POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging

Shishir G. Patil

With Paras Jain, Prabal Dutta, Ion Stoica, Joseph Gonzalez

https://github.com/ShishirPatil/poet

ICML
International Conference
On Machine Learning

riselab
UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging

BERT on edge devices!

Shishir G. Patil

With Paras Jain, Prabal Dutta, Ion Stoica, Joseph Gonzalez

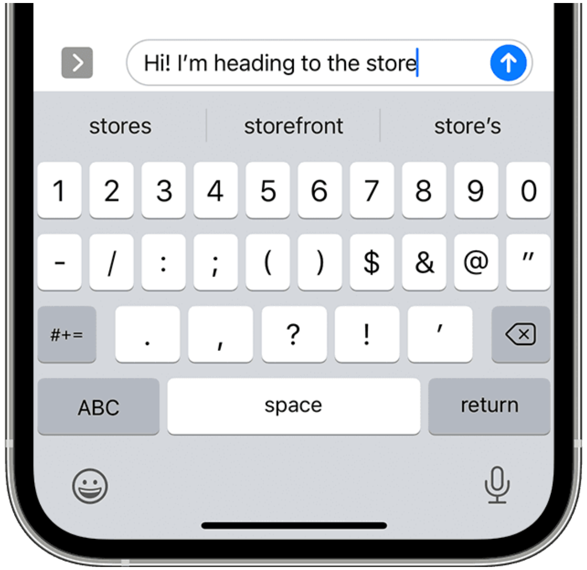
https://github.com/ShishirPatil/poet

ICML International Conference On Machine Learning

riselab UC Berkeley

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Model Personalization Adapts Models by Training on User Data to Improve Accuracy



Hi! I'm heading to the store

Autocompletion

Voice Recognition

Fitness Tracker

Model Fine-tuning – Train on Edge

Fine-tune on-device

Key Challenge: Limited memory for DNN training!

Pros:

+ guarantees user’s privacy as all data stays on their device

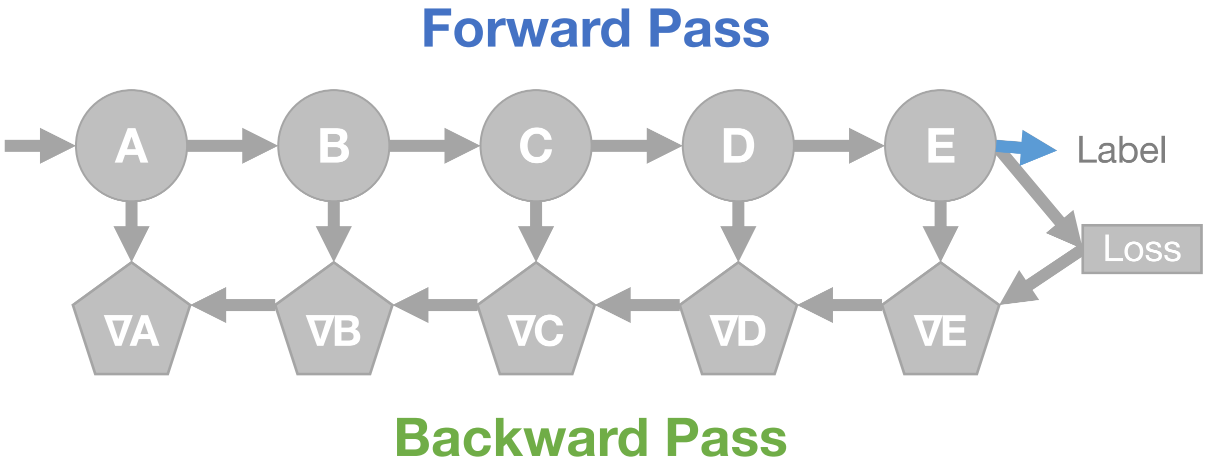
+ enables offline device operation

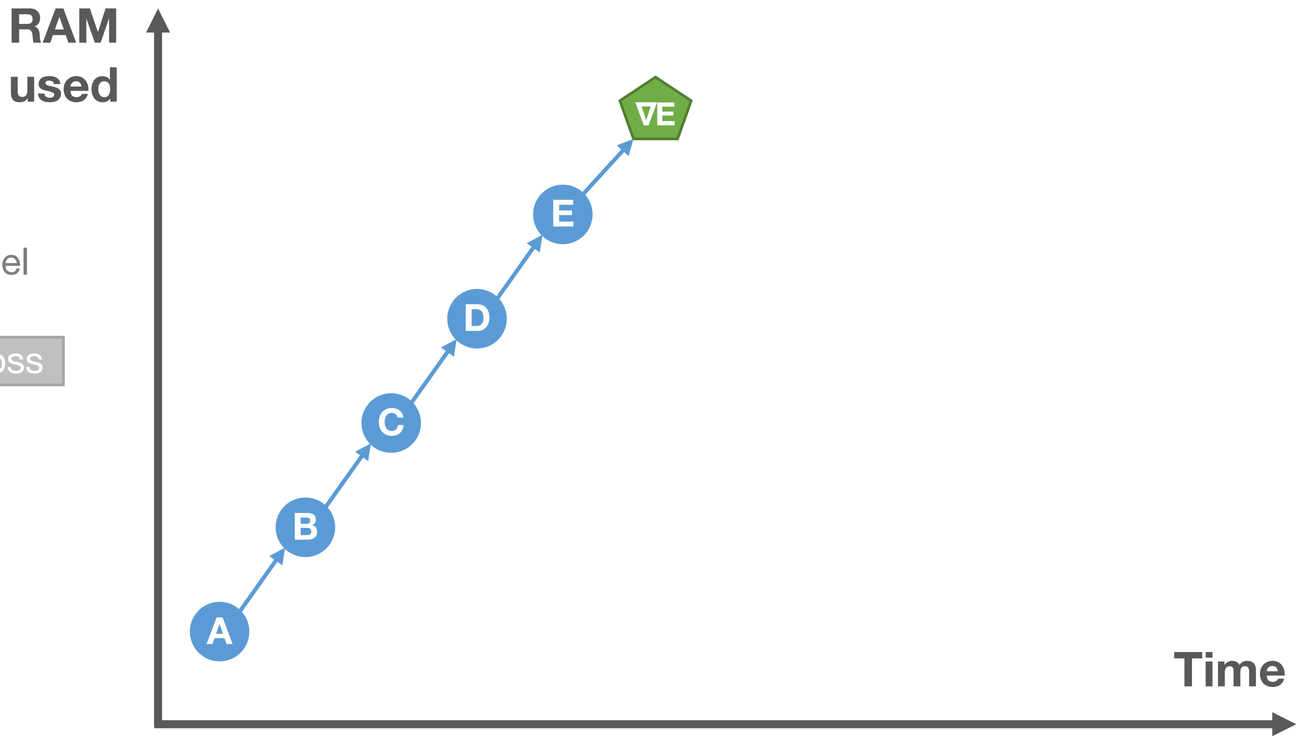
Cons:

