Course Design of Deep Learning Architecture

Kuo-Kun Tseng

kktseng@hit.edu.cn

Harbin Institute of Technology, Shenzhen

School of Computer Science and Technology Shenzhen, China

Abstract—Deep learning brings good prospects for the development of AI, but the complexity and slow operation of deep learning algorithms hinder the development and application of deep learning in AI. So we designed a deep learning architecture course to train people in this area. In addition to imparting relevant theoretical knowledge, this course will also arrange related deep learning architecture design experiments to train talents with deep learning algorithm core knowledge and architecture design ability for future research and development.

Keywords: deep learning architecture; tensor processor; FPGA, SOPC

1 INTRODUCTION

Deep learning has brought development prospects for the development of artificial intelligence, but the complexity and slow operation of deep learning algorithms hinder the development and application of deep learning in artificial intelligence. This course program allows students to understand the core architecture of deep learning and learn to design the hardware architecture of deep learning. The objectives of this course are listed as follows:

- Let the students understand the hardware and software development methods and principles;
- Let the students understand the relevant algorithm design principles of deep learning;
- Let the students understand the deep learning software and hardware architecture design principle.

This course includes theoretical courses and experimental courses, this course will arrange relevant deep learning hardware design experiments to enable students to cultivate the design ability to transform deep learning algorithms into hardware architecture. In a summary, this course should have the following features:

- Innovative courses: This course should be one of the innovative courses. The traditional "computer architecture" can no longer meet the current needs of artificial intelligence technology. This course will combine the latest deep learning technology underlying

ITEI 2022, December 23-25, Harbin, People's Republic of China

DOI 10.4108/eai.23-12-2022.2329194

optimization technology and the new concept of future computer architecture, and design a new characteristic "deep learning architecture " course.

- Theory and experiment teaching: In addition to the teaching of relevant theoretical knowledge, this course will also arrange deep learning architecture design experiments for nearly half of the course hours, so as to cultivate talents with practical ability for future relevant research and development.

- Software and hardware co-design: Cultivate students' comprehensive ability of software and hardware collaboration, so that students can better master the core technologies of AI chip and system.

- This course can provide the remote experiments, which allow school to reduce the damage rate of the experimental board and improve the utilization rate of the experimental board.

The prerequisite courses for students are Python and Digital System Design. This course has also accumulated several years of teaching experience, which can better optimize the course content.

2 RELATED WORK

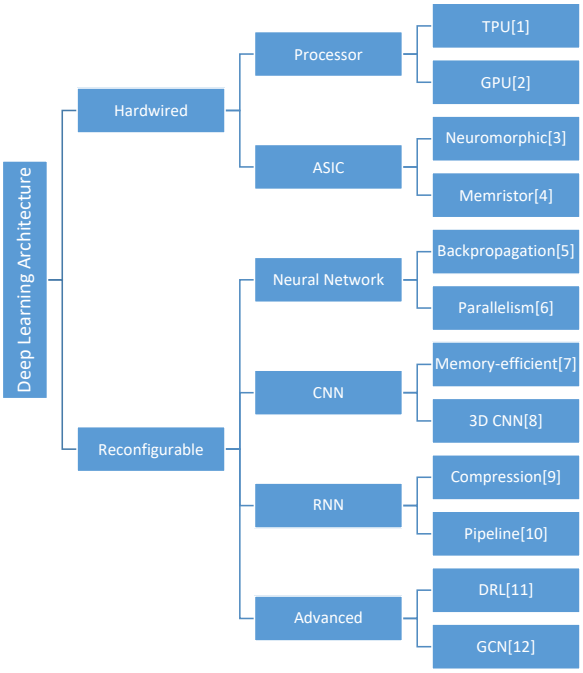

Figure 1 The related researches of this course.

As shown in Figure 1. The first related research branch is Tensor Processing Unit (TPU) [1] and Graph Processing Unit (GPU) [2], and Neuromorphic chip [3] and memritor [4] of

improved hardware underlying components. This branch studies of the generic and underlying elements is not the focus of this course.

The most related work of this course is the architecture design of deep learning algorithm with reconfigurable FPGA platform. we can said that this course focuses on the architecture optimization of the specific deep learning algorithm.

Thus, this related research branch on the implementation of neural network, including the hardware of backpropagation [5] function, with the algorithm of parallism design [6]. Then the next related sub-branch is the convolution function of deep learning, so there are many studies of CNN hardware, including the study of CNN architecture of Memory-efficient[7], and the [8] study of hardware acceleration of 3D CNN. Meanwhile, the hardware of RNN algorithm is also a lot of research, mainly concerned with the compression problem [9] and pipeline issue [10].

Moreover, in recent, the advanced deep learning algorithms are implemented in FPGA as well, such as the hardware of Deep Reinforcement Learning (DRL) [11] and Graph Convolution Network (GCN) [12] can now be implemented.

From these surveys, we can know that the hardware of deep learning algorithms and their architecture optimization are being studied by many researchers, so there should be a great demand for courses in this research area.

