Activity Monitoring System using Deep Learning for People with Dementia

by

Amarzish Qadeer

A Thesis submitted to the Faculty of Graduate Studies

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Biomedical Engineering Program

Faculty of Graduate Studies

University of Manitoba

Winnipeg

Abstract

Dementia is a degenerative condition that affects cognitive abilities and daily functioning. This project aims to explore and evaluate activity recognition algorithms to support the assisted living of people with dementia. The proposed deep learning approach can help to monitor people with dementia and support their caregivers in providing effective care. We tried a new approach for detecting the activities of daily living for people with dementia. We explored ExpansionNet_v2 model and used it to train on the Toyota Smart Home dataset in order to detect the activities od daily living. The dataset was converted into COCO dataset format. Bounding boxes were generated using Faster-RCNN with ResNet backbone pretrained model from pytorch. Captions were generated using scene understanding. This involved analyzing the image or video to extract semantic information about the environment and objects within it, including their relationships and context. Semantic relationships and patterns were extracted, which helped in building a more comprehensive understanding of the scene.

The training process involved two steps - initial training and fine-tuning. During initial training, newly added layers were trained while keeping the pre-trained layers of the Swin-Transformer backbone frozen. Fine-tuning involved training the

entire network, including both the pre-trained backbone and newly added layers, on the dataset. The purpose of using multiple frames from a video during training is to increase the probability of detecting the pose accurately and generating a good caption.

The algorithm to detect ADL was tested on real-life videos of three dementia patients at different stages of dementia. The daily activities of these patients were recorded to test the algorithm after training and validation on the Toyota SmartHome dataset.

Acknowledgements

I would like to begin by expressing my gratitude to Allah Almighty for granting me the opportunity to pursue my Masters in Biomedical Engineering at the University of Manitoba and for providing the support and guidance necessary for me to successfully complete my thesis.

I am deeply indebted to my supervisor, Dr. Mohamed-Amine Choukou, whose inspiring work provided the foundation for my research and for providing me with lots of research ideas. He was instrumental in guiding me throughout the project, sharing the latest research ideas and papers, and offering invaluable feedback to improve my work. His unwavering support, revision of all drafts, and meticulous line-by-line comments were indispensable in helping me present my work in the best possible way.

I would also like to thank my committee members, Dr. Sherif Sherif and Dr. Tabrez Siddiqui, for their encouragement, support, and valuable feedback in enhancing my work and writing. For appreciating my work and giving me valuable feedback to improve my work.

To my parents, I owe a debt of gratitude for their unwavering support, encouragement (even scolding), and prayers, which motivated me to complete my degree successfully. Who did everything in their power to support me with my whole Masters degree.

My deepest appreciation goes to my brother, who stood by me through every step of the process, even when I was at a point where I stuck badly and had to revise the complete work again. He tirelessly supported me, staying up all night to guide me through the process, helping me identify and correct errors, and offering invaluable feedback on my paper.

I would like to extend my sincere thanks to the Toyota Smart Home team for providing a challenging and fantastic dataset that allowed me to work on the real-world challenges of daily living activities.

Finally, I would like to acknowledge myself for coming up with the idea of supporting people with dementia. For all the days and nights working tirelessly to find the right dataset, approach, and algorithms. Despite many challenges, I remained determined to move forward, reading countless research papers and finding a new way to approach this problem. **I am proud of myself.**

Table of Contents

Abstract 2

List of Acronyms / Abbreviations 7

List of Figures 8

List of Table 9

CHAPTER 1 INTRODUCTION 10

1.1. Dementia 10

1.2. Monitoring People with Dementia 13

1.3. What is ADL? 15

CHAPTER 2 BACKGROUND 17

2.1. Challenge for Monitoring Algorithms 17

2.2. Machine Learning 20

2.3. Computer vision 22

2.4. Monitoring ADL 23

2.5. Predicting Movement 24

2.6. Literature 28

CHAPTER 3 OBJECTIVES 46

3.1. Computer vision and dementia 48

CHAPTER 4 METHODOLOGY 53

4.1. Selection of Dataset 53

4.2. Toyota Smart Home Dataset 57

4.3. Dataset challenges 59

4.4. Current algorithms and approaches 61

4.5. Labels 63

4.6. Coarse and Fine Grained lables 64

4.7. Annotation Format 65

4.8. Coco Dataset 65

4.9. Coco Annotations 66

4.10. IMAGE CAPTIONING 66
4.11. CONVERTING OUR DATASET TO COCO ANNOTATION FORMAT 68

CHAPTER 5 APPROACH 70

5.1. SCENE UNDERSTANDING 70
5.2. IMAGE CAPTIONING 72
5.3. CHOICE OF MODEL - EXPANSIONNET V_2 MODEL 73
5.4. BOUNDING BOXES GENERATION 75
5.5. FINE TUNING AND TRAINING ON TOYOTA SMART HOME DATASET 77
..... 84
5.6. LOSS FUNCTION 85

CHAPTER 6 RESULTS & COMPARISON 88

6.1. PERFORMANCE MEASURE 89
6.2. BLEU 90
6.3. METEOR 98
6.4. EXPERIMENTAL SETUP 107
6.5. OUR TESTING SAMPLES 108
6.6. EVALUATION PROTOCOL 110
6.6. QUALITATIVE ANALYSIS 119
6.7. SUMMARY 120

CHAPTER 7 DISCUSSION & FUTURE WORK 124

7.1. DISCUSSION 125
7.2. CASES EXPLAINING WHY THIS RESEARCH IS IMPACTFUL/IMPORTANT 132
7.3. LIMITATIONS 134
7.4. FUTURE STUDIES 136

REFERENCES 138

List of Acronyms / Abbreviations

1- Actvities of Daily Living (ADL)
2- Machine Learning (ML)
3- Deep Learning (DL)
4- Learning Video Pose Embedding (VPN)
5- Unified Franework for Real-World Skeleton based Action-Recognition (UNIK)
6- Selective Spatio-Temporal Aggregation (SSTA)

List of Figures

Figure 1 Single Frame of Action 26
Figure 2 Multiple Frames of Action 27
Figure 3 MSR Daily Activity Dataset 35
Figure 4 MSR Action 3S Dataset 36
Figure 5 Framework 84
Figure 6 Total Loss Across Epochs 87
Figure 7 Person with Dementia 112
Figure 8 Activity Monitoring of Person with Dementia 113

List of Table

Table 1 Literature 46
Table 2 Metrics 107
Table 3 Examples of Caption 115

Chapter 1
Introduction

1.1. Dementia

Dementia is a debilitating condition affecting the brain cells of older individuals, causing a decline in cognitive abilities and interfering with their ability to perform ADLs [1]. The decline in cognitive abilities is not considered a normal part of aging. The cells in the brain of individuals with dementia start to deteriorate, leading to a shrinkage of the brain and an overall decline in the person's ability to perform daily tasks [1].

According to the Alzheimer Society of Canada, there are over 50,000 individuals in Canada suffering from dementia [2]. By 2030, this number is expected to rise to 900,000, with 70,000 people diagnosed each year . The total number of individuals with dementia, including both new and old cases, is referred to as the dementia

prevalence. On average, 7,600 Canadians are diagnosed with dementia each year [3].

As the number of individuals diagnosed with dementia continues to rise, the cost of care for dementia will also increase dramatically. In Canada, it is estimated that by 2040, over 1.2 million individuals will be affected by dementia [1]. The Alzheimer Society of Canada's Landmark Study Report from 2022 found that 1 in 5 Canadians have experience caregiving for someone with dementia [2].

Dementia is a growing concern for healthcare professionals and families across the world. It has a significant impact on the daily lives of people with dementia and those who care for them. The cost of caring for individuals with dementia is projected to rise significantly in the future, due to an increasing number of dementia diagnoses. According to the Alzheimer Society of Canada, the cost of dementia care is expected to reach $16.6 billion by 2031 [2].

Caring for someone with dementia can be a challenging task, both emotionally and physically. This highlights the need for better support and resources for families and caregivers.

A person with dementia needs more care and support due to decline in cognitive abilities and as the condition progresses they need more care [4]. A person with

dementia needs constant monitoring in order to prevent any harms and accidents that might happen such as falls. In most cases a member of the family takes the responsibility of caring for person with dementia. Caring for person with dementia can be exhausting as the person with dementia requires continuous monitoring. Hiring a care specialist to provide the care for person with dementia at home for long and short periods of time allows family caregivers to have some time for themselves. But all this adds to the cost of care for a person with dementia [4].

One of the options for monitoring people with dementia is to have them monitored in a healthcare center where they receive individualized care and attention from trained professionals who monitor their health, behavior, and ADLs. While this type of care can be highly effective, it comes with a high monetary cost, which can be a barrier for many families. However, the system aims to keep people with dementia at home for as long as possible, and institutionalization will only be considered when staying at home is no longer feasible. While this option is often preferred by families, as it allows them to live with their loved one and provide care on a daily basis, it can also present its own challenges. Caregivers must be available 24/7 to meet the needs of the person with dementia, which can be physically and emotionally demanding.

1.2. Monitoring People with Dementia

Monitoring people with dementia is always a challenge. Many caregivers including family caregivers and health professionals face challenges while taking care of people with dementia.

There are several reasons why it's important to monitor people with dementia:

- Safety: People with dementia may become disoriented and wander, putting themselves at risk of accidents or harm. Monitoring can help ensure their safety.
- Health: Monitoring can help detect any changes in the person's health, such as a sudden decline in physical or mental condition. This can lead to early intervention and prompt medical treatment.
- Quality of life: Monitoring can help identify any changes in the person's mood, behavior, or needs, and enable carers to adjust their care plan to ensure the person's quality of life is maintained.
- Caregiver support: Caregiving for someone with dementia can be physically and emotionally demanding, and monitoring can help ensure the caregiver is not overwhelmed and has support when needed.

Monitoring people with dementia is important for ensuring their safety, health, and well-being, as well as the well-being of the caregiver. However, monitoring people with dementia can be challenging for family caregivers. Fortunately, advancements in computer vision technology can help by enabling better monitoring of people with dementia. By tracking their activities and movements, computer vision can detect changes in routines and alert caregivers to potential changes in physical or mental health. This information can help caregivers adjust their care plan to ensure the person's well-being is maintained. While computer vision cannot directly improve the quality of life for people with dementia, it can support caregivers in providing effective care.

One of the key benefits of computer vision technology in the context of dementia care is its ability to detect wandering and falls. People with dementia often wander, putting themselves at risk of injury or harm. Computer vision equipped with cameras can detect when a person is wandering and alert caregivers to intervene and ensure their safety. Similarly, falls can be dangerous for people with dementia, and computer vision can detect falls and alert caregivers to provide assistance.

It can also provide real-time monitoring of the person’s environment and conditions, enabling caregivers to quickly respond to any changes or emergencies.

This can help ensure the person's safety and well-being, providing peace of mind for caregivers.

Computer vision technology has the potential to improve the quality of life for people with dementia and support caregivers in providing effective care. By monitoring ADLs and providing real-time monitoring, computer vision can help ensure the safety, health, and well-being of people with dementia, and provide support for caregivers.

1.3. What is ADL?

ADLs refer to the basic tasks that individuals perform in their daily routine, such as eating, bathing, dressing, grooming, using the bathroom, and transferring (getting in and out of bed or a chair) [5] . Basic ADL include, which are performed by not only a healthy person but are also part of ADL for people with dementia. In the context of a person with dementia, these activities may become more challenging due to the decline in cognitive abilities, memory, and decision-making skills. As the disease progresses, these individuals may require more assistance and support with performing these basic tasks. It is essential for people with dementia to continue performing these activities to the best of their abilities, with proper support and guidance, in order to maintain their independence and quality of life.

The goal of monitoring a person with dementia's ADLs is to classify these actions into different classes, so that their activities can be accurately tracked and reported. This information is useful for both caregivers and healthcare professionals, as it allows them to better understand the individual's daily habits and routines. By using artificial intelligence and neural network technology, the monitoring system can continuously track the person's actions and provide a detailed account of their activities. This information can then be used to make informed decisions about the person's care and support, ultimately improving their quality of life and to address any issues that arise, helping to maintain the independence and well-being of people with dementia.

To achieve this goal, computer vision technology and artificial intelligence can be used to monitor the ADLs of people with dementia.

Chapter 2
Background

2.1. Challenge for Monitoring Algorithms

Action recognition and monitoring algorithms that are trained on scripted and random datasets have limitations in monitoring the ADL of people with dementia [16]. The behavior and activities of people with dementia can be unique and unpredictable, which can pose challenges for these algorithms in accurately recognizing and monitoring their activities. It is important to understand the unique challenges that these individuals face in terms of their abilities and behavior. People with dementia often experience a decline in cognitive abilities, including memory loss, confusion, and difficulty with communication. This can make it difficult for algorithms trained on scripted or random datasets to accurately recognize and monitor their activities.

Traditionally, action recognition algorithms are trained on either scripted or random datasets. Scripted datasets consist of predefined actions and movements,

while random datasets are a compilation of randomly collected actions and movements. However, the ADLs of people with dementia are often unique and unpredictable, making it difficult for these algorithms to accurately identify and track their movements. People with dementia often experience changes in their physical and cognitive abilities, which can further hinder the accuracy of these algorithms. For example, individuals with dementia may experience muscle weakness or stiffness, making it difficult for them to perform specific movements or actions. Furthermore, their cognitive abilities may decline, leading to changes in their ability to perform daily activities and communicate with others.

A person with dementia may perform an ADL in a different way compared to a healthy individual. This can result in a false negative or false positive detection by the algorithm. Furthermore, scripted datasets are limited in their diversity and do not reflect real-world scenarios, making it difficult for algorithms trained on these datasets to generalize to different situations.

One example of this is in the area of falls prevention. Falls are a common concern for people with dementia and their caregivers, as they can lead to serious injuries and further decline in physical and cognitive abilities. However, action recognition algorithms trained on scripted datasets may not accurately capture the unique

movements and behaviors of people with dementia, such as unsteady gait or changes in physical mobility. This can result in false alarms or missed opportunities to intervene and prevent falls.

Another challenge is that people with dementia may engage in repetitive behaviors or routines, which can be difficult for algorithms trained on random datasets to recognize and distinguish from other activities. For example, a person with dementia may repeatedly walk around the house, sit down, stand up, and repeat this pattern throughout the day. An algorithm trained on random datasets may not accurately capture this repetitive behavior and instead classify it as a series of separate activities, such as walking, sitting, and standing.

To address these limitations, researchers have been exploring the development of algorithms that can adapt to the unique behavior of people with dementia. These algorithms use machine learning techniques to learn from the specific behavior of the person with dementia and make more accurate predictions about their activities. Until now, the images have been captured using an RGB camera, and these are detected with 70-75% accuracy [6].

2.2. Machine Learning

The ability to perform daily living activities independently is a critical aspect of maintaining the quality of life for individuals with dementia. As the symptoms of dementia progress, it can become increasingly challenging for individuals to maintain their independence [7]. This can lead to a decline in the individual's ability to perform ADL, such as bathing, dressing, eating, and toileting, without assistance.

To support individuals with dementia and their caregivers, it is essential to develop innovative solutions that can help maintain independence and improve quality of life. One such solution is the use of Artificial Intelligence (AI) and neural network technology to monitor and classify the daily activities of individuals with dementia. This technology can provide valuable information to caregivers and healthcare professionals about the individual's behavior and support in detecting any potential hazards or events that may pose a risk to their safety.

Given the increasing prevalence of dementia and the need for support, many family members choose to care for their loved ones with dementia at home. However, providing 24/7 care can be a significant challenge, especially when the individual's dementia is in its advanced stages. Care homes are an alternative solution, but they also face the challenge of providing adequate care to multiple individuals with

dementia at the same time. The development of AI-powered monitoring solutions can play a significant role in overcoming these challenges and enabling better care for individuals with dementia.

The focus will be to Develop machine learning techniques, a subfield of artificial intelligence that involves training algorithms to make predictions or decisions based on data, to monitor the behavior of the patient to facilitate the physical well-being of people with dementia. This will allow early identification of patients' needs and provide better quality of care for people with dementia.

