



GitHub: <https://github.com/UofT-EcoSystem/hotline>

Demo: <https://danielsnider.ca/hotline/demo>

Usage: `with torch.profiler.profile(
 on_trace_ready=hotline.analyze(model)):`



HOTLINE: Automatic Annotation and A Multi-Scale Timeline for Visualizing Time-Use in DNN Training

Daniel Snider, Fanny Chevalier, Gennady Pekhimenko

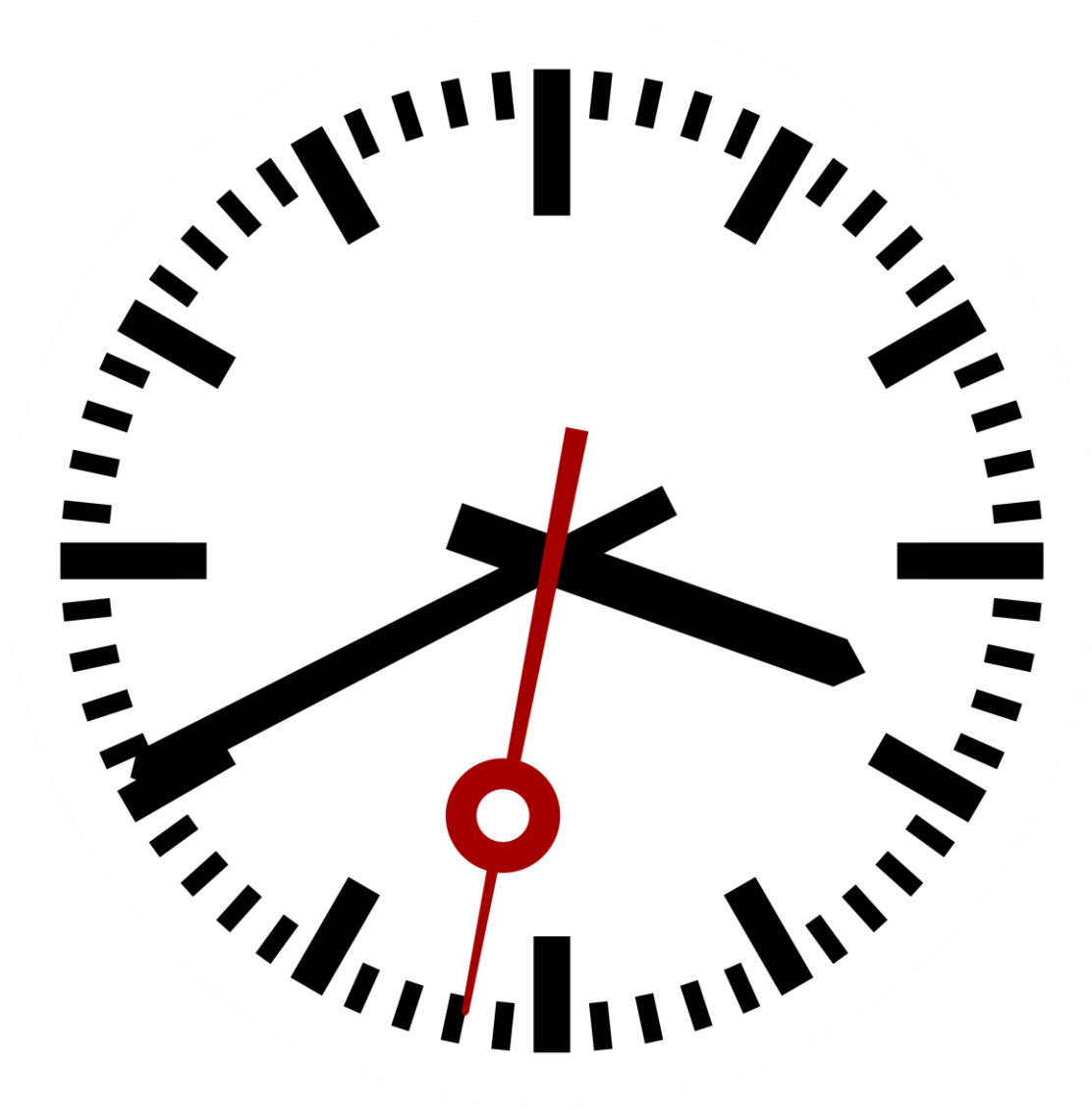
EcoSystem
Efficient Computing Systems




UNIVERSITY OF
TORONTO

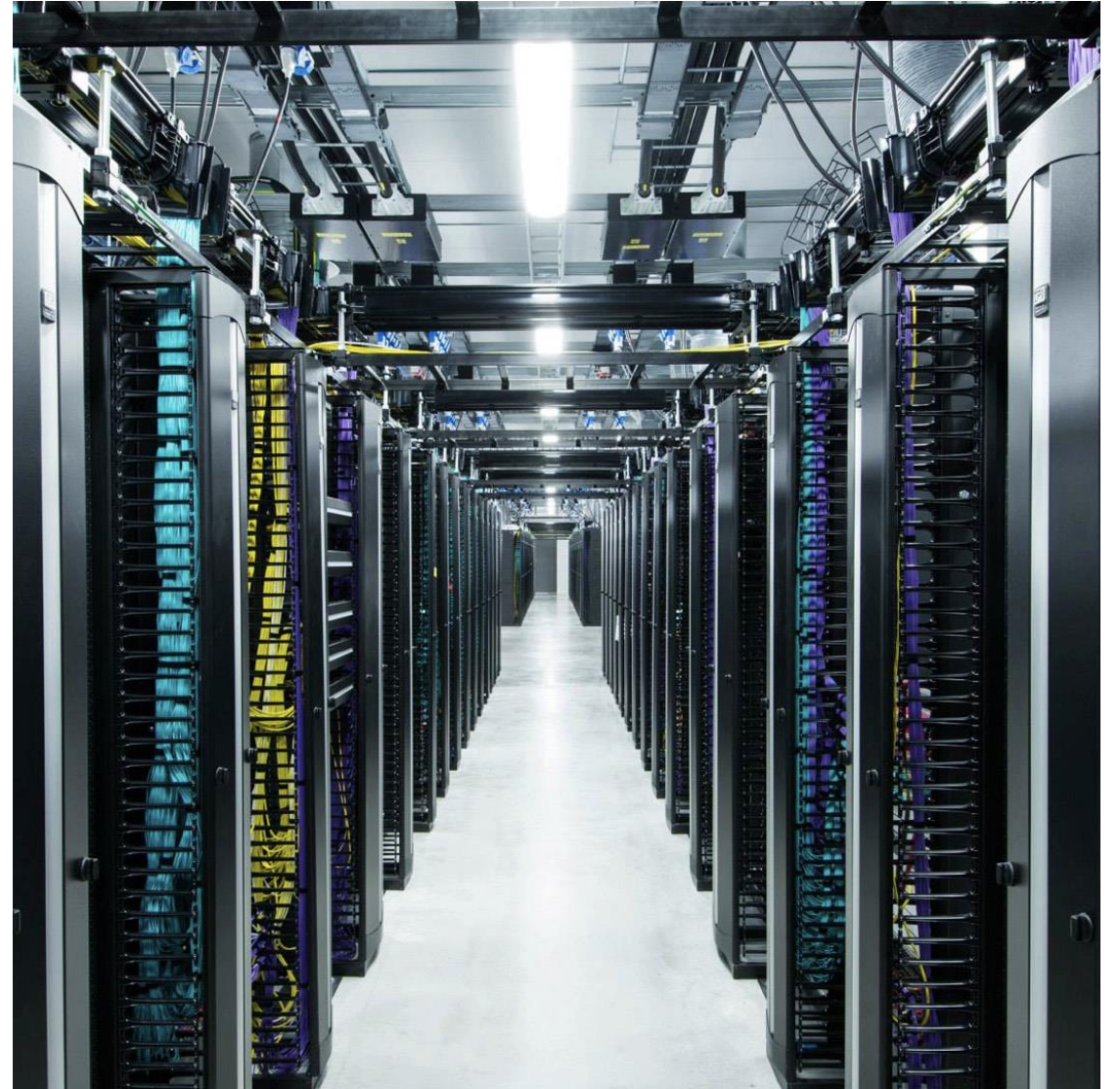


VECTOR
INSTITUTE



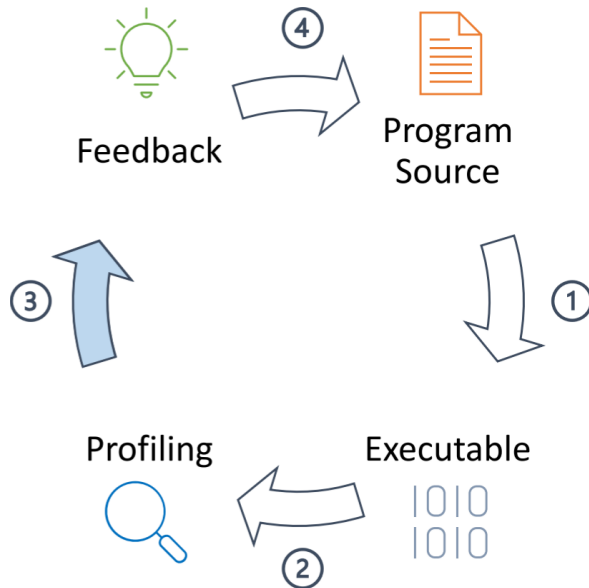


How is ML
spending
it's day?



