Attention Based Multi-Sensor Multi-Task Fusion for Behavior Understanding

Motivation: In 2020, 113,500 people in the US died via preventable, in-home deaths[1]. An in-home monitoring system reporting emergencies could have prevented many deaths, especially during the COVID-19 pandemic when patients were denied care due to overcapacity at medical centers. Current in-home technology is limited to digital assistants, single-task robots, and IoT appliances (e.g., Alexa, Roomba, Nest), yet no solution effectively monitors human behavior and reports anomalies. This is particularly challenging because a human can easily distinguish an emergency **(Fig. 1a)** from a visually similar image **(Fig. 1b)**, but a computer vision (CV) model cannot. Even training a specialized model to detect human falling instances would not support generalizable predictions across other medical emergencies. Integrating data from multiple CV tasks and sensors into one model may enable the enhanced understanding of behavior required for this complex task.

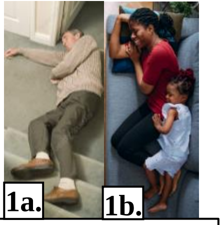
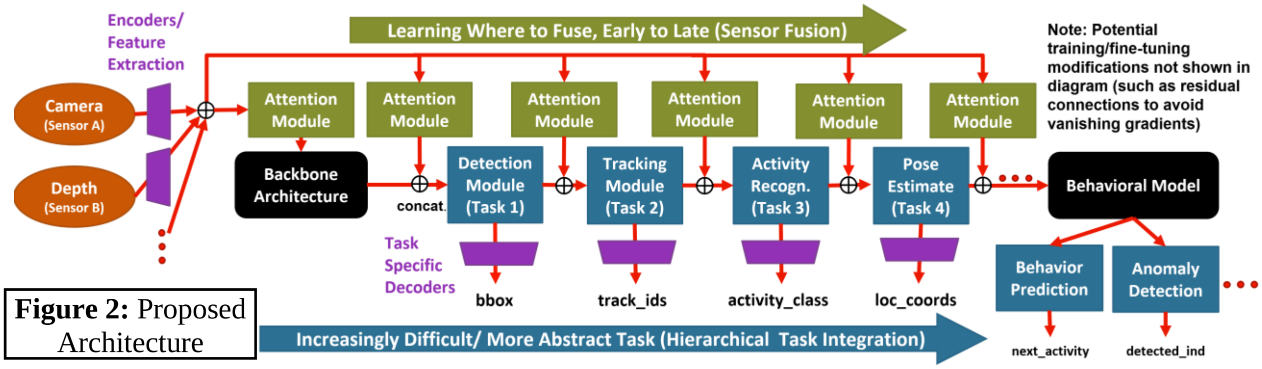


Figure 1: Emergency Detection

State-of-the-art (SOTA) CV performs well on basic tasks (e.g., object detection) but struggles to synthesize basic tasks to perform higher level reasoning (e.g., scene understanding, behavior prediction). Furthermore, integrating sensor data (sensor fusion) is currently underutilized in machine learning. For the discrepancy in **Fig. 1**, a neural network (NN) would benefit from using depth and thermal data to characterize the abnormal body angle and temperature, distinguishing the person in crisis. Enabled by sensor fusion, in-home assistive technology can be equipped to reliably alert authorities of an emergency.

Background: Behavior prediction is currently studied in autonomous driving and robotics (e.g., collision avoidance, robot collaboration), but is less studied for human monitoring. Similarly, autonomous systems fuse camera, lidar and motion sensors (e.g., SLAM, pose estimation), but IoT edge devices rarely use sensor fusion. Sensor data is often combined at a fixed point in the NN (early, middle, or late) but few works experiment with *how* to fuse the data for *various* tasks. Transfer learning and few shot learning adapt a baseline model to different tasks/data with few training samples, and multi-task learning exploits useful information in related tasks for better performance. There are limited works on integrating concepts from sensor fusion to create a new technique of understanding different data for different tasks. Tuning how sensor data is fused may allow a baseline model's output representation to be more effective in downstream tasks, as each task may benefit uniquely from distinct sensor inputs. Foundation models are general multi-modal AI models trained on vast amounts of data in a self-supervised way to obtain a general understanding of the data[2]. The most well-established foundation models are language-based, but there exists emerging interest to extend this focus to CV (e.g., CLIP, iGPT). However, the multimodal capabilities of these models are limited to speech, text and RGB vision; they fail to exploit multimodal hardware sensor data (thermal, depth, motion, etc.) and have not been applied to an embedded system such as a smart home device.