3 COURSE DESIGN

The design of this course is shown in Figure 2, which consists of two parts: theoretical course and experimental course. At the beginning of the course, we will first introduce application and the PYNQ development board, followed by the High-level Synthesis (HLS) development language. Then explain the principles and hardware design methods of basic networks such as Neural Network (NN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN).

Finally, we will arrange YOLO object detection, ECG diagnostic system, driverless car system and federated face recognition system and other application courses, so that students can learn about actual product development.

Starting from the second unit, each lesson unit can have corresponding experiments, and teachers can also choose the class and experimental content according to their own class hours, such as the last three units are advanced courses, and general introductory courses can choose not to attend.

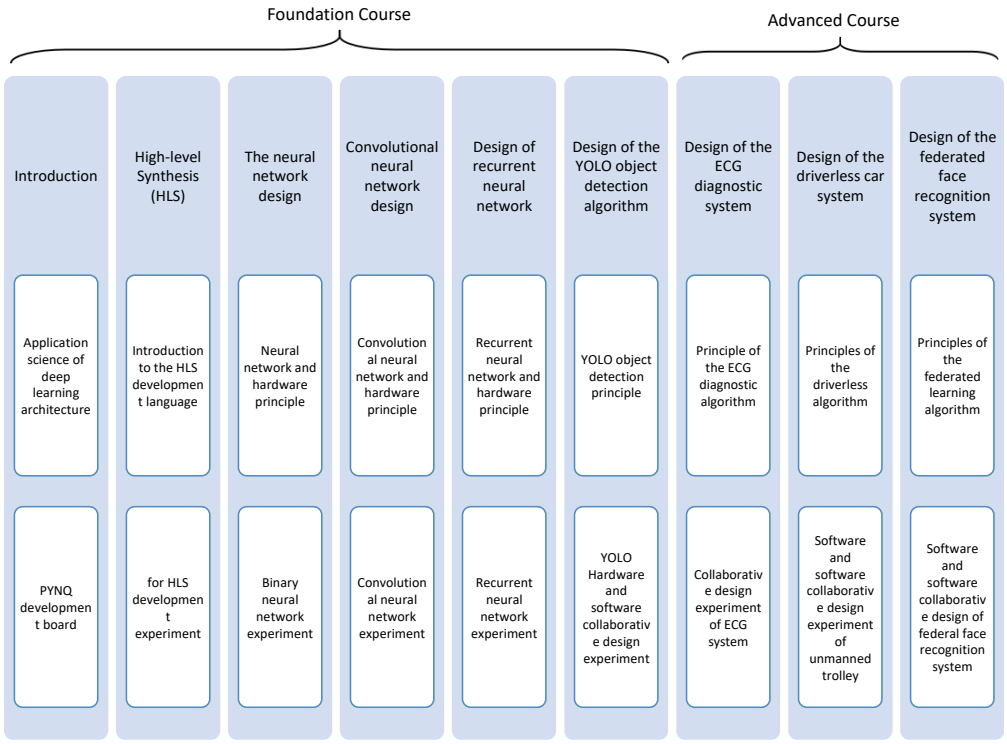

Figure 2 The course outline of deep learning architecture

4 EXPERIMENT DESIGN

For the combination of low-order and high-order dual-platform experiment board, teachers and students should choose low-order PYNQ-Z2 application experiments or high-order PYNQ-ZU applications according to their own funds and course needs. PYNQ-Z2 is cheap but FPGA has less capacity and a slow speed, while PYNQ-ZU has a faster relative capacity and speed, which is suitable for advanced experiments.

Students can use computers to remotely connect to the PYNQ-Z2 or PYNQ-ZU platform through the Internet, which can reduce the damage of the experimental board and improve the utilization rate of the experimental board.

