DNA Sequence Classification using CNN hardware acceleration

Submitted to Dr. Hassan Mostafa

Khaled Mostafa Mostafa	1170427
Ahmed Tarek Ahmed	1170445
Youssef AbdelHamid	1170415
Marwan Walid Ali	1170008
Momen Ehab Khaled	1170559

Acknowledgment:

First of all, we would like express our special thanks of gratitude to our professor Dr. Hassan Mostafa for taking us under his supervision, and for always being there for us, and supporting us constantly. Besides, providing us by all the needed material and valuable comments and advices and throughout our Graduation Project.

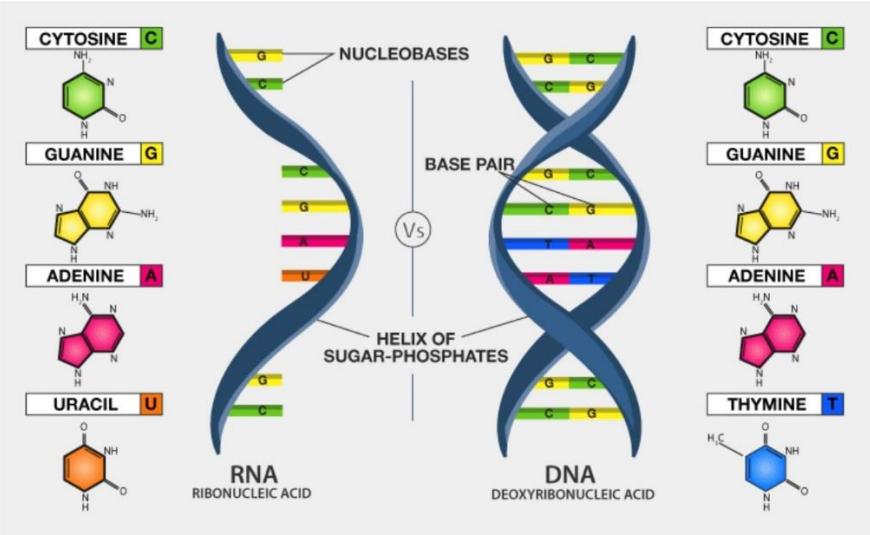
We would also need to thank One-Lab and IP-Valley for giving us the chance to be part of their professional institution, and their efforts in expanding our horizons into the various application that are waiting for us.

And finally, we can't express how lucky we are for being part of the Genome Project, which is the first in Egypt and North African. Genome and the identification of population-specific variation will support genetic research and "precision" medicine in Egypt and North Africa and will help to better characterize "Egyptian" genetic variations and evaluate their relevance for common diseases.

Introduction:

DNA sequencing is one of the most complex and analytical processes that take place in laboratories around the world, here in Egypt it takes a minimum of 5-10 days for results to be produced. The easy part is to acquire the DNA sample using a mouth swab, then comes the hard part of analyzing this data and understanding what it means.

To give a brief idea of the DNA sequencing process, the DNA consists of four types of molecules connecting two strands forming a double helix shape as the following figure (RNA also has four types of molecules but usually have a single strand). From the 4 types there are 2 complimentary molecules (G and C, A and T) so if you know one you already know the second one.



There is a test for each type of molecule that can be done on the sequence resulting in an array indicating all the molecules in the DNA that can be inserted into a computer. This sequence is then analyzed for features using computer software which takes a long time. There is also limited accuracy when it comes to the testing so the sequence is not 100% accurate and the same goes for the subsequent testing results.

This is the place where machine learning techniques, especially the convolutional neural networks are used, as they are dominant with high performance in analyzing the sequences using feature extraction. By transforming the DNA sequence into data and using CNN the results of the classification of DNA is much more accurate and efficient.

We aim to create a hardware-accelerated module that is able to utilize the CNN and perform the task with similar capabilities but with much faster performance than software-based devices.

Project Description:

- ## Problem Definition:

DNA sequencing is a growing field where the method of classification and storing of data for millions of people is both a local and international challenge to be able to classify the DNA data for everyone, which requires the highest speed without sacrificing the accuracy and performance requirement. The hardware acceleration of the CNN classification algorithm is the perfect fit for the solution.

