6-28-2019

State of the Art of Deep Learning Technology and its Next Generation Architecture

Dr. Kuo-Kun Tseng Associate Professor

State of the Art of Deep Learning Technology and its Next Generation Architecture

Presented by : Dr. Kuo-Kun Tseng

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- State of the Art of Deep Learning Technology
  - Our Applications
  - Other New Applications
- Next Generation Architecture for Deep Learning Technology
  - Demand for New Architecture
  - Deep Learning with FPGA Architecture
  - Object Tracking Example
  - OpenCL on FPGA
  - Translation Tool
- Conclusion

About the Speaker

- Kuo-Kun Tseng(email:kktseng@hit.edu.cn), he was born in 1974, and received his doctoral degree in computer information and engineering from National Chiao Tung University of Taiwan in 2008.
- He is currently **a tenure associate Professor at School of Computer Science and Technology in Harbin Institute of Technology (Shenzhen Campus)**, and received **Shenzhen Peacock talent award (B level)**.
- Before he joined HITSZ, he worked as senior software and IC design engineer in the USA and Taiwan for many years. Since 2004, he is working on the research of intelligent algorithm and architecture.
- Furthermore, he has more than 20 research projects, 30 patents and been published more than 80 research articles, of which about half papers are published at SCI/ACM / IEEE journal with high reputation and impact factor. Last but not least, he is an associate editor for Enterprise Information System and International Journal of Engineering Business Management, and the reviewer of many distinguished journals, such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transaction of Internet of Things, IEEE Access, IEEE Sensor, Expert System, Neural Computing and so on.

About Harbin Institute of Technology

- Harbin Institute of Technology (HIT) is a member of top nine University Union (C9) in China with three Campus: Harbin , Weihai, Shenzhen.
- Undergraduate Entrance Examination Score No. 1 among Guangdong Province's Colleges, and 2019 QS Ranking is No. 9.

上榜 2019 QS 世界大学排名（中国大陆高校）			
序号	2019 排名	2018 排名	大学
1	17	25	清华大学
2	30	38=	北京大学
3	44	40	复旦大学
4	59	62	上海交通大学
5	68	87	浙江大学
6	98	97	中国科学技术大学
7	122	114=	南京大学
8	257	282	武汉大学
9	285	325=	哈尔滨工业大学

Harbin Institute of Technology

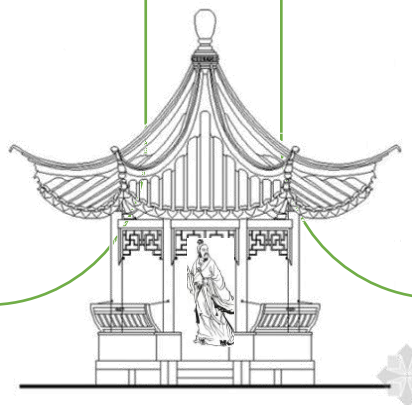
About Shenzhen

- **Developed from "reform and opening-up" policy in 1979.**
- **Actual population to be about 20 million**
- Shenzhen was **one of the fastest-growing cities** in the world
- Has been **ranked second on the list of top 10 cities to visit** in 2019 by Lonely Planet.
- The city is a leading global technology hub, dubbed by media as the **next Silicon Valley**.

Our Lab - Intelligent Architecture Lab

Deep Learning Application

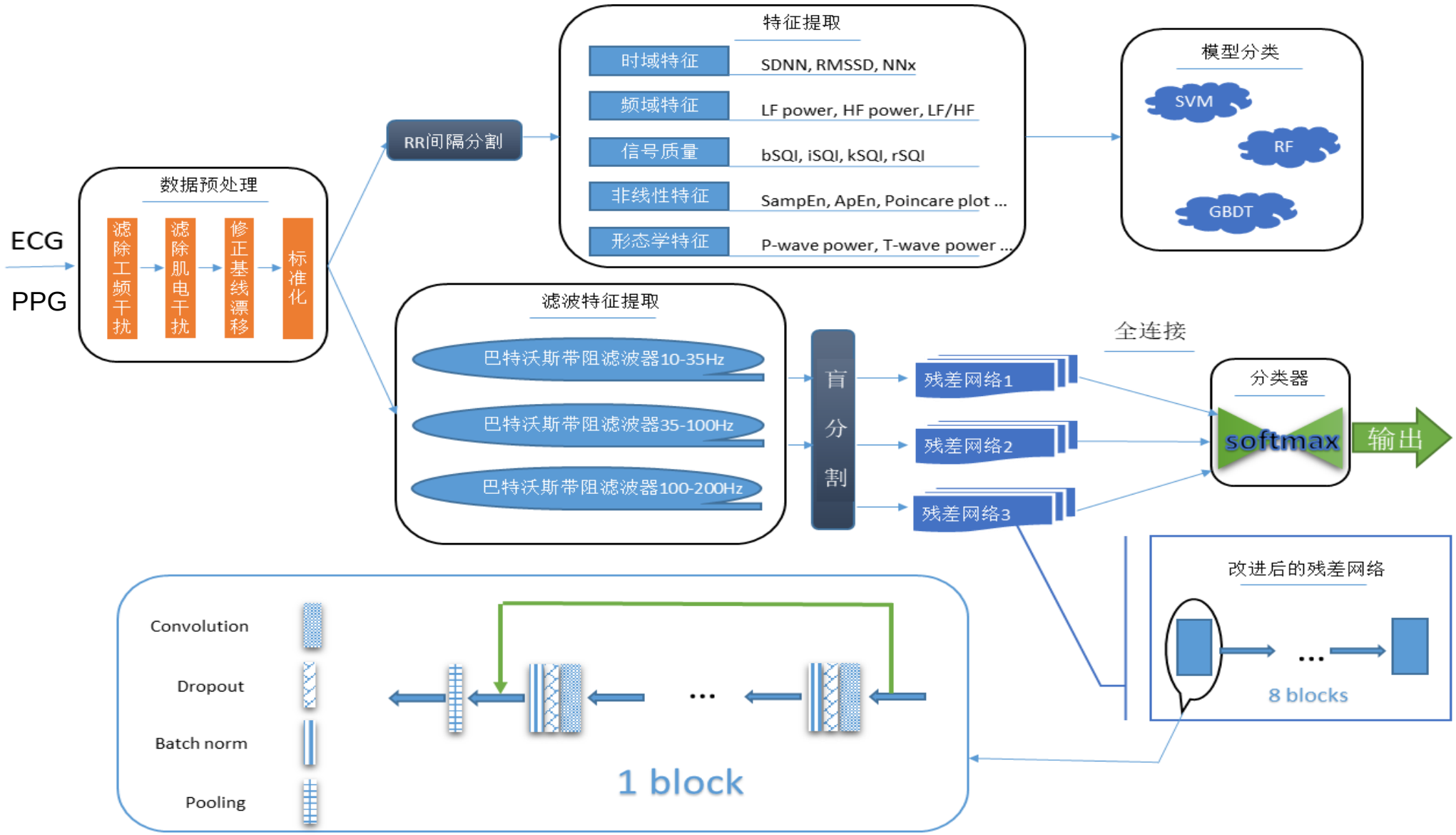
- NLP
  - English to Chinese Translation
  - Specific Domain Q&A Robot
- Signal Processing
  - ECG Abnormal Detection
  - House Price Prediction
- Graphic Processing
  - Image Semantic Segmentation
  - Visual Depth Prediction
  - Medical Image Segmentation



Technology

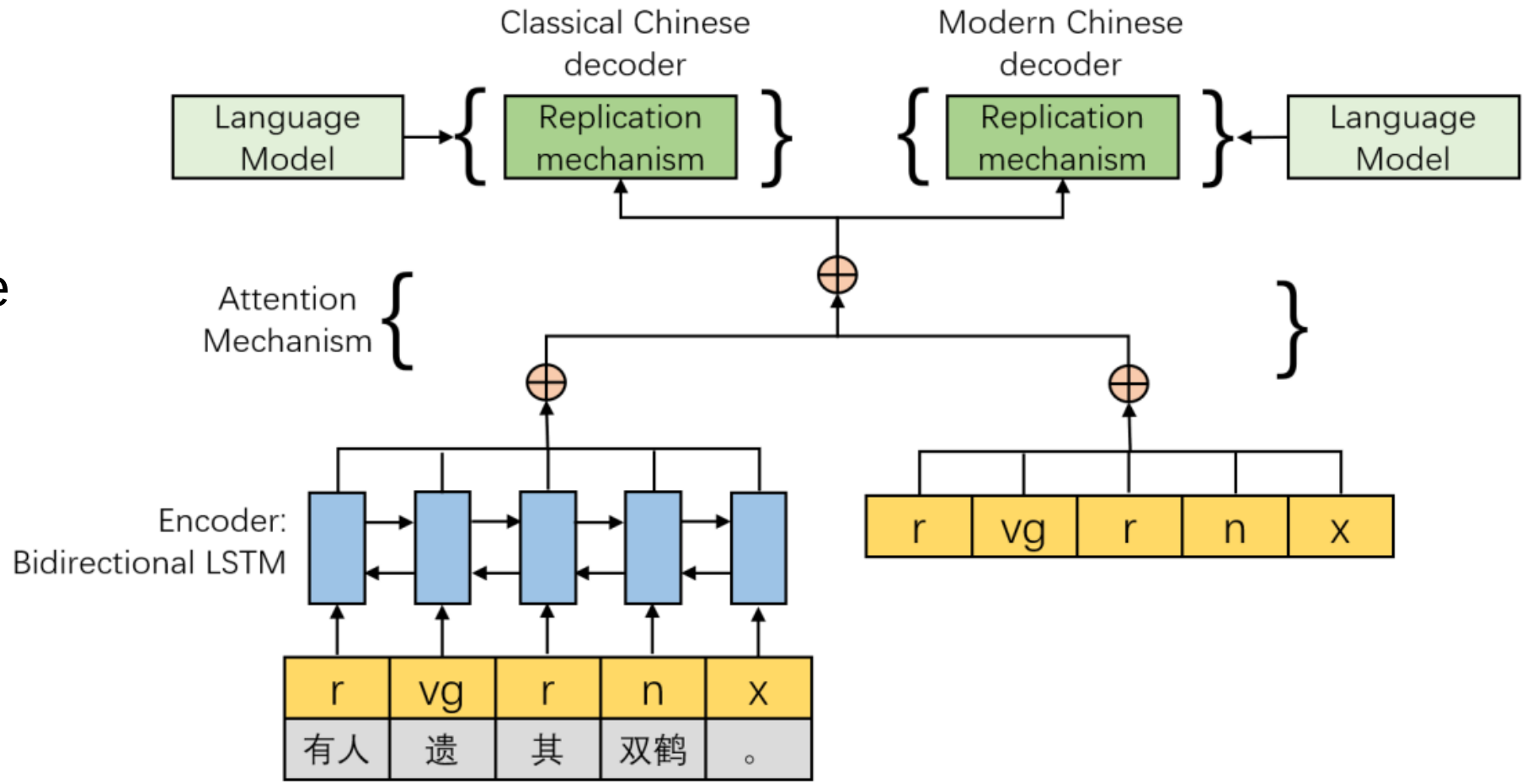
- Algorithm Optimization
  - Design algorithms for deep learning applications.
- Hardware Optimization
  - Based on FPGA and other new hardware, optimize performance for deep learning algorithms.
  - For edge and cloud devices

