COMPUTER SCIENCE

GRADUATE RESEARCH DAY

February 26, 2021

RevBits
Cyber Security Solutions

Stony Brook University

DEPARTMENT OF COMPUTER SCIENCE

AE-OT: A New Generative Framework Based on Optimal Transport

Dongsheng An

Faculty Advisor
Xianfeng Gu

Current generative models like generative adversarial networks (GANs) and variational autoencoders (VAEs) have attracted huge attention due to its capability to generate visual realistic images. However, most of the existing models suffer from the mode collapse or mode mixture problems. In this work, we give a theoretic explanation of the both problems by Figalli's regularity theory of optimal transportation maps. Basically, the generator compute the transportation maps between the white noise distributions and the data distributions, which are in general discontinuous. However, deep neural networks (DNNs) can only represent continuous maps. This intrinsic conflict induces mode collapse and mode mixture. In order to tackle the both problems, we explicitly separate the manifold embedding and the optimal transportation; the first part is carried out using an autoencoder (AE) to map the images onto the latent space; the second part is accomplished using a GPU-based convex optimization to find the discontinuous transportation maps. Composing the extended optimal transport (OT) map and the decoder, we can finally generate new images from the white noise. This AE-OT model avoids representing discontinuous maps by DNNs, therefore effectively prevents mode collapse and mode mixture.

Anthoring Knowledge via Natural Language

Yuheng Wang

Faculty Advisors
Michael Kifer, Paul Fodor

The proposed Knowledge Authoring Logic Machine (KALM) enables domain experts without knowledge representation skills to author knowledge via Controlled Natural Language (CNL) with high accuracy (95%). However, by relying on CNL, KALM only allows restricted syntactic forms as input, which is too burdensome for users. The reliance on CNL also makes KALM unable to parse past and future tense, and therefore, hinders KALM from acquiring time-related knowledge. To address these issues, a natural language parsing toolkit, Stanza, is introduced to generate syntactic analysis instead of the CNL parser. Unfortunately, applying Stanza brings a series of problems on lemmatization, POS tagging, passive voice handling, named entity handling, and so on. These problems significantly lower the knowledge authoring accuracy. Thus, KALM-2 is proposed, which enables users to author knowledge via natural language rather than CNL without any accuracy loss (95%).

ATRIA: Adaptive Streaming of 360-Degree Videos with Reinforcement Learning

Sohee Kim Park

Faculty Advisor
Samir Das

For bandwidth-efficient streaming of 360-degree videos, the streaming technique must adapt both to the changing viewport of the user and variations of the available network bandwidth. The state-of-the-art streaming techniques for this problem attempt to solve an optimization using simplified rules that do not adapt very well to the uncertainties related to the viewport or network.
We adopt a 3D-Convolutional Neural Networks (3DCNN) model to extract spatio-temporal features of videos and predict the viewport. Given the sequential decision-making nature of such streaming technique, we then apply a Reinforcement Learning (RL) based adaptive streaming approach. We address the challenges of using RL in this scenario, such as large action space and delayed reward evaluation. Comprehensive evaluations with real network traces show that the proposed method outperforms three tile-based streaming techniques for 360-degree videos. Compared to the tile-based streaming techniques, the average user-perceived bitrate of the proposed method is 1.3-1.7 times higher and the average quality of experience of the proposed method is also 1.6-3.4 times higher. Subjective user studies further confirm the superiority of the proposed approach.