Machine learning focuses on the development of algorithms and statistical models that enable computers to improve their accuracy and performance on tasks by learning from data [8]. It involves the use of algorithms and statistical models to analyze and understand patterns in large data sets, and then make predictions based on that understanding [9]. Machine learning algorithms are designed to automatically identify patterns and relationships in data and to make decisions or predictions based on that data. The algorithms continuously learn and improve their accuracy over time as they are exposed to more data. Some common applications of machine learning include image recognition, natural language processing, and recommendation systems.

2.3. Computer vision

Computer vision is the “study of visual data” [3]. Various algorithms were developed over time to recognize images and objects in an image. Computer vision is one of the fastest growing fields of technology. Various applications of computer visions are being used from video monitoring of traffic, pedestrians on road to the monitoring of actions and behavior of a person. One of the applications of computer vision is the monitoring behaviors of humans. Medical field is also taking the benefit if this technology and is trying to implement the applications of computer vision in monitoring the behavior of people like people with dementia or Parkinson’s disease. When it comes to monitoring people with dementia, computer vison is being used to monitor various aspects of People with dementia. This vision technology is being used to monitor dangerous behaviors of people with dementia. Also, this vison technology is also being used to detect dangerous objects around patients with dementia.

It is difficult for algorithms to understand what is going on in the visual data. A lot happens in 1 second. A normal video usually has 30 frames per second. Some new algorithms support 200-300 frames per second.

There is not an agreed-upon limit to how many FPS the eye can see. Experts continually go back and forth, but it has been concluded that most people can see 30 – 60 frames per second. [4] And this is the limit of human eyes to monitor the visual data.

2.4. Monitoring ADL

Monitoring the ADL is important to support the independent living of the people with dementia and to support safe and secure “Aging in place”. Many people prefer taking care of the people with dementia, their family members or friends, at their own homes in front of their eyes. But it sometimes gets difficult for them to keep a 24/7 watch over the patient. Also, those patients don’t want company or having the sense of being watched all the time. There are also concerns for privacy when it comes to monitoring patients.

What if we don’t want to see what is going on because of the privacy concerns? What if we just want to know what is going on. For that purpose, the video monitoring technology should be able to provide accurate results. If you are looking at the video, you will be able to know more and if any action is not detected by the video monitoring technology, then your eyes will still be able to see that, and you will feel okay with that, after seeing that the patient is doing okay. But we cannot

watch over the patient all the time. Sometimes we just want to know if the patient is doing okay. For that we will have to believe in the output being shown by the video monitoring technology regarding what activity is being performed. For that purpose, the video monitoring technology must be ablet to detect the activity accurately or the best or closest possible activity that is being performed by the patient.

2.5. Predicting Movement

Activity recognition can be stated as the “problem of predicting movement” [6]. It uses sensors such as video and cameras to capture data and then identify the specific activity, movement, or behavior of an object.

There are many types of action that can be recognized through various activity recognition algorithms such as movement of a person, gesture recognition, gait recognition and it is also being used in video surveillance systems.

When it comes to activity recognition it is divided into steps to recognize any activity performed. First, the action performed in the video is recognized and then the recognized action is classified into an appropriate class [5].

Both action representation and action recognition/classification use different algorithms. To recognize and classify the activity algorithms are combined. Usually spatio-temporal algorithms are used for action recognition, that is, something that is related to both space and time. When it comes to activity recognition it is not possible to identify any activity using just one image. To identify an action, both time and space are required. Machine learning and deep learning algorithms have proved to be efficient in capturing and analyzing the spatio-temporal data. Spatio-temporal algorithms

Any activity can be recognized by analyzing the number of frames regarding where and at what position was the object is in each frame. A single frame gives least information of what activity is being performed.

In the picture given below, Figure 1 we cannot understand whether the man is sitting or standing unless we analyze the other nearby frames as well.

Figure 1 Single Frame of Action

source: https://www.youtube.com/watch?app=desktop&v=0DkMh9ESQsw

But if we see the sequence of frames in a video then it is possible to understand what kind of activity is being performed in the given time.

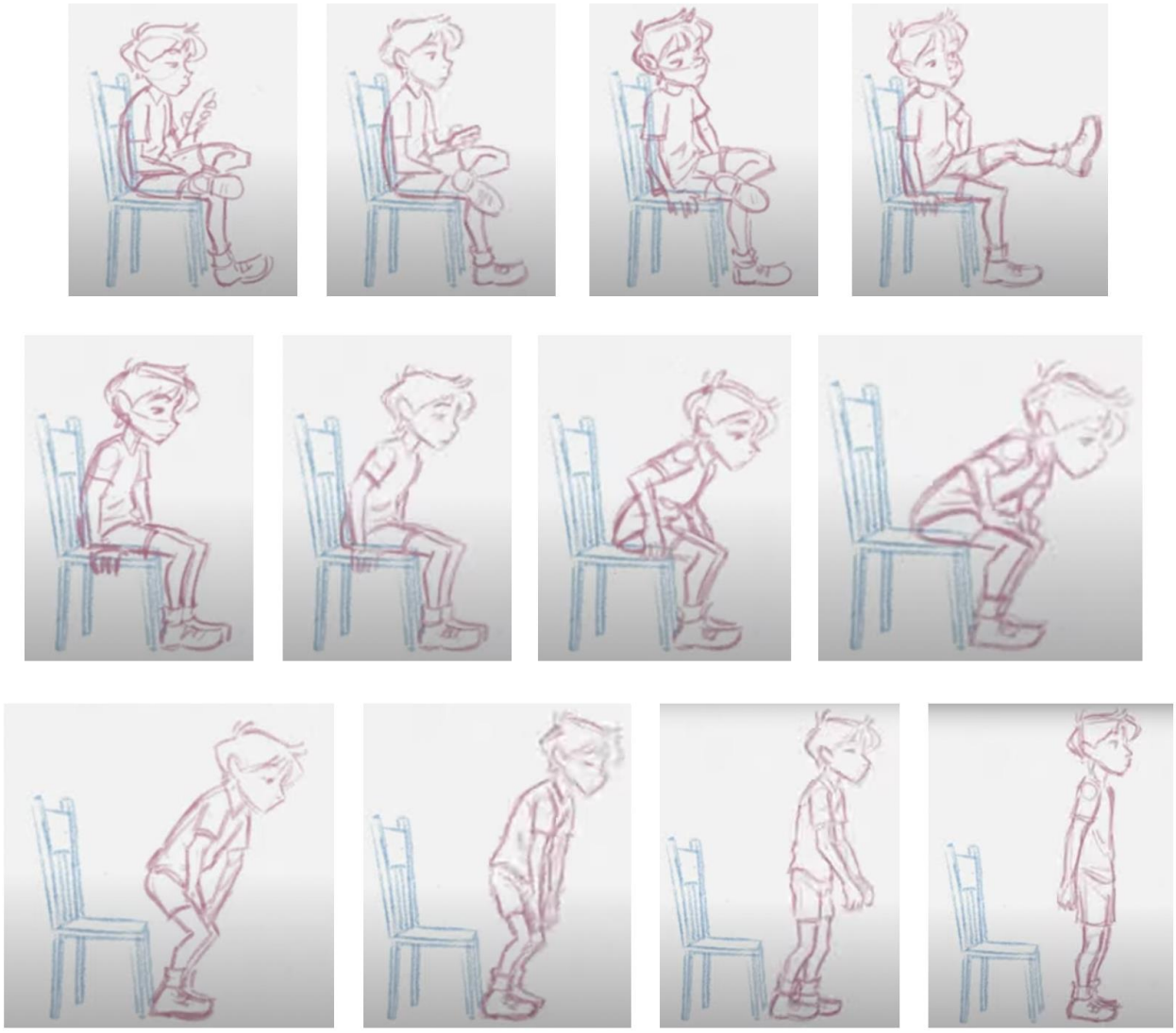


Figure 2 Multiple Frames of Action

source: *https://www.youtube.com/watch?app=desktop&v=0DkMh9ESQsw*

In figure 2 after viewing all the frames of the action performed in the given time it is possible to identify what kind of action is being performed. In this case, the action of "standing up from the chair" is being performed. It is possible to predict the

action being performed after analysing more than 1 frame. 1 single image or frame does not provide complete information of the action being performed.

From a 6 second video [7] with 100+ frames, few frames were extracted to describe the concept and complexity of action recognition in the above figures 1 and 2. A machine learning or deep learning algorithm will analyze all the 100+ frames in the video to recognize the type of action being performed in the video. ML and DL algorithms use spatio-temporal information, that is, the information of time and space at the same time to identify the action being performed.

Then comes the complexity of analyzing number of frames. Some algorithms can analyze small number of frames while others can analyze a greater number of frames per second. Output time also depends on the GPU. A 16GB GPU will perform faster while 8GB GPU will not give faster results.

2.6. Literature

Recognition of AD started with the detection of “Fall”. Many algorithms were developed to detect the fall of a person. During the onset of dementia in a person, falls start to happen. But there’s more to dementia than just the fall.

First trial of experiments to detect fall of a person was conducted by Lee and Mihailidis in 2005. The dataset was created of 126 falls and 189 non-falls [10]. The experiment was conducted in a bedroom scenario (fake scenario). No real patients were recruited, and the environment was also made-up. Subjects performed different 5 scenarios according to the instructions provided to the subjects. The experimental analysis achieved 77% accuracy.

In 2007, Chia Weng and Zhi Hong [11] performed fall detection by recording the dataset of 78 videos which consisted of fall in different scenarios and possible ways. They focused on 3 different parameters and used those parameters of silhouette, histogram, and fall to detect 3 different action types, that is, fall, squat, and walking. They used the collected dataset to differentiate walk, fall and squat conditions.

In 2011, Aertssen et al [12] performed information extraction and tried to use images to extract information for few activities that included walking, getting up or bending down among the elderly people. In the same year fall detection was again performed by the researcher Auvinet et al [13]. They created a dataset of real life and fake falls by installing cameras in few different locations in a room where falls were simulated. Belshaw et al [14] also tested fall detection. The dataset was

created in an office environment rather than the bedroom environment. Fall action were recorded, annotated, and divided into training and testing data. The dataset was scripted and not real time.

In 2015, work recognition of daily activities was done by Abidine et al [15]. This work proposed the methods of “independent component analysis, and linear discrimination analysis features with weighted support vector machines”. In year 2016, another group of researchers, Berlin and John [16], used Harris corner effect to recognize some basic activities such as hand shaking, kicking, punching, pushing, and hugging. They used deep neural nets to recognize these basic ADL. For these small activities they archived the accuracy of 95%. The dataset used was also scripted and the shorts were taken with keen precision. The dataset was not real-time and there were no issues of depth and distance while recording the data. With this reason, the training and testing of the dataset was done rather with ease as compared to the real time dataset. In 2012, posture detection which consisted of sitting, standing, squatting, and lying, was performed. These activities were recognized using fuzzy rules by Brulin et al. This achieved the accuracy of 74.29 percent [17]. In 2015, another research was performed [18] to recognize the ADL. In this research the ADL activities included sitting, standing, squatting, and lying.

For this, 7000 images were used to extract the skeleton data and Recurrent neural networks were applied to determine the posture.

In 2016, Chen et al, performed recognition of ADL by creating 20 different actions. The experiment consisted of cross subject tests and the tests were divided into training and testing sets. Skeleton based features were extracted to recognize the activities. This experiment achieved the accuracy of 96.1 percent [19]. But in this dataset, which was used to extract the skeleton-based features to recognize the activities was scripted dataset and the actions were not performed in real time. The subjects were provided the script as to what actions are supposed to perform. The same scripted dataset was divided into learning, training, and testing sets. In 2016, another research was performed in attempt to recognize the ADL by Huang et al. Lie group feature extraction and lie group networks were used 40 subjects performed 60 different activities [20]. The dataset used in this research was the largest 3D dataset used up until that time. The accuracy achieved by in this research was 89.10 percent [20]. Critical analysis of this research reveals that the actions performed by the subjects were scripted. Scripted actions tend to give more accuracy when trained and tested as compared to non-scripted and real time dataset.

In 2012, another approach was used to detect the fall of a person by using the techniques of “motion feature extraction and moving object extraction” [21]. The dataset that was built and used in this research was small and the dynamics of human body and motion were focused.

Lan et al in 2015 [22], proposed another method to recognize the ADL. Support vector machine technique was applied and for training and testing 4 different datasets were used. The datasets used were UCF101 consists of 101 different categories of action. The action performed in this dataset were extracted from YouTube videos. This dataset has more variation in terms of camera motion which made this dataset as one of the challenging datasets to work on. The 2nd dataset used in this study was UCF50. This dataset consists of 50 different types of categories of action. The actions data was extracted from YouTube videos. UCF101 is the extension of the dataset UCF50. 3rd dataset used in this research was HMDB51 which is a large dataset consisting of different human motions recorded. The actions performed in this dataset have been extracted from movies, web videos, YouTube, and other various sources. This dataset consisted of 51 different categories of action being performed by a human actor.

Lie at al in 2012, [23] performed research on fall detection and in 2016 the research was performed on the recognition of ADL. For the fall detection research in 2012, the dataset was created in laboratory environment and subjects were trained by a nurse to perform fall action to make it look like the real fall. Mattresses were used to fall on during the recording of this dataset. Later some more data was collected in an apartment environment instead of the laboratory environment. Subject performed the actions of fall according to the training and script provided.

In 2016, in action recognition research project a public dataset was which consisted of 16 different activities. Deep learning neural network approach were used to detect the ADL. In 2014, [24] et al used the information of space and time interest points along with optical flow features to recognize the activities if daily living using three different publicly available datasets. The datasets were used were same as used by Lan et al in 2015 [25]. The approach of Peng was different than Lan. Peng used the techniques of Fischer kernels and support vector machines while later Len used the techniques dense activity trajectory techniques with support vector machines. even though the techniques were different but due to the use of the same datasets the accuracy achieved by both the studies and research was around 94 percent with a difference of 2.1 percent.

Mo et al in 2016 [26] used different approach and used deep neural networks to model and recognize 12 different ADL. Shahrouday et al, in 2015 used different dataset to recognize the ADL. Ang used the techniques of robust feature extraction and support vector machines to extract the action features [27]. 3 different datasets were used. MSR Daily Activity dataset Figure 3, which consists of 16 different daily activities performed in a living room environment either while sitting on sofa or while standing near the sofa. The basics activities included in this dataset are sitting, drinking, eating, tossing paper, reading book, playing guitar, suing laptop etc. This dataset not only involves basic human action, but this dataset also involves the interaction with objects also called “Human Object Interaction” which makes this dataset different than other action datasets used to recognize the ADLs.

Figure 3 MSR Daily Activity Dataset

source: https://wangjiangb.github.io/my_data.html

MSR Action 3D (Figure 4) was the 2nd dataset used in the study. This dataset consists of 20 different activities performed by 10 different subjects. In this dataset depth camera was used to record actions. This dataset consists of actions like waving,

hammering, punching, throwing, hand clapping, drawing circle in air, jogging, tennis swing, golf swing, bending etc.

Figure 4 MSR Action 3S Dataset

3D Action Pairs Dataset was the 3rd dataset used in the study of Shahrouday et al. In this dataset actions are in pairs in a way that each pair has similar trajectories and objects. For example, push/pull, pickup/putdown etc. In 2016, Shi et al also attempted the recognition of daily living activities. He used deep learning approach for this purpose using RNN and CNN. Although the datasets used were same as

used by Peng et al and Lan et al. The only different dataset was KTH dataset. This dataset consists of 6 different actions including walking, running, jogging, clapping etc. What makes this dataset different from other datasets is that the actions were performed by the 25 different actors and each action was recorded in different environment/setting. Other 2 datasets used by Shi et al were HMDB51 and UCF101. For that reason, this approach also achieved the accuracy closer to the 2 previous approaches used to recognize ADLs, that is, 96.8 percent (other 2 were 92.3 and 94.4 percent which also used the same datasets).

In 2014, Simonyan and Zissermen used deep learning with temporal streams based on optical flow using 2 different datasets achieving 88 precent accuracy. Veeriah et al in 2015, used RNN on the MSR Action 3D dataset and KTH. These datasets were already used by other authors with different techniques. The accuracy after this technique was also around the accuracy achieved by other techniques due to same dataset.