ECG/PPG Classification

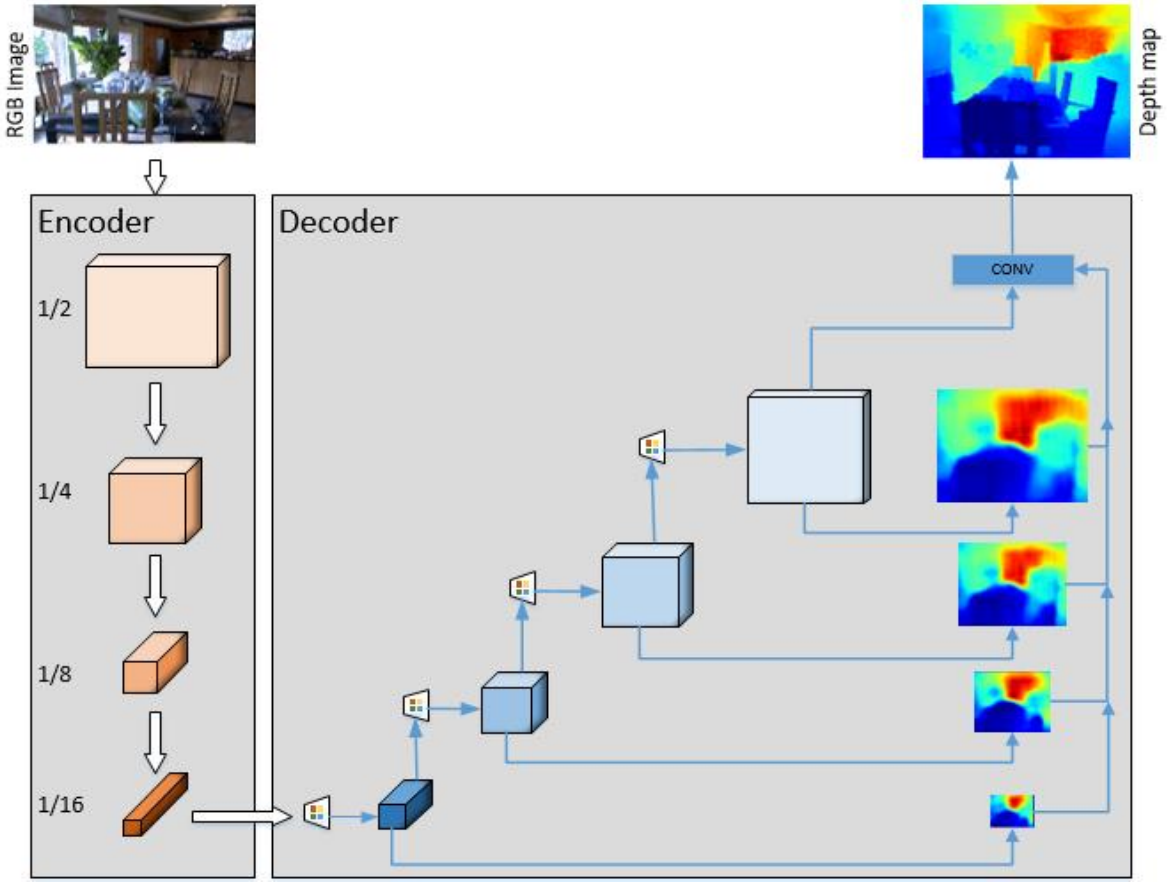


Modern to Classical Chinese Translation

- For Learning Classical Chinese
- Mutual Translation
- Small Training Data

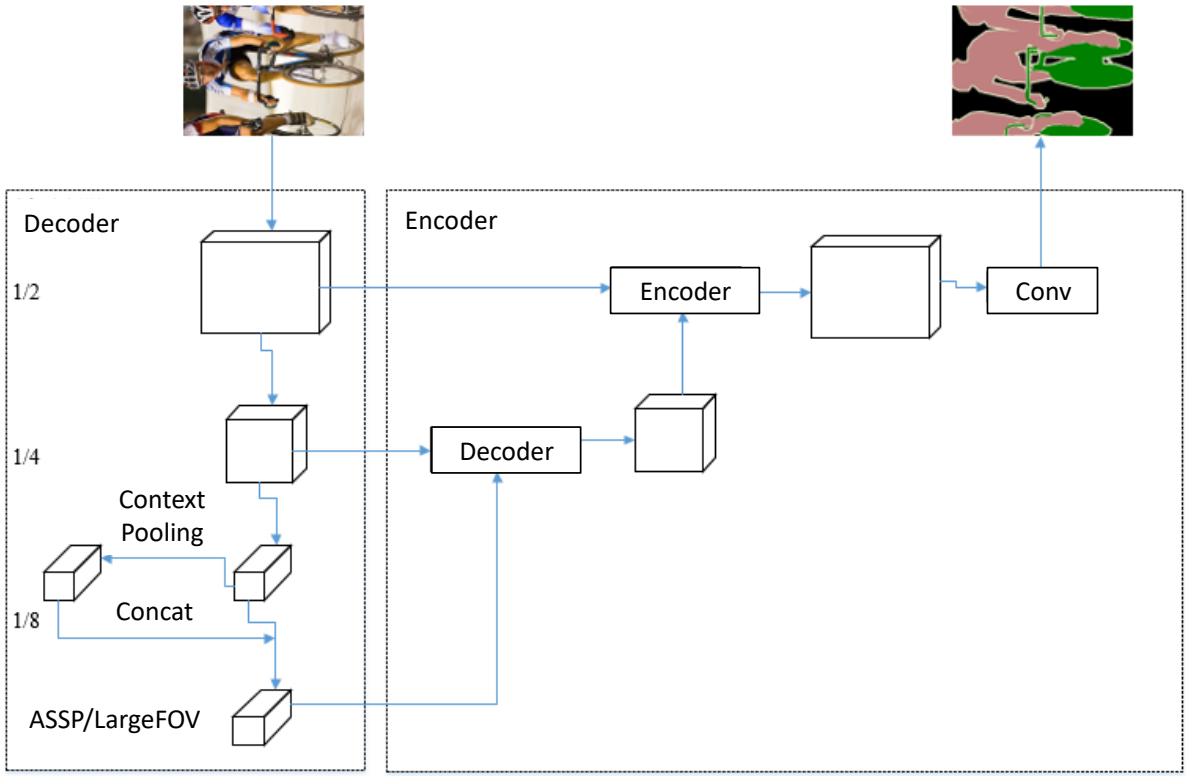


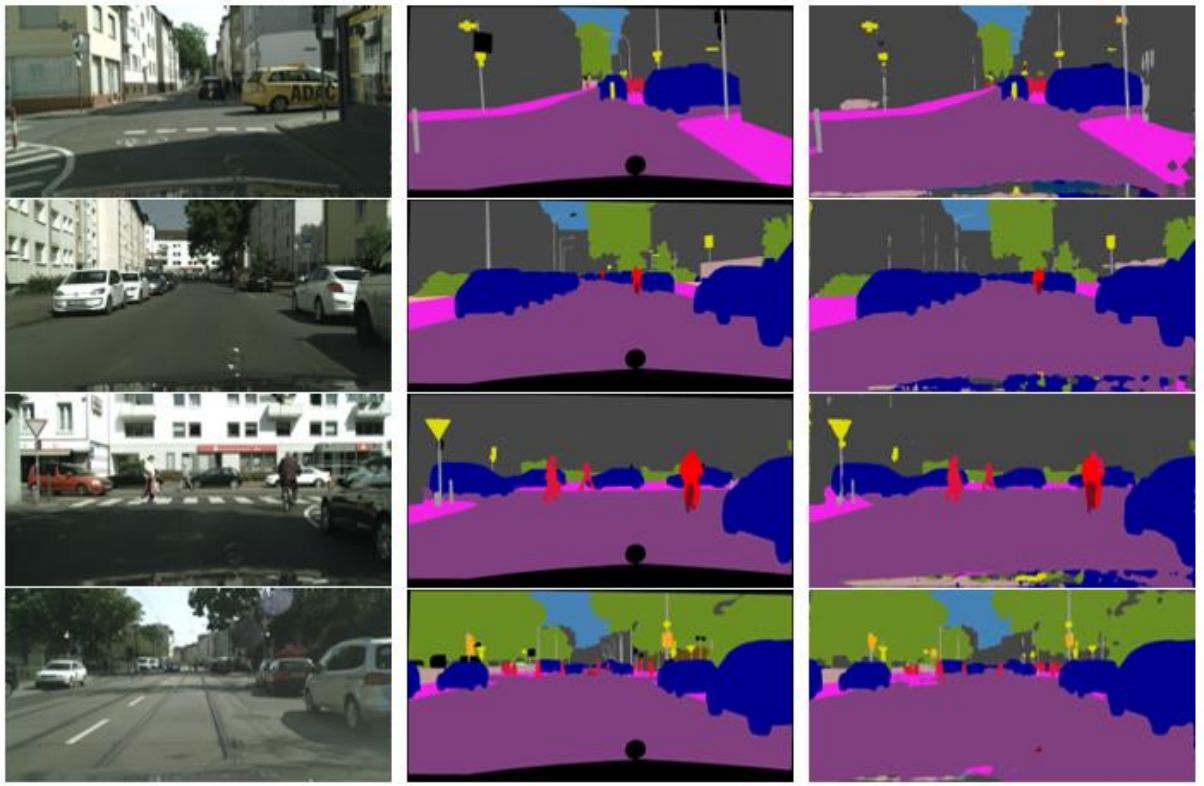
Visual Depth Prediction



Semi-Supervised Learning

Semantic Segmentation

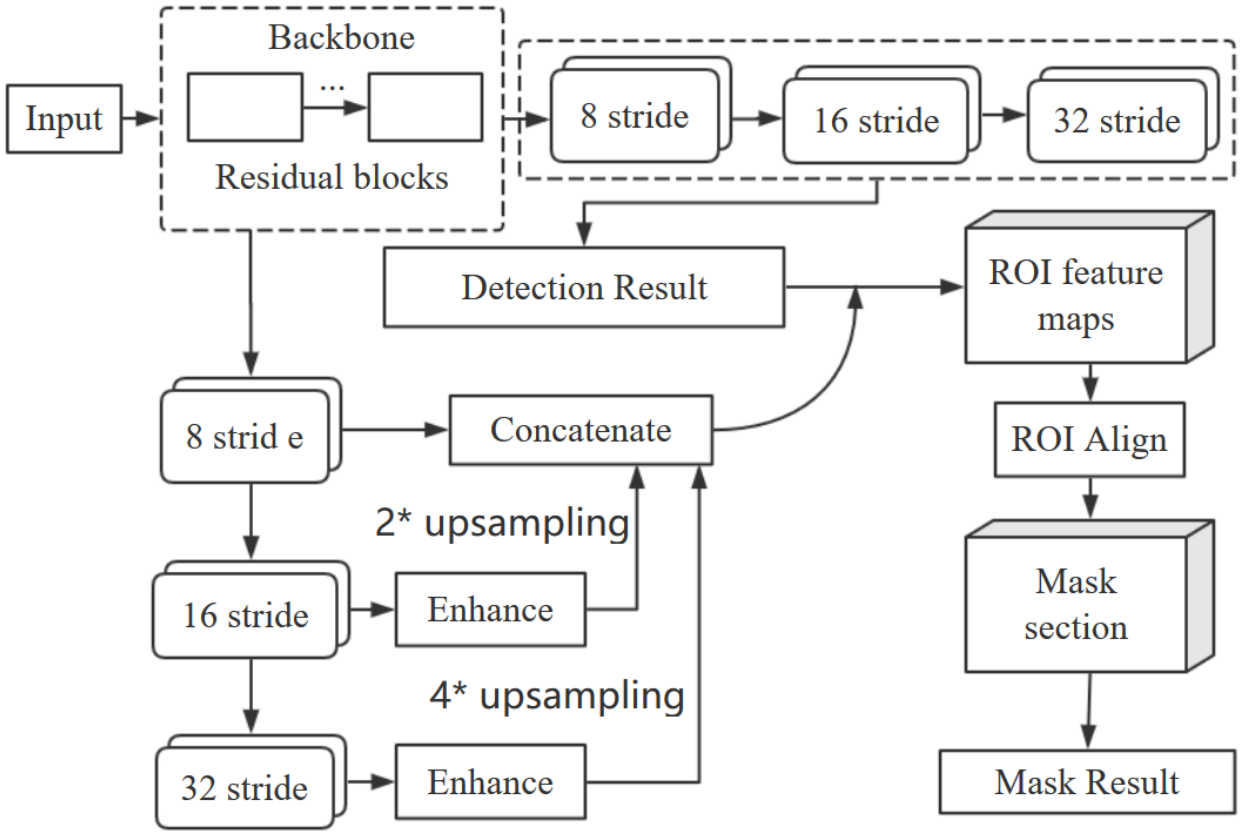




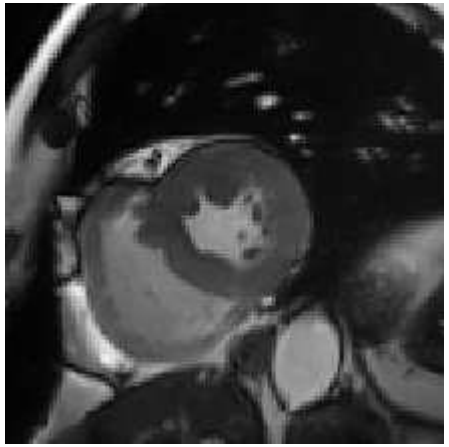
Optimize for accuracy and speed

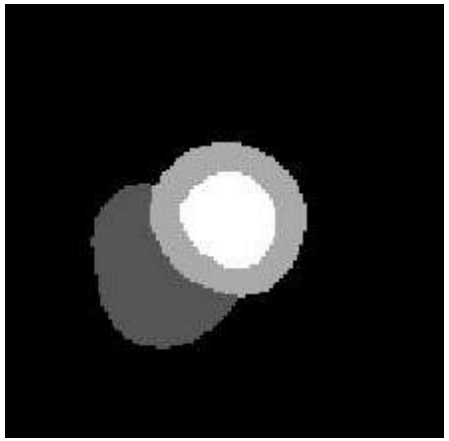
For unmanned driving application

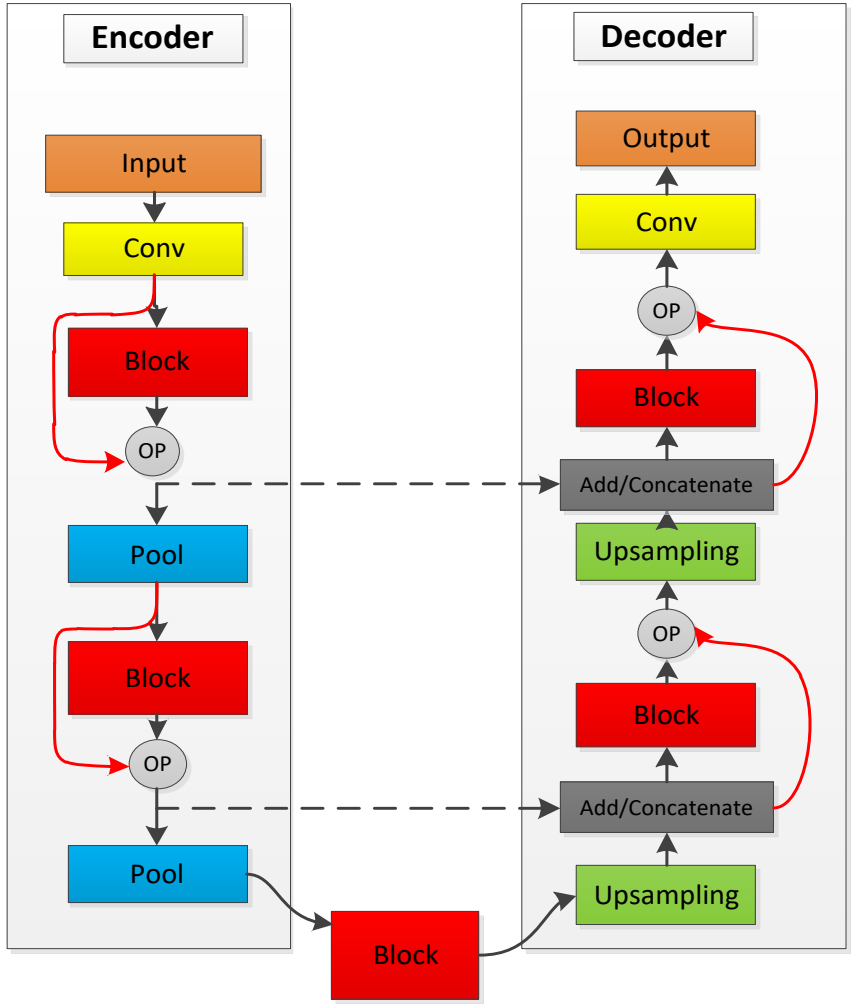
Detection and Segmentation

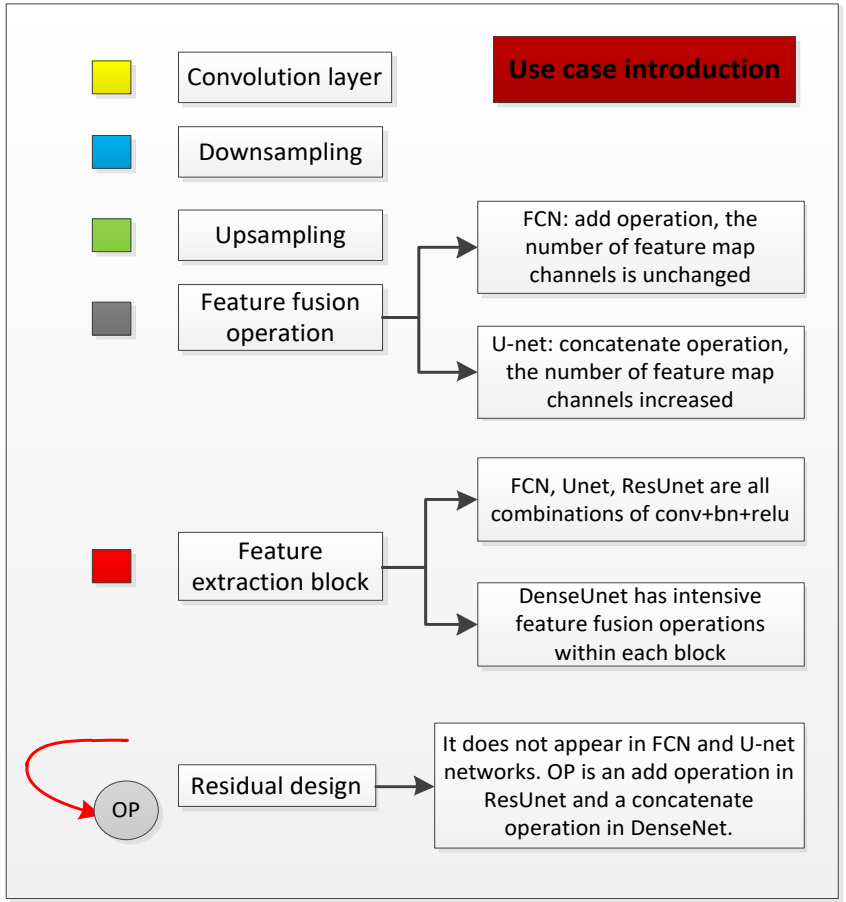


Medical Image Segmentation





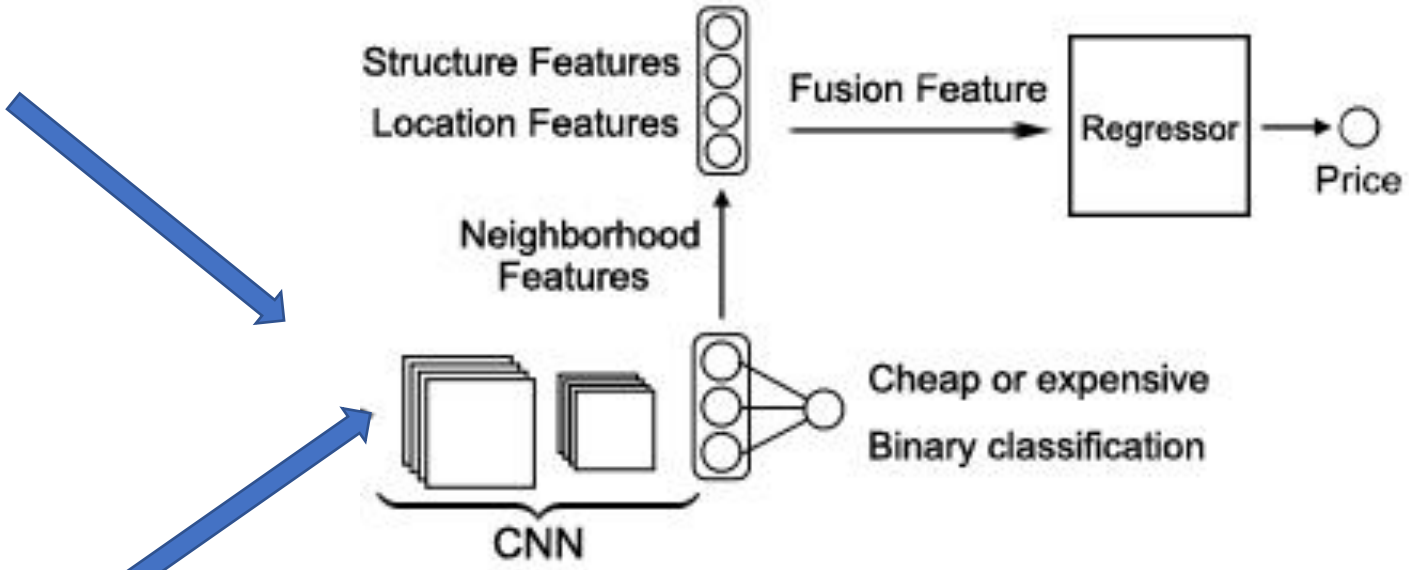




House Price Estimation

Satellite Map

Residential Appearance

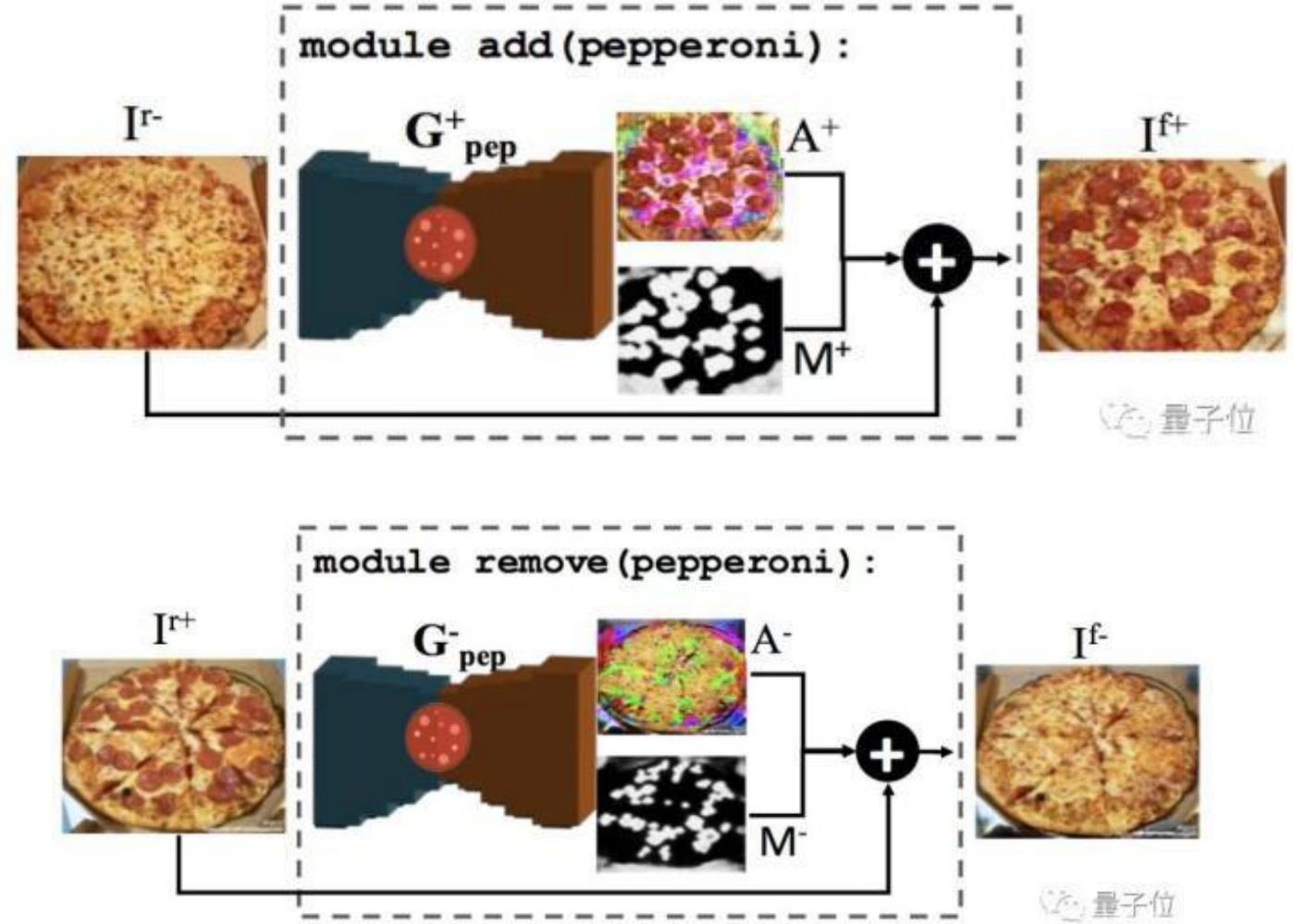


Short Comment

Design deep learning from single task to multi-task network

End to end encoder and decoder model has great applications.