BBR Bufferbloat in DASH Video

BEST PRESENTATION
Session 1

Santiago Vargas
Rebecca Drucker

Faculty Advisors
Aruna Balasubramanian,
Anshul Gandhi

BBR is a new congestion control algorithm and is seeing increased adoption especially for video traffic. BBR solves the bufferbloat problem in legacy loss-based congestion control algorithms where application performance drops considerably when router buffers are deep. BBR regulates traffic such that router queues don't build up to avoid the bufferbloat problem while still maintaining high throughput. Though BBR is able to combat bufferbloat for large download traffic, our analysis shows that video applications experience significantly poor performance when using BBR under deep buffers. In fact, we find that video traffic sees inflated latencies because of long queues at the router, ultimately degrading video performance. To understand this dichotomy, we study the interaction between BBR and DASH video. Our empirical investigation reveals that BBR under deep buffers and high network burstiness severely overestimates available bandwidth and does not converge to steady state, both of which result in BBR sending substantially more data into the network, causing a queue buildup. This elevated packet sending rate under BBR is ultimately caused by the router's ability to absorb bursts in traffic, which destabilizes BBR's bandwidth estimation and overrides BBR's expected logic for exiting the startup phase. We design a new bandwidth estimation algorithm and apply it to BBR (and a still-unreleased newer version of BBR called BBR2). Our modified BBR and BBR2 both see significantly improved video QoE even under deep buffers.

Catching Transparent Phish: Analyzing and Detecting MITM Phishing Toolkits

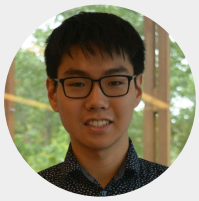
BEST PRESENTATION

Brian Kondracki
Babak Amin Azad

Faculty Advisor
Nick Nikiforakis

Phishing has long been the primary method used by attackers to obtain the login credentials of users. However, the introduction and widespread adoption of two-factor authentication (2FA) mechanisms has raised the bar for attackers, who must now obtain a 2FA code in addition to traditional credentials to compromise a victim's account. This has led to the proliferation of MITM phishing toolkits which act as malicious reverse proxy servers of online services, mirroring live content to users while extracting credentials, 2FA codes, and session cookies in transit. These tools reduce the work required by attackers and substantially increase the believability of phishing web pages. In this paper, we study and measure the current state of MITM phishing toolkits. We develop a machine learning classifier that leverages network timing side-channels to infer the presence of such toolkits in online communications with 98.97% accuracy. We then use our classifier to uncover and measure the use of these toolkits in the wild. To that end, we create a data collection framework that monitors phishing URLs from top phishing blacklist services as well as Certificate Transparency logs to automatically visit and classify phishing web pages as they are created. Using this infrastructure, we detect and capture data on 348 MITM phishing websites. We discover that MITM phishing toolkits occupy a blind spot of phishing blacklists, with only 4.6% of domains and 8.03% of IPs associated with MITM phishing toolkits present on blacklists, leaving unsuspecting users vulnerable to these attacks. Finally, we propose methods that online services can utilize to fingerprint requests originating from these toolkits and stop phishing attempts as they occur.

Chapter Captor: Text Segmentation in Novels



Charuta Pethe
Allen Kim

Faculty Advisor
Steven Skiena

Books are typically segmented into chapters and sections, representing coherent subnarratives and topics. We investigate the task of predicting chapter boundaries, as a proxy for the general task of segmenting long texts. We build a Project Gutenberg chapter segmentation data set of 9,126 English novels, using a hybrid approach combining neural inference and rule matching to recognize chapter title headers in books, achieving an F1-score of 0.77 on this task. Using this annotated data as ground truth after removing structural cues, we present cut-based and neural methods for chapter segmentation, achieving an F1-score of 0.453 on the challenging task of exact break prediction over book-length documents.

Sponsor: NSF

CNN based Framework for Registration of Serial Sections of Whole-slide IHC Images

Mousumi Roy

Faculty Advisor
Fusheng Wang

Registration of whole slide images with Immunohistochemistry (IHC) biomarker of consecutive sections of a tissue block is mandatory for cross-slide analysis. In this work we propose an unsupervised end-to-end deep learning framework HistoRegNet for registration of serial sections of whole-slide IHC images (WSI). The framework consists of an affine and a deformable module for learning the displacement vector field and a spatial transformer network for generating the final warped image from both affine and deformable transformation parameters. The model is trained with image patches extracted from 50 WSIs of IHC stained biomarker for mouse heart images. HistoRegNet is trained end-to-end by an unsupervised optimization of a similarity metric e.g. normalized cross-correlation (ncc) between input image pairs for affine module followed by ncc computed between previously warped image and the reference image. A trained HistoRegNet can be applied non-iteratively to perform registration on unseen image pairs. The performance of this model is evaluated by comparing with several state-of-the-art methods in terms of NCC, MI, SSIM and MSE metrics. The experimental results demonstrate that our proposed model outperforms existing methods suggesting it's efficacy for a quick digital pathology image registration with high accuracy.

