitt.edu

Xun Zhou
University of Iowa
xun-zhou@uiowa.edu

Jun Luo
Logistics and Supply Chain MultiTech R&D Centre Limited

jluo@lscm.hk

CP12

Combining Satellite and Weather Data for Crop Type Mapping: An Inverse Modelling Approach

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Praveen Ravirathinam
University of Minnesota, Twin Cities
pravirat@umn.edu

CP13

Lattice Convolutional Networks for Learning Ground States of Quantum Many-Body Systems

Deep learning methods have been shown to be effective in representing ground-state wave functions of quantum many-body systems. Existing methods use convolutional neural networks (CNNs) for square lattices due to their image-like structures. For non-square lattices, the existing method uses graph neural networks (GNNs) in which structure information is not precisely captured, thereby requiring additional hand-crafted sublattice encoding. In this work, we propose lattice convolutions in which a set of proposed operations are used to convert non-square lattices into grid-like augmented lattices on which regular convolution can be applied. Based on the proposed lattice convolutions, we design lattice convolutional networks (LCN) that use self-gating and attention mechanisms. Experimental results show that our method achieves performance on par or better than the GNN method on spin 1/2 J_1-J_2 Heisenberg model over the square, honeycomb, triangular, and kagome lattices while without using hand-crafted encoding.

Cong Fu, Xuan Zhang, Huixin Zhang, Hongyi Ling, Shenglong Xu, Shuiwang Ji
Texas A&M University
congfu@tamu.edu, xuan.zhang@tamu.edu, zhanghui21@tamu.edu, hongyiling@tamu.edu, slxu@tamu.edu, sji@tamu.edu

CP13

Pretraining Molecules with Explicit Substructure Information

Generative self-supervised learning has recently become popular in molecular modeling because it can improve accuracy and generalization. However, existing generative selfsupervised tasks often have simplified designs that do not effectively use substructure information. Substructure information is important for molecules because it can provide local semantics and capture analogous semantic information on a graph-level scale. For example, -OH, as one of the substructures, is typically associated with hydrophilicity. To address this limitation, we propose a novel pretraining task that incorporates substructure information into generative self-supervised tasks. This integration involves creating a substructure-based vocabulary and fusing structural insights into the representation learning process. We evaluate our approach on 10 publicly available datasets, covering diverse molecular property prediction tasks. Our results consistently show the effectiveness of incorporating substructure information compared with both contrastive and generative self-supervised pretraining methodologies.

Yanming Shen, Yuting Ma, Shuo Yu
Dalian University of Technology
shen@dlut.edu.cn, maytmail@qq.com, shuo.yu@ieee.org

CP13

Rhine: A Regime-Switching Model with Nonlinear Representation for Discovering and Forecasting Regimes in Financial Markets

We investigate the problem of discovering and forecasting regular regime switches in a financial ecosystem comprising multiple time series. Such regime switches, indicative of varying market behaviors across distinct time intervals, are pivotal for a nuanced understanding of market dynamics, which in turn allows informed model selection for forecasting and enhanced interpretability of predictive outcomes. Despite strides in this domain, prevailing methodologies often falter due to: (1) an inability to effectively model the temporal behaviors inherent in financial series; and (2) neglecting the interdependencies among series when discovering regimes. In this paper, we propose RHINE, a **R**egime-switc**HI**ng model with **N**onlinear r**E**presentation. RHINE stands out with its kernel-based representation, adept at capturing the dynamic shifts in market regimes. This representation encapsulates the nonlinear interplay across multiple financial time series. Empirical assessments on both synthetic and real-world stock market datasets underscore RHINE's prowess. The findings illuminate that the inherent structures governing financial market behaviors are dynamic, and harnessing these dynamics via RHINE leads to a regime-based model that outperforms both conventional and state-of-the-art neural network models in predictive capabilities.

Kunpeng Xu
University of Sherbrooke
kunpeng.xu@usherbrooke.ca

Lifei Chen
Fujian Normal University
clfei@fjnu.edu.cn

Jean-Marc Patenaude
Laplace Insights
jeanmarc@laplaceinsights.com

Shengrui Wang
University of Sherbrooke
shengrui.wang@usherbrooke.ca

CP13

Meta-Adaptive Stock Movement Prediction with Two-Stage Representation Learning

Stock movement prediction has always been a tough but attractive task for researchers in data mining and machine learning In this paper, we present ***M**eta-**A**daptive **S**tock movement prediction with two-**S**tag**E** **R**epresentation learning (**MASSER**)*, a framework for stock movement prediction based on self-supervised learning and meta-learning. Specifically, we first design two-stage encoders to learn representations, the first-stage encoder aims to learn unified embeddings, and the second-stage encoder, which is based on the first stage, is used for temporal domain shift detection in the training stage via self-supervised learning. We formalize the problem of stock movement prediction into a standard meta-learning setting. Inspired by

importance sampling, we estimate the sampling probability for tasks to balance the domain discrepancy caused by evolving temporal domains. Extensive experiment results on two open source datasets show that our experimental framework with the classical ResNet as backbone achieves improvements of 5% - 9.5% on average accuracy, compared to state-of-the-art baselines. Furthermore, We extend the standard setting of stock movement prediction to a more challenging online paradigm, which is close to the realistic interday trading scenarios. MASSER outperforms baselines in both online setting and backtesting

Donglin Zhan
Columbia University
dz2478@columbia.edu

Yusheng Dai
University of Science and Technology of China
dalisondys@gmail.com

Yiwei Dong
Renmin University of China
ydong@ruc.edu.cn

Jinghai He
University of California, Berkeley
jinghai_he@berkeley.edu

Zhenyi Wang
University of Maryland, College Park
wangzhenyineu@gmail.com

James Anderson
Columbia University
ja3451@columbia.edu

CP13

MedDiffusion: Boosting Health Risk Prediction Via Diffusion-Based Data Augmentation

Health risk prediction aims to forecast the potential health risks that patients may face using their historical Electronic Health Records (EHR). Although several effective models have developed, data insufficiency is a key issue undermining their effectiveness. Various data generation and augmentation methods have been introduced to mitigate this issue by expanding the size of the training data set through learning underlying data distributions. However, the performance of these methods is often limited due to their task-unrelated design. To address these shortcomings, this paper introduces a novel, end-to-end diffusion-based risk prediction model, named MedDiffusion. It enhances risk prediction performance by creating synthetic patient data during training to enlarge sample space. Furthermore, MedDiffusion discerns hidden relationships between patient visits using a step-wise attention mechanism, enabling the model to automatically retain the most vital information for generating high-quality data. Experimental evaluation on four real-world medical datasets demonstrates that MedDiffusion outperforms 14 cutting-edge baselines in terms of PR-AUC, F1, and Cohen's Kappa. We also conduct ablation studies and benchmark our model against GAN-based alternatives to further validate the rationality and adaptability of our model design. Additionally, we analyze generated data to offer fresh insights into the model's interpretability.