PizzaGAN-Naturally Layered

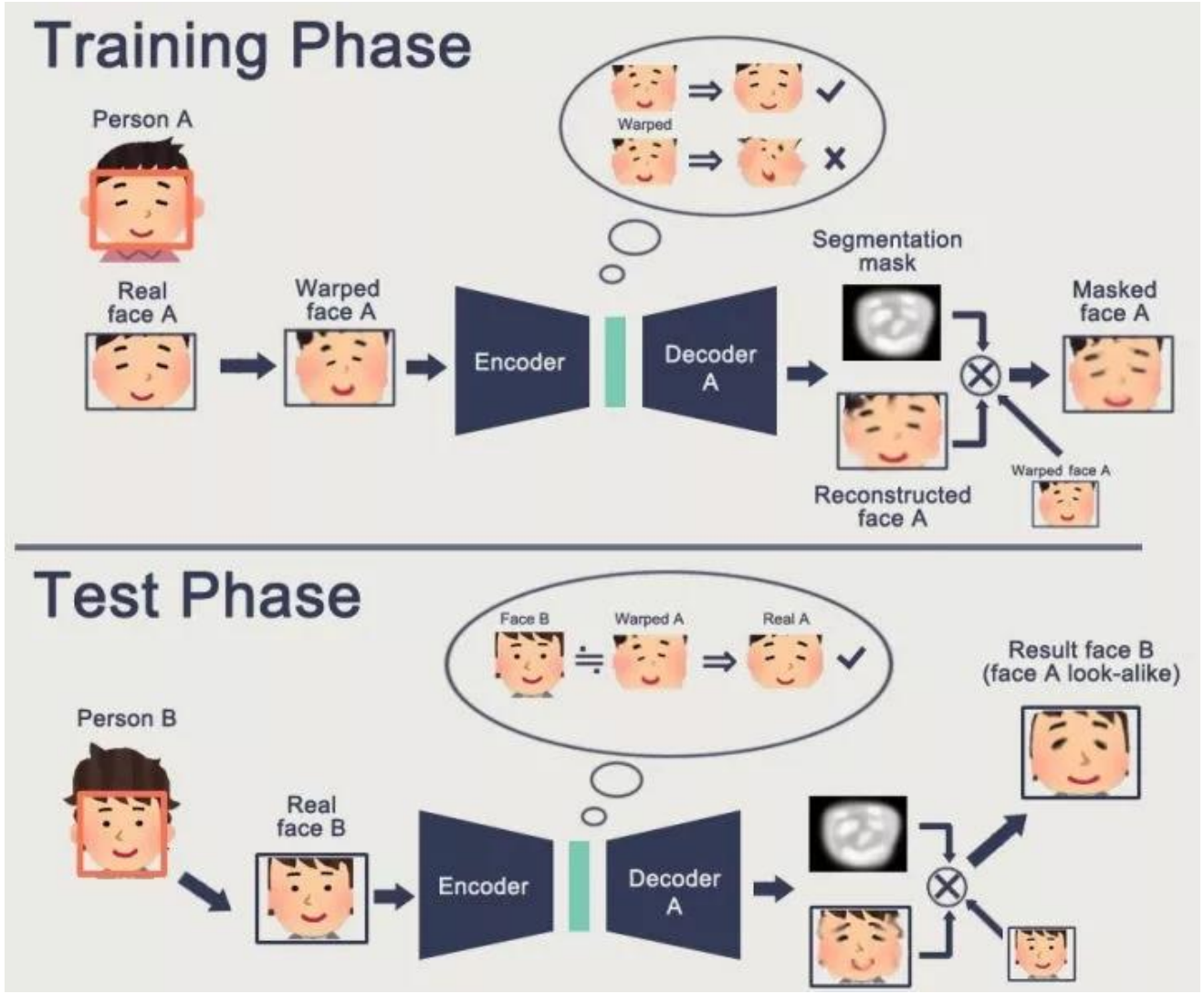


https://arxiv.org/abs/1906.02839

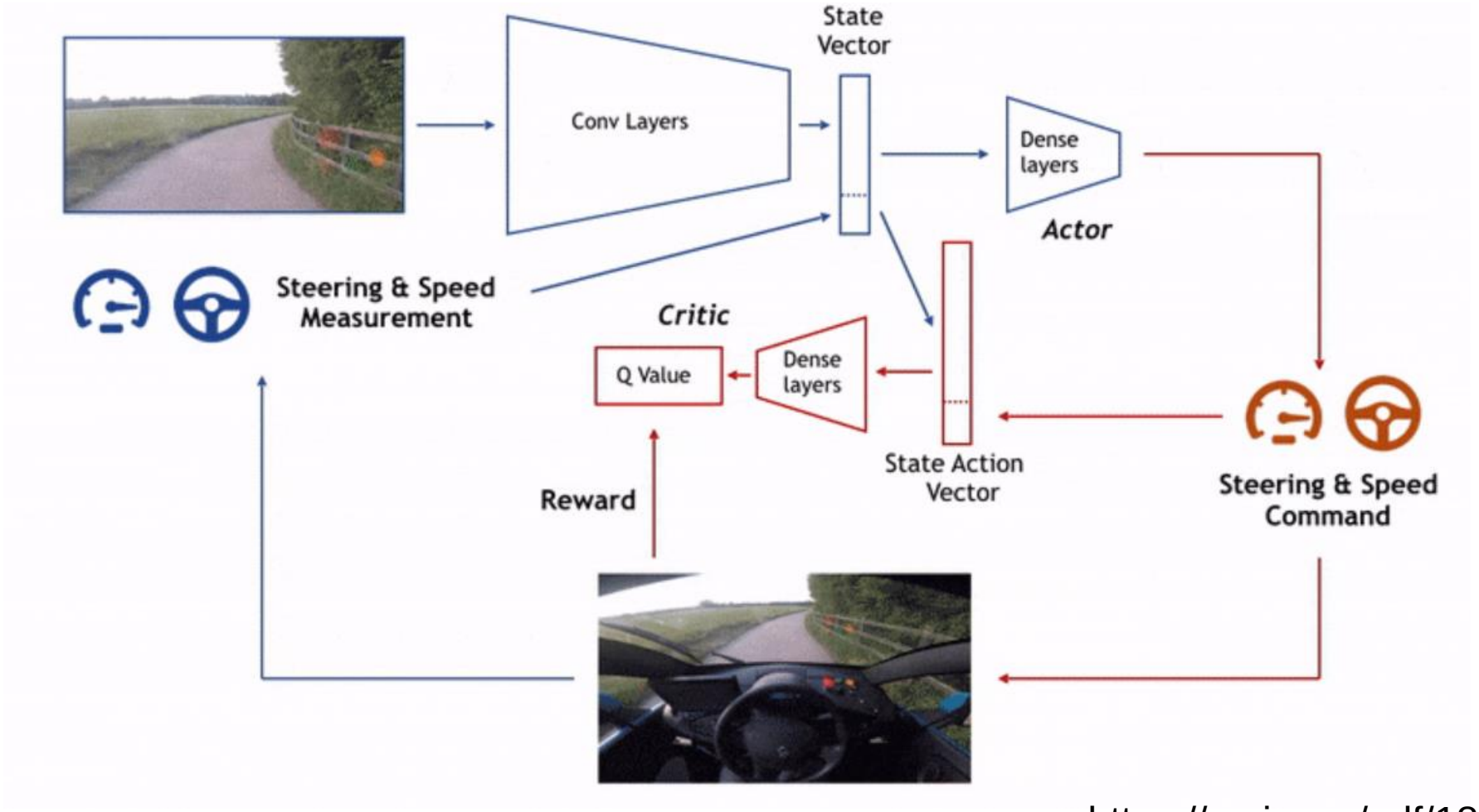
DeepFakes

(Source: shaoanlu/faceswap-GAN)

https://github.com/Fabsqrt/BitTigerLab/tree/master/DeepFake

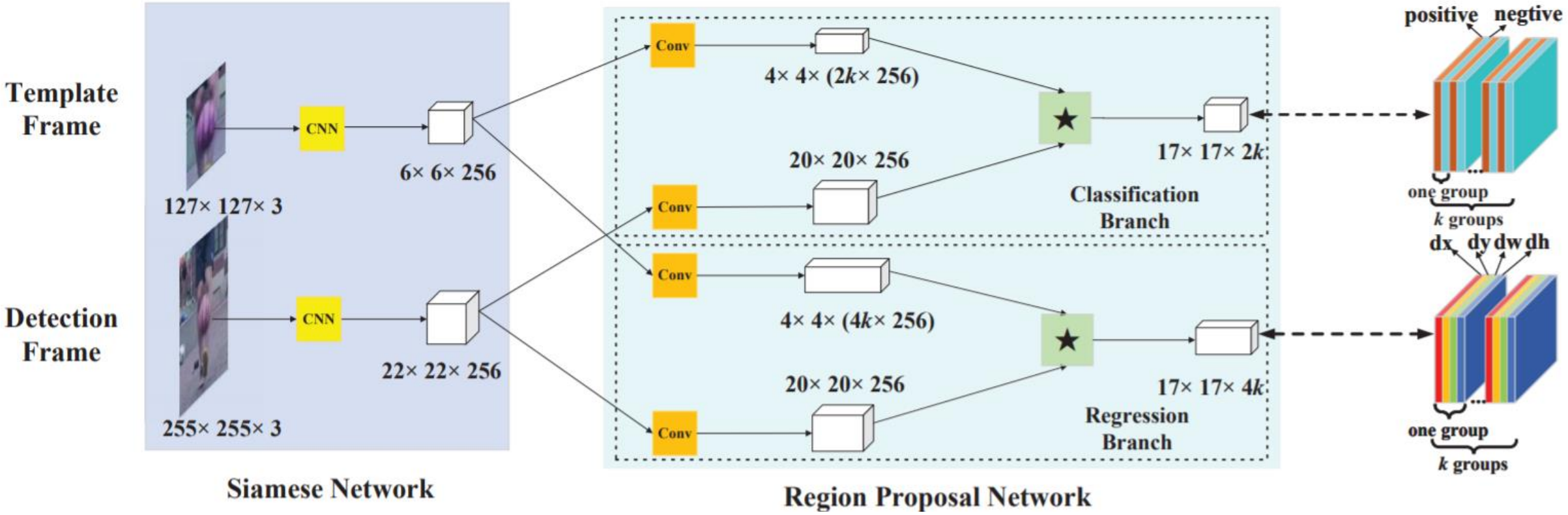


Drive with-Reinforcement-Learning



https://arxiv.org/pdf/1807.00412.pdf

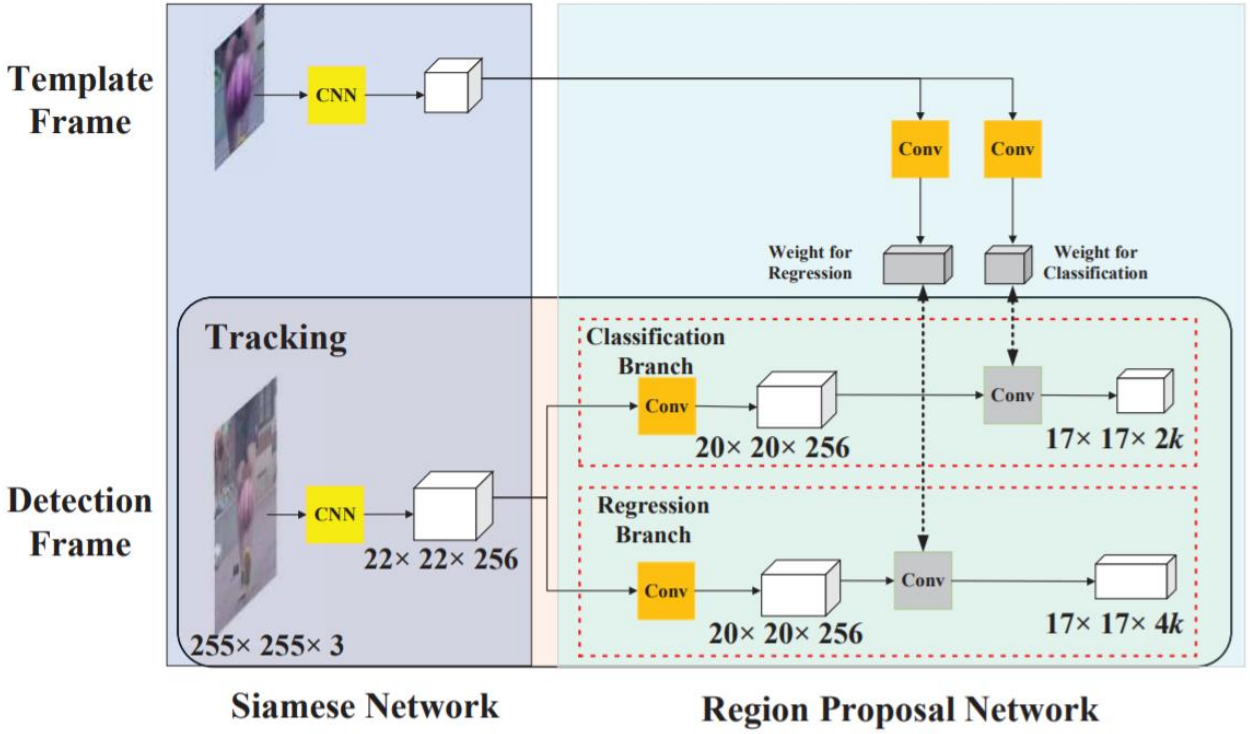
Siamese-RPN - Object Tracking



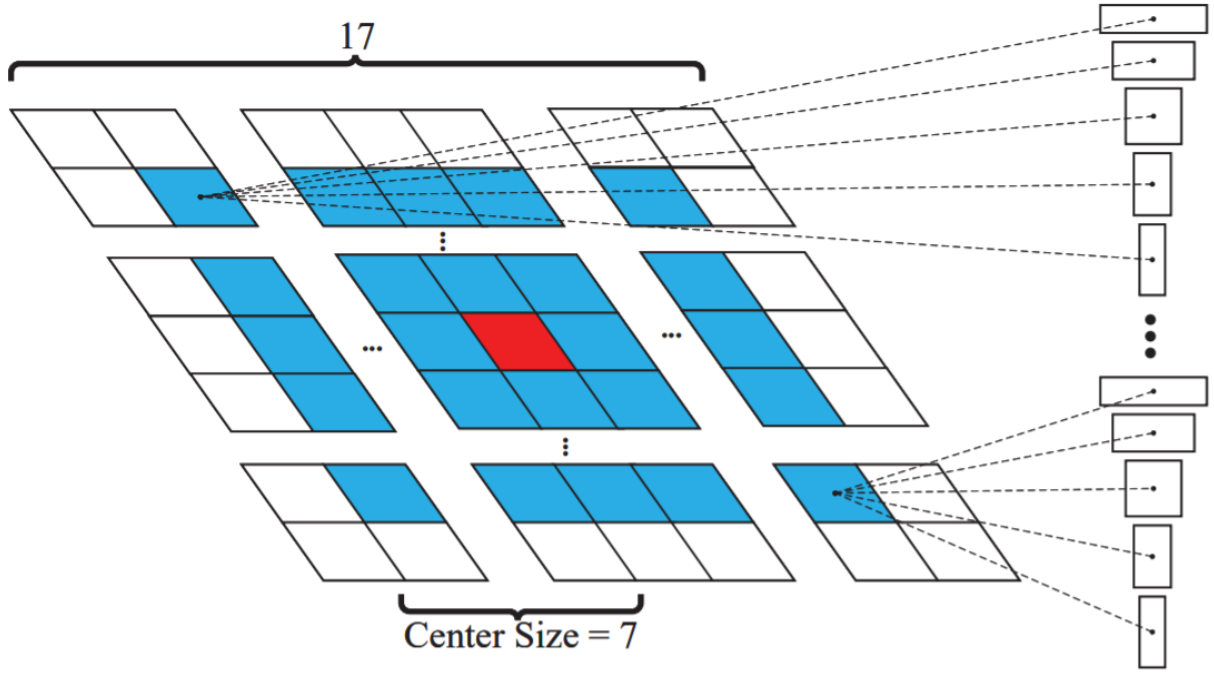
https://github.com/STVIR/pysot/blob/master/demo/bag.avi