Uddin et al in 2017, used random forests techniques to obtain body skeletons and applied hidden Markov model while public dataset consisting of 12 different ADL. They also used Deep neural networks for gait analysis.

Wang el al in the 3 consecutive years tried to recognize ADL using different techniques, algorithms and datasets. In 2014, wan g used local occupancy patterns with support vector machines to recognize 12 different activities. In 2015, wang used pseudo color images with CNN on 4 datasets, that is, MSR Action 3D, MSR Action 3D Ext, UT Kinect, MSR Daily activity. These datasets were also used by other researchers during that time. In 2016, wang used same dataset but used different approach to recognize the activities in the dataset, that is, depth motion maps with CNN.

Yang et al in 2017 also used same dataset as used by wang et al and applied different approach to recognize the activities. Yang used the approach of linear classification. All of these were able to achieve good accuracies. Zhu et al in 2016, deep RNN along with long short-term memory approach and used the datasets of HDM05, SBU Kinect and CMU. SBU Kinect consists of 8 actions performed by 7 subjects. This dataset involves the interaction between 2 persons. CMU is a large dataset with 65 videos and includes the interaction of multiple people in a social environment.

In all the literature discussed above we find that many researchers tried different algorithms and approaches while using various kind of datasets to achieve accuracy

in the recognition of ADL. But those approaches when tested in real time are unable to perform well and are not able to identify or classify the action being performed correctly. This problem is discussed by [5] where they state that all these action datasets consist of simplistic action and are not more realistic. Most of the datasets deals with action like walking, jogging, running, hand waving, clapping, sitting, standing etc. and the variation of backgrounds in those recorded datasets is also less. Some datasets tried to introduce closer to realistic datasets for a set of few simple activities while using different cameras and various backgrounds. In other kinds of datasets, the actions are usually extracted from YouTube videos or Hollywood movies etc. Although variation in background, actors and cameras made those datasets difficult to deal with and new algorithms were developed and applied to extraction action information from those datasets but after training and learning very few algorithms were tested in real environments under realistic conditions. Some other datasets were also recorded to achieve realistic environment goals. They were recorded in kitchen or in living room. But those datasets were small and were depicting only a few activities. For example, preparing a food using single recipe. Regarding the approaches that were used to extract action information included holistic based approaches or local based approaches [5]. Some algorithms detected action based on motion information [5].

When it comes to activity recognition for people with dementia. Data collected for normal actions and the techniques and algorithms used to detect and recognize those actions does not help enough in recognizing the ADL for people with dementia. For that purpose [5], recorded their own dataset of dementia patients at Alzheimer institute. They used sensors in addition to video cameras to record the activities of people with dementia. Sample size in this dataset was "32 patients with mild cognitive impairment" [5] and 1 type of dataset was recorded. Eating, drinking, using phone and reading paper activities were asked to perform by the patients. They performed activities based on scripted scenarios in a kitchen environment which was set up artificially in a room (not the real kitchen). In the 2nd set of datasets recording 35 patients and healthy subjects were asked to perform same activities in addition to performing the activity of opening ang closing of closet. An additional video sensor was attached with their clothes to record the activities performed by the patients, closely. In the analysis of this study, all the activities performed were predefined. Patients were told what to do and what activity to perform. Patents performed each task according to the instructions told. Activities were performed in a set-up environment and not in the real environment. Many cameras and sensors were placed in the room to record the activities. Patients had no liberty of doing whatever they wanted to do instead they

performed provided set of tasks. Activity recognition and classification was performed in 2 steps for this dementia ambient care dataset. For activity recognition Harris 3D model and activity area algorithm was used. For the classification of activities K means clustering algorithms were used. 11 Different classes were introduced for classification purposes, for example, drinking, eating, reading etc.

Activities were not specified. For example, drinking from a glass or drinking from a bottle, both were classified as the action of drinking. The object used for drinking was not detected and specified. In another example scenario, in which patients were asked to read a paper while sitting on a couch they did not use tablet while reading. Hence, the difference between reading from the paper or reading from the phone or tablet was not distinguished. It was simply classified as the action of reading. In Addition [5] faced another issue while detecting the ADL which was many activities were confused with other activities (actions of similar concept). Eating a snack was confused with the activity of drinking beverage [5]. This confusion mostly happened in those conditions where 2 same activities were performed in the same environment, for example, kitchen. But this research [5] was one of the first steps taken to prove that video technology can be used to monitor the ADL of patients, in this case people with dementia.

The table summarizing the different studies on recognition of activities of daily living (ADL) and fall detection is as follows.

Year	Author(s)	Methodology	Dataset	Activities	Accuracy
2005	Lee and Mihailidis	Fall detection algorithms	126 falls, 189 non-falls	-	77%
2007	Chia Weng and Zhi Hong	Silhouette, Histogram, and Fall Parameters	78 videos	Walk, Fall, Squat	89%
2011	Aertssen et al	Information extraction using images	Small dataset	Walking, getting up, bending down etc.	68%
2011	Auvinet et al	Occulusion robust method	Real-life and fake falls	Fall	82%

2014	Belshaw et al	Automated analysis algorithm	Scripted fall detection	Fall	92%
2015	Abidine et al	Independent Component Analysis, Linear Discrimination Analysis, and Weighted Support Vector Machines	TK26M dataset	Random activities	60 – 90% based on activity
2016	Berlin and John	Harris Corner Effect with Deep Neural Nets	Small dataset	Hand shaking, kicking, punching, pushing, hugging	95%
2012	Brulin et al	Fuzzy Rules	NA	Sitting, standing, squatting, lying	74.29%

2015	NA	Skeleton data extraction using 7000 images	7000 images dataset	Sitting, standing, squatting, lying	NA
2016	Chen et al	Skeleton-based features	Scripted dataset with 20 actions	20 different actions	96.1%
2016	Huang et al	Lie Group Feature Extraction and Lie Group Networks	40 subjects, 60 activities	60 different activities	89.10%
2012	NA	Motion feature extraction and moving object extraction	Small dataset focused on dynamics of human body and motion	Random actions	88%

2015	Lan et al	Support vector machine	UCF101, UCF50, HMDB51	101, 50, 51 different categories of action	-
2012	Lie et al	Fall detection	-	-	84%
2016	Lie et al	Motion analysis with support vector machine	-	-	92.5%

Table 1 Literature

Chapter 3
Objectives

The covid-19 pandemic has highlighted the potential of digital health innovations to support older adults and their caregivers. This technology will support caregivers and patients. Not only in healthcare facilities but monitoring technology will also assist in smart homes for dementia patient care and will support independent living. This work will result in real-world solutions as the solution will help the caregivers at the healthcare center to monitor their patients.

The goal of this project is to explore and evaluate the activity of daily living recognition algorithms to support the assisted living of people with dementia in homes or at smart suites. To monitor the presence and activities of a person at home or room and using that information to be able to evaluate the efficiency of the algorithm and the classify and analyze various activities performed by the person with dementia.

This project is an attempt to create more specialized algorithms that can accurately monitor the ADLs of people with dementia. A deep learning approach to monitoring ADLs that utilizes a combination of both scripted and random datasets, with a focus on capturing the unique and unpredictable movements of individuals with dementia. This approach could be more effective in accurately recognizing and tracking the movements of people with dementia compared to traditional action recognition algorithms.

This project involve incorporating specialized datasets, such as those that specifically feature people with dementia, or incorporating advanced machine learning techniques that can better handle the variability and complexities of dementia behaviors. The objective is to propose a deep learning approach to monitoring ADLs that utilizes a combination of both scripted and random datasets, with a focus on capturing the unique and unpredictable movements of individuals with dementia. This approach can be more effective in accurately recognizing and tracking the movements of people with dementia compared to traditional action recognition algorithms. By doing so, we can improve the quality of life for people with dementia and support their caregivers in providing effective care.

3.1. Computer vision and dementia

We cannot replace the human care, but we can help families and caregivers look after their loved ones having dementia.

Monitoring people with dementia has always been a challenge. To support people with dementia and their caregivers, technology can help us in many ways. Various kinds of devices were developed to monitor the activities of the people with dementia. For example, infrared motion sensors were developed to detect the presence and motion of the person in the room. When it comes to monitoring people with dementia it is required to monitor a lot of things in various ways possible for example, falling of the person, interaction, and distance of the person from various objects in room, flushing toilets, opening, and closing of cupboard/cabinet doors, refrigerators, dressers etc.

It is also required to monitor information as the people with dementia have issues with memory, forgetfulness which sometimes leads them to perform activities in a repetitive way and doing things which are not necessary. For example, opening the refrigerator again and again.

To monitor all these activities, many locating, monitoring and Sensor devices are usually placed at different places around the house. The devices include infrared sensors, ultrasonic sensors, photoelectric sensors, vibration sensors, audio sensors, pressure sensors (pressure mats), magnetic switches, wattmeter etc. to monitor various kinds of ADL among people with dementia.

Among these devices, there are 2 types: Wearable devices and non-wearable devices.

When it comes to people with dementia wearable devices cause more problem as compared to non-wearable devices. So, it is necessary to know how the person with dementia feels about wearing the device or having lots of monitoring devices around.

People with dementia require personal independent and freedom and people think that by attaching wearable devices with people with dementia will increase their safety for example attaching a GPS monitor with people with dementia (on clothes or wrist). This is done for the purpose of safety by caregivers and care partners. But while such technologies provide safety for people with dementia and peace of mind to the caregivers and care partners, research has shown that it also hinders in the personal independence and freedom of the people with dementia, and they do not

feel at ease and consider it as invasion in privacy. If the person with dementia is in the early stages, there is possibility that caregivers and care partners can discuss with them about using the devices on and around the for the safety purposes but for a person with middle and late-stage dementia it will be difficult to discuss with them and because of the changes in their mood and behavior they will not understand and act in aggressive way. This aggression was also proved in our experimental setup with the person with dementia in middle stage of the disease. When tried to enter the room and place any monitoring device in the room person with dementia acted in aggressive way and threw everything on the floor. Person with dementia did not appreciate any intruder or disturbance in the room and preferred independence and alone time. At that time, we ended up placing a video monitoring device at one corner of the room. Nothing more than that.

When it comes to wearable devices there are many things that need to be considered before using them especially for people with dementia. Every person is unique similarly each person with dementia will have a different behavior. Behavioral changes among people with dementia evolve over time so there is no standard device to monitor the ADL among people with dementia. In addition, People with dementia forget to wear the devices (wearable device) due to cognitive impairment and issues with forgetfulness and memory. For the device to work, it is

necessary that it is worn on regular basis. But this poses a challenge with people with dementia due to their lack if cognitive abilities and forgetfulness. In addition, people with dementia do not like to wear anything extra on them as they feel uncomfortable with that and end up removing that device from them.

With the advancements in technology, many technological devices have been introduced in healthcare. But regardless of the advancements the accuracy of those technologies and devices is not perfect. “No technology is Fail Proof” as stated by Alzheimer Society Canada [2]. Technology cannot help under all circumstances, and it cannot replace the care, a caregiver or care partner can provide to the people with dementia. Many algorithms were trained to monitor the ADL but those are trained on the ADL of a normal person with normal and random activities. Those trained algorithms are unable to give good results and perform better while monitoring the ADL for people with dementia.

For the non-wearable devices, they are usually placed at many nooks and corners of the home or room. There will a sensor on each door and sensors on refrigerator doors, cabinets, cupboard etc. If there are no doors in the house for the purpose of safety for the people in dementia, then there will be sensors on the entrance of every door to detect the entering and exiting of the room. There will be sensors on

the switches to detect their usage. Use of infrared and ultrasonic sensors at the place. To monitor the standing and sitting time and frequency of the person a sensor is placed on the chair. All this combined makes the people with dementia uncomfortable. In addition, placing a lot of sensors at the caregiving place also increases cost and work to manage those monitoring devices. This also adds to the discomfort of the people with dementia. For example, what if the chair which has a sensor placed on it is replaced with a new chair? What will happen to the sensor in that case? To avoid this, need of a single sensor that is able to detect most of the ADL is crucial.

Instead of using many devices and sensors to monitor the ADL among people with dementia video sensors were introduced. But they need to be trained on an appropriate dataset and in an appropriate way.

In this research we will evaluate the deep learning algorithm to monitor the ADL for people with dementia using video monitoring.

Chapter 4
METHODOLOGY

4.1. Selection of Dataset

Most datasets used for training and testing of various algorithms and techniques developed to recognize the ADL are usually scripted and staged by actors. Scripted dataset collection usually consists of similar background and similar viewpoints. Which makes learning and training easier. But these algorithms trained on such dataset fails in the real time as in real time backgrounds, viewpoints, various features, various camera angles and has more variation which makes the real-time recognition of daily activities a challenge to date.

After all these years of the development of various algorithms and techniques to recognize the ADL and yet it is still difficult to recognize the ADL considering the various and complex behaviors among patient with dementia. Although those trained algorithms and various techniques when tested gave around 80-90 percent results but they are unable to perform good when tested under real conditions.

This is due to many reasons and the main reason was the dataset. Each time the dataset that was used to train the algorithm was scripted and the actions were performed by the staged actors. Actions performed by staged actors are different than the real-time actions. One subject that is pretending to fall will be different than the subject who actually fall while walking.

Another factor that makes the action recognition difficult is the involvement of objects while performing any activity. For example, in the process of drinking from the glass there exist an action of drinking along with the involvement of an object which in this case is glass. Human object interaction in any dataset makes the recognition of the activity process different from the simple activity recognition for example “walking” in which no object is involved except the person himself. Dementia Ambient Care Research [5] was one of the first steps taken to prove that video technology can be used to monitor the ADL of patients, in this case people with dementia.

Toyota Smart home dataset recorded unbiased activities of older adults. Those were not dementia patients. Dementia patients are usually older adults. So, Recording the older adults was a good step in collecting the dataset for recognizing the ADL. Deep learning neural networks perform better with larger datasets.

Performance of these algorithms is greatly affected by the availability of the annotated datasets. AS discussed in the literature review section, most datasets collected for the recognition of activities consisted of videos and shorts takes from web sources, movies, YouTube etc. Such datasets include HDMI, Kinect, UCF, MSR, KTH etc. But these datasets do not contain the challenges of real-life ADL. A scripted action is always different than a real-life action performed in a natural way. In addition, biasness in the dataset was also found in terms of action. Many datasets included various activities which were not the part of the ADL. Many datasets involved actions like playing tennis, throwing ball, playing soccer, or various other sports activities. Training the algorithm on such dataset will not give the accuracy on the activities of real world and daily living activities. the datasets which were used to train the algorithms consisted of video clips of few seconds. A human performing an action for a few seconds to record the video for the dataset. For example, walking phenomenon was recorded for a few seconds and hundreds of walking videos were recorded and extracted from various sources on the web consisting of few seconds. Activities were performed by actors who were instructed what action to perform.

In a nutshell, Datasets which were proposed in the past many years for the purpose of recording the ADL had the following issues.

- Consisting of few seconds
- Unnatural action/activities
- Biased activities
- Similar and simple backgrounds (not complex)
- Also included sports activities along with other basic activities
- Activities recorded were not the ADL but rather random activities were recorded
- Video dataset was recorded/gathered from web sources (e.g., YouTube, movies etc.)
- Inter-class variance (activities like playing tennis and bike riding)
- Activities were recorded using static cameras
- Single point of view of the camera (activities look different from different viewpoints and angles)
- Actors performed activities in front of camera
- Activities were scripted performed by actors
- All the activates performed were instructed
- Voluntary people performed actions for the video dataset
- Activities were performed in an unnatural way
- Many recorded videos and actions were similar

- Lack complex and composite ADL. most of the actions performed were simple actions
- Videos were usually short and atomic action
- Datasets did not reflect the challenges of real life, unscripted activities

After Exploring various datasets, it was concluded to use the Toyota Smart home dataset for this thesis project.

This dataset poses the challenges involved in the real world. Other datasets do not pose the real-world scenario challenges in their dataset videos and shorts. This dataset contains the variety of daily life activities performed in the real environment. In this dataset one video does not depict only one activity instead multiple activities in a single video of the dataset.

The next problem was that the current approaches and algorithm to identify the ADL which were trained on other datasets (not Smarthome dataset) were unable to perform on the Smart home dataset.