Finding Optimization Opportunities

Profiling



https://www.jokeren.tech/assets/CGO21_slides.pdf

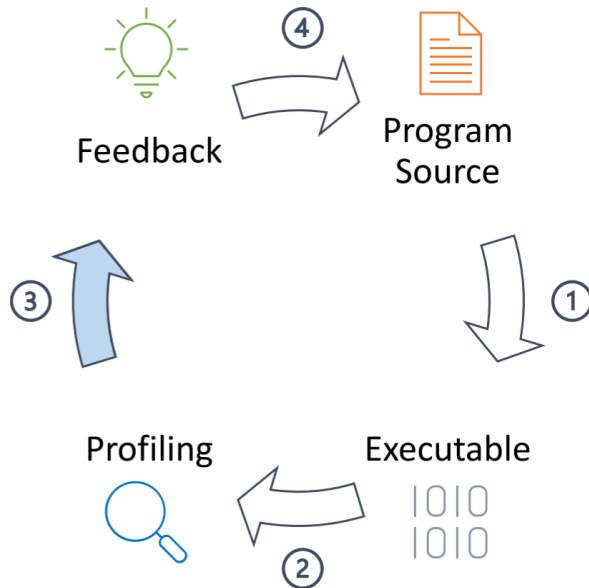
Visualization



<https://www.sigmacomputing.com/blog/what-is-data-visualization/>

Finding Optimization Opportunities

Profiling



https://www.jokeren.tech/assets/CGO21_slides.pdf

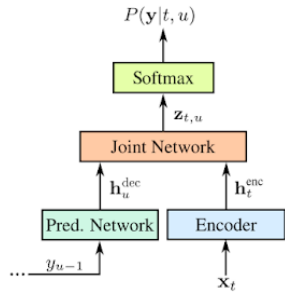
Visualization



<https://www.sigmacomputing.com/blog/what-is-data-visualization/>

Problems: Requires significant time, expertise, and use of expert-level tools.

RNN-T Model (Speech to Text)



Graves, A. "Sequence transduction with recurrent neural networks". ICML, 2012.

+

LibriSpeech Dataset



Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. "LibriSpeech: An ASR corpus based on public domain audio books". ICASSP, 2015.

DNN Profiling Demo



"Where's the
Stall, Yo?"

TensorBoard

 PyTorch Profiler
Runtime Trace

- 81 MB, JSON trace file, standard format.
- Single training iteration after warm up iterations.
- No modification to profiler.



Sinclair, D. "Trace Event Format", 2016. URL <https://docs.google.com/document/d/1CvAClvFfyA5RPhYUmn5OOQtYMH4h6l0nSsKchNAYsU>

RNN-T Model

```
with torch.profiler.profile(  
    on_trace_ready=
```

TensorBoard

PyTorch Profiler
Runtime Trace

HOTLINE

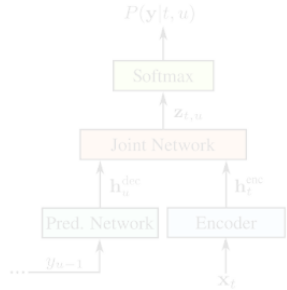
LibriSpeech
Dataset

```
with torch.profiler.profile(  
    on_trace_ready=
```

Panayotov, V., Khudanpur, S. "LibriSpeech: a corpus based on public domain audio books". ICASSP, 2015.

DNN Profiling Demo

RNN-T Model
(Speech to Text)



Graves, A. "Sequence transduction with recurrent neural networks". ICML, 2012.

+

LibriSpeech
Dataset



Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. "LibriSpeech: An ASR corpus based on public domain audio books". ICASSP, 2015.

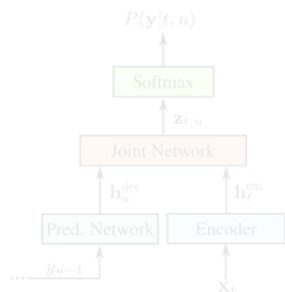
 PyTorch Profiler
Runtime Trace

TensorBoard

HOTLINE

"Where's the
Stall, Yo?"

RNN-T Model (Speech to Text)



Graves, A. "Sequence transduction with recurrent neural networks". ICML, 2012.

+

LibriSpeech Dataset



Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. "LibriSpeech: An ASR corpus based on public domain audio books". ICASSP, 2015.

DNN Profiling Demo

"Where's the
Stall, Yo?"

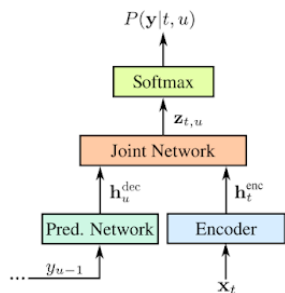
TensorBoard

 PyTorch Profiler
Runtime Trace

HOTLINE

- Q1. Is the **optimizer** a bottleneck?
- Q2. Is the **GPU or CPU** a bottleneck?
- Q3. At a low-level, **what** is the bottleneck?

RNN-T Model (Speech to Text)



Graves, A. "Sequence transduction with recurrent neural networks". ICML, 2012.

+

LibriSpeech Dataset



Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. "LibriSpeech: An ASR corpus based on public domain audio books". ICASSP, 2015.

DNN Profiling Demo



"Where's the
Stall, Yo?"

TensorBoard

PyTorch Profiler
Runtime Trace

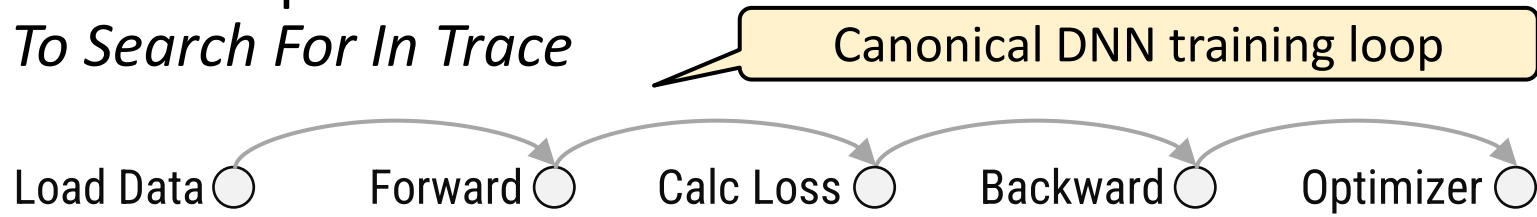
HOTLINE

- Q1. Is the **optimizer** a bottleneck?
Q2. Is the **GPU or CPU** a bottleneck?
Q3. At a low-level, **what** is the bottleneck?

HOTLINE: Automatic Annotation Algorithm

DNN Operations

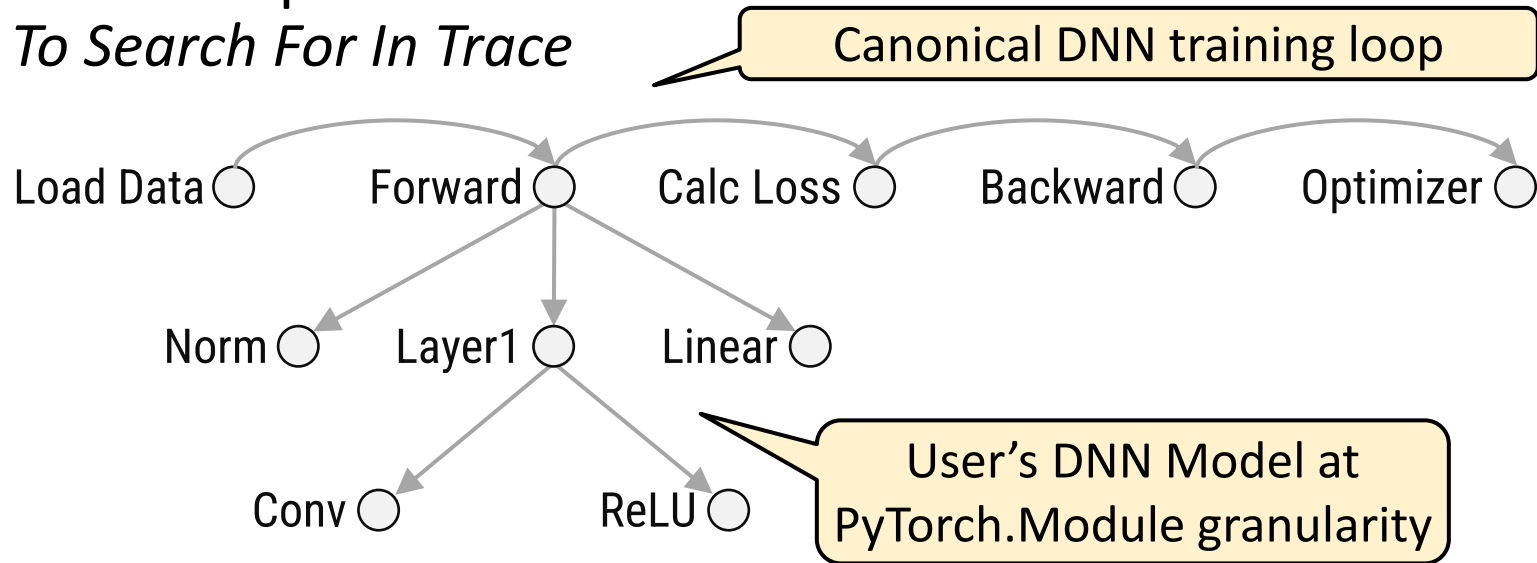
To Search For In Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

