Streszczenie

Projektowanie energooszczędnych systemów wizyjnych o małej latencji i wysokiej skuteczności działania wymaga podejścia charakteryzującego się spójnością między rozwiązaniem algorytmicznym i docelową platformą, na której zostanie ono uruchomione. Głównym wyzwaniem jest implementacja złożonych pamięciowo-obliczeniowo sieci neuronowych w niewielkich urządzeniach małej mocy, jak SoC (System on Chip), FPGA (Field-Programmable Gate Array) czy docelowo ASIC (Application-Specific Integrated Circuit). Stosuje się zatem szereg metod pozwalających na redukcję tej złożoności, poprzez zmniejszenie rozmiarów modelu lub uproszczenie obliczeń, w szczególności operacji mnożąco-akumulujących: jedną z nich jest kwantyzacja parametrów sieci do liczb całkowitych. Pewnym standardem stała się kwantyzacja liniowa 8-bitowa, natomiast to inne, specjalne schematy pozwolą na realizację zaawansowanych systemów o znacznie wyższej wydajności. W ramach przeprowadzonych badań zaproponowano metody uczenia i wydajnej implementacji sprzętowej sieci kwantyzowanych do wag o wartościach potęg dwójki, pozwalając na realizację modeli 4-bitowych o skuteczności porównywalnej do sieci pełnej precyzji i jednocześnie umożliwiając znaczną redukcję złożoności obliczeniowej poprzez zmianę mnożenia na operację przesunięcia bitowego. Ponadto przeanalizowano też możliwości użycia różnych schematów kwantyzacji (liniowej, logarytmicznej, binarnej i mieszanej) dla sieci neuronowych używanych w zaawansowanych systemach wizyjnych, proponując modele dedykowane urządzeniom niewielkiej mocy, dla zadań klasyfikacji, śledzenia oraz detekcji obiektów. W wyniku odpowiedniego projektowania algorytmów i ich implementacji w platformach wbudowanych pokazano, że odpowiednie metody kwantyzacji modeli umożliwiają realizację systemów o wysokiej skuteczności działania, małej latencji i niskim poborze energii.

04.09.2024, Dominika Pretocke-Ruś

Abstract

Designing low-latency, high-accuracy and energy-efficient vision systems requires an approach characterised by consistency between the algorithmic solution and the target platform on which it will run. The main challenge is to implement highly memory and computationally complex neural networks in small low-power devices, such as SoCs (System on Chips), FPGAs (Field-Programmable Gate Arrays) or ultimately ASICs (Application-Specific Integrated Circuits). A number of methods are therefore being used to reduce this complexity, either by reducing the size of the model or by simplifying the computations, particularly multiply and accumulate operations: one of these is the integer quantisation of network parameters. Linear 8-bit quantisation has become a certain standard, while it is other special schemes that will allow to develop advanced systems with much higher performance. This research proposes methods for training and efficient hardware implementation of neural networks quantised to weights of powers of two, allowing the development of 4-bit models with efficiency comparable to full-precision networks and, at the same time, allowing a significant reduction in computational complexity by changing the multiplication to a bit-shift operation. Furthermore, the possibility of using different quantisation schemes (linear, logarithmic, binary and mixed) for neural networks used in advanced vision systems was also analysed, proposing models dedicated to low-power devices, for classification, tracking and object detection tasks. As a result of appropriate algorithm design and implementation in embedded platforms (so-called hardware aware algorithm co-design), it is shown that proper model quantisation methods enable the implementation of complex vision systems with high accuracy, low latency and low power consumption.

04.09.2024, Dominika Przewlocka-Rus