1. Motivation and Related works
2. Approximate Operation to multiplication
3. Building MinConvNets with approximate operation
4. Conclusion
1. Motivation and Related works
USE CASE OF DEEP CONVOLUTIONAL NEURAL NETWORK

Classification: Traffic lights is red!

Object detection: The car is here!

Object tracking: It has to pay a fine!
USE CASE OF DEEP CONVOLUTIONAL NEURAL NETWORK

it's cat
TINY-YOLO [REDMON ET AL.'2016] FOR OBJECT DETECTION
Challenges for embedded systems

- Capacity of computing (multiplicator etc.),
- Memory or bandwidth for loading the data.

\[ FLOP_{\text{multiplication}} = 3.48G \]
How to reduce the computing resources required for convolution which includes a large volume of multiplications?
RELATED WORKS TO REDUCE THE COMPUTING RESOURCES

Original Network  Pruning Network  Quantization Network

- 32 bits floating point
- 8 bits integer
RELATED WORKS TO REDUCE THE COMPUTING RESOURCES

Original Network  Pruning Network  Quantization Network

- 32 bits floating point
- 8 bits integer

Why always multiplication?
2. Approximate Operation to multiplication
USING APPROXIMATE OPERATION INSTEAD OF MULTIPLICATION?

\[ h[n] = x[n]w[n] \]

\[ grad(Conv(x, w)) \]

\[ grad(BN) \]

\[ grad(w)[n] \]
USING APPROXIMATE OPERATION INSTEAD OF MULTIPLICATION?

\[ g[n] = G(x, w) \]

\[ h[n] = x[n]w[n] \]

\[ \text{grad}(\text{Conv}(x, w)) \]

\[ \text{grad}(\text{BN}) \]

\[ \text{grad}(w)[n] \]
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

\[ \rho(h, g) = \frac{\text{cov}(h, g)}{\sqrt{\text{var}(h)\text{var}(g)}} \]

- Similarity of the trends of changes.

\[ L(h, g) = \sum |\frac{g-h}{h}| \]

- Distance between signals.
The similarity between two signals $h$ and $g$.

- $\rho(h, g) = \frac{\text{cov}(h,g)}{\sqrt{\text{var}(h)\text{var}(g)}}$

- Similarity of the trends of changes.

- $L(h, g) = \sum |\frac{g-h}{h}|$

- Distance between signals.
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

Pearson product-moment correlation coefficient (PPMCC)

\[
\rho(h, g) = \frac{\text{cov}(h, g)}{\sqrt{\text{var}(h)\text{var}(g)}}
\]

where:

\[
\begin{align*}
\text{var}(h) &= \sum_n (h[n] - \mu_h)(h[n] - \mu_h) \\
\text{cov}(h, g) &= \sum_n (h[n] - \mu_h)(g[n] - \mu_g)
\end{align*}
\]
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

Pearson product-moment correlation coefficient (PPMCC)

\[
\rho(x, y_1) = 1
\]

\[
\rho(x, y_2) = 0.869
\]

\[
\rho(x, y_3) = 0.193
\]

\[
\rho(h, g) = \frac{\text{cov}(h, g)}{\sqrt{\text{var}(h)\text{var}(g)}}
\]
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

correlation coefficient with multiplication

$g[n] = G(x, w)$

$w[n]$  
$h[n] = x[n]w[n]$  
$BN$  
$grad(x)[n]$  
$grad(Conv(x, w))$  
$grad(BN)$  
$grad(w)[n]$  
$x$
### Correlation with \( h = \xi \cdot \eta \)

<table>
<thead>
<tr>
<th>Correlation with ( h = \xi \cdot \eta )</th>
<th>Min-selector ( g = \min(\xi, \eta) )</th>
<th>Addition ( g = \xi + \eta )</th>
<th>Max-selector ( g = \max(\xi, \eta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \left{ \begin{array}{l} \xi \sim N_f(0,1) \ \eta \sim N_f(0,1) \end{array} \right} )</td>
<td>0.908</td>
<td>0.882</td>
<td>0.673</td>
</tr>
<tr>
<td>( \left{ \begin{array}{l} \xi \sim N_f(0,1) \ \eta \sim N_f(0,10) \end{array} \right} )</td>
<td>0.692</td>
<td>0.683</td>
<td>0.624</td>
</tr>
<tr>
<td>( \left{ \begin{array}{l} \xi \sim U(0,1) \ \eta \sim U(0,1) \end{array} \right} )</td>
<td>0.962</td>
<td>0.926</td>
<td>0.641</td>
</tr>
<tr>
<td>( \left{ \begin{array}{l} \xi \sim U(0,1) \ \eta \sim U(0,1) \end{array} \right} )</td>
<td>0.716</td>
<td>0.717</td>
<td>0.655</td>
</tr>
</tbody>
</table>