4.2. Toyota Smart Home Dataset

The Toyota Smarthome dataset is a collection of videos that were recorded in an apartment, using seven Kinect v1 cameras [28]. The dataset contains videos of 31

daily living activities performed by 18 subjects, all of whom are seniors in the age range of 60-80 years old. The subjects were informed of the recording, but not of the study's purpose, to ensure unbiased activities. Each subject was recorded for 8 hours a day, from morning until afternoon, without any scripted activities. The videos were analyzed and annotated to identify the 31 different activities, resulting in a total of 16,115 video samples. The dataset is available in three modalities, RGB, Depth, and 3D skeleton, with a resolution of 640 × 480. The 3D skeleton joints were extracted from the RGB modality using the LCR-Net. It consists of 16K short RGB+D videos of 31 classes of ADLs. Each video is about 12.5 sec. To protect the privacy of the subjects, the faces in the videos were blurred using the tiny face detection method.

The dataset presents a range of difficulties in accurately recognizing natural and diverse activities. Firstly, Due to the absence of a predetermined script and the subjects engaging in regular daily activities, the samples for different activities are not uniformly distributed. Secondly, the distance between the camera and the subjects varies significantly across the videos, and in some cases, the subjects may be partially obscured. Finally, the dataset comprises a diverse range of activities with varying levels of complexity.

4.3. Dataset challenges

The dataset that is used has a high intra-class temporal variance. High intra-class temporal variance refers to the situation where there is a large variation in the duration of activity within the same class. For example, the activity of cooking can last for different durations, depending on the complexity of the dish being prepared. This variation in the duration of the same activity makes it difficult to recognize and label the activity in a video, leading to a high intra-class temporal variance.

The dataset also has a high class imbalance. High-class imbalance refers to a situation in a classification problem where the number of instances belonging to one class is significantly higher compared to instances belonging to other classes. This can result in a machine learning model that is biased towards the majority class and performs poorly on the minority class. This is because the model is trained to predict the majority class more frequently, and therefore has a higher accuracy when predicting that class. To overcome this issue, techniques such as oversampling, under-sampling, and class weight adjustments can be used to balance the class distribution.

An example of a high-class imbalance in the context of monitoring the ADL of a person with dementia can be if the majority of the activities performed by the person are non-challenging activities, such as sleeping or watching TV, whereas the number of instances of more complex and challenging activities, such as bathing or cooking, are much fewer. This creates an imbalanced distribution of the different activities in the dataset, leading to a biased performance of the activity detection system towards the more frequently occurring activities.

Spontaneous behavior refers to actions that are performed in an unplanned and impulsive manner, without prior intention or deliberate thought. In the context of ADL of a person with dementia, spontaneous behavior would include actions that are not prompted by a specific task or goal but are instead performed without a clear motivation or purpose.

Concurrent activities, on the other hand, refer to actions that are performed in parallel or at the same time as other actions. In the context of ADL of a person with dementia, concurrent activities would include actions performed while performing another main task, such as reaching for an object while walking, or speaking while eating.

4.4. Current algorithms and approaches

Various approaches were then developed and used to detect the ADL using the smarthome dataset.

The algorithms and approaches which were developed to train on the Smarthome dataset were:

- VPN++
- AirStream
- Separable STA
- Assemblenet, Assemblenet, Assemblenet plus lite
- UNIK
- Dense Trajectories

Only a few of those gave accuracy around 70% or more with Toyota Smarthome real world challenges dataset. Among those, which were able to provide a rather higher accuracy required extra training dataset.

VPN++ used the methodology of combining RGB and 3D poses to detect the ADL. But “the cost of computing 3D poses from RGB Stream is high” as stated by [10] So an extension of pose driven mechanism model was proposed and pose knowledge

was transferred to RGB and then pose driven mechanism was mimicked through an attention level distillation. All this combined into a single model called VPN++. But all this involved the usage of extra training data for the algorithm to work and achieving high accuracy.

MMNet used the model based multimodal approach recognition of human actions and activities in RGB videos. MMNet also required extra training dataset for achieving high accuracy.

Assemblenet++ used the methodology of “Modality Representation via Attention Connection” [11]. The accuracy of this approach was around less than 70% but they did not use the extra dataset.

AirStream used the “adaptive intermediate representation” method, but the accuracy was below 70% although they did not use the extra training dataset.

Dense Trajectories used the cross-stream approach but was unable to achieve high accuracy. This approach was one of the few initial attempts to work on the smart home real life dataset. This approach also did not require the extra training dataset.

It was noticed that those algorithms and approaches which were able to achieve a rather high accuracy as compared to others required extra training dataset. Those

which did not use the extra training dataset were unable to achieve very high accuracy.

For this purpose, the goal of this thesis was to develop an approach that will not require the extra dataset but is able to achieve higher accuracy in recognizing the ADL. For training and validation of the dataset, SmartHome Dataset with various challenges involved was used instead of the other regular datasets.

4.5. Labels

Labeling videos is a crucial step in monitoring the ADLs of a person with dementia as it provides important information about their behavior and habits that can help in the early detection and management of the condition. By accurately labeling the videos, we can identify changes in the person's ADLs, such as an increase in repetitive behaviors or a decrease in mobility, which can be early indicators of a decline in cognitive function. The labeled videos are used to train machine learning algorithms to automatically detect and monitor the ADLs of a person with dementia.

4.6. Coarse and Fine Grained lables

The activities in the dataset are labeled with both coarse and fine-grained labels. Coarse-grained labels and fine-grained labels refer to the level of specificity of the labels in a classification task.

Coarse-grained labels are broader, more general categories that encompass a range of similar or related items. For example, in a classification task for animal species, the coarse-grained label could be "mammal," which would encompass multiple specific species such as "dog," "cat," and "human."

Fine-grained labels, on the other hand, are more specific, providing detailed subcategories that distinguish between similar or related items. In the animal species example, fine-grained labels could be "Golden Retriever," "Siamese," and "Homo sapiens."

The choice between coarse-grained and fine-grained labels depends on the goal of the classification task, the amount of available data, and the computational resources available. Coarse-grained labels can be easier to work with, as they require fewer data and computational resources, but fine-grained labels can

provide more detailed information, allowing for a higher degree of accuracy in classification.

4.7. Annotation Format

The COCO (Common Objects in Context) annotation format is a standardized way of representing object instance annotations in an image dataset. To convert a dataset into the COCO format, the object instance information, such as the bounding box coordinates and the class labels, for each image in the dataset are extracted. The information is then formatted as a JSON file following the COCO annotation format specifications, which include fields such as image information, annotations, and object categories.

4.8. Coco Dataset

The COCO (Common Objects in Context) dataset is a large-scale image recognition dataset that contains over 330,000 images, each annotated with 80 object categories and captions describing the objects and their relationships within the images. It is widely used for training and evaluating computer vision models, especially for tasks such as object detection, instance segmentation, and image captioning.

It was created by the Microsoft COCO team and contains over 330,000 images, each annotated with 80 object categories and various attributes, such as object location, size, and occlusion. The COCO dataset also includes more than 2.5 million object annotations and 330,000 caption annotations, making it one of the largest and most comprehensive datasets for training and evaluating computer vision models.

4.9. Coco Annotations

The Microsoft Common Objects in Context (COCO) dataset was annotated using Amazon's Mechanical Turk, a platform that provides a way to hire human workers to perform tasks that are difficult to automate [29]. For the captioning task, workers were shown an image and asked to describe the image in natural language. The captions were then reviewed by another set of workers to ensure they were accurate and grammatically correct. The captions were then integrated into the COCO dataset using a specific annotation format.

4.10. Image captioning

In the COCO dataset, the dataset is annotated with 5 captions per image [26] and the captions are free-form natural language descriptions of the objects in the

images. The captions in the COCO dataset are used to train and evaluate image captioning algorithms.

The annotation format of the COCO dataset is in JSON (JavaScript Object Notation) format [26], which is a lightweight data-interchange format. The JSON file contains the annotations of the images, including the captions, the object instances, and the categories of the objects.

The basic structure of the COCO annotation file is as follows:

- The file starts with a header that includes information about the version of the COCO dataset, the copyright information, and the contributors.
- The file then includes a list of all the images in the dataset, including the file name, the height and width of the image, and the ID of the image.
- The file also includes a list of all the annotations in the dataset, including the captions, the object instances, and the categories of the objects.
- Finally, the file includes a list of all the categories of the objects, including the ID and the name of the category.

In conclusion, the COCO dataset annotation format for image captioning is a well-structured and standardized format that provides a comprehensive and consistent

representation of the annotations of the images. The JSON format of the annotations makes it easy to parse and process the annotations, making it a widely used and well-accepted format for image captioning.

4.11. Converting our dataset to Coco Annotation format

COCO dataset uses JSON format to annotate their images for image captioning. JSON stands for JavaScript Object Notation, and it is a lightweight data-interchange format that is easy for humans to read and write, and easy for machines to parse and generate. It is a text format that is completely language-independent but uses conventions that are familiar to programmers of the C family of languages. In the COCO dataset annotation files, the JSON format is used to store information about the objects and their properties in the images, as well as the annotations for the captions of the images. The JSON format is a commonly used data format in computer science, and it is widely supported by programming languages and tools.

The JSON files contain information about the images, including their unique IDs, file names, and sizes, as well as the annotations for each image. The annotations include the object categories and instance IDs for each object in the image, as well as the captions for the image, which are represented as a list of sentences. Each

sentence is represented as a string of words, and the captions are associated with their corresponding image IDs.

The annotations include information such as image ID, caption ID, caption text, and other image and caption-specific information. The JSON format makes it easy to store and access this information for machine learning models.

In our approach, we converted the Toyota SmartHome Dataset annotations into Coco annotation format.

Chapter 5
APPROACH

In our approach we used ExpansionNet v_2 Model for Image captioning for the Toyota Smart Home Dataset.

5.1. Scene Understanding

Scene understanding in computer vision is the process of extracting semantic information from an image or a sequence of images to understand the environment and the objects present in it. It involves analyzing the image at various levels, such as low-level features like edges and textures, mid-level features like segments and regions, and high-level features like objects, scenes, and their relationships. Scene understanding helps in tasks like object recognition, object tracking, scene classification, and scene reconstruction. The output is a complete sentence or a short paragraph that describes the objects and their relationships within the image. Image captioning is used in applications such as image and video retrieval, assistive technology for visually impaired individuals, and content creation.

It involves extracting semantic relationships and patterns.

Extracting semantic relationships and patterns refers to the process of analyzing data to identify meaningful connections between different entities or concepts. This involves identifying patterns or trends in the data and understanding how different data points are related to one another.

In computer vision, extracting semantic relationships and patterns often involves analyzing visual features of images or videos to identify objects, people, or other elements within a scene, and then using machine learning algorithms to identify patterns and relationships between these elements. This can be useful for a variety of applications, such as image recognition, object tracking, or video monitoring.

Scene understanding in computer vision involves analyzing an image or video to identify the objects, their properties, and their interactions with the environment, as well as the context and layout of the scene. Extracting semantic relationships and patterns is a key part of scene understanding, as it helps to build a more comprehensive and accurate understanding of the scene.

To extract semantic relationships and patterns, computer vision algorithms use techniques such as object recognition, object detection, and semantic segmentation to identify and label the objects in the scene. They can also use

spatial and temporal analysis to identify patterns in the arrangement and movement of objects, such as trajectories or groupings, which can reveal important information about the scene and its context.

In terms of monitoring ADL for a person with dementia, scene understanding can involve identifying the objects and people present in a scene, as well as the activities being performed. This can be achieved by analyzing the semantic relationships between the objects and people, as well as the patterns of activity occurring in the scene.

For example, a scene understanding system may be able to detect when a person with dementia is preparing a meal in the kitchen by identifying the objects in the scene (e.g., pots, pans, and utensils) and the actions being performed (e.g., chopping vegetables, boiling water, or stirring a pot). By extracting semantic relationships and patterns from the scene, the system can monitor the person's ADL.

5.2. Image Captioning

Image captioning and scene understanding are related but distinct tasks in computer vision.

Image captioning involves generating a natural language description of an image, which requires a deep understanding of the content and context of the image. It involves not only recognizing objects and their attributes but also understanding their relationships, scene semantics, and the overall context of the image.

Image captioning is a form of scene understanding that involves generating a textual description of an image, whereas scene understanding is a more general task that encompasses a wide range of image analysis tasks.

5.3. Choice of model - ExpansionNet v_2 Model

The ExpansionNet V2 model is an image captioning model that uses the standard encoder-decoder architecture, which consists of an image encoder and a language decoder [30]. In this model, the image encoder is based on the Swin Transformer, a transformer-based model that uses a hierarchical structure to process images. The Swin Transformer is used to encode the input image and generate a set of features that are used by the language decoder to generate a natural language description of the image. The language decoder is an autoregressive model that generates the caption one word at a time, conditioned on the previous words

generated. The model also includes additional modules for semantic expansion and attention, which help improve the quality of the generated captions.

The training objective of the ExpansionNet V2 model for image captioning is to minimize the negative log-likelihood of the ground-truth caption given the image. The model is trained in an end-to-end manner, using a combination of stochastic gradient descent and the Adam optimizer, to optimize the cross-entropy loss between the predicted and ground-truth captions.

The training algorithm used for the ExpansionNet V2 model is the standard backpropagation algorithm with Adam optimizer. During training, the model learns to optimize the loss function, which measures the discrepancy between the predicted captions and the ground truth captions. The model is trained end-to-end, meaning that both the encoder and decoder are updated during training to optimize the loss function. The training process involves feeding the model with input images and corresponding captions and adjusting the parameters of the model to minimize the loss. The process is repeated over multiple iterations until the model converges to a satisfactory solution.

Adam Optimizer:

Adam (Adaptive Moment Estimation) is a popular optimization algorithm used in deep learning. It combines the benefits of two other popular optimization algorithms, Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSprop), to provide an efficient method for updating the weights in a neural network during training.

Adam optimizer calculates a moving average of the gradient and the squared gradient and then uses this to update the weights. It helps to speed up the training process and allows the model to converge faster.

5.4. Bounding Boxes Generation

Bounding boxes were generated while detecting the ADLs. Those bounding boxes were generated using the faster-RCNN model. A Faster R-CNN model with a ResNet-50-FPN backbone was used.

Faster R-CNN with a ResNet-50-FPN:

Faster R-CNN with a ResNet-50-FPN model is a popular object detection model that combines the Faster R-CNN architecture with a

ResNet-50 feature extraction backbone that uses a Feature Pyramid Network (FPN).

The ResNet-50 architecture is a deep convolutional neural network that has 50 layers and is used to extract features from the input image. The FPN is added on top of ResNet-50 to address the issue of scale invariance in object detection. FPN generates a feature pyramid with multi-scale features that can be used to detect objects of varying sizes.

Faster R-CNN is an object detection model that consists of two modules: a Region Proposal Network (RPN) and a Fast R-CNN detector. The RPN is responsible for generating object proposals, which are regions of the image that are likely to contain objects. The Fast R-CNN detector is responsible for classifying the proposals generated by the RPN and refining their bounding boxes. The RPN and Fast R-CNN detector are trained together in an end-to-end manner.

By combining the ResNet-50-FPN backbone with the Faster R-CNN architecture, the model can achieve state-of-the-art performance on object detection tasks and generating bounding boxes.

The Faster R-CNN model with a ResNet-50-FPN backbone was primarily designed for object detection in images, and it can be trained to recognize different objects or activities, including those related to ADL. We used it for ADL detection in older adults and people with dementia, Toyota Smart Home dataset of videos that show different ADLs of older adults was used and annotated with bounding boxes that indicate the location of different objects and/or activities in the scene.

5.5. Fine tuning and training on Toyota Smart Home Dataset

We used the ExpansionNet v_2 model which was trained on the COCO dataset and did end-to-end model fine-tuning. Fine-tuning is a common technique used in transfer learning, where a pre-trained model is further trained on a new dataset to improve its performance on a specific task. The idea behind fine-tuning is to leverage the knowledge learned by the ExpansionNet V_2 model from the COCO dataset and use it to learn the monitoring of ADL task on the Toyota SmartHome dataset and caption generation. Toyota dataset annotations were converted to coco image captioning annotation format.

