作者

Pieter van der Wolf

Sw l mnqwq l l a.

简介

k

:

•

k

•

k

!

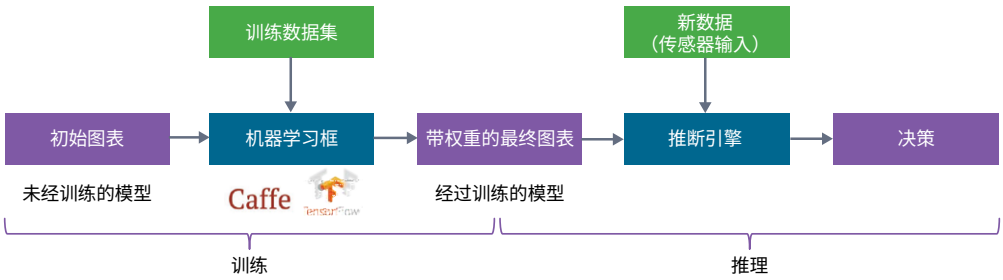
! ! k

k

k

Y1[k 1

k



1

k

k C_ddc Tc l qmpFjm u

!

k

网络处理

k

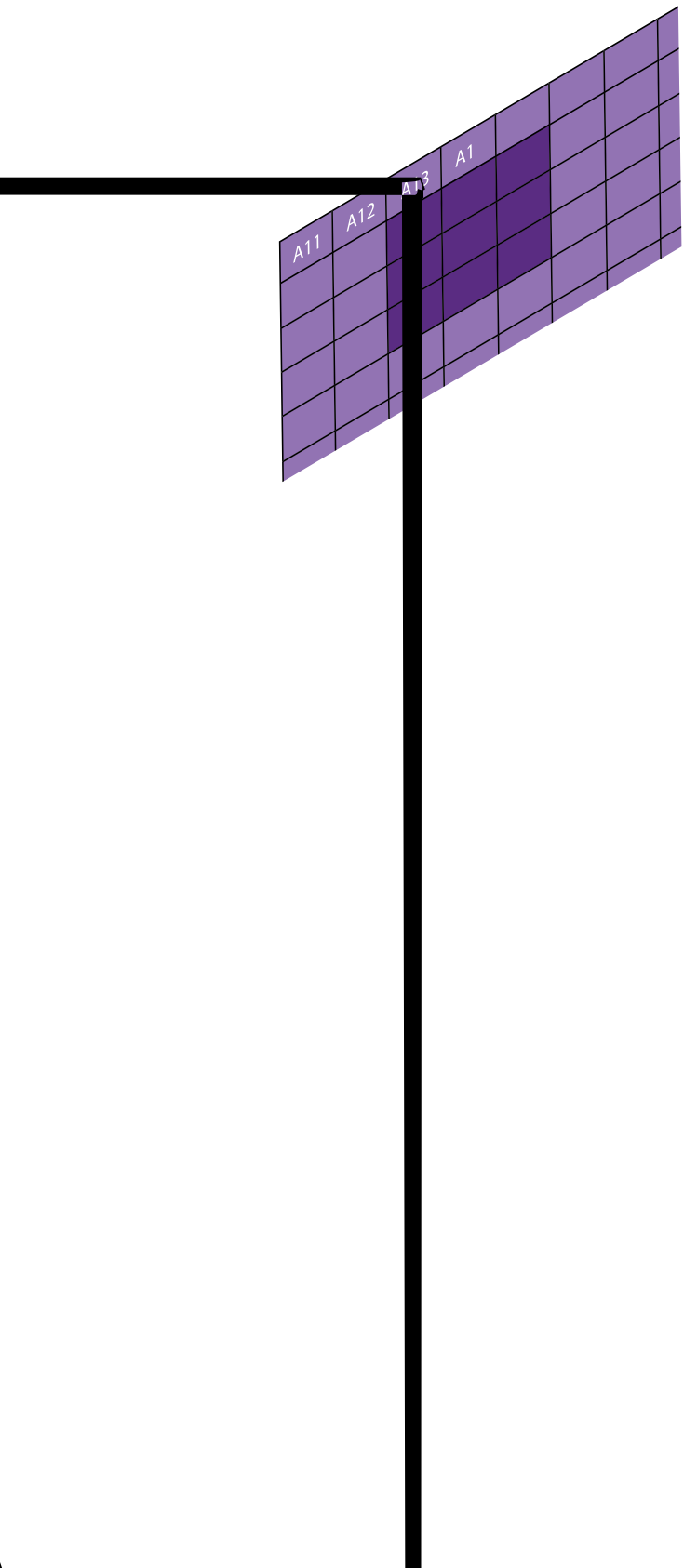
k (CNN) k

3 k

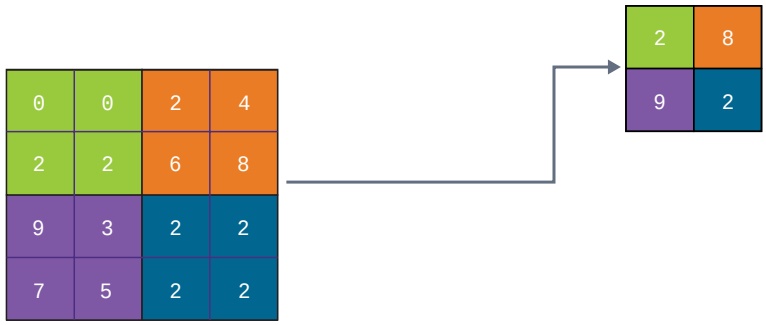
k (M23) k (2, 3) k

k

A24-A26 A34-A36 A44-A46 M24 k



M23 = W11 x A23 + W12 x A24 + W13 x A25
+ W21 x A33 + W22 x A34 + W23 x A35
+ W31 x A43 + W32 x A44 + W33 x A45


0 0 2 4
2 2 6 8
9 3 2 2
7 5 2 2
2 8
9 2

低功耗机器学习推理的处理器功能要求

/ 32 DSP ARC EM9D

(MAC)

MAC 5 ARC EM9D MAC

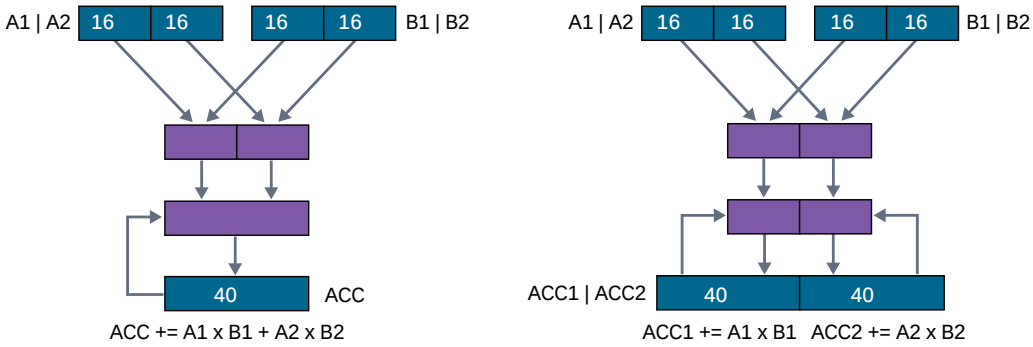


图5 ARC EM9D处理器的两类矢量MAC指令

MAC 2v16 DMAC MAC A1 A2
B1 B2 ARC EM9D 32 8 DMAC
MAC
5 TMAC -MAC
1 A1 A2 B1 B2 A1 A2

MAC MAC
ARC EM9D VW VW
6
(AGS) AGS AGS
AGS
AGS

VW AGS ! !
k AGS k
VW AGS k
k
