5-2024

Pedestrian Pathing Prediction Using Complex Contextual Behavioral Data in High Foot Traffic Settings

Laurel Bingham
Utah State University, laurel.bingham@usu.edu

PEDESTRIAN PATHING PREDICTION USING COMPLEX CONTEXTUAL BEHAVIORAL DATA IN HIGH FOOT TRAFFIC SETTINGS

by

Laurel Bingham

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

_____	_____
Mario Harper, Ph.D.	Vladimir Kulyukin, Ph.D.
Major Professor	Committee Member
_____	_____
Shuhan Yuan, Ph.D.	D. Richard Cutler, Ph.D.
Committee Member	Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2024

ABSTRACT

Pedestrian Pathing Prediction Using Complex Contextual Behavioral Data in High Foot Traffic Settings

by

Laurel Bingham, Master of Science

Utah State University, 2024

Major Professor: Dr. Mario Harper
Department: Computer Science

Predicting the movement and intention of pedestrians in real life settings is one of the many great challenges researchers must address while developing fully autonomous vehicles. The earlier, and more accurately a pedestrian's intent to cross in front of a given vehicle is recognized, the pedestrian's movements can be accommodated more safely by the vehicle. While a pedestrian's line of motion can often be computed based solely on their previous motion- pedestrians can make sudden or erratic choices depending on the external factors at play.

This project examines Honda's TITAN dataset, which contains 700 unique clips from moving vehicles. This dataset shows various scenes in Tokyo, Japan, in areas with high foot traffic. Each person and vehicle in each scene has five human labeled contextual tags, with attributes such as age, motion status, or communicative actions. The dataset also includes the vehicle speed and trajectory information from the vehicle that filmed the clip. This project examines the impact of these contextual clues on model accuracy.

For this purpose, an LSTM was then trained on various combinations of this contextual data, in addition to basic bounding box coordinates. From the best of these models, their MSE for each prediction was used to train decision trees- which attempted to classify which, if any, pieces of the data consistently resulted in high or low error. This work ultimately suggests which pieces of contextual behavioral data are most important to classify for in a pedestrian pathing prediction setting.

(49 pages)

PUBLIC ABSTRACT

Pedestrian Pathing Prediction Using Complex Contextual Data in High Foot Traffic Settings

Laurel Bingham

Ensuring the safe integration of autonomous vehicles into real-world environments requires a comprehensive understanding of pedestrian behavior. This study addresses the challenge of predicting the movement and crossing intentions of pedestrians, a crucial aspect in the development of fully autonomous vehicles.

The research focuses on leveraging Honda's TITAN dataset, comprising 700 unique clips captured by moving vehicles in high-foot-traffic areas of Tokyo, Japan. Each clip provides detailed contextual information, including human-labeled tags for individuals and vehicles, encompassing attributes such as age, motion status, and communicative actions. Long Short-Term Memory (LSTM) networks were employed and trained on various combinations of contextual data, along with basic bounding box coordinates. The best-performing models were identified based on mean squared error (MSE) for each prediction. Subsequently, decision trees were trained using the MSE data to classify the contextual features that consistently contributed to high or low prediction errors.

This project sheds light on the significance of contextual behavioral data in predicting pedestrian motion and intention. By analyzing the impact of age, motion status, communicative actions, and other factors on prediction accuracy, the study offers valuable insights into the key elements that autonomous vehicles should consider when

anticipating pedestrian movements in real-world settings. Ultimately, this research contributes to advancing the development of robust and safe autonomous vehicle systems by identifying crucial contextual cues for accurate pedestrian pathing predictions.

ACKNOWLEDGMENTS

I would like to thank Dr. Mario Harper for making available to me the Honda Research Institute's (HRI) TITAN data set for the research in this thesis. I would especially like to thank my committee members, Drs. Mario Harper, Shuhan Yuan, and Vladimir Kulyukin, for their support and assistance through- out the entire process.

I would like to give thanks to my family, friends, and colleagues for their encouragement, moral support, and patience as I have worked towards finishing this document. I would like to give special thanks to my mother and my brother, who have supported me wholeheartedly as I've worked these past two years on my masters degree. I could not have done it without all of you.

Laurel Bingham

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As automatic assistive features on automobiles have become more advanced, the push to develop fully autonomous vehicles has grown substantially. One of the greatest challenges in autonomous driving is interpreting and predicting the actions of observed cars and pedestrians on the road. Proactively reacting to take preventative measures, such as slowing down or stopping is imperative to preventing dangerous or life threatening situations, particularly when reacting to pedestrian motion.

The actual task of pedestrian trajectory prediction can be broken down into three major steps. First, sensory data is gathered from the vehicle. In the simplest cases, this consists of a dashboard camera, but can include multiple cameras, or even vehicle telemetry, such as the vehicle's heading or speed. From this point, the images received from the camera are passed through some form of object detection. This is where key information is extracted from the image- namely drawing bounding boxes around humans and vehicles visible in the scene. This is also where key environmental features, such as weather, the presence of crosswalks, or traffic signs are identified. This information is then stored in a frame by frame sequence. Finally, the third step of trajectory prediction is to use k frames of this sequence to predict position or path of each identified pedestrian in the sequence, typically with a time series model through an LSTM, RNN, or GRU.

Balancing which features are detected and considered when forming a predictive model is one of the greatest challenges in designing a model that can predict quickly and

accurately. While only using bounding box coordinates has shown some success [1][2], most models incorporate additional contextual data from the scene to gain additional accuracy. However because pedestrian pathing prediction is a model which runs in real time, taking in only a few seconds of data and predicting the next few seconds- the prediction model must be fast. Attempting to identify too many features in a frame can make the prediction useless- if the second or two it predicts has passed while the model was processing the frames. While environmental factors such as the presence of a crosswalk or traffic control signs are the most common supplement to a model's input [3], the recently released TITAN dataset offers a rich amount of pedestrian centric contextual labeling that it purports to be valuable.

This work dives into the TITAN dataset, focusing specifically on identifying which of the behavioral annotations are both feasible to train on with the given dataset, and meaningfully impactful on the accuracy of predicting pedestrian motion. The dataset is dissected on a frame by frame basis, with an emphasis on discovering how frequently highly specific contextual actions are present across all clips.

The following chapter of this work will act as a general overview of the current trends in pedestrian modeling and prediction. It describes common model constructions, as well as common inputs and methods for framing the problem of pedestrian prediction. The most common datasets used for training these models are also discussed, emphasizing identifying each dataset's strengths and weaknesses. The next chapter dives into the TITAN dataset specifically. The number of pedestrian sequences, the length of their presence, and the distribution of each label are analyzed in greater depth. How this dataset was transformed to be read by an LSTM is also discussed. Following this, a

chapter that discusses the experimental methods used by this work is included. This is where the various choices in model construction are discussed. Following that description is a chapter on this paper's major findings. This includes comparisons of each trained model, as well as the output decision trees developed to find behavioral tags that consistently resulted in high error. Finally, these results are discussed, with this paper’s final recommendations on which pedestrian behaviors are most important to identify to improve model accuracy.

CHAPTER 2

LITERATURE REVIEW

Pedestrian intent prediction is a multistage process, which begins with identifying where in an image a pedestrian is located. Typical approaches to the problem of object detection utilize variants of the CNN, or convolutional neural network to detect key features. However, the majority of studies in the field of pedestrian intention prediction have utilized adapted versions of the YOLO algorithm to quickly look once over an image [4][5][6]. The YOLO approach is preferable due to its quick runtime and high accuracy. Its major advantage over other image detection algorithms, such as the Fast R-CNN, is that it is able to accomplish a detection task with only one pass through the network, without relying on sliding windows or repeated passes over the image. [7]. The output from a YOLO algorithm is a labeled, normalized bounding box around each identified class in the image.

Despite the speed the YOLO algorithm offers, some pedestrian detection algorithms have sought to preserve deeper context of the pedestrian by utilizing pose estimation. Tools such as Openpose can provide up to 25 key skeleton points on a human body- offering a much more detailed representation of pedestrians on the screen [8]. Several works have examined including pose estimation in intent prediction algorithms[8][9], however the focus of these works is on creating suitable training datasets, and on measuring pose estimation accuracy, rather than examining the impact of pose estimation on pedestrian intention prediction accuracy.

Due to the visual elements in pedestrian intention prediction, sources of training

data are critical to assessing and training new algorithms. The following section examines the currently available pedestrian intention prediction datasets. Through these datasets, key insights can be gained on which features are considered universally important, as well as where key factors may be ignored or under-labeled. A paper published in 2022, for example, highlights the bias of all pedestrian datasets towards highly developed, western countries, noting that pedestrian behavior in the streets of New Delhi could not be predicted well with the current datasets [10]. Thus, understanding the construction and limitations of pedestrian datasets is very important when considering the application of prediction algorithms in diverse settings.

2.1 Joint Attention in Autonomous Driving (JAAD) Dataset

The JAAD dataset was developed in 2017, with the stated purpose of focusing on "Pedestrian and driver behaviors at the point of crossing and factors that influence them" [11]. This dataset provides 346 short video clips, each around five to ten seconds in length. They are filmed across both North American and European cities, and feature a variety of contextual annotations. These include weather and location tags, behavioral labels, such as walking or looking, demographic information, such as age or gender, and visible traffic elements in the scene. Each Pedestrian is also identified with both a bounding box and an occlusion tag.