https://github.com/STVIR/pysot

Siamese-RPN - Object Tracking

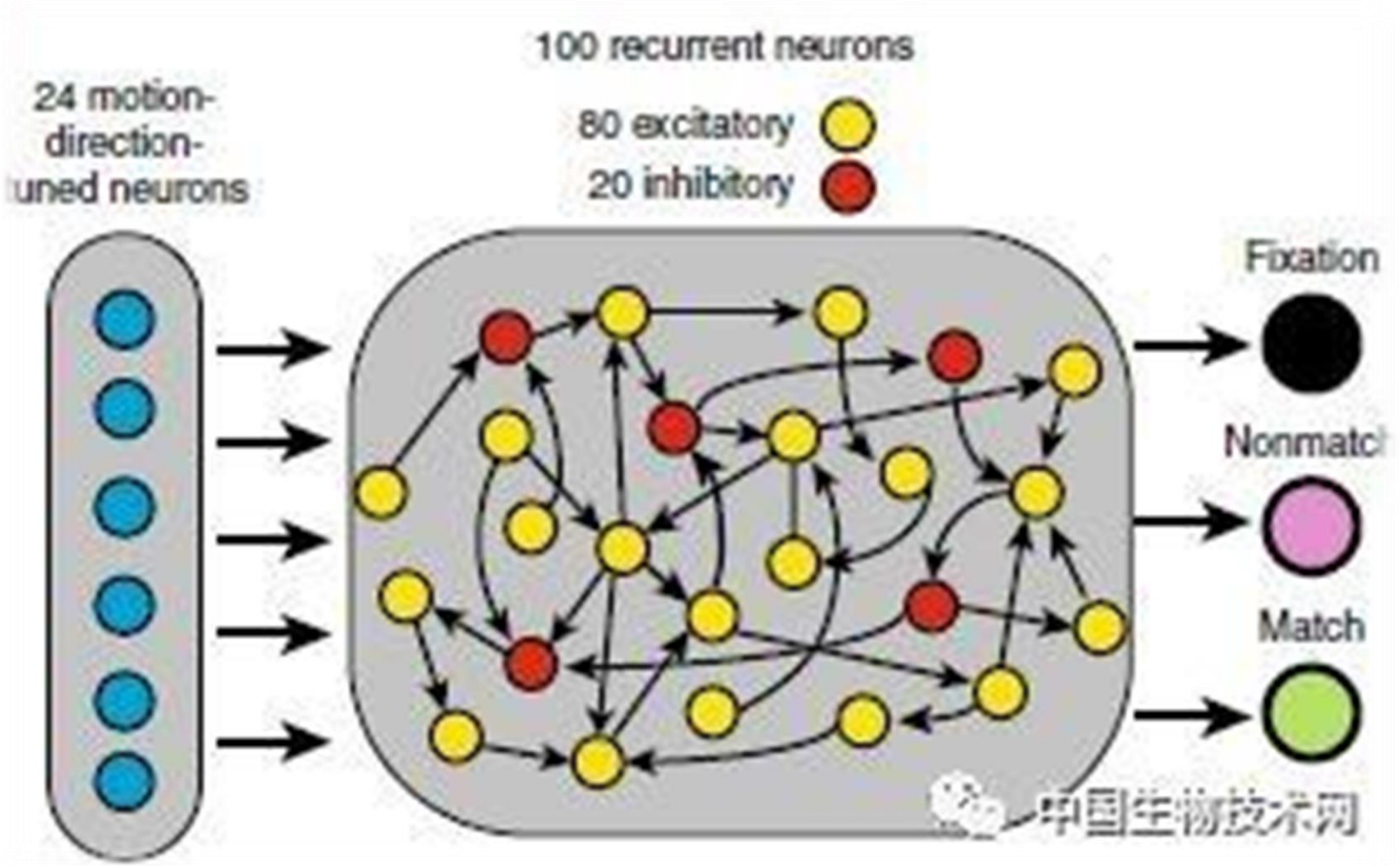


One-Shot Detection



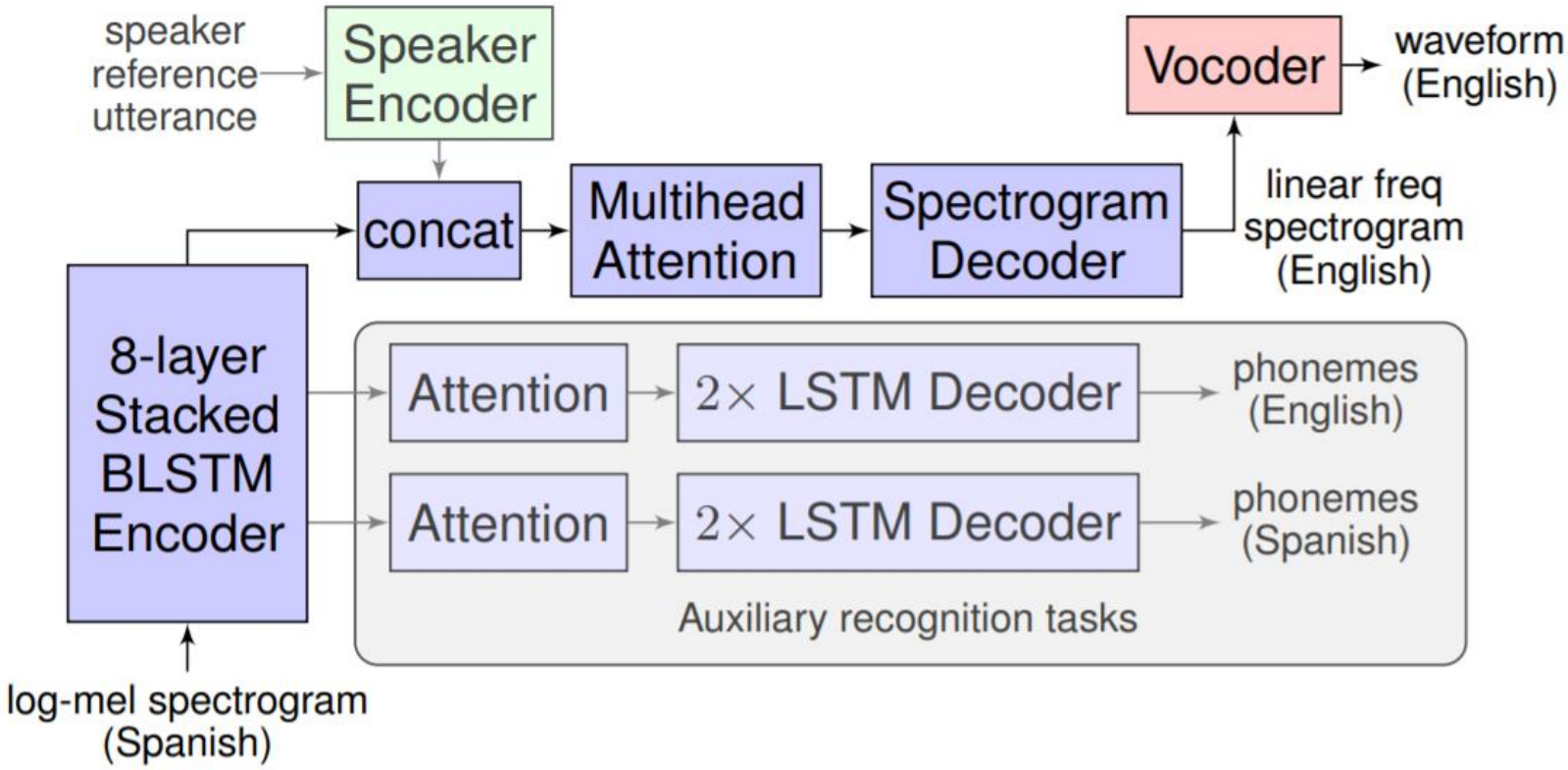
RPN feature map

Inhibitory for Shorterm Memory



- This AI model shows that during the silent period of memory, the brain can use the short-term plasticity of synaptic connections between neurons to memorize information.
- These two forms of short-term memory last from a few seconds to a few minutes. Some of the information used in short-term memory may eventually be stored for a long time, but most of the information will disappear over time.

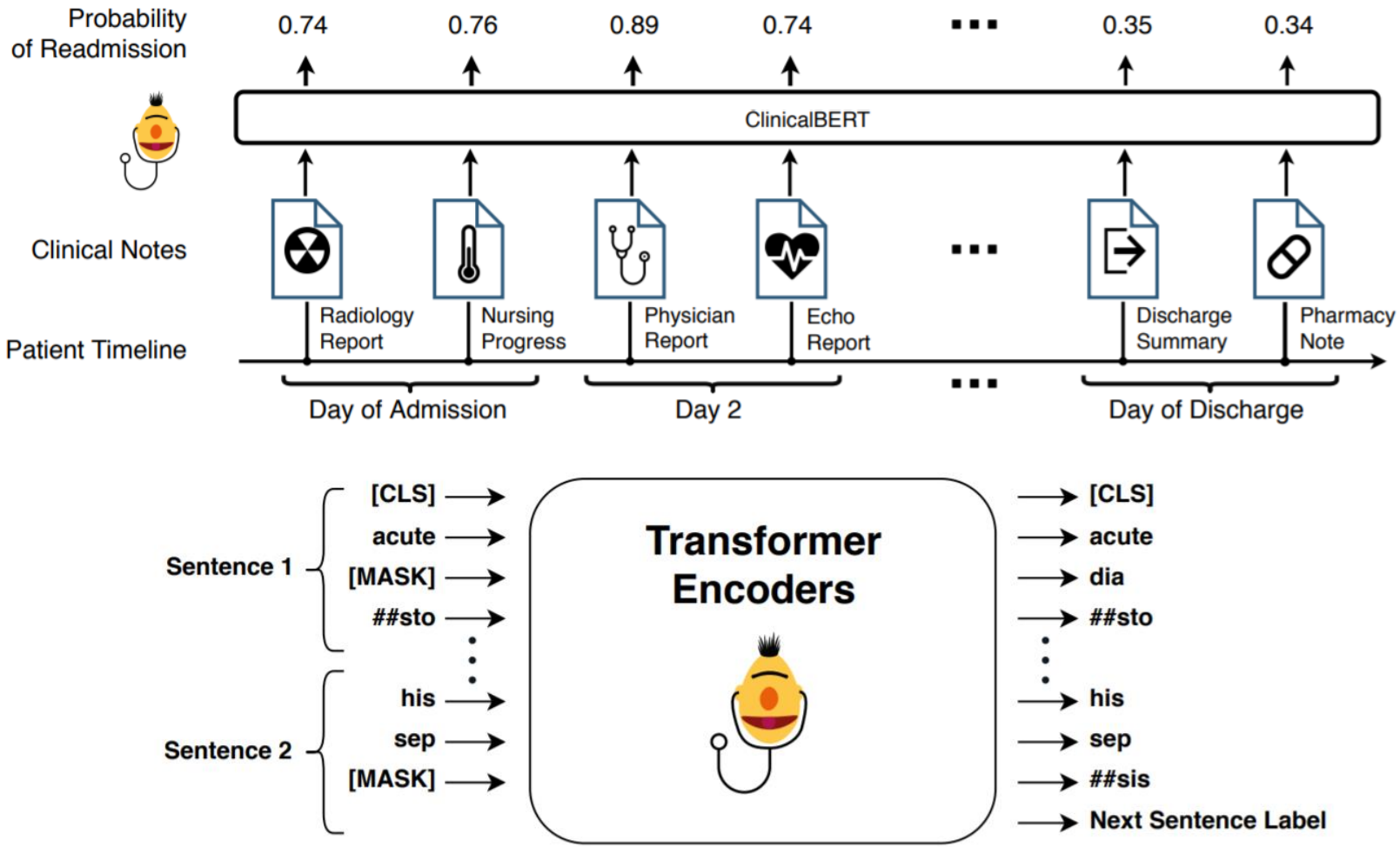
Direct Speech-to-Speech Translation



https://arxiv.org/abs/1904.06037

Clinical BERT - Readmission Prediction

Sparse text information



Short Comment

Many novel architectures

Many new applications

Would Deep Learning Model be important as Programming Language Model?

Comparison for Hardware Architectures

- CPU: Insufficient Energy Efficiency
- GPU: High efficiency in training, but low efficiency in reasoning (batch size = 1)
- DSP: Low hit rate of cache
- ASIC has high NRE: Large-scale application market has not yet formed
- ASIC has a long input period and neural network is developing

- FPGA (Reconfigurable Architecture):
  - Acceptable energy consumption and performance
  - Accept flexible architecture
  - On-chip storage with high bandwidth
  - Short Market Cycle

Demand for Low Power and High Performance Hardware

UAV
Client

Demand
Real-time scene recognition

Limitations
Limited battery capacity

Video surveillance
Edge

Demand
Real-time image analysis

Limitations
Low cost and high performance hardware

Speech recognition
Cloud

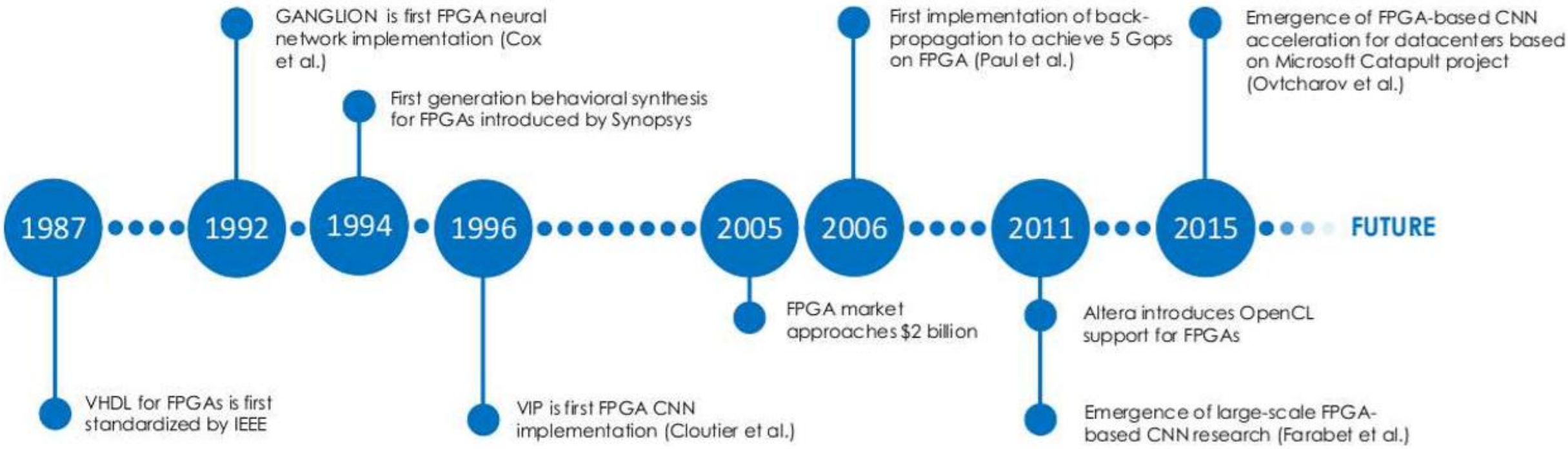
Demand
Processing delays are lower

Limitations
Higher maintenance and cooling costs

Problem of Current Architecture

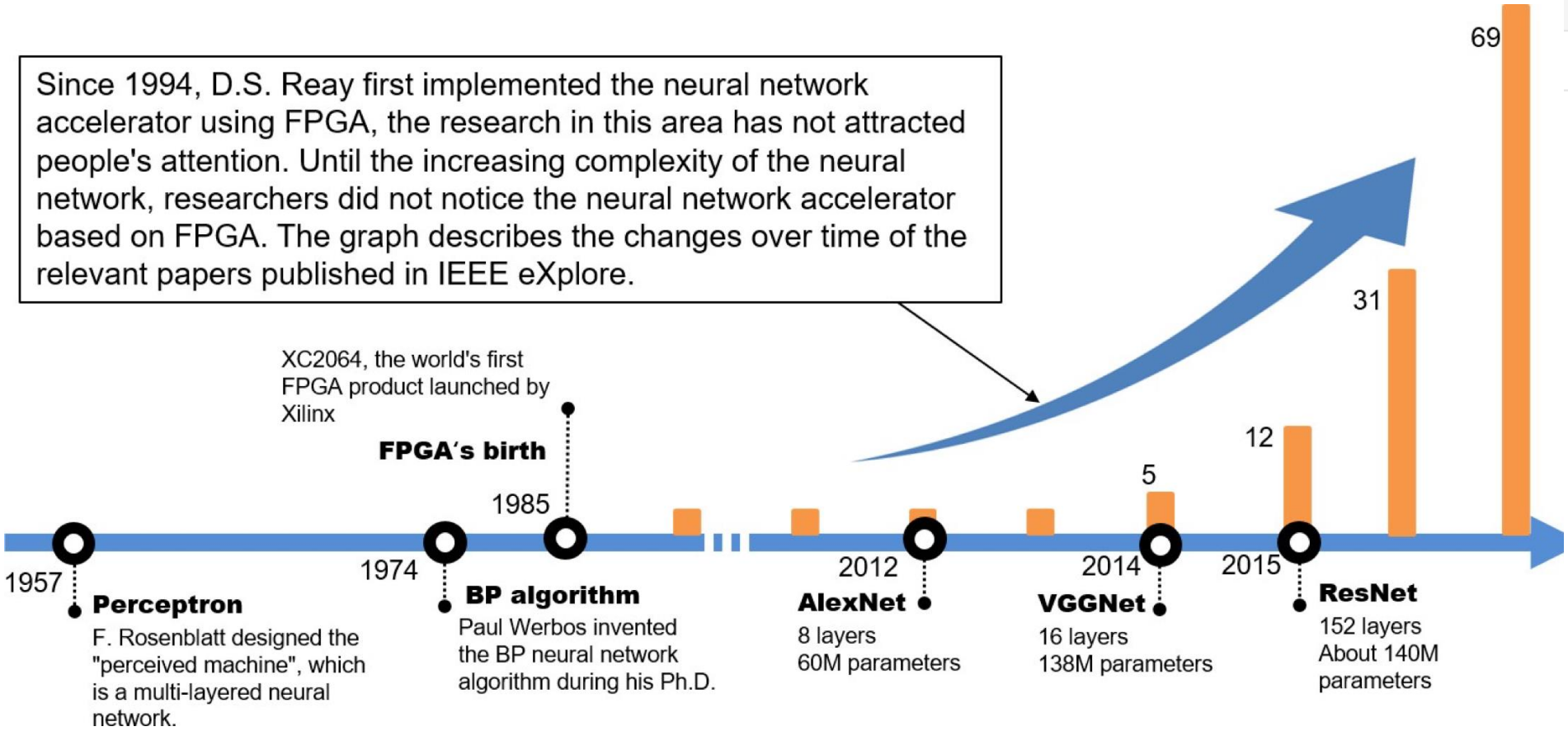
- High Redundancy in Neural Networks
  - VGG16 network can be compressed from 550 MB to 11.3 MB
- The limited bandwidth of BRAM and DDR in FPGA
- Different neural networks have different computational models
  - CNN: Frequent data reuse, high density
  - DNN/RNN/LSTM: No data reuse, data sparseness
- Different architectures need to adapt to different neural networks
  - With the rapid development of neural networks, the architecture should be adapted to the new algorithm.

