Research Statement

M. G. Sarwar Murshed
murshem@clarkson.edu
https://people.clarkson.edu/ murshem/

My research interests are in the area of developing algorithms for efficient and scalable deep learning. Despite the remarkable success of deep learning algorithms across various applications, including image and speech recognition and natural language processing, there remain several fundamental challenges that hinder their wide adaptation. Such challenges include the difficulties in precisely detecting objects with arbitrary orientations in images, limitations in robustly classifying objects that belong to similar classes, and the need for substantial computational resources to run deep learning algorithms. In my Ph.D. research, I have addressed these challenges by proposing novel solutions and making improvements to existing deep learning algorithms. My research contributions are:

- introducing oriented-regional proposal network (O-RPN) which helps deep learning algorithms to accurately detect objects with arbitrary orientations in images
- introducing discriminative localization to mitigate class overlap and to enhance the ability of AI models to effectively differentiate between findings in images that belong to classes with significant overlap.
- develops slap fingerprint segmentation models for both contact-based and contactless fingerprints using O-RPN based Faster R-CNN architecture,
- makes deep learning models scalable to deploy on resource-constrained devices.

My research has resulted in multiple book chapters, journals, and conference papers which have been published in some of the top publication venues including ACM Computing Surveys, USENIX, ICMLA, OpML, and ISVLSI. My original research on the development of resource-aware deep learning algorithms, improving the ability of AI models to discriminate highly similar objects, and the development of improved biometric segmentation and authentication technology have been extensively cited by my peers and received interest and support from leading government and industry organizations including Department of Homeland Security (DHS), Verizon Communications, and CITeR (an NSF Industry/University Cooperative Research Center IUCRC). My publications have been cited a total of 240 times according to Google Scholar, thereby demonstrating that these publications are widely recognized and relied upon in the field of artificial intelligence. I am regularly invited to conduct peer reviews for authoritative journals and conferences in the field of AI, including the Journal of Network and Computer Applications, Discover Artificial Intelligence, the Joint International Conference on Data Science & Management of Data, IEEE World Congress on Services (IEEE Services), and the IEEE International Conference on Web Services.

Doctoral Research

I conducted several fundamental research during my Ph.D. This research includes developing novel algorithms for deep learning and developing new methods to improve the performance of biometric template security algorithms. I have applied these algorithms to various domains including biometrics, computer-aided diagnosis (CADx), and supermarket safety to make practical impacts.

Oriented bounding boxes for precise object detection

It is a common practice to generate horizontal bounding boxes using deep learning algorithms to detect target objects in an image. A minimum of four parameters are required to locate a fingerprint in a slap image using a bounding box. These parameters are (x_c, y_c, w, h), where (x_c, y_c) represents the center coordinates, w represents the width and h represents the height of a horizontal bounding box. However, in real-world applications, the presence of rotated objects or features is very common and it is difficult to describe the object outline precisely with horizontal bounding boxes. We need an additional parameter θ to represent the orientation of the object relative to the vertical axis.

In this work, **we introduced an oriented-regional proposal network (O-RPN) which can calculate the θ and generate oriented bounding boxes around the target object [tBSH23].** O-RPN supports all possible angles from (0 to 360 degrees) which covers all types of possible positions in an image. We introduced orientation loss to the overall loss function so that the algorithm considers that loss during training and makes the model learn about the orientation of different objects. When we implemented O-RPN in Faster R-CNN architecture and train it on slap fingerprint images for both contactless and contact-based, we achieved better results than the state-of-the-art Faster R-CNN.

Slap fingerprint segmentation and matching

Fingerprint-based identification systems achieve higher accuracy when a slap containing multiple fingerprints of a subject is used instead of a single fingerprint. However, segmenting or auto-localizing fingerprints of different age groups is a challenging task and most of the existing segmentation models such as NFSEG and VeriFinger failed to segment them due to different reasons. Those reasons include the rotated orientations of fingerprints, different sizes and special characteristics of fingerprints in different age groups, noisy backgrounds, and the smaller size of fingertip components.

In collaboration with the Department of Homeland Security (DHS), I developed a deep learning-based slap segmentation model named CRFSEG (Clarkson Rotated Fingerprint segmentation Model) which overcomes the limitations of existing state-of-the-art fingerprint segmentation systems [tKB+21]. We build CRFSEG by introducing O-RPN to the Faster R-CNN algorithm. After training the CRFSEG algorithm on a slap dataset containing images collected from both adult and children subjects, we found the CRFSEG model outperformed NFSEG and VeriFinger in terms of segmentation (precise localization of the fingerprints), correct angle prediction of the fingerprints, and fingerprint matching. The CRFSEG model is robust against age-invariant and achieved a matching accuracy of 97.17% in a dataset containing 133000 slap images of both children and adult subjects.

