M.Tech Seminar

HARDWARE IMPLEMENTATION OF IMAGE CLASSIFIER MODEL

Pradumn Kumar
[213070055]

Supervisor : Prof Sachin B. Patkar

Department of Electrical Engineering
Indian Institute of Technology, Bombay

- As we are moving towards the era of AI, there a need of ML model on architecture.
- Due to the scope of inherent parallelism in the application-specific hardware (e.g., FPGA, ASIC, etc.).
- There has been growing interest in exploring hardware based solutions to solve some of the fundamental problems associated with ML techniques.
- As the size of datasets and complexity of machine learning models grows, the need for faster computation and lower power consumption becomes extremely critical.

- They receive one or more input signals.
- They perform some calculations and apply Non Linear Transformation.
- They send some output signals to neurons deeper in the neural net through a synapse.

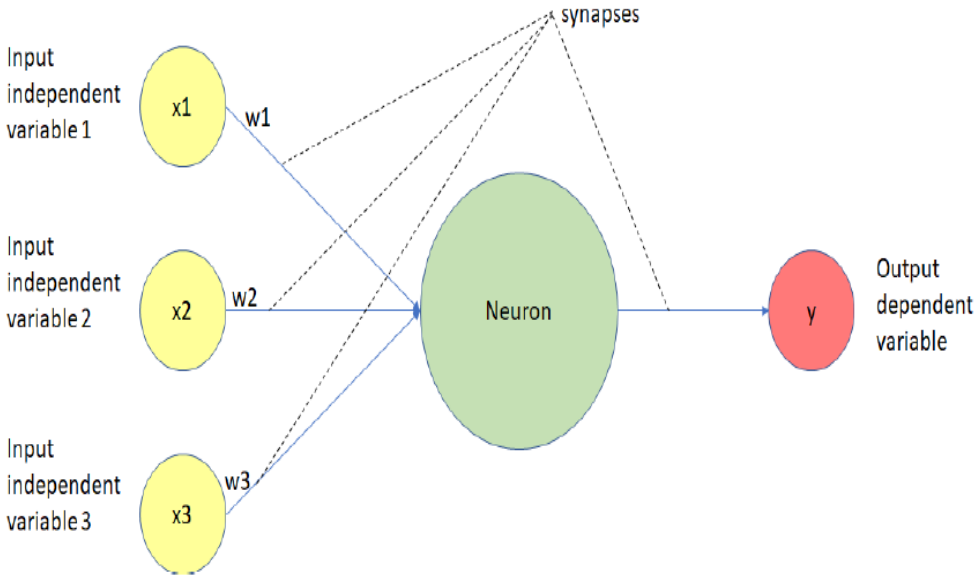


Fig. Input & output of Neuron

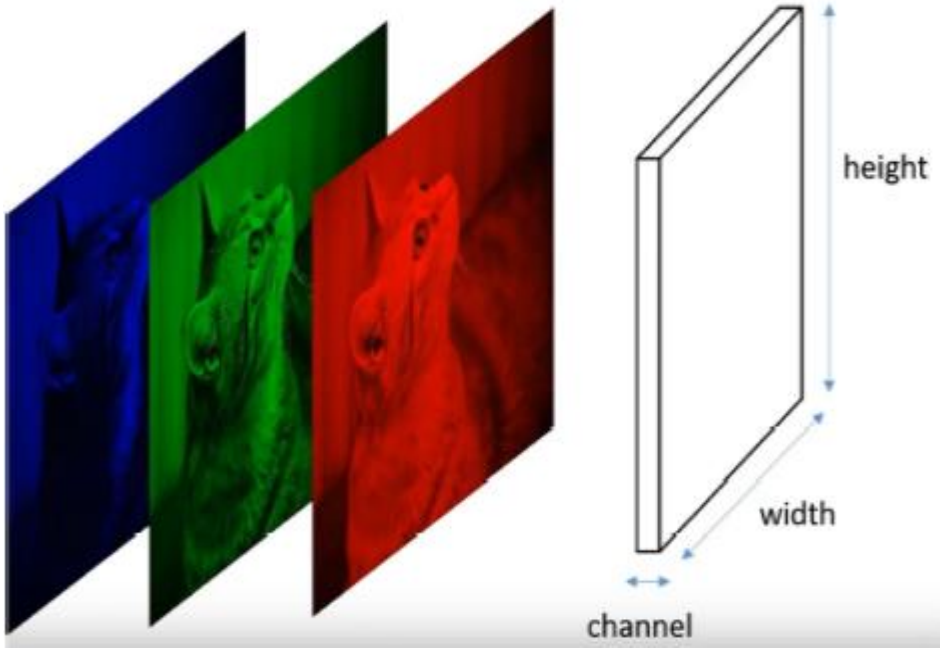
[1] B. Hershberg, “Ring amplifiers for switched capacitor circuits”

- Connections between the nodes don't form a loop.
- Information moves in only one direction that is forward.
- f(.) is a non linear activation function in the case of classification and is the identity in case of regression.

