Review of design approaches toward artificial intelligence system-on-chip (SoC)

SECTION I. INTRODUCTION

1) Various types of machine intelligence algorithm

   A. Multilayer perceptron (MLP)

   B. Fuzzy inference system (FIS)

   C. Neuro-fuzzy (NF) system

2) Computer vision applications

   A. Image classification

   B. Object detection

   C. Autonomous vehicles

3) Variant of DNN used for different applications

   A. Convolutional Neural Network (CNN) used for image processing

   B. Recurrent neural network (RNN) used for natural language processing

4) DNN consists of hundreds of layers requiring hundreds of layers that they require huge amount of computations and memory footprints

   A. Most of applications target at bulky systems with high performance GPUs or servers rather than mobile applications although the needs of mobile DNN are increasing nowadays

   B. More advanced DL researches have given promise of success on mobile applications by providing new user experience or useful functionality with sensory data such as image, video, and voice that can be easily collected on mobile devices

   C. However, the computationally expensive DL algorithms suffer from satisfying processing speed requirements on mobile applications with limited hardware resources and power budget

D. In spite of algorithmic efforts to lighten DNN architectures, they could not resolve fundamental problem of heavy computational cost

E. Thus, the importance of hardware accelerator design is gaining more attention worldwide and they are actively investigated to efficiently run DNNs in real time with low power.

5) Neuromorphic Chips

A. Aim at modeling the operations of biological brain to achieve high energy efficiency since human brain is well-known as the most energy-efficient computer

B. The development of neuromorphic chip also gained lots of attention among researchers

6) There have been many efforts to develop dedicated processors from the early age of neural networks with various hardware design approaches and it is important to analyze the previous approaches to develop advance System on Chips (SoCs).

7) There are three different hardware design approaches with their own pros and cons

A. Analog

Analog Neural Network (NN) VLSIs enable low-cost parallelism with low-power computation, but their inaccurate circuit parameters induced by noise and low precision degrade accuracy

B. Digital

Digital Neural Network (NN) processors have advantages that they can achieve high accuracy, flexibility, and programmability, but they consume huge power and area due to the large amount of data transaction and fast operation speed

C. Mixed-mode implementations

Several mixed-mode SoCs took the advantages of both analog and digital implementations obtaining low-power consumption within small area, but it suffers from domain conversion overhead cost

8) In this paper, we provide a review of the design methodologies by introducing various processors with versatile design approaches for NN/FIS/NF/DNN acceleration

9) Paper organization:

A. **Section II:** related works about NN processor design including today's DNN and neuromorphic processors will be explained as previous works

B. **Section III:** review several NN/NF processors used as a functional block of an intelligent computer vision SoC. They are classified into three categories according to their design methodology: analog, digital, and mixed-mode implementation

- Each processor is dedicated to different functional blocks of the vision SoC pipeline, e.g. visual attention module, classification, workload prediction

C. **Section IV:** Recent DNN processors that incorporate with fully-digital circuit implementation is explored

D. **Section V:** Neuromorphic Processors

E. **Section VI:** Insights and perspectives on future research directions

F. **Section VII:** Conclusion

SECTION II. RELATED WORKS – NEUTRAL NETWORK PROCESSORS

1) Many researchers investigated on developing DNN and neuromorphic SoCs recently

A. Neuromorphic Designs

- Lu [22] Jan 2015 implemented clustering algorithm in analog domain with floating-gate non-volatile memory
- Zhang [23] Feb 2015 implemented matrix-multiplying ADC that enables multiplications with input samples, which is used for feature extraction in classification algorithm
- Kim [24] 2014 proposed sparse coding ASIC to enable training of sparse representation of images for feature detection and recognition using spiking neural network
- Kim [25] 2015 also developed a simple object re cognition system that is

composed of spiking neural network inference module

- Lee [26] Feb 2016 designed energy-efficient matrix multiplier with switched capacitor scheme for classification applications on analog front-end
- Ambrogio [27] Jun 2016 used resistive switching memory, so-called RRAM, to emulate the function of spiking neural network
- Zhang [28][29] proposed in-memory computation scheme using standard 6-T SRAM array Jun 2016