Contactless fingerprint segmentation and matching

The use of contactless fingerprint identification and verification system is very important in situations where physical contact is not desirable or possible, such as patients with a contagious disease, or forensic applications. **We developed a deep learning-based contactless fingerprint segmentation model which precisely extracts fingerprints from a finger-photo image taken at a distance, without the need for physical contact with a fingerprint scanner.**

Discriminative localization using Class Activation Maps (CAMs)

It is challenging to explain the results of a deep learning model due to the non-linearity and high complexity of deep networks, with many layers and millions of parameters. Class Activation Maps (CAMs) are a technique used to explain the result of a deep learning model by visualizing the regions of the input image that are most important for generating that result. However, in the case of overlapping classes where multiple classes share similar features or characteristics for example images of car and car-wheel, CAMs suffer from neuron co-adaptation problems and visualize large regions resulting in less accurate explainability.

In collaboration with the air force lab, we introduced a method that mitigates this problem by computing directed differences in scaled activation of overlapping classes CAMs, amplifying these differences, and projecting the results onto the original image in order to precisely visualize the most important regions in an input image [VtK+21]. Our method clearly delineates spatial regions of an image that the classifier model considers as the most relevant for a given classification, even when standard CAMs for multiple classes within the model exhibit significant spatial overlap due to similarity in feature activations.

Mitigating the class overlap problem in deep learning

Early detection and diagnosis of the SARS-CoV-2 virus which causes COVID-19 remain a difficult process due to data scarcity and class overlap between X-Ray images of COVID-19 and pneumonia patients. **In joint work with Edward Verenich from Air Force Research Laboratory, Rome, NY, USA, we applied our novel discriminative localization approach based on the computation of Class Activation Maps (CAMs) that addresses the problem of class overlap and imbalance within training data [VtK+21]**. In particular, we train a Convolutional Neural Network (CNN) on viral pneumonia and COVID-19 labels and compute the difference in scaled activations of the features in the last convolutional layer of the CNN for both classes. In doing so, we exploit the feature similarity between the two diseases for training purposes, thereby mitigating issues arising due to the scarcity of COVID-19 data. In models with overlapping classes, this can yield a neuron co-adaptation problem wherein the model is less confident in the outputs of similar classes, resulting in large CAMs with decreased localization certainty. Our approach mitigates this by computing directed differences in scaled activations of pneumonia and COVID-19 CAMs.

Resource-aware deep learning on ROS (Robot Operating System) environment

Autonomous robots are often deployed in applications to continually monitor changing environments such as supermarket floors or inventory monitoring, patient monitoring, and autonomous driving. With the increasing use of deep learning techniques in robotics, a large number of robot manufacturing companies have started adopting deep learning techniques to improve the monitoring performance of autonomous robots. The Robot Operating System (ROS) is a widely-used middleware platform for building autonomous robot applications. However, the deployment of deep learning models to autonomous robots using ROS remains an unexplored area of research. Most recent research has focused on using deep learning techniques to solve specific problems (e.g. shopping assistant robots, autopilot systems, automatic annotation of 3D maps for safe flight). However, integrating the data collection hardware (e.g. sensors) and deep learning models within ROS is difficult and expensive in terms of computational power, time, and energy(battery). To address these challenges, in collaboration with Badger Technologies, I have developed **EasyDLROS, a novel framework for the robust deployment of pre-trained deep learning models on robots** [tCKH22]. Our framework is open-source, independent of the underlying deep learning framework, and easy to deploy. To test the performance of EasyDLROS, we deployed seven pre-trained deep learning models for hazard detection on supermarket floors in a simulated environment and evaluated their performances. Experimental results show that our framework successfully deploys the deep learning models on the ROS environment.

Machine learning at the network edge

Deep learning (DL) has achieved enormous success in a variety of application domains, such as intelligent machines, image processing, speech processing, and medical diagnostics. Efforts to deploy deep learning models on cyber-physical systems to make real-time decisions based on fast-changing environmental conditions have increased in the last few years. However, resource limitations are a key bottleneck because modern deep learning models need a lot of computational power, a large memory, and dedicated hardware to build and function at a reasonable rate. To address the issue, efforts have been made to reduce the computational requirement of deep learning models and make them suitable for real-time applications in a resource-constrained setting. However, Such compression methods often have a negative impact on the accuracy of DL models.