- ## Overview of system modules:

1. CNN algorithm for DNA classification.
2. Verilog code for modules used.
3. FPGA for hardware testing.

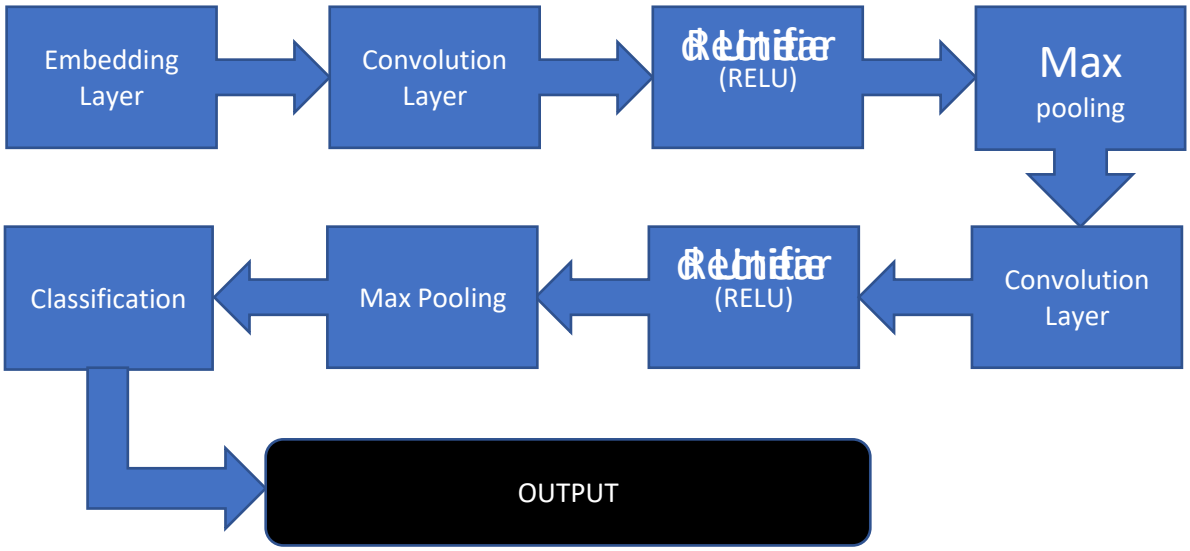
Impact on community, market and end-user:

The new national genome bank that will open in Egypt will create opportunities to use the genetic data of the public to do research and make studies of genetic patterns. This will open the door for new technologies concerning genetics, and create an increasing need for new technologies that increase the efficiency and speed of data enquiry and classification. Moreover, it will be required to remotely sequence DNA in inaccessible parts of our country especially in rural areas where the advanced medical centers are not available and its price is currently unattainable. Moreover, the latest COVID-19 pandemic also showed the need for more DNA studies and how essential the role of genome bank is to produce a medical background for the general public. Another crucial advantage to genome bank is security and making a DNA profile for everyone. Our project allows more efficient DNA sequencing and classification devices that have higher performance and speed while working on a mobile device. It can easily take part in the collection and classification of the DNA sequences and make the process easier, cheaper and more accessible to everyone.

Project's Final Outcome:

Our project required outcome is to produce a prototype of a DNA sequencing FPGA device that has acceptable performance compared to labs while being efficient, fast and accurate to produce meaningful results. The device would be easily fed the DNA data and is able to accurately classify it and produce results which in turn can be stored for each member of the population and take part in forming the Egyptian genome bank. This can also make Egypt one of the leading countries in the region in DNA sequencing and spread the use of similar devices to surrounding countries as they start forming genome banks.
If this succeeds it will allow the process of sampling the populace DNA and classifying to go much faster, much more efficient and be able to detect any genetic illness simply and efficiently. It can also be able to classify the RNA of diseases and detect the patient's type of illness. This device will also use a lot less power, a lot less space, be able to classify faster and cheaper than a regular computer which would allow it to be used even in the rural or the more economically challenged areas that can't afford the more expensive regular computer. This could be also be potentially used for households to analyze the illness of the patient if it is RNA or DNA based and can be used as over the counter appliance in pharmacies just as blood pressure and sugar level measuring device.

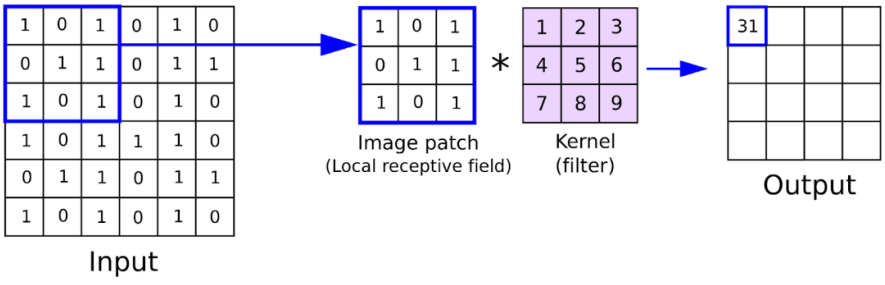
After analyzing the Given description code we settled for the hardware implementation to take the following form



Convolutional Layer:

Convolution layer is the heart of the CNN, it is the most important layer in the neural networks architectures, it is designed to work on two dimensional matrices of data, where at each layer there is introduced two sources of data one of them is called the filter (Kernel), the filter is the source of data that is carrying the features that are ought to be extracted from the overall data.

