MedAssist: An Automated Solution for The Assessment of Medication Intake

Adem Kaya, George Lalidis, Giedrius Mirklys, Tam Van

21-06-2025

1 Introduction

For the course AI in the Professional Workfield (SOW-MKI76), we were challenged to develop a project for a company in a selection of companies provided on the Masters Challenge platform. Given our shared background and passion for societal impact and healthcare, we ended up with a company called MedAssist.

1.1 MedAssist

MedAssist is a company that is concerned with building medication dispensary devices. Not only do their devices feature automatic release of medication, such that it capable of helping its patients to take their medication on time, but the devices also feature a build in camera that is capable of recoding videos of the patients whenever they are exactly in front of the device.

1.2 The Problem

As of now the video material that is collected by the deviced is manually reviewed by humans to check whether the patient in question has sucessfully taken their medication. The problem lays in the time consuming nature of this process. In order to provide a solution to this time consuming approach, MedAssist has reached out to us to build an AI which is capable of automatically assessing whether the patient has taken their medication or not.

1.3 The Goal

In order to provide a solution to the problem, the goal is to build an AI which is capable of automatically assessing whether the patient has taken their medication or not to its best extent. Not only should we aim to maximize the accuracy of the AI, but we should also carefully aim to minimize the amount of false positives as we do not want the AI to make it seem like the patient has taken their medication even though they have not.

2 Methods

2.1 Dataet

The dataset that was provided to us by MedAssist, initially consisted of 85 unlabeled videos taken by the medication dispensary devices. We were told beforehand that there were not privacy concerns applicablel to the dataset, and that we were allowed to use it freely. The videos had small variations in lenghts, subjects, surroundings; but did include a large variety of different lightning conditions. Each of the 85 videos was watched and manually annotated whether the subject took their medication or not. Of these 85 videos, a total 71 cases where found to contain content of the subject taking their medication.

A few weeks before the end of the project, another suplementary dataset was provided to us. This dataset consisted of another 29 videos, of which 23 were cases in which the subject was taking their medication. Again, the previous mentioned variations are applicable to this part of the dataset as well, with the addition of a slight increase in subject variation.

Combining these two datasets, we ended up with a total of 114 videos with a class imbalance of 82%. From this combined dataset we split up the data into a training-, test- and validation set, using a split of 70%, 20% and 10% respectively.

2.2 Human Activity Detection (HAD)

Human activity detection (HAD) is a subfield of machine learning that focuses on recognizing and classifying human activities from different possible data modalities. For this specefic task, we are interested in the application of HAD based on video data. HAD models often make use of a list of different possible activities, which they are trained on to recognize and classify. In our case, we can break this down into two classes:

I. The subject has taken their medication at some point in the video.

II. The subject has **not** taken their medication at any point in the video.

2.2.1 SligLIP 2

Because the datset that we were provided was substanstially smaller than desired to train a well performing HAD model from scratch. We made use of a pre-trained model based on the SligLIP 2 transformer architecture [cite]. The implemnentation of this model was made publically available and retrieved through the HuggingFace model hub [cite]. The pre-trained model was build and made available by Prithiv Sakthi, and reported great performance in detecting and classifying a total of 15 different human activities for still images.

Even though the model was not directly trained on the task of medication intake assessment, *drinking* as well as *eating* which we considered to be especially useful for our task. In specific these two activities had F1 scores of 0.89 and 0.93 respectively. For a full summary of the model's performance ono the 15 different activities we refer to the confusion matrix in Figure 1.

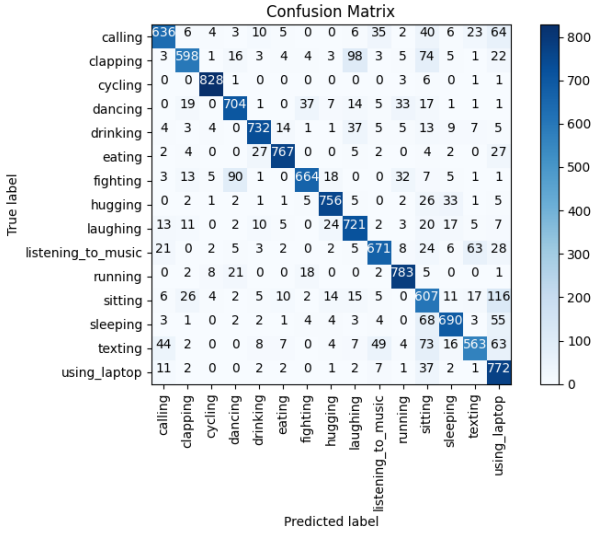

Figure 1: SligLIP 2 Confusion Matrix for Activity Classification

2.2.2 Finetuning

Because the pre-trained model was not directly trained for the task of medication we needed to finetune the model. In order to do so, we manually annotated individual frames of the videos in our training set. We annotated a total of X frames, of which Y were cases where the subject was taking their medicaiton, and Z were cases where the subject was not taking their medication. This resulted ina class imbalance of XY% which we found to suffice for the task of finetuning.

The task of finetuning improved the model's test performance on the task of medication intake from an F1 score of 0.53 to 0.75, suggesting a succesful finetuning process.

2.3 Working with Videos

Because the fine-tuned model was trained to classify individual frames, rather than entire videos, we needed to find a way to efficiently process videos recorded by the devices. In order to do so, we extracted a total of 10 frames, evenly spaced out over the entire length of any video. By the application of this approach, we were able to process any video while only processing a small fraction of the frames. This way we were able to process the entire video, while onlyprocessing a small fraction of the frames. Assuming that in at least one of these frames the subject would be taking their medication, for the videos that were annotated as such, this limited amount of frames is desirable since it hugely drives back computational costs.

3 Results

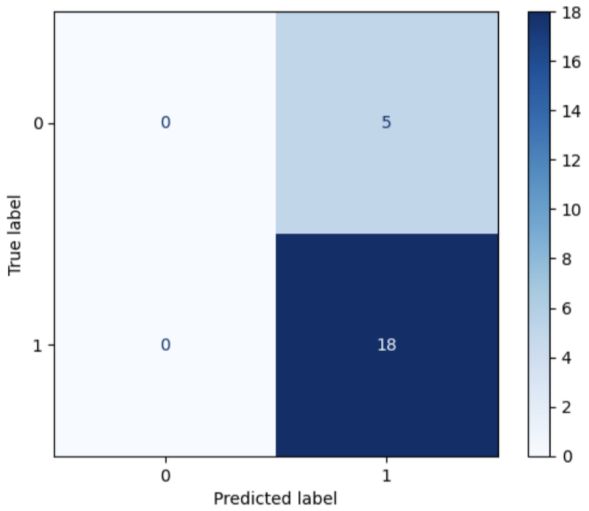

Figure 2: SligLIP 2 Confusion Matrix for Activity Classification

4 Conclusion

5 Discussion and Recommendations

5.1 Data Suggestions

References