Deep Learning for Vision Project Milestone

30093813

ACM Reference Format:
30093813. 2023. Deep Learning for Vision Project Milestone. 1, 1 (March 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

In recent times, remote communication has become a more significant part of our lives. People are interacting more frequently using video conferencing tools, which has led to the need for better communication and interaction during online video calls. This project aims to develop an image classifier that can identify the hand gestures "thumbs-up," "thumbs-down," and "okay," in a video call participant's video feed, and to classify the case of none of the aforementioned gestures being made. The classifier can be used to provide real-time natural feedback during video calls, which can increase interaction, engagement, and the overall effectiveness of the video call.

As an anecdotal example, at the University of Calgary students rarely have their cameras turned on, which can make teaching more challenging as teachers don't receive feedback from students. Similarly, remote workers may not share their cameras during meetings, leading to communication difficulties. In larger meetings, it may not be possible to have everyone's camera on due to bandwidth limitations, and it can be challenging to gauge reactions from individual video feeds.

The proposed classifier for common "reaction" gestures would enable video call participants to provide feedback naturally and effortlessly. Although some video call software includes built-in emoji "reactions," they can be slow and unnatural, failing to replicate the type of real-time feedback people expect in face-to-face interactions. The proposed classifier offers a compromise between the desire to maintain privacy by not sharing camera feeds and the need for real-time natural feedback in video calls.

Hand gesture recognition has been the subject of several research studies in recent years. There are multiple areas of research in hand gesture recognition areas, with the most prominent being recognition via kinetics (data collected from wearable devices), and recognition via computer vision, sometimes using special cameras with depth sensing capabilities. We will focus on research for recognition via computer vision without special depth sensing equipment, i.e. recognition with just RGB images.

Kapitanov et al., the paper introducing the dataset used by this project, provided a baseline exploration of hand gesture recognition. The authors trained multiple models for both classification and detection tasks. In classification of 19 gestures, ResNeXt-101 achieved an F1-score of 99.28, and MobileNetV3 Large + SSDLite achieved 71.49 mAP score in the object detection task. Notably for this project, MobileNetV3 small and large achieved 96.78 and 97.88 F1 scores respectively in gesture classification, with respective inference times 8.9 and 16.9 milliseconds. As the proposed classifier would ideally run on personal computing devices, a smaller model such as MobileNetV3 would benefit the

Author's address: 30093813.

Manuscript submitted to ACM

project. Further, faster inference times with less expensive computations would be best as the inferences need to run frequently over long periods of time.

Miah et al., published January of this year (2023), proposed a novel multistage spatial attention-based neural network for hand gesture recognition. The authors did not give the model a name, so I will refer to it as AttentionHGR. The proposed model utilizes three stages: in the first two stages, concurrent convolutional "attention" and "feature-extractor" modules are trained and combined, with the intent to have the attention modules help the rest of the network identify the important regions of the image; the last stage is a convolutional classification module. For hand gesture recognition the attentions modules are important, as the hand gesture in an image may only be one small part of the whole image. The attention modules are based on the spatial attention module A_T from Mohla et al., which introduced the FusAtNet model for classification from multi modal data. Further details of the architecture are discussed below, as this project will attempt to utilize this architecture. AttentionHGR outperformed the existing state-of-the-art models on the tested datasets.

GOAL

The project aims to develop a classifier which can be used by video conferencing software to enable real-time, natural feedback and interactions from participants via three "reaction" types: "thumbs-up," "thumbs-down," and "okay," which correspond to emojis 👍, 👎, and 👌, respectively.

Taking into account the performance achieved by Kapitanov et al. and the limited computational resources available to me, I expect the model to achieve a minimum average F1-score of 90, with average inference time under 200 milliseconds. I expect the error across classes to be approximately even, given the size and balance of the dataset. Given the nature of the application, some error is acceptable, and speed should be a priority to achieve the goal of the feedback feelings as natural as possible. It would also be best to optimize for model size, but for my own learning experience I may forgo this as most interesting models seem to be over 10 megabytes.

DATASET

As mentioned previously, Kapitanov et al. introduced this project's dataset, hereby referred to as Hagrid. The Hagrid dataset was designed for hand gesture recognition systems, with the classes selected with the intent of their use in human computer interaction systems, particularly for device control. The dataset is the result of crowd-sourcing in Russia. The data collection was a four stage process:

(1) Mining: crowd-source workers were tasked to take photos of themselves making various hand gestures under various conditions (lighting, locations), and some preliminary filtering was performed to remove bad images.
(2) Validation: crowd-sourcing is again used to validate the mined images, this time with trained crowd-source workers. A majority vote system is used to identify high quality images.
(3) Filtration: ethically questionable images are removed. A similar crowd-sourcing process was again used for this, however this time only employees of the company developing the model were used.
(4) Annotation: crowd-sourcing is again used to annotate the images with bounding boxes and classes. Crowd-workers were again trained, and a majority vote system was used. The dataset was annotated two times, and the annotations were aggregated for each image.

