
Workout Assistant

Ákos Morvai, Csaba Mészáros
BME AUT
morvaiakos1998@gmail.com, csaba591@gmail.com

Abstract

Our selected project was a training helper application based on deep learning technologies. The program uses the device's camera to recognise three of the most common exercises people do in their workout routine. There are push ups, pull ups and squats. Aside from this the app detects up and downward motion to determine the number of repetitions for each exercise. Throughout the development we held the lightweight of the program a key factor as we want the app to reach all kind of users with laptops and mobile phones. For this purpose we also created a web app.

A választott feladatunk egy testmozgás végzést segítő alkalmazás elkészítése volt, deep learning technológiák felhasználásával. Az alkalmazás az eszköz kamerája segítségével felismer három gyakori erősítést, ezek a fekvőtámasz, guggolás és húzockodás. Emelett a program a fel és le mozgásokból megállapítja az ismétlésszámot is. Az elkészítés során törekedtünk rá, hogy az alkalmazás széles körben elérhető legyen, így azt egy webappként készítettük el. Ennek feltétele, hogy a felhasznált modellek kis erőforrásigényűek legyenek. Így akár egy telefonos vagy laptopos böngészőben, a hétköznapi felhasználó számára is elérhető legyen.

1 Introduction

2020 was a strange year for all of us. Our lives were changed drastically by the pandemic. We spent most of our at home as we couldn't go to public places like cinemas, theatres or to the gym. However we all know the importance of doing exercises.

So we decided to create an application that can help everyday people train better on their own without the help of a trainer.

It included a smart web camera system that detects the kind of motion one does. So far it supports push ups, pull ups and squats. In addition to this, it counts the number of repetitions.

Our goal with this application is to bring the perfect gym experience for everyone in their homes.

2 Background and past work

Computer vision challenges have been in the spotlight for a long time now, especially since the great success of AlexNet in the 2012 ImageNet challenge, where it showed the true potential of deep neural networks in the field. Our task at hand is similar regarding the computer vision aspect, but still it's something more specific and it's for everyday use.

When first starting out we looked for similar solutions and projects. There are a handful of blogposts, repositories and example projects, but none were exactly the same. Either their goal or the way they work differed. For example when looking at the problem of counting exercise repetitions we found

35 some great ideas using optical flow or proximity sensors. Although they gave us some points to start
36 from, they were mostly small projects implementing only one functionality.

37 Another way of looking at the problem is using devices such as smart watches that track body statistics
38 while working out. Using these live datapoints streamed to a computer through bluetooth or a wi-fi
39 connection can work great. This means that we would have to deal with mostly numerical data which
40 comes with the benefit of simpler implementations and model architectures as well compared to using
41 live video footage from a camera. These devices are although widely available nowadays, we still
42 can't say that the average person has them. To tackle this problem we stuck with using live camera
43 data which is very widely available for anyone that has a smartphone or laptop.

44 2.1 Data

45 Gathering good quality training data seemed to be an easy task when first thinking about it, but later
46 on it proved to be a bit of a challenge. There is a large selection of datasets consisting of images and
47 videos nowadays, but most of them consist of a great variety of categories. This is great for challenges
48 such as ImageNet where the goal is to make a model with a great ability of generalisation as to be
49 able to recognize as many different scenes and objects as possible. Our case was different, since our
50 goal was to recognize only a handful of exercise types: pushups, pullups and squats. This meant that
51 we needed a dataset with large amount of videos or images about people doing these exercises.

52 After looking through a couple of dataset catalogues, such as Kaggle and Google's OpenImages we
53 found out that the number of videos falling into our 3 classes was rather small. We ended up using
54 UCF101 which consists of 101 action categories belonging to 25 groups. For each action type 4-7
55 videos are provided with a large variation of camera movement, lighting conditions and recording
56 equipment. These videos were not recorded by actors, but in a real environment. This is especially
57 useful in our case, since every person has their own different room with its varying lighting conditions
58 and color palette. In conclusion the dataset is somewhat similar to YouTube videos, but they're
59 categorized and selected by hand for this exact purpose of doing deep learning projects, experiments
60 and challenges. The dataset is provided by the University of Florida's computer vision research lab.

61 3 Model selection and training

62 Dealing with a computer vision task we had many resources for learning about ways of creating
63 applicable models. Our choice of framework was Tensorflow, which comes with the advantage of
64 a Javascript implementation which was absolutely necessary for our use case. There are many CV
65 challenges these days that have led to the creation and research of many different deep learning
66 methods, algorithms and models.