Fenglong Ma, Yuan Zhong, Suhan Cui, Jiaqi Wang, Ziyi Yin, Xiaochen Wang
Pennsylvania State University
fenglong@psu.edu, yfz5556@psu.edu, suhan@psu.edu, jqwang@psu.edu, zmy5171@psu.edu, xmw5190@psu.edu

Yaqing Wang
Purdue University
wang5075@purdue.edu

Houping Xiao
Georgia State University
hxiao@gsu.edu

Mengdi Huai
Iowa State University
mdhuai@iastate.edu

Ting Wang
Stony Brook University
twang@cs.stonybrook.edu

CP14

Dualvae: Dual Disentangled Variational AutoEncoder for Recommendation

Learning precise representations of users and items to fit observed interaction data is the fundamental task of collaborative filtering. Existing studies usually infer entangled representations to fit such interaction data, neglecting to model the diverse matching relationships between users and items behind their interactions, leading to limited performance and weak interpretability. To address this problem, we propose a Dual Disentangled Variational AutoEncoder (DualVAE) for collaborative recommendation, which combines disentangled representation learning with variational inference to facilitate the generation of implicit interaction data. Specifically, we first implement the disentangling concept by unifying an attention-aware dual disentanglement and disentangled variational autoencoder to infer the disentangled latent representations of users and items. Further, to encourage the correspondence and independence of disentangled representations of users and items, we design a neighborhood-enhanced representation constraint with a customized contrastive mechanism to improve the representation quality. Extensive experiments on three real-world benchmarks show that our proposed model significantly outperforms several recent state-of-the-art baselines. Further empirical experimental results also illustrate the interpretability of the disentangled representations learned by DualVAE.

Zhiqiang Guo
Huazhong University of Science and Technology
zhiqiangguo@hust.edu.cn

Guohui Li, Jianjun Li
Huazhong University of Science and Technology
guohuili@hust.edu.cn, jianjunli@hust.edu.cn

Chaoyang Wang
Wuhan Digital Engineering Institute
sunwardtree@outlook.com

Si Shi
Guangdong Laboratory of AI and Digital Economy (SZ)

shisi@gml.ac.cn

CP14

DimReg: Embedding Dimension Search via Regularization for Recommender Systems

Modern recommender systems aim to identify items that are most pertinent to a particular user and are particularly useful when an overwhelming number of items are present. Feature embedding is essential to deep recommender systems, which constructs memory-efficient and semantically meaningful representations by mapping high-dimensional sparse feature vectors into low-dimensional dense vectors. Most existing systems assign a unified dimension to all feature fields, regardless of the diverse importance of different features, which usually results in sub-optimal performance and high memory usage. In this paper, we propose a low-cost embedding dimension search approach named DimReg for recommender systems, by assessing information overlapping between the dimensions within each feature field and pruning unimportant and redundant dimensions progressively during model training via a two-level polarization regularizer, while introducing minimum overhead. Moreover, our method does not require retraining after embedding dimension search, which significantly reduces the computational cost and is more friendly to deployment in real-world recommender systems. Extensive experiments conducted on multiple CTR (Click Through Rate) prediction tasks demonstrate that our method can efficiently reduce the model parameters up to 98.6%, and achieve strong recommendation performance outperforming existing automated embedding dimension search methods.

Mingjun Zhao, Liyao Jiang, Yakun Yu
University of Alberta
zhao2@ualberta.ca, liyao1@ualberta.ca,
yakun2@ualberta.ca

Xinmin Wang, Yi Yuan, Zheng Wei
Tencent
symonwang@tencent.com, robertyuan@tencent.com,
hemingwei@tencent.com

Di Niu, Qikai Lu
University of Alberta
dniu@ualberta.ca, qikai@ualberta.ca

CP15

Towards Spatially-Lucid Ai Classification in Non-Euclidean Space: An Application for Mxif Oncology Data

Given multi-category point sets from different place-types, our goal is to develop a spatially-lucid classifier that can distinguish between two classes based on the arrangements of their points. This problem is important for many applications, such as oncology, for analyzing immune-tumor relationships and designing new immunotherapies. It is challenging due to spatial variability and interpretability needs. Previously proposed techniques require dense training data or have limited ability to handle significant spatial variability within a single place-type. Most importantly, these deep neural network (DNN) approaches are not designed to work in non-Euclidean space, particularly point sets. Existing non-Euclidean DNN methods are limited to one-size-fits-all approaches. We explore a spatial ensemble framework that explicitly uses different training strategies, including weighted-distance learning rate and spatial domain adaptation, on various place-types for spatially-lucid classification. Experimental results on real-world datasets (e.g., MxIF oncology data) show that the proposed framework provides higher prediction accuracy than baseline methods.

Majid Farhadloo
University of Minnesota, Twin Cities
farha043@umn.edu

Arun Sharma, Jayant Gupta
University of Minnesota, Twin Cities
Department of Computer Science and Engineering
sharm485@umn.edu, gupta423@umn.edu

Alexey Leontovich, Svetomir Markovic
Mayo Clinic
leontovich.alexey@mayo.edu,
markovic.svetomir@mayo.edu

Shashi Shekhar
University of Minnesota
shekhar@umn.edu

CP15

Spatial-Temporal Augmented Adaptation Via Cycle-Consistent Adversarial Network: An Application in Streamflow Prediction

Accurate prediction of water flow is of utmost importance, particularly for ensuring water supply and informing early actions for floods and droughts. Existing flow prediction methods rely on the input of weather drivers, which hinders their applicability to monitoring small headwater streams due to the limited spatial resolution of existing weather datasets. This paper introduces a new dataset with frequent imagery on streams for water monitoring tasks. We aim to automatically predict streamflow for each stream site using frequent images taken at a sub-hourly scale. To overcome the challenge of limited labels for certain stream sites, we employ knowledge transfer from well-observed sites to poorly-observed sites via domain adaptation. As each stream site involves highly variable time series data over long periods, we introduce a novel method STCGAN (Spatial-Temporal Cycle Generative Adversarial Network), which incorporates temporal context by conditioning on the sequence's time and learns overall trends of stream flow variation. It integrates the predictive modeling of streamflow with the cyclic generative process and enhances the prediction with data augmentation using generated synthetic samples. Our experiments demonstrate superior performance of the proposed method using data collected from the West Brook area in western Massachusetts, US. The proposed method can be extended to selectively combine information from multiple stream sites, improving overall performance.