Sponsors: NSF, NIH

Combating Dependence Explosion in Forensic Analysis Using Alternative Tag Propagation Semantics

Nahid Hossain
Sanaz Sheikhi

Faculty Advisor
R. Sekar

We are witnessing a rapid escalation in targeted cyber-attacks called Advanced and Persistent Threats (APTs). Carried out by skilled adversaries, these attacks take place over extended time periods, and remain undetected for months. A common approach for retracing the attacker's steps is to start with one or more suspicious events from system logs, and perform a dependence analysis to uncover the rest of attacker's actions. The accuracy of this analysis suffers from the dependence explosion problem, which causes a very large number of benign events to be flagged as part of the attack. In this paper, we propose two novel techniques, tag attenuation and tag decay, to mitigate dependence explosion. Our techniques take advantage of common behaviors of benign processes, while providing a conservative treatment of processes and data with suspicious provenance. Our system, called MORSE, is able to construct a compact scenario graph that summarizes attacker activity by sifting through millions of system events in a matter of seconds. Our experimental evaluation, carried out using data from two government-agency sponsored red team exercises, demonstrates that our techniques are (a) effective in identifying stealthy attack campaigns, (b) reduce the false alarm rates by more than an order of magnitude, and (c) yield compact scenario graphs that capture the vast majority of the attacks, while leaving out benign background activity.

Sponsors: DARPA NSF, ONR

Deep Video Compression

Mallesham Dasari

Faculty Advisor
Samir Das

We explore deep video compression (DVC), a data-driven, deep learning-based approach that is fundamentally different from the current generation video coding (e.g., MPEG H.265, VP9) for Internet video delivery. Our goal in this paper is to seek a first principles approach to understand the potential benefits of DVC for Internet video applications as well as for the underlying networking infrastructure. We discover that DVC can address several limitations of the current practice of video streaming while bringing additional benefits. Specifically, DVC can 1) resurrect the benefits of scalable video compression while eliminating cross layer compression overheads and dramatically decreasing the encoding and decoding latencies, 2) provide the flexibility to integrate codec features on-demand, support agile codec development, provide royalty-free codecs, and eliminates compatibility issues by enabling software defined video compression, 3) provide fine-grained rate adaptation capabilities by marrying the transport layer and application layer protocols. In this work, we design and implement a DVC method, and demonstrate its superiority over traditional video coding methods. While there are still many open questions and challenges, our preliminary results suggest that DVC can significantly benefit various stakeholders across the Internet video delivery path.

Detecting Hands and Recognizing Physical Contact in the Wild

BEST PRESENTATION:
Session 3

Supreeth Narasimhaswamy

Faculty Advisor
Minh Hoai

We investigate a new problem of detecting hands and recognizing their physical contact state in unconstrained conditions. Hand contact recognition has potential applications in contamination prevention, contact tracing, and harassment detection. This is a challenging inference task given the need to reason beyond the local appearance of hands. The lack of training annotations indicating which object or parts of an object the hand is in contact with further complicates the task. We propose a novel convolutional network based on Mask-RCNN that can jointly learn to localize hands and predict their physical contact to address this problem. The network uses outputs from another object detector to obtain locations of objects present in the scene. It uses these outputs and hand locations to recognize the hand's contact state using two attention mechanisms. The first attention mechanism is based on the hand and a region's affinity, enclosing the hand and the object, and densely pools features from this region to the hand region. The second attention module adaptively selects salient features from this plausible region of contact. To develop and evaluate our method's performance, we introduce a large-scale dataset called ContactHands, containing unconstrained images annotated with hand locations and contact states. The proposed network, including the parameters of attention modules, is end-to-end trainable. This network achieves significant relative improvement over a baseline network that was built on the vanilla Mask-RCNN architecture and trained for recognizing hand contact states.