- \( \xi \) and \( \eta \) are non-negative value.
- \( N_f(\mu, \sigma^2) \): folded normal distribution with expected value \( \mu \), variance \( \sigma^2 \).
- \( U(a, b) \): a uniform distribution in an interval \([a, b]\).
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

correlation coefficient with multiplication

\begin{align*}
y_1 &= x_1 \cdot x_2 \\
y_2 &= \min(x_1, x_2) \\
y_3 &= x_1 + x_2 \\
y_4 &= \max(x_1, x_2)
\end{align*}
### THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

correlation coefficient with multiplication

<table>
<thead>
<tr>
<th>Correlation with $h = \xi \cdot \eta$</th>
<th>Min-selector $g = \min(\xi, \eta)$</th>
<th>Addition $g = \xi + \eta$</th>
<th>Max-selector $g = \max(\xi, \eta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\begin{cases} \xi \sim N_f(0,1) \ \eta \sim N_f(0,1) \end{cases}$</td>
<td>0.908</td>
<td>0.882</td>
<td>0.673</td>
</tr>
<tr>
<td>$\begin{cases} \xi \sim N_f(0,1) \ \eta \sim N_f(0,10) \end{cases}$</td>
<td>0.692</td>
<td>0.683</td>
<td>0.624</td>
</tr>
<tr>
<td>$\begin{cases} \xi \sim U(0,1) \ \eta \sim U(0,1) \end{cases}$</td>
<td>0.962</td>
<td>0.926</td>
<td>0.641</td>
</tr>
<tr>
<td>$\begin{cases} \xi \sim U(0,1) \ \eta \sim U(0,1) \end{cases}$</td>
<td>0.716</td>
<td>0.717</td>
<td>0.655</td>
</tr>
</tbody>
</table>

$h = \xi \cdot \eta$ and $g = \min(\xi, \eta)$ have the similar trends of changes, if:

- $\xi$ and $\eta$ follow similar distribution:
  - They have the same expected values, noted as $\mu_{|\xi|} = \mu_{|\eta|}$
  - They are distributed in similar intervals, noted as $\sigma_\xi \sim \sigma_\eta$
THE SIMILARITY BETWEEN TWO SIGNALS $h$ AND $g$

- $\rho(h, g) = \frac{\text{cov}(h, g)}{\sqrt{\text{var}(h)\text{var}(g)}}$
- Similarity of the trends of changes.

- $L(h, g) = \sum \left| \frac{g-h}{h} \right|$
- Distance between signals.
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

\[ L(h, g) = \sum |\frac{g-h}{h}| \]

- Find the constraints to make $L$ as small as possible.
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

Let inputs $\xi$ and $\eta$ random variables with probability distribution $p_x(\xi)$ and $p_w(\eta)$, and outputs $g$ and $h$ are calculated as:

$$\begin{cases} h = H(\xi, \eta) = \xi \cdot \eta \\ g = G(\xi, \eta) = \min(\xi, \eta) \end{cases}$$

Then the distance between signals is calculated as:

$$L(h, g) = \int_{\xi} \int_{\eta} \frac{H(\xi, \eta) - G(\xi, \eta)}{H(\xi, \eta)} \cdot p_x(\xi)p_w(\eta)d\xi d\eta$$
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

Let inputs $\xi$ and $\eta$ random variables with probability distribution $p_x(\xi)$ and $p_w(\eta)$, and outputs $g$ and $h$ are calculated as:

$$\begin{align*}
&h = H(\xi, \eta) = \xi \cdot \eta \\
g = G(\xi, \eta) = \min(\xi, \eta)
\end{align*}$$

Then the distance between signals is calculated as:

$$L(h, g) = \int_{\xi} \int_{\eta} \left| \frac{H(\xi, \eta) - G(\xi, \eta)}{H(\xi, \eta)} \right| \cdot p_x(\xi)p_w(\eta) d\xi d\eta$$
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

Let inputs $\xi$ and $\eta$ random variables with probability distribution $p_x(\xi)$ and $p_w(\eta)$, and outputs $g$ and $h$ are calculated as:

\[
\begin{align*}
    h &= H(\xi, \eta) = \xi \cdot \eta \\
    g &= G(\xi, \eta) = \min(\xi, \eta)
\end{align*}
\]

