Machine Learning based GPT for the differently abled

Abstract:

The main idea of this project was inspired from seeing the advancement in technology and the ability to simplify tasks and receive answers in an instance that the vast number of abled people are able to take advantage of, but there is also a large group of people who are unable to utilize normal methods of communication. These people utilize Sign Language to communicate on a daily basis, and the numbers of people affected by it are in the billions. Thus came the idea to integrate machine learning, computer vision and neural networks to be able to create a GPT for the differently abled, to answer their questions, and to remove their dependencies on people around them.

Implementation method:

The aim is to create a responsive machine learning model that is trained and is able to understand different hand symbols and be able to create a sentence through the sign language inputs of the user. This code will predominantly be in python and utilize python's vast number of libraries like Open CV, Tensorflow, Mediapipe to utilize and train the model through computer vision, what different symbols stand for and mean utilize them. To train the model through computer vision, Google has a very useful and advance library called Mediapipe that helps vectorize the human joints in computer vision input that helps train the model far quicker and more efficiently than it would be to train it based on pixel movement.

For the chatbot part of the system we aim to create many functions and a main function. The main function would contain the general structure of the chat, and it would take input from the user. The input from the user would be sent to input processing function that would utilize Spacy library that is a Natural Language Processing(NLP) module to tokenize and Vectorize the input from user and be able to call from the different functions created to be called form to get a satisfactory reply from the user. The system would act as a responsive actual chatting and answering GPT. It would have features like fetching current weather in different cities and the Top news when asked. It would be able to do all this through API keys.

The system would also have a general function where the prompt entered would be able to give an in depth solution by utilizing the OpenAI's ChatGPT API key to give an answer. This option has been provided as the computation required to create a responsive and in-depth system like ChatGPT is unmatched and would take years of machine learning and training to give all the answers a user would enter.

What is Computer Vision?

Computer vision (CV) is a subfield of artificial intelligence (AI) that enables computers to process, analyse, and interpret visual data. The goal of computer vision is to enable computers to identify objects and people in images and videos, and then take appropriate action.

Computer vision applications mimic the functioning of the human visual system by utilising data from sensing devices, artificial intelligence, machine learning, and deep learning. Algorithms used in computer vision applications are trained on vast volumes of visual data or cloud images. They identify patterns in this visual input and make inferences about the meaning of other images based on those patterns.

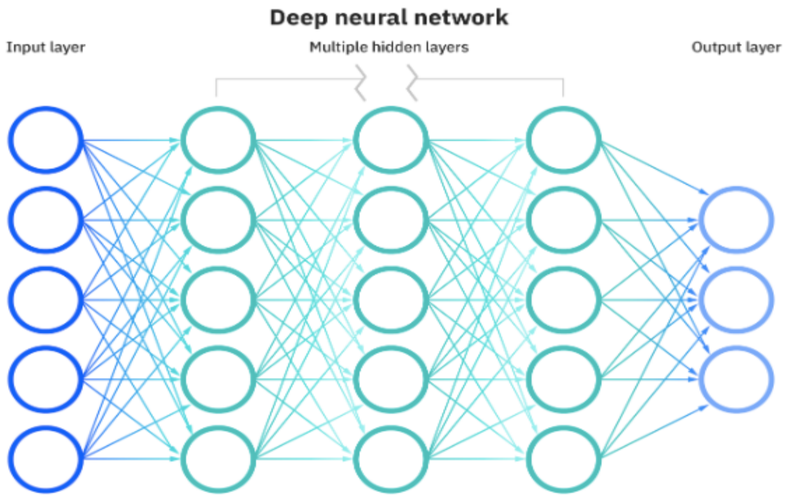
Neural networks

A neural network is defined as a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.
It generally consists of three or more layers to train and re-establish trends and similarity between the data fed into the model and rectify what the expected answer is for each permutation and combination of data.

How does the node work?

Each node, or artificial neuron, in the neural network connects to another and has an associated weight and threshold assigned to it. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network.

The system rely on training data to learn and improve their accuracy over time and can become extremely powerful computational mechanism if their accuracy is perfected.



Through the layers of deep neural network, segmentation and classification takes place by trained previous trend, moving the known information from node to node, thus giving a classified output in output layer

Once an input layer is determined, weights are assigned. These weights help determine the importance of any given variable, all inputs are then multiplied by their respective weights and then summed. Then the output is passed through an activation function, which determines the output. If that output exceeds a given threshold it activates the node of the next layer, thus making the neural network act as a feedback loop.