- cannot train modern DNNs on edge devices

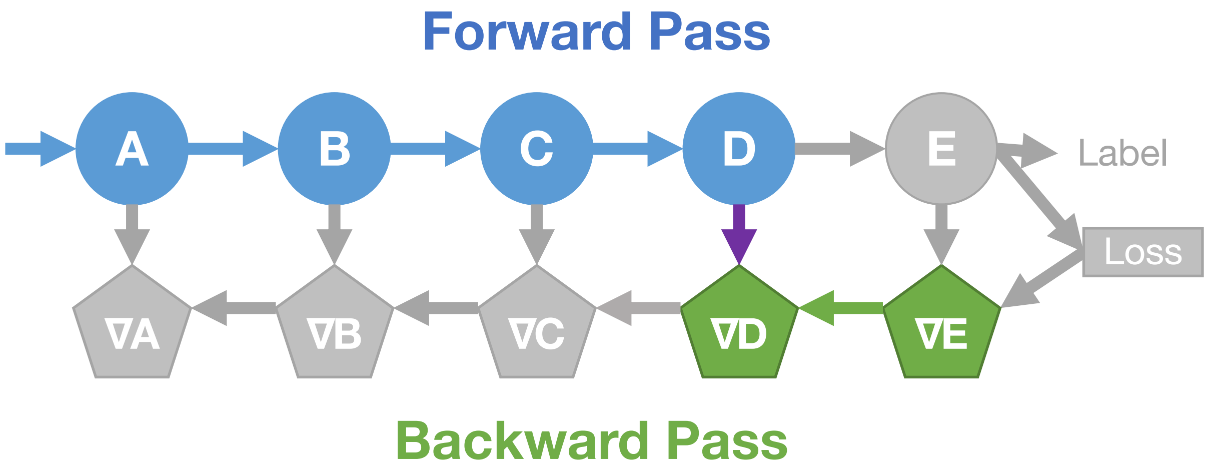
How to reduce the memory and energy requirements of ML training for modern DNN architectures within the constraints of edge devices?

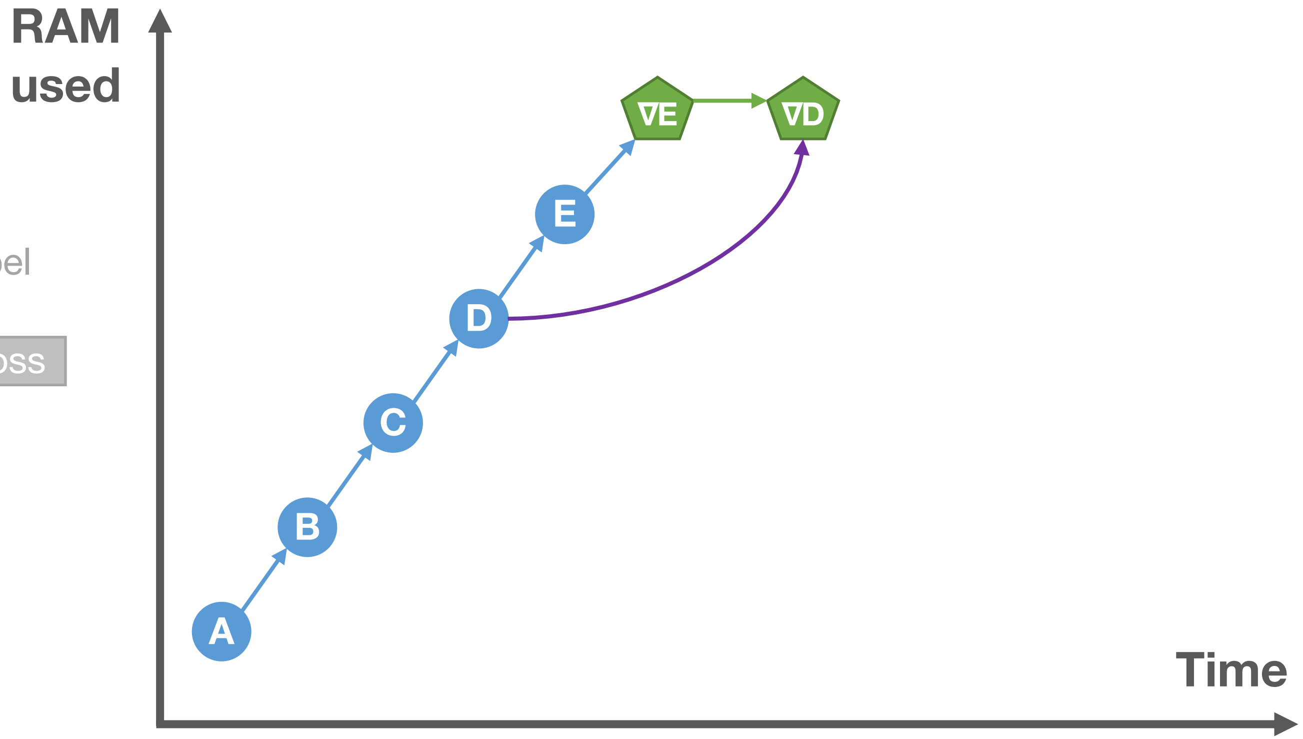
Training is Memory Intensive since Activation from Forward Pass Need to be Stored for Backpropagation