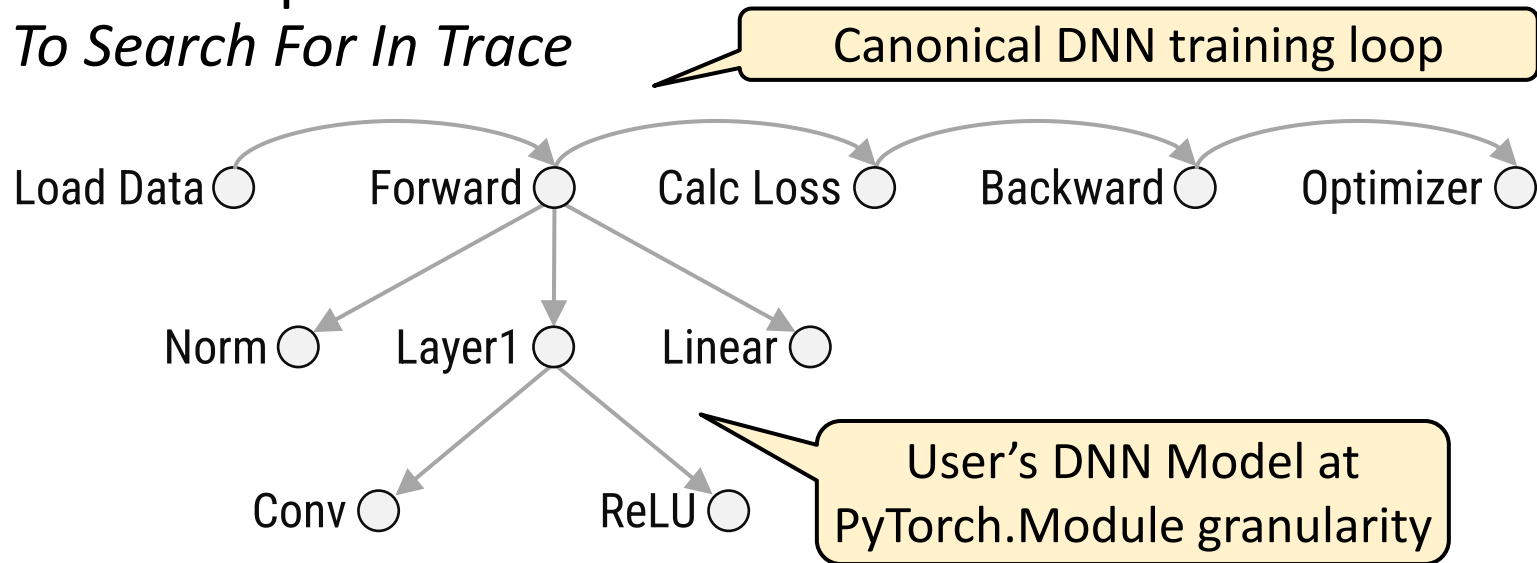
To Search For In Trace



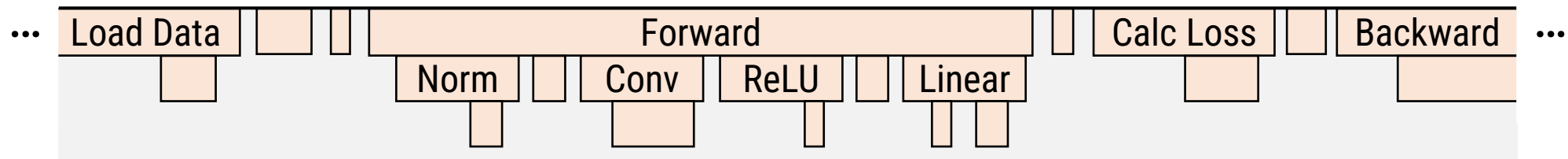
HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



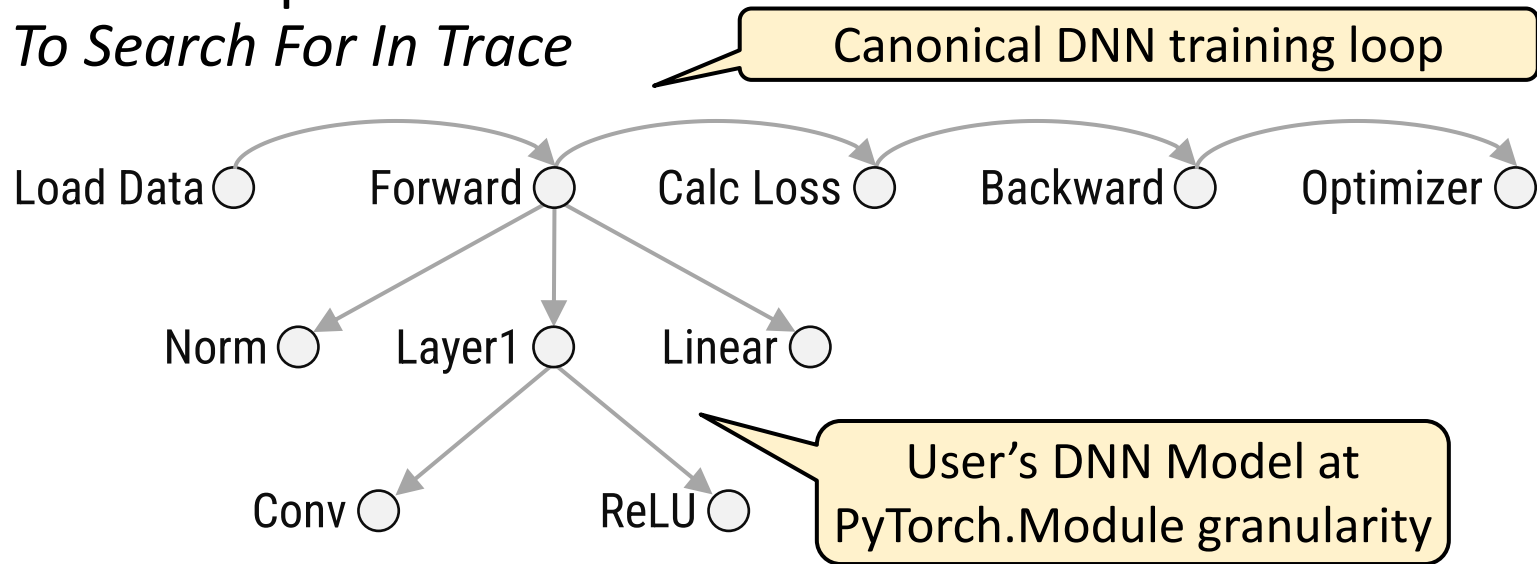
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