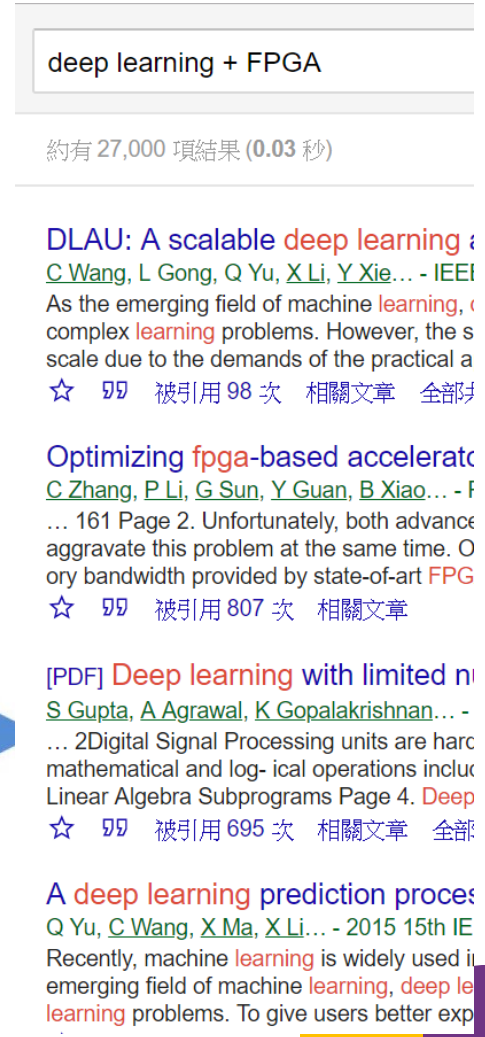
Development of FPGA CNN



Research Trend

Year 2016, the number of neural network accelerators based on FPGA published on IEEE eXplore had reached 69, and it has been increasing. This is enough to illustrate the research trend in this direction.

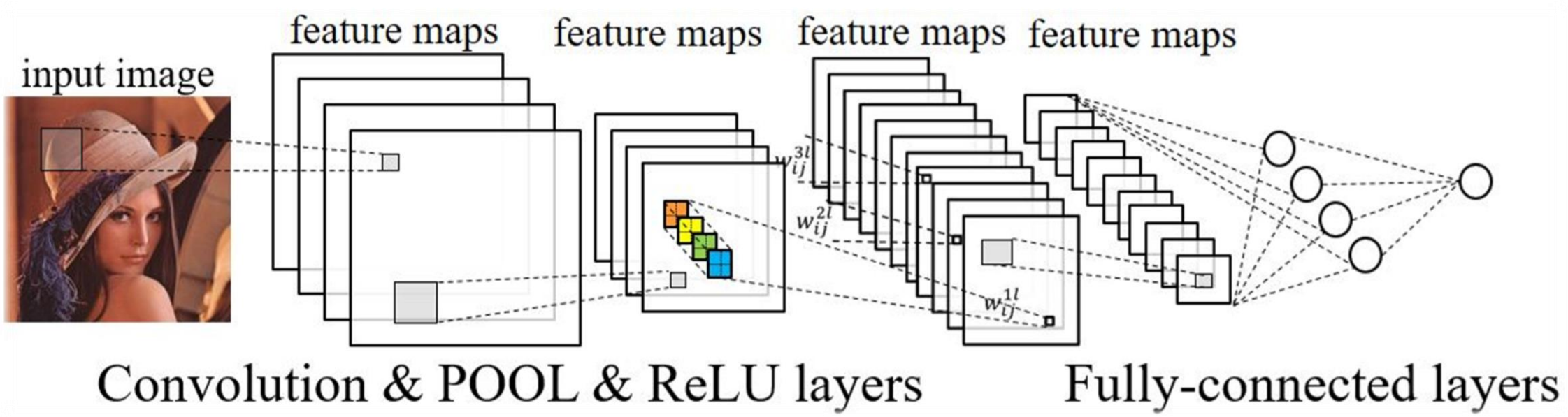




Level of Deep Learning Hardware Design

- **Designing Accelerators for Specific Applications**
- **Designing Accelerators for Specific Algorithms**
- **Designing Accelerators for Common Features of Algorithms**
- **Designing a Universal Accelerator Framework with Hardware Templates**

Structure and Complexity of CNN



	CONV	POOL	ReLU	FCN
Comput. ops(10^7)	$3E3(99.5\%)$	$0.6(0\%)$	$1.4(0\%)$	$12.3(0.4\%)$
Storage (MB)	$113(19.3\%)$	$0(0\%)$	$0(0\%)$	$471.6(80.6\%)$
Time% in pure sw	96.3%	0.0%	0.0%	3.7%
after CONV acc	**48.7%**	**0.0%**	**0.0%**	**51.2%**

Hardware Accleration for CNN

```
1. for(o=0; o< To ; o++){
2.     for(i=0; i< Ti ; i++){
3.         for(r=0; r< Tr ; r++){
4.             for(c=0; c< Tc ; c++){
5.                 for(p=0; p< K1 ; p++){
6.                     for(q=0; q< K2 ; q++){
                        cache_output[o][r][c] +=
                        cache_weights[o][i][p][q] * cache_input[i][ S *r+p][ S *c+q];
} } } } } } } } } }
```

Fig. 5: Pseudo code of original on-chip computation

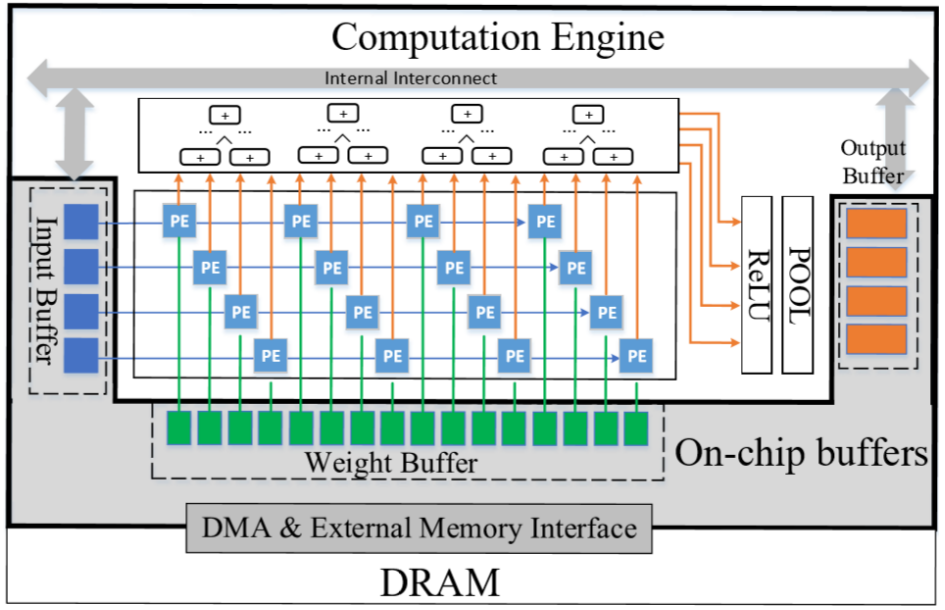
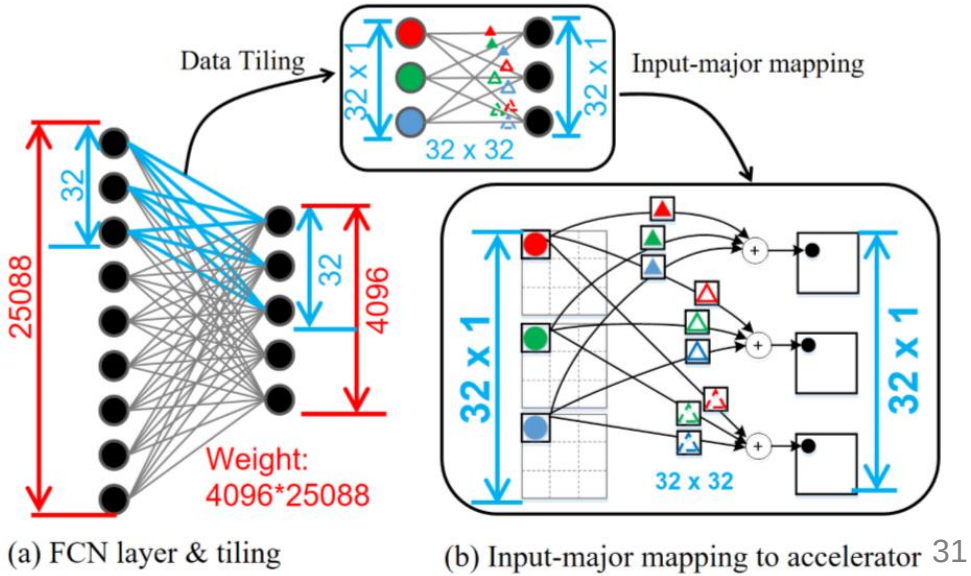


Fig. 7: Scalable accelerator architecture design

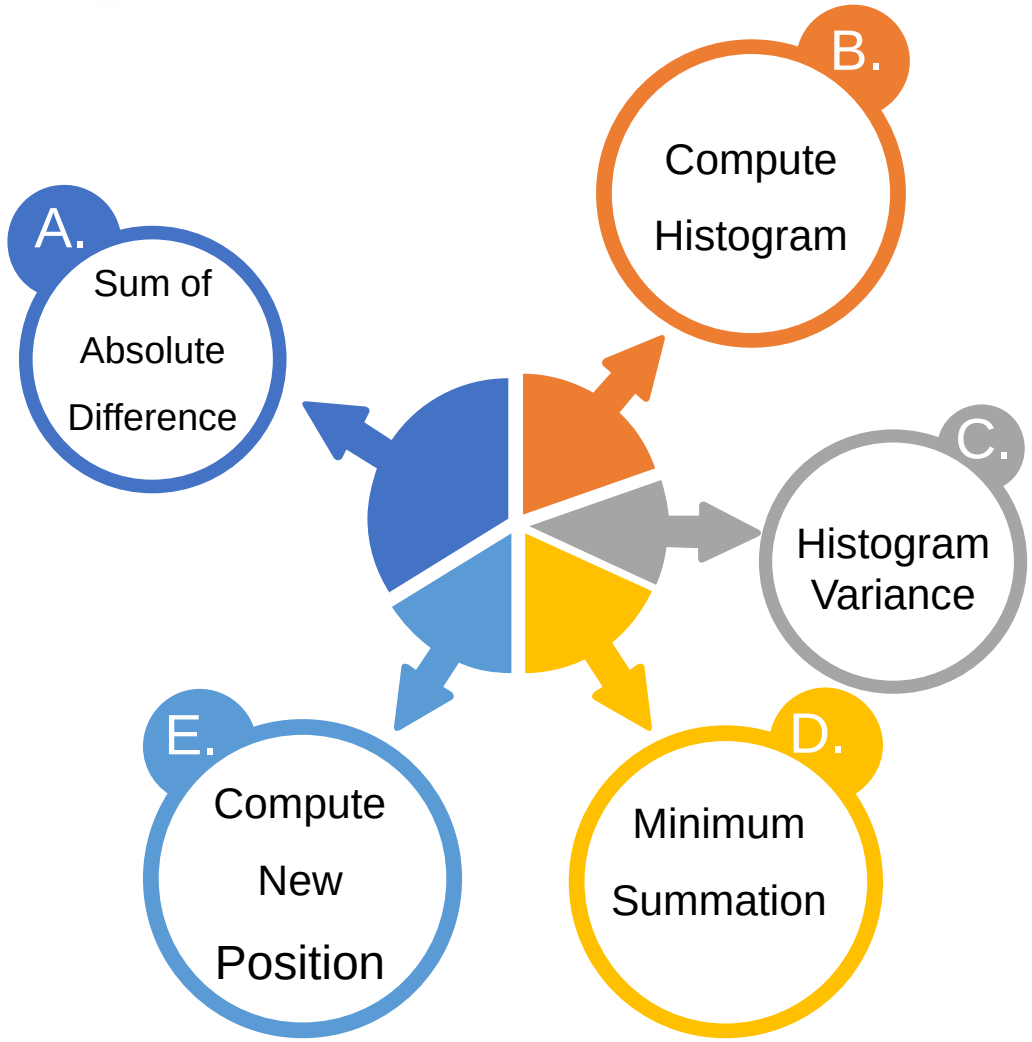
```
1. for(p=0; p< K1 ; p++){
2.     for(q=0; q< K2 ; q++){
3.         for(r=0; r< Tr ; r++){
4.             for(c=0; c< Tc ; c++){
#pragma for LOOP PIPELINE (Fine-grained Pipeline Parallelism)
5.                 for(o=0; o< To ; o++){
#pragma for LOOP UNROLL (Multilevel Data Parallelism)
6.                     for(i=0; i< Ti ; i++){
#pragma for LOOP UNROLL (Multilevel Data Parallelism)
                        cache_output[o][r][c] +=
                        cache_weights[o][i][p][q] * cache_input[i][ S *r+p][ S *c+q];
} } } } } } } } } }
```

Fig. 6: Pseudo code of optimized on-chip computation



(a) FCN layer & tiling (b) Input-major mapping to accelerator

Object Tracking FPGA Architecture



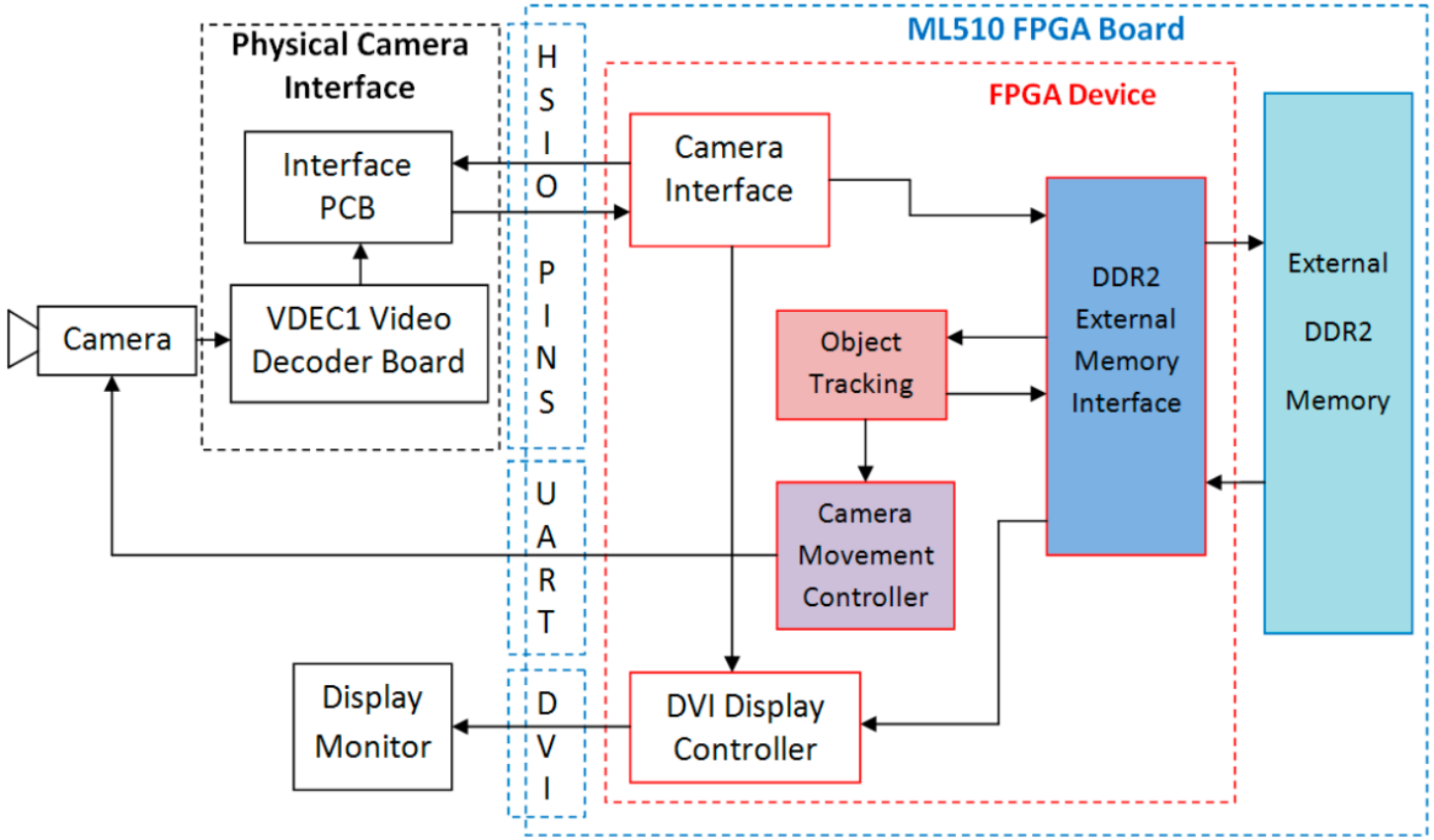


Figure 2. Dataflow diagram of the proposed and developed object tracking system.