Nasrin Kalanat
University of Pittsburgh
nak168@pitt.edu

Yiqun Xie
University of Maryland
xie@umd.edu

Yanhua Li
Worcester Polytechnic Institute
Worcester, MA
yli15@wpi.edu

Xiaowei Jia
U. of Pittsburgh
xiaowei@pitt.edu

CP15

Unified Modeling and Clustering of Mobility Trajectories with Spatiotemporal Point Processes

In various application domains like transportation, urban planning, and public health, analyzing human mobility, represented as a sequence of consecutive visits (aka trajectories), is crucial for uncovering essential mobility patterns. Current practices often discretize space and time to model trajectory data with sequence-analysis techniques like Transformers and LSTM, but this discretization tends to obscure the intrinsic spatial and temporal characteristics inherent in trajectories. Recent work shows the effectiveness of modeling trajectories directly in continuous space and time using the spatiotemporal point process (STPP). However, these approaches often assume that all observed trajectories originate from a single underlying dynamic. In reality, real-world trajectories exhibit varying dynamics or moving patterns. We hypothesize that grouping trajectories governed by similar dynamics into clusters before trajectory modeling could enhance modeling effectiveness. Thus, we present a novel approach that simultaneously models trajectories in continuous space and time using STPP while clustering them. Our method leverages a variational Expectation-Maximization (EM) framework to iteratively improve the learning of trajectory dynamics and refine cluster assignments within a single training phase. Extensive tests on synthetic and real-world data demonstrate its effectiveness in clustering and modeling trajectories.

Haowen Lin
University of Southern California
haowenli@usc.edu

Yao-Yi Chiang
University of Minnesota
yaoyi@umn.edu

Li Xiong
Emory University
lxiong@emory.edu

Cyrus Shahabi
University of Southern California Speaker:
shahahbi@usc.edu

CP15

Bridging Semantics: Mobility Analytics Framework for Knowledge Transfer

This paper introduces MoveInsight, a novel framework, leveraging a Mobility Knowledge Graph and deep learning architecture to analyze individuals' GPS traces from sensor-equipped smartphones for extracting trip purposes and understanding spatio-temporal mobility patterns. Unlike traditional information retrieval methods, MoveInsight deciphers the motivations behind travels by examining relations among individuals' movement behaviors, locations, and semantic contexts. The framework employs a multi-task learning approach for annotating trajectories and a transfer learning method for extending analysis to different regions, utilizing insights from comparable areas. Through real-world dataset testing, MoveInsight outperformed baseline methods in trip-purpose extraction and Point-of-Interest annotations by around 18% to 30%, showcasing its promise in enhancing location-centric services by providing deeper insights into human mobility dynamics.

Shreya Ghosh
The Pennsylvania State University
The Pennsylvania State University
spg5897@psu.edu

Prasenjit Mitra
The Pennsylvania State University, USA
pmitra@psu.edu

CP15

Prescribed Fire Modeling Using Knowledge-Guided Machine Learning for Land Management

In recent years, the increasing threat of devastating wildfires has underscored the need for effective prescribed fire management. Process-based computer simulations have traditionally been employed to plan prescribed fires for wildfire prevention. However, even simplified process models are too compute-intensive to be used for real-time decision-making. Traditional ML methods offer computational speedup but struggle with physically inconsistent predictions, biased predictions due to class imbalance, biased estimates for fire spread metrics, and limited generalizability in out-of-distribution wind conditions. This paper introduces a novel machine learning (ML) framework that enables rapid emulation of prescribed fires while addressing these concerns. To overcome these challenges, the framework incorporates domain knowledge in the form of physical constraints, a hierarchical modeling structure to capture the interdependence among variables of interest, and also leverages pre-existing source domain data to augment training data and learn the spread of fire more effectively. Notably, improvement in fire metric (e.g., burned area) estimates offered by our framework makes it useful for fire managers, who often rely on these estimates to make decisions about prescribed burn management.

Somya Sharma, Kelly Lindsay
University of Minnesota
sharm636@umn.edu, lind0436@umn.edu

Neel Chatterjee
University of Minnesota - Twin Cities
neel.chatterjee95@gmail.com

Rohan Patil, Ilkay Altintas De Callafon
University of California - San Diego
rpatil@ucsd.edu, ialtintas@csd.edu

Michael Steinbach
University of Minnesota
stei0062@umn.edu

Daniel Giron
Colorado State University
danielrgiron95@gmail.com

Mai Nguyen
University of California - San Diego
mhnguyen@ucsd.edu

Vipin Kumar
University of Minnesota

kumar001@umn.edu

CP16

Ada-VAD: Domain Adaptable Video Anomaly Detection

Video anomaly detection (VAD) aims at identifying unusual behaviors from videos. Most of the existing video anomaly detection methods can achieve promising performance in the scenarios where training and test samples are drawn from the same distribution. In real-world situation, however, it is intractable to collect and label sufficient training video samples that cover many possible test scenarios, and existing methods demonstrate limited generalization ability. Focusing on this issue, we present the few-shot cross-domain video anomaly detection (FC-VAD) problem, which aims to adapt anomaly detection model to target samples, with access to only a few target video frames. To solve the FC-VAD problem, we propose an adaptive video anomaly detection framework named Ada-VAD, which contains a pretraining stage and an adaptation stage. In the pretraining stage, we synthesize abnormal samples and design a self-supervision based prediction task to pretrain a domain invariant model. In the adaptation stage, we adapt the pre-trained model to target domain with few-shot samples by mitigating the distribution shift with an adversarial training approach. We conduct extensive experiments on three benchmark datasets, and results show that our Ada-VAD approach outperforms the state-of-the-art VAD methods in most cases. Our code is available at https://github.com/donglgcn/ADA-VAD

Dongliang Guo
University of Virginia
dongliang.guo@virginia.edu

Yun Fu
yunfu@ece.neu.edu
northeastern university

Sheng Li
University of Virginia
shengli@virginia.edu

CP16

Dimensionality-Aware Outlier Detection

We present a nonparametric method for outlier detection that takes full account of local variations in intrinsic dimensionality within the dataset. Using the theory of Local Intrinsic Dimensionality (LID), our 'dimensionality-aware' outlier detection method, DAO, is derived as an estimator of an asymptotic local expected density ratio involving the query point and a close neighbor drawn at random. The dimensionality-aware behavior of DAO is due to its use of local estimation of LID values in a theoretically-justified way. Through comprehensive experimentation on more than 800 synthetic and real datasets, we show that DAO significantly outperforms three popular and important benchmark outlier detection methods: Local Outlier Factor (LOF), Simplified LOF, and kNN.

