SONAR
Direct Architecture and System Optimization Search

Elias Jääsaari, Michelle Ma, Ameet Talwalkar, Tianqi Chen
Constrained deployment

Model

Objectives
- Accuracy
- Latency
- Memory consumption
- Energy use
- Number of parameters
- FLOPS
- Fairness
- ...
The ML pipeline

- Data Processing
- Model Training
- Model Evaluation
- Model Deployment
The ML pipeline

System optimization search

```
for yo in range(1024 / ty):
    for xo in range(1024 / tx):
        [[yo*ty:yo*ty+ty][xo*tx:xo*tx+tx]] = 0
for k in range(1024):
    for yi in range(ty):
        for xi in range(tx):
            C[yo*ty,yi][xo*tx,yi] +=
                A[k][yo*ty,yi] * B[k][xo*tx,xi]
```

```
for yo in range(128):
    for xo in range(128):
        intrin.fill_zero(C[yo*8:yo*8+8][xo*8:xo*8+8])
for ko in range(128):
    intrin.fused_gemm8x8_add(
        C[yo*8:yo*8+8][xo*8:xo*8+8],
        A[yo*8:yo*8+8][xo*8:xo*8+8],
        B[yo*8:yo*8+8][xo*8:xo*8+8])
```
Indirect search

Model training provides accuracy feedback

Architecture search

Latency proxy provides latency feedback

Target

Data
Indirect search

Model training provides accuracy feedback

Architecture search

Latency proxy provides latency feedback

Sequential stages

System optimization search improves latency

Latency proxy might not match deployment
Direct search

Model training provides accuracy feedback

Architecture search

System optimization search provides latency feedback

No second stage needed!
It is better to be direct
Efficient direct search

How to perform efficient direct search?

Use early stopping!
Early stopping for accuracy

Different networks

Corresponding training curves
Early stopping for accuracy

Accuracy vs Latency

- Eliminate network with bad accuracy

Accuracy vs Training iterations

- Allocate budget to promising networks
Early stopping for accuracy

Accuracy vs. Latency

Graph showing accuracy and latency with training iterations on the x-axis.

Final selected network highlighted.
Early stopping for latency

- Precision
- Latency

Program evaluations

Different networks

Corresponding tuning curves
Early stopping for latency

Eliminate network with bad latency

Allocate budget to promising networks
Early stopping for latency

![Diagram showing accuracy vs. latency with a final selected network highlighted. The right side shows program evaluations vs. latency, with a steep drop indicating efficiency.](image)
Different networks
SONAR applies early stopping to both objectives simultaneously.
SONAR finds near optimal models
Thank you