A great example of how weights are assigned in neural networks given by IBM is given below:

We can apply this concept to a more tangible example, like whether you should go surfing (Yes: 1, No: 0). The decision to go or not to go is our predicted outcome, or y-hat. Let’s assume that there are three factors influencing your decision-making:

1. Are the waves good? (Yes: 1, No: 0)
2. Is the line-up empty? (Yes: 1, No: 0)
3. Has there been a recent shark attack? (Yes: 0, No: 1)

Then, let’s assume the following, giving us the following inputs:

- X1 = 1, since the waves are pumping
- X2 = 0, since the crowds are out
- X3 = 1, since there hasn't been a recent shark attack

Now, we need to assign some weights to determine importance. Larger weights signify that particular variables are of greater importance to the decision or outcome.

- W1 = 5, since large swells don't come around often
- W2 = 2, since you're used to the crowds
- W3 = 4, since you have a fear of sharks

Finally, we'll also assume a threshold value of 3, which would translate to a bias value of –3. With all the various inputs, we can start to plug in values into the formula to get the desired output.

If we adjust the weights or the threshold, we can achieve different outcomes from the model based on current inputs.

Neural networks normally leverage sigmoid neurons, which are different by having weights between 0 and 1. Most deep neural networks are feedforward, meaning they flow in one direction only, from input to output.

Types of Neural Networks:

Neural networks can be divided into two categories: Convolutional neural networks (CNNs) similar to feedforward networks, are used for image recognition, pattern identification, and/or computer vision. To find patterns in a picture, these networks use ideas from linear algebra, particularly matrix multiplication.

Recurrent neural networks (RNNs) where presence of feedback loops identifies When using time-series data to predict future events, such as stock market projections or sales forecasting, these learning algorithms are generally used.

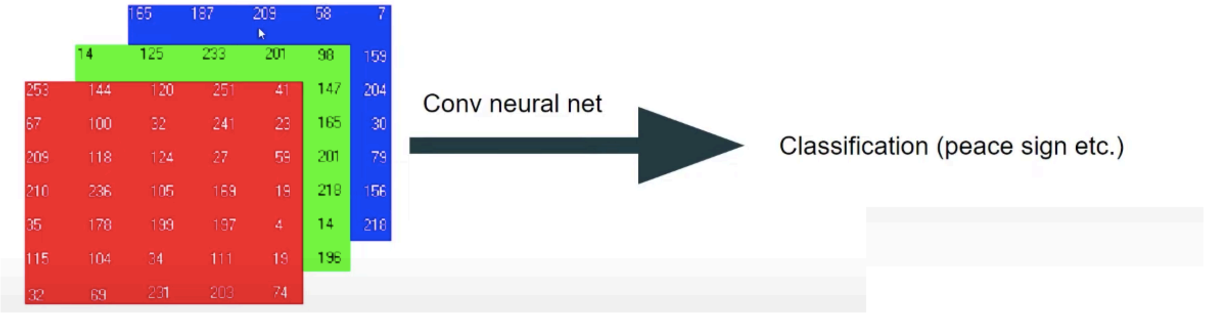
Creating the ML Sign language model:

The two possible approaches to train the model in understanding the hand gestures where :