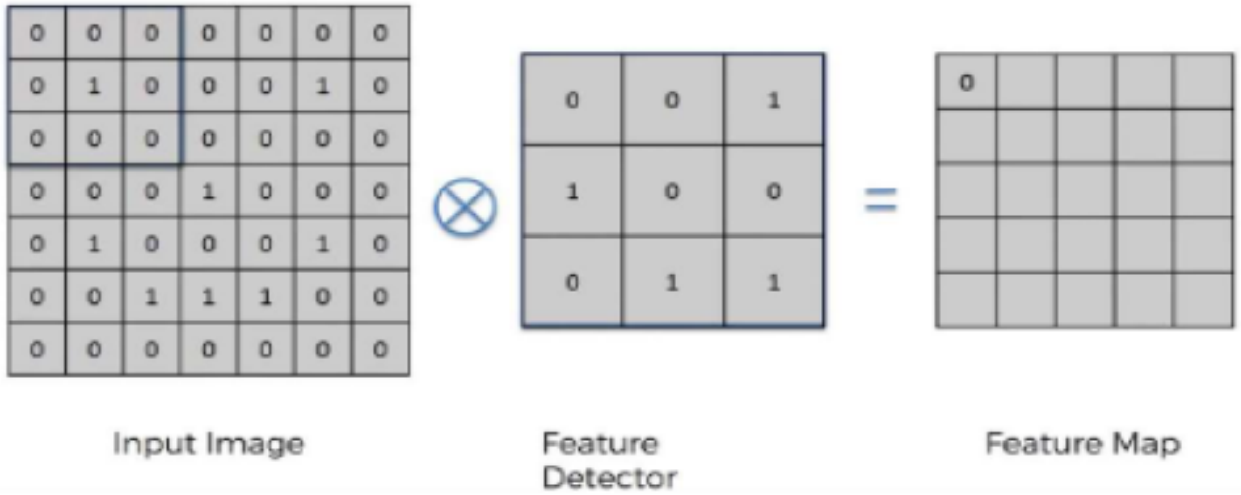
$$y(X,W) = f\left(\sum_{j=1}^{M} w_j \emptyset_j(X)\right) where\ X = input\ vector\ , W = weight\ vector$$

$$y_k(X,W) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)}\, h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

- 3 channels namely Red, Green and Blue will show 3 matrices.

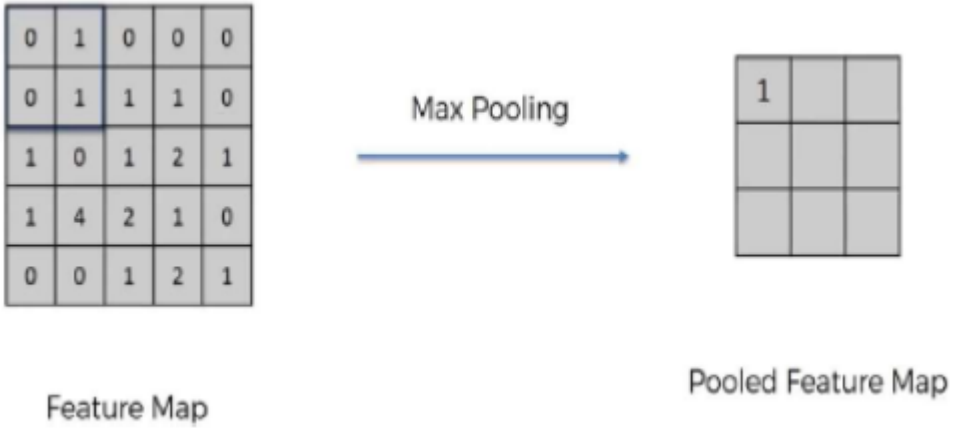


- Features are localised in some dataset.
- Computationally less expensive.
- Predicts more accurately.