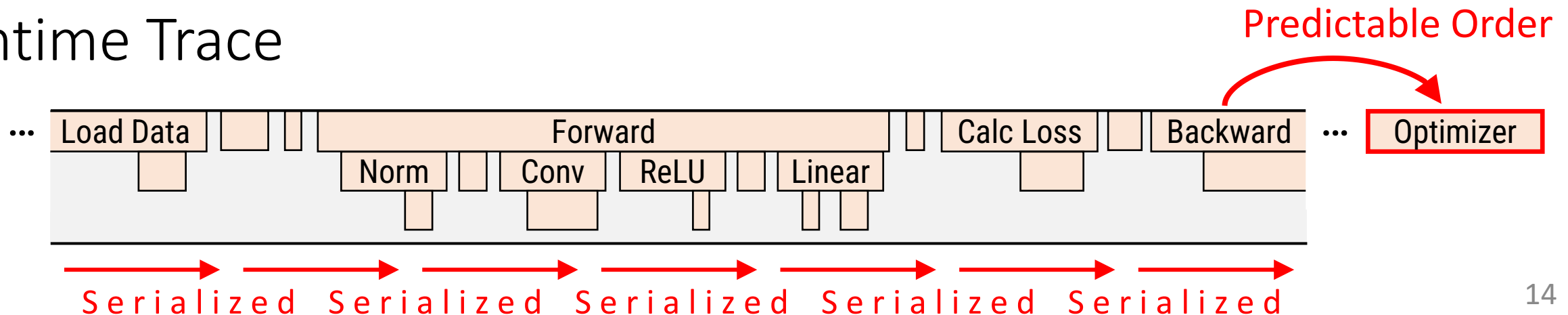
Observations

To Enable Trace Annotation

||||| Serialized

↻ Repeatable / Predictable

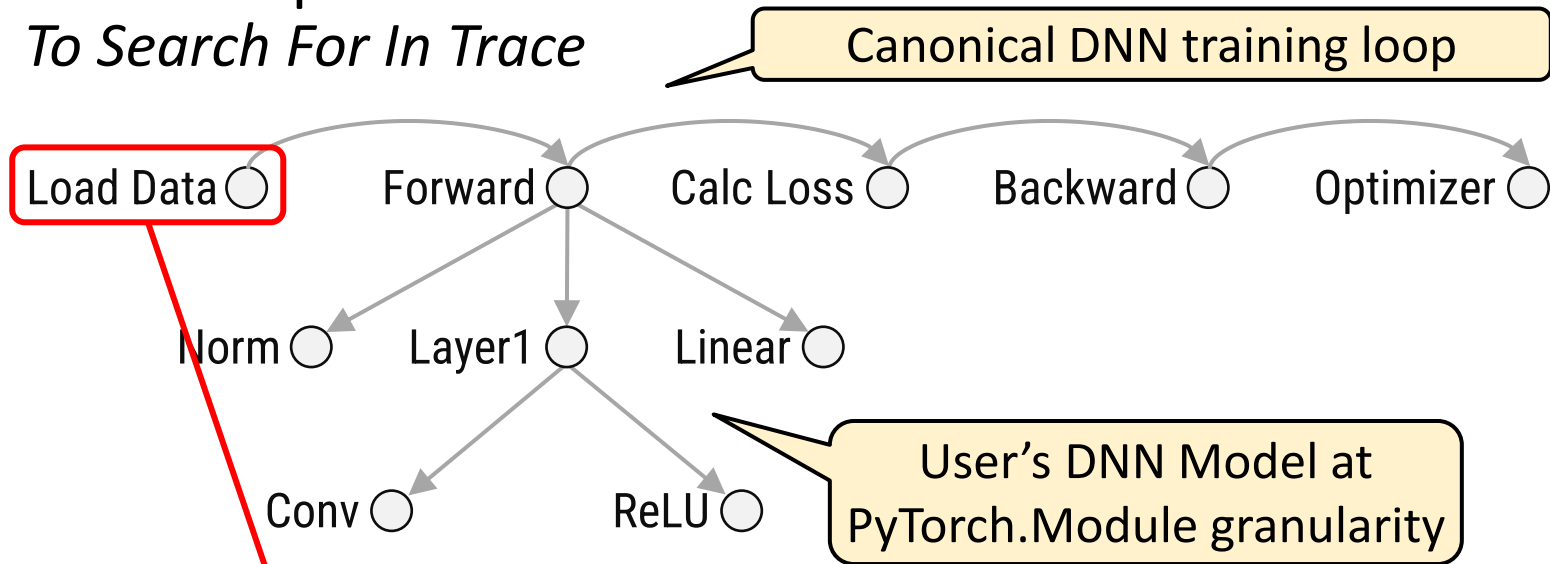
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

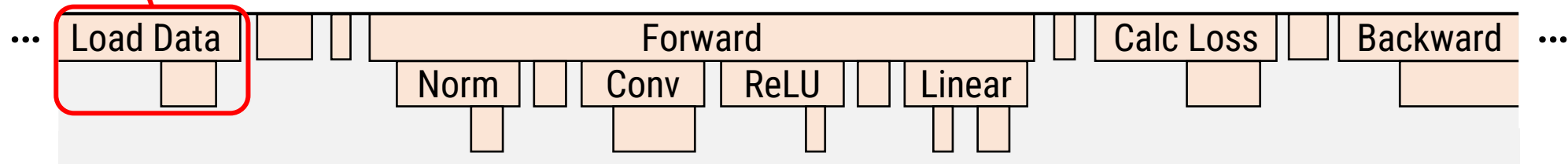
To Enable Trace Annotation

||||| Serialized

↻ Repeatabile / Predictable

👁 Recognizable Names

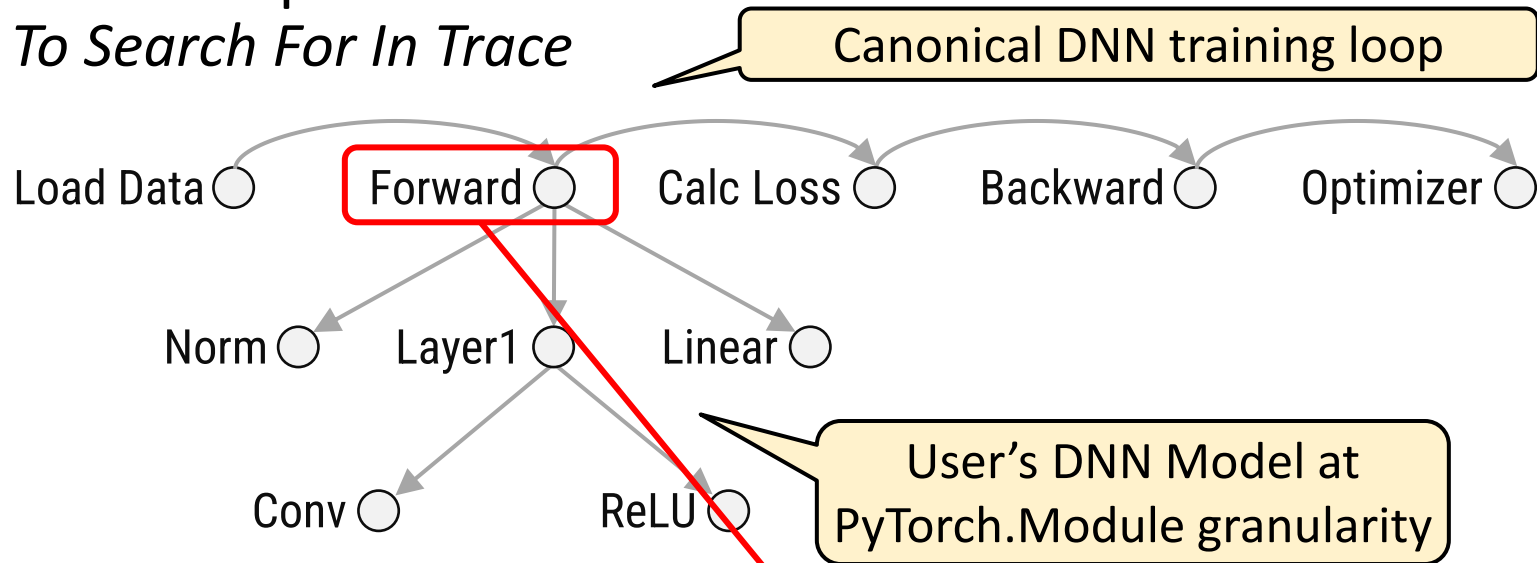
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

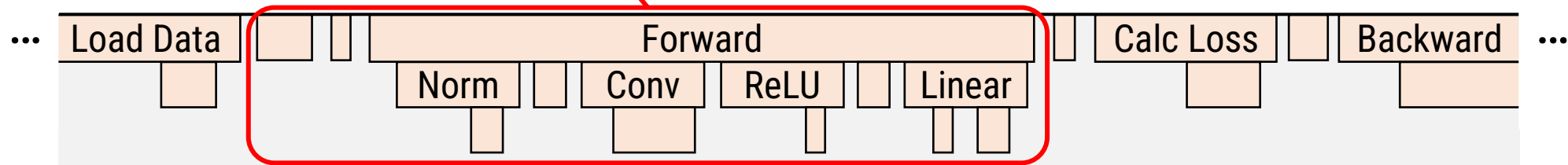
To Enable Trace Annotation

||||| Serialized

↻ Repeatabile / Predictable

👁 Recognizable Names

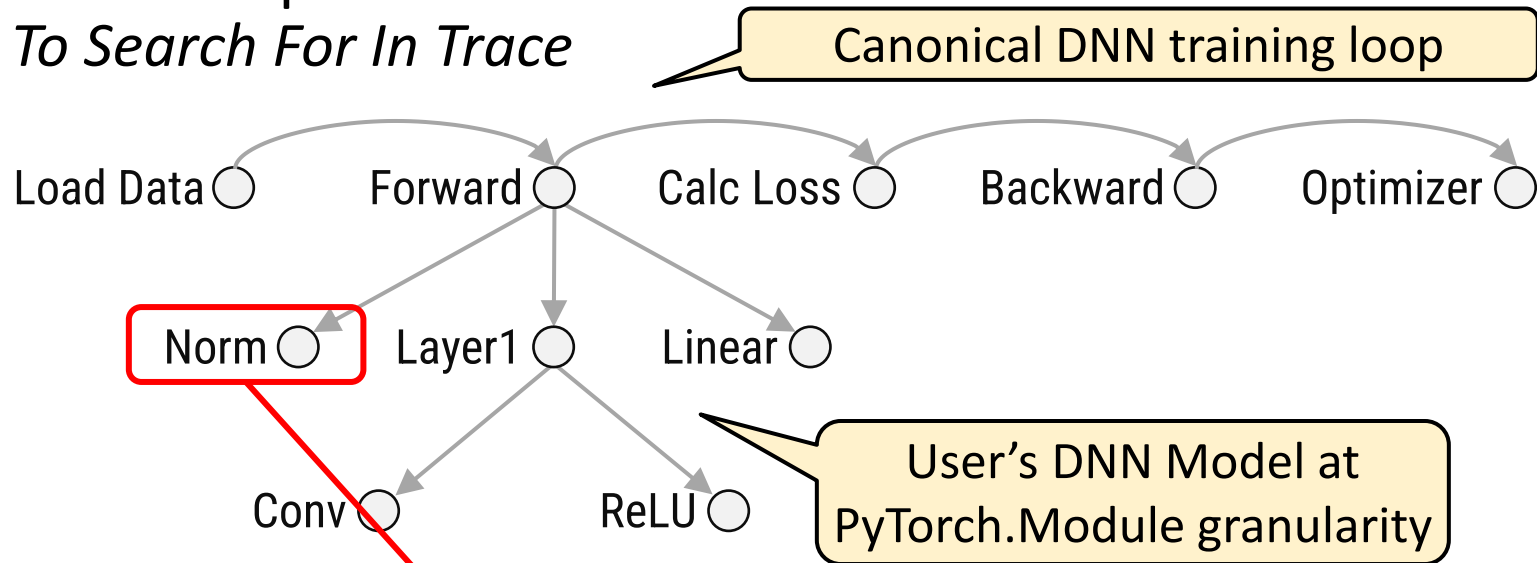
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