Distributed Flight Formations

Our work primarily focuses on designing distributed controllers for multi-agent systems. First we design a cost function based Declarative Flocking (DF) controller which uses Model Predictive Control to generate its control actions. The DF controller is capable of various control actions including flocking, collision/obstacle/predator avoidance and target seeking. We then extend our work to synthesizing Distributed Neural Flocking controllers which are learnt using deep neural networks from flock trajectories generated by the DF controllers. Finally we also introduce an Acceleration-weighted Neighborhooding (AWN) methodology which designs a distributed controller capable of executing high-speed flock maneuvers.

Shouvik Roy

Faculty Advisor
Scott Smolka

Efficient and Practical Neural Question Answering for Heterogeneous Platforms

My research has made language processing systems and applications more energy-efficient, privacy-preserving, and run faster and more widely applicable to heterogeneous hardware. I have focused on question answering (QA) systems that power many real-world applications ranging from intelligent personal assistants(like Alexa or Siri) to commercial search engines such as Google and Bing. However, QA systems use complex deep learning models that run in the cloud and require expensive energy and computing resources. This also means they cannot run on mobile devices, making on-device, privacy-preserving QA impractical. My work combines systems principles with a deep understanding of NLP models. I have shown how to run complex NLP models on mobile devices using fine-grained bottleneck and critical path analysis and exploring data caching and reuse opportunities [MobiSys'19, EMDL'17]. Earlier in my research, I worked on the UIWear [MobiCom'17] project that made mobile applications more practical and accessible. I have also contributed to efficient NLP research by developing efficient QA model architecture variants that identify and remove the representation dependency in the Transformer attention blocks [ACL'20]. More recently, I have focused on modeling the energy consumption of large NLP models; preliminary results [SustaiNLP'20] show existing software energy measurements without calibration are problematic, and using hardware power meters provide more accurate energy measurements. We further develop a multi-level regression approach to provide accurate energy estimation and interpretable energy analysis for the NLP models [ACL'21 under review].

Qingqing Cao

Faculty Advisors
Aruna Balasubramanian,
Niranjan Balasubramanian

Efficient Audit Logging with eBPF

BEST POSTER: Systems

Despite several defense mechanisms, large enterprises continue to be marred by stealthy and long-term cyber-attacks, commonly known as Advanced Persistent Threats (APTs). The only way to detect and prevent such attacks is forensic analyses. The system audit logs provide crucial information for such analyses. For a successful forensic analysis, we need to capture all dependencies and information flow at the granularity of system calls. But today's approaches are either not deployable, or have poor performance. We propose a lightweight eBPF based audit logging system that requires no modifications in the Kernel. Our system can capture system events at the granularity system calls and report the data to the user-space by a high performance ring buffer.

Rohit Aich

Faculty Advisor
R. Sekar

Framework for Synthesizing Attacks on ICDs

This work seeks to evaluate the robustness of algorithms often used to deliver treatment in medical cyber-physical systems. Prior work has been mainly concerned with demonstrating that specific attacks are feasible along certain channels; while many of these devices have been in use for decades, the full extent of attacker capabilities and limits on resulting damage has yet to be defined. In this work, we investigate the robustness of an Implantable Cardioverter-Defibrilator (ICD) that monitors heart signals (EGM) to administer therapy. We design a method to identify short-lived radio-frequency attacks against one of these models: the RythymID algorithm used in Boston Scientific Devices. Minimal attack parameters can be devised via a multi-objective optimization: balancing the strength of the attack with its stealthiness. Future work explores the range of these attacks under various threat models, including adversaries with knowledge of patients' specific ailments, as well as adversaries that may have access to additional biological data from commercial wearables. This understanding is vital for the practical deployment of closed-loop medical systems since the unforseen consequences of an insecure system can be dire

Veena Krish

Faculty Advisor
Amir Rahmati

Graphics Languages & Tools: A Survey

BEST POSTER: Theory

Matthew Castellana

Faculty Advisor
Y. Annie Liu

The process of developing computer graphics applications has been changing drastically ever since it started decades ago, from simple languages that could produce only limited graphics to a plethora of languages, libraries, and interactive environments capable of creating photo-realistic graphical experiences. Each tool comes with its advantages, but also disadvantages that make graphics programming tedious and time consuming.