In collaboration with Microsoft, SRC Inc., and the University of Punjab, my advisor and I published a paper investigating the current methods for reducing the computation and memory requirements of deep algorithms [tMH+21]. We identified three primary techniques for making resource-aware algorithms without significant accuracy losses. Those techniques are (i) the development of lighter and faster architectures, (ii) model compression through quantization and model pruning, (iii) distributed training and inference techniques such as federated learning. Finally, we propose methods to extend compression techniques and find better solutions for making deep learning algorithms efficient in terms of resource usage. Our paper has received increasing attention from the deep-learning research community and has been widely recognized and cited.

EdgeLight model, a small deep network for resource-constrained devices

Supermarkets need to implement safety measures to create a safe environment for shoppers and employees. Many of these injuries, such as falls, are caused by a lack of safety precautions. Such incidents are preventable by timely detection of hazardous conditions such as undesirable objects on supermarket floors. In this project, my colleague and I developed **EdgeLite, a new lightweight deep learning model specifically designed for local and fast inference on edge devices that have limited memory and computing power** [tVG+20, tCKH20, VVtH20]. To build the EdgeLight model, I inherited the architecture of the GoogleNet model, then utilized the layer pruning and model compression technique. In layer pruning, we first train the GoogleNet model on our Grocery hazard dataset. This training helps the model learn useful features of the hazards. Then we performed "sensitivity analysis" which is able to evaluate the contribution of each layer to the performance of the model. After that, we removed the least important layers from the base GoogleNet model which increases the efficiency of this model and makes it resource-aware. This pruning helps us reduce 13% of the parameters used in the base GoogleNet model. We named the new architecture EdgeLight. Finally, we fine-tuned the EdgeLight model on the hazard dataset. This fine-tuning helps the model adapt to changes in the architecture. On our dataset of supermarket floor hazards, EdgeLite outperformed six state-of-the-art object detection models in terms of accuracy when deployed on the three small devices. Our experiments also showed that the energy consumption, memory usage, and inference time of EdgeLite were comparable to that of the baseline models. Based on our experiments, we provide recommendations to practitioners for overcoming resource limitations and execution bottlenecks when deploying deep learning models in settings involving resource-constrained hardware.

Data layout organization for resistive computing systems

Resistive Computing Systems (RCSs) are a type of computing system that uses resistive switches, also known as memristors, as its building blocks. RCSs are projected to be leveraged as inference engines for Deep Neural Networks (DNNs). Unfortunately, limited device yield due to immature fabrication processes may severely degrade the DNN's classification accuracy. **In collaboration with the University of Central Florida, we propose a framework for fast resilient-aware data layout organization to enable large DNNs to be deployed on RCSs with defects [ZtHE20]**. The framework contains three speed-up mechanisms: (i) sparse defect indexing: (ii) weight range characterization (iii) linear programming formulation The experimental results demonstrate that the proposed framework is capable of achieving software-level classification accuracy in resistive hardware without any use of retraining.

Compared with the previous work, the run-time is reduced by 89% on average.

Fingerprint template security

Biometric recognition relies on the comparison of a template of biometric features stored upon enrollment with a subsequent capture of biometric data. The stored template becomes a source of vulnerability, as a template that is stolen could be used through a replay attack or converted to a fake biometric to be used through a presentation attack. In this project, we work together with Verizon Inc. to develop new metrics to analyze the security of stored biometric templates. **I developed the harness for assessing the existing biometric PKI (public key infrastructure) software developed by Verizon.** We found that this system proved to be only 74% accurate. **To improve this, we overcame several problems including that the algorithm was incapable of working with incomplete images and that the parameters of the algorithm were not dynamic.** By resolving these issues, we were able to increase the accuracy to 96%.

Data collection

A clean, rich, and balanced dataset is what drives complex and sophisticated deep-learning algorithms to deliver state-of-the-art performance and therefore an invaluable resource for building robust deep-learning models. During my Ph.D. in the collaboration with CITeR, I help to build a biometric dataset containing 236000 slap fingerprint images. This unique dataset is the first effort to build a dataset containing both adult and children's fingerprints. I build data annotating and augmentation tools to annotate this huge number of images. We ended up with half a million annotated fingerprints which can be used to train deep-learning models. I also help to build a dataset containing images of grocery floor hazards.