The pre-trained model was used as a starting point to learn the new task of monitoring ADL. The pre-trained model was originally trained on a large COCO

dataset and learned features for the new task of monitoring ADL. We used this model as a starting point for a new task of image captioning.

Instead of training a new model from scratch, we used the pre-trained model as a starting point and then fine-tuned it for the new task by updating the weights of the model's output layer to fit the new task of monitoring ADL on the SmartHome dataset. In image captioning, the model needs to not only recognize objects but also generate a textual description of the image

We used the pre-trained model as a starting point and then fine-tuned the entire model to generate captions instead of object labels. This was to teach the network to extract useful features from the input data and learn how to perform the new task of monitoring ADL.

Fine-tuning the entire model without adding any new layers was beneficial in learning the new task of monitoring ADL because it allowed for the modification of parameters of the pre-trained layers that captured general features and patterns in the data. This technique is particularly useful when the new task is related to the original task, and the features learned by the pre-trained layers can be repurposed for the new task.

For that purpose, we used ExpansionNet v_2 pre-trained neural network with weights represented as W. We wanted to fine-tune the model for a new task, but we didn't want to add any new layers. So, we decided to adjust the weights of the existing layers to optimize for the new task of ADL.

L(W) is the loss function for the pre-trained network on the original task. We wanted to optimize a new loss function L'(W) for the new task of ADL by fine-tuning the weights W. It was done by using stochastic gradient descent (SGD) to iteratively adjust the weights in the direction of the negative gradient of the new loss function with respect to the weights:

W' = W - learning_rate * gradient(L'(W), W)

Here, the learning rate is a hyperparameter that determined the step size of each update, and the gradient is computed using backpropagation through the network.

During training, we updated the weights W' by iterating over the training data and computing the gradient of L'(W') with respect to W'. We then applied the update rule above to get the new weights W'' = W' - learning_rate * gradient(L'(W'), W'), and the process was repeated until convergence.

The key idea behind fine-tuning without adding new layers was to leverage the pre-trained weights as a starting point for the new task of ADL, and then adjust the weights to better fit the new data. This approach was particularly useful as it allowed us to transfer knowledge from the pre-trained model to the new task.

During training, the researcher followed a rule where three frames from a single video were selected, namely the first frame, the middle frame, and the last frame. These frames were used to generate proper captions and to detect the pose of the subject in the video.

By using more frames from a video, there is a higher probability of detecting the correct pose and generating a more accurate caption. This is because using only a single frame from a video may not capture the entire motion of the subject and using multiple frames can help to capture the different poses and movements of the subject over time. Additionally, using frames from different parts of the video (i.e., the first, middle, and last frames) can help to capture the different stages of the activity being performed.

The purpose of using multiple frames is to increase the probability of detecting the pose accurately and generating a good caption. By using more frames from a video,

the model has more information to work with and can potentially produce better results.

Fine Tuning involved training the entire network, on the pre-trained backbone, on the dataset. During this phase, the weights of the backbone were updated. In this process, the weights of the pre-trained network were used as the starting point, and the network was further trained with the SmartHome dataset to improve its performance on the new task of monitoring ADLs. The choice of batch size, beam size, and learning rate can affect the performance of the fine-tuned model.

The batch size determines how many samples are processed in each training iteration. Larger batch size can speed up the training process but may require more memory to store the gradients. A smaller batch size may lead to more stable training but may take longer to converge. In our case, we chose a batch size of 4, which is a relatively small value. This was based on the available memory on the machine we used for training.

The beam size is a parameter used in beam search, which is a common method for generating captions in image captioning tasks. It determines how many candidate captions are considered at each step of the decoding process. A larger beam size can lead to more accurate captions but may also increase the computational cost.

We chose a beam size of 3, which is a relatively small value. This was based on empirical results from previous work on image captioning.

The learning rate determines how much the weights of the model are updated at each training iteration. A smaller learning rate can lead to more stable training, but may also result in slower convergence. A larger learning rate can speed up the training process, but may also lead to unstable training and result in the model getting stuck in suboptimal solutions. When fine-tuning a model without adding any new layers, the learning rate can be chosen using techniques such as learning rate schedulers or by performing a grid search over a range of learning rates.

We chose an initial learning rate of 2e-4, which is a relatively small value. This was based on empirical results from previous work on fine-tuning pre-trained models. We also used a learning rate scheduler that reduces the learning rate by a factor of 0.8 every 3 epochs to help the model converge to a good solution.

A learning rate scheduler adjusts the learning rate dynamically during training based on some pre-defined criteria. For example, it could decrease the learning rate as the training progresses or when the loss function plateaus. We used the learning rate annealing approach. Annealing is a technique used in deep learning to adjust the learning rate of an optimization algorithm during training [31].

anneal_coeff was used to determine how quickly the learning rate should be reduced as the training progressed to enable the model to converge more efficiently. The anneal_coeff was adjusted based on the performance of the model during training to achieve the best results.

It's important to note that the choice of learning rate can have a significant impact on the performance of the fine-tuned model. A learning rate that is too high can cause the model to diverge, while a learning rate that is too low can result in slow convergence or get stuck in a suboptimal solution.

In summary, the model is fine-tuned using batch size 4, beam size 3, annealing coefficient 0.8, and an initial learning rate of 2e-4.

To perform the fine-tuning Pytorch library, numpy, PIL Image library, OpenCV, and Torch vision were used.

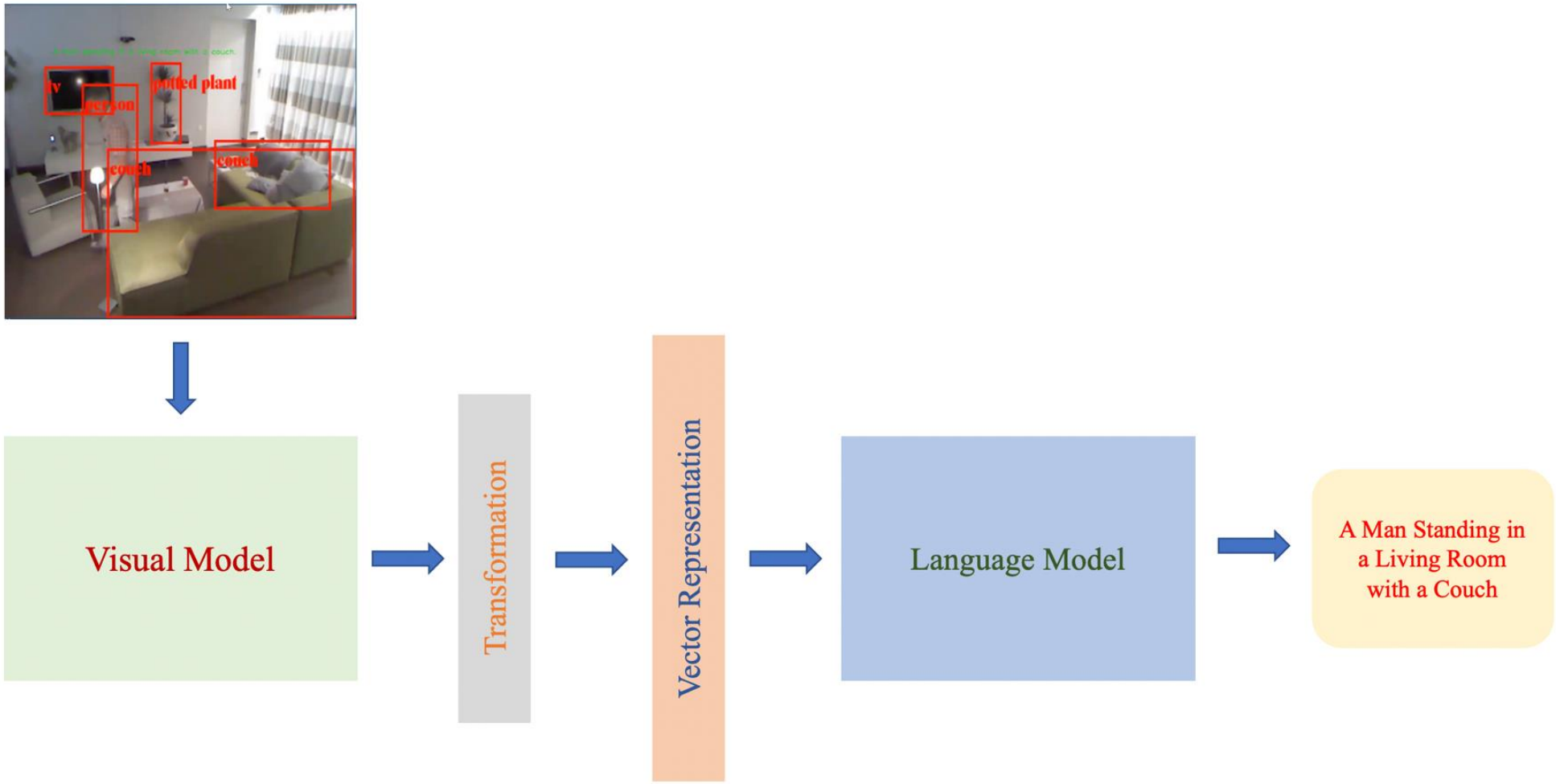


Figure 5 Framework

5.6. Loss Function

Cross-entropy loss function is commonly used in deep learning for classification tasks. It measures the difference between the predicted probability distribution and the true probability distribution for a given set of inputs.

In the initial training of ExpansionNet v_2 model on the Toyota Smart Home dataset, the cross-entropy loss function was used to optimize the weights of the model. The objective was to minimize the cross-entropy loss between the predicted and ground-truth labels of the training data.

In fine-tuning, the same loss function was used to optimize the weights of the entire model, including the previously frozen backbone layers. The objective was to further adjust the weights of the entire network to better fit the new dataset, in this case the Toyota Smart Home dataset.

Cross-Entropy loss function was used to provide a measure of the difference between the predicted and true labels and to guide the optimization process to improve the model's performance.

Cross entropy loss function is a popular loss function used in classification problems, particularly in deep learning. It measures the difference between the predicted probability distribution and the actual probability distribution.

In binary classification, the cross entropy loss function is defined as:

$$L(y, \hat{y}) = -[y * \log(\hat{y}) + (1-y) * \log(1-\hat{y})]$$

where y is the actual label (either 0 or 1), and $\hat{y}$ is the predicted probability of the positive class (between 0 and 1).

In multi-class classification, the cross entropy loss function is defined as:

$$L(y, \hat{y}) = - \sum y * \log(\hat{y})$$

where y is a one-hot encoded vector representing the actual class label, and $\hat{y}$ is a vector of predicted probabilities for each class.

To calculate the cross entropy loss, we first computed the predicted probability distribution for each input sample using the model's output layer. We then compared this predicted distribution to the actual distribution using the cross entropy loss function. The cross entropy loss for each sample was then averaged over the entire dataset to get the overall loss value for the model. The goal was to minimize this loss value during the training process.

After training, the model was used to detect ADLs in videos in real-time. Some test videos were also used for this purpose.

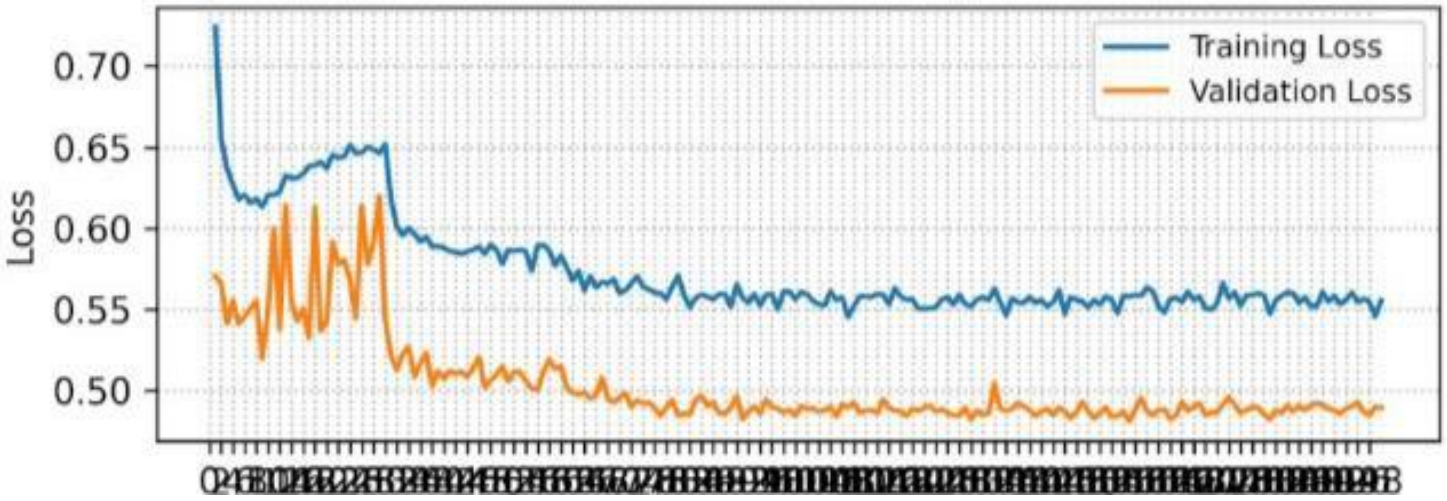


Figure 6 Total Loss Across Epochs

Chapter 6
Results & Comparison

The method involves using the machine learning and deep learning technology. Public dataset relevant to the aging and dementia have been identified for the project. In this case we used Toyota SmartHome Dataset.

Toyota Smart home dataset is a real-world dataset of ADL. The activities were performed by senior people in the age range 60-80 years old [32]. The activities performed by those senior people were not scripted. The activities were recorded while senior people performed their daily activities in a natural manner. This prevented the biasness of the scripted activities in the dataset.

Toyota SmartHome dataset was divided into 3 sets. The division of the dataset is as follows

- 70% training
- 20% validation
- 10% testing

For the purposes of the detection of ADL and the actions being performed.

The goal was to train the algorithm first on Toyota Smart home dataset and then testing on the Real-world scenario action dataset.

Training of the proposed approach was done, and 77% accuracy was achieved after training from scratch.

6.1. Performance Measure

Cross-entropy is a commonly used loss function in machine learning that measures the difference between predicted and actual probability distributions. It can also be used as a performance measure to evaluate the performance of a machine learning model.

In the context of image captioning, cross-entropy can be used to measure the quality of generated captions by comparing the predicted caption distribution to the ground-truth caption distribution. A lower cross-entropy value indicates a better match between the predicted and actual caption distributions, which indicates better performance of the model.

During training, the goal was to minimize the cross-entropy loss between the predicted and actual distributions. In other words, the model learned to predict

captions that were as close as possible to the ground-truth captions, resulting in a lower cross-entropy value.

After training, the cross-entropy value was used as a performance measure to compare the quality of generated captions across different models. A lower cross-entropy value indicates better performance in generating captions that are similar to the ground-truth captions.

6.2. BLEU

One widely used method for evaluating the quality of machine translation is the Bilingual Evaluation Understudy (BLEU) method. BLEU compares each translated segment with a set of reference translations with good quality, then calculates each segment score to estimate the overall quality of the translation [33]. In the field of image description, BLEU uses an n-gram matching rule as a similarity measurement method. To evaluate the BLEU metric, the co-occurrence frequency of n-grams in the predicted caption and label can be analyzed. The precision of the sentence for a given n-gram is expressed as follows when considering the predicted caption (Candidates) and the label (Reference).

