Deep Learning – Final Project.

Idan Kelman, Doron Chapnitsky:

Info About Us:

Idan kelman 207190828, kelmanidan@gmail.com

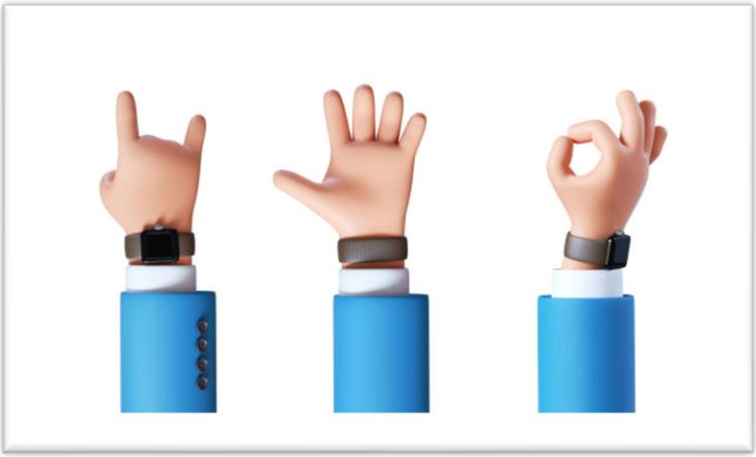
Doron Chapnitsky 316231190, chapnitskydoron@gmail.com

Introduction:

An approach to hand gesture classification via convolutional neural network.

The problem:

Since the last four decades, almost all forms of human gestures have been studied and used either as a natural or intuitive method to interact with computational devices. To compliment, all input–output technologies have been supportive of gesture-oriented interactions.



The use of gestures acts as a much more attractive yet effective alternative of complex interface devices.

Gestures are as natural as computers completely integrated into our life. We were curious if we would be able to classify between different hand gestures, and the applications are enormous:

If we would be able to correctly classify between the basic hand gestures that we are training our model on, then the applications can be directly reduced to other problems such as :

Translating the sign language, replacing a keyboard, replacing other forms of interaction between a human and a machine and so on .

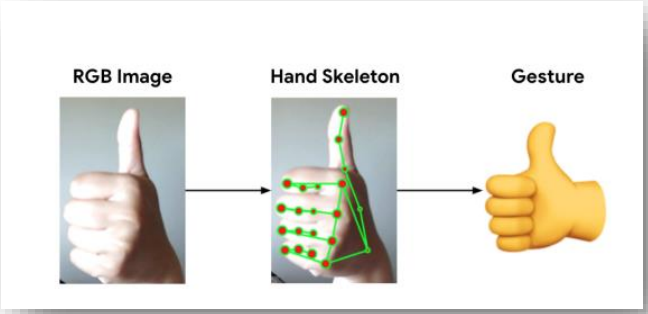
Predecessors to this work:

The problem that we are trying to face is pretty common , therefore there had been a lot of research on this topic .

Even though there had been a lot of research , they often solve a small sub-set of hand gestures , for example : one might detect finger indices , the other might detect swipe hand movements , and so on : as an example : we found the following paper by Google , called On-device Real-time Hand Gesture Recognition [https://arxiv.org/pdf/2111.00038.pdf]. in their work ,they suggest a small range of emogi type of symbols as follows :

The main differences between our work and Google’s is in the steps of the classification:

Google suggest a classification of the key points of the hand as the basis of the hand skeleton tracker, which will be later fed onto a Neural Network to predict the hand gesture.



Our method is a bit different; we are using a CNN network to classify directly from the input image. another key difference between the two works, is that our model is predicting more types of hand gestures, and different kinds of them.

Our Dataset:

We used a custom dataset that find online, that is containing 10 different classes of hand gestures, 20,000 different greyscale images that were shot by 10 different people .

- We normalized our data in order for it to fit into the CNN, by resizing the images to 128x128.
- Since we want to be able to recognize the hand gestures from a video, we had to make sure that we have all of the angles of the hands, therefor we also applied some horizontal and vertical random rotations to the images
- On some of the images, we also made it harder for our CNN to recognize what's the hand gesture inside the image, by merging two images into one, containing a distraction.
- In order to correctly use the dataset, we then finally organized the batches in custom classes for our model.

Methods:

The method that we used in order to solve this problem is using a convolutional neural network. This way, we were able to train a network that will be able to predict the correct classification of what occurs on the image with a good precision.

There are many other ways of trying to classify the outcome, but this way we were able to skip a stage in comparison with the previous predecessor offered by google. The alternative method for trying to predict is using an encoder-decoder CNN network, naming a few options: GAN's , DDPM and so on. This way we would also be able to predict where exactly the hand gesture is placed inside the image , but this is currently an information that we don't need , since we just want to classify one hand gesture , and we don't care about the placement, therefor our method can predict faster than other method , since the forward process will take less time (less variables) .