Michael E. Houle
New Jersey Institute of Technology
michael.houle@njit.edu

Alastair Anderberg
The University of Newcastle
Australia
anderberg.alastair@gmail.com

James Bailey
The University of Melbourne
baileyj@unimelb.edu.au

Ricardo Campello, Henrique Marques
University of Southern Denmark
Denmark
campello@imada.sdu.dk, oli@imada.sdu.dk

Miloš Radovanovic
University of Novi Sad
Serbia
radacha@dmi.uns.ac.rs

Arthur Zimek
University of Southern Denmark,
zimek@imada.sdu.dk

CP16

Combood: A Semiparametric Approach for Detecting Out-of-Distribution Data for Image Classification

Identifying out-of-distribution (OOD) data at inference time is crucial for many machine learning applications, especially for automation. We present a novel unsupervised semi-parametric framework `COMBOOD` for OOD detection with respect to image recognition. Our framework combines signals from two distance metrics, nearest-neighbor and Mahalanobis, to derive a confidence score for an inference point to be out-of-distribution. The former provides a non-parametric approach to OOD detection. The latter provides a parametric, simple, yet effective method for detecting OOD data points, especially, in the *far OOD* scenario, where the inference point is far apart from the training data set in the embedding space. However, its performance is not satisfactory in the *near OOD* scenarios that arise in practical situations. Our `COMBOOD` framework combines the two signals in a semi-parametric setting to provide a confidence score that is accurate both for the near-OOD and far-OOD scenarios. We show experimental results with the `COMBOOD` framework for different types of feature extraction strategies. We demonstrate experimentally that `COMBOOD` outperforms state-of-the-art OOD detection methods on the OpenOOD (both version 1 and most recent version 1.5) benchmark datasets (for both far-OOD and near-OOD) as well as on the documents dataset in terms of accuracy.

Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Breglind, Supratik Mukhopadhyay
Louisiana State University
mrajas1@lsu.edu, msajol1@lsu.edu,
frej.berglind@icloud.com, mmukho1@lsu.edu

Kamalika Das
Intuit Inc.
kamalika_das@intuit.com

CP16

Semi-Supervised Isolation Forest for Anomaly Detection

Anomaly detection algorithms attempt to find instances that deviate from the expected behavior. Because this is often tackled as an unsupervised task, anomaly detection models rely on exploiting intuitions about what constitutes anomalous behavior. These typically take the form of data-driven heuristics that measure the anomalousness of each instance. However, the effectiveness of unsupervised detectors are limited by the validity of their intuition. Because these are not universally true, one can improve the detectors' performance by using a semi-supervised approach that exploits a few labeled instances. This paper proposes a novel semi-supervised tree ensemble based anomaly detection framework. We compare our proposed approach to several baselines and show that it achieves comparable performance to state-of-the-art neural networks on six real-world and 14 benchmark datasets.

Luca Stradiotti, Lorenzo Perini, Jesse Davis
KU Leuven
luca.stradiotti@kuleuven.be, lorenzo.perini@kuleuven.be, jesse.davis@kuleuven.be

CP16

Multinetad: Multiplex Network-Based Anomaly Access Detection Featuring Semantic Hierarchies

Conventional anomaly access detection frameworks typically utilize all attribute fields to collectively embed them into a unified space to detect various types of anomaly accesses. However, attributes inherently contain varying semantic hierarchies, and different anomaly types exhibit inconsistent characteristics at different semantic levels. Therefore, the unified embedding results in a blending of attributes that either exhibit or do not exhibit anomaly characteristics, impacting the detection performance. To address this issue, we conduct a formal analysis of the attribute blending problem and propose MultiNetAD, a novel multiplex network-based framework designed for anomaly access detection. By introducing the multiplex network, we partition the semantic hierarchy of attributes, thereby mitigating attribute blending and consequently achieving hierarchical and unified anomaly access detection. In experiments targeting intrusion and anonymous traffic detection scenarios, MultiNetAD solves the attribute blending problem, surpasses state-of-the-art methods, and remains adaptable even with minimal proportions of anomaly accesses and labeled anomalies. Further case studies provide in-depth insights into the hierarchy and detection results.

Ziqi Yuan, Qingyun Sun
Beihang University
yuanzq@buaa.edu.cn, sunqy@buaa.edu.cn

Haoyi Zhou
Beihang University
Zhongguancun Laboratory
haoyi@buaa.edu.cn

Zukun Zhu
Beihang University
zhuzk@buaa.edu.cn

Jianxin Li
Beihang University
Zhongguancun Laboratory
lijx@buaa.edu.cn

CP17

Ge-Advgan: Improving the Transferability of Adversarial Samples by Gradient Editing-Based Adversarial Generative Model

Adversarial generative models are widely applied for generating various types of data. Accordingly, its promising performance has led to the GAN-based adversarial attack methods in the white-box and black-box attack scenarios. The importance of transferable black-box attacks lies in their ability to be effective across different models and settings. However, it remains challenging to retain the performance in terms of transferable adversarial examples for such methods. Meanwhile, we observe that some enhanced gradient-based transferable adversarial attack algorithms require prolonged time for adversarial sample generation. Thus, in this work, we propose a novel algorithm named GE-AdvGAN to enhance the transferability of adversarial samples whilst improving the algorithm's efficiency. The main approach is via optimising the training process of the generator parameters. With the functional and characteristic similarity analysis, we introduce a novel gradient editing (GE) mechanism and verify its feasibility in generating transferable samples on various models. Moreover, by exploring the frequency domain information to determine the gradient editing direction, GE-AdvGAN can generate highly transferable adversarial samples while minimizing the execution time in comparison to the state-of-the-art transferable adversarial attack algorithms. The performance of GEAdvGAN is comprehensively evaluated by large-scale experiments on different datasets.