B. DNN Processor Design

- Tsai [30] Jun 2014 proposed a digital DL processor for big-data applications such as data filtering and data estimation kernels
- Park [31][32] Feb 2015 attempted to implement DNN training on silicon with scalable architecture and massively parallel thread-level parallelisms. They expanded their work to develop a user experience glass system using embedded DL SoC [33]. Feb 2015 It is capable of both simple NN training and inferencing
- Lee [34] Feb 2016 developed an SoC for advanced driver assistance system using RNN and FIS accelerators for energy-efficient automotive applications This work also supported simple on-line learning of RNN using dedicated MAC processing elements (Pes) with SIMD extension cores
- Chem [35] Feb 2016 proposed a CNN accelerator utilizing data-reuse pattern and its dedicated hardware architecture
- Sim [36] Feb 2016 designed a digital CNN processor with multi-range MAC unit and kernel compression scheme to reduce off-chip memory access
- Moon [37] Jun 2016 designed a CNN processor with 2-D MAC array and scalable bit precision, and enhanced the energy-efficiency with zero guarding and wide-range voltage-frequency scaling [38]
- Knag [39] June 2016 proposed a convolution restricted Boltzmann machine processor to infer support vector machine classifier with integrated sparse convolution unit
- Bang [40] Feb 2017 and Whatmough [41] introduced efficient accelerators for fully-connected (FC) DNN, with their hardware architectures optimized for matrix

multiplication

- Desoli [42] Feb 2017 integrated low-power DNN SoC with other hardware components such as DSP or dedicated direct memory access
- More works were exposed to integrate digital circuits for different type of DNNs from FC-DNNs to CNN and RNN [43] – [48] Feb 2017 – Feb 2018
- The trend of recent design techniques show high-performance DNN SoCs are implemented in digital while ultra-low-power SoCs are mixed-mode implementation

SECTION III. NEURAL NETWORK / NEURO-FUZZY PROCESSORS AS A FUNCTIONAL BUILDING BLOCK OF SYSTEM ON CHIP

There have been many researchers to deploy NN and NF algorithms as a functional block of a machine vision system. The details of such building block processors are introduced

1) Fully Analog Implementation

- There were several attempts to analog circuit implementation of NN and spiking neural network (SNN)
- These fully-analog circuit implementations were proposed for functional blocks used in NN/NF such as synapse or sigmoid activation generation
- Analog circuit lacks of programmability while NN/NF operations requires training and setting many parameters by nature
- Most of analog-based neural networks were assisted by digital circuits for flexible control of the analog arrays
- Full integration of NN/NF could be classified as mixed-mode and digital designs

2) Mixed Mode Implementation

- Most of mixed-mode processors utilized analog circuits for feedforward operation of NN and NF due to tis low power consumption
- Analog design saves area since analog multipliers are usually smaller than digital multipliers

- Current-mode analog circuit employs massively parallel architecture with simple current summation based on Kirch-hoff's current law, therefore, no additional adder circuit design is required
- Although voltage-mode circuits have some advantages over current-mode circuits, it requires both of multipliers and adders for NN/NF operation. Thus, current-mode circuits are preferably designed in many cases
- Digital circuits are used for training and controlling the analog parameters since they provide high programmability and accurate calculations with high bit precision
- First NF mixed-mode circuit for object detection [49] was designed as a functional block of the whole object recognition SoC [50], to achieve high performance and low power with small area overhead. It is fabricated in 0.13 um CMOS technology and reduced area and power by 59% and 44% compared with the fully-digital implementation in the same process technology, respectively.
- A versatile adaptive neuro-fuzzy inference system (VANFIS) hardware is designed for multiple purposes of classification [52]. The VANFIS was used for object classification and dynamic workload prediction to increase energy efficiency of the object recognition SoC by adapting to different types of input vector. Implemented with current-mode circuits, the mixed-mode VANFIS processor that is fabricated in 0.13 um CMOS technology saved area and power by 56% and 85%, respectively, compared with the equivalent digital implementation [53]
- Another mixed-mode Intelligent Reconfigurable Integrated System (IRIS) SoC is introduced for multi-purpose application of NN and FIS [55] The mixed-mode design reduced power and area by 71.2% and 54% compared with equivalent digital design; the SoC fabricated in 0.13 um CMOS technology achieves 1mJ/frame energy efficiency and consumes 57 mW on average for object recognition
- RBFNN is widely used as a classifier for its high accuracy [57], a mixed-mode RBFNN classifier deploying current-mode circuits is proposed for low-power yet highly-accurate scene classification in [58]. To make the mixed-mode processor tolerant to noise, the proposed RBFNN classifier contains temperature and supply voltage variation compensation circuits, which outputs stable current despite the variations. Fabricated in 0.13 um CMOS technology, saves area and power by 84% and 82%, respectively, compared with fully digital implementation