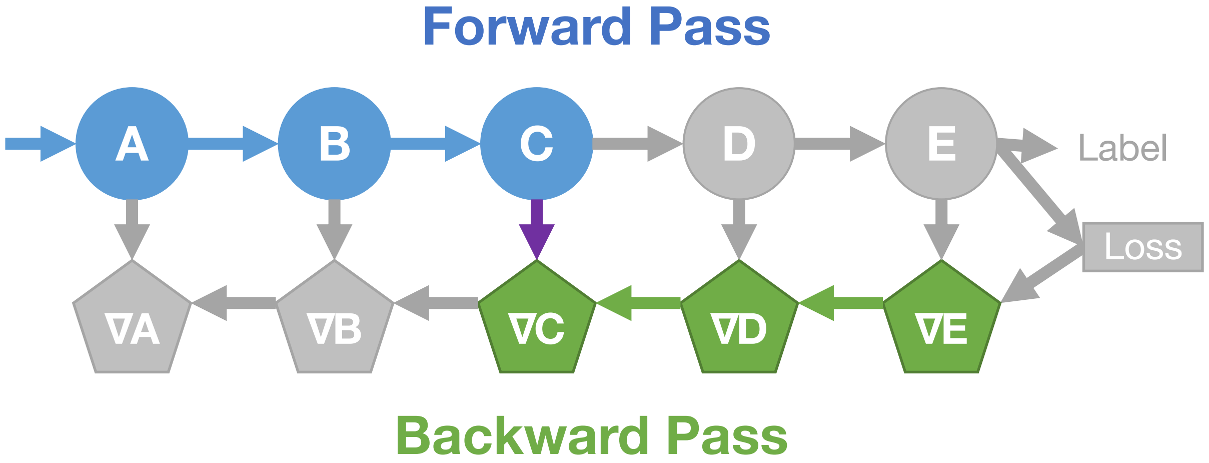


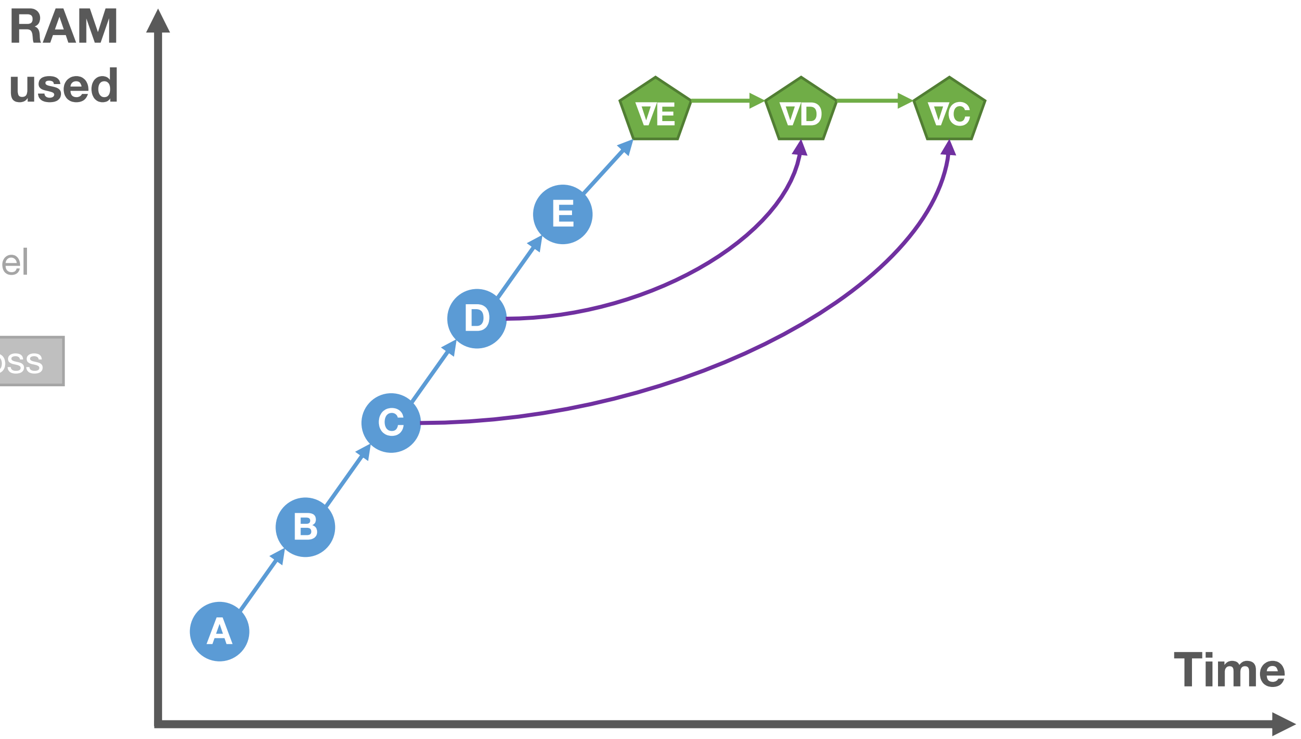
Training is Memory Intensive since Activation from Forward Pass Need to be Stored for Backpropagation