This report presents an overview of the state of the art in languages and tools for graphics programming, analyzing their different features and use cases. We identify and examine four key features needed for an easy-to-use, powerful high-level language---rich graphics primitives, ease of scripting, concurrent objects, and declarative constraints---and conclude with the need for an integrated programming language and environment that seamlessly supports all four features as a direction for future research.

Sponsors: National Science Foundation, GAANN

JawSense: Recognizing Unvoiced Sound using a Low-cost Ear-worn System

BEST SOCIAL IMPACT

Prerna Khanna
Tanmay Srivastava

Faculty Advisors
Aruna Balasubramanian,
Shubham Jain

This project explores a new wearable system, called JawSense, that enables a novel form of human-computer interaction based on un-voiced jaw movement tracking. JawSenseallows its user to interact with the computing machine just by moving their jaw. We study the neurological and anatomical structure of the human cheek and jaw to design JawSense so that jaw movement can be reliably captured under the strong impact of noises from human artifacts. In particular, JawSense senses the muscle deformation and vibration caused by unvoiced speaking to decode the unvoiced phonemes spoken by the user. We model the relationship between jaw movements and phonemes to develop a classification algorithm to recognize nine phonemes. Through a prototyping implementation and evaluation with six subjects, we show that JawSense can be used as a form of hands-free and privacy-preserving human-computer interaction with a 92% phoneme classification rate.

Sponsor: National Science Foundation

Learning-based Self-adaptive System for Forensic Analysis

Sanaz Sheikhi
Nahid Hossain

Faculty Advisor
R.Sekar

Advanced Persistent Threats (APTs) targeting different organizations have turned into a serious challenge for enterprise security. Skilled adversaries combine social engineering and advance exploit techniques to breach enterprises' networks. A common approach for retracing the attacker's steps is to start with one or more suspicious events from system logs, and perform a dependence analysis to uncover the rest of attacker's actions. The accuracy of this analysis suffers from the dependence explosion problem, which causes a very large number of benign events to be flagged as part of the attack.
In this research, we propose a novel technique to mitigate dependence explosion.
Our technique works based on learning system behavior to generate adaptive tags while leaving out benign background activities by assigning them benign tags and consequently decrease false positive alarms. Our system is able to construct a compact scenario graph that summarizes attacker activity by sifting through millions of system events in a matter of seconds. Our initial experimental evaluation, carried out using data from two government-agency sponsored red team exercises, demonstrates that our technique is (a) effective in identifying stealthy attack campaigns, (b) reduces the false alarm rates , and (c) yield compact scenario graphs that capture the vast majority of the attacks, while leaving out benign background activity.

Localization in the Crowd with Topological Constraints

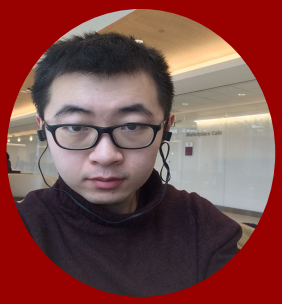
BEST PRESENTATION
Session 2

Shahira Abousamra

Faculty Advisors
Chao Chen,. Dimitris Samaras

We address the problem of crowd localization, i.e., the prediction of dots corresponding to people in a crowded scene. Due to various challenges, a localization method is prone to spatial semantic errors, i.e., predicting multiple dots within a same person or collapsing multiple dots in a cluttered region. We propose a topological approach targeting these semantic errors. We introduce a topological constraint that teaches the model to reason about the spatial arrangement of dots. To enforce this constraint, we define a persistence loss based on the theory of persistent homology. The loss compares the topographic landscape of the likelihood map and the topology of the ground truth. Topological reasoning improves the quality of the localization algorithm especially near cluttered regions. On multiple public benchmarks, our method outperforms previous localization methods. Additionally, we demonstrate the potential of our method in improving the performance in the crowd counting task.