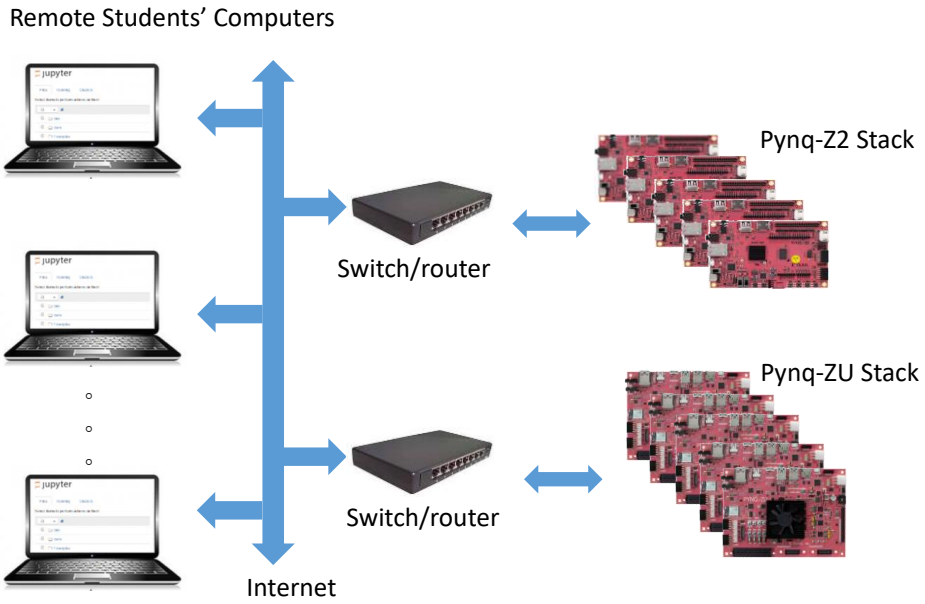

Figure 3 The Experiment setup of deep learning architecture

5 DISSCUSION

5.1 Differences from the deep-learning courses?

There are obvious differences between this course and the deep learning course. This deep learning architecture course focuses on allowing students to understand the core architecture of deep learning, and cultivating the architecture design ability of transforming deep learning algorithms into hardware and underlying software. The key differences are as follows:

- The experimental courses conducted in this course are combined with deep learning courses. All experiments were related to FPGA, and the previous deep learning course was a general CPU / GPU.
- This course will take up a considerable amount of time to introduce the FPGA HLS hardware language and development platform, but no deep learning course.
- This course will take up a considerable amount of time to teach deep learning hardware technology, not deep learning courses.

5.2 Why not use GPU, TPU and AI chip for experiments ?

We chose SOPC FPGA as a development platform for three main reasons:

- GPU, TPU and AI chip are often general-purpose architecture processors, if students use these as experimental platforms, students usually only learn the software instruction

optimization that controls these chips, and cannot advance to the low-level architecture optimization.

- Students can convert software algorithms into hardware functions more straight, and run faster.

- If the designed algorithm has a wide range of applications, it is convenient to convert the FPGA design to ASIC design, and if the application amount is not large, you can still use the SPOC chip to directly apply.

6 CONCLUSION

This article introduces the course design of deep learning architecture, and although deep learning has brought development to artificial intelligence, it has now also encountered a bottleneck in speed. If the algorithm architecture of deep learning can be optimized by combining software and hardware, it should bring better prospects to deep learning algorithms.

This course is properly planned to allow students to understand the core of deep learning architecture and allow students to conduct relevant deep learning algorithm development experiments, which should cultivate more outstanding talents in the field of artificial intelligence.

REFERENCES

[1] Pandey, P., Basu, P., Chakraborty, K., & Roy, S..(2020).Greentpu: predictive design paradigm for improving timing error resilience of a near-threshold tensor processing unit.IEEE Transactions on Very Large Scale Integration (VLSI) Systems, PP(99), 1-10.

[2] Pal, S., Ebrahimi, E., Zulfiqar, A., Fu, Y., Zhang, V., & Migacz, S., et al.(2019).Optimizing multi-gpu parallelization strategies for deep learning training.IEEE Micro, 39(5), 91-101.

[3] Miyashita, D., Kousai, S., Suzuki, T., & Deguchi, J..(2017).A neuromorphic chip optimized for deep learning and cmos technology with time-domain analog and digital mixed-signal processing.IEEE Journal of Solid-State Circuits, 52(10), 2679-2689.

[4] James, A.P..(2019).A hybrid memristor–cmos chip for ai.Nature Electronics, 2(7), 268-269.

[5] Gadea, R., Cerda, J., Ballester, F., & Macholi, A..(2000).Artificial neural network implementation on a single FPGA of a pipelined on-line backpropagation.International Symposium on System Synthesis (pp.225-230).

[6] Yu, J., Lukefahr, A., Palframan, D., Dasika, G., & Mahlke, S..(2017).Scalpel: customizing dnn pruning to the underlying hardware parallelism.ACM SIGARCH Computer Architecture News, 45(2), 548-560.

[7] Li, G., Liu, Z., Li, F., & Cheng, J..(2022).Block convolution: toward memory-efficient inference of large-scale cnns on fpga.IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems: A publication of the IEEE Circuits and Systems Society(5), 41.

[8] Shen, J., You, H., Wang, Z., Qiao, Y., & Zhang, C..(2018).Towards a Uniform Template-based Architecture for Accelerating 2D and 3D CNNs on FPGA.the 2018 ACM/SIGDA International Symposium.ACM.

[9] Han, S., Kang, J., Mao, H., Hu, Y., Li, X., & Li, Y., et al.(2016).ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA.Acm/sigda International Symposium on Field-programmable Gate Arrays.ACM.

[10] Bank-Tavakoli, E., Ghasemzadeh, S.A., Kamal, M., Afzali-Kusha, A., & Pedram, M.. (2019). Polar: a pipelined/overlapped fpga-based lstm accelerator.IEEE Transactions on Very Large Scale Integration (VLSI) Systems, PP(99), 1-5.

[11] Laserna, J., Otero, A., & Torre, E..(2022).A Multi-FPGA Scalable Framework for Deep Reinforcement Learning Through Neuroevolution.Springer, Cham.Springer, Cham.

[12] R Hungjosé, LiChao, WangPengyu, ShaoChuanming, GuoJinyang, & WangJing, et al. (2021). Ace-gcn: a fast data-driven fpga accelerator for gcn embedding.ACM Transactions on Reconfigurable Technology and Systems (TRETS).