: Compile-Time   : Runtime

1. Computation Graph (forward)

Data   Weight

MatMul

Out

2. AutoDiff
$f(x) \rightarrow f'(x)$

3. Computation Graph (backward)

dy/dx   dy/dw

MatMul'

grad

4. Execution Engine

Runtime graph optimizations lead to **training overhead**.

(a). Conventional training frameworks

1. Computation Graph (forward)

Data   Weight

MatMul

Out

2. AutoDiff
$f(x) \rightarrow f'(x)$

3. Computation Graph (backward)

dy/dx   dy/dw

MatMul'

grad

4. Execution Engine

Extensive graph optimizations during compile-time.

(b). Our proposed Tiny Training Engine