Opioid Epidemic Study



Xinyu Dong

Faculty Advisor
Fusheng Wang

The United States is experiencing an opioid epidemic. In recent years, there were more than 10 million opioid misusers annually. Identifying patients at high risk of opioid use disorder (OUD) can help to make early clinical interventions to reduce the risk of OUD. Our goal was to predict OUD for patients on opioid medications using electronic health records data and deep learning methods. The resulting models help us to better understand OUD, providing new insights on the opioid epidemic. Electronic health records of patients who have been prescribed with medications containing active opioid ingredients were extracted from Cerner's Health Facts database for encounters between January 1, 2008 and December 31, 2017. Long Short-Term Memory (LSTM) models were applied to predict opioid use disorder risk based on five recent prior encounters before the target encounter, and compared to logistic regression, random forest, decision tree and dense neural network. Prediction performance was assessed using F-1 score, precision, recall, and AUROC. Further, these models provide a foundation for clinical tools to predict OUD before it occurs, permitting early interventions. Our sequential deep learning model provided promising prediction results which outperformed other methods, with an F1 score of 0.8023 and AUROC of 0.9369. LSTM based sequential deep learning models can accurately predict OUD using a patient's past history of electronic health records data, with minimal prior domain knowledge. This tool has the potential to improve clinical decision support for early intervention and prevention to combat the opioid epidemic.

OS Support for File System Model Checking

Yifei Liu
Wei Su

Faculty Advisor
Erez Zadok

After decades of development, file systems are still not bug-free. Hand debugging is time-consuming; existing regression suites cover only a fraction of possible states. Model checking has shown promise but there is room for further improvement. We propose MCFS, an architecture for model checking file systems comprehensively, effectively, and efficiently. MCFS uses new techniques to capture and restore a file system's in-memory and on-disk states, enabling checking of a much larger state space than previously possible. We believe comprehensive model checking of file systems requires OS APIs to capture and restore file system states efficiently. In this work we describe our earlier attempts—including unsuccessful or inefficient ones—at model checking file systems, which eventually led us to develop MCFS. We have developed a simple FUSE-based file system, VeriFS, to illustrate MCFS's model-checking principles via the proposed APIs; MCFS has already identified a few bugs that sped VeriFS's development.

Sponsor: National Science Foundation

Practical Fine-Grained Binary Code Randomization

Huan Nguyen
Soumyakant Priyadarshan

Faculty Advisor
R. Sekar

Despite its effectiveness against code reuse attacks, fine-grained code randomization has not been deployed widely due to compatibility as well as performance concerns. Previous techniques often needed source code access to achieve good performance, but this breaks compatibility with today's binary-based software distribution and update mechanisms. Moreover, previous techniques break C++ exceptions and stack tracing, which are crucial for practical deployment. In this paper, we first propose a new, tunable randomization technique called LLR(k) that is compatible with these features. Since the metadata needed to support exceptions/stack-tracing can reveal considerable information about code layout, we propose a new entropy metric that accounts for leaks of this metadata. We then present a novel metadata reduction technique to significantly increase entropy without degrading exception handling. This enables LLR(k) to achieve strong entropy with a low overhead of 2.26%.

Sponsors: Office of Naval Research, National Science Foundation

Progress Imbalance in Multi-Process Performance

Arghya Bhattacharya

Faculty Advisors
Michael A. Bender, Rezaul A. Chowdhury

Most modern systems have multi-core, multi-threaded, and time-shared architecture and processes run on a shared cache. Understanding the behavior of a cache that several concurrent processes share is crucial for application designers. Multiple homogeneous threads, threads running copies of the same program, may suffer from an imbalance of cache-residency.

We observe an interesting phenomenon: if we run multiple copies of the same program (homogeneous instances), each has private-data, and share a given shared memory, we observe a progress imbalance of the copies of the program; the program instances finish at different times. We run up to six concurrent instances of two cache-oblivious divide-and-conquer cubic matrix multiplication algorithms. We compare the differences in the running time of the program instances. More particularly, we observe the relative standard deviation of the running time of the program instances. The more concurrent programs we run, the more progress imbalance among them we get to observe."