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Conut_{clip}(n-gram)}{\sum_{\mathscr{C}' \in \{Candidates\}} \sum_{n-gram' \in \mathscr{C}'} Conut(n-gram')}$$

The evaluation of the BLEU metric involves several parameters. One such parameter is $Conut_{clip}$ (n-gram), which refers to the minimum number of times an n-gram should appear in the predicted caption (Candidates) and the reference caption. Another parameter is Conut (n-gram'), which denotes the frequency of occurrences of n-gram in the Candidates. When the number of tuples increases, the probability of the n-gram statistics (P_n) decreases, resulting in a shorter sentence. To avoid this, the final evaluation formula incorporates a brevity penalty factor.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

The value of the brevity penalty factor (BP) can be determined based on the length of the predicted caption (Candidates) and the reference corpus (Reference). Specifically, let c be the length of the Candidates and r be the length of the Reference corpus. The formula for BP is as follows:

$$BP = \begin{matrix} 1 & if\ c < r \\ e^{1-\mathrm{r}/\mathrm{c}} & if\ c > r \end{matrix}$$

To evaluate the quality of machine-translated sentences, the BLEU method was used, which involved dividing the sentences into words and counting how many

times the words in the translation appear in the reference sentence. The minimum number of times this occurs is recorded, and the ratio is calculated with the translation sentence. To avoid bias towards short sentences, a penalty factor was multiplied with the ratio to obtain the final result. In the case of BLEU-2, 2-tuples of words were used for statistical calculations, and up to 4-tuples are usually considered.

Lets take an example of ADL in our case. The calculation of BLEU score will be as follows:

Step 1: Tokenization

The first step is to tokenize both the reference and machine-generated sentences. Tokenization means splitting the sentences into words or subwords to prepare them for comparison. In this example, the tokenized reference and machine-generated sentences are:

Reference: ['A', 'man', 'sitting', 'in', 'a', 'living', 'room', 'with', 'a', 'couch']

Machine-generated: ['A', 'man', 'sits', 'on', 'a', 'couch', 'in', 'a', 'living', 'room']

Step 2: N-gram matching

Next, we compare the n-grams (contiguous sequences of n words) in the machine-generated sentence with those in the reference sentence. We compute the number of matching n-grams between the machine-generated sentence and the reference sentence for different values of n, ranging from 1 to 4 (or higher). For example, for n=1, the unigrams (single words) in the machine-generated sentence that match with those in the reference sentence are:

[‘A’, ‘man’, ‘in’, ‘a’, ‘living’, ‘room’]

Note that the word 'sitting' in the reference sentence does not match with 'sits' in the machine-generated sentence.

	A	man	sitting	in	a	living	room	with	a	couch
A	**1.00**	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	**1.00**	0.00
man	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sitting	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
on	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
couch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**
in	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
a	**1.00**	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	**1.00**	0.00
living	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
room	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00
watching	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Similarly, we compute the matching bigrams (n=2), trigrams (n=3), and 4-grams (n=4) for both sentences. For instance, for n=2, the bigrams that match between the two sentences are:

Reference: ['A man', 'man sitting', 'sitting in', 'in a', 'a living', 'living room', 'room with', 'with a', 'a couch']

Machine-generated: ['A man', 'man sits', 'sits on', 'on a', 'a couch', 'couch in', 'in a', 'a living', 'living room']

Counting matching pairs of words, we get:

['A man'] - 2 matches

['man sits'] - 0 matches

['sits on'] - 1 match

['on a'] - 1 match

['a couch'] - 1 match

['couch in'] - 0 matches

['in a'] - 1 match

['a living'] - 1 match

['living room'] - 1 match

Step 3: Counting

We count the number of matching n-grams in the machine-generated sentence for each value of n. For example, in our case, the number of unigrams (n=1) that match between the two sentences is 6, the number of bigrams (n=2) is 5, the number of trigrams (n=3) is 2, and the number of 4-grams (n=4) is 0.

Step 4: Precision and brevity penalty

The precision of the machine-generated sentence was computed as the total number of matching n-grams divided by the total number of n-grams in the machine-generated sentence, across all values of n. Precision measures the ratio of correct n-grams generated by the machine to the total number of n-grams in the machine-generated sentence. However, BLEU also applies a brevity penalty to avoid favoring shorter sentences. The brevity penalty is a modification factor applied to the precision score, which reduces it for shorter sentences. In our example, the brevity penalty is 1, as the length of the machine-generated sentence is the same as the reference sentence.

We calculate the precision for each n-gram length by dividing the total number of matching n-grams by the total number of n-grams in the machine-generated

sentence. We also keep track of the maximum count of each n-gram length in any reference sentence.

In our example, the precision scores for n=2 are:

Precision for n-gram length 2 = (2+1+1+1+1)/9 = 0.67

Step 5: Final score

Finally, the BLEU score is computed as a weighted geometric mean of the precision scores across different values of n. The weight for each precision score is determined by the harmonic mean of the lengths of the machine-generated sentence and the reference sentence. The formula for BLEU is:

BLEU = brevity_penalty * exp(1/N * sum(log(precision_n)))

In our example, N=4

Natural Language Toolkit (nltk) library in Python was used to calculate the BLEU score. Following is a simple example code to calculate the BLEU-1 score using the nltk library:

```
import nltk

from nltk.translate.bleu_score import sentence_bleu

# reference sentence
```

```
ref = [['A', 'man', 'sitting', 'on', 'the', 'couch']]

# candidate sentence

cand = [‘a’, 'man', 'sitting', 'in', 'living', 'room', 'on’, 'a', 'couch']

# compute BLEU score

score = sentence_bleu(ref, cand, weights=(1, 0, 0, 0))

print(score)
```

6.3. METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is an evaluation index used for machine translation systems. It was developed to overcome some of the limitations of BLEU, which mainly focuses on lexical similarity and does not consider the semantic similarity of the translated sentences.

METEOR, on the other hand, combines multiple sources of information, including word overlap, word order, and synonymy, to compute a score that reflects the overall quality of the translation. It uses a combination of precision, recall, and alignment-based measures to evaluate the quality of the translation.

The calculation of the METEOR evaluation index involves the following steps:

- Tokenization: The sentences are tokenized into words, and the words are stemmed using a porter stemmer or other stemmer.
- Compute unigram, bigram, and trigram precision and recall: The unigram, bigram, and trigram precision and recall scores are calculated for the candidate sentence and reference sentence.
- Compute F-mean: The precision and recall scores are combined using the harmonic mean (F-measure) to get the unigram, bigram, and trigram F-measures.
- Compute the meteor score: The F-measures are combined using weighted geometric mean, with a coefficient that is determined based on the length of the candidate sentence and the number of matching words with the reference sentence.
- Compute penalty score: A penalty score is applied for extra words in the candidate sentence and missing words from the reference sentence.
- Compute final meteor score: The final meteor score is computed by subtracting the penalty score from the meteor score.

Suppose we have a candidate sentence and a reference sentence. Let's denote the number of words in the candidate sentence by wt, and the number of words in the reference sentence by wr. Now, let m be the number of words that are common

between the candidate and reference sentences. To compute precision and recall, we can use the following formulas: Precision = m / wt and Recall = m / wr. The harmonic mean of precision and recall is often used to get a single score that represents how well the candidate sentence matches the reference sentence. This can be expressed as the F-measure, which is defined as the harmonic mean of precision and recall. Harmonic mean is expressed as:

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R}$$

When evaluating machine translation, we might believe that the longer the longest common subsequence matched, the better the translation quality. But, the evaluation metric only takes into account single words, thus, to account for this limitation, a penalty factor called "pen" is introduced. To calculate the METEOR metric, the precision and recall values of unigram, bigram, and trigram are combined using the harmonic mean. Afterward, the weighted geometric mean of the resulting F-measures is computed, where the weight coefficient is determined based on the candidate sentence's length and the number of matching words with the reference sentence. Finally, the penalty score is applied to the metric to account for any extra or missing words in the candidate and reference sentences.

The METEOR can be calculated as:

$$METEOR = (1 - pen) \, x \, F_{mean}$$

In the above equation penalty factor,

$$pen = \gamma \left(\frac{ch}{m} \right)^{\theta}$$

where ch is the number of chunks. The hyperparameters α, θ, and γ are determined based on the specific datasets being evaluated. Unlike BLEU, which indirectly measures recall and word matching through high-order n-grams, METEOR directly measures accurate word-to-word matching. This addresses a weakness in BLEU's approach. However, despite its benefits, some researchers choose not to use METEOR due to its somewhat complex calculation steps and the large number of parameters that must be considered.

To evaluate the sentences using METEOR, we need to first tokenize the reference and hypothesis sentences and calculate their unigram, bigram, and trigram matches. Then, we calculate a harmonic mean of precision and recall using a parameterized formula. Finally, we apply several other penalties to adjust the score based on various factors such as stemming, synonymy, and paraphrasing.

METEOR takes into account both the precision and recall of the machine-generated translation when compared to a reference or human-generated translation. Here

is an example of how the METEOR scores were calculated for the reference and machine-generated sentences for ADL.

Step 1: Preprocessing

Both the reference and machine-generated sentences are preprocessed to remove any special characters, extra spaces, and convert all letters to lowercase. The resulting preprocessed sentences are:

Reference: a man sitting in a living room with a couch

Machine-generated: a man sits on a couch in a living room

Step 2: Tokenization

Both sentences are tokenized into individual words, resulting in the following tokens:

Reference: ['a', 'man', 'sitting', 'in', 'a', 'living', 'room', 'with', 'a', 'couch']

Machine-generated: ['a', 'man', 'sits', 'on', 'a', 'couch', 'in', 'a', 'living', 'room']

Step 3: Matching and Alignment

Each token in the machine-generated sentence is compared with each token in the reference sentence to find the best matching token. Matching is done using exact string matching and synonyms. Then, the best matching tokens are aligned

between the two sentences. The matching tokens and their alignment are shown in the following table:

Unigrams: "a", "man", "couch", "in", "living", "room"

Bigrams: "a man", "man sits", "sits on", "on a", "a couch", "couch in", "in a", "a living", "living room"

Trigrams: "a man sits", "man sits on", "sits on a", "on a couch", "a couch in", "couch in a", "in a living", "a living room"

Unigram, bigram, and trigram matches:

Unigram Matches: 6
Bigram Matches: 3
Trigram Matches: 1

For uni-gram

Reference Sentence	Machine-generated sentence
a	a
man	man
Sitting	-
in	in
a	a
living	living
room	room
with	-
a	a
couch	couch
	on

Step 4: Calculation of Precision and Recall

Precision is the number of matching tokens divided by the number of machine-generated tokens, while recall is the number of matching tokens divided by the number of reference tokens.

Precision and recall are calculated based on the matched tokens and their alignment. The precision and recall are then combined using a weighted harmonic mean, the harmonic mean is computed as the weighted average of precision and recall. For METEOR, the harmonic mean is weighted by a parameter with a weighting parameter β = 3:

Precision = 6/10 = 0.6

Recall = 6/9 = 0.67

Harmonic mean = (1 + 0.5) * (0.6 * 0.67) / (0.6 * 0.5 + 0.67 * 0.5) = 0.61

F-measure = (1+β^2)PrecisionRecall / (β^2*Precision + Recall) = 0.63

Step 5: Penalty for Unmatched Words

METEOR also applies a penalty for unmatched words in the machine-generated sentence that do not have any corresponding token in the reference sentence. The penalty is calculated as follows:

$$\text{Penalty} = \gamma * (\log(1 - (m/mg)))^{\eta}$$

where γ, η, m, and mg are constants. In the default implementation, γ = 0.5 and η = 0.5. m is the number of unmatched words in the machine-generated sentence, and mg is the total number of words in the machine-generated sentence. In this case, there are no unmatched words, so the penalty is 0.

Step 6: Final Score

The final METEOR score is the combination of the F-measure and penalty for unmatched words:

$$\text{METEOR} = (1 - \lambda) * \text{F-measure} + \lambda * \text{Penalty}$$

where λ is a constant weighting parameter between 0 and 1. In the default implementation, λ = 0.5. In this case, since there are no unmatched words, the final score is simply the F-measure:

$$\text{METEOR} = 0.63$$

Over all the scores of evaluation metrics in our case of ADLs are as follows

Metric	Score
B@1	79.4
B@4	36.7
METEOR	28.9

Table 2 Metrics

6.4. Experimental Setup

An experiment was performed to test the effectiveness of the learning algorithm to detect ADL in a patient with dementia. Various scenarios were tested on both healthy people and dementia patients. The experiment was setup as a comparison study between scripted actions vs real-time actions. First group characteristics were

- Healthy person (researcher herself)
- Scripted Actions
- Partially unnatural actions

- Multiple points of view of the camera (at least 2 different points of views for most actions)
- Instructed actions (Researcher knew what to do and what not to do)
- Researcher was the “Actor”
- Researcher could not perform actions in a way a person with dementia would perform with different cognitive and action performing challenges.

6.5. Our Testing samples

Our testing samples involves the following

- Unbiased, non-scripted activities
- Older adults
- Real Life Dementia patients (Early, middle, Early Late stages)

Another testing sample was recorded which included, Scripted activities performed by the researcher. The researcher had a clear idea of what kind of activities needed to be performed and recorded.

The researcher tried to perform those activities naturally as best as possible. But since the researcher recorded herself, we can clearly say that the videos were biased.

Those videos were also part of the test samples as it was tested, how the algorithm will perform on the biased vs unbiased activities.

The testing dataset that we collected for testing purposes consisted of real-time people with dementia who were diagnosed with either early, middle, or late-stage dementia. We will not be making our testing videos available to the public. The videos were collected only for this thesis purposes and all the people with dementia whose videos were recorded were all researcher's family members.

The researcher herself took the responsibility of keeping the collected data for the testing of the thesis project and were not made available and shared with anyone else for keeping, and storing purposes. Only the researcher worked on testing the proposed algorithmic approach on those collected videos.

For the comparison between scripted actions vs real-time actions, under general circumstances and testing, video monitoring algorithms and approaches give more accuracy on scripted actions while less accuracy on real-time actions performed naturally.

In our project, we also tested this fact and tested our designed approach on both scripted and unscripted test videos.

Scripted actions were performed by the researcher and unscripted actions were performed by the people with dementia.

6.6. Evaluation Protocol

The developed algorithm and methodology to detect the ADL was tested on videos which involved the challenges of real life with Real dementia patients.

We had 3 different dementia patients at 3 different stages of dementia. Every person of dementia is unique with challenges different than others, there is no standard or one way of defining and explain the actions, behavior, and activities of any dementia patient.

So, the daily activities of 3 different dementia patients and different stages of dementia were recorded. And these videos were used as the final testing of the algorithm after training and validation of the Toyota SmartHome dataset.

Toyota SmartHome dataset was split randomly into 2 sets while avoiding the biasness. Training and validation sets. Training data is the set of datasets which is used by the model for learning purposes and to fit the parameters of the learning model.

Validation data is the set of datasets which is used for the purpose of unbiased evaluation of the model which trained on the training data. During validation we also fine tune the hyperparameters of the model.

After the training and evaluation of the dataset. The model was tested on the real world recorded videos of the dementia patients. Figure shows the dementia patient at the early stage of dementia who was diagnosed by dementia almost a year ago.

The decline in the cognitive ability of the dementia patient and the change in their behavior always vary. Some patients have a slow decline in their cognitive ability while others have sudden changes of their cognitive state. One action performed by any person with dementia will never be similar to the other person with dementia.

Figure 7 Person with Dementia

Toyota SmartHome dataset consisted of 18 subjects (older people, 60-80 years old). This dataset was divided into training and validation sets.