While the contextual information provided in the JAAD dataset is extensive, the JAAD dataset notably has few frames where pedestrians show intention to cross the street. Of the 82,032 annotated frames in the dataset, there are only 686 pedestrians with behavior annotations of crossing or not crossing. Of those 686 pedestrians, 495 cross the

surveystreet, while 191 do not cross [11]. In terms of machine learning, the relatively small sample of pedestrians should be kept in mind when attempting to develop generalizations around pedestrian behavior.

2.2 Pedestrian Intention Estimation (PIE) Dataset

The PIE dataset was released in 2019, and advertises itself as the largest publicly available dataset for studying pedestrian behavior in traffic. [12] . This dataset provides information about the vehicle that recorded the annotated traffic scenes in addition to the annotated clips. Namely, it recorded the vehicle speed, heading, direction, and GPS coordinates. Much like the JAAD dataset, pedestrian behavioral annotations were provided. Additionally, the dataset includes intention annotations. These intention annotations are the result of a "Large scale human experiment to determine early crossing intention of the 1842 annotated pedestrians" [12]. The output of this experiment is the aggregated intention probability from the surveyed human subjects.

Unlike the JAAD Dataset, the PIE dataset consists of much larger segments of footage. Six hours of 30 FPS video footage is included in ten minute chunks. A total of 909,480 frames are provided, with 293, 437 of them being annotated. Across this footage, there are 1842 pedestrians with behavioral annotations. This breaks down further into 519 who intended to cross, and did cross, 894 who intended to cross and did not cross, and 429 who did not intend to cross [12].

A notable limitation of this dataset is that it was gathered in a single North American city, Toronto. Another important consideration with this dataset is that approximately two thirds of the provided frames are not annotated, which should be kept in mind when

training on this dataset.

2.3 Stanford-TRI Intent Prediction (STIP) Dataset

The STIP dataset was released in 2020 as a collaboration between Stanford University and the Toyota Research Institute. This dataset consists of annotated sequences of three cameras- mounted on the front, left, and right of the vehicle. The dataset consists of approximately 15 hours of 20 FPS footage. Approximately 350,000 pedestrian bounding boxes are manually annotated at a rate of 2 FPS [13]. The other frames have interpolated annotations using instance segmentation results and a JPDA (Joint Probabilistic Data Association) based tracker [13].

Key aspects of this dataset are its wide field of view provided by the multiple image feeds. It additionally provides pedestrian bounding boxes, as well as a ‘crossing’ and ‘not crossing’ label. However, it does not provide information about the movement of the recording vehicle, or notable traffic feature bounding boxes like the previous PIE and JAAD datasets. While the dataset provides fewer ground truths to train on than previous datasets, the wide array of camera feeds offer an interesting perspective and additional context for pedestrian trajectory prediction.

2.4 Trajectory Inference using Targeted Action priors Network (TITAN) Dataset

The TITAN dataset was released in 2020 by the Honda Research Institute. All footage was gathered in Tokyo, Japan, “ from a moving vehicle on highly interactive urban traffic scenes” [14]. The dataset provides the bounding boxes of all pedestrians and vehicles in each scene, vehicle telemetry information, and rich contextual labels

associated with each pedestrian. The TITAN dataset is unique, in that it has developed 5 overarching categories- Atomic Actions, Simple Contextual Actions, Complex Contextual Actions, Communicative Actions, and Pedestrian Age, by which each pedestrian was assigned a label in each category [14]. The dataset contains 700 clips of 10 to 20 second duration, which are annotated at a rate of 10 HZ. The dataset contains in total 75,262 frames with 395,770 persons, 146,840 4-wheeled vehicles, and 102,774 2-wheeled vehicles [14]. Notably, this implies an average of five pedestrians in any given frame.

One of the greater limitations to this dataset is that it is not immediately publicly available. However, university researchers can apply to gain access to this dataset. Other important considerations when using this dataset are the frequencies of provided labels. Some of the more unique labels, such as ‘jumping’ or ‘lying down’ occur in only a double digit number of frames, while more universal actions, such as walking, occur over three hundred thousand times in the dataset.

2.5 Common Network Inputs for Pedestrian Intention Prediction

Most models for intent prediction utilizes the pedestrian location. This is most typically represented by a bounding box, but can also take the form of a single point of coordinates of the form (x,y) [3]. Many recent papers, especially those which view pedestrian intent prediction from a multitask perspective, only utilize pedestrian location as an input feature[1][2]. Other pedestrian centric-inputs can come from pose estimation, which can help identify changes in a pedestrian's action, such as walking to stopping [3]. Finally, human assigned behavioral annotations are frequently included as input. This can take the

form of gaze direction, body orientation, or state of movement [3].

2.5.1 Common Environmental Network Inputs

There are two major approaches to tagging and utilizing environmental context in pedestrian intention algorithms. One involves finding the local environmental context by cropping the image around a pedestrian's bounding box- and identifying key nearby features, such as a crosswalk under the pedestrian, or the presence of other nearby pedestrians or cars. The second method regards global features, such as the number of car lanes, or the presence of traffic signs or lights [3].

Additionally, some networks account for the position or speed of the vehicle. Some datasets provide this information natively, while other datasets require additional calculation to estimate the speed of the filming vehicle [3]. In either case, vehicle speed and the estimated distance between the vehicle and passenger can be important inputs. Pedestrians are more likely to cross if they believe they have a lead time of 3 to 7 seconds before collision. Further, studies have shown that somewhat unintuitively, pedestrians are more likely to attempt to cross with smaller lead times if oncoming vehicles are moving faster [15].

2.6 Common Architectures

The architecture of a pedestrian intention prediction algorithm can be divided into two phases. The first phase consists of processing the camera feed into a time series object representing a series of pedestrian positions, along with other identified contextual data. The second phase consists of a time series model, typically an LSTM [3], which predicts

the next k frames of the scene. These predictions rarely exceed a timeframe of 1 to 3 seconds. In addition to LSTMs, researchers have applied RNNs and GRUs to the problem, all with relatively similar reported performances. Frequently, these time series models are layered on each other, with a final fully connected layer to predict [3]. Another common construction is an encoder-decoder style LSTM or RNN. For image processing, CNN's are a popular choice, though depending on the type of data being gathered for the time series model, these preliminary feature extraction models can be varied. Notably, adding CNN style components can add an interpreted spatial aspect to the time series prediction, allowing a model to predict across both time and space.

2.7 Common Network Outputs

The actual prediction from a pedestrian intention prediction algorithm is frequently framed in one of two ways. The first is a binary classification problem, where the question asked is simply, 'will the pedestrian cross?'. The second framing is a question of the pedestrian's expected trajectory. Many models ask the question, "Where on the screen will the pedestrian be in k time steps?". Many models answer both questions, first providing a yes/no or probability to the first question, and then provide a region or coordinates that the pedestrian is expected to move to. Some papers have additional outputs, which may take the form of pose estimation, predicted type of action, or behavioral tags. However these outputs are far less common[3].

CHAPTER 3

TITAN DATASET DESCRIPTION AND ANALYSIS

The TITAN dataset is leveraged in this study due to the contextual pedestrian information. As the TITAN dataset is provided under NDA with Honda Research Institute, specifics may not be disclosed. However, the data is described in some detail.

3.1 Examining TITAN Dataset Construction

Firstly, Table 1 shows what the TITAN dataset considers a significant piece of pedestrian contextual data.

Age of Pedestrian	**Atomic Actions**	**Communicative Actions**	**Complex Contextual Actions**	**Simple Contextual Actions**
Adult Child Senior over 65	Bending Jumping Laying Down None of the Above Running Sitting Squatting Standing Walking	Looking into phone None of the above Talking in group Talking on phone	None of the above Loading Unloading Getting on 2 wheeled vehicle Getting on 4 wheeled vehicle Getting off of 2 wheeled vehicle Getting out of 4 wheeled vehicle	Biking Cleaning an Object Closing Crossing street at pedestrian crossing Entering a building Exiting a building Jaywalking Motorcycling None of the above Opening Waiting to cross the street Walking on the road Walking along side of road

Table 1: Pedestrian Contextual Data Categories Provided by the TITAN Dataset

As seen above, the contextual data is broken into five major categories. Atomic Actions consist of basic movement and position actions, such as walking, sitting, or squatting. Age is divided into three major categories, Adult, Child, and Senior over 65.

Communicative actions involve pedestrian attention and communication, indicating if they're interacting with a group, looking at their phone, or otherwise speaking. Complex Contextual actions are actions which involve pedestrians interacting with two or four wheeled vehicles. This includes loading or unloading these vehicles, or getting in or out of them. Finally, Simple Contextual actions are actions which broadly include pedestrians interacting with the environment around them. This is where street crossing intention is noted, divided into legal and illegal crossing intents. Further this category includes categories for entering and exiting buildings, as well as behaviors like walking in or on the side of the road. Shown below is a sample frame of the dataset, with the bounding boxes of each vehicle or pedestrian shown in red.

Figure 1: Image from the Titan Dataset

3.2 Examining Pedestrian Presence Duration across the Dataset

The TITAN dataset features 786 clips, with csv files that track each pedestrian or vehicle on a frame by frame basis. Each object in the frame is assigned a unique object tracking id the first time it appears on screen, which persists until the object leaves the field of

view. Figure 2 below shows the calculated length of time of each Pedestrian's presence captured by the cameras during data collection. These statistics guided the sequence length used for training the various LSTM tested in this research.