On every convolutional layer and there might be more than one for further detailed exactions, there could exist multiple filters that uses same mechanism, where the filter is to overlap the beginning of the data and moved by a certain step (Stride) to make sure it scans the main data matrix then, element wise multiplication is and then elements is summed up where the number produced represent how similarly the is the overlapped data part resemble the feature in the filter being used

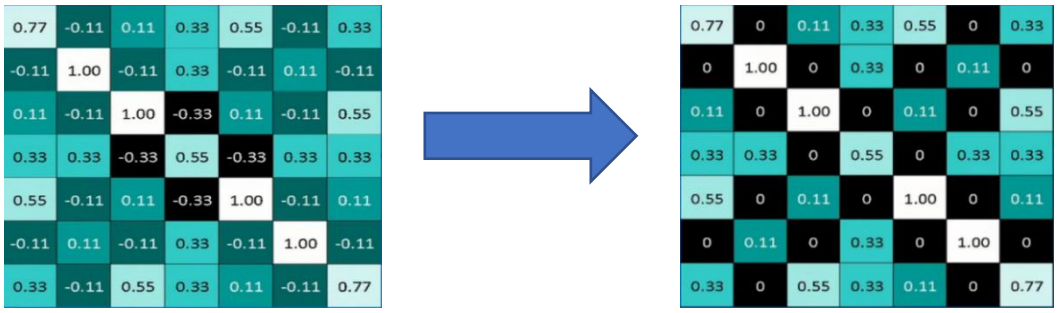


Embedded Layer:

The word embedding layer is used in this study to transform the K-mer sentence into a dense feature vector matrix. It is an improvement over more the traditional bag-of-word model encoding schemes where large sparse vectors were used to represent each word or to score each word within a vector to represent an entire vocabulary. These representations were sparse because the vocabularies were vast and a given word or document would be represented by a large vector comprised mostly of zero values. Instead, in an embedding, words are represented by dense vectors where a vector represents the projection of the word into a continuous vector space. The position of a word within the vector space is learned from text and is based on the words that surround the word when it is used. The position of a word in the learned vector space is referred to as its embedding. The label encoding and K-mer techniques are used to encrypt the DNA sequence, which preserves the position information of each nucleotide in the sequence. The embedding layers is used to embed the data from the above two techniques.

Rectifier Linear Unit

In the Rectified Linear Unit, it takes the input matrix and sweep all the numbers in it and change the negative values to zero, so the output of the RELU is similar to the input matrix but with no negative numbers.

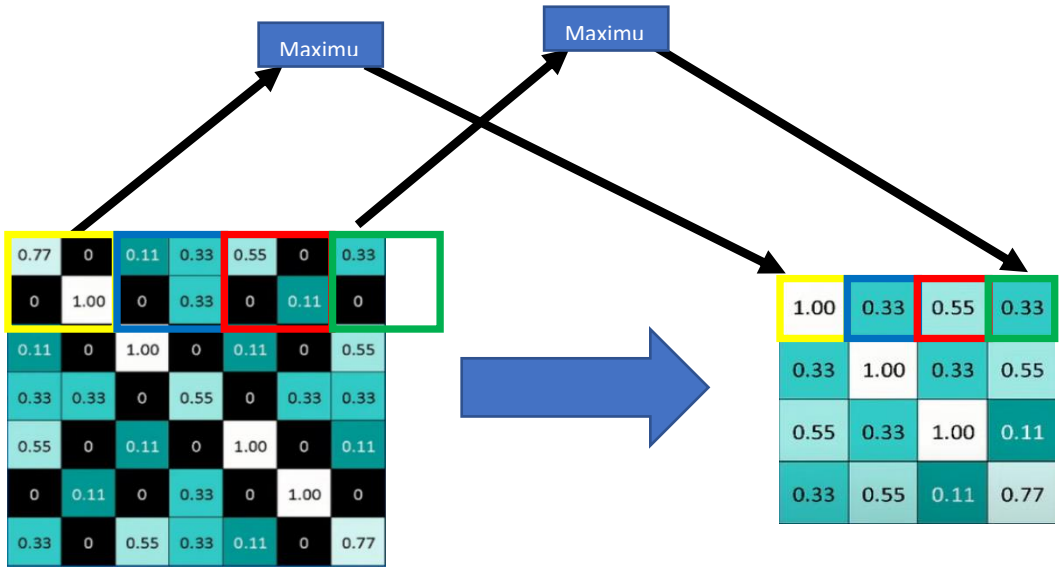


Max Pooling

In this stage we will have reduce the size of the matrix layer using the following steps

1. Pick a window size (usually 2 or 3)
2. Pick a stride (usually 2)
3. Walk the window across the matrix
4. From each window, take the maximum value

If the window exceeds the end of the matrix, we will put zeros as padding in these areas.



Classification

First, we take all the output matrices and flattens them into one array.
When we feed the system decision 1, there are values in the array tends to be high that's mean they predict very strong that is decision 1 they got a lot of vote for decision 1 outcome.