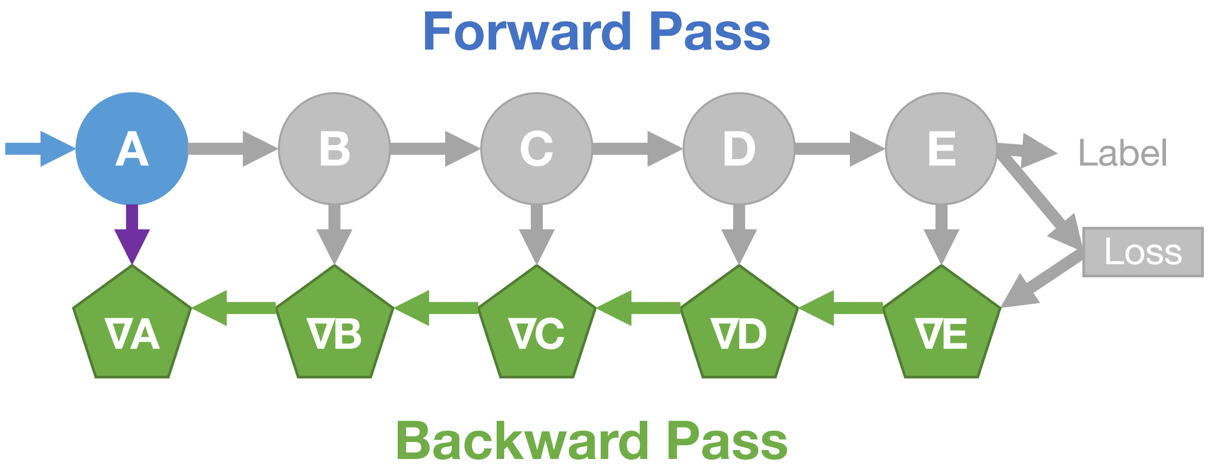


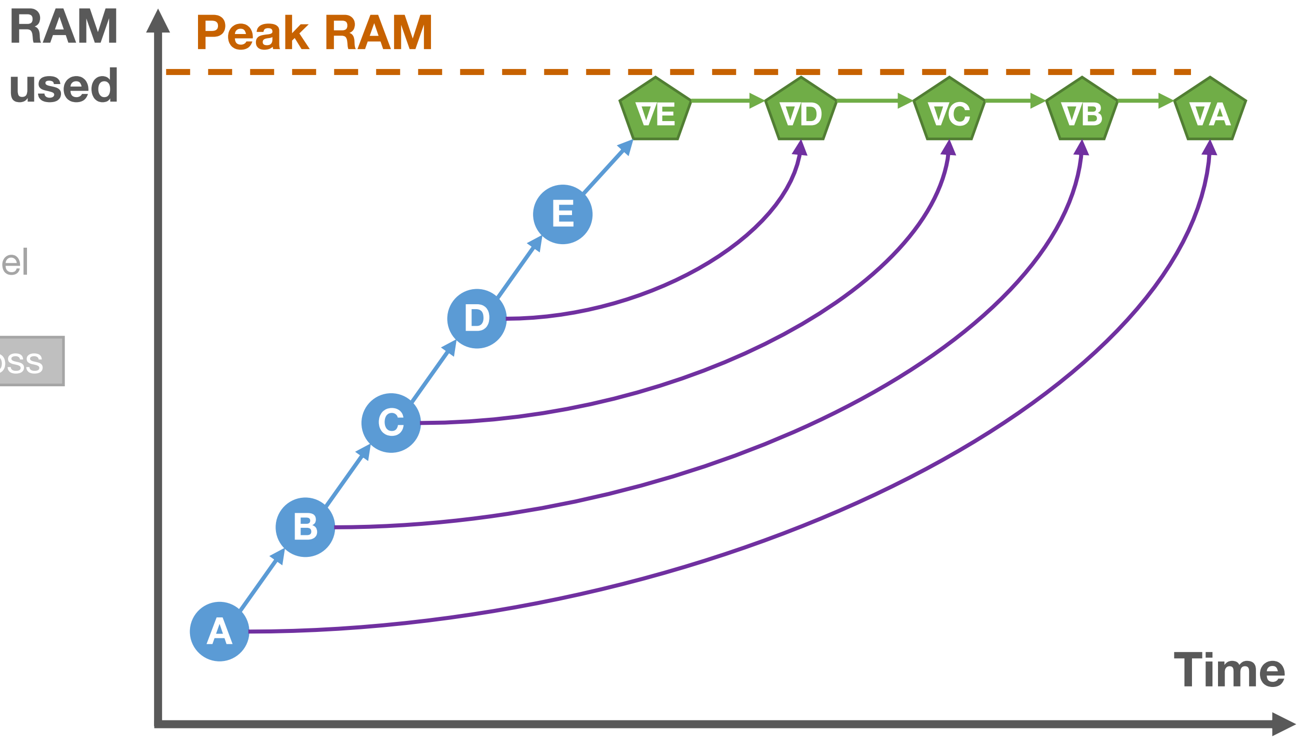
Training is Memory Intensive since Activation from Forward Pass Need to be Stored for Backpropagation



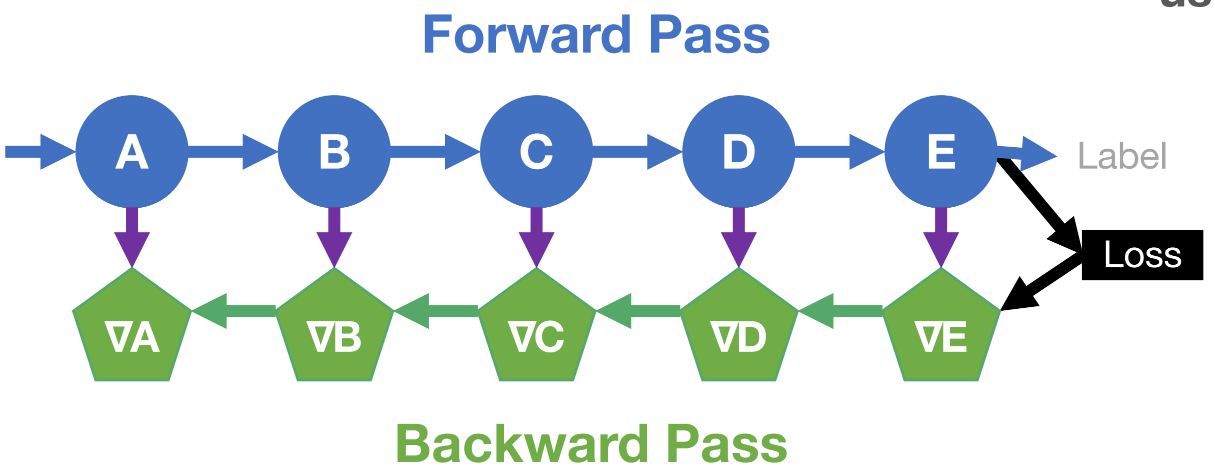


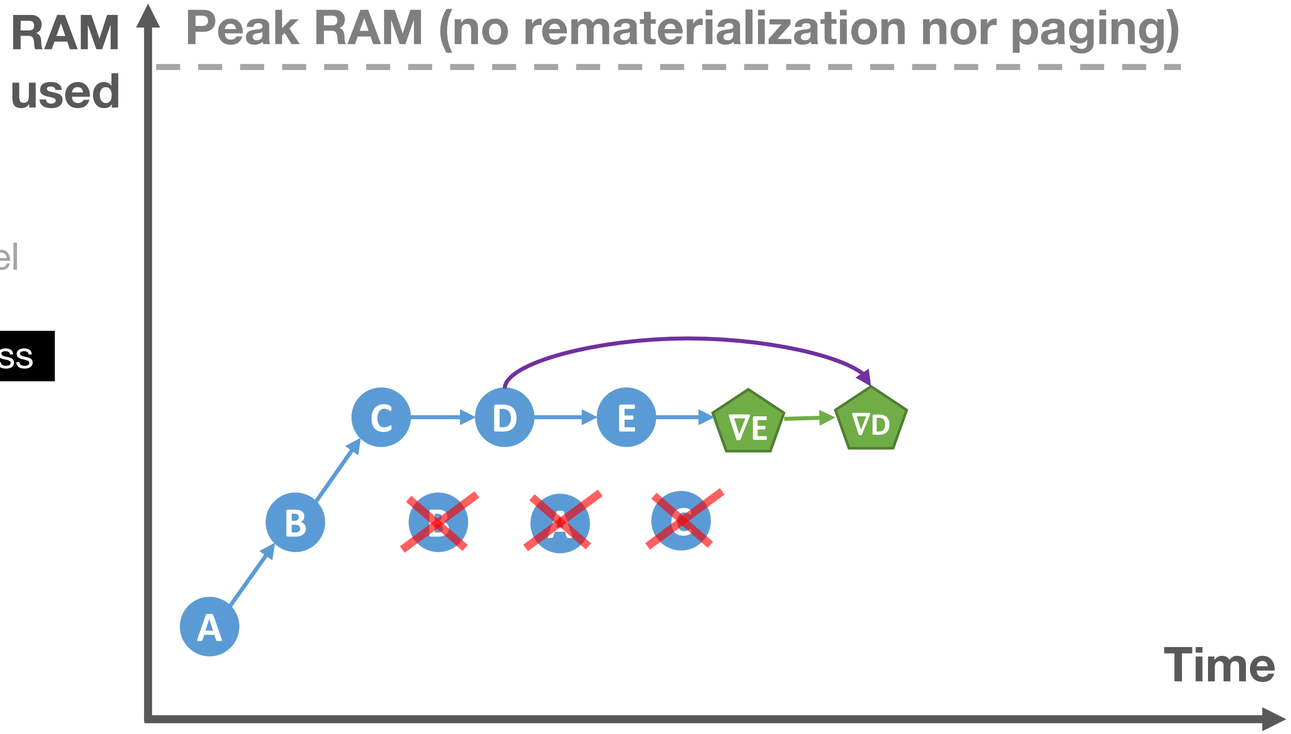
Training is Memory Intensive since Activation from Forward Pass Need to be Stored for Backpropagation





