Deep Learning for Vision Project Proposal

30093813

ACM Reference Format:
30093813. 2023. Deep Learning for Vision Project Proposal. 1, 1 (March 2023), 3 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.

? , 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 project. Further, faster inference

Author's address: 30093813.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

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

? , published January of this year (2023), proposed a novel multistage spatial attention-based neural network for hand gesture recognition. 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. For hand gesture recognition this is important, as the hand gesture may only be one small part of the whole image. The attention modules are based on the spatial attention module A_T from **?** , which introduced the FusAtNet model for classification from multi modal data. Further details of the architecture are discussed below, as the project will attempt to utilize this architecture.

METHODOLOGY

- Perhaps pretrain model using bounding boxes somehow
- Utilize attention

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.

DATASET

- 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.

CHALLENGES

- Large dataset
  - Running out of ram
- Heterogeneity of dataset
- Limited computational resources
- Lack of skin tone variety?

CHATGPT1

The HaGRID dataset is a collection of more than half a million high-resolution images, divided into 18 classes of gesture signs. The dataset was created to design hand gesture recognition (HGR) systems for human-computer interaction. The gestures are chosen for their semiotic functional role, which helps communicate information between people, and in this case, between humans and computers. The dataset includes a small lexicon of functional gestures to reduce HGR system complexity and avoid unnecessary cognitive load on the device user. The dataset also includes an extra class with samples of natural hand movements called "no gesture."

The annotation process for the dataset involved using two crowdsourcing platforms, Yandex.Toloka and ABC Elementary, to collect the images with different backgrounds, lighting, scenes, and subjects. The images were collected in the RGB format to ensure heterogeneity in the dataset.

The distribution of samples per class is not specified in the information provided. However, Figure 1 in the paper shows the 18 gesture classes included in the dataset. These gesture classes are not language-oriented and were selected as the most useful for designing a device control system. The classes include “no gesture,” “stop,” “go,” “next,” “previous,” “like,” “dislike,” “call,” “hang up,” “volume up,” “volume down,” “volume mute,” “scroll up,” “scroll down,” “zoom in,” “zoom out,” “inv. peace,” and “inv. rock.”

The paper also provides a review of related work on existing HGR datasets, a description of the process of creating the dataset, and several models trained on the dataset with experimental results. The paper concludes with a description of future directions for work in this area.

CHATGPT2

The Hagrid dataset consists of approximately half a million FullHD RGB images with 18 different gestures and a “no gesture” class. There are at least 34,730 unique scenes in the dataset. The dataset was created to combine high-resolution images, heterogeneity across the image scene, and a high number of samples. The dataset also includes some gestures in two positions, allowing for dynamic gestures to be interpreted using two static gestures. The dataset can be used for hand detection problems, gesture/non-gesture binary classification problems, and right/left leading hand binary classification problems.

The dataset was collected in four stages: mining, validation, filtration, and annotation. The mining stage involved crowd workers taking a photo of themselves with the particular gesture indicated in the task description. The validation stage was implemented to achieve high confidence in the images collected in the mining stage. The filtration stage involved filtering out inappropriate images, and the annotation stage involved marking up bounding boxes and leading hands. Two Russian crowdsourcing platforms, Yandex.Toloka and ABC Elementary, were used for the data collection process.

The workers were required to complete a training process before performing the tasks, and the mining tasks were accompanied by instructions with a warning about the further publication of the crowd workers’ photos. The validation stage established that some users tried to cheat the system during the mining stage. The goal of the validation stage was to select correctly executed images at the mining stage, and only “correct” images were added to the dataset. For each image at the validation stage, the system set a dynamic overlap of 3 to 5 performers, and some photos were rejected based on the majority rule.