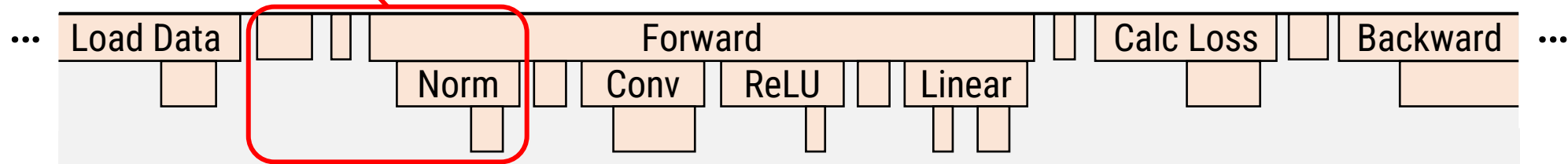
To Enable Trace Annotation

||||| Serialized

↻ Repeatable / Predictable

👁 Recognizable Names

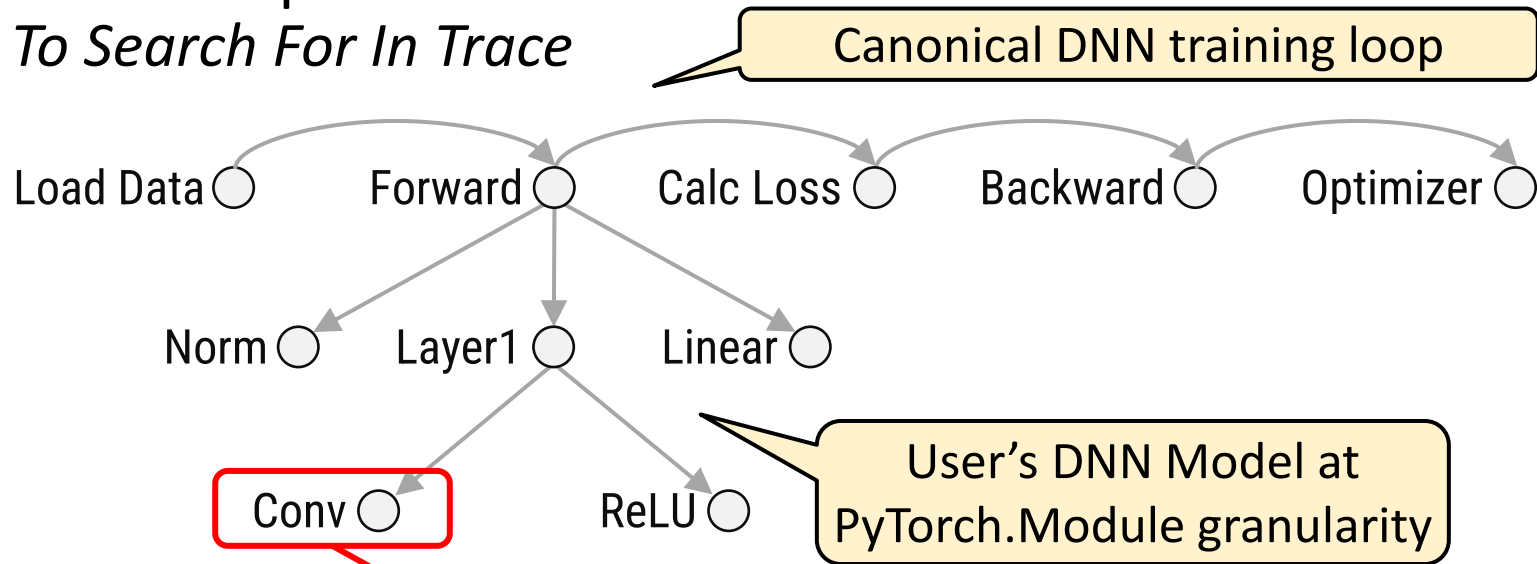
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

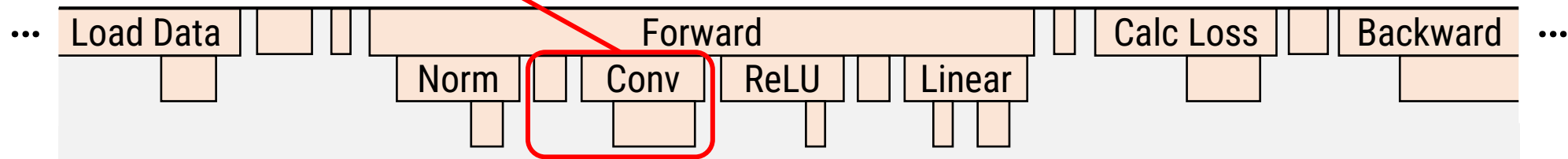
To Enable Trace Annotation

||||| Serialized

↻ Repeatabile / Predictable

👁 Recognizable Names

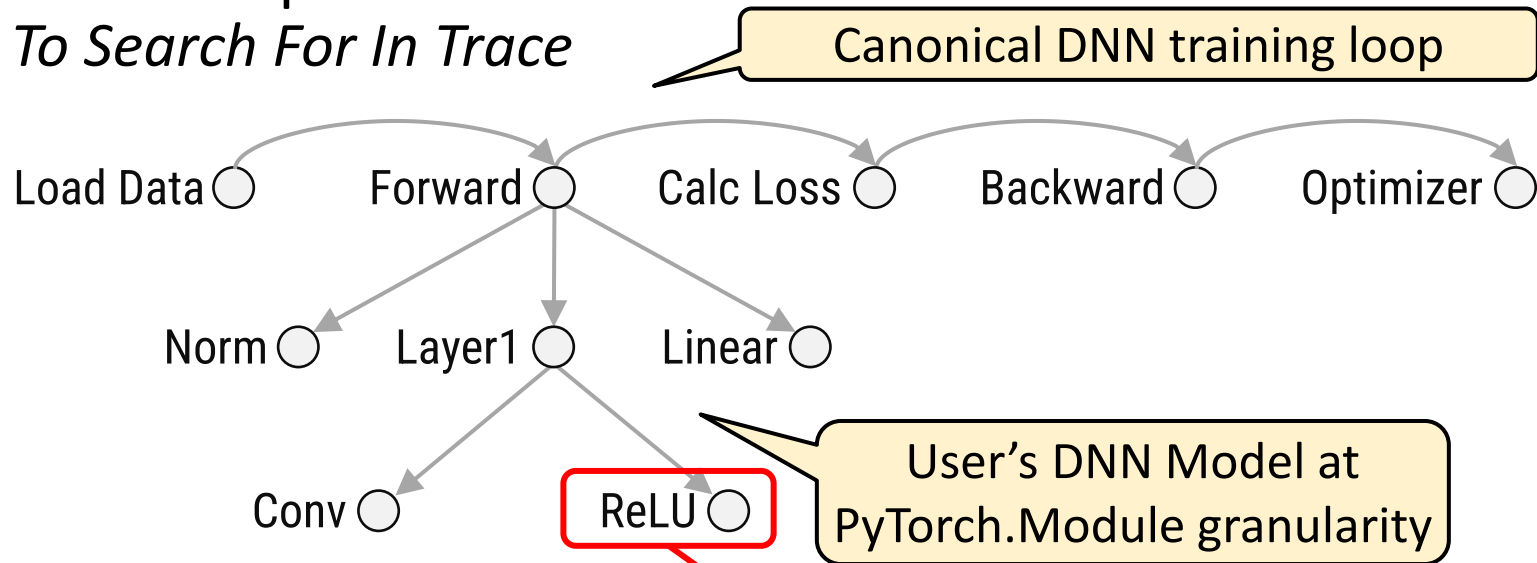
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

