Scaling Back-Propagation by Parallel Scan Algorithm

Shang Wang¹,² | Yifan Bai¹ | Gennady Pekhimenko¹,²

¹Computer Science, University of Toronto
²Vector Institute

Back-propagation (BP)'s Strong Sequential Dependency

\[ \nabla_x f(\hat{x}) = \nabla_x \nabla_{\hat{x}} f(\hat{x}) \]

- Linear
- ReLU
- Conv2d
- MaxPool2d

Strong Sequential Dependency along layers.

Model Parallel Training

Strong sequential dependency limits scalability on parallel systems.

- Linear per-device space complexity.
- “Bubble of idleness” vs. convergence affect.

What is a Scan Operation?

Binary, associative op.: +, input:

- Worker (p): an instance of execution; e.g., a core in a multi-core CPU.
- Number of Elements (n).

Reformulate BP as a Scan Operation

Key Insight: matrix multiplication in BP is also binary & associative!

G = \nabla_x f(\hat{x})

Define op.: A \cdot B = BA, input:

- On a single worker: scan linearly: n steps.
- With more workers: sublinear steps?

Jacobians are Memory & Compute Hungry

A full Jacobian: prohibitively expensive.

- e.g., \( ^1 \text{st} \) convolution in VGG-11 on CIFAR-10 images:

\[
\begin{align*}
\nabla_x f(\hat{x}) &= \begin{bmatrix}
3072 & 65536
\end{bmatrix}^T
\end{align*}
\]

Generated by Op_Grad (basis vectors) one by one.

Leverage the Sparsity in the Jacobians

- Deterministic pattern.
- Known ahead of training time.
- Potentially better SpGEMM performance.

Generated directly into CSR:

Evaluation

- BPPSA scales with \( n \) until being bounded by \( p \); and scales with \( p \).

BPPSA reconstructs the original BP exactly.

Hardware Sensitivity

- # of SMs(2080Ti) > # of SMs(2070)
- Latency(2080Ti) < Latency(2070)

Scan BP by Bellowcho Scan

What is a Scan Operation?

A worker (p): an instance of execution; e.g., a core in a multi-core CPU.

Definition: Scan

- Exclusive scan:

\[ \begin{bmatrix} G_1, G_2 \end{bmatrix} \]

- Parallel scan:

\[ \begin{bmatrix} G_1, G_2, G_3 \end{bmatrix} \]

What is a Scan Operation?

A worker (p): an instance of execution; e.g., a core in a multi-core CPU.

Definition: Scan

- Exclusive scan:

\[ \begin{bmatrix} G_1, G_2 \end{bmatrix} \]

- Parallel scan:

\[ \begin{bmatrix} G_1, G_2, G_3 \end{bmatrix} \]

What is a Scan Operation?

A worker (p): an instance of execution; e.g., a core in a multi-core CPU.

Definition: Scan

- Exclusive scan:

\[ \begin{bmatrix} G_1, G_2 \end{bmatrix} \]

- Parallel scan:

\[ \begin{bmatrix} G_1, G_2, G_3 \end{bmatrix} \]