MAC kTLIU MAC
k k
kARC EM9D
sDMA k
k
ARC EM9D DSP FFT k

-
-
-
- 2v16 4v8
- 16+16 `srrcpdjw
-

DSP ARC EM9D / k

用于机器学习推理的软件库

k
ARC EM9D k

c k `ARC (MLI) Y5[C k
/ k MLI k 2
MLI k ! k
! ! k

c k `ARC MLI k k
k
Y6[k c k `ARC MLI LSTM RNN k GRS PpmhcarcbLSTM RNN RNN ! ! k

MLI k k
k TclqmpFjmu C_ddc `jm` k
k kjg]rclqmp MLI k
! ! ! k k
2D ! k
k

- k
- k
-

ck`ARC MLI Q-lmr_rgml Y7[k / 8 16
k k
16 8 MLI 16 8
k

ARC EM9D AGS k
AGS k k 16 8 k
16 ARC EM9D MAC
5 k 16v8 MAC 2v16 2v8 8
k 7 C ! 16v8 MAC
k

```
            ; AGU 0 is used for input data. AGU 1 is used for weights
            ; setup AGU 0 for loading next 2x16-bit vector
            ; setup AGU 1 for loading next 2x8-bit into lower bytes of 32-bit word


     A  B   0, %agu_u0, %agu_u1 ; DMAC + 2 loads + 2 pointer updates
    :
```