- Another NF processor with RNN-FIS is proposed for automotive Black Box in [61] The algorithm alerts rivers to the risky objects that are about to the collided with the vehicle in driving-mode while it triggers surveillance recording when object is getting closer to harm the vehicle in parked-mode to extend life-time of recording with limited capacity of battery. The SoC is fabricated in 65 nm CMOS technology and achieves high performance (502 GOPS) in driving-mode and lower power in parked-mode. Thanks for the mixed-mode implementation, the total area and power consumption are reduced by 64% and 39% respectively, compared with the fully-digital implementation

- Martinez [82] designed mixed-mode neural network by proposing digitally programmable multipliers for linearization of the NTC thermistor. The processor designed with 0.18 um CMOS technology consumes 0.538 mW with 100 MHz

- Larras [86] proposed a mixed-signal IC for encoded binary neural network. The chip consists of analog neuron computing nodes for lower power and cost design with digital network for communication among the nodes. The chip is fabricated in 64 nm CMOS technology and achieves 68 fJ ultra-low-energy operation per each synaptic operation.

- Although analog/mixed mode design facilitates low-power and energy efficient designs [88], it is often complicated due to PVT variation as well as the domain conversion overhead in terms of conversion speed, area, and power. High end ADCs with fast sampling rate [94]-[96] consumes ~10mW under ~10,000 um^2 area. Since digital domain runs with ~250 MHz, ADC could be the bottleneck in operating speed, thus, analog and digital domains must be carefully divided and balanced

3) Digital Implementation

- Although mixed-mode design has advantages in low power and small area implementation, digital processor has advantages in its speed and high precision as the technology gets smaller in addition to the removal of data conversion overhead in image processing

- Cellular neural network processor [62] is designed for visual attention in the brain-inspired object recognition pipeline. The hardware architecture shown in Figure 9 combines the flexibility of digital approach with high performance of fully parallel

cell topology of analog approach. The processor is implemented in 0.13 um CMOS technology and occupies 4.5 mm². It consumes 84 mW running at 200 MHz. With the help of processor, incorrect local features are drastically reduced to increase frame rate by 83% and reduce energy/frame by 45% without degradation in recognition accuracy

- In another aspect of applying NN as a functional block, neural network task scheduler [66] is designed for workload prediction to enhance energy efficiency by reducing time overhead on core-to-core allocation in a multi-core vision processor [67] Many multi-core systems today adopt network-on-chip as their communication fabric for fast and parallel data transmission, but data transaction delay caused by network congestion directly degrades energy efficiency because processing cores must run faster to meet the overall system latency. The neural network task scheduler contains a RISC scheduler and an 8-way neuron array that is capable of reconfigurable precision. The processor is fabricated in 65 nm CMOS technology. It consumes 4.9 mW while achieving 12.7 mJ/frame energy efficiency by reducing 24.4% of network latency on average. To enhance the performance, an advanced version of neural network scheduler using RNN (Figure 12) with new network-on-chip architecture that is dedicated to SIFT-based object recognition [68] is proposed [69]. It is also designed as a part of the object recognition SoC [64]. It improved the workload prediction accuracy to 91.4% and system throughput by 50.2%

SECTION IV. DEEP LEARNING / DEEP NEURAL NETWORK SYSTEM-ON-CHIPS

The processors explained in Section III imply that utilizing NN/NF algorithms as a functional block of SoC brings great improvements if they are used in right place with dedicated hardware architecture. Now recent SoC implements for accelerating the entire DL/DNN algorithm will be explored.

1) TPU by Goolge [44]

- Targeted for general DNN acceleration, it was not suitable for mobile applications due to its huge power consumption (>40W) and memory footprint (>28 M-Byte)

2) DNN Processing Unit (UNPU) [70]

- The DNN processing unit (DNPU) [70] was introduced to accelerate both CNN and RNN in mobile environment with following features to optimize both CNN and RNN on a single chip:

  1. Heterogeneous multi-core architecture
  2. Mixed channel division scheme for CNN computation
  3. Quantization-table (Q-table) for efficient matrix product

- The heterogeneous multi-core architecture of DNPU (Figure 13) aims at fulfilling the two different peculiarities. The CNN core is dedicated to maximize data reusability with high computation performance over 300 GOPS, and the RNN core was designed to minimize the amount of external memory access with algorithmic support and dedicated PEs. The SoC is fabricated in 65 nm CMOS technology and achieves 8.1 TOPS/W at 50 MHz, 0.77 V where peak power is 279 Mw