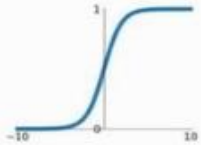
- Feature extracting using Feature Detector.
- Output of convolution operation goes to Pooling Layer.

- Some data from convolution layer is still redundant.
- So we either take average value or max value.



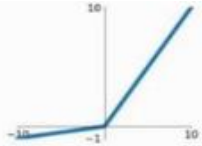
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



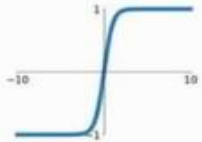
Leaky ReLU

$$\max(0.1x, x)$$



tanh

$$\tanh(x)$$



Maxout

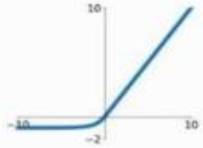
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ReLU

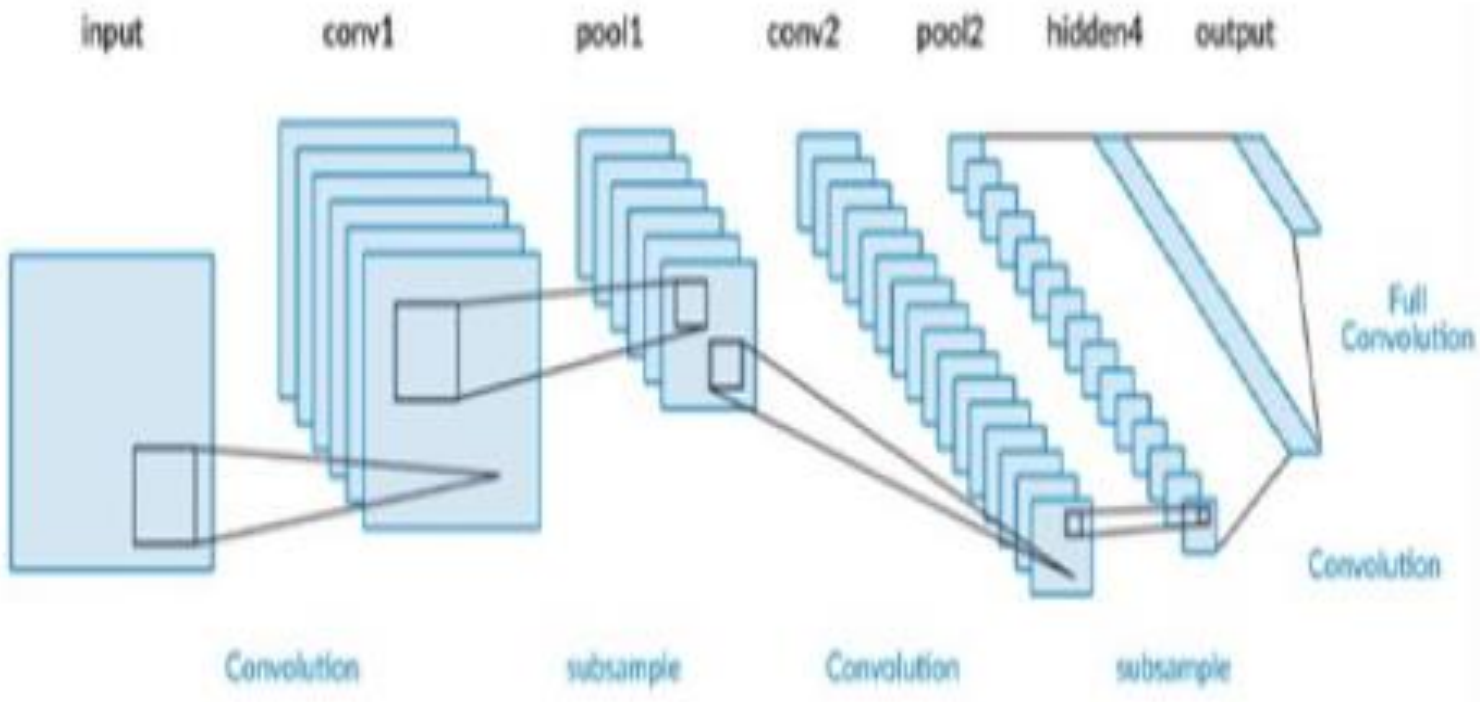
$$\max(0, x)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