To balance the number of subjects and activities in both categories the dataset was divided as follows

- Training = 11 subjects
- Validation = 7 subject

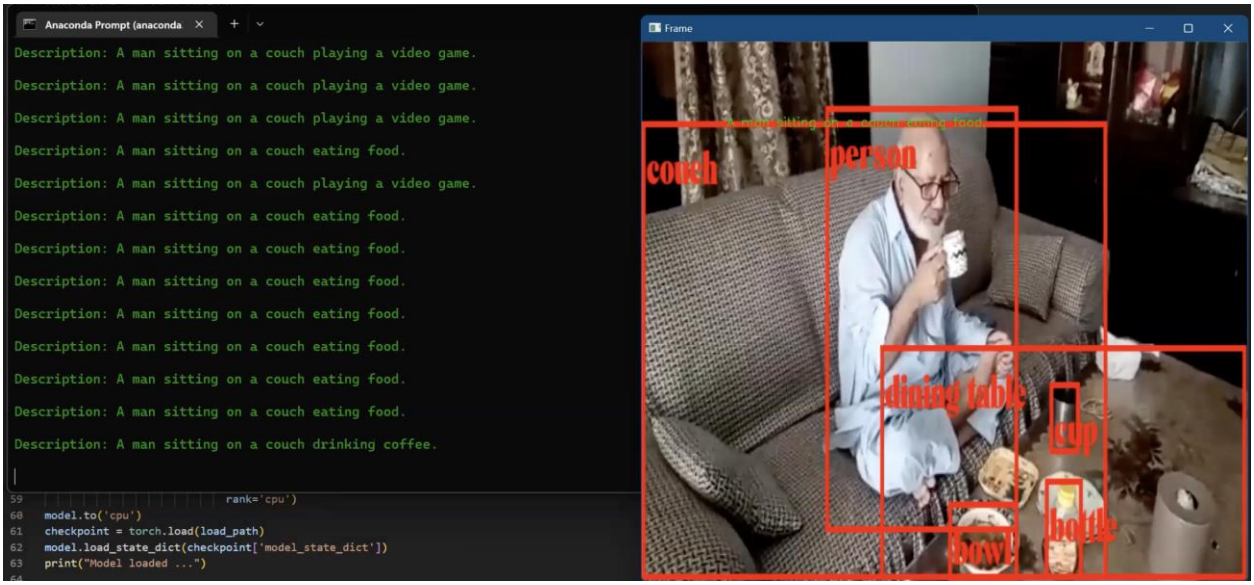


Figure 8 Activity Monitoring of Person with Dementia

Then after training and validation of the dataset consisted of ADL of older people in a smart home, another sample was introduced.

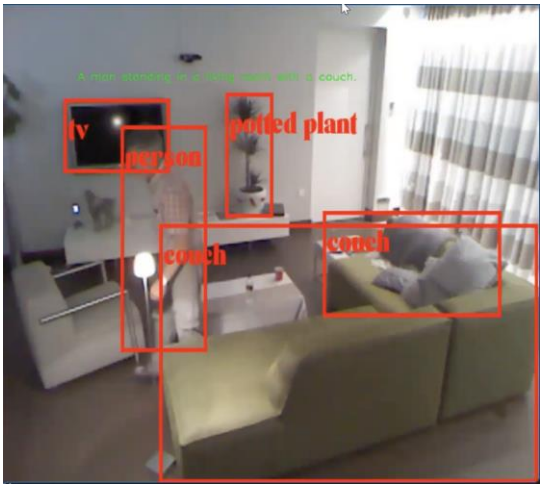
This testing sample consisted of the following:

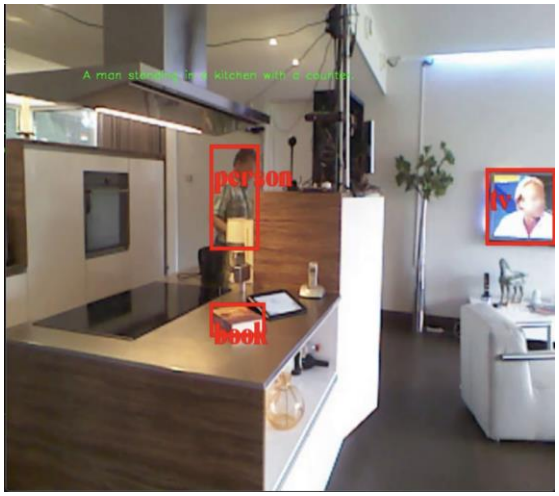
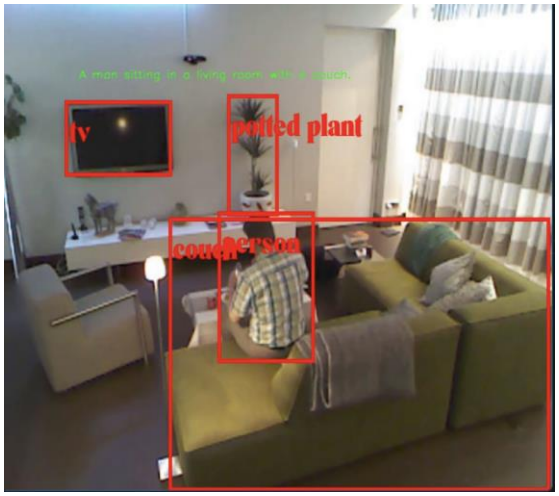
- Scripted activities of the researcher
- Unscripted activities of people with dementia

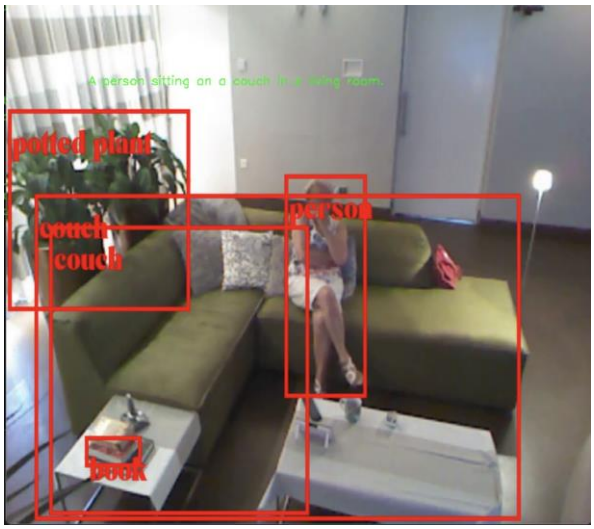
    ⇒ Early stage = 1 subject

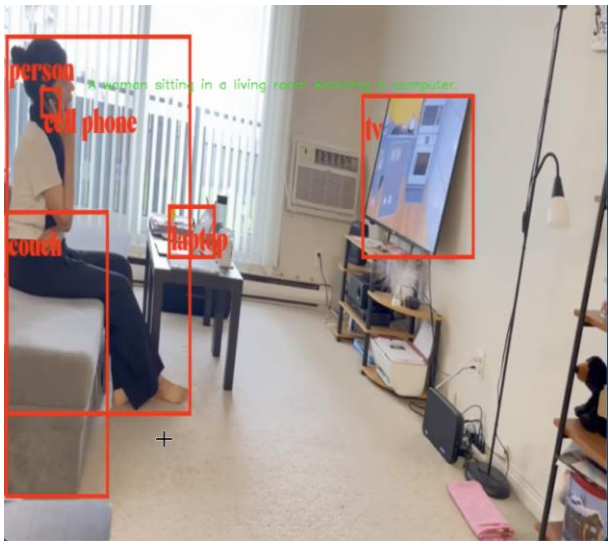
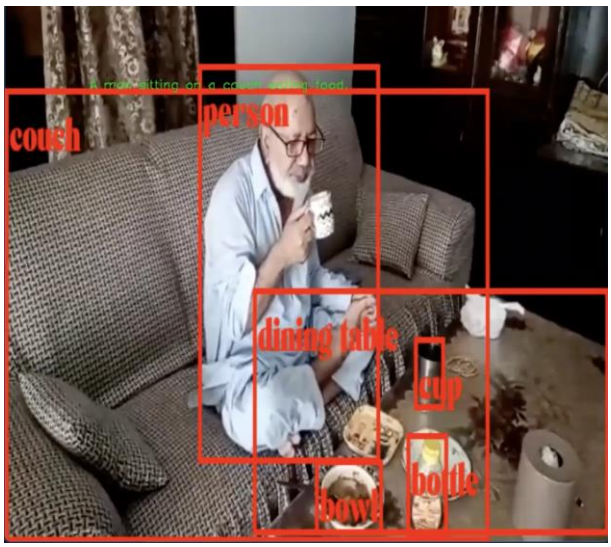
    ⇒ Middle stage = 1 subject

    ⇒ Late stage = 1 subject

ID	Image	Network Prediction	Ground Truth
1		A Man Standing in a Living Room with a Couch.	Gt 1: A man walking in living room. Gt 2: An old man standing besides a table. Gt 3: A man walking in front of couch. Gt4: A person standing between table and couch Gt5: Man standing up from the couch.
2		A woman standing in a kitchen with a sink.	Gt 1: A woman standing near a sink Gt 2: A woman standing near sink washing dishes. Gt 3: A woman washing dishes Gt4: A woman standing in kitchen near sink with bottle on top Gt5: A woman holding a saucepan while standing in kitchen near sink

3		A man standing in a kithen with a counter	Gt 1: A man standing in kitchen behind counter Gt 2: A man standing in kitchen with book placed on counter top. Gt 3: A man standing in kitchen watching TV. Gt4: A man behind counter Gt5: A man leaving the kitchen.
4		A man sitting in a living room with a couch.	Gt 1: A man sitting on a couch Gt 2: A man sitting on couch in a living room Gt 3: A man sitting on couch watching TV. Gt4: A man sitting on green color couch with TV hanging on wall in front. Gt5: A man sitting on couch doing something.

5		A man standing in kitchen next to a sink	Gt 1: A man standing in kitchen Gt 2: A man standing in kitchen holding something Gt 3: A man standing in kitchen making coffee Gt4: A man standing in kitchen besides sink. Gt5: A man standing in kitchen with a sink making coffee.
6		A person sitting on a couch in a living room	Gt 1: A woman sitting on a couch in aivving room Gt 2: A woman sitting on couch Gt 3: A woman sitting on a green couch with a table in front of her. Gt4: A woman sitting in aliving with a couch, table, plants and book on table. Gt5: A woman sitting in a living room on a green couch talking over the cell phone.

7		A woman sitting in a living room watching a computer.	Testing
8		A man sitting on a couch eating food.	Testing

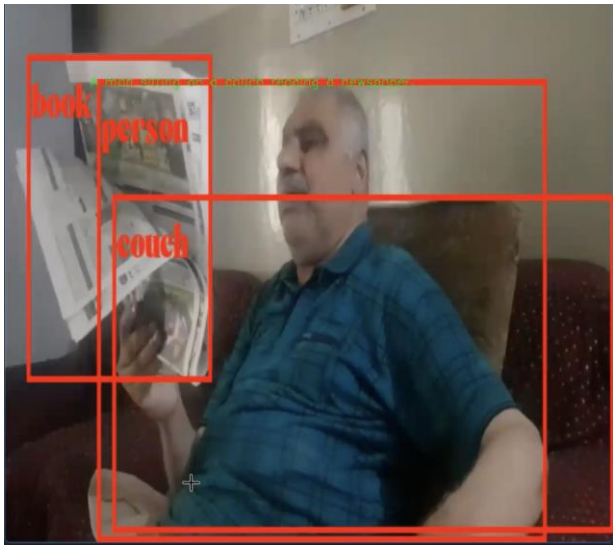
9		A man sitting on a couch reading a newspaper	Testing
10		A man sitting on a bed eating spoon.	Testing

Table 3 Examples of Caption

6.6. Qualitative Analysis

A decline in the accuracy was noticed in case of person with dementia at the late stage. This is due to the activates performed by the person with dementia. Person with dementia at the late stage had rather different activates which were not normal.

For example, it was noticed that in the “Eating” action performed by the person with dementia at the late stage. While the person was eating rice with a spoon a deviation from the normal action was noticed. The person took the rice from the plate but as the spoon reached the mouth for putting food in mouth the spoon was tilted. The action of the person with dementia at the late stage were not normal but rather deviated from the normal.

The walking action was also noticed. A normal person walks straight heads up. But the person with dementia at all the stages were not walking straight but rather in a bent position and in some cases while holding the wall for the support. The person with dementia at the late stage showed more of this behavior.

This can be concluded as the action might also be because of the age as all the person with dementia were 70-90 years old. Dementia could have also caused this change in behavior.

6.7. Summary

In a nutshell,

- Explored and obtained access to the Toyota SmartHome dataset:

  To begin with, the first step was to explore and get access to the Toyota SmartHome dataset. This dataset was crucial to the project as it served as the basis for the image captioning task that was being tackled. The dataset was explored thoroughly, and its annotation format was also examined to understand how the images and captions were labeled.

- *Downloaded the dataset and explored its annotation format:*

  After getting access to the dataset, the next step was to download it and explore its annotation format. This was necessary as the annotations contained vital information about the images and captions, such as the location of objects in the image, the caption itself, etc. The annotation

format was thoroughly examined and understood before moving on to the next steps.

- *Explored and cloned the ExpansionNet v_2 repository:*

  To build the image captioning model, the ExpansionNet v_2 model was chosen as it was already trained on the Coco dataset. The repository was explored, and the model was cloned to begin working on it.

- *Installed necessary libraries by creating a new environment:*

  Before working on the model, a new environment was created, and all the required libraries and dependencies were installed. This was done to ensure that there were no conflicts with existing libraries, and the project ran smoothly.

- *Converted the SmartHome annotations to the COCO annotation format:*

  The annotations in the SmartHome dataset were in a different format from the Coco dataset. Therefore, they needed to be converted into the Coco annotation format to work with the ExpansionNet v_2 model. This was done by mapping the SmartHome dataset annotations to the Coco dataset annotations.

- *Implemented bounding box detection using ResNet:*

Before fine-tuning the model for image captioning, object detection was performed by implementing bounding boxes using Resnet. This was necessary as it helped to identify the objects in the image and their locations.

- *Fine-tuned the ExpansionNet v_2 model on the converted SmartHome dataset for image captioning:*

  The next step was to fine-tune the ExpansionNet v_2 model on the converted SmartHome dataset. This involved adjusting the model's weights and hyperparameters such as batch size, epochs, and learning rate. This step was crucial as it helped to optimize the model for the image captioning task.

- *Adjusted the model's weights, batch size, epochs, and learning rate during fine-tuning:*

  During the fine-tuning process, we adjusted various parameters of the ExpansionNet v_2 model to optimize it for the Toyota Smart Home dataset. This included adjusting the model's weights, which determine the importance of each input feature to the output. We also modified the batch size, which determines the number of training examples used in each iteration of the training process. Additionally, we changed the number of epochs, which determine how many times the model sees the entire dataset during training, and the learning rate, which controls how quickly the model

learns from the data. By carefully tuning these parameters, we aimed to improve the performance of the model and ensure that it could accurately generate captions for images in the Smart Home dataset.

- *Evaluated the loss function of the fine-tuned model:*

  After fine-tuning the model, the loss function was evaluated to ensure that the model was learning and making progress. This was done by computing the loss on a validation set, and the results were examined.

- *Computed evaluation metrics, including BLEU and METEOR, to assess the model's performance:*

  To assess the performance of the model, evaluation metrics such as BLEU and METEOR were used. These metrics helped to measure the quality of the generated captions by comparing them to human-generated captions.

- *Tested the model on real-life people with dementia to evaluate the robustness of our approach for detecting and generating captions:*

  To evaluate the robustness of the approach, the model was tested on real-life images of people with dementia. The results were analyzed and compared with the ground truth captions to assess the performance of the model.

Chapter 7
Discussion & Future Work

The recognition of ADL is a challenging task due to the complexity and variation in human behavior. The difficulty is compounded by the fact that most datasets used for training and testing of ADL recognition algorithms are scripted, consisting of similar backgrounds and viewpoints, and involving staged actors. These algorithms may perform well when tested on these datasets but may not be effective in real-life conditions due to the variation in backgrounds, viewpoints, and features. In addition, the involvement of objects in activities further complicates recognition. While some datasets, such as the Toyota Smart Home dataset, have attempted to address these challenges by recording unbiased activities of older adults, most datasets still do not capture the challenges of real-life ADL. The accuracy of ADL

recognition algorithms is greatly affected by the availability of annotated datasets, with deep learning neural networks performing better with larger datasets.

The ExpansionNet V_2 model was trained and fine-tuned on the Toyota SmartHome dataset to improve its performance on the ADL monitoring task. The training process consisted of fine-tuning. In the fine-tuning phase, the entire network was fine-tuned.

The algorithm for detecting ADL was tested on real-life videos of people with dementia at different stages of dementia. There is no standard way of defining the behavior and activities of dementia patients as each patient is unique. The Toyota SmartHome dataset was split randomly into training and validation sets, and the model was fine-tuned on the validation set. The trained model was then tested on the recorded videos of the dementia patients. The cognitive ability and behavior of dementia patients vary, and each patient is unique.

7.1. Discussion

Dementia is a neurological disorder that affects the cognitive abilities of an individual, including memory, thinking, and behavior. However, it also affects the motor abilities of the individual, including the way they walk, perform actions, and

interact with their environment. This is because dementia affects the motor areas of the brain that are responsible for controlling movements, such as the basal ganglia and the cerebellum.