67 Considering this our first thought was using a pre-trained convolutional neural network that could
68 work within a browser environment. This means that it must be small and should only require low-end
69 hardware to run. Our first choice was EfficientNet which promises great performance under low
70 hardware specs compared to other similar models such as Inception or Xception for example. This
71 idea was quickly changed when we found out that Tensorflow.js doesn't support all layer types present
72 in the Python version of TF that were needed for EfficientNet. Looking for alternatives we found
73 another subject, MobileNet. This model was especially engineered for extra low resource hardware,
74 such as mobile phones, while still performing relatively well.

75 After our choice was made we applied some transfer learning methods to retrain the model for our
76 purpose of predicting exercise categories, instead of the 1000 others found in the ImageNet dataset
77 which the model was originally trained on. We left the base models layers untouched and only added
78 a couple fully connected output layers for classification using a categorical-crossentropy loss function
79 and the softmax activation function at the output.

80 Training on the UCF101 dataset - which we first converted to individual frames for each video - the
81 model's performance was looking great. Doing some further experiments and data augmentation for
82 even better results we did some final training rounds and had our model ready for use.

83 Since we also needed something for counting repetitions we also added a second model. Our choice
84 was PoseNet which is another model based on MobileNet that predicts the position of 26 points of
85 the human body from and RGB image input. We used this model for locating the person's upper body

86 while moving. Counting was implemented as follows: a repetition is considered completed when the
87 person has moved downwards, then upwards and then stops or starts moving down again. This can be
88 done using some simple heuristics based on the y coordinates of the upper body.

89 4 Testing the model

90 After building the model the next important step was to test it.

91 First we performed tests with the validation data set. The accuracy there was very high, it went all
92 the way up to 99%.

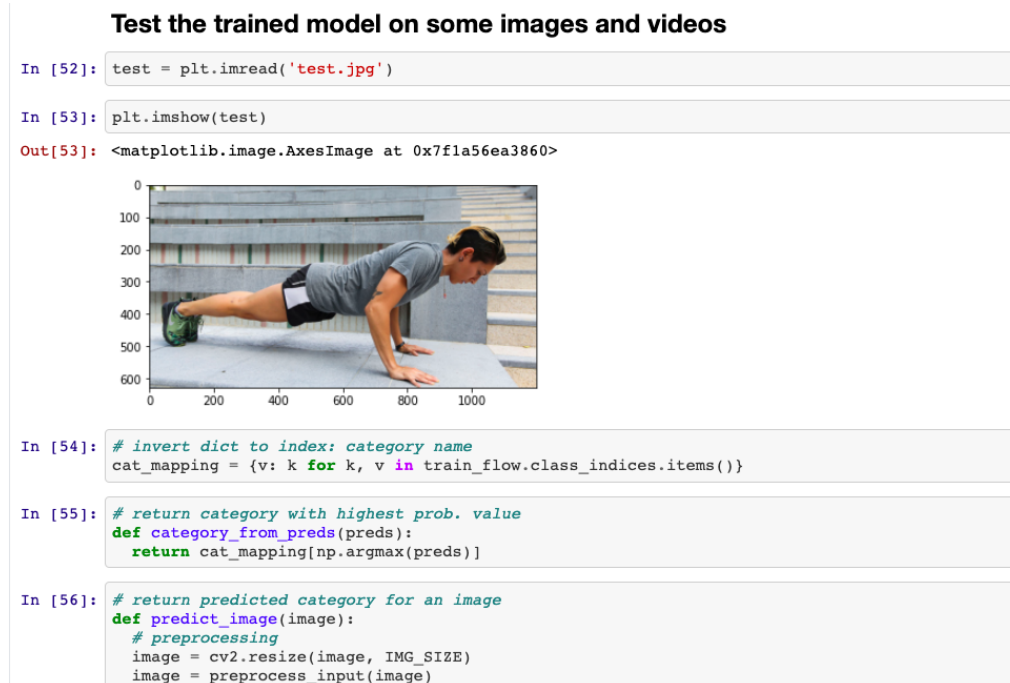
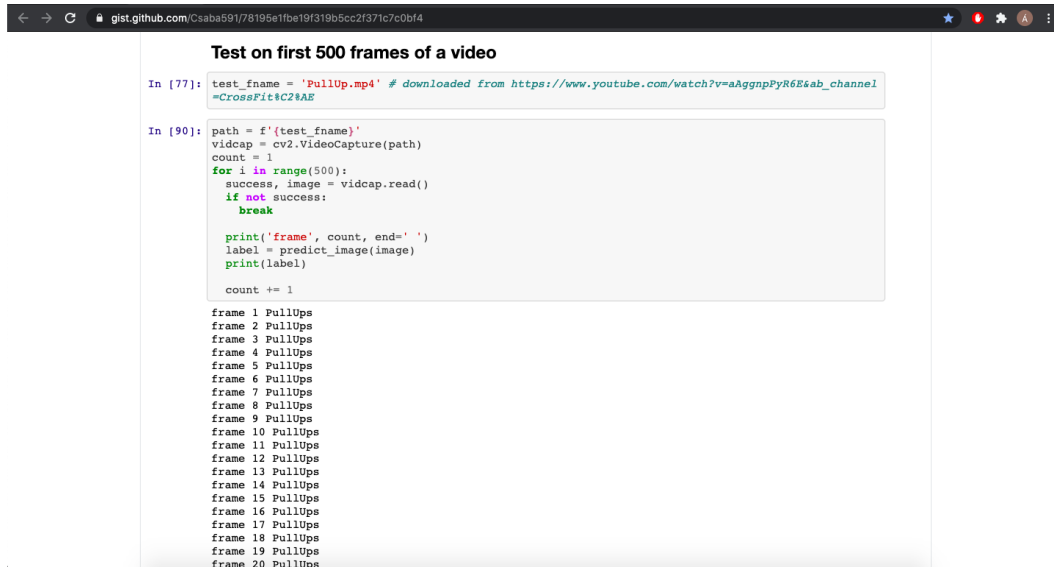