Rematerialization and Paging: Two Techniques to Reduce Memory Consumption

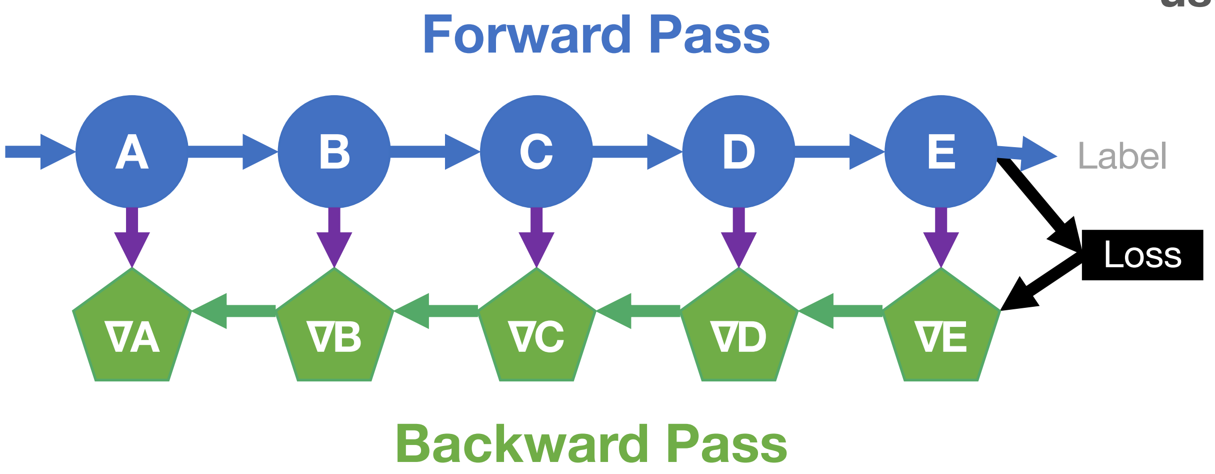


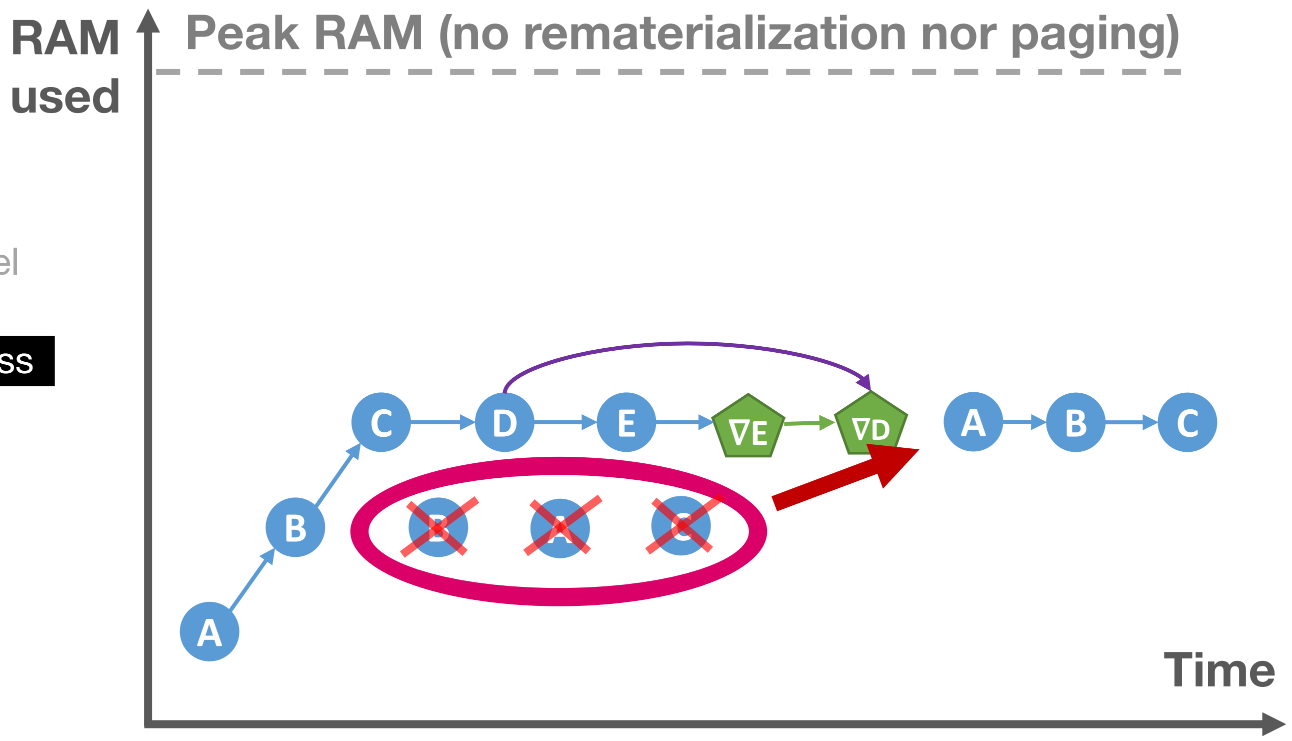


Rematerialization:

