# Convolutional Neural Networks & Specialized Hardware Accelerators

## Introduction

Deep neural networks like convolutional neural networks (CNNs) have revolutionized computer vision and image recognition tasks. However, neural network algorithms are computationally intensive. This has led to new deep learning accelerators designed to reduce power consumption and latency [1]. This technical paper is a review of the state-of-the-art CNN algorithm, the applications of CNNs, the underlying technology of hardware accelerators for CNNs, and some commercially available hardware accelerators for CNNs.

## State-Of-The-Art CNN Algorithm

Current CNN algorithm consists of an input layer, an output layer, and three major layers in the middle: a filtering layer, a ReLU layer, and a pooling layer. The input layer is fed with a class of image matrices. In the filtering layer, a filter matrix with a specific visual feature is convolved with the image matrices. The convolution process emulates the response of an individual neuron to visual stimuli in biology [2]. The output of the filtering layer is then fed into the ReLU layer, which applies rectified linear units to the image matrices. The ReLU layer trains the neural network several times faster without losing accuracy [3]. The output of the ReLU layer is then fed into the pooling layer, in which the input image matrices are partitioned into a set of non-overlapping rectangles. For each rectangle, the pooling layer outputs the maximal value in that region. The pooling layer reduces the size of the image, the number of parameters, and the amount of computation in the neural network [4]. The output of the pooling layer is then fed into the output layer, which is a fully connected layer. This layer outputs the probability that each image has the same visual feature as the filter [5].

## Applications of CNNs

Compared with traditional methods like multilayer perceptions (MLPs), CNNs gives better accuracy and boosts the performance of the system due to unique features like shared weights and local connectivity [6]. Therefore, CNNs are widely used in computer vision and image recognition. Currently, the security industry is using CNN techniques to design intelligent security cameras which can automatically detect suspicious activities. In this way, police department can decrease the need for human supervision in surveillance systems [7]. In addition, CNNs are used in drug discovery. In 2015, the company Atomwise introduced AtomNet, the first CNN-based method for drug design. AtomNet integrated small chemical features into large chemical structures and produced drugs for multiple diseases such as Ebola and sclerosis [8]. Moreover, CNNs have also been used in natural language processing tasks such as prediction, classification, query retrieval, and semantic parsing.

**Underlying Technology of Hardware Accelerators for CNNs**

Recently, researchers and tech companies have investigated and built special purpose hardware accelerators for CNNs to decrease execution time and power consumption [9]. The tensor processing units (TPUs) designed by Google reveal the state-of-the-art technology in current deep learning accelerators. A TPU's prototype is an 8-bit engine for matrix multiplication. It is manufactured on a 28 nm process with a die size smaller than 331 $mm^2$. The clock speed is 700 MHz with a thermal design power of 28 to 40 W. It has 28 MB on the chip memory. There are 4 MB of 32-bit accumulators taking the results of a systolic array of 8-bit multipliers. The current generation of TPU supports floating point arithmetic and further accelerators the process of activation, matrix multiplication, and convolution in CNNs by adding 16 GB of high bandwidth memory to increase bandwidth to 600 GB/s and performance to 45 TFLOPS. The TPU is arranged into four chip modules with a performance of 180 TFLOPS. 64 of these modules are assembled into 256 chip pods with 11.5 PFLOPS of performance [10].

**Commercially Available Hardware Accelerators for CNNs**

Google's TPUs are currently proprietary and not commercially available. In industry, most commercially available hardware accelerators for CNNs are graphics processing units (GPUs). Over 60% of them are manufactured by Nvidia [11]. The Tesla K40 designed by Nvidia is used to facilitate neural networks for embedded systems, game platforms, mobile phones, personal computers, and workstations. Its cost ranges from $3000 to $3500, which is the lowest among Nvidia Tesla GPUs. Tesla K40 is also compact, lightweight, and energy-efficient. It is 26.7 cm long and 11.15 cm wide. Its weight is 826 g. Its power consumption is 235 W in the running state and 16 W in the idle state. Additionally, Tesla K40 gives full control to end-users to select the core clock frequency that fits their workload the best [12]. Therefore, users can choose a specific clock frequency for the core depending on the program they are running and the amount of computation in their CNNs. For a CNN with large input matrices, users may increase the clock frequency, consuming more power while decreasing the amount of time for computation.

There are other companies like AMD and ARM which also manufacture GPUs. Different brands of GPUs differ in their costs, performance, and functionalities. The cost of AMD Radeon GPUs ranges from $2000 to $2500, lower than Nvidia Tesla GPUs. In addition, AMD Radeon GPUs do not give users full control of clock frequency. Instead, it relies on a dynamic voltage and frequency scaling (DVFS) system to control the clock frequency on the core. Moreover, AMD Radeon GPUs are designed to run various types of neural networks such as CNNs, recurrent neural networks, and residual neural networks, while Nvidia's Tesla K40 is more specific to CNNs.

## References


[1] C. Lin. E6895. Class Lecture, Topic: "Advanced Big Data Analytics." Department of Electrical Engineering and Computer Science, Columbia University, New York, NY, Mar. 29, 2018.

[2] C. Gulcehre, "Convolutional Neural Networks (LeNet)," Dec, 2015. [Online]. Available: http://deeplearning.net/tutorial/lenet.html. [Accessed: Oct. 18, 2018].

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, pp. 3-4, 2012.

[4] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," *arXiv preprint*, pp. 2, 2015.

[5] A. Deshpande, "A Beginner's Guide To Understanding Convolutional Neural Networks," Jul. 20, 2016. [Online]. Available: https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/. [Accessed: Oct. 19, 2018].

[6] A. Bhandare, M. Bhide, P. Gokhale, and R. Chandavarkar, "Applications of Convolutional Neural Networks," *International Journal of Computer Science and Information Technologies*, vol. 7, no. 5, pp. 8, 2016.

[7] S, Kwiatkowski, "Deep Surveillance," Mar. 23, 2018. [Online]. Available: https://towardsdatascience.com/deep-surveillance-6b389abeaf95. [Accessed: Oct. 19, 2018].

[8] A, Heifets, I. Wallach, and M. Dzamba, "Systems and methods for applying a convolutional network to spatial data," U. S. Patent 9,373,059, 5 May., 2014.

[9] K. Kiningham, M. Graczyk, and A. Ramkumar, "Design and Analysis of a Hardware CNN Accelerator," *small*, pp. 1, 2017.

[10] P. Bright, “Google brings 45 teraflops tensor flow processors into its compute cloud,” May. 17, 2017. [Online]. Available: https://arstechnica.com/information-technology/2017/05/google-brings-45-teraflops-tensor-flow-processors-to-its-compute-cloud/. [Accessed: Oct. 20, 2018].

[11] H. Mujtaba, “NVIDIA Gained Major Discrete GPU Market Share With GeForce 10 Series GPUs Over AMD Radeon RX Series in Q2 2018,” Sep. 6, 2018. [Online]. Available: https://wccftech.com/nvidia-amd-discrete-gpu-market-share-q2-2018/. [Accessed: Oct. 20, 2018]

[12] Nvidia, “Tesla K40 GPU Active Accelerator,” Tesla K40 datasheet, Nov. 2013 [Revised Oct. 2014].