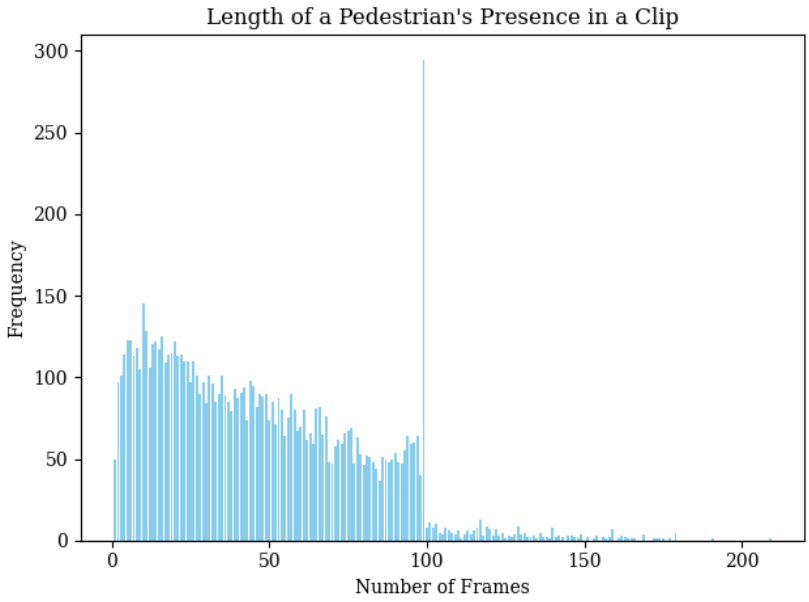


Figure 2: Frequency of Length of a Pedestrian's Presence

As shown above, it is important to note the high density of pedestrians that appear in low single and double-digit numbers of frames. The peak at 100 frames occurs because many of the clips end at exactly 100 frames, with several pedestrians that were in view for the entire clip. Of the 8588 uniquely identified pedestrians, the calculated average duration of their presence was 46.49 frames. Further, 1565 of these pedestrians were on screen for less than 20 frames. The shortest pedestrian appearance in a clip was one frame, and the longest pedestrian presence was 209 frames.

3.3 Examining Distribution of Pedestrian Contextual Labels

While the TITAN dataset provides a wide variety of behavioral and contextual tags, it is

important to understand the frequency by which each label occurs. Within each category, there is a notable imbalance in the dataset, with one category significantly outweighing the others. This section explores those distributions, and makes recommendations on which categories are sufficiently populated to train a classifier.

3.3.1 Distribution of Age Across All Frames

As can be seen below in Figure 3, the adult category is by far the most prevalent.

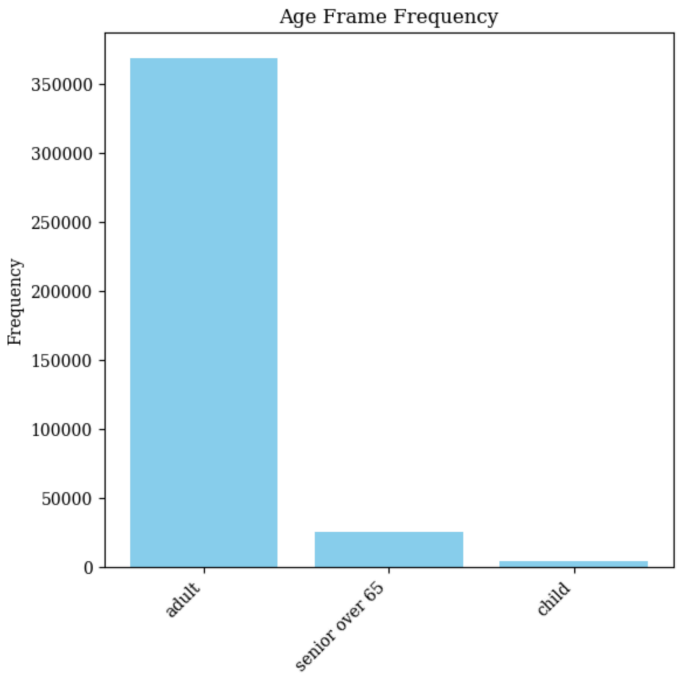


Figure 3: Distribution of Age Across all Frames

Adults appear in nearly 369,000 frames, while seniors appear in only 26,000 frames, with children appearing in just under 5,000 frames. Though 5,000 is a small number of frames, this dataset is sufficiently sized to train a classifier on, provided that the training dataset was balanced around the smallest category in the dataset.

3.3.2 Distribution of Atomic Actions Across All Frames

Of all behavioral categories in the TITAN dataset, Atomic Actions are one of only two categories where a named action is the most frequent label. In this case, the most frequent action by a wide margin, as can be seen in Figure 4, is walking.

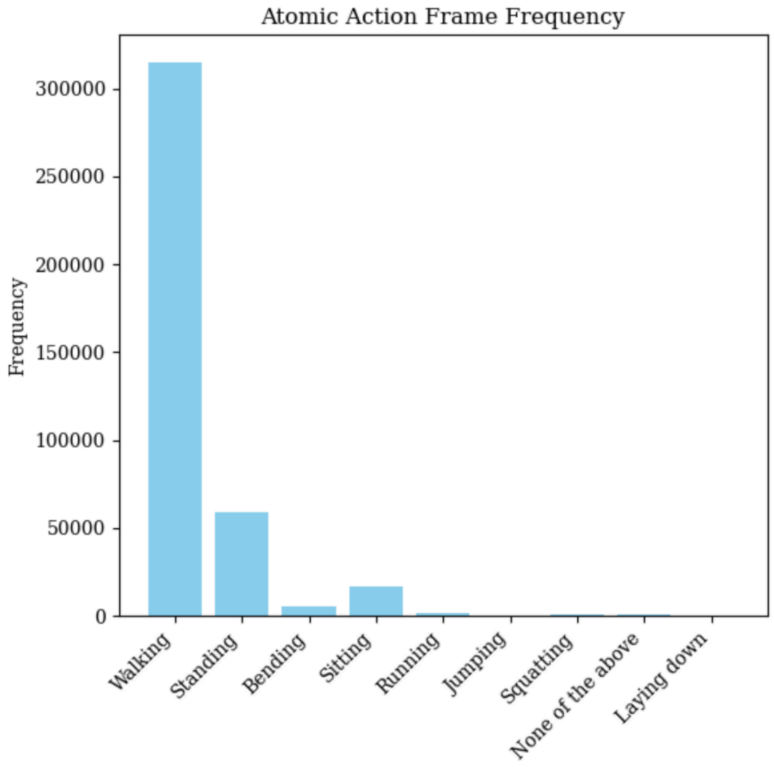


Figure 4: Distribution of Atomic Actions Across All Frames

The actions standing and sitting also occur in a significant number of frames, with standing appearing in approximately 59,000 frames, while sitting occurs in approximately 16,800 frames. Notably, the actions, jumping, laying down, squatting and none of the above occur in fewer than 1000 frames, with laying down and jumping occurring in fewer than 100. It is therefore not feasible to train a model to classify all nine of these actions from this dataset alone. However, the most common actions are well represented.

3.3.3 Distribution of Communicative Actions Across All Frames

As will be the case with most categories going forward, the classification, none of the above is the most frequent classification by a large margin. This is expected, as the categories going forward refer to much more specific attributes. However, as Figure 5 shows, even though "None of the Above" outweighs other categories by a significant margin, all categories are well represented in the dataset.

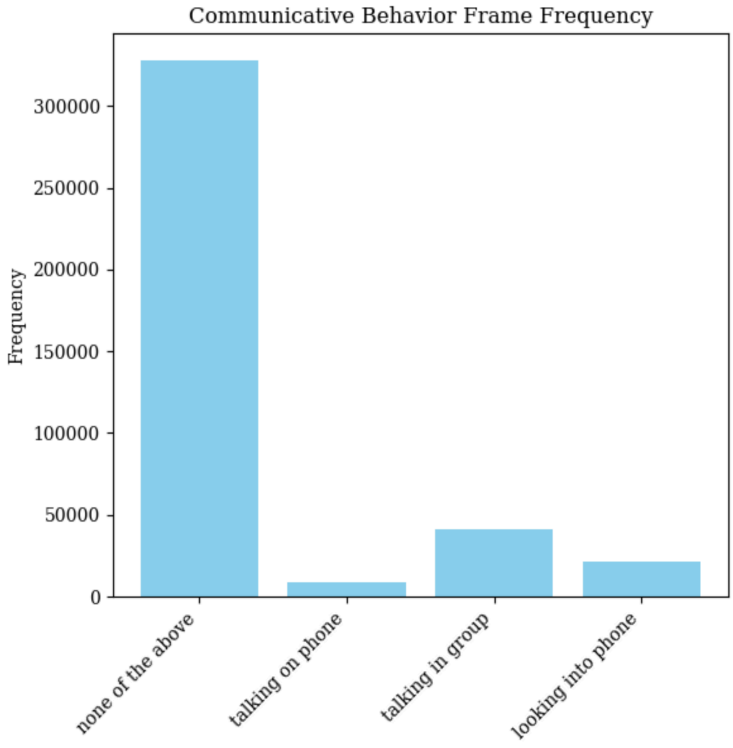


Figure 5: Distribution of Communicative Actions Across All Frames

Talking in a group is the most frequent meaningful label, with approximately 42,200 frames. Looking into a phone and talking on a phone both occur in 21,800 frames and 8500 frames respectively. Though it would require balancing, this dataset is sufficiently large enough to train a classifier on.

3.3.4 Distribution of Complex Contextual Actions Across All Frames

Of the categories, complex contextual data is the most imbalanced. It is important to note that Figure 6 is shown in a log scale to give the other categories visibility on the chart.