图7 :使用MLI C代码为具有16位输入数据和8位权重的全连接层生成的汇编代码

MAC RNN k k
3v3 - - (CHU) k 2D k
MAC MAC k 1
k MAC 5 8
k

```
          ; AGU 0 and AGU 1 are used for input data (one data pointer with two modifiers)
          ; AGU 2 is used for weights
          ; setup AGU 0 for loading next two 16-bit input values
          ; setup AGU 1 for loading two 16-bit input values at next row of input data
          ; setup AGU 2 for loading next 8-bit value with sign extension & replication
…

   A 2       0, %agu_u0, %agu_u2 ; VMAC + 2 loads + sign ext + repl + 2 pointer updates
   A 2       0, %agu_u0, %agu_u2 ; VMAC + 2 loads + sign ext + repl + 2 pointer updates
   A 2       0, %agu_u1, %agu_u2 ; VMAC + 2 loads + sign ext + repl + 2 pointer updates
  :
```

图8 :使用MLI C代码为具有16位输入数据和8位权重的2D卷积而生成的汇编代码

AGS 2D k

- C_ddc - (g k 2amj) k

kARC EM9D AGS k

c k `ARC MLI k !

k k c k `ARC MLI

k

k 3v3 ! 1 ! Tc l qmpFjm u SAME

```
mli_krn_depthwise_conv2d_chw_fx8w16d_k3x3_str1_krnpad(...)
```

c k `ARC MLI k

C++ (up_nncp) k kc k `ARC MLI ARC

EM9D Mcr_U_pc C/C++ k DSP

k

ARC EM9D c k `ARC MLI ARC kc k `ARC MLI k

```
mli_krn_permute_fx8(&input,▫&permute_hwc2chw_cfg,▫&ir_Y);
▫
ir_X.el_params.fx.frac_bits▫=▫CONV1_OUT_FRAQ;
mli_krn_conv2d_chw_fx8_k5x5_str1_krnpad(&ir_Y,▫&L1_conv_wt,▫&L1_conv_b,▫&conv_cfg,▫&ir_X);
mli_krn_maxpool_chw_fx8_k3x3(&ir_X,▫&pool_cfg,▫&ir_Y);
▫
ir_X.el_params.fx.frac_bits▫=▫CONV2_OUT_FRAQ;
mli_krn_conv2d_chw_fx8_k5x5_str1_krnpad(&ir_Y,▫&L2_conv_wt,▫&L2_conv_b,▫&conv_cfg,▫&ir_X);
mli_krn_avepool_chw_fx8_k3x3_krnpad(&ir_X,▫&pool_cfg,▫&ir_Y);
▫
ir_X.el_params.fx.frac_bits▫=▫CONV3_OUT_FRAQ;
mli_krn_conv2d_chw_fx8_k5x5_str1_krnpad(&ir_Y,▫&L3_conv_wt,▫&L3_conv_b,▫&conv_cfg,▫&ir_X);
mli_krn_avepool_chw_fx8_k3x3_krnpad(&ir_X,▫&pool_cfg,▫&ir_Y);
▫
ir_X.el_params.fx.frac_bits▫=▫FC4_OUT_FRAQ;
mli_krn_fully_connected_fx8(&ir_Y,▫&L4_fc_wt,▫&L4_fc_b,▫&ir_X);
▫
ir_Y.el_params.fx.frac_bits▫=▫FC5_OUT_FRAQ;
mli_krn_fully_connected_fx8(&ir_X,▫&L5_fc_wt,▫&L5_fc_b,▫&ir_Y);
mli_krn_softmax_fx8(&ir_Y,▫&output);
▫
```

图10 :CIFAR-10推理应用程序的MLI代码

10 c k `ARC MLI k c k `ARC MLI
(ncp k src) RGB CHU k
gp]V gp]W k

CIFAR-10 CNN Y10[

3 ARC EM9D

A 3 RISC-T B 1.7 k ARC EM9D

k

结语

k

k / k ARC EM9D MAC

VW k DSP ARC EM9D k

c k ` ARC MLI ARC EM9D / k CIFAR-10

k ! Swlmnqwq

k

References

[1] Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3): 210–229.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[3] www.synopsys.com/ev

[4] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. “Quantization and training of neural networks for efficient integer-arithmetic-only inference”. arXiv preprint arXiv:1712.05877, 2017.

[5] https://embarc.org/

[6] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., “Deep speech 2: End-to-end speech recognition in english and mandarin”, in International Conference on Machine Learning, 2016, pp. 173–182.

[7] https://en.wikipedia.org/wiki/Q_(number_format)

[8] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images." 2009.

[9]