- Provides non linearity.
- ReLu for classification is better choice.


input
conv1
pool1
conv2
pool2
hidden4
output
Full Convolution
Convolution
Convolution
subsample
Convolution
subsample

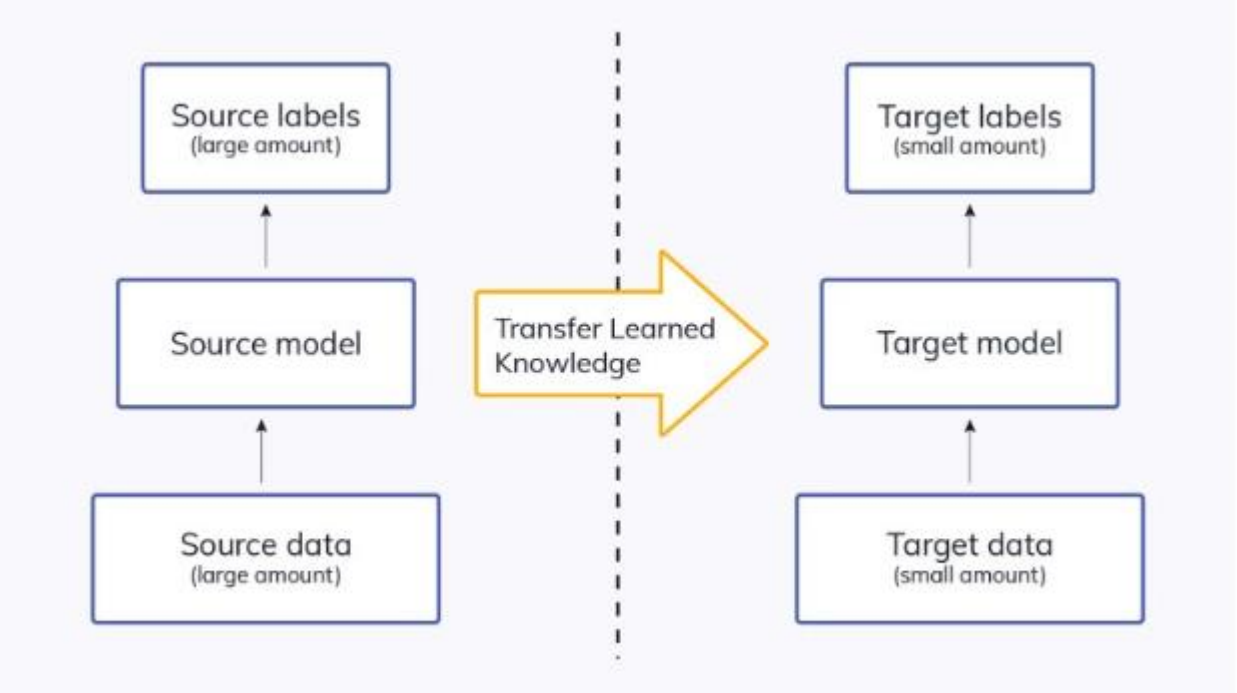
- An ASCII Text file that specifies the initial content of a memory block (RAM or ROM).

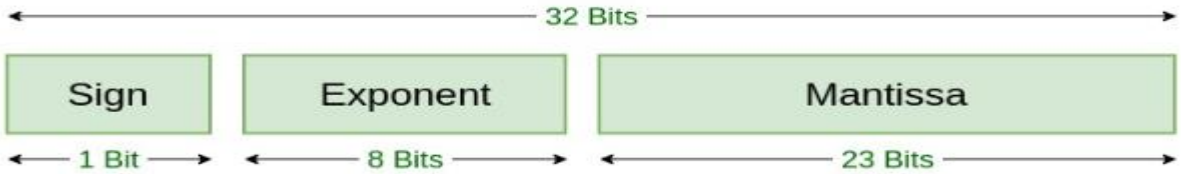
ip1 - Notepad

File Edit View

```
WIDTH=32;
DEPTH=65536;
ADDRESS_RADIX=HEX;
DATA_RADIX=HEX;
CONTENT BEGIN
        00000 : 00000000;
        00001 : 423543CA;
        00002 : 00000000;
        00003 : 00000000;
        00004 : 42506F65;
        00005 : 00000000;
        00006 : 00000000;
        00007 : 43569D75;
        00008 : 4407F1F0;
        00009 : 00000000;
        0000A : 00000000;
        0000B : 00000000;
        0000C : 00000000;
        0000D : 00000000;
        0000E : 4313267E;
        0000F : 4382D090;
        00010 : 42B74FC3;
```