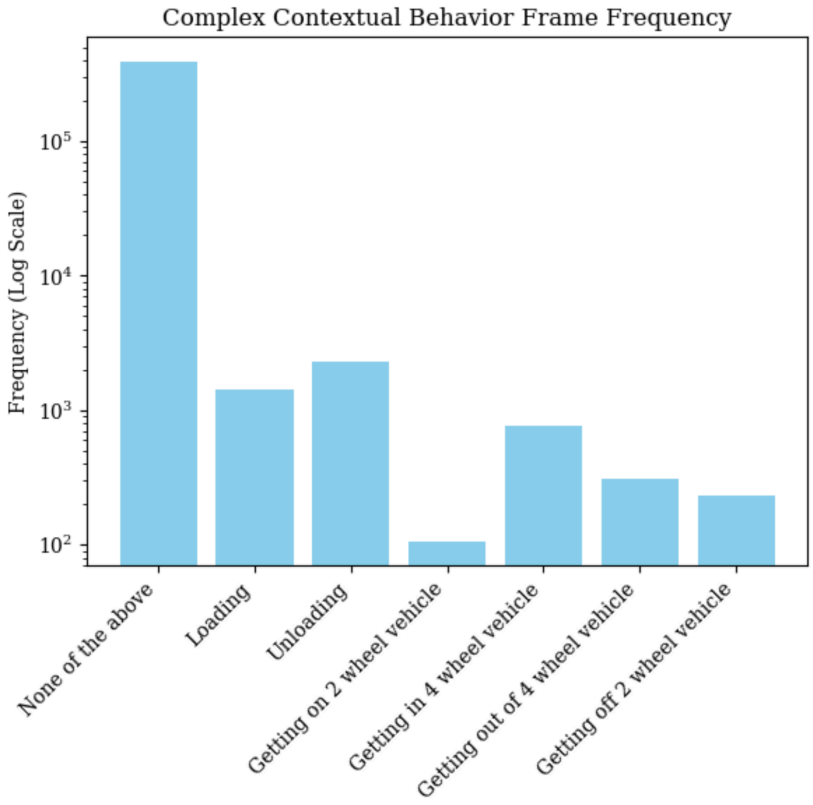


Figure 6: Distribution of Complex Contextual Actions Across All Frames. Note that this is a log-axis.

None of the above occurs in 394100 frames. However, the next most frequent tag, unloading, only occurs in 2300 frames. Loading occurs in 1400, while all other tags occur in only a few hundred frames. Excluding none of the above, the dataset is small and highly specialized. It is not an ideal dataset to train a classifier on, as there are so few examples of each specialized tag.

3.3.5 Distribution of Simple Contextual Actions Across All Frames

Simple contextual actions is the other behavioral category which has a category other than none of the above as the most common tag. As can be seen in Figure 7, both walking along the side of the road and walking on the road occur more frequently than none of the above, with well over 100,000 frames with each tag.

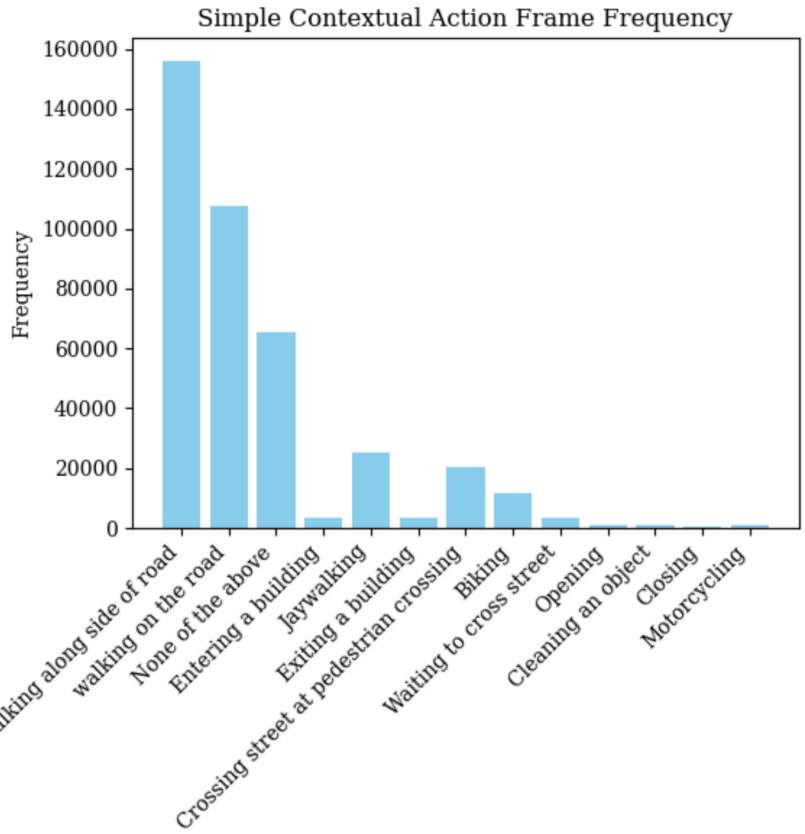


Figure 7: Distribution of Simple Contextual Actions Across All Frames

Also occurring in a significant number of frames is jaywalking, at nearly 25,000 frames, crossing at a pedestrian crossing at 20,000 frames, and biking at nearly 12,000 frames. While this category has some sparsely populated categories, such as opening or closing, which occur in under 1,000 frames, the majority of categories in this dataset are large enough to train a classifier on.

3.4 Dataset Preprocessing and Transformation

Most implementations of an LSTM require both a constant, known number of input features, as well as a constant sequence length. To facilitate this, each csv file provided by the TITAN dataset was split into subsets equal to the number of unique objects observed in the clip. The dataset was then filtered to remove any object ids that represented vehicles rather than pedestrians. Next, each sequence was set to a uniform length.

To include as many clips as possible, sequences as small as 20 frames were included in the transformed dataset, which resulted in only 1565 of the 8588 sequences being dropped from the dataset. Any sequence length less than 35 frames was forward padded, with the value -1 put in place of all columns of the dataset for each missing frame. The (x,y) position of the pedestrian bounding box was included in each step of the sequence, as well as the calculated diagonal of the box. All categorical values were one hot encoded, resulting in 39 total features provided per frame per person.

CHAPTER 4

EXPERIMENT CONSTRUCTION

The structure of the experimentation completed in this work can be broken down into three major portions. First, a baseline pedestrian trajectory prediction, coordinates-only model was developed and trained on the sequence data. Next, a series of LSTMs were trained, each with one category of the behavioral tagging included with the coordinate data. The “All Categories” model was also included. Based on the results of those models, any category which performed better than the baseline accuracy of coordinates was combined together to train a final LSTM. Figure 8 shows this construction in more detail.

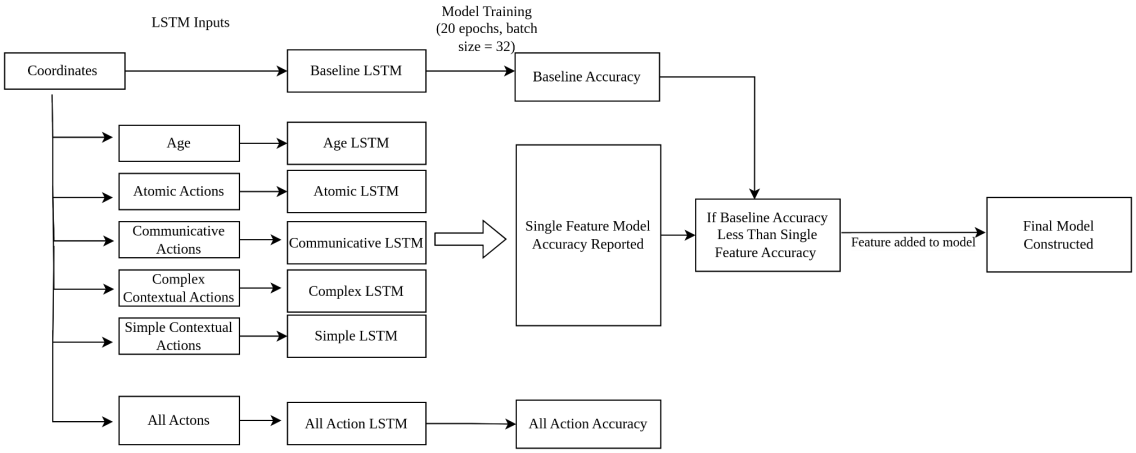


Figure 8: LSTM Data Feeds

Finally, in an attempt to better understand which features of the data were most pivotal in increasing accuracy, a decision tree classifier used the error of the baseline model to classify if a prediction would have high or low error based on the present behavioral tags.

4.1 Baseline Model Construction

In order to determine which behavioral tags are most important in pedestrian prediction, a baseline is established to compare against. In this work, an LSTM encoder/decoder model was trained on bounding box coordinates that acts as the point of comparison. This LSTM is trained on a sequence length of size 28, and predicts out 7 steps into the future. This translates roughly into a model which uses the previous four seconds of input to predict the next second of output. The model was trained for 20 epochs on a train/test split of 80/20. An Adam optimizer was used, with mean square error as the loss function and batch size of 32. The layers of the base LSTM can be seen in Figure 9 below.