Object Tracking Result

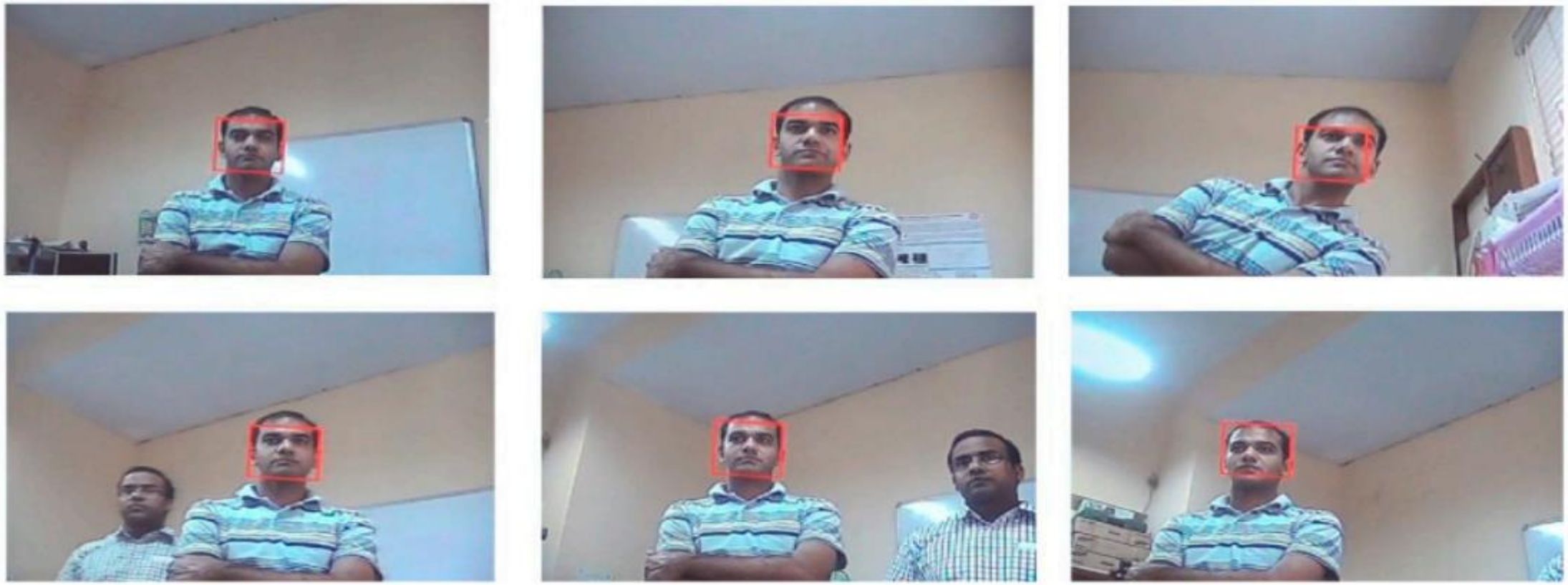


Figure 16. Frame sequence showing object tracking in scenes that change due to camera movement and presence of other moving objects in the scene.

OpenCL FPGA Framework

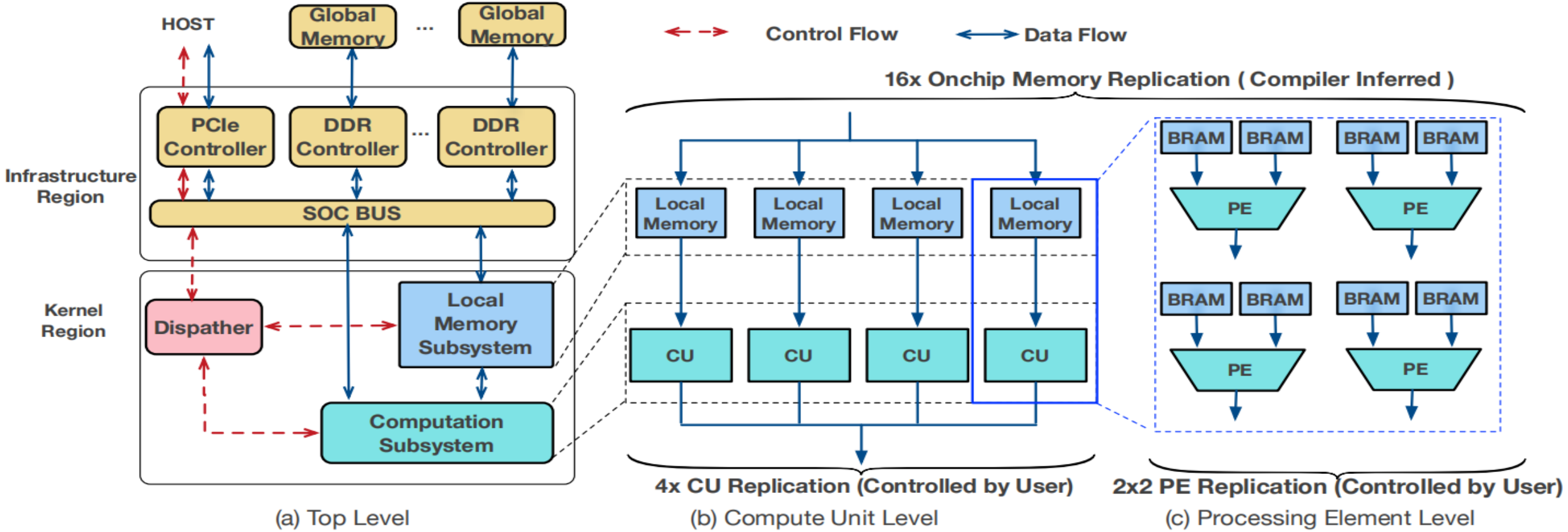
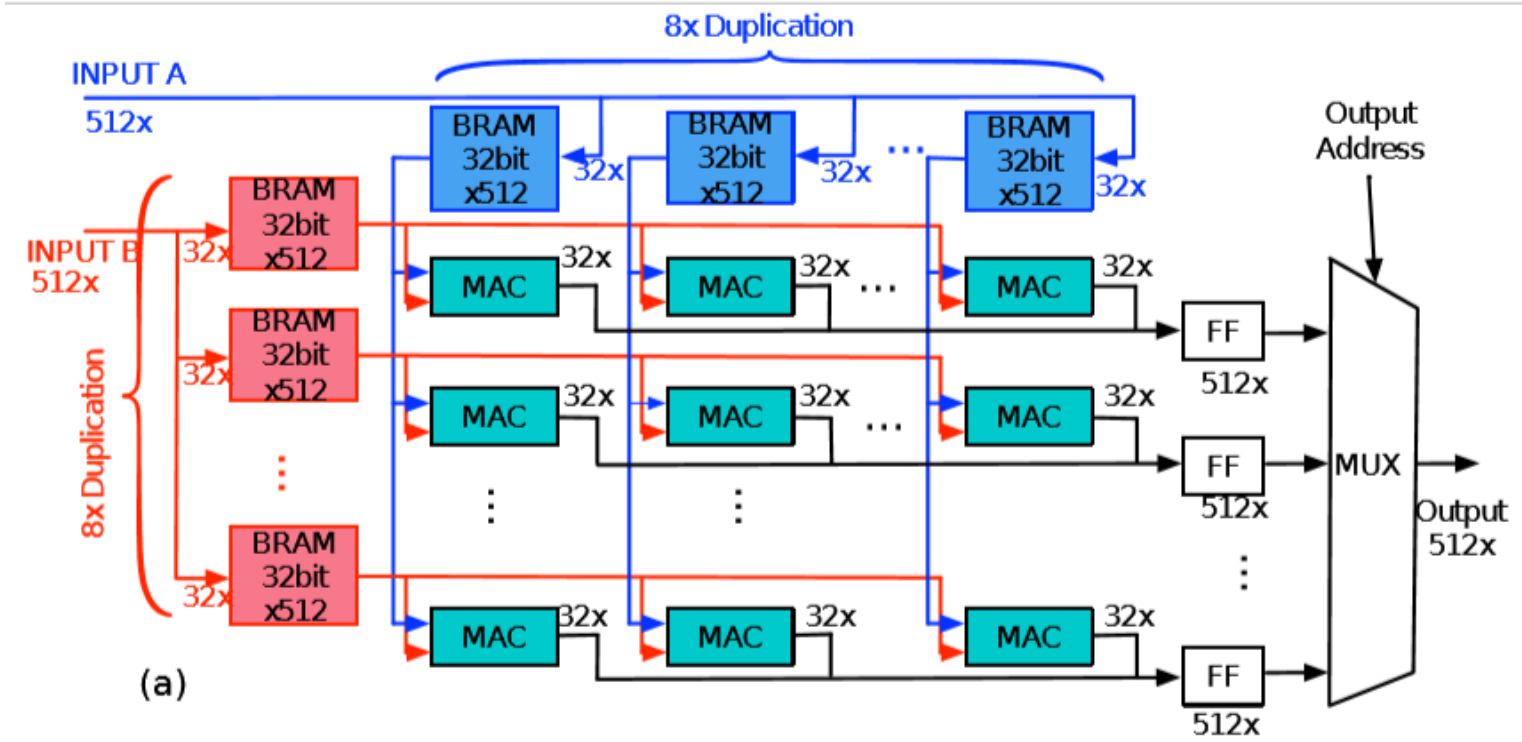


Figure 2: OpenCL FPGA framework:(a) Top level ;(b) Compute unit (CU); (c) Processing element(PE)

OpenCL FPGA Framework

- Processing element for Convolution:
  - A table tennis mechanism (similar to pipeline) is introduced to transmit data and operations to hide latency of external memory access.
  - A computing unit has 256 DSP chips, which can parallel 256 computations at a time by reusing the storage ports of 16 rows and columns.

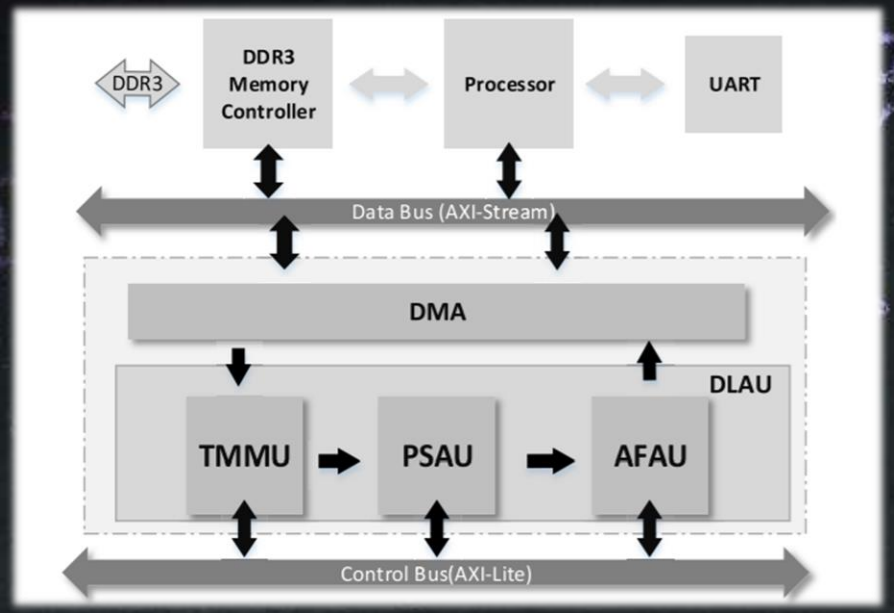


DLAU : SLICE TECHNIQUES

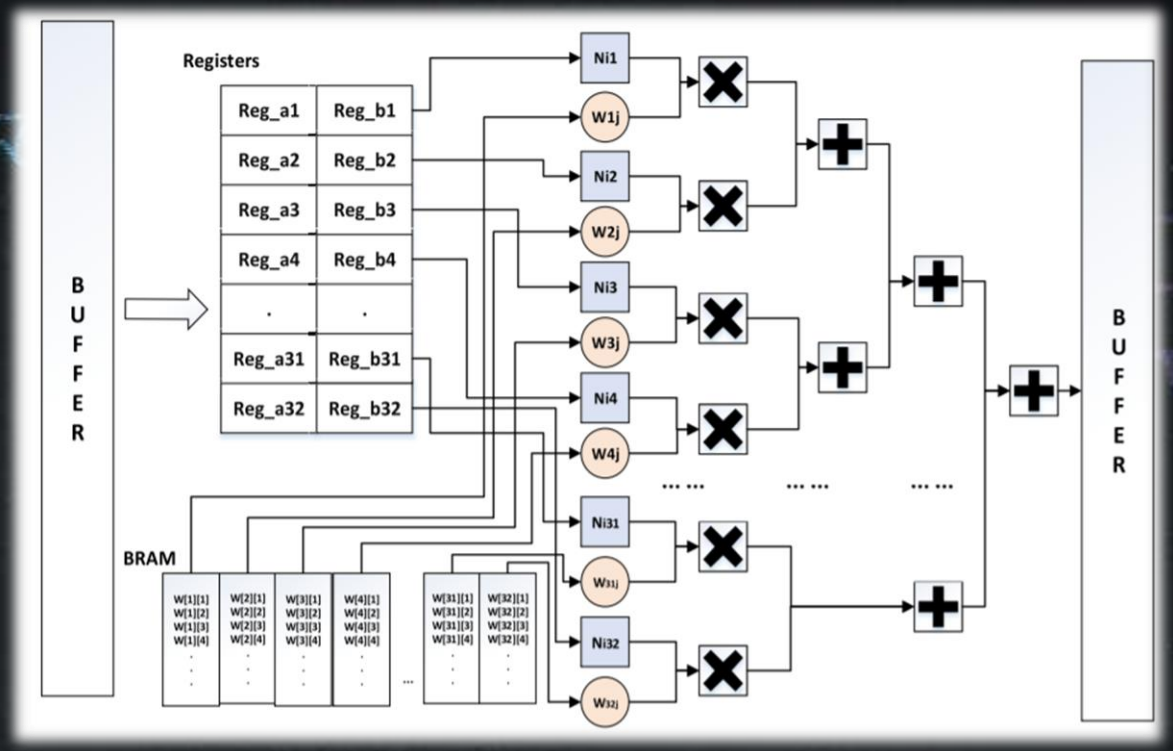
- No matter how big the data in the input neuron is, it can be sliced into several data subsets of the same size.
- The weight matrix is also divided into slices of the same size according to the size of data slices.
- Multiply the slice with its corresponding weight matrix to get partial sum
- The above operations are performed on each slice until the data is processed.

```
Require:
  Ni: the number of the input neurons
  No: the number of the output neurons
  Tile_Size: the tile size of the input data
  batchsize: the batch size of the input data
  for n = 0; n < batchsize; n++ do
    for k = 0; k < Ni; k+ = Tile_Size do
      for j = 0; j < No; j++ do
        y[n][j] = 0;
        for i = k; i < k + Tile_Size&&i < Ni; i++ do
          y[n][j]+ = w[i][j] * x[n][i]
          if i == Ni − 1 then
            y[n][j] = f(y[n][j]);
          end if
        end for
      end for
    end for
  end for
```