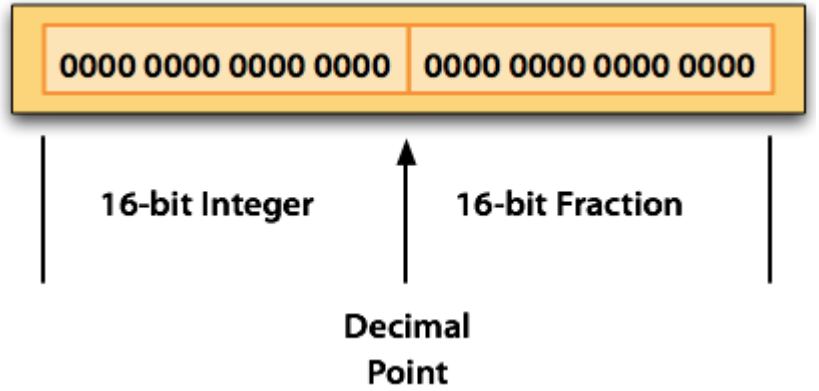
- Usually this technique is used when we are short of dataset.
- We use a PreTrained model for our task.
- Transfer learning is computationally efficient
- This PreTrained model is fine-tuned with available dataset.
- It achieve optimal performance faster than the traditional ML models.


Source labels
(large amount)
Source model
Source data
(large amount)
Transfer Learned
Knowledge
Target labels
(small amount)
Target model
Target data
(small amount)

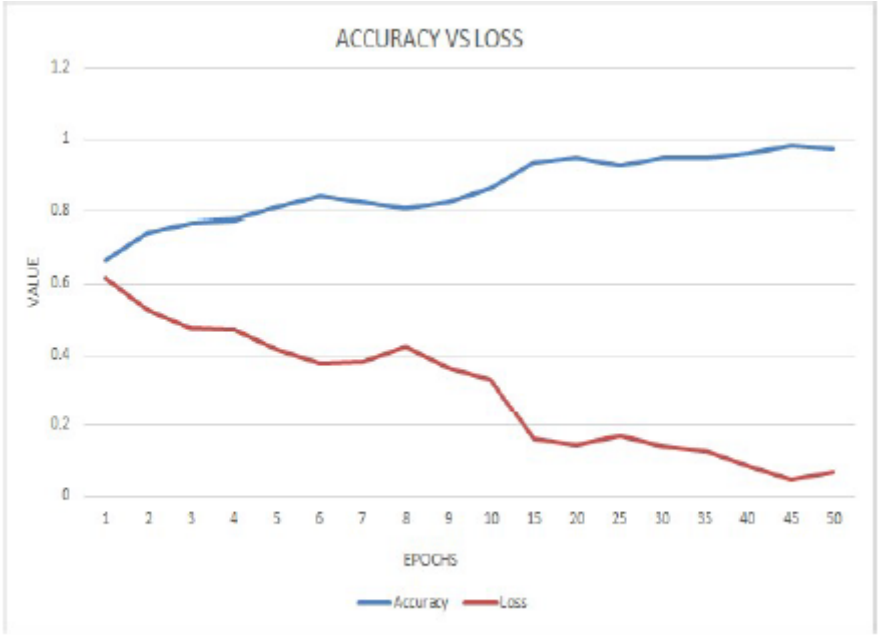


Single Precision
IEEE 754 Floating-Point Standard

Vs.



- Training was done on OpenSource dataset of 23000 Images.