3) Unified Neural Processing Unit (UNPU)

- However, heterogeneous architecture of DNPU limited hardware utilization. For example, if a DL application requires only one of CNN or RNN, the unused part of the DNPU wastes hardware resources. The unified neural processing unit (UNPU) was presented [72] to resolve the limitations of the DNPU by obtaining high hardware utilization on energy-optimal point. It was designed with new microarchitecture with the following features:

  1. Unified DNN core architecture
  2. Fully-variable weight bit precision

  As described in [72], feature map becomes dominant when the weight bit precision is reduced on the accuracy-energy optimal point, so the UNPU shown in Figure 15 utilizes feature map reuse datapath while DNPU deal with weight reuse. This unified data-path enabled to combine the separate CNN and RNN cores of the DNPU into a single DNN core, which achieved 1.15x and 13.8x higher peak-performance with the same hardware footprint.

4) CNN Processor (CNNP)

- Contrary to ASICs for general purpose DNN acceleration, the CNN Processor (CNNP) introduced in [78] targeted user identification application on always-on sensor node, which essentially requires highly accurate face recognition with low-power consumption.
- To satisfy the requirement, the CNNP proposed two main features for low power operation:

  1. Distributed memory CNN architecture
  2. Separable filter approximation

  As shown in Figure 17, it utilized distributed memory architecture without separate global routing and memory to lower routing complexity, and brings the external memory into multiple local on-chip memories for lower power CNN operation
- To reduce additional power consumption, the CNNP utilized separable filter approximation technique that estimates the 2D convolution with two cascaded 1D vertical and horizontal convolutions as depicted in Figure 18.
- The CNNP is fabricated in 65 nm CMOS technology and achieved 5.3 mW

5) CNN-Stereo Engine (CSE)

- CNNP suffers from poor hardware utilization when the size of input feature map is mall or when the image size assigned to the core is not multiples of the number of words that is computed at once in the core. It is not sufficient for stereo matching CNN algorithm without pooling layers [81] of which the image size gets smaller in deeper layers, leading to degradation in hardware utilization. Therefore, the CNN-Stereo Engine (CSE) [80] is published to maximize the hardware utilization with two main features:

  1. Channel-wise parallel 1-D MAC operation
  2. Core-to-Core data balancing
- The CSE increased parallelism in the channel direction rather than spatial direction, and it was designed to handle 1-D operations to remove the redundancy caused by the 2D workload assignment method of the CNNP

- The overall hardware utilization is improved by reducing the number of words computed at once. The CSE hides both the accumulation latency among multiple channels and data fetch latency by pipelining
- It achieved 20% higher energy efficiency than the CNNP. Moreover, the FWD/BWD units support exchanging adjacent boundaries of two local feature maps to guarantee the balanced workload allocation (Figure 20). This operation was hidden behind CNN pipeline and improved the overall performance by 23.9%

Section V. NEUROMORPHIC PROCESSORS

Neuromorphic processors that aim at ultra-low-power consumption by mimicking neurons are investigated along with the advance of DNN accelerators. The representative processors would be Neurogrid [84], IBM's TrueNorth [99] and Intel's Loihi [100].

1) Neurogrid

- Mixed mode multichip system of which neuron array is designed in analog while spike TRX and memory are designed in digital with 0.18 um CMOs

2) TrueNorth

- TrueNorth implemented 4,096 neurosynaptic cores of SNN that are connected via scalable routing network as shown in Figure 21
- Each core consists of 256 digital axons and neurons which are connected by 256 x 256 crossbar for dendrite design
- The large-scale chip (4.3 cm^2) fabricated in 28 nm CMOS was applied to multiobject detection with 63 mW with 400 x 240.30 fps video

3) Loihi

- Also is fully digital architecture implementing 128 cores of 1,024 SNN units and supports sparsity as its performance is evaluated with LASSO optimization
- The chip has a high level of completion that is also contained three x86 cores for message control for the neuromorphic cores

- The shole SoC is implemented in 14 nm FinFET technology and occupies 60 mm²
- Although their die area were very large, both TrueNorth and Loihi are energy efficient consuming only ~10mW

4) Time domain Neural Network Processor

- Miyashita [101] designed time domain neural network processor in part of Binarized NN with analog MAC circuits in CMOS technology
- Employ differential signaling to prevent precision degradation from PVT variation
- The chip fabricated in 65 nm CMOS technology achieved high energy efficiency of 12.9 fJ per synaptic operation