Zhiyu Zhu, Huaming Chen
University of Sydney
zzhu2018@uni.sydney.edu.au,
huaming.chen@sydney.edu.au

Xinyi Wang
University of Malaya
22103906@siswa.um.edu.my

Jiayu Zhang
Suzhou Yierqi
zjy@szyierqi.com

Zhibo Jin
University of Sydney
zjin0915@uni.sydney.edu.au

Kim-Kwang Raymond Choo
University of Texas at San Antonio
raymond.choo@fulbrightmail.org

Jun Shen

University of Wollongong
jshen@uow.edu.au

Dong Yuan
dong.yuan@sydney.edu.au
dong.yuan@sydney.edu.au

CP17

DualToken-ViT: Position-Aware Efficient Vision Transformer with Dual Token Fusion

Self-attention-based vision transformers (ViTs) have emerged as a highly competitive architecture in computer vision. Unlike convolutional neural networks (CNNs), ViTs are capable of global information sharing. With the development of various structures of ViTs, ViTs are increasingly advantageous for many vision tasks. However, the quadratic complexity of self-attention renders ViTs computationally intensive, and their lack of inductive biases of locality and translation equivariance demands larger model sizes compared to CNNs to effectively learn visual features. In this paper, we propose a light-weight and efficient vision transformer model called DualToken-ViT that leverages the advantages of CNNs and ViTs. DualToken-ViT effectively fuses the token with local information obtained by convolution-based structure and the token with global information obtained by self-attention-based structure to achieve an efficient attention structure. In addition, we use position-aware global tokens throughout all stages to enrich the global information. Position-aware global tokens also contain the position information of the image, which makes our model better for vision tasks. We conducted extensive experiments on image classification, object detection and semantic segmentation tasks to demonstrate the effectiveness of DualToken-ViT. On the ImageNet-1K dataset, our models of different scales achieve accuracies of 75.4% and 79.4% with only 0.5G and 1.0G FLOPs, respectively.

Zhenzhen Chu
East China Normal University
51215903091@stu.ecnu.edu.cn

Jiayu Chen
Alibaba Group
yunji.cjy@alibaba-inc.com

Cen Chen
East China Normal University
cenchen@dase.ecnu.edu.cn

Chengyu Wang, Ziheng Wu, Jun Huang
Alibaba Group
chengyu.wcy@alibaba-inc.com,
zhoulou.wzh@alibaba-inc.com,
huangjun.hj@alibaba-inc.com

Weining Qian
East China Normal University
wnqian@dase.ecnu.edu.cn

CP17

Knowledge Guided Machine Learning for Extracting, Preserving, and Adapting Physics-Aware Features

Training machine learning (ML) models for scientific problems is often challenging due to limited observation data. To overcome this challenge, prior works commonly pretrain ML models using simulated data before having them fine-tuned with small real data. Despite the promise shown in initial research across different domains, these methods cannot ensure improved performance after fine-tuning because (i) they are not designed for extracting generalizable physics-aware features during pre-training, (ii) the features learned from pre-training can be distorted by the fine-tuning process. In this paper, we propose a new learning method for extracting, preserving, and adapting physics-aware features. We build a knowledge-guided neural network (KGNN) model based on known dependencies amongst physical variables, which facilitate extracting physics-aware feature representation from simulated data. Then we fine-tune this model by alternately updating the encoder and decoder of the KGNN model to enhance the prediction while preserving the physics-aware features learned through pre-training. We further propose to adapt the model to new testing scenarios via a teacher-student learning framework based on the model uncertainty. The results demonstrate that the proposed method outperforms many baselines by a good margin, even using sparse training data or under out-of-sample testing scenarios.

Erhu He
University of Pittsburgh
erh108@pitt.edu

Yiqun Xie
University of Maryland
xie@umd.edu

Licheng Liu, Zhenong Jin
University of Minnesota
lichengl@umn.edu, jinzn@umn.edu

Dajun Zhang
Carleton University
dajunzhang@cmail.carleton.ca

Xiaowei Jia
U. of Pittsburgh
xiaowei@pitt.edu

CP17

Identification and Uses of Deep Learning Backbones via Pattern Mining

Deep learning is extensively used in many areas of machine learning and data mining as a black-box method with impressive results. However, understanding the core mechanism of how deep learning makes predictions is a relatively understudied problem. Here we explore the notion of identifying a backbone of deep learning for a given group of instances. A group here can be instances of the same class or even misclassified instances of the same class. We view each instance for a given group as activating a subset of neurons and attempt to find a sub-graph of neurons associated with a given class/group. We formulate this problem as a set cover style problem and show it is intractable and presents a highly constrained integer linear programming (ILP) formulation. As an alternative, we explore a coverage-based heuristic approach related to pattern mining, and show it converges to a Pareto equilibrium point of the ILP formulation. Experimentally we explore these backbones to identify mistakes and improve performance, explanation, and visualization. We demonstrate application-based results using several challenging data sets, including Bird Audio Detection (BAD) Challenge and Labeled Faces in

the Wild (LFW), as well as the classic MNIST data.

Ian Davidson
University of California, Davis
davidson@cs.ucdavis.edu

MICHAEL Livanos
UC Davis
mjlivanos@ucdavis.edu

CP17

Dual-Disentangled Deep Multiple Clustering

Multiple clustering has gathered significant attention in recent years due to its potential to reveal multiple hidden structures of the data from different perspectives. Most of multiple clustering methods first derive feature representations by controlling the dissimilarity among them, subsequently employing traditional clustering methods (e.g., k-means) to achieve the final multiple clustering outcomes. However, the learned feature representations can exhibit a weak relevance to the ultimate goal of distinct clustering. Moreover, these features are often not explicitly learned for the purpose of clustering. Therefore, in this paper, we propose a novel Dual-Disentangled deep Multiple Clustering method named DDMC by learning disentangled representations. Specifically, DDMC is achieved by a variational Expectation-Maximization (EM) framework. In the E-step, the disentanglement learning module employs coarse-grained and fine-grained disentangled representations to obtain a more diverse set of latent factors from the data. In the M-step, the cluster assignment module utilizes a cluster objective function to augment the effectiveness of the cluster output. Our extensive experiments demonstrate that DDMC consistently outperforms state-of-the-art methods across seven commonly used tasks. Our code is available at https://github.com/Alexander-Yao/DDMC.

Jiawei Yao, Juhua Hu
University of Washington
jwyao@uw.edu, juhuah@uw.edu

CP18

Vietoris-Rips Complex: A New Direction for Cross-Domain Cold-Start Recommendation

Cross-domain recommendation (CDR) has emerged as a promising solution to alleviating the cold-start problem by leveraging information from an auxiliary source domain to generate recommendations in a target domain. Most CDR techniques fall into a category known as *bridge-based methods*, but many of them fail to account for the structure and rating behavior of target users from the source domain into the recommendation process. Therefore, we present a novel framework called Vietoris-Rips Complex for Cross-Domain Recommendation (VRCDR), which utilizes the Vietoris-Rips Complex (a technique from computational geometry) to understand the underlying structure in user behavior from the source domain, and includes the learned information into recommendations in the target domain to make the recommendations more personalized to users' niche preferences. Extensive experiments on large, real-world datasets demonstrate that VRCDR consistently improves recommendations compared to state-of-the-art bridge-based CDR methods.

