Análisis de desempeño de capas de CNN para arquitecturas heterogéneas basadas en FPGAs usando HLS [

2021

text (article)

Analítica

Context: Convolutional neural networks (CNNs) are currently used in a wide range of artificial intelligence applications. In many cases, these applications require the execution of the networks in real time on embedded devices. Hence the interest in these applications achieving excellent performance with low power consumption. CNNs perform operations between the input data and the network weights, with the particularity that there is no dependence between most of the operations. Thus, the inherent parallelism of Field Programmable Gate Arrays (FPGAs) can be used to perform multiple operations in parallel, maintaining the good performance per watt that characterizes these devices. This paper focuses on evaluating the convolution algorithm for a convolutional layer of neural networks by exploring parallelization directives using VIVADO HLS, and it aims to evaluate the performance of the algorithm using optimization directives. Method: The methodology consists of an exploration of the design space of a convolutional neural network layer implementation using VIVADO HLS. Performance verification of the FPGA was performed by comparing the output data with the same convolution algorithm implemented in MATLAB. A layer of the commercial version Xilinx DNNK was used as a reference for performance measurements of the different implementations obtained during the exploration of the design space. In this work, multiple variations of optimization directives are used, such as pipeline, array partition and unroll. Results: This paper presents the results of a reference implementation (without optimization directives) of the convolution algorithm concerning algorithm latency and the hardware resources used by the FPGA. The results are compared with the implementations of the algorithm, including different combinations of two optimization directives (pipeline and partition array). Conclusions: This work explores the design space of a convolution algorithm for a convolutional neural n

exploration of the design space of a convolutional neural network layer implementation using VIVADO HLS. Performance verification of the FPGA was performed by comparing the output data with the same convolution algorithm implemented in MATLAB. A layer of the commercial version Xilinx DNNK was used as a reference for performance measurements of the different implementations obtained during the exploration of the design space. In this work, multiple variations of optimization directives are used, such as pipeline, array partition and unroll. Results: This paper presents the results of a reference implementation (without optimization directives) of the convolution algorithm concerning algorithm latency and the hardware resources used by the FPGA. The results are compared with the implementations of the algorithm, including different combinations of two optimization directives (pipeline and partition array). Conclusions: This work explores the design space of a convolution algorithm for a convolutional neural n

---

Título: Análisis de desempeño de capas de CNN para arquitecturas heterogéneas basadas en FPGAs usando HLS electronic resource]

Editorial: 2021

Tipo Audiovisual: Convolution convolutional neural network FPGA high-level synthesis optimization directives Convolución directivas de optimización FPGA red neuronal convolucional síntesis de alto nivel

Documento fuente: Ingeniería, ISSN 0121-750X, Vol. 26, Nº. 1, 2021 (Ejemplar dedicado a: January-April), pags. 62-76

Nota general: application/pdf

Restricciones de acceso: Open access content. Open access content star

Lengua: Spanish

Enlace a fuente de información: Ingeniería, ISSN 0121-750X, Vol. 26, Nº. 1, 2021 (Ejemplar dedicado a: January-April), pags. 62-76