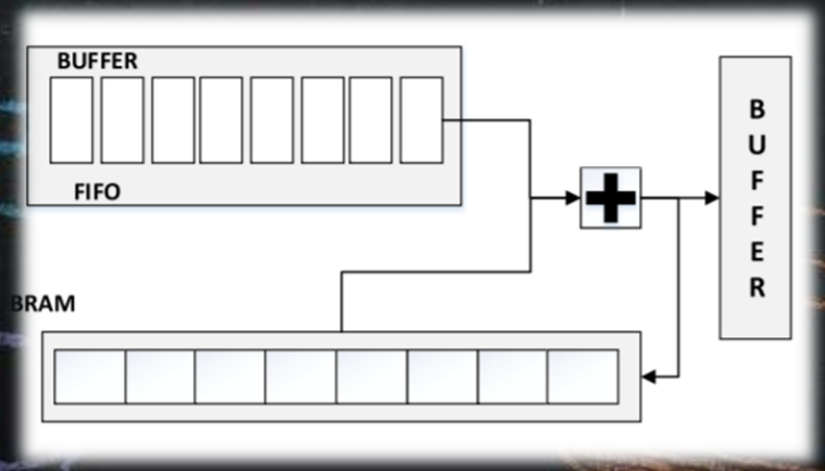
TILE TECHNIQUES



TMMU for Tile and weight sum



PSAU for Accumlation

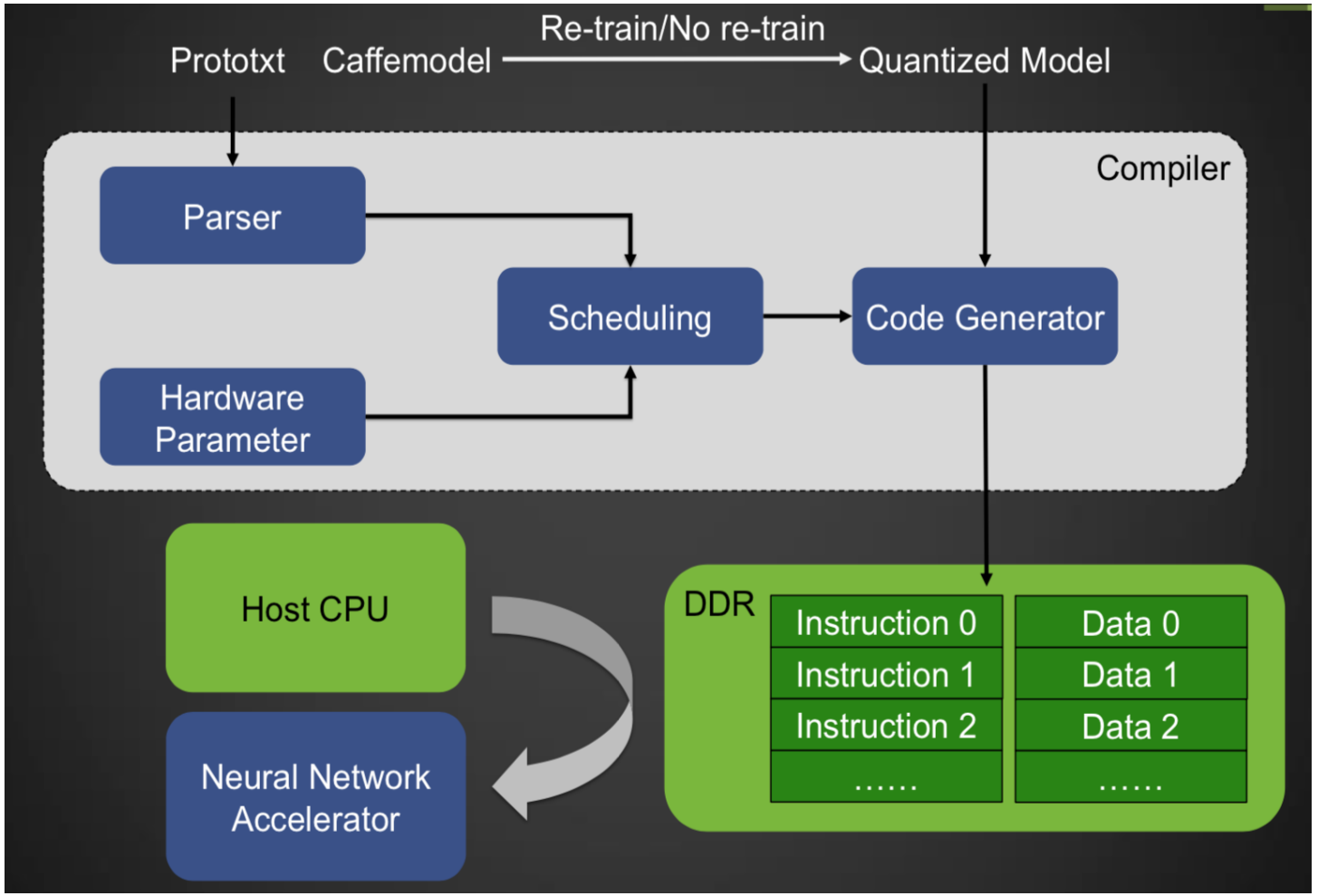


AFAU for Activation Funtion

$$f(x) = \begin{cases} 0 & \text{if } x \le -8 \\ 1 + a\left[\left\lfloor \frac{-x}{k} \right\rfloor\right]x - b\left[\left\lfloor \frac{-x}{k} \right\rfloor\right] & \text{if } -8 < x \le 0 \\ a\left[\left\lfloor \frac{x}{k} \right\rfloor\right]x + \left[\left\lfloor \frac{x}{k} \right\rfloor\right] & \text{if } 0 < x \le 8 \\ 1 & \text{if } x > 8. \end{cases}$$

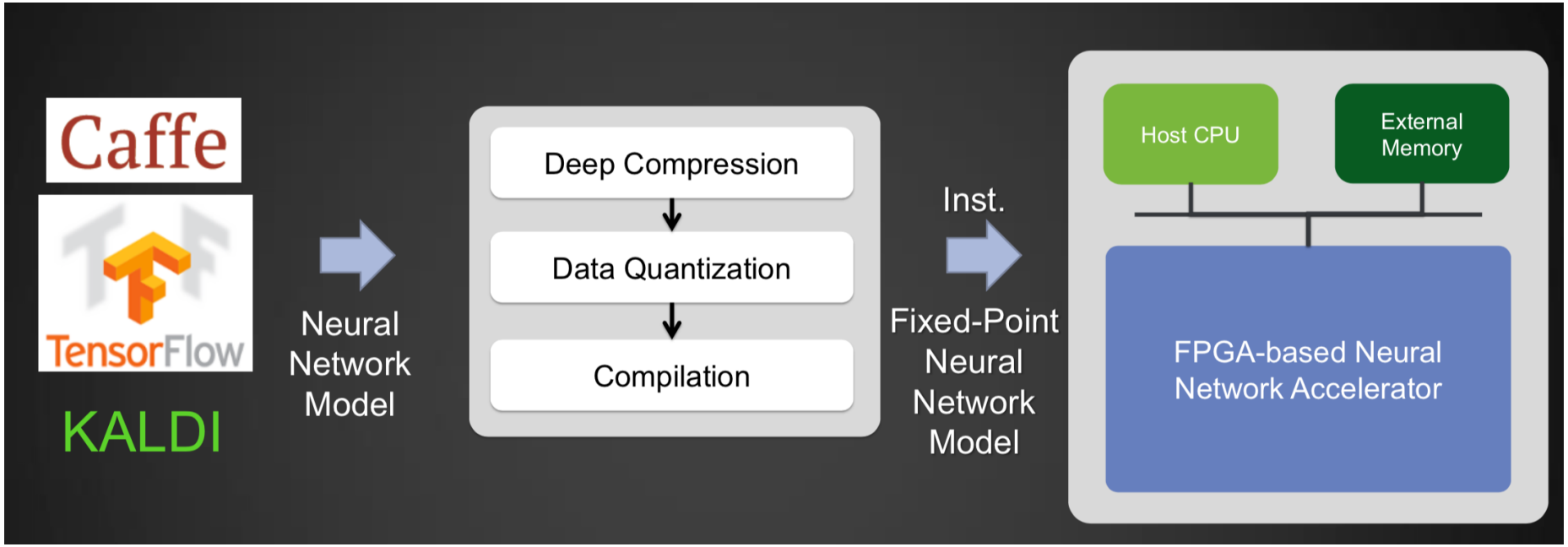
DeepPhi : Architecture

- Compiler + Framework Replaces OpenCL
- Algorithmic developers do not need to understand hardware architecture
- Generate instructions instead of RTL code
- Compile in one minute
- Better performance and lower energy consumption