The changes in motor movement in people with dementia can be observed in their gait, which is the way they walk. Their gait may become slow, unsteady, and irregular. They may also take shorter steps, have difficulty initiating movement, and have a higher risk of falls. This is because dementia affects the nervous system in the body, including the motor nervous system that controls muscle movements.

In addition to gait changes, dementia can also affect the way an individual performs various tasks, such as dressing, eating, and grooming. They may have difficulty with fine motor movements, such as buttoning a shirt or tying shoelaces. They may also have trouble with gross motor movements, such as reaching for objects or getting out of a chair.

Overall, the changes in motor movement in people with dementia can significantly impact their daily activities and quality of life.

The study of recognizing activities of daily living (ADL) is essential for people with dementia. Since people with dementia have difficulties in performing their daily

activities, recognizing and monitoring their ADL can provide important information to caregivers and medical professionals.

The impact of dementia on the motor movement of individuals has been widely studied. Therefore, recognizing ADL of people with dementia requires the use of specialized techniques that can account for these differences. Our study shows that image captioning techniques can be effective in recognizing ADL of people with dementia, even with dataset like the Toyota Smart Home dataset.

The early study of recognizing activities of daily living (ADLs) initially focused on the detection of falls. Early experiments were conducted to identify falls in individuals, but these experiments were performed in simulated environments. In these experiments, no real patients were involved, and participants were asked to act out various scenarios of falling according to the instructions provided by researchers. But theres more to ADL then just the detection of fall. There are a range of activities that individuals engage in on a daily basis, and the ability to accurately recognize and monitor these activities can have significant implications for healthcare and quality of life.

The development of datasets for activities of daily living (ADLs) began with the focus on detecting falls in individuals. However, early experiments to detect falls

were conducted in a simulated environment where subjects performed different scenarios, acting out the act of falling according to instructions provided by the researchers. This approach did not involve real patients and did not represent real-world scenarios.

Later, datasets were developed for ADLs, but these were also recorded by subjects who performed activities according to specific scripts provided by the researchers. In other words, the actions were scripted and did not represent real-world situations. However, these scripted actions tend to provide greater accuracy when trained and tested as compared to non-scripted and real-time datasets.

Many of the ADL datasets were also performed in a laboratory environment and not in a real environment, which can present challenges when trying to detect the action performed in terms of the background scenario. Furthermore, in real-life situations, there is never a similar background while performing different actions, which can pose a challenge to activity recognition algorithms. Therefore, it is important to develop datasets that accurately represent real-world situations to improve the accuracy of ADL recognition algorithms.

More reseraches were perfomed usng the large datasets but those datasets consisted of random activities and not the activities related to the person with

dementia. For example the datasets included activities like squatting, hitting a tennis ball or other actions that are not necessarily related to daily activities of individuals with dementia. These datasets were not focused on the ADL of people with dementia. As a result, such datasets are not particularly useful for detecting and understanding the ADL of people with dementia

In order to overcome the limitations of existing datasets that focused on random activities, a new dataset was developed called the "Toyota SmartHome" dataset. This dataset involved older adults performing unscripted activities in a real home environment. Amongst the numerous ADL and activity datasets, we identified the "Toyota SmartHome" dataset as a suitable dataset for our research. We gained access to this dataset and used it to detect ADLs in people with dementia. This dataset provided a more realistic and challenging environment for the activity recognition algorithms to perform accurately.

Several algorithms and models were proposed to identify the Activities of Daily Living (ADL) in the Toyota Smart Home dataset, which categorized the actions into fixed and limited classes such as sitting, standing, eating, and walking. However, the focus of the research was not just to classify the actions but to generate a comprehensive description of the entire scenario in which the person was

performing the activity. Therefore, instead of relying solely on the dataset for classification, we used it to create captions that described the position, environment, and activity being performed by the individual. This approach aimed to provide a more detailed and holistic understanding of the ADL being performed in a real-world scenario.

In our research, we aimed to address the challenges faced in detecting ADLs in real-world environments by taking a novel approach. Instead of relying solely on pre-recorded datasets, we introduced real-time people with dementia and involved the scenarios performed by them in testing our approach for detecting ADLs and captioning and providing a description of the action being performed. This is a significant step forward in the history of ADL research as no single study has involved actions performed by people with dementia before.

Moreover, we identified three people at three different stages of dementia, as the cognitive decline of people with dementia affects their behavior differently at each stage. Thus, we not only focused on ADLs performed by people with dementia but also accounted for the stage of dementia they were in. By doing so, we aimed to provide a more comprehensive understanding of the challenges of ADL detection in people with dementia.

The motor movement differences and changes in behavior in people with dementia were captured and analyzed through the video dataset of their daily activities. By analyzing these differences, we gained a better understanding of recognizing and understanding the behaviors of people with dementia

Moreover, we conducted a qualitative analysis of our results by testing on real-life people with dementia. This evaluation showed that our approach was able to recognize the ADL of people with dementia accurately, despite their differences in motor movements and behavior. Our study opens up new opportunities for developing assistive technologies for people with dementia that can improve their quality of life. Overall, our work contributes to the field of image captioning and recognition of ADL of people with dementia.

While conducting our research, we did not have a dataset specifically focusing on the ADLs of people with dementia. Therefore, we trained our model on the ADLs performed by healthy older adults. However, we did use a few videos and scenarios involving people with dementia in the testing stage to evaluate the effectiveness of our model in real-life situations. We recognize the need for a dataset specifically tailored to the ADLs of people with dementia to overcome the unique challenges associated with their behavior. Therefore, we suggest that future studies should

focus on developing such a dataset and use it to train and test models for ADL detection in people with dementia. By doing so, we can improve the accuracy and effectiveness of ADL recognition systems for people with dementia, ultimately leading to better care and support for this vulnerable population.

7.2. Cases Explaining why this research is impactful/important

Care partners of people with dementia often need breaks from the 24/7 care they provide, but at the same time, they worry about leaving their loved ones alone. Acting happy and positive in front of people with dementia is important, as their emotions often mirror those of their caregivers. The constant stress and worry can lead to exhaustion for care partners. Activity monitoring technology can provide relief for care partners by allowing them to check on their loved ones and make sure they are doing well, even when they are not physically present. This technology can help care partners take much-needed breaks while ensuring the safety and well-being of their loved ones.

Activity monitoring technology can be very helpful for care partners of people with dementia as it can provide them with the reassurance and peace of mind that their

partner is safe and doing well, even when they are not physically present. This technology typically involves the use of sensors placed throughout the home that can track the movement and activity of the person with dementia.

In addition to providing reassurance and peace of mind, activity monitoring technology can also help care partners identify potential issues or problems that may need attention. For example, if the person with dementia has not opened the refrigerator or eaten a meal for an extended period of time, this could indicate that they are not eating properly or may be experiencing other health issues.

Moreover, activity monitoring technology can be used to identify patterns or changes in behavior that may indicate a decline in cognitive function or other issues related to dementia. For example, if the person with dementia begins to wander or becomes more agitated than usual, this could indicate that their condition is worsening and that additional support or care may be needed.

Overall, activity monitoring technology can be a valuable tool for care partners of people with dementia, providing them with the reassurance and support they need to take time for themselves while also ensuring the safety and well-being of their loved one.

7.3. Limitations

some limitations of the activity monitoring system that uses the expansionNet_v2 model trained on Toyota smart home dataset are:

Slow Processing: The model may require a high processing GPU, which can make it slower and less practical for use in real-time applications.

Limited Dataset: The Toyota smart home dataset used to train the model may not be representative of all possible scenarios and variations that may occur in the daily activities of people with dementia. This dataset consisted of older adult as sujects, not people with dementia. Additionally, the dataset did not have a sufficient amount of data to capture all possible variations of daily activities for people with dementia.

Need for More Dementia-Specific Data: Dementia patients are unique, and their daily activities and behavior can vary significantly from person to person. The model's accuracy may be limited by the need for more data from dementia-specific scenarios and individuals to better train the model to recognize different patterns and variations in daily activities.

Dataset Labeling Challenges: Labeling the dataset for people with dementia can be challenging due to the unpredictability of dementia patients' behavior and the need for specialized knowledge to label and interpret the dataset accurately.

Practicality for Real-World Use: The model's accuracy in detecting ADLs may be limited in real-world settings due to environmental factors such as changes in lighting, furniture layout, or other factors that can affect the performance of the model.

Ethical Considerations: The use of an activity monitoring system in the context of people with dementia raises ethical concerns about privacy and dignity. Careful consideration must be given to how the system is implemented to ensure that it respects the rights and dignity of the individuals being monitored.

Dependence on Video Data: The model's performance may be limited by the quality and quantity of video data available for training and evaluation. Depending solely on video data to detect ADLs may not be practical in all settings, especially in situations where privacy concerns limit the use of video recording.

Sensitivity to False Positives and Negatives: The model's performance may be impacted by false positives and negatives, where the system might have detected activity that is not occurring or failed to detect an activity that is occurring. This can

lead to inaccurate monitoring and may impact the usefulness and acceptability of the system.

Cost and Maintenance: The high cost of GPU and the need for regular maintenance and updates can make the activity monitoring system expensive and difficult to sustain in the long run.

7.4. Future Studies

A future study should be conducted with a larger sample size of people with dementia. Although gathering video monitoring data can be challenging, it can help us understand the behavior of people with dementia, which in turn will enable us to monitor them more accurately. In our case, the sample size for testing was small, so the results and evidence might not be as strong. We suggest conducting tests with a larger sample size to determine the accuracy of monitoring people with dementia at different stages of the disease. Behaviors, actions, and activities of daily living change as cognitive abilities decline and dementia progresses. No research has been conducted to monitor people with dementia at different stages. This study has taken the first step in monitoring people with dementia at various stages of the disease, but more evidence needs to be collected, and more work needs to be done. Before we start detecting dangerous behavior, it is essential that

we begin by monitoring and testing our algorithms and approaches with larger, real-world scenario datasets based on natural, unscripted behavior, not scripted behavior with actors performing various actions. A lot of work needs to be done in this field. Monitoring repetitive behavior can also be achieved by generating captions repeatedly at some interval of time, which can help predict the repetition of the behavior.

References

[1] H. Brodaty. &. Marika. Donkin, "Family caregivers of people with dementia," *Dialogues in Clinical Neuroscience,* pp. 217-228, 2022.

[2] B. Goldman, "Navigating the path forward for dementia in canada," Alzheimer society canada, 2022.

[3] Public Health Agency of Canada, "Dementia in Canada, including Alzheimer's disease," 2017.

[4] C. Broony, M. Larry W. Chambers, "Prevalence and Monetary Costs of Dementia in Canada," The Alzheimer Society of Canada in collaboration with the Public Health Agency of Canada, Toronto, 2016.

[5] Zoha. A. Markides. A. Skillman. S. Acton. S. Elsaleh. T. e. a. Enshaeifar S, "Health management and pattern analysis of daily living activities of people with dementia using in-home sensors and machine learning techniques," *PLoS ONE 13(5): e0195605.,* 2018.

[6] J. Brownlee, "Deep Learning Models for Human Activity Recognition," Machine Learning mastery, 2018. [Online]. Available: https://machinelearningmastery.com/deep-learning-models-for-human-activity-recognition/. [Accessed February 2023].

[7] T. Yamasaki and S. Kumagai, "Nonwearable Sensor-Based In-Home Assessment of Subtle Daily Behavioral Changes as a Candidate Biomarker for Mild Cognitive Impairment," *Journal of personalized medicine,* vol. 11, 2022.

[8] William. K, T. Md. Zia Uddin, "Ambient Sensors for Elderly Care and Independent Living: A Survey," *MDPI,* vol. 18, 2018.

[9] L. Annica Kristoffersson, "A Systematic Review ofWearable Sensors for Monitoring Physical Activity," *MDPI,* vol. 22, no. 573, 2022.

[10] T. Lee and Mihailidis, "A. An intelligent emergency response system: Preliminary development and testing of automated fall detection," *Journal of Telemed,* vol. 11, p. 194–198, 2005.

[11] C. Lin and Z. Ling, "Automatic Fall Incident Detection in Compressed Video for Intelligent Homecare," *In Proceedings of the 16th International Conference on Computer Communications and Networks,* p. 1172–1177, 2007.

[12] J. Aertssen, M. Rudinac and P. Jonker, "Fall and Action Detection in Elderly Homes," in *AAATE*, Maastricht, 2011.

[13] E. Auvinet, L. Reveret, A. St-Arnaud, J. Rousseau and Meunier, "Fall detection using multiple cameras," in *In Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vancouver, 2008.

[14] M. Belshaw, B. Taati, D. Giesbrecht and Mihailidis, "A. Intelligent vision-based fall detection system: Preliminary results from a realworld deployment.," *Rehabil. Eng. Assist. Technol,* pp. 1-4, 2011.

[15] M. Abidine and B. Fergani, "News Schemes for Activity Recognition Systems Using PCA-WSVM, ICA-WSVM, and LDA-WSVM," *Information,* vol. 6, p. 505–521, 2015.

[16] S. Berlin and John, "M. Human interaction recognition through deep learning network.," in *IEEE International Carnahan Conference on Security Technology (ICCST),*, Orlando, 2016.

[17] D. Brulin, Y. Benezeth and E. Courtial, "osture Recognition Based on Fuzzy Logic for Home Monitoring of the Elderly," *IEEE Trans,* vol. 16, p. 974–982, 2012.

[18] Y. Du, W. Wang and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015.

[19] H. Chen, G. Wang, J. Xue and L. He, "A novel hierarchical framework for human action recognition.," *Pattern Recognition,* vol. 55, p. 148–15, 2016.

[20] Z. Huang, C. Wan, T. Probst and L. Gool, "Deep Learning on Lie Groups for Skeleton-Based Action Recognition," *arXiv Prepr,* 2016.

[21] M. Krekovic´, P. C˘ eric´, T. Dominko, M. Ilijaš, K. Ivanc˘ic´, V. Skolan and J. Šarlija, "A method for real-time detection of human fall from video," in *35th International Convention MIPRO*, Opatija, 2012.

[22] Z. Lan, M. Lin, X. Li, A. Hauptmann and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015.

[23] Y. Li, K. Ho and M. Popescu, "A Microphone Array System for Automatic Fall Detection.," *IEEE Trans,* vol. 59, p. 1291–1301, 2012.

[24] X. Peng, C. Zou, Y. Qiao and Q. Peng, "Action recognition with stacked fisher vectors," in *Computer Vision–Asian Conference on Computer Vision*, Berlin, 2014.

[25] Z. Lan, M. Lin, X. Li, A. Hauptmann and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015.

[26] L. Mo, F. Li, Y. Zhu and Huang, "Human physical activity recognition based on computer vision with deep learning model," in *IEEE International Instrumentation and Measurement Technology Conference Proceedings*, Taipei, 2016.

[27] A. Shahroudy, T. Ng, Q. Yang and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 38, p. 2123–2129, 2016.

[28] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond and Gianpiero, "Toyota Smarthome: Real-World Activities of Daily Living," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, seoul, south korea, 2019.

[29] M. Marie. S. Belongie. L. Bourdev. R. Girshick. J. Hayes. P. Perona. D. Ramanan. C. Lawrence. Zitnik. P. Dollar. Tsung-Yi Lin, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, ECCV 2014.

[30] R. Cavicchioli. A. Capotondi. Jia Cheng Hu, "ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning," *archive,* no. Computer Vision and Pattern Recognition, 19 August 2022.

[31] V. Lendave, "analytics india mag," 28 November 2021 . [Online]. Available: https://analyticsindiamag.com/how-to-use-learning-rate-annealing-with-neural-networks/. [Accessed 2023].

[32] S. Das. S. Sharma. L. Minciullo. L. Garattoni. F. Bremond. G. Fransesca. Rui Dai, "Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection," *Computer Vision and Pattern Recognition (cs.CV),* vol. v1, 2020.

[33] L. Cheng. C. Jing. C. Zhao. G. Song. Gaifang Luo, "A thorough review of models, evaluation metrics, and datasets on image captioning," *IET Image Processing,* vol. 16, no. 2, pp. 311-332, 2022.