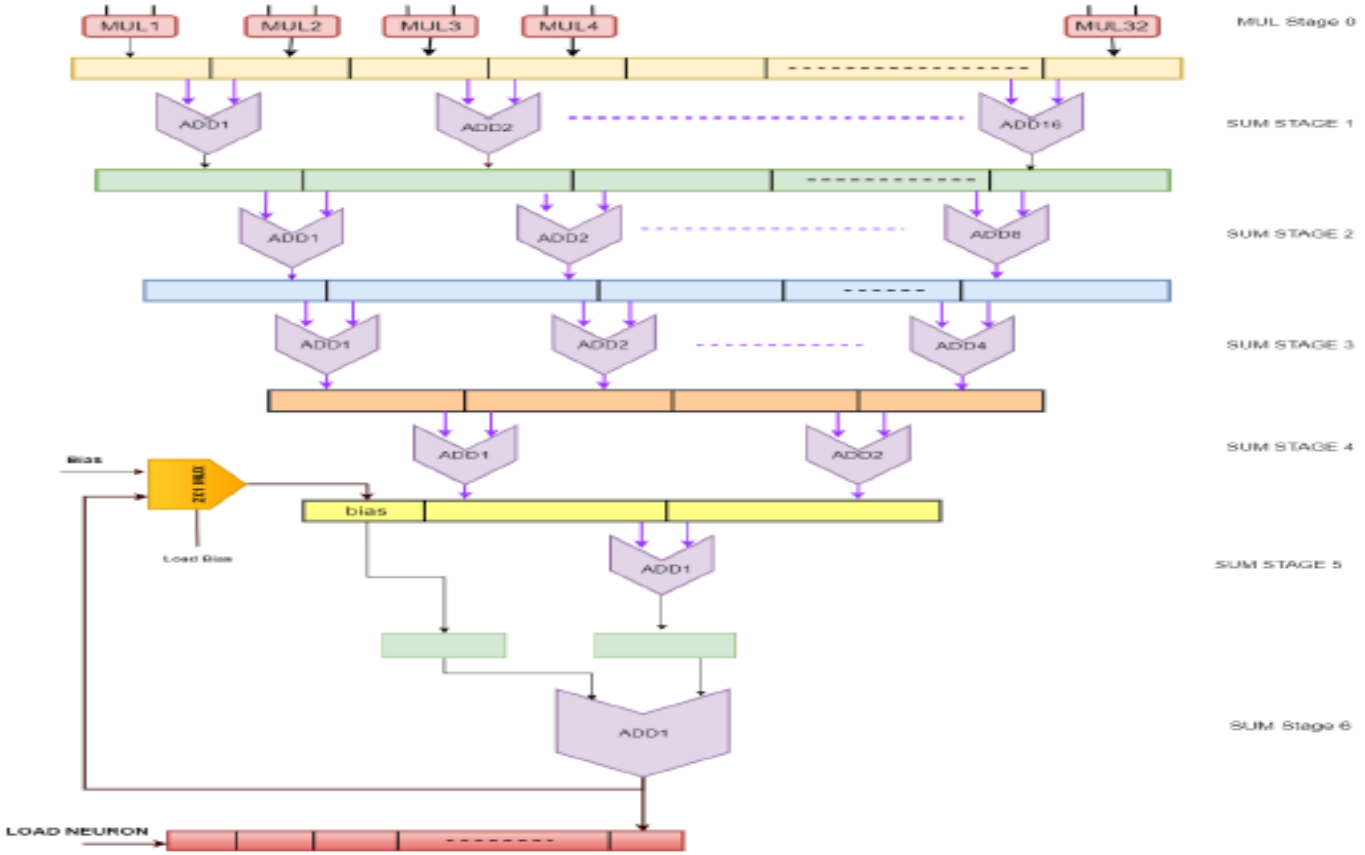
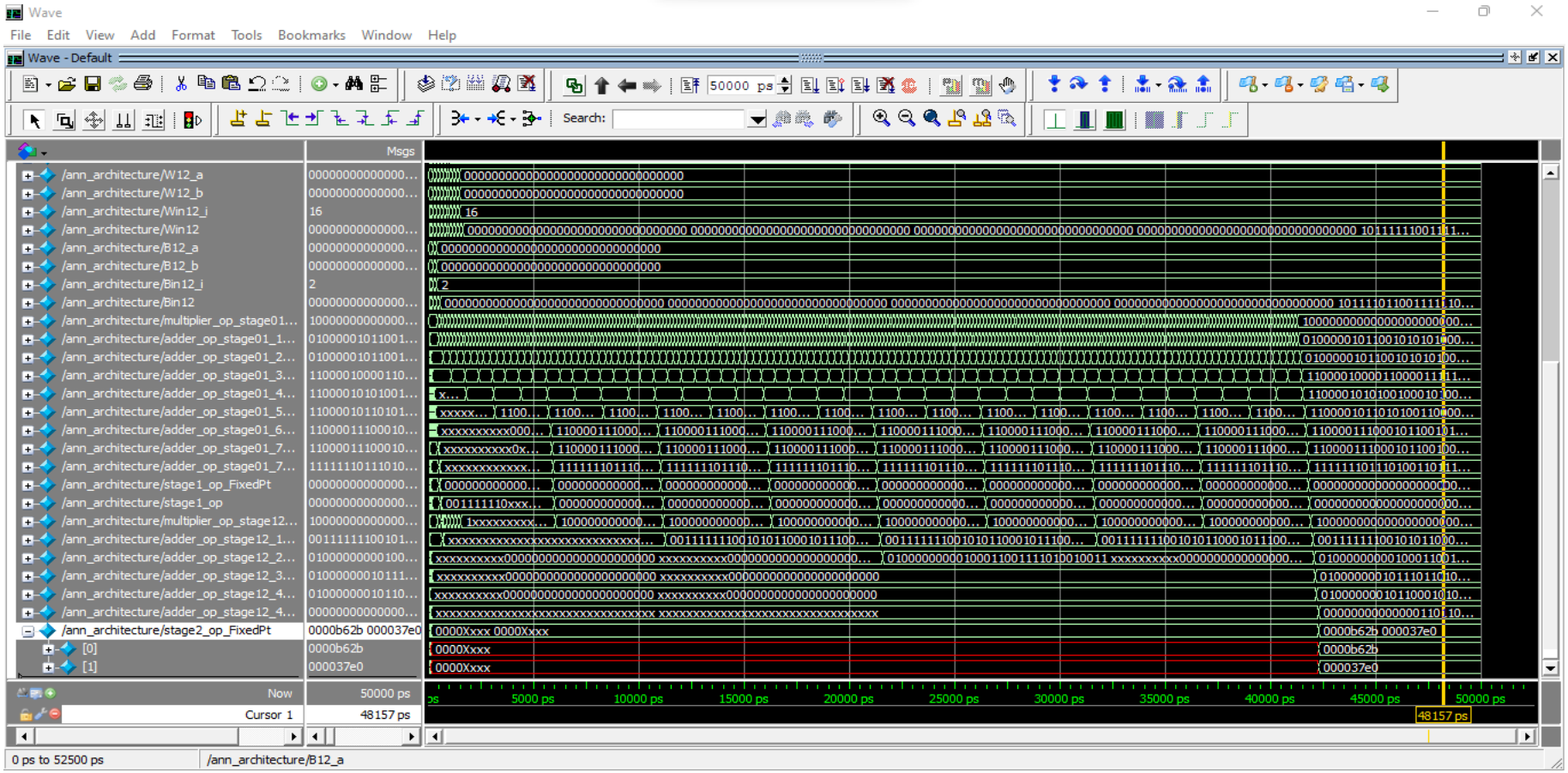