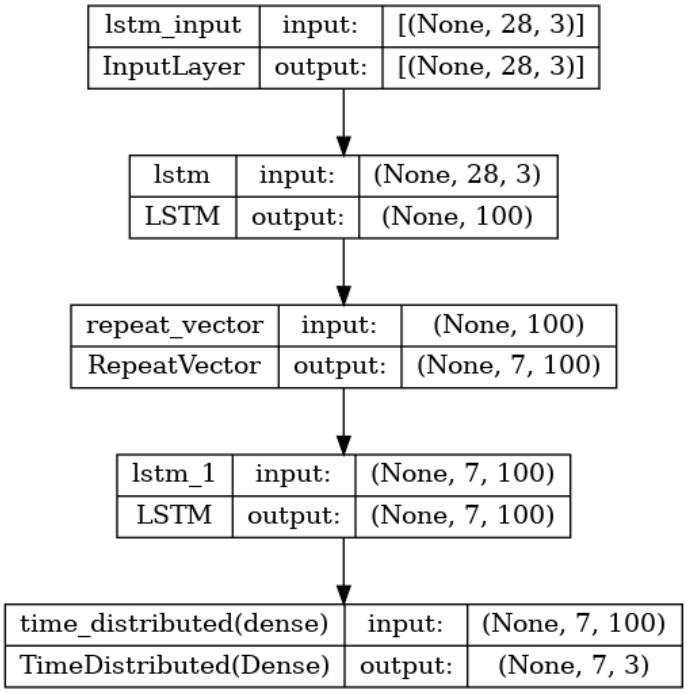


Figure 9: LSTM Layer Construction

4.2 Experimental Model Construction

All experimental models followed the same structure as the baseline model, with one key difference- the size of the input features. Each included the three bounding box coordinates, along with a one hot encoded set of columns representing the contextual or behavioral action present in each frame. The categories tested were Age (6 inputs), Atomic Actions (12 inputs) , Communicative Actions (7 inputs), Complex Contextual Actions(10 inputs), Simple Contextual Actions (16 inputs), and finally, All Categories (39 inputs). Despite the varying number of inputs, all models produced the same number of outputs, predicting only the bounding box coordinates for each future frame. The models were built in python 3.9.12 using the tensorflow keras version 2.11's implementation of the LSTM layer. As will be discussed in the following section, the sklearn library was imported to build the decision tree classifiers utilizing version 1.0.2.

4.3 Use of Decision Trees for Error Prediction

As previously discussed in this work, many actions only occur in a few dozen to a few hundred frames. While these actions are not frequent enough in the data to train a classifier to identify them, this work seeks to answer the question: Are those tags important to identify? Are these rarer tags impactful on a prediction?

To answer this question, a decision tree classifier is leveraged which takes a behavioral tag, or set of tags, as inputs to estimate if the LSTM model prediction will have high error. In this case, high error is defined as a prediction which has an average error across all predicted frames greater than or equal to the value of the third quartile of error of all predictions. The error used as the predictive values came from the baseline

LSTM. The use of a decision tree was motivated by the highly categorical nature of this data. Similarly to the LSTMs generated above, decision tree classifiers were trained to each consider one category of behavioral data. This means that separate decision trees were used to generate prediction errors from Age, Atomic Actions, Communicative Actions, Complex Contextual Actions, and Simple Contextual Actions. A decision tree classifier with all behavioral tags included was also trained.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Comparing LSTM Model Accuracy

Firstly, the training graph of the baseline model can be seen below in Figure 10:

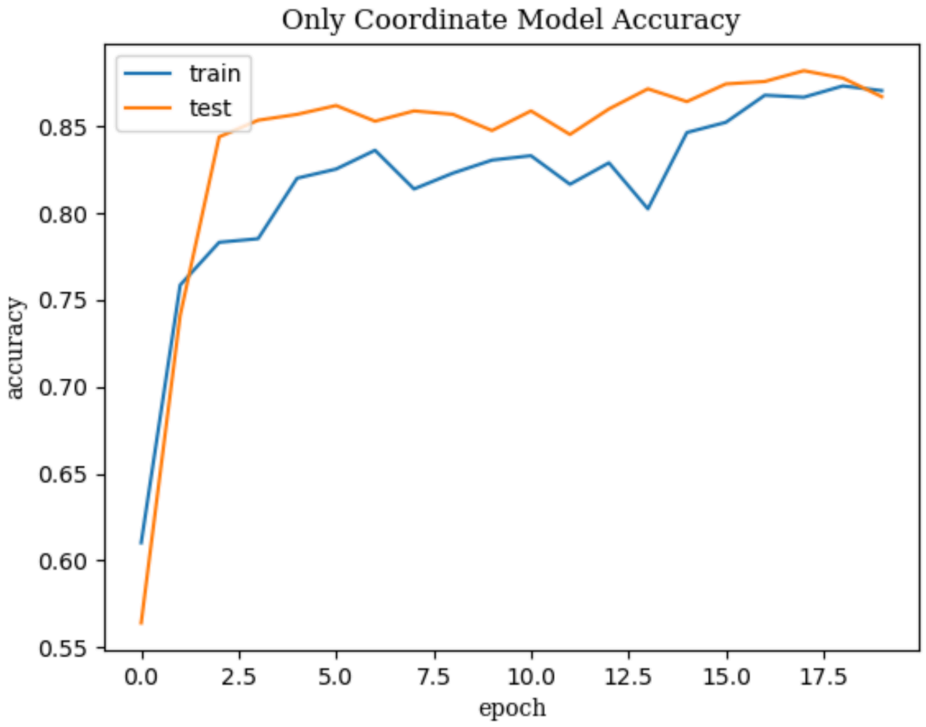


Figure 10: Baseline Model Training and Validation Accuracy

As shown above, the coordinates only model trained to above 85% accuracy, with a converging train and validation accuracy at around 87%. The individual training graphs for each model can be found in Appendix A. The final accuracies of each LSTM model can be seen in Figure 11 below.

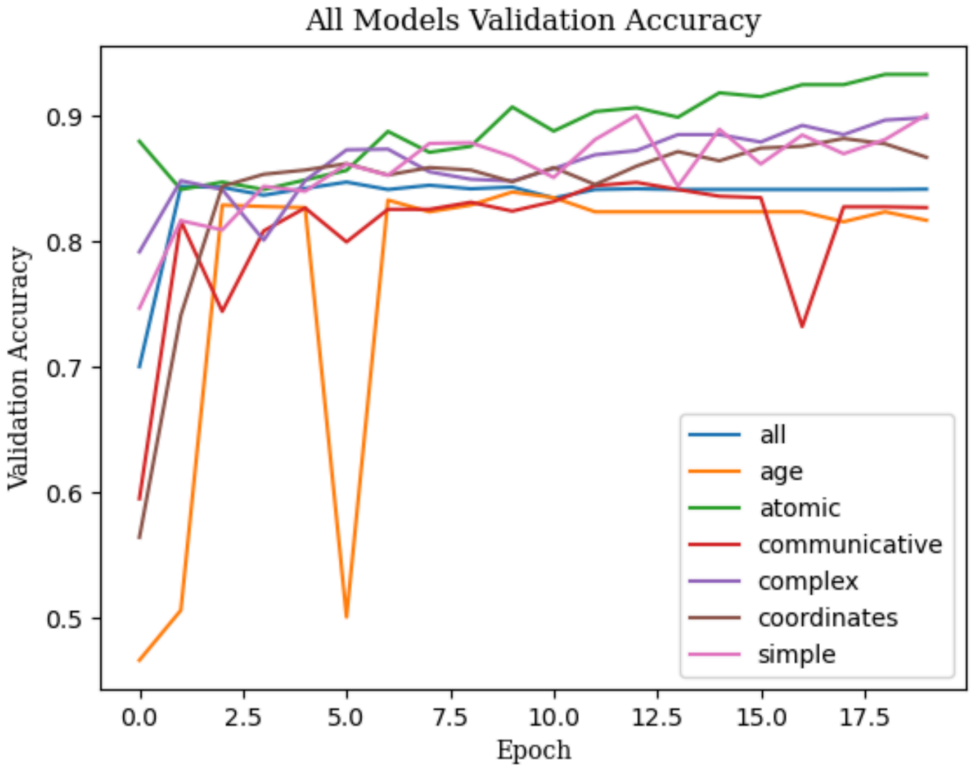


Figure 11: Compared Accuracies of all LSTM Models.

As can be seen above, the baseline coordinate model falls in the middle of the pack. Of the LSTM models that perform worse were the All Categories model, the Communicative Actions model, and the Pedestrian Age model. Scoring higher than the baseline model was the Atomic Action model, the Complex Contextual model, and the Simple Contextual model. The final reported accuracies of each model is shown in Table 2.

Category	Training Accuracy	Validation Accuracy
All Categories	0.8309	0.8416
Atomic Actions	**0.9252**	**0.9331**
Communicative Actions	0.8319	0.8269
Complex Contextual Actions	0.8898	0.8986
Coordinates Only	0.8705	0.867
Simple Contextual Actions	0.8823	0.9009

Pedestrian Age	0.8289	0.8167

Table 2: Compared LSTM Final Accuracies

Notably, in nearly half of the models, the training accuracy was found to be lower than the validation accuracy.

Finally, the best performing categories, Atomic Actions and Simple Contextual Actions were combined to train one final model. The training and validation accuracies of the combined model can be seen below in Figure 12.