Figure 1: Testing on images

93 Then came tests with real life inputs. In the beginning, we applied images about people performing
94 various exercises. The model confidently guessed all them correctly.

95 For testing videos we used OpenCv's read() function to capture the individual frames. On each frame
96 we called the model to predict the input and print out the label. Here we got into some problems,
97 mainly with finding videos that didn't have writings on top of the image as it could confuse the model.



The screenshot shows a Jupyter Notebook interface with a browser window at the top displaying a GitHub Gist. The notebook has two cells. The first cell, labeled 'In [77]:', contains a comment: `test_fname = 'PullUp.mp4' # downloaded from https://www.youtube.com/watch?v=aAggnpPyR6E&ab_channel=CrossFit%26AE`. The second cell, labeled 'In [90]:', contains a Python script that uses `cv2.VideoCapture` to read frames from a video file. It loops through the first 500 frames, prints the frame number and the prediction 'PullUps', and increments a counter. The output of the script shows the first 20 frames, all correctly predicted as 'PullUps'.

```
In [77]: test_fname = 'PullUp.mp4' # downloaded from https://www.youtube.com/watch?v=aAggnpPyR6E&ab_channel=CrossFit%26AE

In [90]: path = f'{test_fname}'
vidcap = cv2.VideoCapture(path)
count = 1
for i in range(500):
    success, image = vidcap.read()
    if not success:
        break

    print('frame', count, end=' ')
    label = predict_image(image)
    print(label)

    count += 1

frame 1 PullUps
frame 2 PullUps
frame 3 PullUps
frame 4 PullUps
frame 5 PullUps
frame 6 PullUps
frame 7 PullUps
frame 8 PullUps
frame 9 PullUps
frame 10 PullUps
frame 11 PullUps
frame 12 PullUps
frame 13 PullUps
frame 14 PullUps
frame 15 PullUps
frame 16 PullUps
frame 17 PullUps
frame 18 PullUps
frame 19 PullUps
frame 20 PullUps
```

Figure 2: Testing on images

98 5 Web client Nguyen et al. [2020]

99 As mentioned above our purpose was to develop an application we could apply test the model in real
100 life situations. Initially the decision fell on an Angular web app but because of compatibility issues
101 with Tensorflow Js, we decided to stick with a Vanilla Js webpage.

102 Tensorflow Js is a great tool which helped us convert the python model into a JSON file with several
103 .bin files for weights. In javascripts we could use methods provided by TfJs to load the model from
104 the JSON file and use it to make predictions.