Figure 1: Multiplier & Accumulator Data Path Block Diagram

- It includes Multiplier and Adder implemented in pipeline.
- Addition is implemented through various layer.
- Convertor are added before and after each Neuron

ModelSim Simulation



- 70% accuracy is achieved.

Image No.	Simulation Value	True Value	Remarks
1	Cat	Cat	Correct Result
2	Dogs	Dogs	Correct Result
3	Dogs	Cats	InCorrect Result
4	Dogs	Dogs	Correct Result
5	Dogs	Dogs	Correct Result
6	Dogs	Dogs	Correct Result
7	Cats	Dogs	InCorrect Result
8	Dogs	Dogs	Correct Result
9	Dogs	Cats	InCorrect Result
10	Dogs	Dogs	Correct Result

- Cyclone IV has resource constrained.
- A more sequential implementation is possible.
- There a scope of CNN along with ANN in Verilog.
- Accuracy can be improved with more fine-tuning.

References

- [1] https://github.com/sudhamshu091/32-Verilog-Mini-Projects/tree/main/Floating
- [2] https://keras.io/gettingstarted/faq/#how-can-i-obtain-the-output-of-anintermediate-layer
- [3] https://www.intel.com/content/www/us/en/programmable/quartushelp/13.0
- [4] https://gist.github.com/Towdium/1a2fad63dd3665c064df48b39b41ab01
- [5]https://www.researchgate.net/publication/265179275_Implementation_of_Back_Propagation_Algorithm_in_Verilog

- [6] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6024574
- [7] https://yycho0108.github.io/CompArchNeuralNet/

Thank You