5) MLP Processor with ReRAM Crossbar

- Some neuromorphic processors make use of nonvolatile memory (NVM) such as ReRAM and MRAM in their analog arrays to replace floating-point MAC with parallel analog processing array [83], [85], [102]
- The conceptual diagram of analog computing is depicted in Figure 22
- It was capable of limited MLP application with only 10 hidden neurons because number of neurons must fit to the array size
- The analog accelerators show significant performance improvement than digital processors
- According to [102], analog processor achieved 270x energy, 1040x latency, and 1.8x area than digital ReRAM accelerators; 430x, 34x, and 11x compared to SRAM-based digital accelerators, respectively
- However, it is important to ensure that implementation details and material properties of the NVM must be aligned with the requirements of NN algorithms [85]
- These processors still have problems of their limited functionality. Most of the works were capable of only simple classification tasks with MNIST or CIFAR-10 dataset while DNN/DL SoCs provide higher-level intelligence with much complex dataset

SECTION VI. DISCUSSION AND FUTURE DIRECTION

Various circuit design approaches toward NN/NF/DL were explained in this paper.

1) Mixed Mode NN/NF Processors

- **Power Reduction**

  - Table I shows the summary table of the mixed-mode NN/NF processors, each of which was state-of the art designs by the time they were proposed
  - They employed current-mode mixed-mode circuits for area and power reduction compared with the fully-digital implementations
  - However, the amount of reduction becomes less significant as the technology node goes down to nm
  - The average power reduction using 0.13 um technology was 70%, while that of 65 nm technology was only 40%. This is because the analog circuit cannot be scaled down with the same ratio of technology scaling
  - But this does not mean that mixed-mode circuits no longer have advantages over digital circuits. It still can take the advantage of having smaller but flexible digital controller as well as utilizing natural parallelism of circuit operations
  - Also, analog cores can be reconfigurable with some design techniques to provide versatile functions even through it can be less accurate than digital computations
  - By having shared datapath to ADC with the digital learning controller, the RNN-FIS processor [61] can handle four different NN/NF topologies. This shows that analog circuits can also be programmable assisted by digital controller

- **Domain Conversion Overhead Cost (Speed, area, power)**

  - However, analog and digital domains must be carefully divided in mixed-mode design to reduce the domain conversion overhead cost in terms of conversion speed, area, and power
  - Overhead can be resolved when the processing comes into sensors and computes with analog data before A-D conversion

    1. In [103], pupil edge detection and glint corner detection circuits were deployed into CMOS image sensor, each of which circuit configures pixel array

to extract edge in analog voltage and compares charge difference, respectively

2. In [104], matrix multiplication was integrated into ADC that multiplies analog input with digital signal to reduce additional energy required for A-D conversion

3. Kim [105] designed an image sensor dedicated to stereo matching. Its reconfigurable pixel array computes sparse image rectification to align two different inputs and census transformation is executed in analog domain with simple switch network and comparators

- Since those values are computed with analog signal before ADC the digital processor integrated in the sensor excluded additional blocks for rectification and census transformation

- All of the works achieved high energy efficiency by putting computations into the sensors, and showed new design paradigm

2) Digital NN/NF Processors

- Table II shows the summary of digital NN/NF processors as a functional block of the machine vision systems

- They reduced the amount of overall computation with visual attention and improved system throughput with workload prediction. It is noticeable that taking NN/NF functions for specific purpose reduces energy efficiency of the whole machine vision SoC, significantly

- This implies that the processor does not always have to be dedicated for accelerating end-to-end DNN algorithm but applying small NN/NF system to right places with dedicated hardware architecture for that specific purpose brings performance enhancement

3) DNN SoCs

- **Power Consumption**

- Although the majority of computing engine for general-purpose DNN acceleration are traditional CPU and GPU, the consume too much power. For example, AlexNet

running on Intel Xeon CPU and NVIDIA Titan X consumes 130 W and 250 W [97]

1. Even one FC layer of AlexNet consumes lots of power
    i. 73 W with Intel Core-i7 CPU
    ii. 159 W with Titan X GPU
    iii. 5.1 W with Tegra K1 mobile GPU [98]

- In contrast, ASIC designs reduce power consumption to the order of mW, which makes ASICs suitable for mobile and embedded system
- Table III shows the summary of the DNN SoCs
    1. AlexNet running on [72] consumes only 290 mW at 346 fps maximum, while [48] consumed 2.7 W with 76.7% accuracy