Free early & recompute

Rematerialization and Paging: Two Techniques to Reduce Memory Consumption

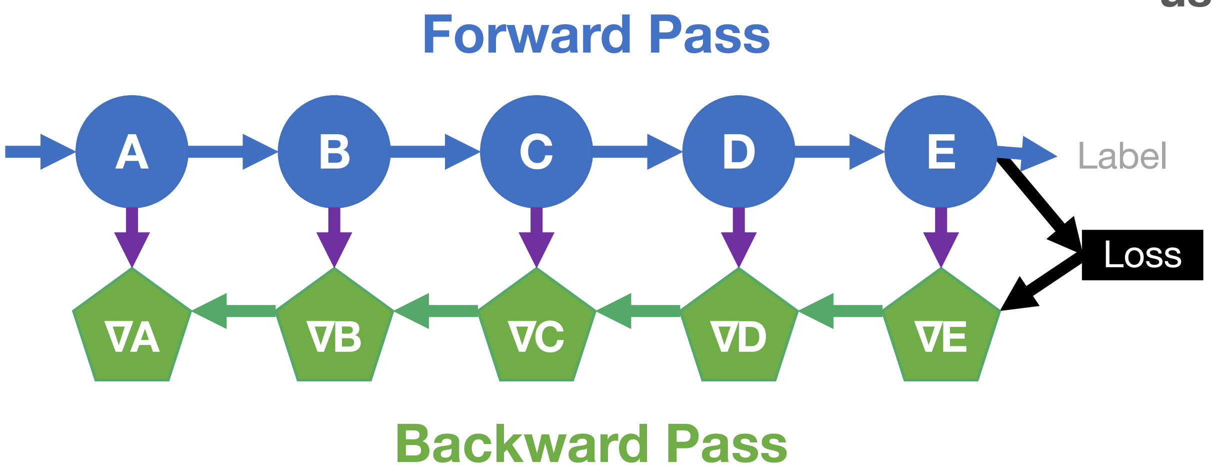


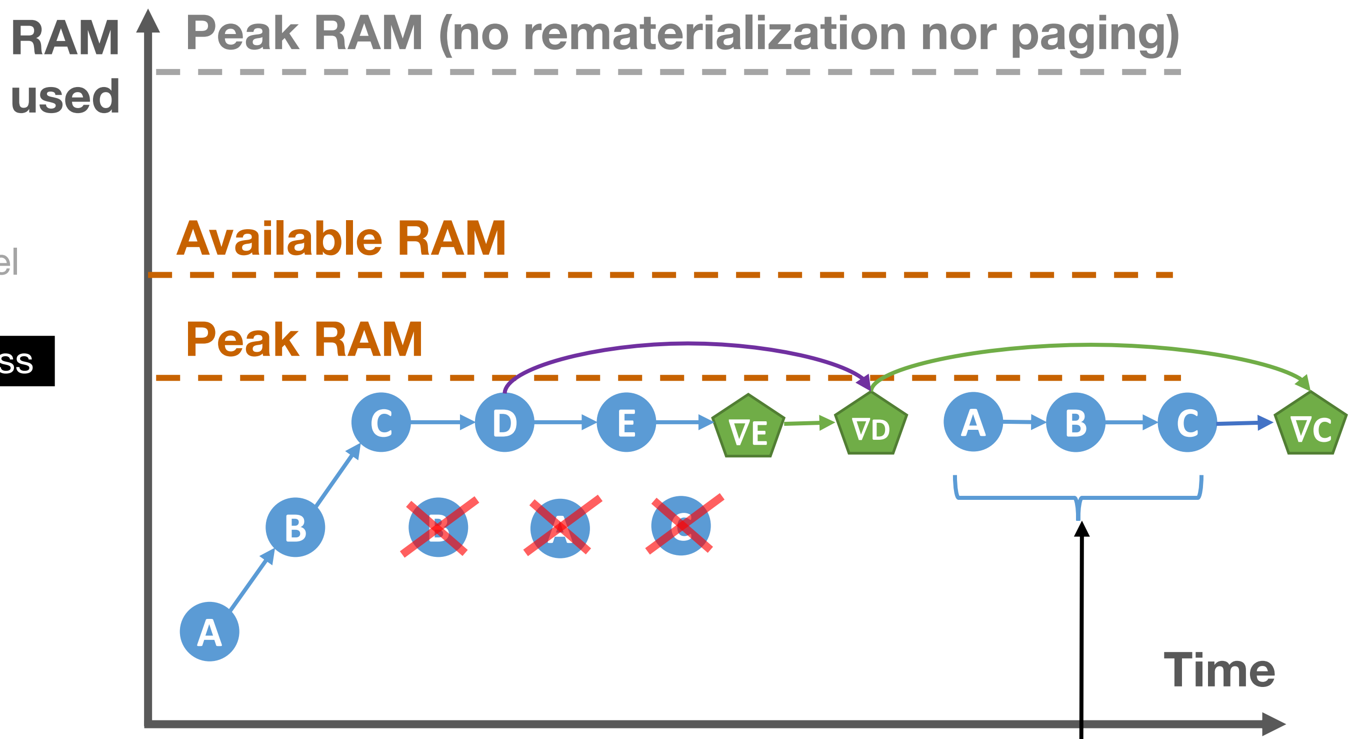


Rematerialization:

Free early & recompute

Rematerialization and Paging: Two Techniques to Reduce Memory Consumption



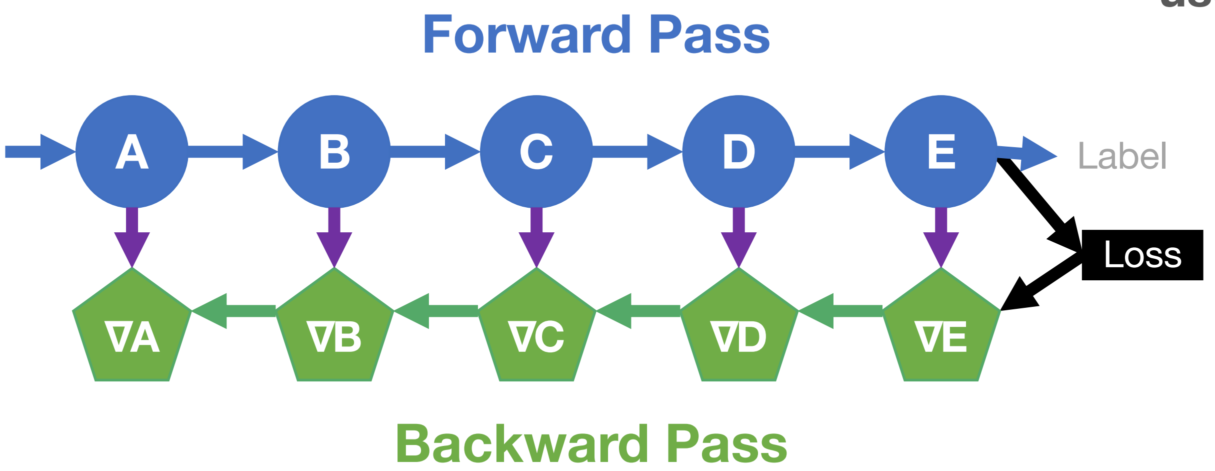


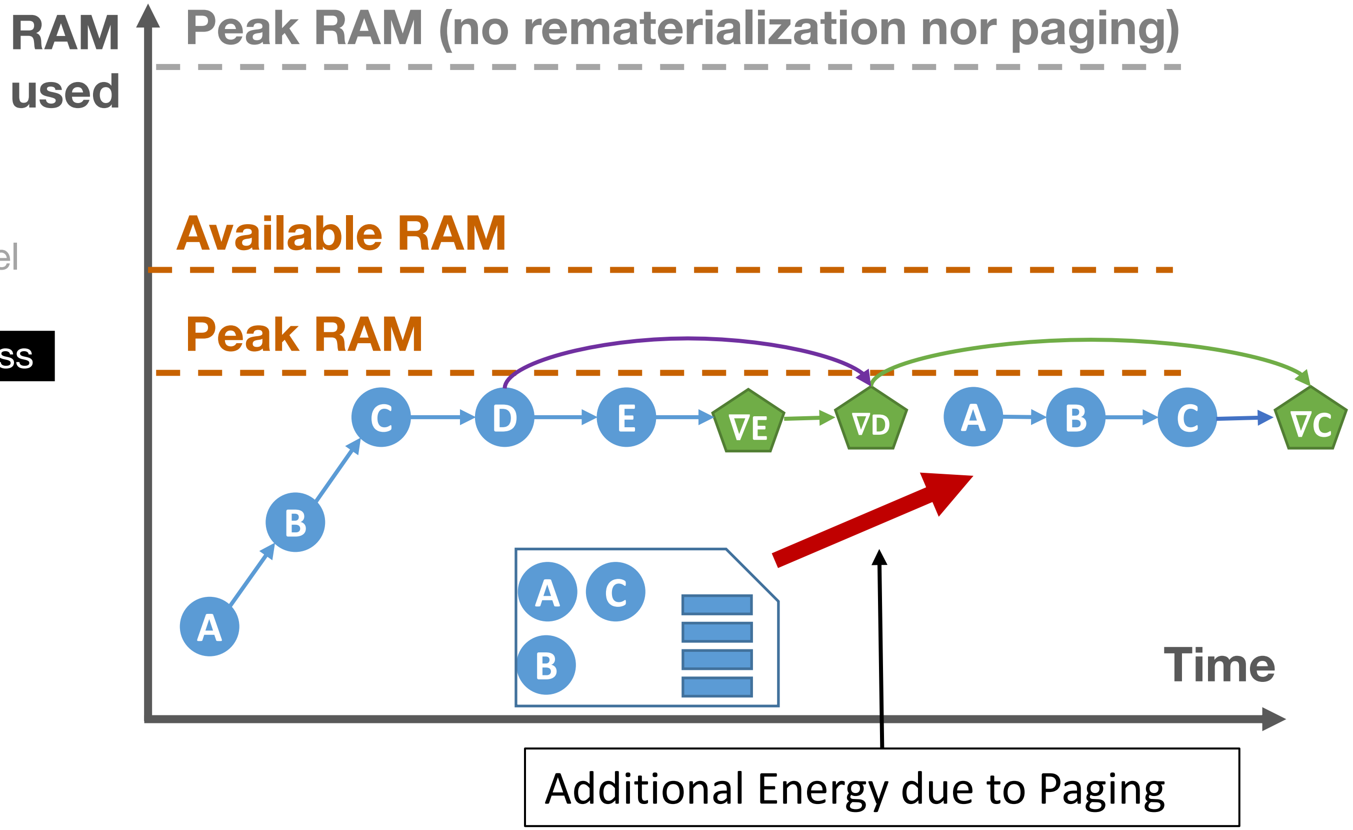
Additional Energy and runtime due to recomputation!

Rematerialization:

Free early & recompute

Rematerialization and Paging: Two Techniques to Reduce Memory Consumption



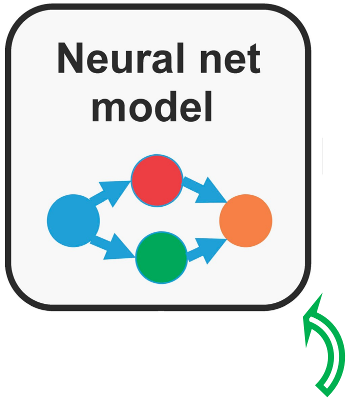


RAM used
Peak RAM (no rematerialization nor paging)
Available RAM
Peak RAM
A
B
C
D
E
∇E
∇D
A
B
C
∇C
A
C
B
Time
Additional Energy due to Paging

Paging:

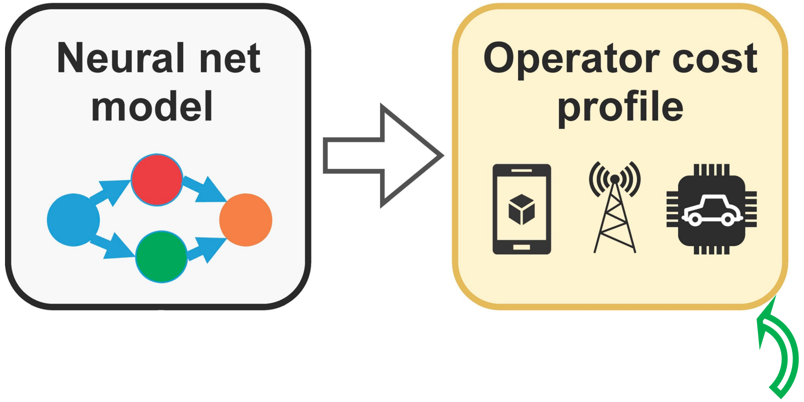
Page-out to secondary storage and page-in Just-in-Time!