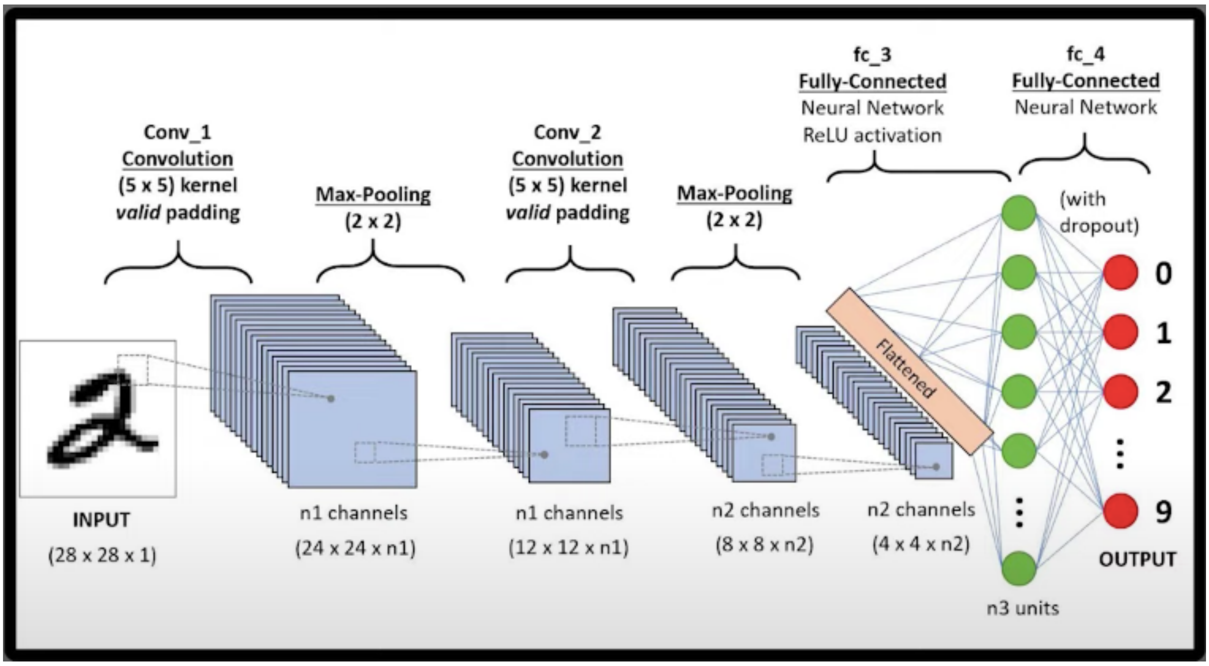
- Pixel by Pixel training:

Where the creator would require the model to be trained by showing the same hand gesture many-many more times in different areas of the camera and different backgrounds for it to understand the symbol and register the particular symbol. It would also require a large amount of computation and can be said is a more orthodox way to do it.

Here the pixels are sent to a different kind of a convolutional neural network where the system vectorizes the points for easier processing over training regressive pixels. The system would have to go through regression training in order to process data fed to it.



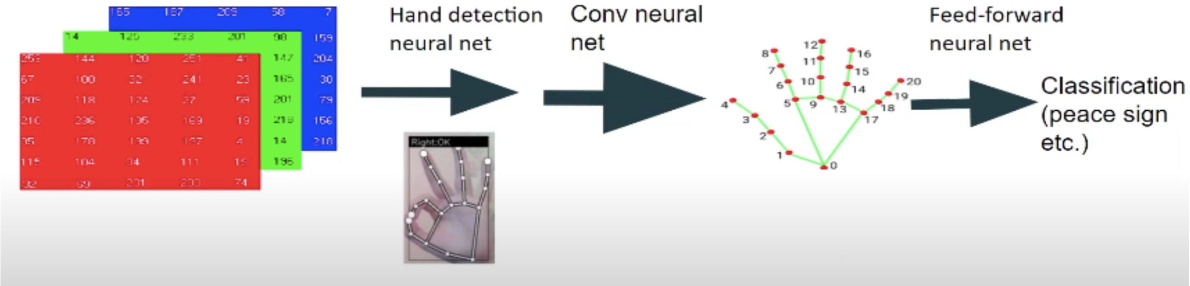
This system would then have to be put into a pretty complex convolutional neural network like so that is generally optimized and used for image processing and recognition:



- Mediapipe

The implementation when using Mediapipe, is to understand that it is trained to vectorize joints in the body, it is done by taking the pixel inputs and putting the input into a Mediapipe's hand detection neural network which is able to track body joints accurately in real time.

The vectors in the hand look something like the image given above and these vector points are fed into a neural network to get out the desired trained model. The neural network working is something as the below image.



The model system provided by Mediapipe eliminates problems like skin colour, camera focus, background light and other such factors to be eliminated because they are trained and have already collected a lot of data through training.

What Mediapipe provides?