Then the distance between signals is calculated as:

\[
L(h, g) = \int_{\xi} \int_{\eta} \left| \frac{H(\xi, \eta) - G(\xi, \eta)}{H(\xi, \eta)} \right| \cdot p_x(\xi)p_w(\eta) d\xi d\eta
\]

\[
= f_1(p_x(\xi), p_w(\eta))
\]
Let inputs $\xi$ and $\eta$ random variables with probability distribution $p_x(\xi)$ and $p_w(\eta)$, and outputs $g$ and $h$ are calculated as:

$$
\begin{align*}
    h &= H(\xi, \eta) = \xi \cdot \eta \\
    g &= G(\xi, \eta) = \min(\xi, \eta)
\end{align*}
$$

Then the distance between signals is calculated as:

$$
L(h, g) = f_1(p_x(\xi), p_w(\eta))
$$
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

Let inputs $\xi$ and $\eta$ random variables with probability distribution $p_x(\xi)$ and $p_w(\eta)$, and outputs $g$ and $h$ are calculated as:

\[
\begin{align*}
    h &= H(\xi, \eta) = \xi \cdot \eta \\
    g &= G(\xi, \eta) = \min(\xi, \eta)
\end{align*}
\]

Then the distance between signals is calculated as:

\[
L(h, g) = f_1(p_x(\xi), p_w(\eta))
\]

If $\xi$ and $\eta \sim N_f(k, \nu)$:

\[
L(h, g) = f_2(k, \nu)
\]

where $k$ represents the expected values of $\xi$ and $\eta$, and $\nu$ represents the variance of $\xi$ and $\eta$. 
THE DISTANCE BETWEEN TWO SIGNALS $h$ AND $g$

To make $L(k, \nu)$ as small as possible:

- **C1**: $k$ that minimizes $L$ is around 1, noted as $\mu_{|\xi|} = \mu_{|\eta|} = 1$.
- **C2**: $\nu$ should be as small as possible.
3. Building MinConvNets with approximate operation
BUILD THE APPROXIMATE CONVOLUTION

with C1: $\mu_{|\xi|} = \mu_{|\eta|} = 1$.

Let matrix multiplication arbitrary:

$$|z| = |x| \cdot |w|$$

be transformed as:

$$\frac{|z|}{\mu_{|x|}\mu_{|w|}} = \frac{|x|}{\mu_{|x|}} \cdot \frac{|w|}{\mu_{|w|}}$$

That meets constraint $\mu_{|\xi|} = \mu_{|\eta|} = 1$, therefore:

$$\frac{|z|}{\mu_{|x|}\mu_{|w|}} \approx \min\left(\frac{|x|}{\mu_{|x|}}, \frac{|w|}{\mu_{|w|}}\right)$$

So:

$$|z| = \mu_{|w|} \cdot \min(|x|, \frac{\mu_{|x|}}{\mu_{|w|}} \cdot |w|)$$
Remove excessively large values:

\[
\text{clip}(w, \alpha) = \begin{cases} 
\alpha & \text{if } w > \alpha \\
-w & \text{if } w < -\alpha \\
w & \text{otherwise}
\end{cases}
\]

- In these works, \( \alpha = 2\mu_{|w|} \) shared by each filter.
- Weights and inputs are both clipped during training.
- Only weights are pre-clipped for inferring.
BUILD THE APPROXIMATE CONVOLUTION
with approximate multiplication composed by min-selector
VALIDATION OF MINCONVNET

Top-1 accuracy of LeNet applied to Cifar10

Approximate computing for embedded machine learning
VALIDATION OF MINCONVNET
Top-1 accuracy of mini-Cifar applied to Cifar10
4. Conclusion
CONCLUSION
MinConvNets: A new class of multiplication-less Neural Networks

- Approximate Multiplication is proposed.
- MinConvNets are built by using Approximate Multiplication.
- Transfer Learning is used to optimize the training.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>LeNet-MNIST</th>
<th>LeNet-Cifar10</th>
<th>Mini_cifar-Cifar10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Network</td>
<td>99.06%</td>
<td>75.26%</td>
<td>77.30%</td>
</tr>
<tr>
<td>Approximate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>170 epoch</td>
<td>98.42%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>512 epoch</td>
<td></td>
<td>64.18%</td>
<td>71.46%</td>
</tr>
<tr>
<td>2048 epoch</td>
<td></td>
<td>65.54%</td>
<td>72.89%</td>
</tr>
<tr>
<td>Transfer Learning</td>
<td></td>
<td>74.92%</td>
<td>77.01%</td>
</tr>
<tr>
<td>512 epoch</td>
<td></td>
<td>75.10%</td>
<td>77.26%</td>
</tr>
<tr>
<td>1024 epoch</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>