人工智慧晶片設計與應用

AI-ON-CHIP FOR MACHINE LEARNING AND INFERENCE

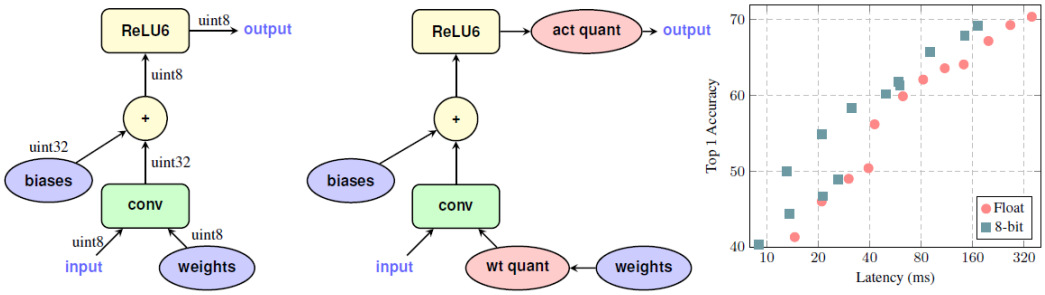
PRA 2

E14096724

鄭喆嚴

我選擇閱讀的論文是 Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference 這篇論文。這篇論文的動機是要解決在移動設備上的深度學習需要大量的運算，目前最先進的卷積神經網絡並不適合在移動設備上使用，這篇論文要改善準確性和設備上的延遲之間的平衡，並且在盡可能不損失準確度下減小 CNN 的模型大小和推斷時間。

這篇論文中主要有四點來改善 MobileNets 在移動設備硬體的延遲與準確性之間的平衡，第一點是使用一種量化方案，這種量化方案是把權重和激勵函數都量化成 8-bit 整數，只留一些像是偏差向量量化成 32-bit 整數；第二點是量化推論架構，可以有效率在整數運算的硬體像是 Qualcomm Hexagon 上實用；第三點是量化訓練架構，與量化推論架構共同設計來最小化在真實模型準確度上的損失；第四點是把這些框架運用在使用 MobileNets 的高校分類和偵測系統，把結果提供在流行的 ARM CPU 上，顯示對於 MobileNet 架構在延遲與準確性之間有了顯著的改進，這在 ImageNet 分類、COCO 物體偵測等等上得到了證明。



(a) Integer-arithmetic-only inference (b) Training with simulated quantization (c) ImageNet latency-vs-accuracy tradeoff

我認為這篇論文提到的量化方案相當巧妙，並且從這個量化方案中能夠看出可以減少大量的執行時間，但如何在丟失一些資料的情況下，不丟失太多的精確度或是能提升精確度才是要解決的問題。這篇論文也有將這個量化方案進行一些實驗，論文中將量化的方法套用在 ImageNet 上使用 Inception v3 和 ResNets 在沒有丟失太多精確度的情況下在執行時間上有額外的進步，還有將量化的方法套用在 ImageNet、COCO、Face detection 和 Face attributes 上使用 MobileNets，其中在 ImageNet 上使用 MobileNets 時，在相同的執行時間上整數運算量化的 MobileNets 可以比浮點數運算的 MobileNets 的準確率含可以高出大約 10%。

我覺得這篇論文在未來的發展可以往更高效的量化方法前進，在論文中提出的量化方案有一定的成績，但未來的研究還是可以往更高效的量化方法，來更加提高量化後模型的準確度和效率；或是往在其他的模型或任務取得更好的成績，研究特殊的設計來解決其他的模型或任務；或是研究更多樣量化的方法，並且可以根據模型結構、數據集和硬體的特性來自動改變量化的方法。