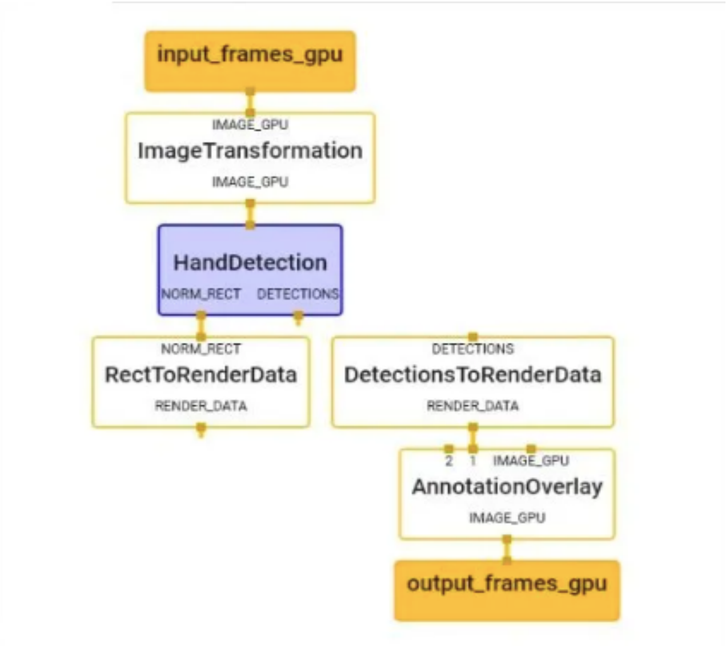
It is an open-source framework for building machine learning (ML) pipelines. It allows developers to create complex processing graphs for audio, image, and other sensor data

MediaPipe Framework consists of three main elements:

1. A framework for inference from sensory data (audio or video)
2. A set of tools for performance evaluation, and
3. Re-usable components for inference and processing (calculators)

The main components of MediaPipe:

- Packet: The basic data flow unit is called a “packet.” It consists of a numeric timestamp and a shared pointer to an immutable payload.
- Graph: Processing takes place inside a graph which defines the flow paths of packets between nodes. A graph can have any number of input and outputs, and branch or merge data.
- Nodes: Nodes are where the bulk of the graph’s work takes place. They are also called “calculators” (for historical reasons) and produce or consume packets. Each node’s interface defines a number of in- and output ports.
- Streams: A stream is a connection between two nodes that carries a sequence of packets with increasing timestamps



How the model will understand the sign language:

The model uses Mediapipe holistic function which is able to simultaneously estimate multiple aspects of the human body, including:

- Face Landmarks: It can detect and track facial landmarks like eyes, nose, and mouth.
- Hand Landmarks: It can estimate the landmarks of the hand, allowing for hand gesture recognition.
- Body Pose: It can track the key points of the human body, such as shoulders, elbows, hips, knees, and ankles.

This thus can be used to track the points in our hand and our posture, to train the model to understand different movements and associate to particular models.
Then Key point extraction occurs and we track the movement and are able to assign it to a label.

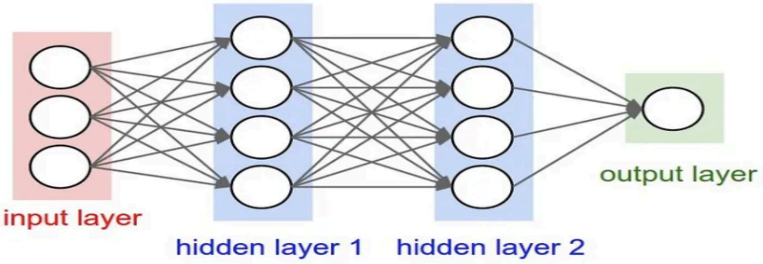
We first store labels in an empty list and make folders using OS library in local directory with name of each label.

When the training model is ran the model starts collecting data for 30 frames and using landmark extraction is able to record the movements and store it in NumPy array format. This loop of collecting data occurs 30 times so the model would have sufficient data to allocate and recognise certain movements with labels assigned to it.

Feed Forward Neural Network

The Neural network used in training the model is TensorFlow's layers. A Sequential Neural network is used as it is a simple and straightforward way to define a linear stack of layers.

A Sequential neural network is a neural network that passes data through a series of neural layers, one after the other, from input to output. A sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.

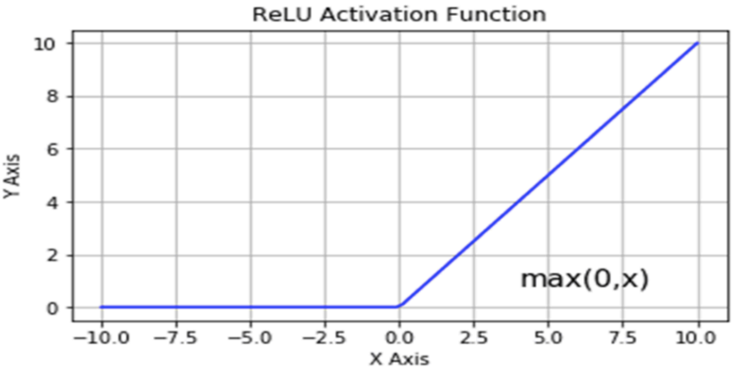


The model was trained using three LSTM models and three dense layer networks. A Long short-term memory(LSTM) network used in deep learning. It is a type of Recurrent Neural Network. It is used mainly in time series and sequence data. LSTM has feedback connections, which means it can process entire sequences of data at once rather than just a single data point such as an image.