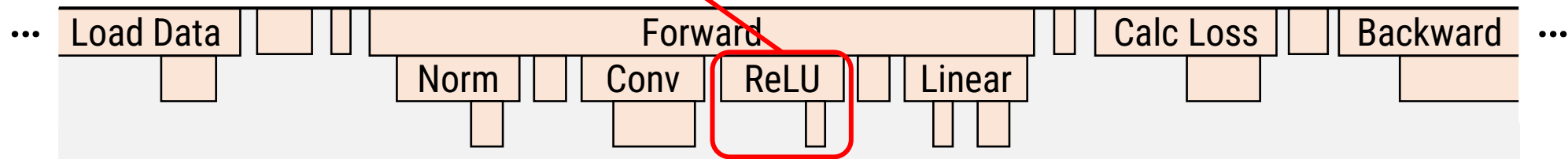
To Enable Trace Annotation

||||| Serialized

↻ Repeatabile / Predictable

👁 Recognizable Names

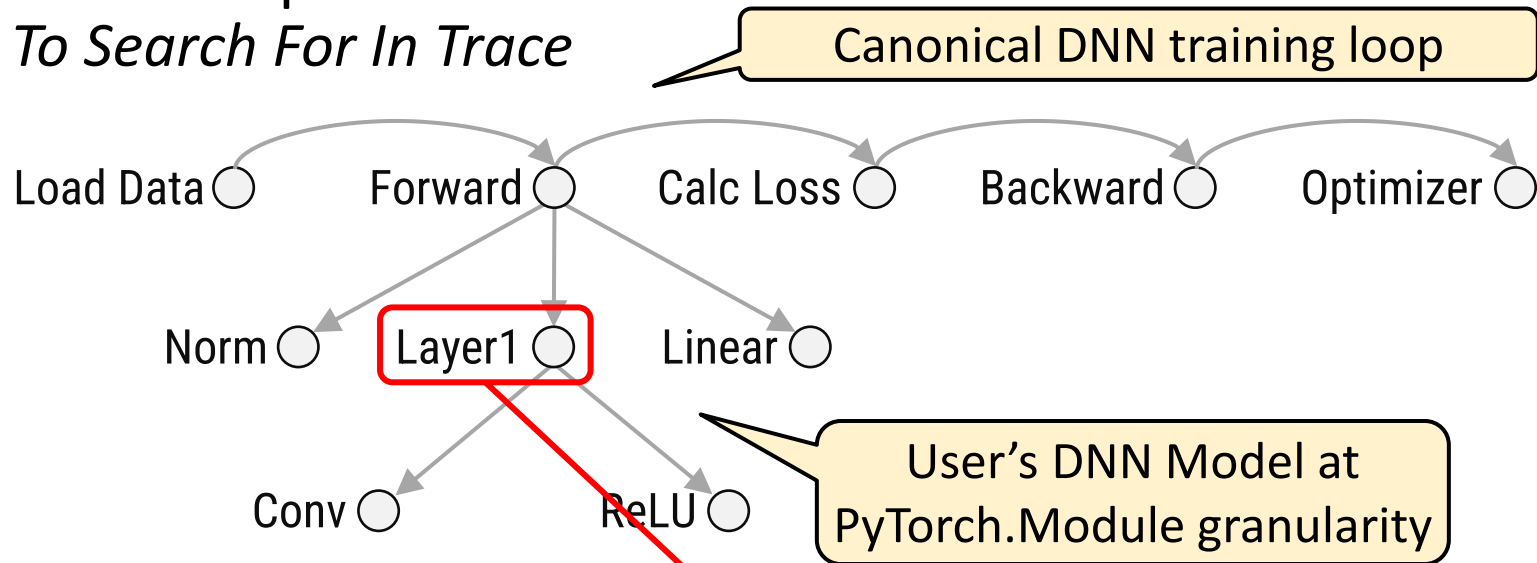
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

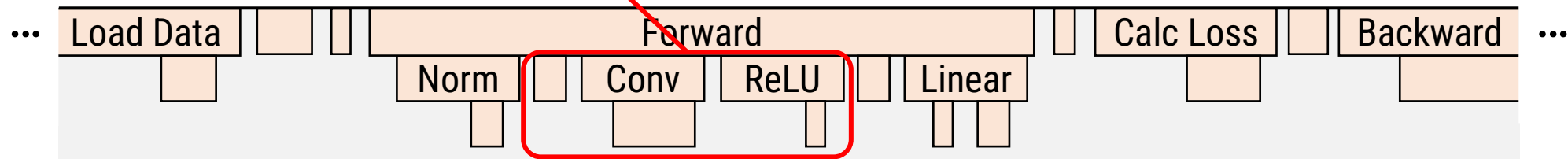
To Enable Trace Annotation

||||| Serialized

↻ Repeatable / Predictable

👁 Recognizable Names

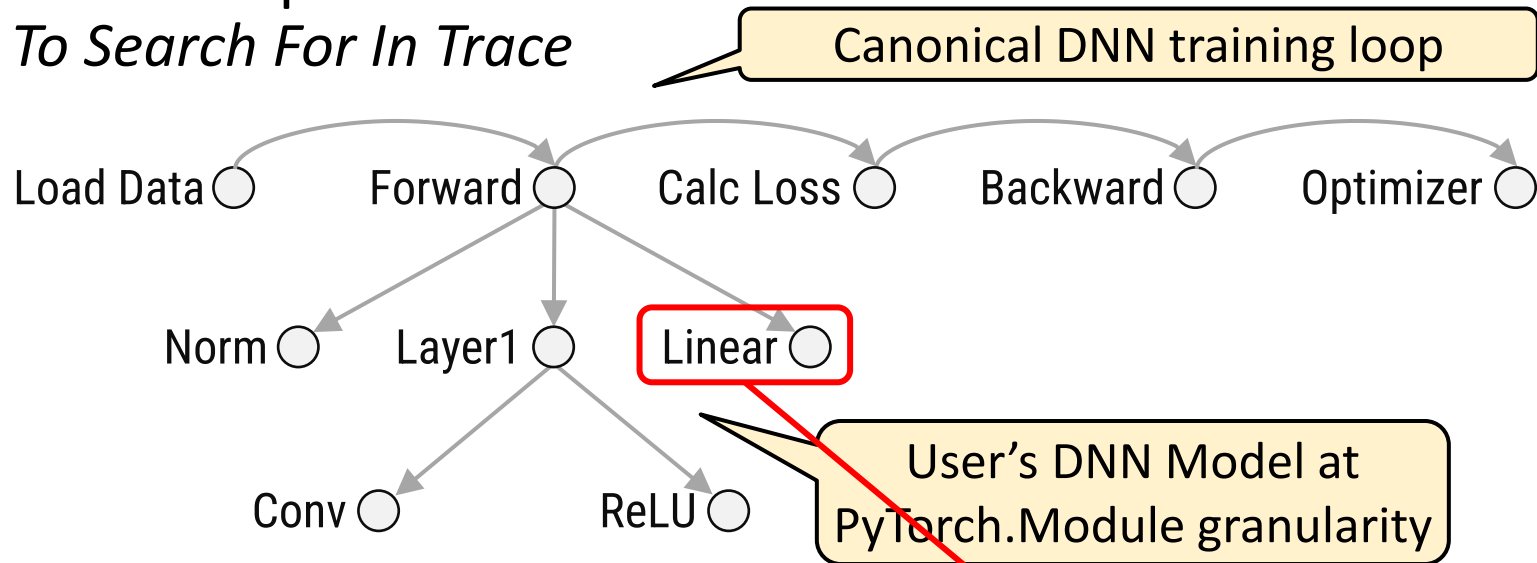
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

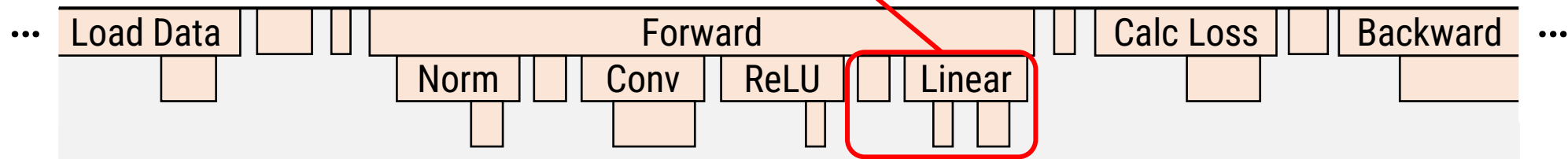
To Enable Trace Annotation

||||| Serialized

↻ Repeatable / Predictable

👁 Recognizable Names

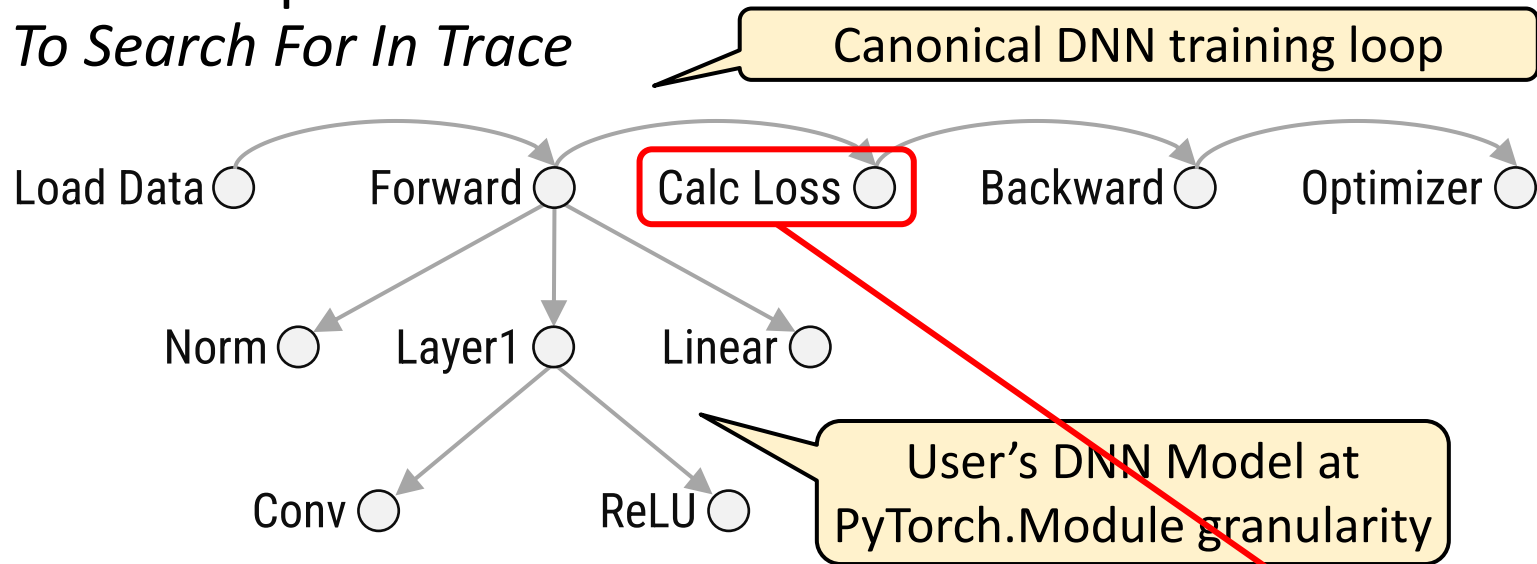
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