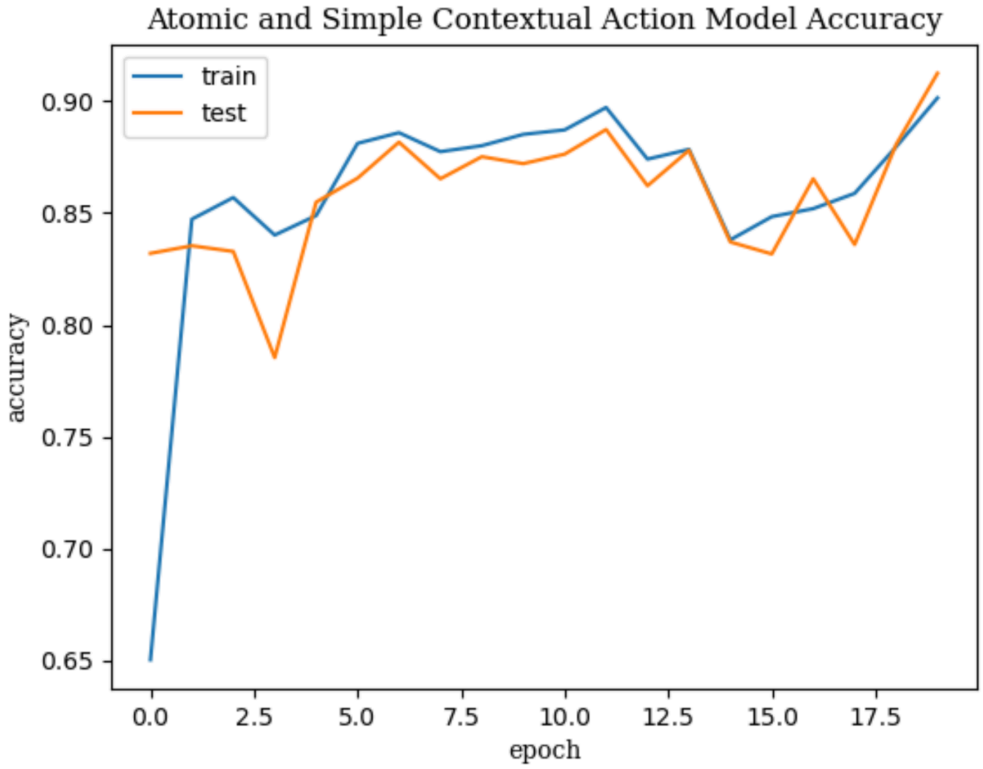


Figure 12: Combined Atomic and Simple Contextual Model

Notably, the accuracy of the combined model was found to be lower than the accuracy of the model trained on Atomic Actions only.

5.2 Decision Tree Classifier Results

Shown below in Table 3 are the confusion matrices of all trained decision trees.

Category	Confusion Matrix		
All Categories		Low Error (Predicted)	High Error (Predicted)
	Low Error (True Value)	1026	2
	High Error (True Value)	237	0
Atomic Actions		Low Error (Predicted)	High Error (Predicted)
	Low Error (True Value)	1037	0
	High Error (True Value)	228	0
Communicative Actions		Low Error (Predicted)	High Error (Predicted)
	Low Error (True Value)	1037	0
	High Error (True Value)	228	0
Complex Context		Low Error (Predicted)	High Error (Predicted)
	Low Error (True Value)	1012	2
	High Error (True Value)	250	1
Simple Context		Low Error (Predicted)	High Error (Predicted)
	Low Error	1032	0

	(True Value)		
	High Error (True Value)	233	0
Pedestrian Age		Low Error (Predicted)	High Error (Predicted)
	Low Error (True Value)	1014	0
	High Error (True Value)	251	0

Table 3: Decision Tree Confusion Matrices

Across all matrices, it is evident that the models did not learn to classify high error based on any features of the data. Though the models all reported approximately 75% accuracy, this stemmed from the construction of the dataset, which consisted of 75% low error data points, and 25% high error data points.

CHAPTER 6

DISCUSSION AND FUTURE WORK

Overall, we see that including Atomic Actions provides the greatest gains to performance, with a 6.6 point gain in accuracy compared to the baseline model. However, complex and simple actions both showed an approximate 3 point gain in accuracy compared to the baseline. Attempting to combine multiple higher performing categories was shown to be detrimental to accuracy, as the Atomic and Simple Context model was outperformed by the Atomic only model by approximately 3 points. It was shown to be detrimental to include either the age of the pedestrian, or the communicative actions of the pedestrian, resulting in a 5 point drop in accuracy in the case of pedestrian age, and a 4 point drop in the case of pedestrian communication actions. The All Categories model clearly suffered from the inclusion of these two categories, also performing 2 points worse than the baseline accuracy.

These findings demonstrate that while it is possible to identify a wide array of contextual information regarding a pedestrian, that most of it is not necessary to gather. The Atomic Actions category consists primarily of the classifications running, walking, standing, and sitting. Essentially classifying if a pedestrian is in a pose which is in motion or stationary is the simplest, and most effective piece of contextual data that can be gained from the pedestrian alone.

It is therefore not necessarily required to train on the TITAN dataset to develop an effective pedestrian crossing intention prediction. Both the JAAD dataset and the PIE dataset provide this information, along with other environmental features to train on, such

as bounding boxes for traffic indicators and pedestrian crossings. This does not necessarily mean that the TITAN dataset should be ignored when training for pedestrian intent prediction, namely because the TITAN dataset provides additional diversity in location. As the TITAN dataset was gathered in Tokyo, Japan, the TITAN dataset provides an opportunity to counteract the western bias present in most currently available datasets [10].

6.1 Discussion of Methods

The use of the LSTM architecture in this paper was motivated by the prevailing industry standard. As discussed earlier in this paper, the majority of studies published on pedestrian intent prediction utilize the LSTM architecture [3]. This work showed similar baseline accuracies to other prevailing literature, further validating its findings. While other nonlinear regressors exist, such as a random forest with a different number of trees, and could have been utilized in this work, the goal of this work was not to construct a unique architecture for pedestrian intent. Rather, the goal was to identify if utilizing the prevailing methods in pedestrian intention prediction, if the information from the TITAN dataset offers meaningful improvements to a prediction.

A slightly unusual aspect of the results from the generated LSTM networks was the tendency of a trained network to perform better in validation accuracy than while training. The difference was a notable 1 to 2 point gain in accuracy on nearly half of the trained categories. A possible explanation for this discrepancy comes from the class imbalances in the dataset. As discussed early on in the paper, there is a notable tendency in the dataset for one class to be extraordinarily dominant compared to the others. In

cases where one of the smaller classes of data is randomly disproportionally large or small in the validation dataset, the validation accuracy can fluctuate. Ideally, this would be solved by augmenting the classes with low representation, or trimming the number of frames in the most dominant class. However, given the results of the feature importance evaluation discussed below, it seems that the difficulty of creating an augmented dataset exceeded the possible benefits of building a more balanced version of this dataset.

The use of decision trees in this paper to evaluate feature performance provide some of the most compelling results. The idea behind the decision trees is simple. If a pedestrian's behavior frequently causes high error, the decision tree should be able to identify it, even if there are few examples of the given behavior.

This analysis was intended to explore the poorly represented classes in the data. Since the LSTM would likely not be able to learn to account for poorly represented behaviors, these decision trees were a second chance to examine if these additional data features should be included for consideration. However, as discussed briefly in Chapter 5, the decision tree classifiers were never able to identify a behavior that caused high error.

Regardless of the depth of the tree, the classifiers were unable to categorize the high error. They simply predicted low error for every feature. While not included in the results section, a decision tree regressor was made first to not just classify the error as high or low, but directly predict it. That decision tree attained a negative R^2 value- which, according to the sklearn library, means that the developed regression was worse at predicting the error than a horizontal line. From these results, it is safe to conclude that there is no feature that directly causes the LSTM to predict poorly. While some features

may help improve accuracy, no behavioral feature specifically causes an LSTM to fail to predict well.

6.2 Future Work and Final Conclusions

The most effective contextual behavioral action was shown to be the set of Atomic Actions. Essentially the set of actions which describe the motion status of the pedestrian visible on the screen. Numerous studies have shown the ability to train on this feature such that it is identified accurately, and computes quickly enough to be useful in real time predictions. The other potentially helpful attributes of simple contextual actions and complex contextual actions, both showed lower performance than simply identifying the Atomic Action, and proved to be detrimental to overall accuracy if included alongside the Atomic Actions. Future work should focus on identifying and utilizing other contextual elements of a scene, such as the weather, or presence of traffic control signs, rather than continuing to attempt to extract features from the pedestrians themselves.

REFERENCES

[1] S. A. Bouhsain, S. Saadatnejad, and A. Alahi, ‘Pedestrian Intention Prediction: A Multi-task Perspective’, CoRR, vol. abs/2010.10270, 2020.

[2] A. Poibrenski, M. Klusch, I. Vozniak, and C. Müller, ‘Multimodal Multi-Pedestrian Path Prediction for Autonomous Cars’, SIGAPP Appl. Comput. Rev., vol. 20, no. 4, pp. 5–17, Jan. 2021.

[3] T. Chen and R. Tian, ‘A Survey on Deep-Learning Methods for Pedestrian Behavior Prediction from the Egocentric View’, in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2021, pp. 1898–1905.

[4] C. -F. Hsieh, Q. Lin, W. -J. Jhang and C. Hsu, "The Implementation of Pedestrian Detection Based on YOLOv5s," 2023 9th International Conference on Applied System Innovation (ICASI), Chiba, Japan, 2023, pp. 106-108, doi: 10.1109/ICASI57738.2023.10179530.

[5] J. -M. Lin, W. -L. Lin and C. -P. Fan, "Deep Learning Based Efficient Pedestrian Age Group Estimation and Tracking Techniques for Autonomous Mover," 2023 5th International Conference on Computer Communication and the Internet (ICCCI), Fujisawa, Japan, 2023, pp. 15-19, doi: 10.1109/ICCCI59363.2023.10210174.