Ajay Krishna Vajjala, Dipak Meher, Shrunal Pothagoni, Ziwei Zhu, David Rosenblum
George Mason University
akrish@gmu.edu, dmeher@gmu.edu, spothago@gmu.edu, zzhu20@gmu.edu, dsr@gmu.edu

CP18

Towards More Robust and Accurate Sequential Recommendation with Cascade-Guided Adversarial Training

Sequential recommendation models, models that learn from chronological user-item interactions, outperform traditional recommendation models in many settings. Despite the success of sequential recommendation, their robustness has recently come into question. Two properties unique to the nature of sequential recommendation models may impair their robustness - the cascade effects induced during training and the model's tendency to rely too heavily on temporal information. To address these vulnerabilities, we propose Cascade-guided Adversarial training, a new adversarial training procedure that is specifically designed for sequential recommendation models. Our approach harnesses the intrinsic cascade effects present in sequential modeling to produce strategic adversarial perturbations to item embeddings during training. Experiments on training state-of-the-art sequential models on four public datasets from different domains show that our training approach produces superior model ranking accuracy and superior model robustness to real item replacement perturbations when compared to both standard model training and generic adversarial training.

Juntao Tan
Rutgers University
chrisjtan23@gmail.com

Shelby Heinecke
Salesforce Research
shelby.heinecke@salesforce.com

Zhiwei Liu
Salesforce AI Research
zhiweiliu@salesforce.com

Yongjun Chen
Salesforce Research
yongjunchen1995@gmail.com

Yongfeng Zhang
Rutgers University
zhangyf07@gmail.com

Huan Wang
Salesforce Research
huan.wang@salesforce.com

CP18

Variational Invariant Representation Learning for Multimodal Recommendation

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Wei Yang
University of Chinese Academy of Sciences

weiyangvia@gmail.com

CP19

Stable Synthetic Control with Anomaly Detection for Causal Inference

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Qiang Li
Wilfrid Laurier University
qli@wlu.ca

CP19

Analysis of Causal and Non-Causal Convolution Networks for Time Series Classification

Applications of neural networks like MLPs and ResNets in temporal data mining has led to improvements on the problem of time series classification. Recently, a new class of networks called Temporal Convolution Networks (TCNs) have been proposed for various time series tasks. Instead of time invariant convolutions they use temporally causal convolutions, this makes them more constrained than ResNets but surprisingly good at generalization. This raises an important question: How does a network with causal convolution solve these tasks when compared to a network with acausal convolutions? As the first attempt at answering these questions, we analyze different architectures through a lens of representational subspace similarity. We demonstrate that the evolution of input representations in the layers of TCNs is markedly different from ResNets and MLPs. We find that acausal networks are prone to form groupings of similar layers and TCNs on the other hand learn representations that are much more diverse throughout the network. Next, we study the convergence properties of internal layers across different architecture families and discover that the behaviour of layers inside Acausal network is more homogeneous when compared to TCNs. Our extensive empirical studies offer new insights into internal mechanisms of convolution networks in the domain of time series analysis and may assist practitioners gaining deeper understanding of each network.

Uday Singh Saini
University of California Riverside
usain001@ucr.edu

Zhongfang Zhuang, Chin-Chia Michael Yeh, Wei Zhang
Visa Research
zzhuang@visa.com, miyeh@visa.com, wzhan@visa.com

Evangelos Papalexakis
University of California, Riverside
epapalex@cs.ucr.edu

CP19

Camu: Disentangling Causal Effects in Deep Model Unlearning

Machine unlearning requires removing the information of forgetting data while keeping the necessary information of remaining data. Despite recent advancements in this area, existing methodologies mainly focus on the effect of removing forgetting data without considering the negative impact this can have on the information of the remaining data, resulting in significant performance degradation after data removal. Although some methods try to repair the performance of remaining data after removal, the forgotten information can also return after repair. Such an issue is due to the intricate intertwining of the forgetting and remaining data. Without adequately differentiating the influence of these two kinds of data on the model, existing algorithms take the risk of either inadequate removal of the forgetting data or unnecessary loss of valuable information from the remaining data. To address this shortcoming, the present study undertakes a causal analysis of the unlearning and introduces a novel framework termed Causal Machine Unlearning (CaMU). This framework adds intervention on the information of remaining data to disentangle the causal effects between forgetting data and remaining data. Then CaMU eliminates the causal impact associated with forgetting data while concurrently preserving the causal relevance of the remaining data. Empirical results suggest that CaMU enhances performance on the remaining data and effectively minimizes the influences of forgetting data.

Shaofei Shen, Chenhao Zhang, Alina Bialkowski
The University of Queensland
shaofei.shen@uq.edu.au, chenhao.zhang@uq.edu.au,
alina.bialkowski@uq.edu.au

Weitong Chen
University of Adelaide
t.chen@adelaide.edu.au

Miao Xu
University of Queensland
miao.xu@uq.edu.au

CP19

Robust Estimation of Causal Heteroscedastic Noise Models

Distinguishing the cause and effect from bivariate observational data is the foundational problem that finds applications in many scientific disciplines. One solution to this problem is assuming that cause and effect are generated from a structural causal model, enabling identification of the causal direction after estimating the model in each direction. The heteroscedastic noise model is a type of structural causal model where the cause can contribute to both the mean and variance of the noise. Current methods for estimating heteroscedastic noise models choose the Gaussian likelihood as the optimization objective which can be suboptimal and unstable when the data has a non-Gaussian distribution. To address this limitation, we propose a novel approach to estimating this model with Student's t-distribution, which is known for its robustness in accounting for sampling variability with smaller sample sizes and extreme values without significantly altering the overall distribution shape. This adaptability is beneficial for capturing the parameters of the noise distribution in heteroscedastic noise models. Our empirical evaluations demonstrate that our estimators are more robust and achieve better overall performance across synthetic and real benchmarks.