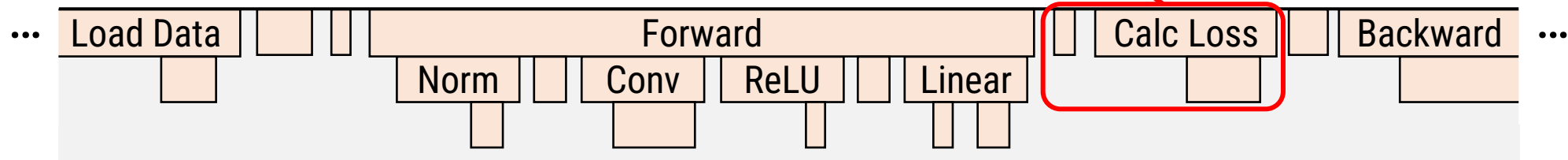
To Enable Trace Annotation

||||| Serialized

↻ Repeatabile / Predictable

👁 Recognizable Names

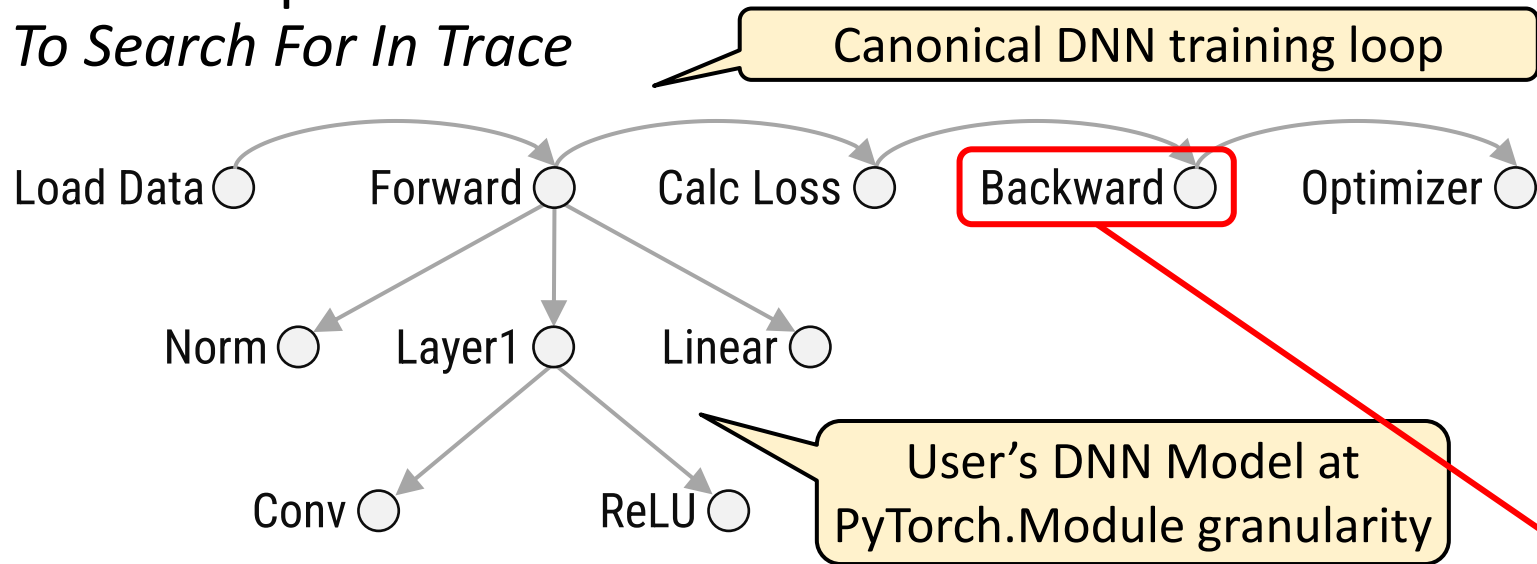
Runtime Trace



HOTLINE: Automatic Annotation Algorithm

DNN Operations

To Search For In Trace



Observations

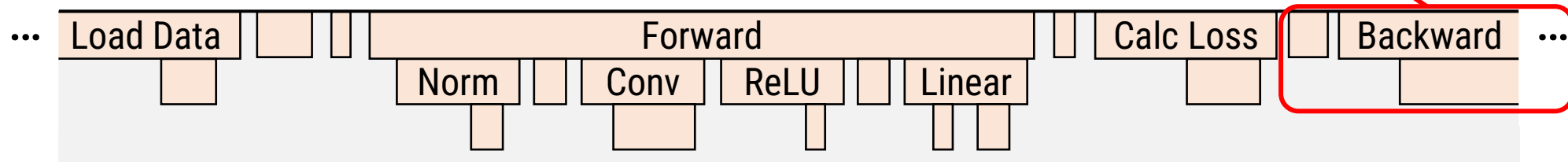
To Enable Trace Annotation

||||| Serialized

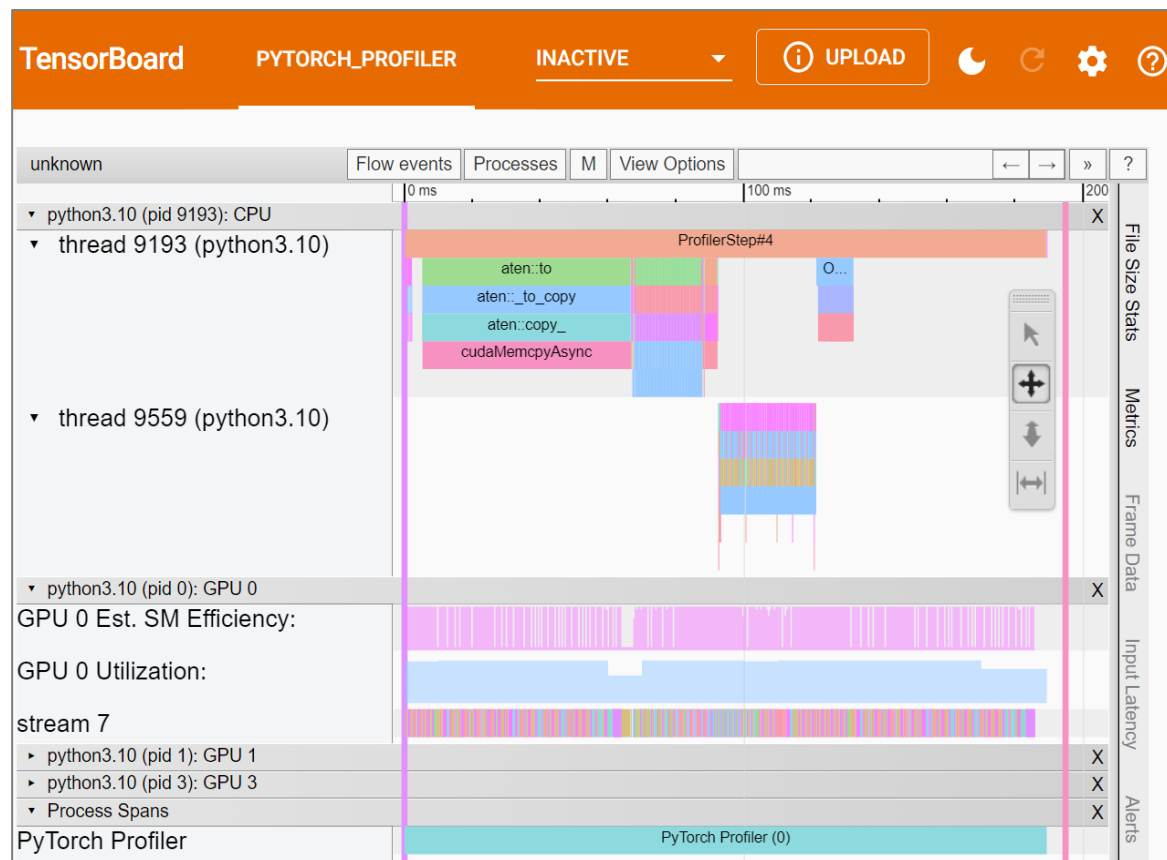
↻ Repeatabile / Predictable

👁 Recognizable Names

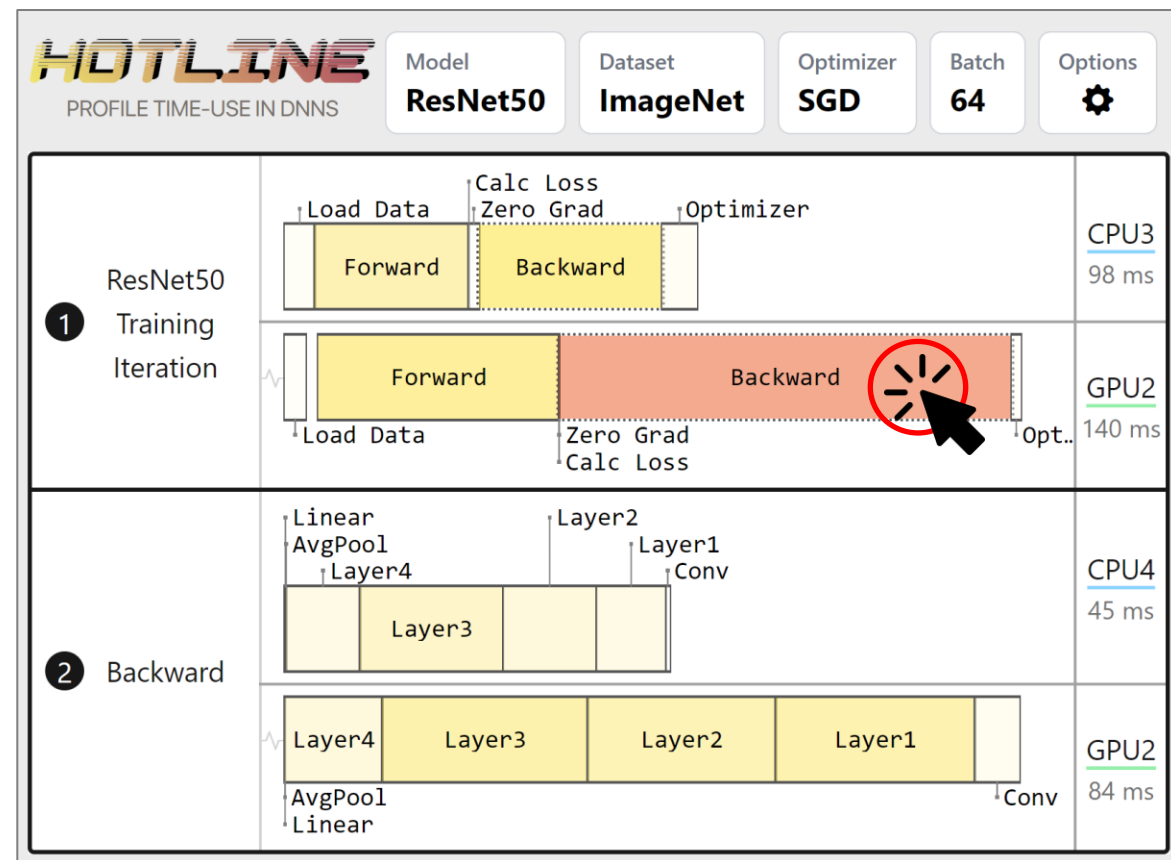