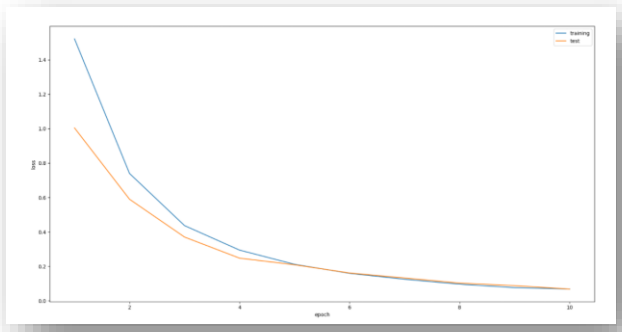
Results:

We made several test on how good the network performs, and we have a lot of data and 10 classes to predict from. We train our model on the custom data that we described on the previous steps, and plotted the results:

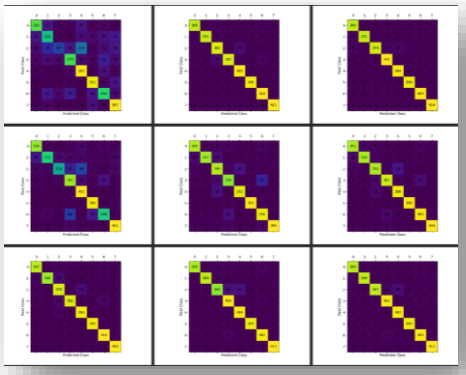
Google's Model

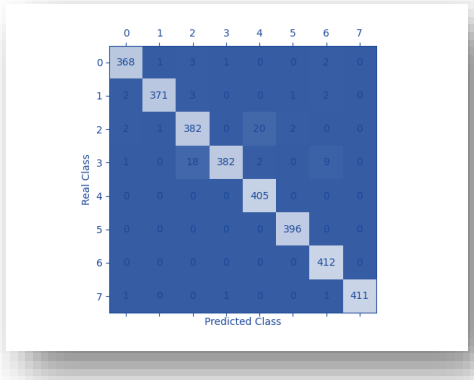
The NN classifier achieves an average recall rate of 87.9% across the 6 static gesture classes at a false positive rate of 1%.

Our Model: 98% acc

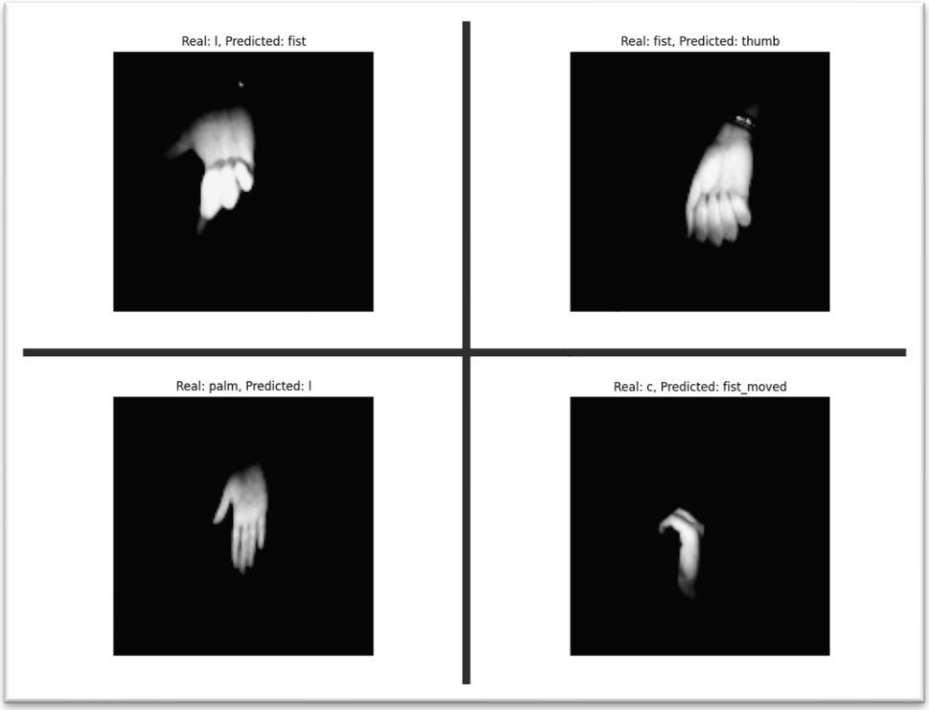


We used different batch sizes, different normalization methods (changing the rotation by [-x,x]) different image sizes , and we found out that the best way to train our model (using an neuro-evolutional neural network) is using a fixed image size of 128x128 pixels , with rotation range of +-15 degrees . this way , we were able to train our model with an accuracy of 98 % : we made various changes to the architecture of the model, the variables, the methods that we used,etc : and were able to come up with better and better results :





Not that the images that our network fails to classify are very debatable hand gestures, that even most people would not be able to classify correctly:



Where does our model lack in performance?

Our model works really well when an image of a hand alone is being shown, other scenarios, like outdoors, or in cases where the hand is not in front of plan background.

There are various methods that might be good in order to try and classify the hand gestures. It might be smart to use some transfer learning from recent encoder-decoder CNNs like DDPM (Denoising Diffusion Probabilistic Models) that excels in predicting multiple instances of the training data, and where they are located, or using a transfer learning from an already pretrained model that was trained on a data that closely resembles the data that we worked on.

Conclusions:

In conclusion, our custom trained model performs greatly on a very specific scenario of cases. There are many other issues that can be solved by reducing from our implementation, e.g.: classifying the single language, playing video games with hand gestures etc.