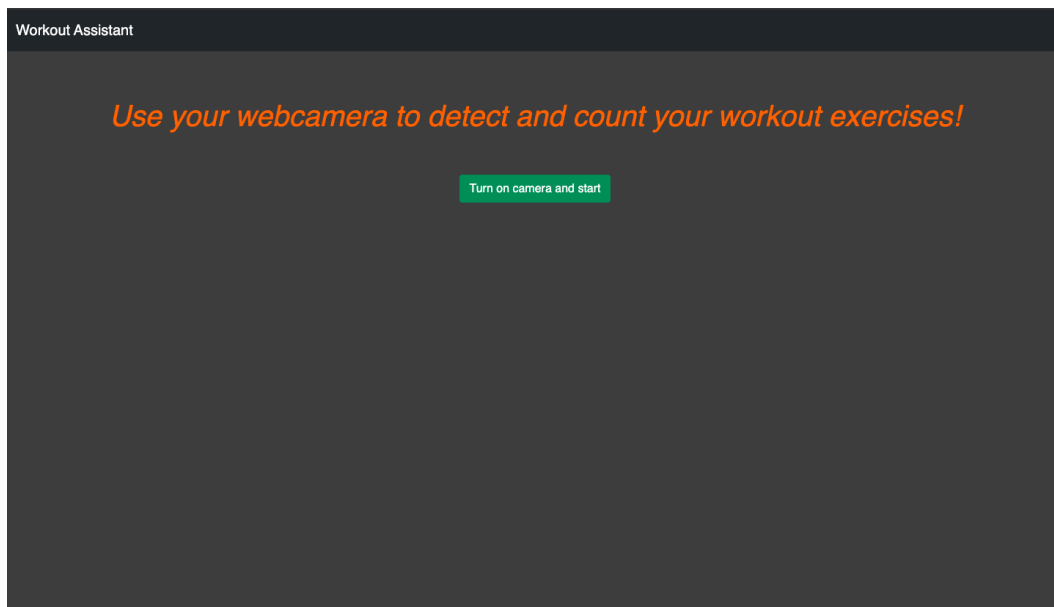


Figure 3: Testing on images

105 In our simple web application you can enable the web camera on your device to provide input for the
106 testing. While the camera is on the application counts the repetitions of each type of exercises the
107 model recognises.

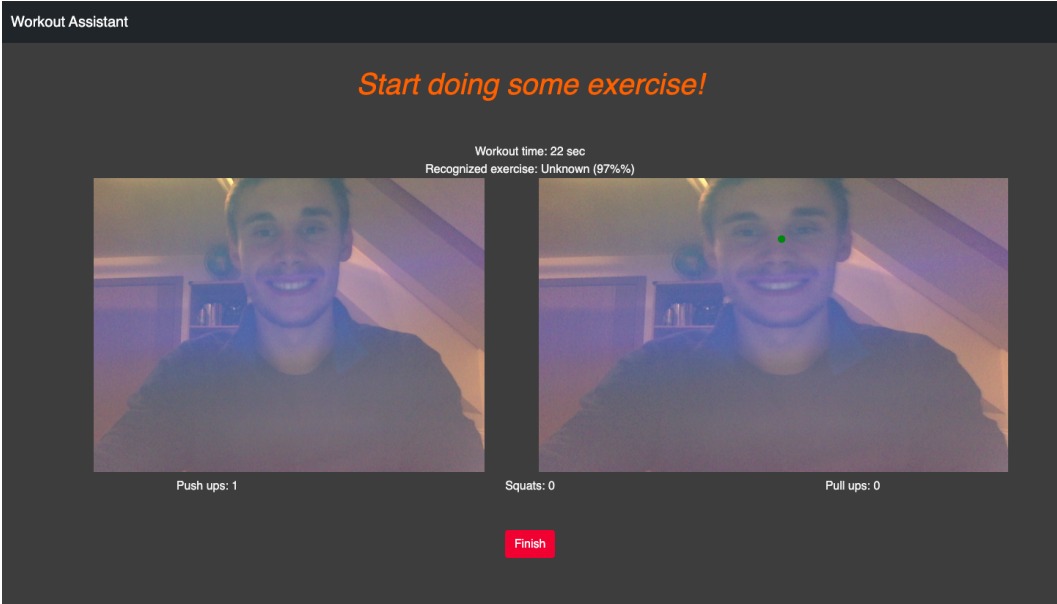


Figure 4: Testing on images

108 After hitting the Finish button it shows a summary of the workout time in which it shows pie diagrams
109 about the number and time of each exercise you did.