- **Design Challenges**
    - **Power Consumption**
    - There are several design challenges in DNN SoC. Although they are designed for low power consumption, memory bandwidth becomes crucial problem.
    - The processors [38], [43], [47], [79] tried to deploy data locality for maximizing on-chip data reuse in order to achieve lower power yet high throughput
        1. Moons [38] and Chen [79] brought with systolic array to reuse data within Pes as many as possible
        2. Shin [43] and Lee [47] utilized 2-D MAC array with dedicated logics for data management and scheduler to maximize data reusability
        3. They are on the basis of global buffer to filter the amount of external memory bandwidth that involves with global routing and aggregation path between global buffer and parallel PEs
        4. This approach requires a large memory bandwidth and complex global routing to high parallelism, which is not suitable for low-power consumption
    - So [78] utilized distributed memory architecture without global routing and memory to lower routing complexity and resolve memory bandwidth bottleneck
        1. Such distributed memory architecture extends to Processing-in-Memory (PIM)

architecture

- Kang [106] employed deep in-memory architecture that computes in the periphery of the memory cell array

    1. This technique minimizes the costs of data access and processing similar to the concept of mixed computing in sensors
    2. Data stored in SRAM bitcell are computed and selected before passing through ADC
    3. The processor achieved 4.9x energy saving and 2.4 throughput improvement compared to a conventional von-Neumann architecture

- **Energy efficiency**

- In [107], results of census transform are stored in stereo-SRAM and they are XORed in the peripheral between bitcell arrays to generate costs before propagated to digital block
- In [108], Yang inserted pulse width modulation and shift memory into the bitcell array to efficiently accelerate binary-weight network
- In [109], Biswas utilized 1-b multiplication and average operation in analog domain right next to the 10T SRAM cell array to accelerate binary weight network.
- Yin proposed SRAM macro that computes ternary-XNOR-and-accumulate operation
- These approaches achieve significant improvement in energy efficiency. But the architecture is very primitive that they could only support small neural network with limited dataset

- **External Memory Bandwidth**

    - External memory bandwidth is another bottleneck as on-chip data reuse technique advances
    - 3D stacking technique with DNN accelerators as reported in [48] will bring immense improvement in external bandwidth overhead

- **Dependency on DNN architecture and accuracy**
  - Unlike CPU and GPU, the DNN SoCs introduced so far are dedicated to specific system and application which induces dependency on DNN architecture and accuracy
    1. In [78], target application was always on face recognition for user identification where the number of output class is less than ImageNet dataset, or the task is simpler.
    2. This fact implies that small CNN architecture is enough to achieve 97.4% of accuracy while only 0.2% degradation is not crucial for the application
    3. Also, SoCs utilized reduced bit precision to resolve computational complexity that directly harms the accuracy even a small amount
  - GPU and TPU are preferable than the ASICs introduced in this paper is programmability of DNN algorithm is versatile or the target application requires precise computation for high accuracy such as autonomous driving

- **Technology Scaling**
  - Thanks to the technology scaling, we foresee the future researches on hardware accelerators for high performance DNN applications will be digital to maintain high throughput and precise computation and it will bring integration of higher-level intelligence with complicate DNN architectures with energy-efficient digital implementations

- **Ultra-low-power Applications**
  - Neuromorphic processors will be investigated for ultra low-power applications with simple intelligence by utilizing energy-efficient analog computations

- **Reliance on Development of New Technologies**
  - Many neuromorphic processors highly rely on the development of new technologies such as ReRAM or MRAM
  - These technologies are still being researched but its time-to-market until mass

production is unpredictable

- In the meantime, it is believed that mixed-mode implementations and PIM architectures of neural network processor that uses current CMOS technology is promising for ultra-low-power DNN applications until the neuromorphic devices become in mass production

SECTION VII. CONCLUSION

A. This paper provides a review of dedicated processors for neural network, neuro-fuzzy system, deep learning and deep neural network processors in various design approaches of analog, digital, and mixed-mode implementation over decades

B. Provide both mixed-mode and fully-digital implementations of neural network / neuro-fuzzy processors that are deployed as a functional building block of machine vision systems

C. Provide the comparisons upon the different design approaches

D. The most recent designs of deep learning processors were introduced

E. Neuromorphic chips are highly dependent on the development of new technologies, but no one can assure how long it will take for the technologies to be in mass production

F. Before then, it is believed that mixed-mode implementation of neural network processor is promising for ultra-low-power applications while digital implementation will fit for high-performance applications