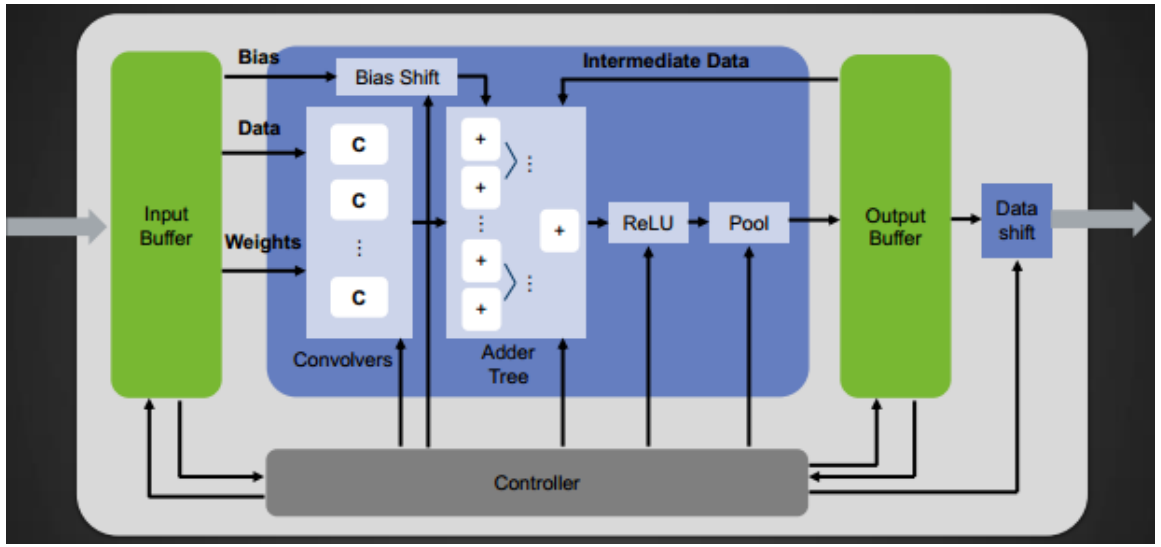


DeepPhi:Workflow and Proessor Eelement

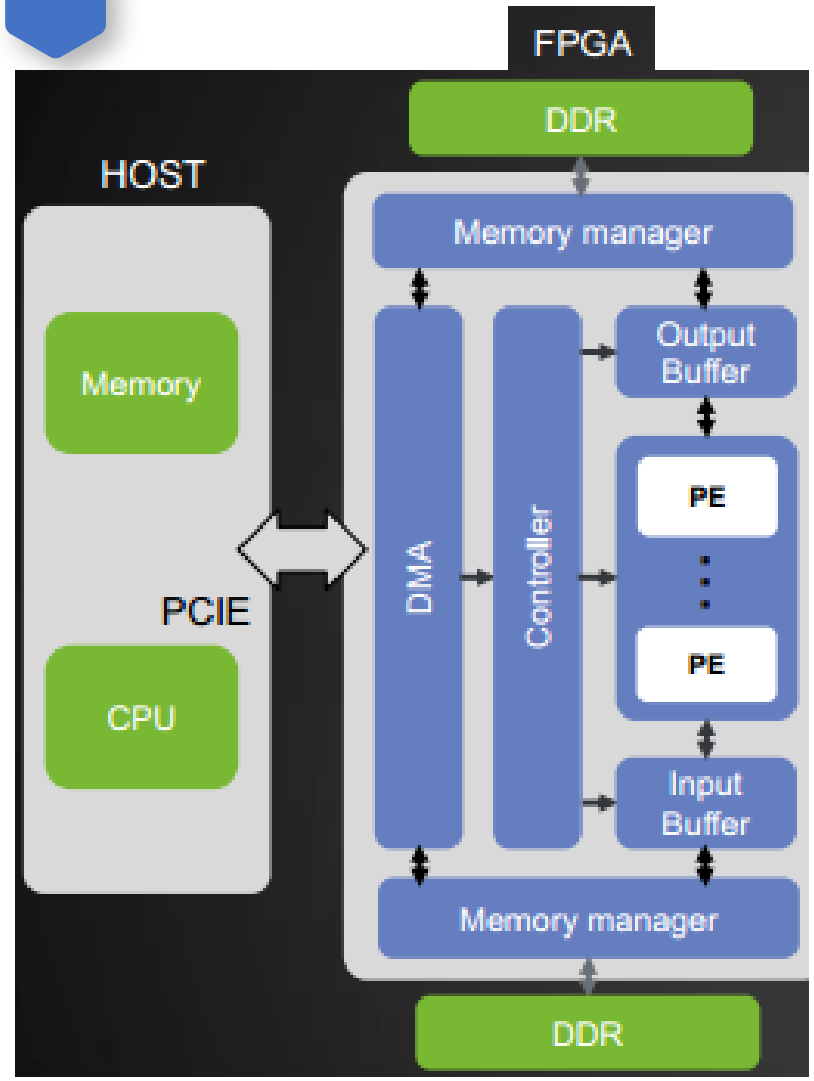
Workflow

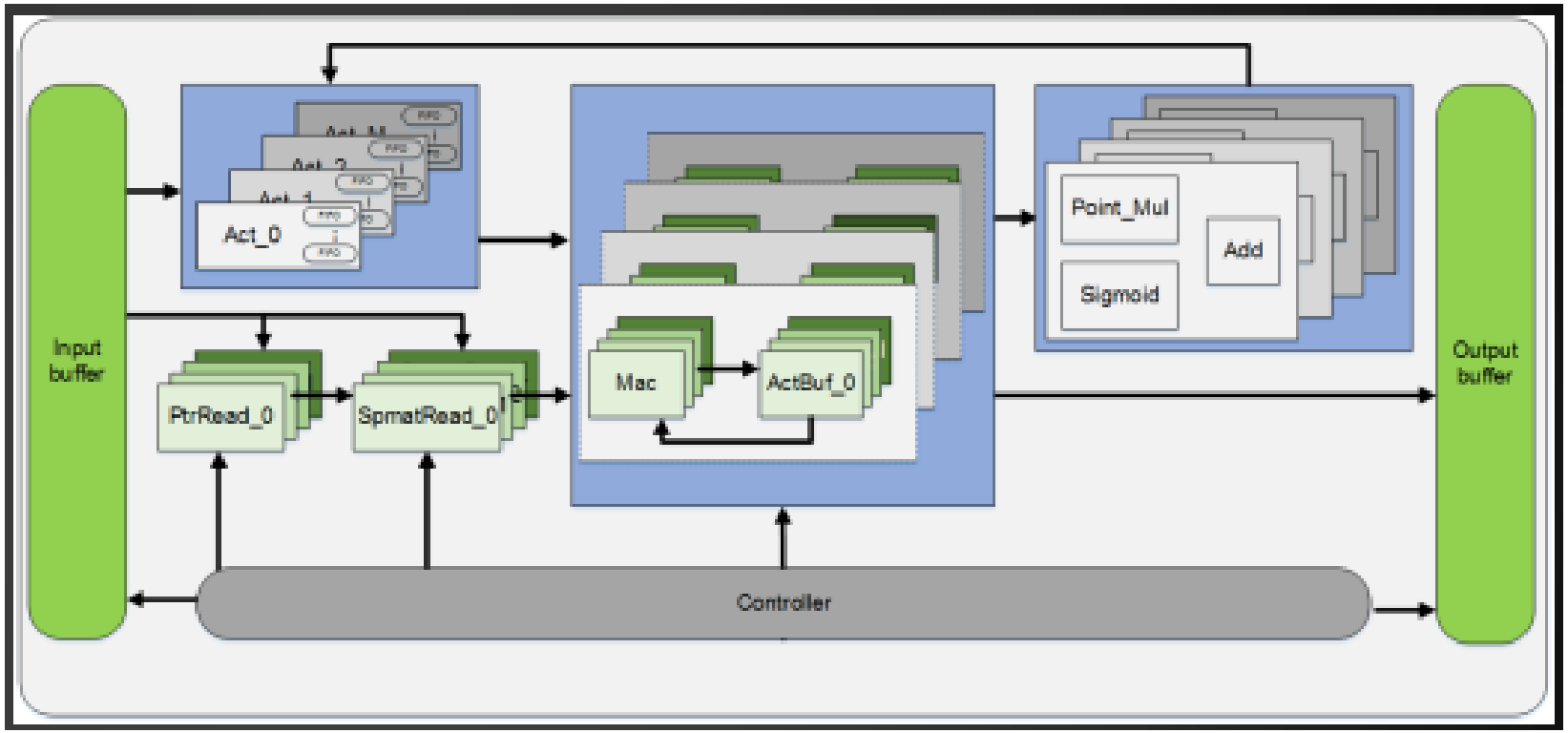


Proessor Eelement

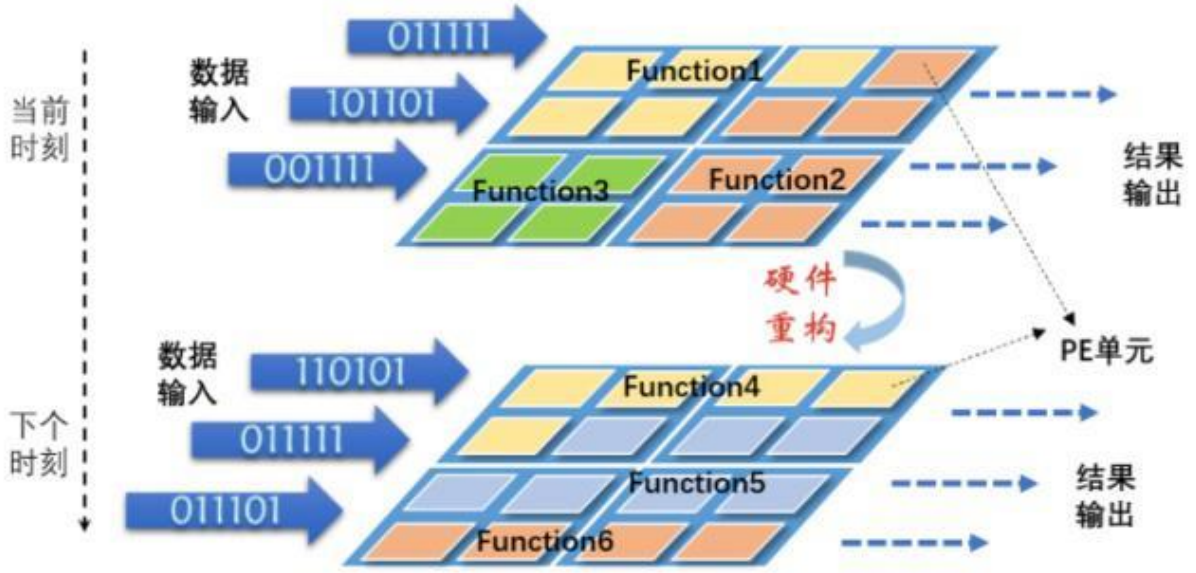


DeepPhi:RNN/LSTM Architecture



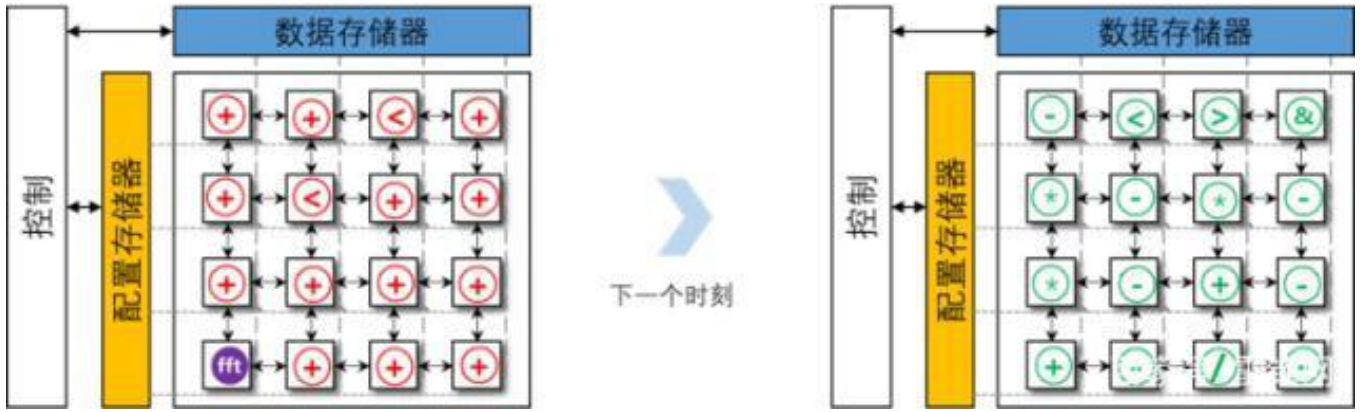


Coarse grained Reconfigurable Architecture

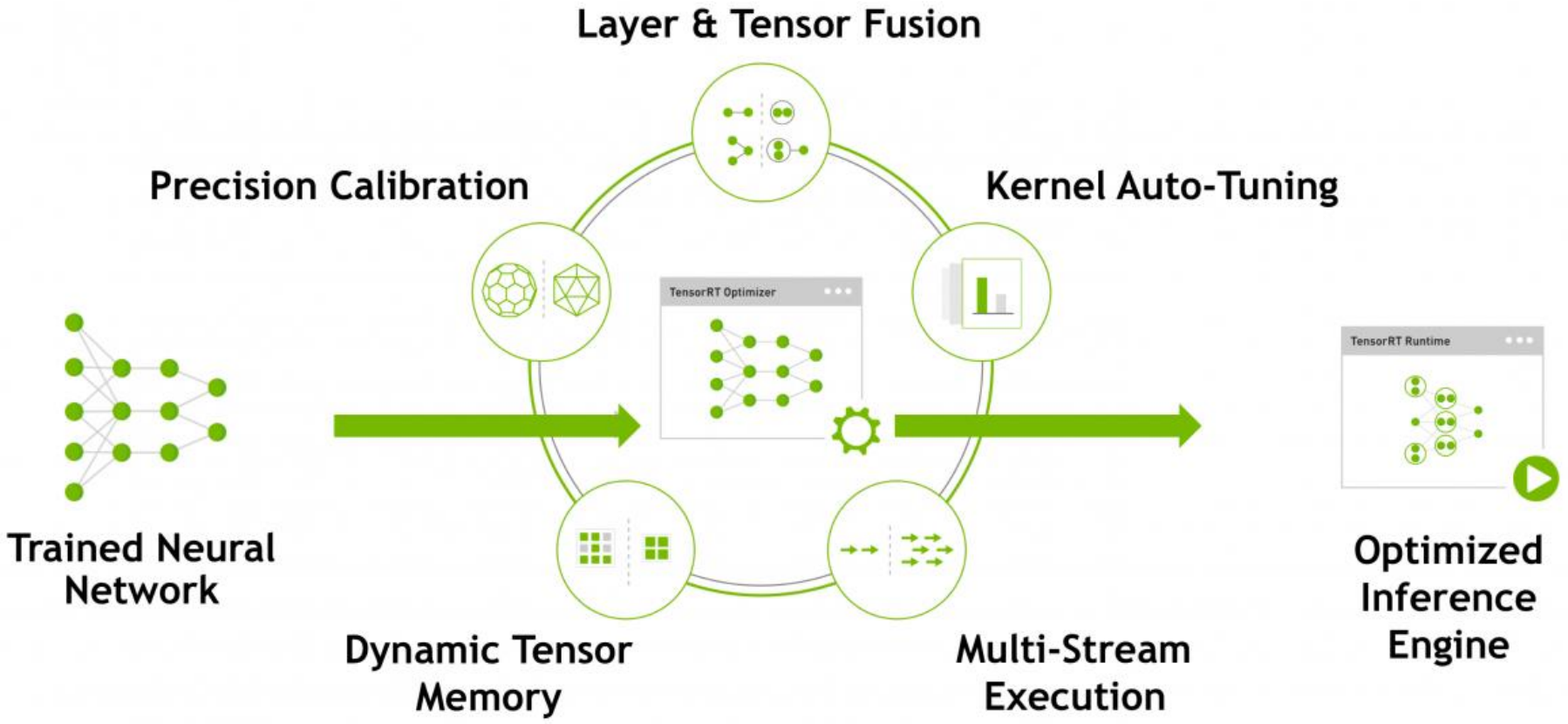


Reconfigurable Array

- CGRA computing energy efficiency can reach 1000 times of CPU computing architecture,
- 100-1000 times of GPU computing architecture, and more than 100 times of FPGA computing architecture.
- Compared with NPU, CGRA can improve performance more than 10 times.
- CGRA is based on configuration mode, and its execution efficiency can be comparable to ASIC, but its flexibility is much better than ASIC.



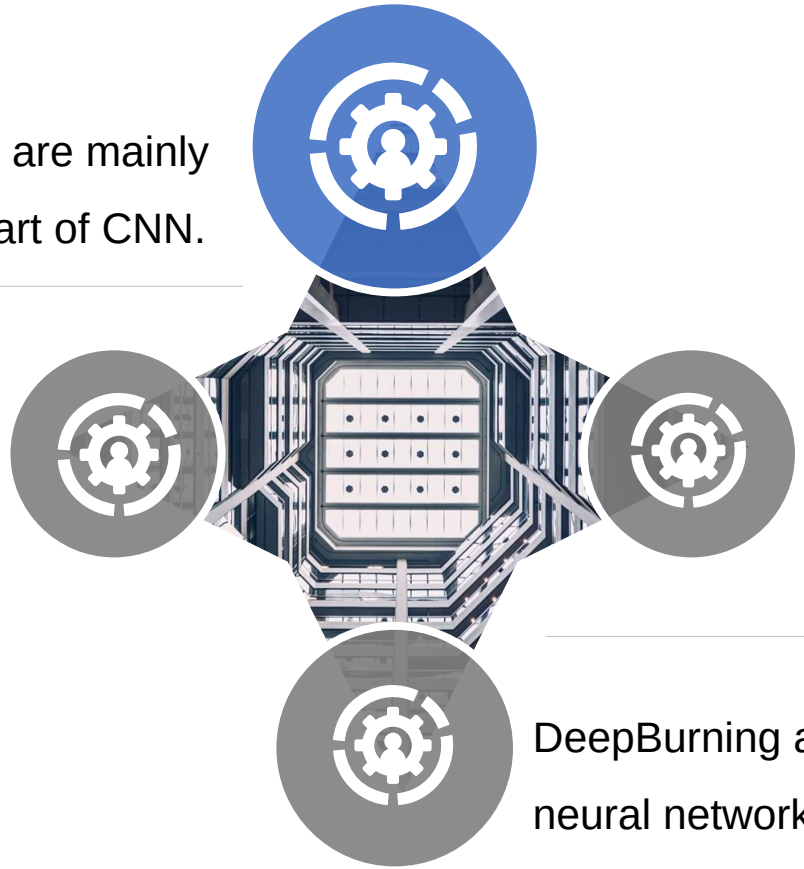
Nvidia NVIDIA TensorRT



Other CNN-TO-FPGA Tools

FpgaConvNet, ALAMO and Snowflake are mainly concerned with the feature extractor part of CNN.

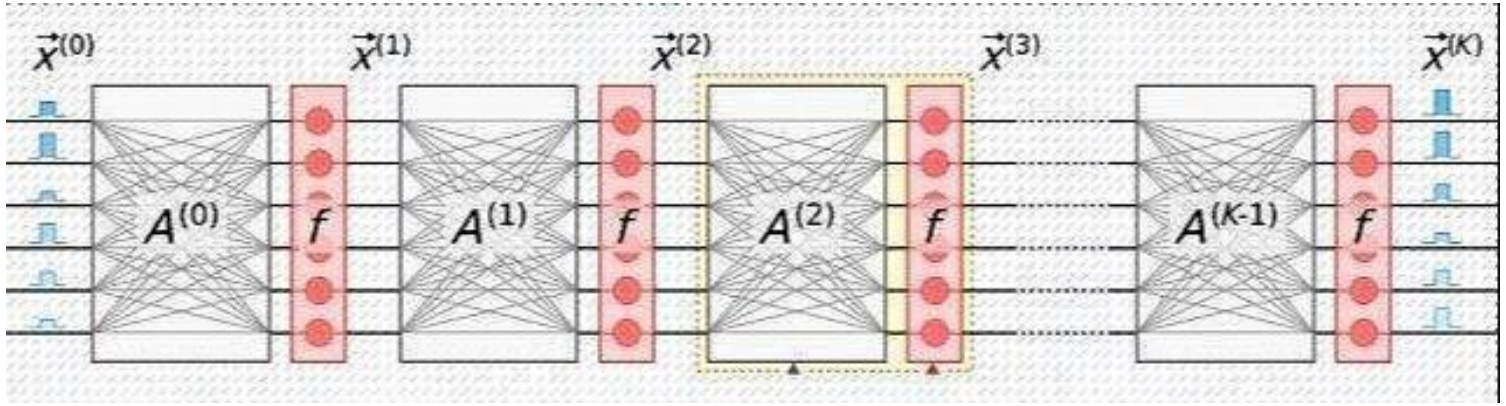
Inception module of fpgaConvNet, Snowflake and dense module of fpgaConvNet support irregular CNN building module and other modules

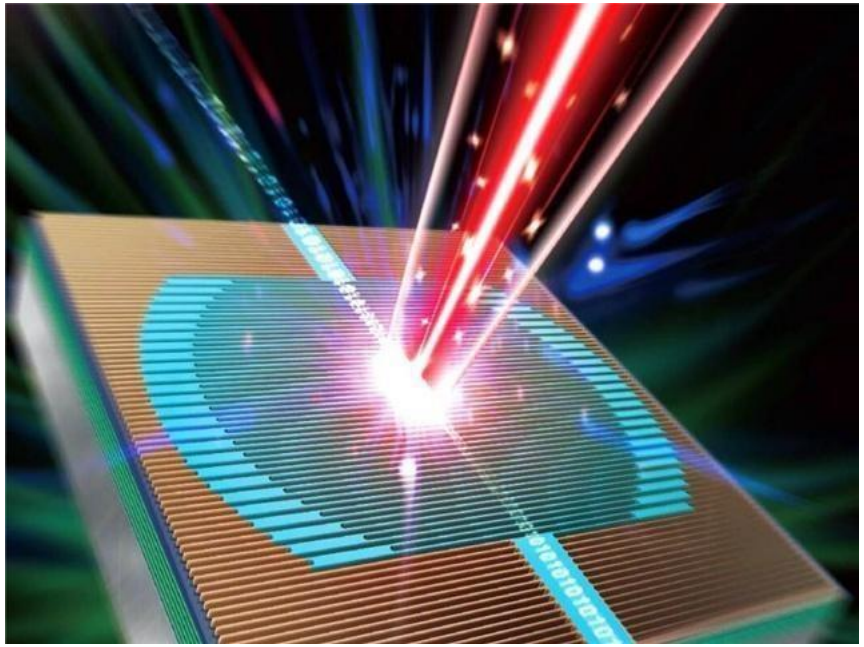


Haddoc2 requires all weights to be stored on the chip, so the size of the supported model is limited by the storage resources of the target device.

DeepBurning and FP-DNN support recurrent neural network (RNN) and long-term and short-term memory (LSTM) networks.

New Photonic AI Chips





In a paper in Physical Review X, MIT researchers describe a new photon accelerator that uses optical components and optical signal processing technology to reduce chip size, which will allow the chip to **expand to neural networks several orders of magnitude larger than electrical chips**.

Near Future Archtecture

Support Next Generation Network

- Increase in Model Depth
- Increase the workload of reasoning
- Introducing new components (e.g. enhancing the CNN layer by introducing complex blocks)

Support Compression, Sparse

- Post-training
- Training-time methods

Support Low Accuracy

- Angel-Eye, ALAMO, DnnWeaver, DeepBurning and AutoCodeGen support dynamic quantization with fixed, uniform word length and different scaling across layers

Integrate with popular frameworks

- Such as Google's TensorFlow, seem to be of interest to academia and industry because of the variety of machine learning models supported and the flexibility to deploy across different heterogeneous systems.

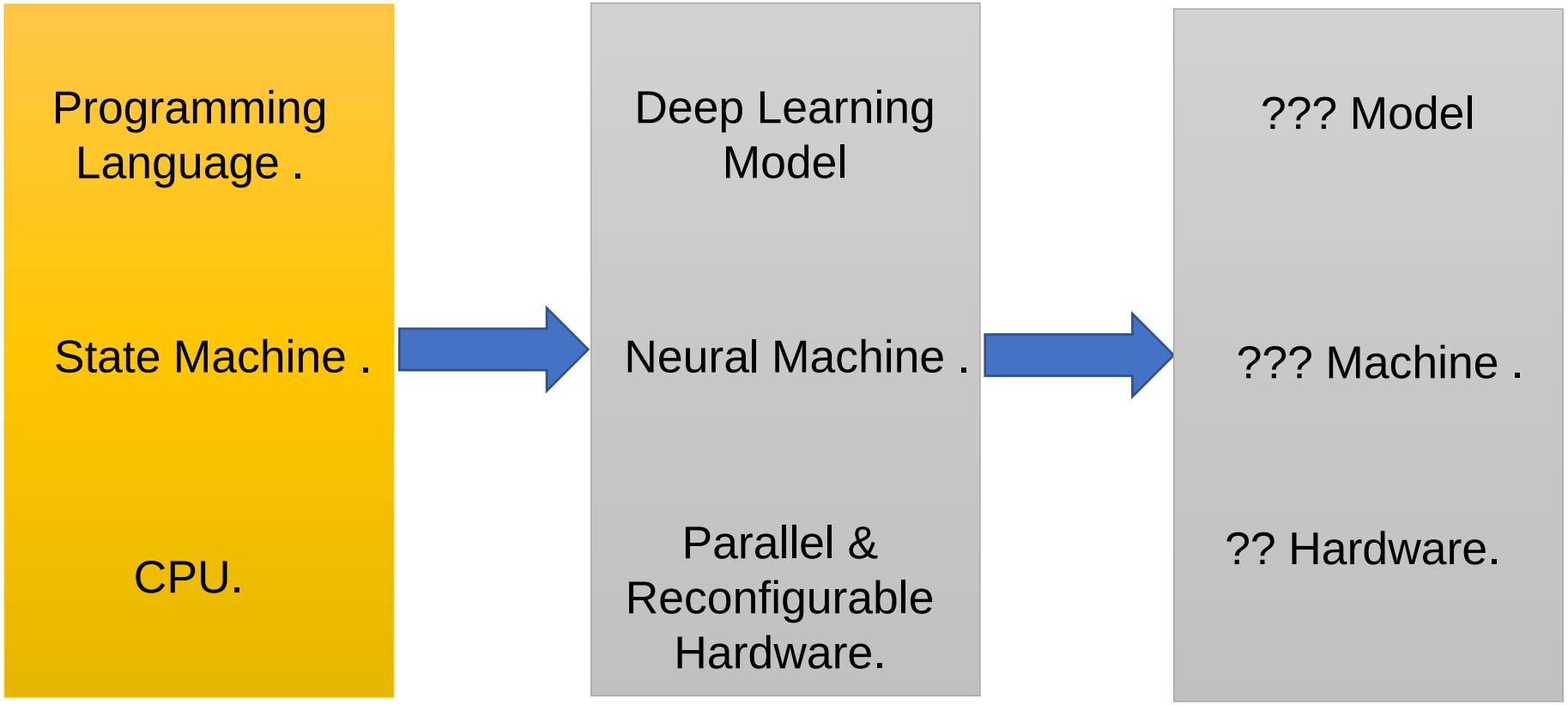
Support Hardware Unit

- Such as deploying Tensor Processing Unit (TPU) ASIC in its servers for the training and reasoning stages of machine learning models.

Hardware-Network Codesign

- By taking hardware performance and power consumption as indicators in the training phase, hardware adjustable parameters, model weight and topology will be jointly modified in the optimization process to jointly optimize the application-level accuracy and the required reasoning execution time and power consumption.

Conclusion



Artificial intelligence with deep learning architecture is still in infancy.
But it has already brought a lot of help to mankind.