[6] M. Zaman, S. Saha, N. Zohrabi and S. Abdelwahed, "Deep Learning Approaches for Vehicle and Pedestrian Detection in Adverse Weather," 2023 IEEE Transportation Electrification Conference & Expo (ITEC), Detroit, MI, USA, 2023, pp. 1-6, doi: 10.1109/ITEC55900.2023.10187020.

[7] J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO: From YOLOv1 and Beyond," arXiv preprint arXiv:2304.00501, 2023, cs.CV.

[8] X. Zhu, W. Fu and X. Xu, "Intent Prediction of Pedestrians via Integration of Facial Expression and Human 2D Skeleton for Autonomous Car-like Mobile Robots," 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 2021, pp. 1775-1780, doi: 10.1109/ICIEA51954.2021.9516159.

[9]A. P. Samant, K. Warhade and K. Gunale, "Pedestrian Intent Detection using Skeleton-based Prediction for Road Safety," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, 2021, pp. 238-242, doi: 10.1109/ACCESS51619.2021.9563293.

[10] A. Gagneja, M. Bansal, A. Arora and B. Lall, "On the efficacy of Pedestrian Detection in Indian Road Scenario," 2022 1st International Conference on Informatics (ICI), Noida, India, 2022, pp. 92-97, doi: 10.1109/ICI53355.2022.9786893.

[11] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, ‘Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior’, in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 206–213.

[12] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, ‘PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction’, in International Conference on Computer Vision (ICCV), 2019.

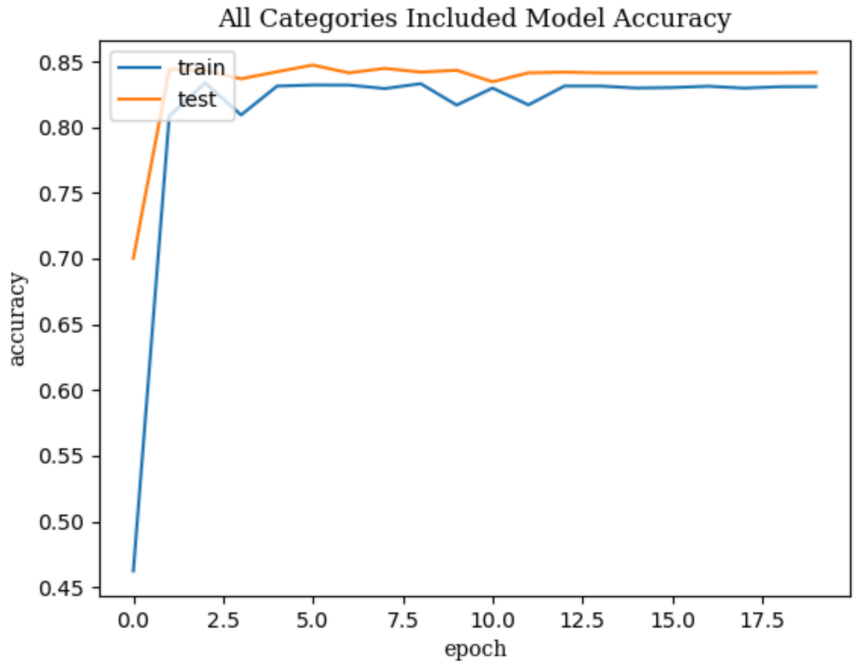
[13] B. Liu et al., ‘Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction’, IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 3485–3492, 2020.

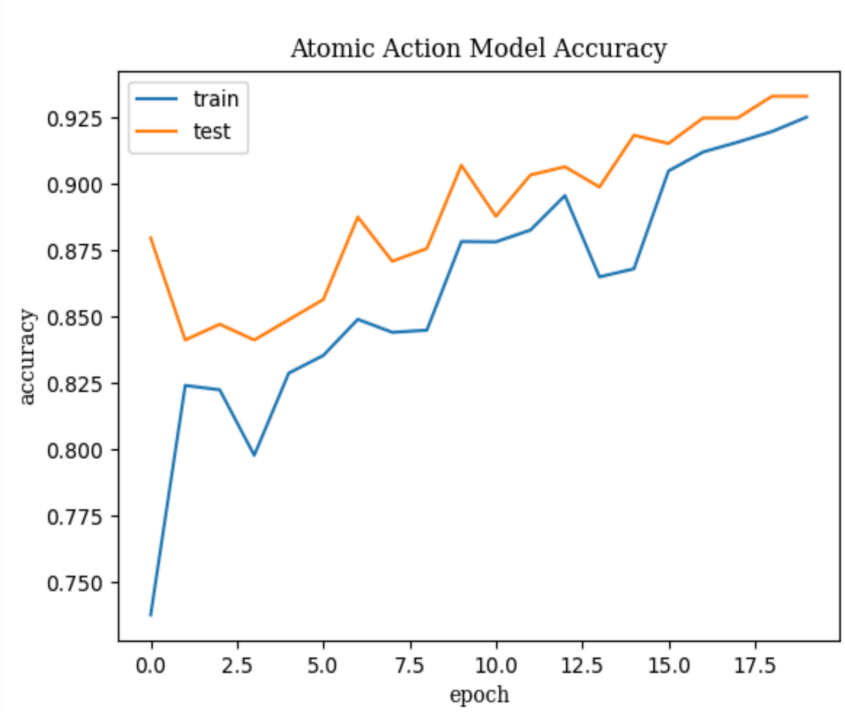
[14] S. Malla, B. Dariush, and C. Choi, ‘TITAN: Future Forecast using Action Priors’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11186–11196.

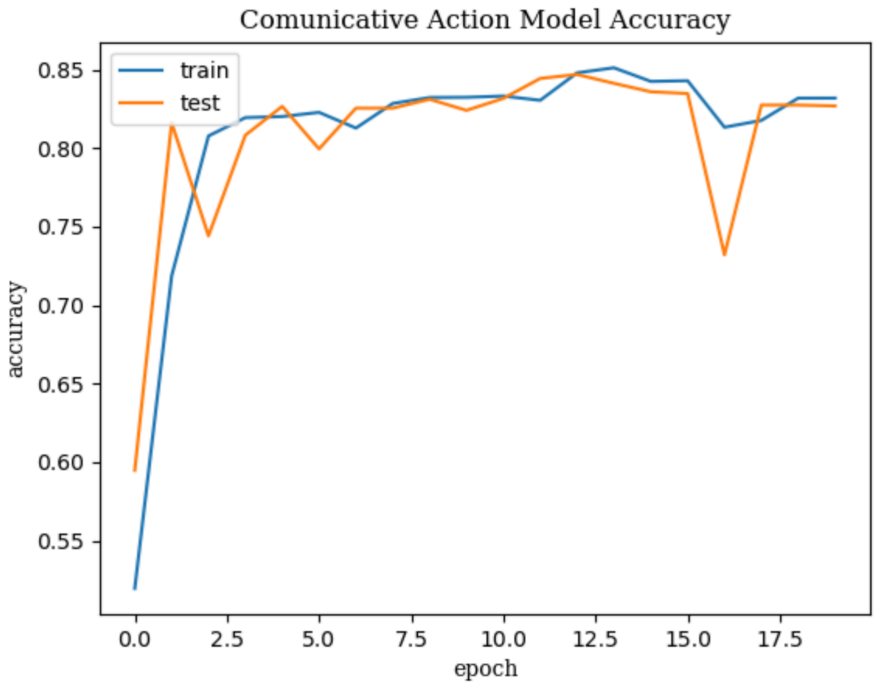
[15] S. Schmidt and B. Färber, ‘Pedestrians at the kerb – Recognising the action intentions of humans’, Transportation Research Part F: Traffic Psychology and Behaviour, vol. 12, no. 4, pp. 300–310, 2009.