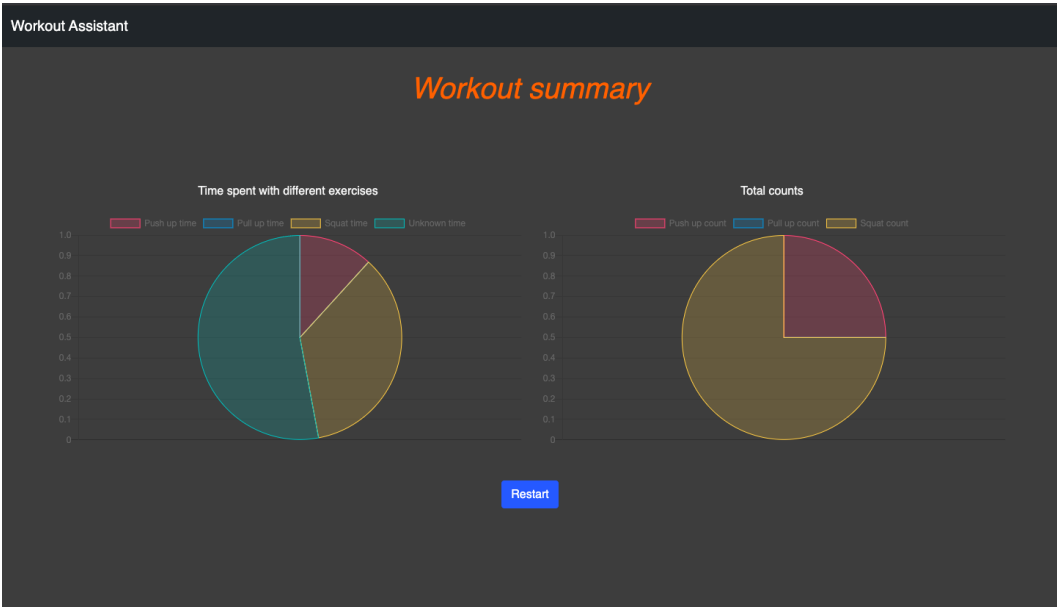


Figure 5: Testing on images

6 Summary and future plans

We found this project quite challenging, but fun as well, meaning that it was interesting to learn about ways of tackling such problems while creating something useful as well. It was not only a machine learning, but an engineering problem as well. Another issue we faced was having a team of two instead of the original three. This made the whole project harder as well, but regarding the final result we believe that it's safe to say that we were successful in creating our original goal, even if not in it's most perfect or complete form.

For the future we planned to use the app and show it to friends as well. Of course it can still be improved upon with new and better features, which we hope to continue working on together soon.

7 References

The following articles were used for the project:

- web based deep learning: Taheri et al. [2017], Smilkov et al. [2019], Nguyen et al. [2020], Nguyen [2020]
- computer vision: Nagarkoti et al. [2019], ATI [2021]
- dataset: Soomro et al. [2012]
- physical exercise: Soro et al. [2019], Islam et al. [2017], Kothari [2020]

for web based machine learning:

References

- H. Nguyen, M. Nguyen, Q. Nguyen, S. Yang, and H. Le. Web-based object detection and sound feedback system for visually impaired people. In *2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2020. doi: 10.1109/MAPR49794.2020.9237770.
- Sajjad Taheri, Alexander V. Veidenbaum, A. Nicolau, and M. Haghighat. Opencv . js : Computer vision processing for the web. 2017.
- Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viégas, and Martin Wattenberg. Tensorflow.js: Machine learning for the web and beyond, 2019.
- Kha Nguyen. Real-time fashion items classification using tensorflowjs and zalandomnist dataset, 2020.
- A. Nagarkoti, R. Teotia, A. K. Mahale, and P. K. Das. Realtime indoor workout analysis using machine learning computer vision. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1440–1443, 2019. doi: 10.1109/EMBC.2019.8856547.
- Plant leaf disease classification using efficientnet deep learning model. *Ecological Informatics*, 61:101182, 2021. ISSN 1574-9541. doi: <https://doi.org/10.1016/j.ecoinf.2020.101182>. URL <http://www.sciencedirect.com/science/article/pii/S1574954120301321>.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.
- Andrea Soro, Gino Brunner, Simon Tanner, and Roger Wattenhofer. Recognition and repetition counting for complex physical exercises with deep learning. *Sensors*, 19(3):714, Feb 2019. ISSN 1424-8220. doi: 10.3390/s19030714. URL <http://dx.doi.org/10.3390/s19030714>.
- Muhammad Usama Islam, Hasan Mahmud, Faisal Ashraf, Iqbal Hossain, and Md Kamrul Hasan. Yoga posture recognition by detecting human joint points in real time using microsoft kinect. pages 668–673, 12 2017. doi: 10.1109/R10-HTC.2017.8289047.
- Shruti Kothari. Yoga pose classification using deep learning, 2020.