Quang-Duy Tran, Bao Duong, Phuoc Nguyen, Thin Nguyen
Deakin University
q.tran@deakin.edu.au, duongng@deakin.edu.au,
phuoc.nguyen@deakin.edu.au,

thin.nguyen@deakin.edu.au

CP20

Distributed Collapsed Gibbs Sampler for Dirichlet Process Mixture Models in Federated Learning

Dirichlet Process Mixture Models (DPMMs) are widely used to address clustering problems. Their main advantage lies in their ability to automatically estimate the number of clusters during the inference process through the Bayesian non-parametric framework. However, the inference becomes considerably slow as the dataset size increases. This paper proposes a new distributed Markov Chain Monte Carlo (MCMC) inference method for DPMMs (DisCGS) using sufficient statistics. Our approach uses the collapsed Gibbs sampler and is specifically designed to work on distributed data across independent and heterogeneous machines, which habilitates its use in horizontal federated learning. Our method achieves highly promising results and notable scalability. For instance, with a dataset of 100K data points, the centralized algorithm requires approximately 12 hours to complete 100 iterations while our approach achieves the same number of iterations in just 3 minutes, reducing the execution time by a factor of 200 without compromising clustering performance. The code source is publicly available at https://github.com/redakhoufache/DisCGS.

Reda Khoufache
Paris-Saclay University, UVSQ, David Lab
reda.khoufache@uvsq.fr

CP20

Hyperflora: Federated Learning with Instantaneous Personalization

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Di Niu
University of Alberta
dniu@ualberta.ca

CP20

Personalized Federated Learning with Contextual Modulation and Meta-Learning

Federated learning has emerged as a promising approach for training machine learning models on decentralized data sources while preserving data privacy. However, challenges such as communication bottlenecks, heterogeneity of client devices, and non-i.i.d. data distribution pose significant obstacles to achieving optimal model performance. We propose a novel framework that combines federated learning with meta-learning techniques to enhance both efficiency and generalization capabilities. Our approach introduces a federated modulator that learns contextual information from data batches and uses this knowledge to generate modulation parameters. These parameters dynamically adjust the activations of a base model, which operates using a MAML-based approach for model personalization. Experimental results across diverse datasets highlight the improvements in convergence speed and model performance compared to existing federated learning approaches. These findings highlight the potential of incorporating contextual information and meta-learning techniques into federated learning, paving the way for advancements in distributed machine learning paradigms.

Anna Vettoruzzo, Mohamed-Rafik Bouguelia, Thorsteinn Rögnvaldsson
Halmstad University
anna.vettoruzzo@hh.se, mohamed-rafik.bouguelia@hh.se, thorsteinn.rognvaldsson@hh.se

CP20

Deep Efficient Private Neighbor Generation for Subgraph Federated Learning

Behemoth graphs are often fragmented and separately stored by multiple data owners as distributed subgraphs in many realistic applications. Without harming data privacy, it is natural to consider the subgraph federated learning (subgraph FL) scenario, where each local client holds a subgraph of the entire global graph, to obtain globally generalized graph mining models. To overcome the unique challenge of incomplete information propagation on local subgraphs due to missing cross-subgraph neighbors, previous works resort to the augmentation of local neighborhoods through the joint FL of missing neighbor generators and GNNs. Yet their technical designs have profound limitations regarding the utility, efficiency, and privacy goals of FL. In this work, we propose FedDEP to comprehensively tackle these challenges in subgraph FL. FedDEP consists of a series of novel technical designs: (1) Deep neighbor generation through leveraging the GNN embeddings of potential missing neighbors; (2) Efficient pseudo-FL for neighbor generation through embedding prototyping; and (3) Privacy protection through noise-less edge-local-differential-privacy. We analyze the correctness and efficiency of FedDEP and provide theoretical guarantees on its privacy. Empirical results on four real-world datasets justify the clear benefits of the proposed techniques.

Ke Zhang
ClusterTech Limited
cszhangk@connect.hku.hk

Lichao Sun
Legh University
lis221@lehigh.edu

Bolin Ding
Alibaba Group
bolin.ding@alibaba-inc.com

Siu Ming Yiu
The University of Hong Kong
smyiu@cs.hku.hk

Carl Yang
Department of Computer Science
Emory University
j.carlyang@emory.edu

CP20

UPFL: Unsupervised Personalized Federated Learning Towards New Clients

Personalized federated learning has gained significant attention as a promising approach to address the challenge of data heterogeneity. In this paper, we address a relatively unexplored problem in federated learning. When a federated model has been trained and deployed, and

an unlabeled new client joins, providing a personalized model for the new client becomes a highly challenging task. To address this challenge, we extend the adaptive risk minimization technique into the unsupervised personalized federated learning setting and propose our method, FedTTA. We further improve FedTTA with two simple yet highly effective optimization strategies: enhancing the training of the adaptation model with proxy regularization and early-stopping the adaptation through entropy. Moreover, we propose a knowledge distillation loss specifically designed for FedTTA to address the device heterogeneity. Extensive experiments on five datasets against eleven baselines demonstrate the effectiveness of our proposed FedTTA and its variants. The code is available at: https://github.com/anonymous-federated-learning/code.

Tiandi Ye, Cen Chen
East China Normal University
52205903002@stu.ecnu.edu.cn, cenchen@dase.ecnu.edu.cn

Yinggui Wang
Ant Group
wyinggui@gmail.com

Xiang Li, Ming Gao
East China Normal University
xiangli@dase.ecnu.edu.cn, mgao@dase.ecnu.edu.cn

CP20

Aedfl: Efficient Asynchronous Decentralized Federated Learning with Heterogeneous Devices

Federated Learning (FL) has achieved significant achievements recently, enabling collaborative model training on distributed data over edge devices. Iterative gradient or model exchanges between devices and the centralized server in the standard FL paradigm suffer from severe efficiency bottlenecks on the server. While enabling collaborative training without a central server, existing decentralized FL approaches either focus on the synchronous mechanism that deteriorates FL convergence or ignore device staleness with an asynchronous mechanism, resulting in inferior FL accuracy. In this paper, we propose an Asynchronous Efficient Decentralized FL framework, i.e., AEDFL, in heterogeneous environments with three unique contributions. First, we propose an asynchronous FL system model with an efficient model aggregation method for improving the FL convergence. Second, we propose a dynamic staleness-aware model update approach to achieve superior accuracy. Third, we propose an adaptive sparse training method to reduce communication and computation costs without significant accuracy degradation. Extensive experimentation on four public datasets and four models demonstrates the strength of AEDFL in terms of accuracy (up to 16.3% higher), efficiency (up to 92.9% faster), and computation costs (up to 42.3% lower).