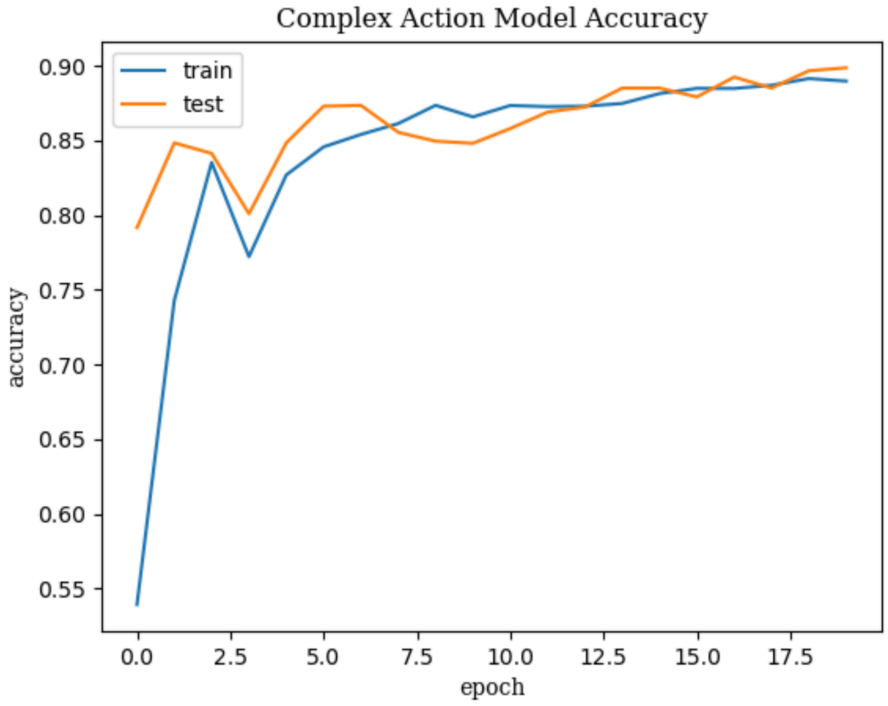
APPENDIX A

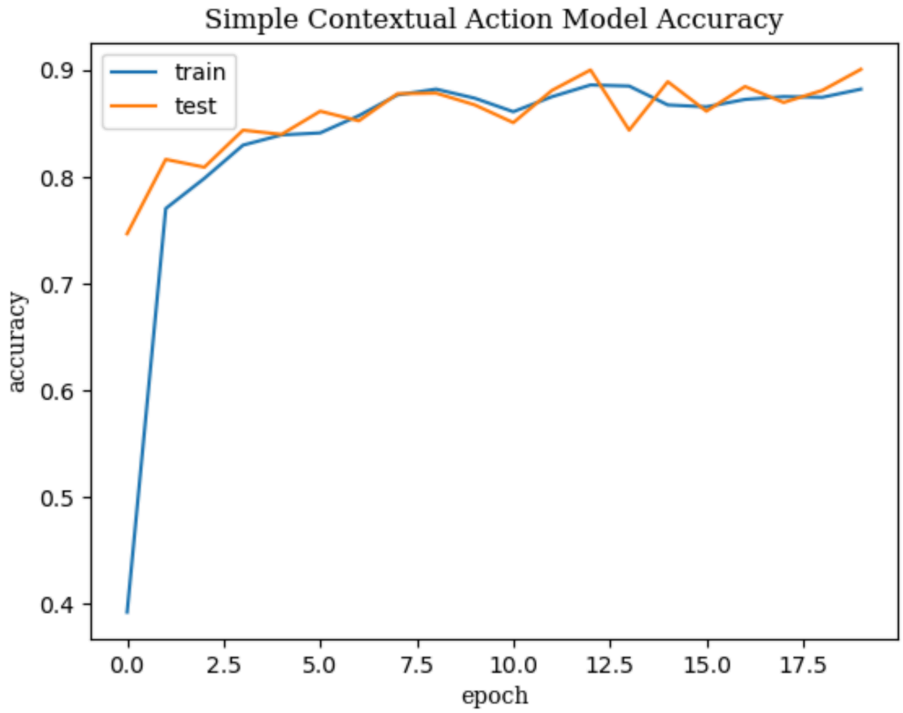
LSTM TRAINING GRAPHS

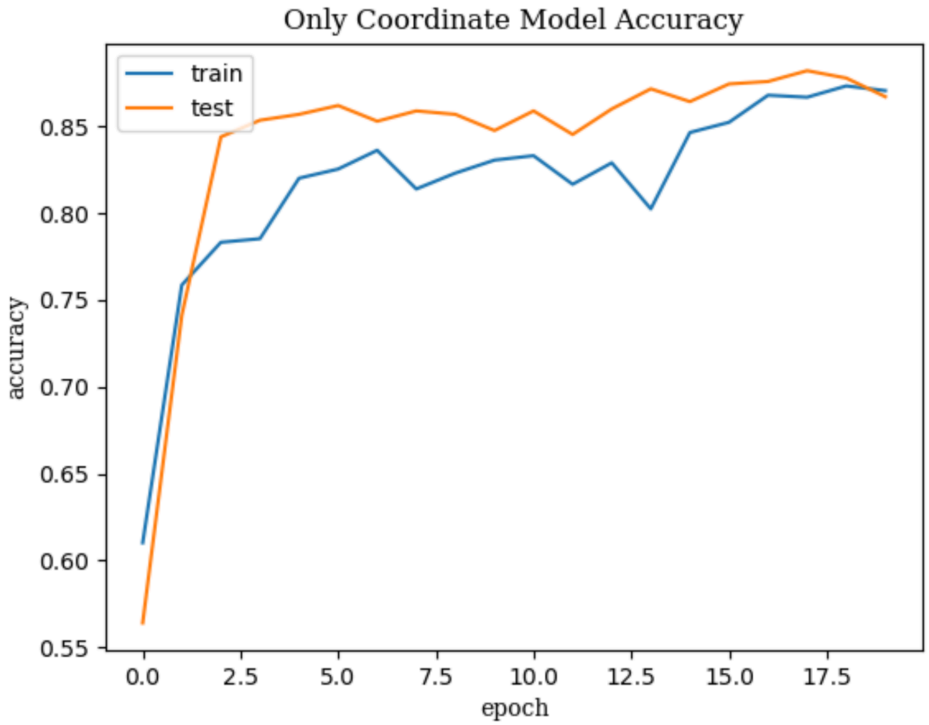


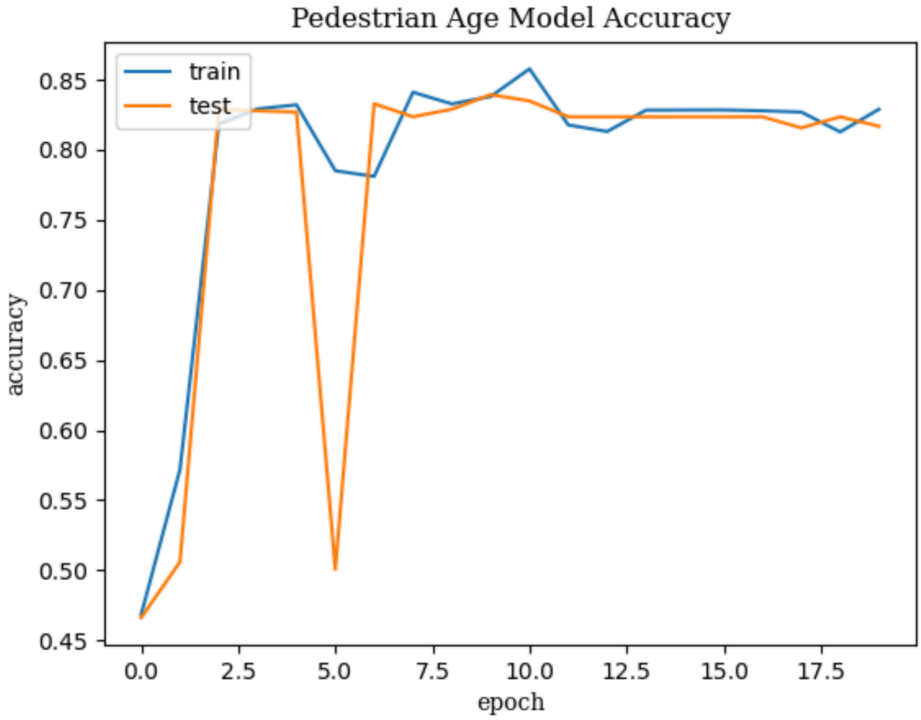


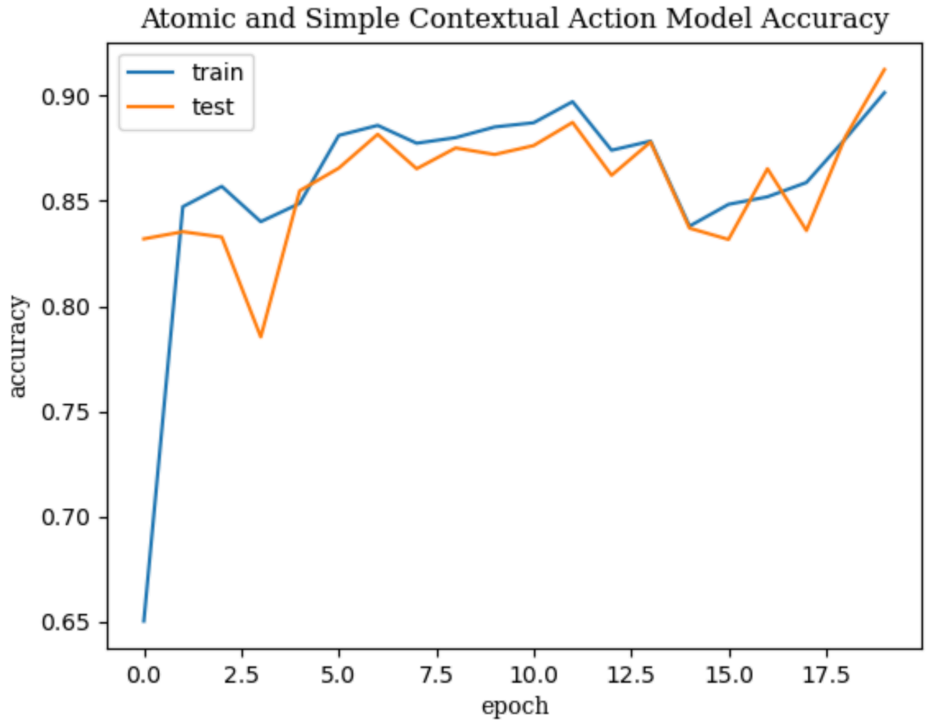

Comunicative Action Model Accuracy
train
test
accuracy
epoch
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Complex Action Model Accuracy
train
test
accuracy
epoch
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Simple Contextual Action Model Accuracy
train
test
accuracy
epoch
0.9
0.8
0.7
0.6
0.5
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Only Coordinate Model Accuracy
train
test
accuracy
epoch
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Pedestrian Age Model Accuracy
train
test
accuracy
epoch
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Atomic and Simple Contextual Action Model Accuracy
train
test
accuracy
epoch
0.65
0.70
0.75
0.80
0.85
0.90
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5