An Energy-efficient Memory-based High-throughput VLSI Architecture for Convolutional Networks

Mingu Kang, Sujan K. Gonugondla, Min-Sun Keel, Naresh R. Shanbhag

(mkang17@illinois.edu)

Apr 22, 2015
Motivation: Pattern Recognition

• Convolutional Network (ConvNet)
  - State of the art performance in computer vision applications
    (Hand-written character recognition: 99.2 % [Y. LeCun 98’])
  - Software/FPGA platform
    -> power hungry, difficulty in mobile platform
Convolutional Network (ConvNet)

ConvNet data flow

ConvNet structure

(Hand-written character recognition)

ConvNet computation kernel

\[
\begin{bmatrix}
    y_1 \\
    \vdots \\
    y_N
\end{bmatrix} = \phi \left\{ \begin{bmatrix}
    w_{11} & \cdots & w_{1M} \\
    \vdots & \ddots & \vdots \\
    w_{N1} & \cdots & w_{NM}
\end{bmatrix} \ast \begin{bmatrix}
    x_1 \\
    \vdots \\
    x_M
\end{bmatrix} + \begin{bmatrix}
    b_1 \\
    \vdots \\
    b_N
\end{bmatrix} \right\}
\]

- \( w_{mn} \): kernel function
- \( x_m \): input feature map
- \( b_n \): bias term
- \( y_m \): input feature map

*: convolutional operator

- Large storage required for \( w_{mn} \) in complex applications
- Complex connection between \( w_{mn} \) and \( x_m \)
  - Due to moving window of kernel function
Compute Memory

- Conventional system
  - SRAM array for storage
  - Explicitly separated processor
  - Computation w/ high voltage swing
Compute Memory

- Conventional system
  - SRAM array for storage
  - Explicitly separated processor
  - Computation w/ high voltage swing

- In-memory computing platform
  - Pattern matching (ICASSP 14’)
  - Multi-row (MR) READ
  - Embedded analog processing w/ low voltage swing
Multi-row Read (MR-READ) (ICASSP 14’)

- Energy efficiency: low-swing and multi-row access per precharge
- High throughput: read-out word (instead of bit) per BL

\[
\Delta V_{BLB} \propto \sum_{i=0}^{N-1} T_i d_i \\
\text{(} T_i \ll R_{BL} C_{BL} \text{)}
\]

\[
D = \sum_{i=0}^{N-1} 2^i d_i
\]

\[
D = 0000b' \text{ or } 1010b'
\]
Mixed-signal Capacitive Multiplier

\[ \Delta V_m = (0.5)^{B_P} P \Delta V_{B_{LB}}(D) = \alpha P D \]

Mixed-signal (capacitive) multiplier

- Multiplication of digital (P) and analog (D) values
- \( B_P \): bit precision of P

※ HSPICE simulation w/ Industry 45nm SOI process
ConvNet on Compute Memory

- Mixed-signal multiplication in every BL in parallel
- No BW limitation between logic and memory
Behavioral model with circuit non-idealities

MR-READ non-linearity (ICASSP 14’):

\[ \Delta V_{BLB}(D) = \sum_{k=0}^{4} c_k D^k \]

Process variation in MR-READ:

\[ \Delta \hat{V}_{BLB}(D) \sim \mathcal{N}(\Delta V'_{BLB}(D), \sigma_D^2) \]

Capacitive multiplier behavior:

\[ \Delta V_m = f_0 \Delta V_{BLB}(D)P + f_1 \Delta V_{BLB}(D) + f_2 P + f_3 \]

• Coefficients obtained from circuit simulations
• LetNet5 w/ K = 5, L = 32, back propagation training with 80 iterations
• Detection rate: 99.13% (0.02% degradation as compared to conv.) w/ MNIST dataset (60000 training hand-written number dataset)
Throughput and Energy

Relative delays

\[ E_{\text{conv}} = K^2 E_{\text{read}} + E_{\text{leak}} + (L - K + 1)^2 K^2 E_{\text{MAC}} + E_{\text{reg}} \]

\[ E_{\text{CM}} = (L - K + 1)^2 K^2 E_{\text{MAC}_{\text{read}}} + E_{\text{leak}_{\text{CM}}} + (L - K + 1)^2 K^2 E_{\text{MAC}_{\text{add}}} + E_{\text{reg}} \]

- 4.9X delay reduction
- 5X energy saving

※ HSPICE simulation w/ Industry 45nm SOI process
Conclusion

• In-memory computing platform (ICASSP14’) applied for Convolutional network
• Behavioral and energy model proposed
• 99.13% w/ MNIST dataset
• Delay reduction: 4.9X
• Energy saving: 5X
Q & A