Saliency based 360° Video Streaming

BEST POSTER

Duin Baek

Faculty Advisors
Samir Das, Jihoon Ryoo

With the recent enhancement of display technology, users demand a higher quality of streaming service, which escalates the bandwidth requirement. Considering the recent advent of high FPS (frame per second) 4K and 8K resolution 360^o videos, such bandwidth concern further intensifies in 360^o Virtual Reality (VR) content streaming at a larger scale. However, the currently available bandwidth in most of the developed countries can hardly support the bandwidth required to stream such a scale of content. To address the mismatch between the demand on higher quality of streaming service and the saturated network improvement, we propose SALI360 that practically solves the mismatch by utilizing the characteristics of the human vision system. By rendering a set of regions where viewers are expected to fixate on 360^o VR content in higher quality than the other regions, SALI360 improves viewers' quality of perception (QoP) while reducing content size with geometry-based content encoding. In our experiment, we compare the performance of SALI360 to the existing 360^o content-encoding techniques based on 20 viewers' head movement and eye gaze traces. To evaluate viewers' QoP, we propose FoL (field of look) that captures viewers' quality perception area in the visual focal field (8^o) rather than a wide (90^o) field of view (FoV). Results of our experimental 360^o VR video streaming show that SALI360 achieves 53.3% of PSNR improvement in FoL, while gaining 9.3% of PSNR improvement in FoV. In addition, our subjective study on 93 participants verifies that SALI360 improves viewers' QoP in the 360^oVR streaming service.

Sponsors: National Research Foundation of Korea (NRF), Institute of Information & communications Technology Planning & Evaluation (IITP), MSIT(Ministry of Science and ICT), Korea

Semantic Distortion in Medical Information

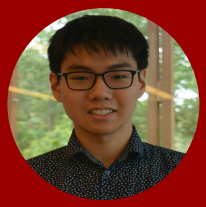
Aakash Bhatia

Faculty Advisor
Ritwik Banerjee

Today, the internet is a major source of news for most people. The claims made by news articles may be misrepresented or the meaning of these claims may change as they pass through multiple sources on the internet. Changes in the news claims could lead to widespread misinformation. This provides Natural Language Processing (NLP) researchers an opportunity to study the changes that news claims may be subject to, as well as methods to quantify these changes. Our goal is to introduce a novel dataset with semantically distorted medical news claims. This dataset will be created using 10,507 unique medical news claims obtained from 11 major news sources. To build a labelled dataset, we will perform a crowd-sourcing activity via Amazon Mechanical Turk (AMT). Each of these 10,507 news claims will be semantically distorted using 7 distortion types – Paraphrase, Generalization, Specification, Hyperbole, Meiosis, Negation and Unrelated change in entity or relation. The final dataset will thus contain approximately 70,000 distorted medical news claims. We will perform a multi-class classification using standard and state-of-the-art NLP models to understand whether these distortion types can be learned. We believe this problem has a number of real-world applications in journalism, fake-news detection, and limiting the spread of misinformation via the internet.

What time is it? Temporal Analysis of Novels

BEST POSTER:
Intelligent Systems



Allen Kim,
Charuta Pethe

Faculty Advisor
Steven Skiena

Recognizing the flow of time in a story is a crucial aspect of understanding it. Prior work related to time has primarily focused on identifying temporal expressions or relative sequencing of events, but here we propose computationally annotating each line of a book with wall clock times, even in the absence of explicit time-descriptive phrases. To do so, we construct a data set of hourly time phrases from 52,183 fictional books. We then construct a time-of-day classification model that achieves an average error of 2.27 hours. Furthermore, we show that by analyzing a book in whole using dynamic programming of breakpoints, we can roughly partition a book into segments that each correspond to a particular time-of-day. This approach improves upon baselines by over two hours. Finally, we apply our model to a corpus of literature categorized by different periods in history, to show interesting trends of hourly activity throughout the past. Among several observations we find that the fraction of events taking place past 10 P.M jumps past 1880 - coincident with the advent of the electric light bulb and city lights.