In the layers of the LSTM model, there are 64 nodes in the first layer, 128 nodes in the second layer and 64 in the third layer, this is done so that the neural network is able to be trained more effectively.

The three dense layers also have similar number of nodes. The dense layer is a single simple layer of neuron in which each neuron receives input from all the neurons of previous layer. It is a fully connected layer in which each neuron in the layer is connected to every neuron in the preceding layer and feeds all the outputs from the previous layer to all its neurons providing one output to the next layer.

The activation functions used to get output from the neural network is Rectified linear unit (ReLU). The ReLU activation function is defined as f(x) = max (0,x) where x is input of the neuron. Introduces non-linearity into the model, which is crucial for networks ability to learn complex patterns in data.



Training the Sign language model

The aim of the project was to train the model on word by word rather than letter by letter as it would take user longer to input what they wanted to get answers to.

YES
NO
THANK YOU
HELLO
I LOVE YOU
GOODBYE
YOU ARE WELCOME
PLEASE
SORRY

The process of capturing and processing data includes utilizing the Google's media pipe library to vectorize the joints in our body and collect data based on the change in position from one point to another.

The NumPy array library is very useful in this as we are able to save these vectorized points in array matrix format for ease of the model to compare movement an identify the particular sign.

Once the model has been trained, the model while being used shows the labels being predicted at the top of the screen thus making it visible to the user.
These predicted words are appended to a list called sentence, so that the user can keep track of the predicted words that once entered is sent to the chatbot

This sentence list is then sent to the chatbot where it is able to attain back a response and convert this text-to-speech so any user is able to get back an answer and interpret it regardless of their disability.

Working of Chatbot

The chatbot for this project works on the basis of Natural Language Processing(NLP), which is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language.

The library that is mainly used in the implementation of NLP in the project is SpaCy. It is a widely used NLP library that provides a variety of linguistic annotations to give you insights into a text's grammatical structure.

SpaCy has many useful and important functions like tokenizer, and vectorizers to interpret sentences and give feedback to model.

Tokenizer:

- The tokenizer function breaks a string of text into tokens, which are the basic units of NLP.

Part-of-speech tagging:

- The part-of-speech tagger assigns a part-of-speech tag to each token, which indicates the function of the word in the sentence.

The Tokenizer function in Chatbot is used in the model to interpret the input from the user and provide them with an appropriate reply. The input from the users are converted into tokens and the model is trained to call certain functions when certain words are in the input.

The functions that the chatbot can do:

- Weather: Returns back the weather in any place in the world using API Key, and it is able to take input location from the user using tokens in SpaCy. For the API request we the

Request library. The website that is used to fetch the data was NewsApi.org due to its ease and accuracy in fetching data.

- News: This function is able to return the top 10 news in any city, and is able to provide a description for each news. The function also utilizes API requests and the website it uses to retrieve data is NewsApi.org .

- The model is able to reply to “Hello” and “what is your name” and is able to ask the user about their day and reply according to the users response.

The main function of the chatbot gives structure to the operation and calling of functions in the chatbot. When the chatbot is activated the main function calls the sign-language model, switching the camera on, to start taking input from the user. Once the input is decided the user presses ‘q’ to send the input to the chatbot.

Conclusion

This project was created keeping in mind the troubles of the differently abled, hoping to make their life just a little bit easier and making them just a little more self-reliant. The project was created with the help of multiple resources and a lot of help and guidance from the team at Wipro, who gave me this opportunity at picking any topic I choose, that would solve a problem and make a positive impact. They helped me, and pushed me to think and further my horizon through the multiple problems faced during the project. The model is just the first step in using computer vision in a good way to help out people who are differently abled, just making them feel a little more included in modern day advancing society.

Once again thanking Digraj Sir, for giving me this opportunity to grow and learn, from different mentors and to push me in taking up a project that would not only make a good impact, but help me hone a new skillset.