Neural Network Project - MobileNet v3

Salvatore Cognetta 1874383

January 2022

Contents

1 Introduction

Nell'ultimo decennio la potenza di device portatili di piccole dimensioni, come gli smartphone, è cresciuta in modo esponenziale, quasi a raggiungere potenze di low-budget computer. Questo ha fatto sì che si studiasse la possibilità di utilizzo di reti neurali su questi devices.
Come sappiamo, utilizzare tecniche di machine learning è molto heavy load, per tale motivo è necessario fare delle assumptions prima di addentrarsi in questo modo. Allo stato attuale è impensabile effettuare il train su un singolo device mobile con così poca potenza wrt workstation e big datacenter utilizzate generalmente per il train delle reti. Anche se l'edge computing e la possibilità di utilizzare un'insieme di devices contemporaneamente anche per il train di reti neurali si fa sempre più insistente, in questo lavoro ci focalizziamo su un altro tipo di studio: nello specifico nell'utilizzo di lightweight reti neurali, appositamente ingegnerizzate, che vengono trainate a priori ma che possono essere utilizzate per fare inference proprio sui device mobili, come gli smartphone.
In questo senso, abbiamo deciso di implementare quella che è allo stato attuale la state of the art per questo tipo di reti e cioè MobileNetV3 [1]. Gli autori propongono due tipologie di reti, che vengono chiamate MobileNetV3 Small and Large. La prima si è un'evoluzione della rete MobileNetV2 [2], mentre la seconda si basa sul lavoro contenuto in MnasNet-A1 [3].

The goal of the reference paper is to develop the best possible mobile computer vision architectures optimizing the accuracy-latency trade off on mobile devices. To accomplish this they introduce (1) complementary search techniques, (2) new efficient versions of nonlinearities practical for the mobile setting, (3) new efficient network design, (4) a new efficient segmentation decoder.
In questo lavoro viene implementata la soluzione proposta in MobileNetV3 e applicata a due dataset differenti per image detection and classification: MNIST [4] and CIFAR10 [5].

2 Related works and solution

Tale lavoro, come già detto, è un evoluzione di MobileNetV2 alla ricerca di un modello quasi-ottimo per soluzioni con poca potenza computazionale. Per realizzare questa soluzione è ovviamente necessario raggiungere un punto di trade-off In MobileNetV2 gli autori

SqueezeNet[22] extensively uses 1x1 convolutions with squeeze and expand modules primarily focusing on re- ducing the number of parameters. More recent works shifts the focus from reducing parameters to reducing the number of operations (MAdds) and the actual mea- sured latency. MobileNetV1[19] employs depthwise sepa- rable convolution to substantially improve computation ef- ficiency. MobileNetV2[39] expands on this by introducing a resource-efficient block with inverted residuals and lin- ear bottlenecks. ShuffleNet[49] utilizes group convolution and channel shuffle operations to further reduce the MAdds. CondenseNet[21] learns group convolutions at the training stage to keep useful dense connections between layers for feature re-use. ShiftNet[46] proposes the shift operation in- terleaved with point-wise convolutions to replace expensive spatial convolutions.

2.1 Network search

3 Implementation

3.1 Large

3.2 Small

3.3 Adroid app

4 Results

5 Conclusion

References

[1] Andrew Howard et al. "Searching for MobileNetV3". In: (2019). arXiv: 1905.02244 [cs.CV].

[2] Mark Sandler et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: (2019). arXiv: 1801.04381 [cs.CV].

[3] Mingxing Tan et al. "MnasNet: Platform-Aware Neural Architecture Search for Mobile". In: (2019). arXiv: 1807.11626 [cs.CV].

[4] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[5] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009.