The dataset is approximately 716 GB including more than 550 thousand high-resolution images divided into 18 classes of gestures. For this project only a subset of the dataset will be used, including only the gesture classes “like”, “dislike”, and “ok”. The classes are fairly evenly distributed, with each class containing between 30-32 thousand examples.

The dataset contains many unique faces and scenes. The subjects are people from 18 to 65 years old. The ratio of women to men is about 27 to 20. The dataset was collected mainly indoors with considerable variation in lighting, including artificial and natural light. The dataset includes images taken in extreme conditions such as facing and backing to a window. The subjects also had to demonstrate gestures at different distances from the camera. Data on the distribution of skin-tones of the subjects is not available, but given the singular location of the crowd-sourcing and some scrolling through the images, the dataset does not appear to contain much diversity of skin-tone.

Figure 1 shows the types of images in the dataset, and figure 2 shows the different hand gestures.

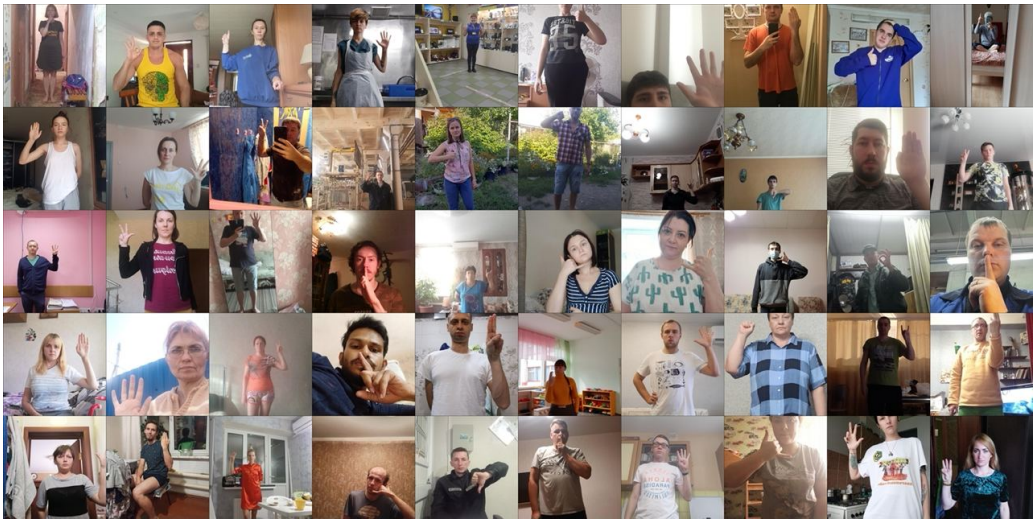

Fig. 1. A sample of images from Hagrid

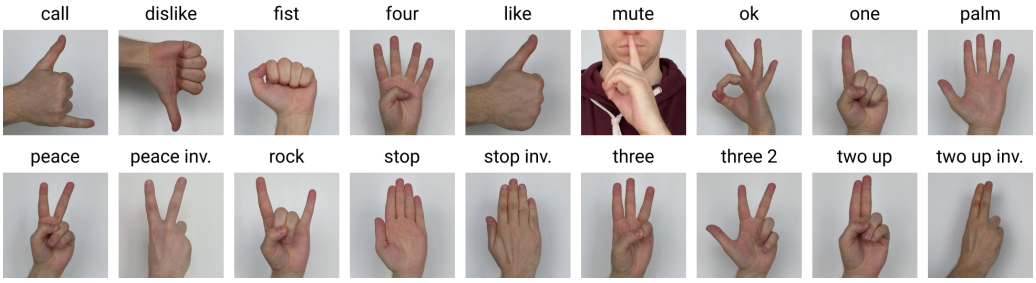

Fig. 2. Classes of Hagrid

METHODOLOGY

Preprocessing

To deal with storage space limitations, the images are resized on the file system, as most models do not require the large image sizes provided by the dataset. This resizing enables the dataset to be uploaded to Colab.

Further preprocessing and data augmentation will be used to suit the particular model architectures tested:

(1) For pretrained MobileNet, images will be resized, center-cropped, and normalized to match ImageNet.
(2) For AttentionHGR, the authors appear to use very minimal image preprocessing, only resizing the images to size 160x160. Further experimentation may be required with preprocessing to adapt the architecture to the Hagrid dataset.

In addition to the above general preprocessing, various random augmentations will be experimented with, including for example color jitter and gray-scaling to hopefully help the model generalize to different skin-tones.

Model Architecture

I will first attempt to train MobileNetV3 as a sort of "MVP", so that a demo can be performed using live video feed. I will then attempt to implement AttentionHGR in Pytorch.

For MobileNet, I will try using the pretrained model available with Pytorch, possibly unfreezing layers as the dataset is quite large.