Research Plan: My research goal is to better understand how different forms of perceiving the environment (sensor data) can combine and aid in AI sub-tasks to build a model that reasons about human behavior. The objectives of this research are to 1) evaluate the effectiveness of self-attention mechanisms on sensor fusion, 2) understand how decomposing a higher-level task (behavior understanding) to lower vision tasks (object tracking, action recognition, etc.) could aid the construction of a model, and 3) understand how such a system could be implemented in the smart home setting. Attention mechanisms, widely used in transformers[3], weigh embeddings in relation to each other and using different attention



Note: Potential training/fine-tuning modifications not shown in diagram (such as residual connections to avoid vanishing gradients)

Figure 2: Proposed Architecture

modules before each task would allow a network to weigh embeddings differently per task i.e., learn where to use which sensor data. Furthermore, data representations used for basic CV tasks should be reused by more difficult abstract tasks (i.e., the same model that detects a human should further be able to identify them in a crowd, track them, and recognize their activity simply by adding more layers to the previous task). I hypothesize that this attention-based sensor fusion and hierarchical task integration **(Fig. 2)** will result in a fundamental behavior model that will outperform current fusion methods and language-vision based foundation models in a smart home setting. Performance will be measured by the accuracy metrics of downstream tasks, ability to adapt to novel data/tasks, and computational resources needed.

Methods: **1)** ***Sensor Level Fusion:*** My research lab has deployed prototype smart home devices running separate detection and activity recognition models with an RGB camera. I propose to integrate depth and thermal sensors into our existing model and measure accuracy improvements over the camera-only model and public multi-sensor models[4] on open-source datasets including our lab's smart home dataset. **2)** ***Task Level Integration:*** Following the design in **Fig. 2**, starting with detection, I will individually add the task modules, changing the modules' parameters and architectures accordingly to maintain competitive task specific accuracies. Furthermore, following the extrinsic evaluation paradigm[2] of foundation models, I will compare my model's task specific accuracies and computation requirements with SOTA models. **3)** ***Smart Home Constraint Optimization: a) Security and Privacy:*** For privacy, I will deidentify training data (e.g., blurring or contour extraction). For security, I will only transfer irreversible data to a server for computation, i.e., data after a nonlinear activation in the NN. If these constraints hinder accuracy, the model and the deidentifying approach can be tuned accordingly. A potential extension of this work could include evaluating the effectiveness of this approach through legal and community engagement (e.g., surveys measuring the smart home user's comfort with this technology). ***b) Resource Utilization:*** I will profile the model and its components against common metrics (e.g., latency, memory, power, etc.), use embedded AI tools (e.g., TensorRT) to optimize the NN in software, and consider heterogeneous computing for system level design, i.e., which components of the model should run on the CPU, FPGA, GPU, or the cloud.

Intellectual Merit: This research will bridge the gap between sensor fusion and multipurpose AI models to perform higher level tasks, contributing knowledge to robotics, graphics (e.g., virtual/augmented avatar engagement), and philosophy (e.g., theory of mind in robots). The hierarchical multi-task AI system I have devised in **Fig. 2** seamlessly provides points of interpretability between tasks, allowing for robust, accountable, trustworthy AI. Modeling human behavior can also help understand 1) sociological or psychological phenomena to diagnose patients in medicine 2) consumer behavior to market products in business 3) energy usage in grid/renewables planning. Moreover, my work will advance current understanding of how large AI models can be deployed on hybrid cloud/edge environments. Physical experiments in a smart home will provide unique insights on design and practical constraints on how AI technology can be implemented in society that were underexplored by previous research efforts.

Broader Impacts: Smart home technology that processes behavioral data can enable safety and comfort, especially for those who society typically neglects to design for. The disabled and elderly could be assisted and monitored around the house independently, and patient care centers could also benefit from the extra help. Furthermore, understanding human consumption behavior can help optimize household appliances accordingly and provide data to public officials to design efficient energy infrastructure to combat climate change. Additionally, studying robust sensor fusion would allow autonomous systems to perceive their environment more accurately. Robots could aid agriculture and manufacturing giving the US technological leadership, and autonomous vehicles may decrease accidents and taxpayer infrastructure expenses. As an informed scientist eager to make a difference, I will take steps to bring these impacts to life by engaging with UIUC's Office of Access and Equity about my work, collaborating with our industry smart home partners (Foxconn Interconnect and Belkin) about the feasibility of such a device, open sourcing my work, and engaging with related disciplines and applications of sensor fusion technology.

References: [1] National Safety Council, "NSC Deaths in Home" https://injuryfacts.nsc.org/home-and-community/deaths-in-the-home/introduction/data-details/ [2] Bommasani, R., et al. "On the opportunities and risks of foundation models." arXiv:2108.07258 (2021). [3] Vaswani, A., et. al. Attention is all you need. NeurIPS. (2017). 30. [4] Chen, C., et. al. 2015 IEEE Internat. Conf. on Img. Proc. (2015). 168-172