Ji Liu
Hithink RoyalFlush Information Network Co., Ltd.
jiliuwork@gmail.com

Tianshi Che, Yang Zhou
Auburn University
tzc0029@auburn.edu, yangzhou@auburn.edu

Ruoming Jin
Kent State University
rjin1@kent.edu

Huaiyu Dai
North Carolina State University
hdai@ncsu.edu

Dejing Dou
Boston Consulting Group
dejingdou@gmail.com

Patrick Valduriez
INRIA
patrick.valduriez@inria.fr

CP21

Feature Interaction Aware Automated Data Representation Transformation

Recent advancements in automated feature engineering (AutoFE) have made significant progress in addressing various challenges associated with representation learning, issues such as heavy reliance on intensive labor and empirical experiences, lack of explainable explicitness, and inflexible feature space reconstruction embedded into downstream tasks. However, these approaches are constrained by: 1) generation of potentially unintelligible and illogical reconstructed feature spaces, stemming from the neglect of expert-level cognitive processes; 2) lack of systematic exploration, which subsequently results in slower model convergence for identification of optimal feature space. To address these, we introduce an interaction-aware reinforced generation perspective. We redefine feature space reconstruction as a nested process of creating meaningful features and controlling feature set size through selection. We develop a hierarchical reinforcement learning structure with cascading Markov Decision Processes to automate feature and operation selection, as well as feature crossing. By incorporating statistical measures, we reward agents based on the interaction strength between selected features, resulting in intelligent and efficient exploration of the feature space that emulates human decision-making. Extensive experiments are conducted to validate our proposed approach.

Ehtesamul Azim, Dongjie Wang
University of Central Florida
ehtesamul.azim@ucf.edu, dongjie.wang@ucf.edu

Kunpeng Liu
Portland State University
kunpeng@pdx.edu

Wei Zhang
University of Central Florida
wzhang.cs@ucf.edu

Yanjie Fu
Arizona State University
yanjie.fu@asu.edu

CP21

Word Embedding with Neural Probabilistic Prior

Input your abstract, including TeX commands, here. The abstract should be no longer than 1500 characters, including spaces. Only input the abstract text. Don't include title or author information here.

Ping Li
Baidu Research USA

liping11@baidu.com

CP21

Spatial-Aware Deep Reinforcement Learning for the Traveling Officer Problem

The traveling officer problem (TOP) is a challenging stochastic optimization task. In this problem, a parking officer is guided through a city equipped with parking sensors to fine as many parking offenders as possible. A major challenge in TOP is the dynamic nature of parking offenses, which randomly appear and disappear after some time, regardless of whether they have been fined. Thus, solutions need to dynamically adjust to currently fineable parking offenses while also planning ahead to increase the likelihood that the officer arrives during the offense taking place. Though various solutions exist, these methods often struggle to take the implications of actions on the ability to fine future parking violations into account. This paper proposes SATOP, a novel spatial-aware deep reinforcement learning approach for TOP. Our novel state encoder creates a representation of each action, leveraging the spatial relationships between parking spots, the agent, and the action. Furthermore, we propose a novel message-passing module for learning future inter-action correlations in the given environment. Thus, the agent can estimate the potential to fine further parking violations after executing an action. We evaluate our method using an environment based on real-world data from Melbourne. Our results show that SATOP consistently outperforms state-of-the-art TOP agents and is able to fine up to 22% more parking offenses.

Niklas Strauss
Munich Center for Machine Learning, LMU Munich
strauss@dbs.ifi.lmu.de

Matthias Schubert
Ludwig-Maximilians University Munich
schubert@dbs.ifi.lmu.de

CP21

Neural Locality Sensitive Hashing for Entity Blocking

Locality-sensitive hashing (LSH) is a fundamental algorithmic technique widely employed in large-scale data processing applications. However, its applicability in some real-world scenarios is limited due to the need for careful design of hashing functions that align with specific metrics. Existing LSH-based Entity Blocking solutions primarily rely on generic similarity metrics such as Jaccard similarity, whereas practical use cases often demand complex and customized similarity rules surpassing the capabilities of generic similarity metrics. Consequently, designing LSH functions for these customized similarity rules presents considerable challenges. In this research, we propose a neuralization approach to enhance locality-sensitive hashing by training deep neural networks to serve as hashing functions for complex metrics. We assess the effectiveness of this approach within the context of the entity resolution problem, which frequently involves the use of task-specific metrics in real-world applications. Specifically, we introduce NLSHBlock (Neural-LSH Block), a novel blocking methodology that leverages pre-trained language models, fine-tuned with a novel LSH-based loss function. Through extensive evaluations, we demonstrate the superiority of NLSHBlock over existing methods, exhibiting significant performance improvements. Furthermore, we showcase the efficacy of NLSHBlock in enhancing the performance of the entity matching phase.

Runhui Wang
Rutgers University
wangrunhui.pku@gmail.com

Luyang Kong, Yefan Tao, Andrew Borthwick, Davor Golac, Henrik Johnson, Shadie Hijazi
Amazon.com Services, Inc.
uyankon@amazon.com, tayefan@amazon.com, andborth@amazon.com, dgolac@amazon.com, mauritz@amazon.com, shijazi@amazon.com

Dong De, Yongfeng Zhang
Rutgers University
dong.deng@rutgers.edu, yongfeng.zhang@rutgers.edu

MS0

Session Speaker: Cornelia Caragea, National Science Foundation

Dr. Cornelia Caragea is serving as a program director in the division of Information and Intelligent Systems at NSF. Her research interests are in artificial intelligence, machine learning, data mining, information retrieval, and natural language processing, with applications to text and image analysis, and social network analysis. The overarching goal of her research is to effectively and efficiently mine and discover knowledge from large amounts of data. She is particularly interested in information extraction, supervised and semi-supervised learning, privacy analysis, knowledge integration, and information and social network analysis.

Cornelia Caragea
National Science Foundation
ccaragea@nsf.gov

MT1

See Tutorial Webpage for Tutorial Information

Tutorial Attendees
.
N/A

MT2

See Tutorial Webpage for Tutorial Information

Tutorial Attendees
.
N/A

MT3

See Tutorial Webpage for Workshop Information

Tutorial Attendees
,
NA

MT4

See Tutorial Webpage for Tutorial Information

Tutorial Attendees

,
NA

MT5
See Tutorial Webpage for Tutorial Information

Tutorial Attendees
,
NA

MT6
See Tutorial Webpage for Tutorial Information

Tutorial Attendees
,
NA

MT7
See Tutorial Webpage for Tutorial Information

Tutorial Attendees
,
NA

MT8
See Tutorial Webpage for Tutorial Information

Tutorial Attendees
.
N/A