For AttentionHGR, upon successful implementation of the model if time permits, I will attempt to pretrain the model using the bounding box annotations of the Hagrid dataset, with the aim of providing a better starting point for the attention modules.

As MobileNet is more well known and is already implemented in Pytorch, I will focus here on the architecture of AttentionHGR.

The model expects 160x160 RGB images as input. The architecture is shown in Figure 3. Note that the grey boxes are definitions of named modules used within other modules. As described previously, the model consists of two "attention" stages and a classification stage. In the attention stages, feature extraction and attention modules are combined. The original image is concatenated with the output of the first stage, so the second stage has the output of the previous layer and the image data as input, with the idea being that the first stage extract spatial features from the image to aid the second stage in highlighting key areas of the image.

The feature extraction modules are basic six-layer 2D convolutional networks with each layer followed by batch normalization. The attention modules are six-layer convolutional networks with residuals after the second and fourth layers, and again each convolutional layer is followed by batch normalization. Matrix multiplication is used to combine the concurrent attention and feature extraction modules. Finally, the output of the second stage is fed to a fairly standard convolutional classifier module utilizing ReLU activation, batch normalization, max pooling, and softmax for the final output. The kernel size of all convolutional layers is 3x3.

Postprocessing

Time-permitting, I may attempt to ensemble the MobileNet and AttentionHGR, and possibly other models. However, given the results of the previously mentioned papers, I don't anticipate postprocessing for performance purposes will be required.

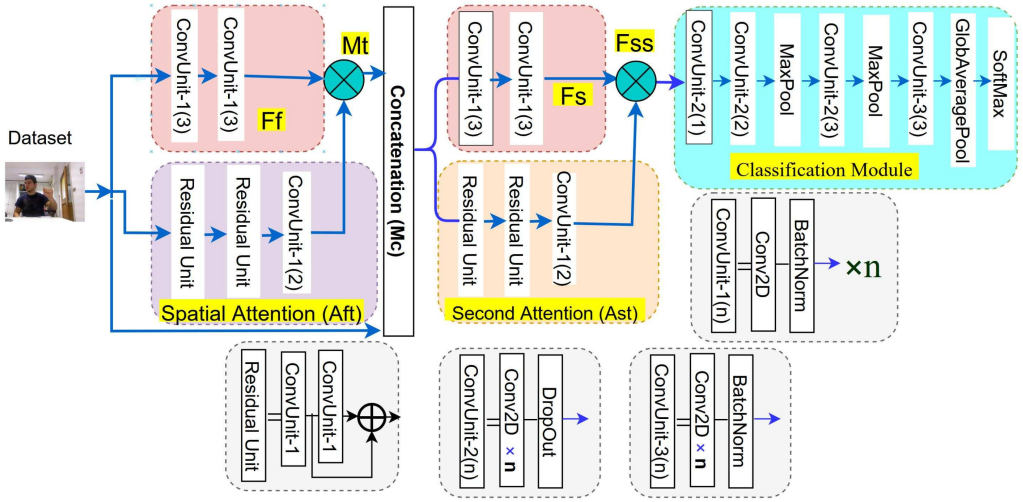

Fig. 3. AttentionHGR flow diagram

CHALLENGES AND SOLUTIONS

One of main challenges encountered so far is the size of the dataset. To deal with this, I have augmented the images on the file system to have size 300x300, reducing the training dataset size from almost 100GB to just under 1GB. Another issue I have encountered is running out of RAM when attempting to train a model on the entirety of the training set. A possible solution is to train only on a subset, as the dataset is very large so even a subset may be sufficient.

Another possible challenge will be handling the heterogeneity of the data. The dataset contains many different images with different backgrounds, people, lighting, and distances from the camera. However the result from Kapitanov et al. indicate that it should be possible to train a model that generalizes well.

Another possible issue previously mentioned is the lack of skin-tone diversity in the dataset. Hopefully data augmentation can help with this, and if not, my skin-tone luckily seems to match the majority of the images so a demonstration should be possible.

UPDATED TIMELINE

By March 27th I will plan to have a working demo using MobileNet, where live video feed from my laptop's camera can be classified. The remaining time before the project submission deadline will be used for preparation of the presentation, the final report, and for further experimentation, including development of the Pytorch implementation of AttentionHGR. (sorry a Gantt chart seemed like overkill for this, I'm guessing it was more intended for groups with multiple people)

REFERENCES

[1] Alexander Kapitanov, Andrew Makhlyarchuk, and Karina Kvanchiani. 2022. HaGRID - HAnd Gesture Recognition Image Dataset. https://doi.org/10.48550/ARXIV.2206.08219

[2] Abu Saleh Musa Miah, Md. Al Mehedi Hasan, Jungpil Shin, Yuichi Okuyama, and Yoichi Tomioka. 2023. Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition. *Computers* 12, 1 (2023). https://doi.org/10.3390/computers12010013

[3] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. 2020. FusAtNet: Dual Attention Based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.