Runtime Trace



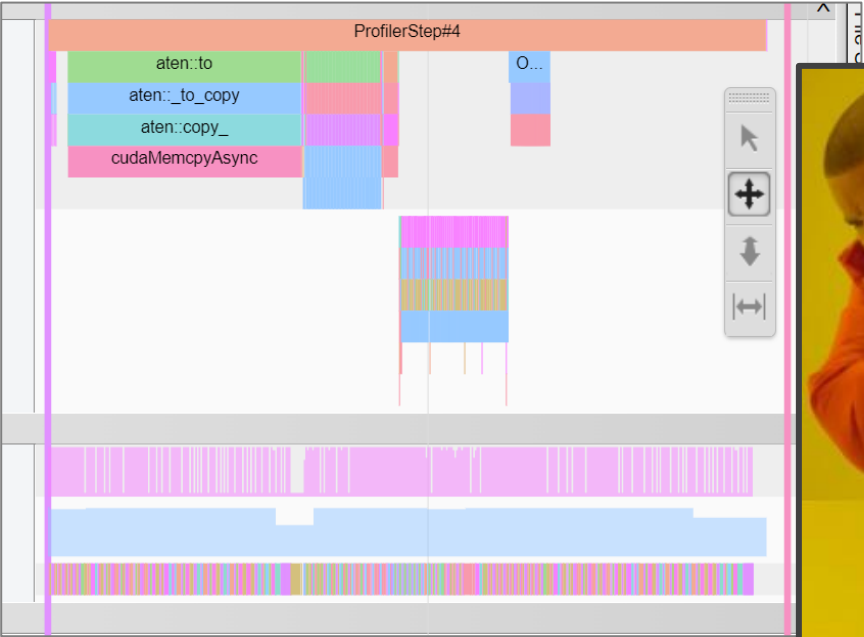
TensorBoard



HOTLINE



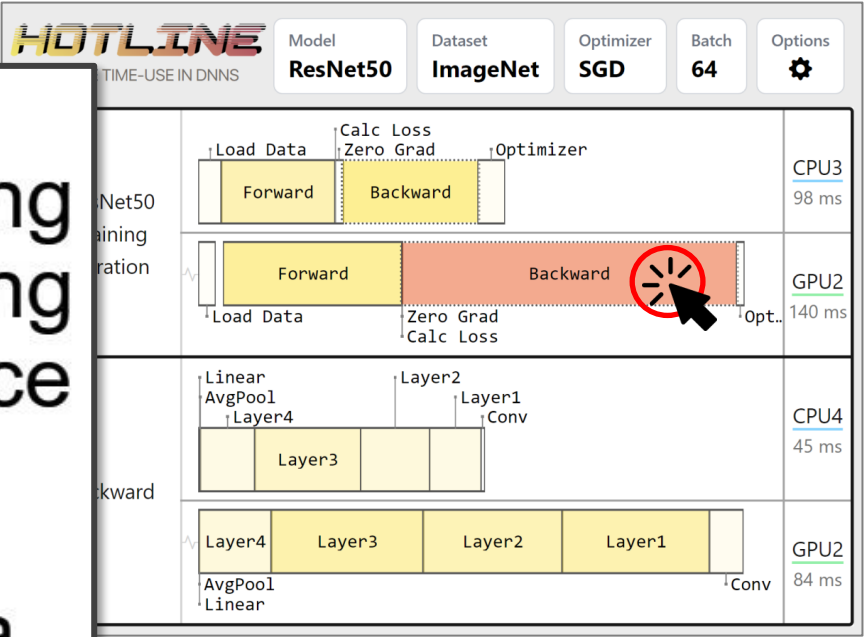
TensorBoard



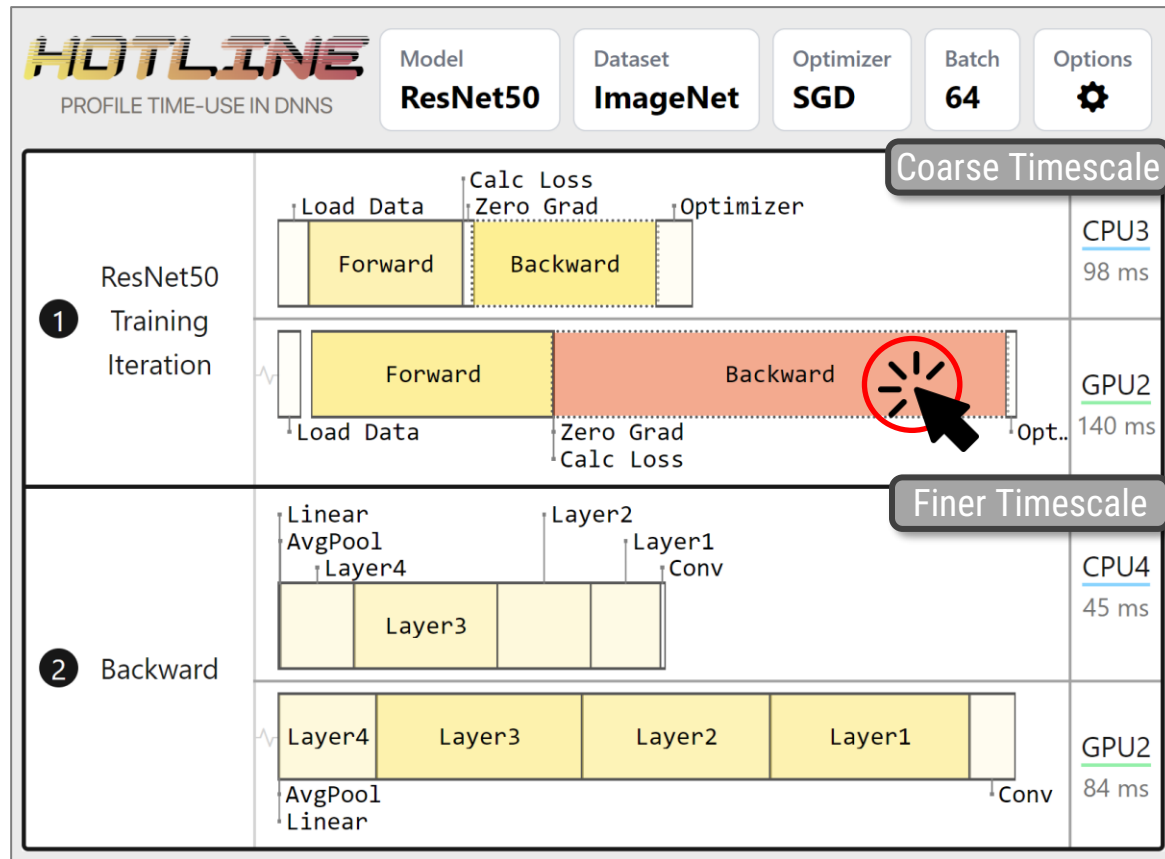
Displaying everything all at once



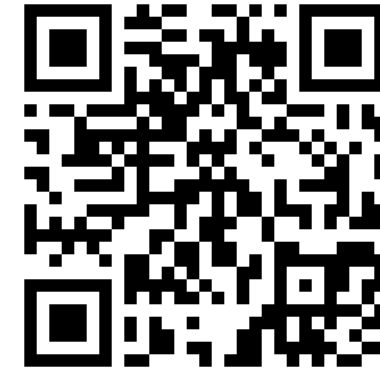
Using a multi-scale timeline



HOTLINE: Automatic Annotation and A Multi-Scale Timeline for Visualizing Time-Use in DNN Training



Try the demo!



<https://danielsnider.ca/hotline/demo>

EcoSystem
Efficient Computing Systems