Future Work

Explainable face and fingerprint matching via improved localization

Deep learning-based fingerprint, face, and iris recognition systems suffer from the class overlap problem and produce inaccurate matches. Class overlap refers to the situation where two or more individuals have similar or overlapping biometric features such as ridges, slopes, and valley patterns for fingerprints and doppelgangers for faces. **My plan is to use discriminative localization in such scenarios to distinguish features and highlight regions of a biometric image that are most relevant to a particular prediction task, such as biometric matching.**

Deep learning for data mining on edge devices

Due to the explosive growth of wireless communication technology, the number of Internet of Things (IoT) devices has increased dramatically in recent years. These resource-constrained IoT devices, such as sensors and actuators, have produced large quantities of data in real time, which is an appealing target for data mining and AI systems. Deep learning algorithms are capable of automatically learning complex patterns and relationships within data, which makes them ideal for data mining tasks. However, deploying deep learning models on such end devices is nearly impossible. **My plan involves the continuation of my research on deploying deep learning in resource-constrained environments [tVG+20, tCKH20, VVtH20]. Additionally, I aim to develop novel resource-aware deep-learning algorithms that can be used for analyzing unstructured data, including text and images.** Specifically, I am interested in exploring the area of deep learning and edge computing. By combining these two areas, I am confident that we can develop powerful data mining systems that can help us better understand complex data and make better decisions.

Detecting liveness of fingerprint images

Liveness detection of a fingerprint using a mobile device is a challenging problem. Depth measurement plays a key role to determine if an image is real or fake. Image quality analysis is another key factor in liveness detection. The quality of the image can be analyzed using various metrics, such as image resolution, contrast, and noise levels. My current research on contactless fingerprint segmentation using deep learning will help me in this area. **I will continue my research in this domain to introduce new methods for detecting the liveness of fingerprints.**

Road health measuring using deep learning

Surface cracks are the earliest sign of road distress which could lead to more serious problems if left unaddressed. Hence, the distress detection, classification, and quantification of surface cracks are essential steps in the road management system. Nowadays, traditional manual inspection methods have been replaced by computer vision techniques. Although these techniques are generally superior, they struggle to detect thin, irregular dark-lined cracks buried into the textured backgrounds. In the collaboration with Sunny Canton university, I published a paper describing how hybrid deep-learning models can be used for analyzing road images. **I aim to continue this research to improve automated road health detection systems with the help of deep learning-based algorithms.**

Deep learning for cybersecurity

Cyber security is extremely important in today's digital world because everything we do is connected to the Internet and other digital devices. Many industries are still vulnerable to various cyber attacks such as fraudulent authentication, malware attacks, phishing attacks, and zero-day exploits. Such vulnerability has a tremendous negative economic impact on any institution. Deep learning can be used for biometric authentication, such as facial recognition or fingerprint recognition, which can help prevent unauthorized access to systems and data. **My plan is to continue my current research on fingerprint recognition systems with modern deep learning algorithms and make the system efficient enough to run on resource-constrained devices like Raspberry Pi and IoT devices.**

References

[tBSH23] **M. G. Sarwar Murshed**, Keivan Bahmani, Stephanie Schuckers, and Faraz Hussain. Deep age-invariant fingerprint segmentation system. arXiv, 2023.

[tCKH20] **M. G. Sarwar Murshed**, James J Carroll, Nazar Khan, and Faraz Hussain. Resource-aware on-device deep learning for supermarket hazard detection. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 871–876. IEEE, 2020.

[tCKH22] **M. G. Sarwar Murshed**, James J Carroll, Nazar Khan, and Faraz Hussain. Efficient deployment of deep learning models on autonomous robots in the ros environment. *Deep Learning Applications, Volume 3*, pages 215–243, 2022.

[tKB+21] **M. G. Sarwar Murshed**, Robert Kline, Keivan Bahmani, Faraz Hussain, and Stephanie Schuckers. Deep slap fingerprint segmentation for juveniles and adults. In *2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4, 2021.

[tMH+21] **M. G. Sarwar Murshed**, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain. Machine learning at the network edge: A survey. *ACM Computing Surveys (CSUR)*, 54(8):1–37, 2021.

[tVG+20] **M. G. Sarwar Murshed**, Edward Verenich, Conrad Gende, James J Carroll, Nazar Khan, and Faraz Hussain. Hazard detection in supermarkets using deep learning on the edge. In *3rd USENIX Workshop on Hot Topics in Edge Computing*. USENIX, 2020.

[VtK+21] Edward Verenich, **M. G. Sarwar Murshed**, Nazar Khan, Alvaro Velasquez, and Faraz Hussain. Mitigating the class overlap problem in discriminative localization: Covid-19 and pneumonia case study. *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications*, pages 125–151, 2021.

[VVtH20] Edward Verenich, Alvaro Velasquez, **M. G. Sarwar Murshed**, and Faraz Hussain. Flexserve: Deployment of pytorch models as flexible rest endpoints. In *2020 USENIX Conference on Operational Machine Learning.* USENIX, 2020.

[ZtHE20] Baogang Zhang, **M. G. Sarwar Murshed**, Faraz Hussain, and Rickard Ewetz. Fast resilient-aware data layout organization for resistive computing systems. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 72–77. IEEE, 2020.