POET: Private Optimal Energy Training



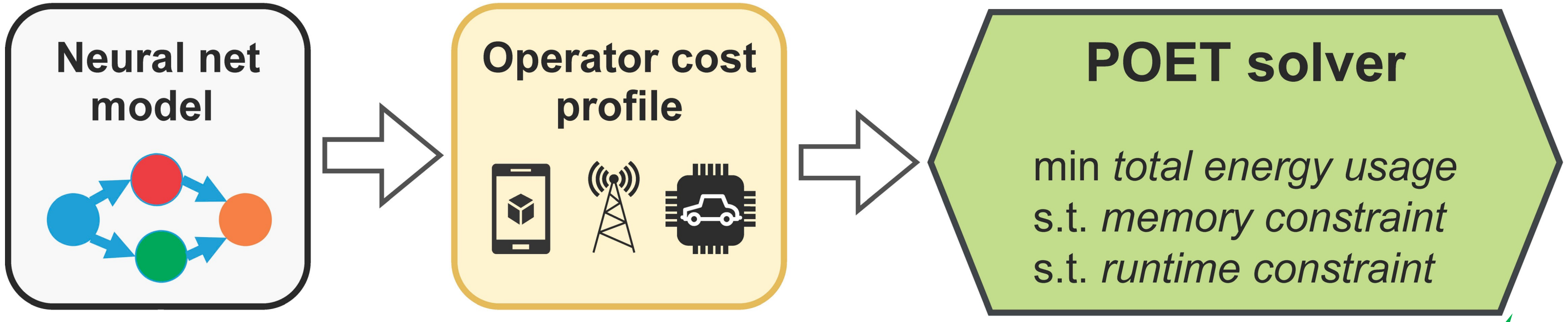
{SOTA ML model,
memory and runtime
constraints}

POET: Private Optimal Energy Training



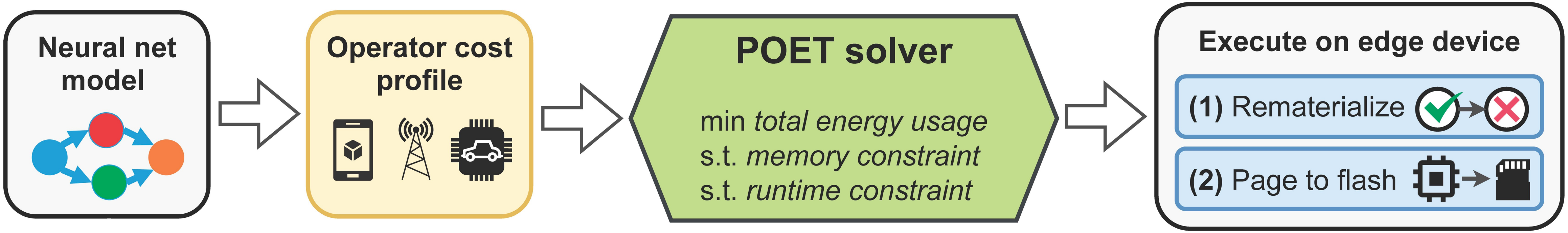
Accurate cost profile of ML operators on target edge platform!

POET: Private Optimal Energy Training



Incorporate *memory* and *runtime* constraints into a Mixed Integer Linear Program (MILP) formulation

POET: Private Optimal Energy Training



Neural net model

Operator cost profile

POET solver

min *total energy usage*
s.t. *memory constraint*
s.t. *runtime constraint*

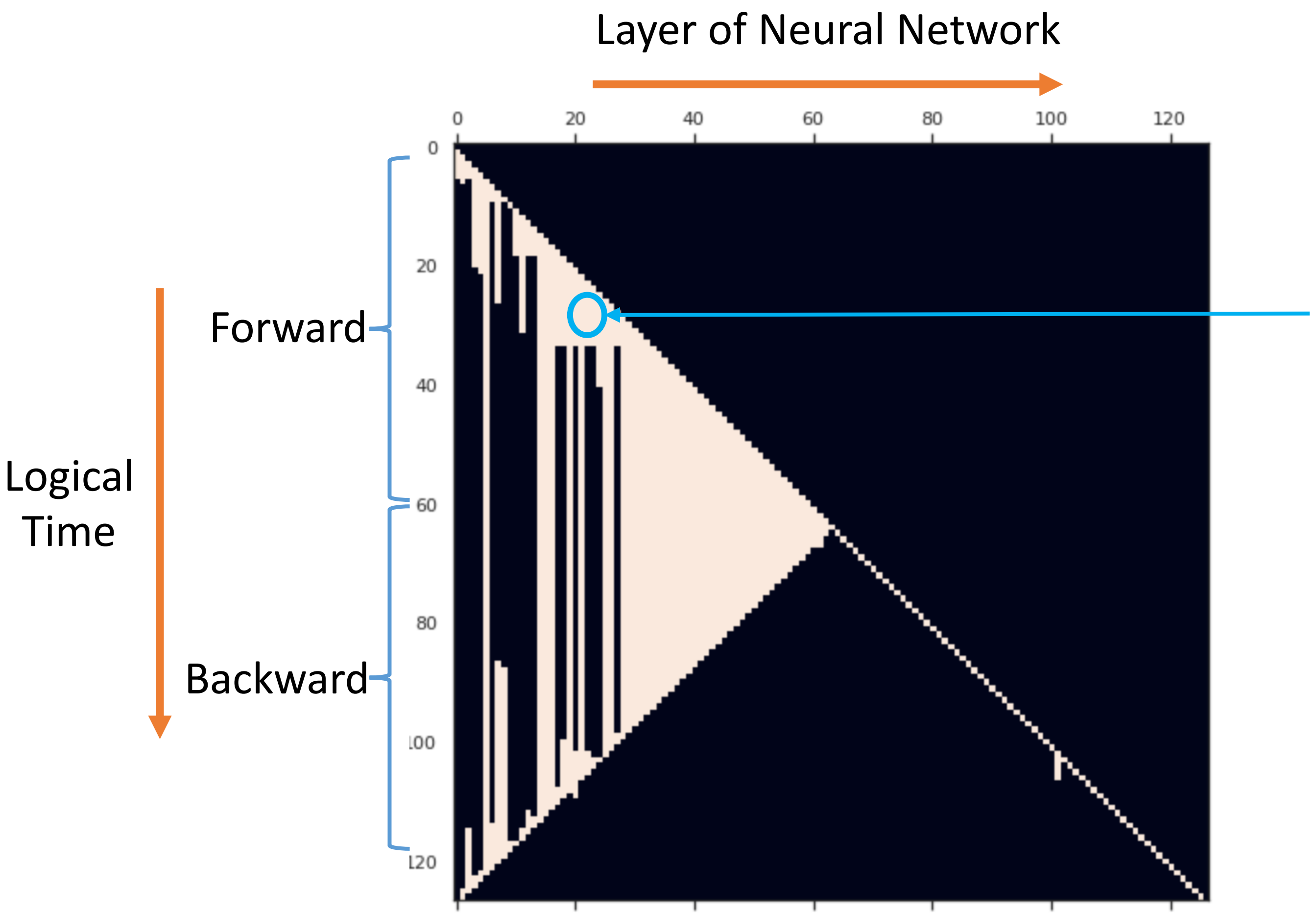
Execute on edge device

(1) Rematerialize

(2) Page to flash

POET finds a provably optimal solution through integrated rematerialization and paging.

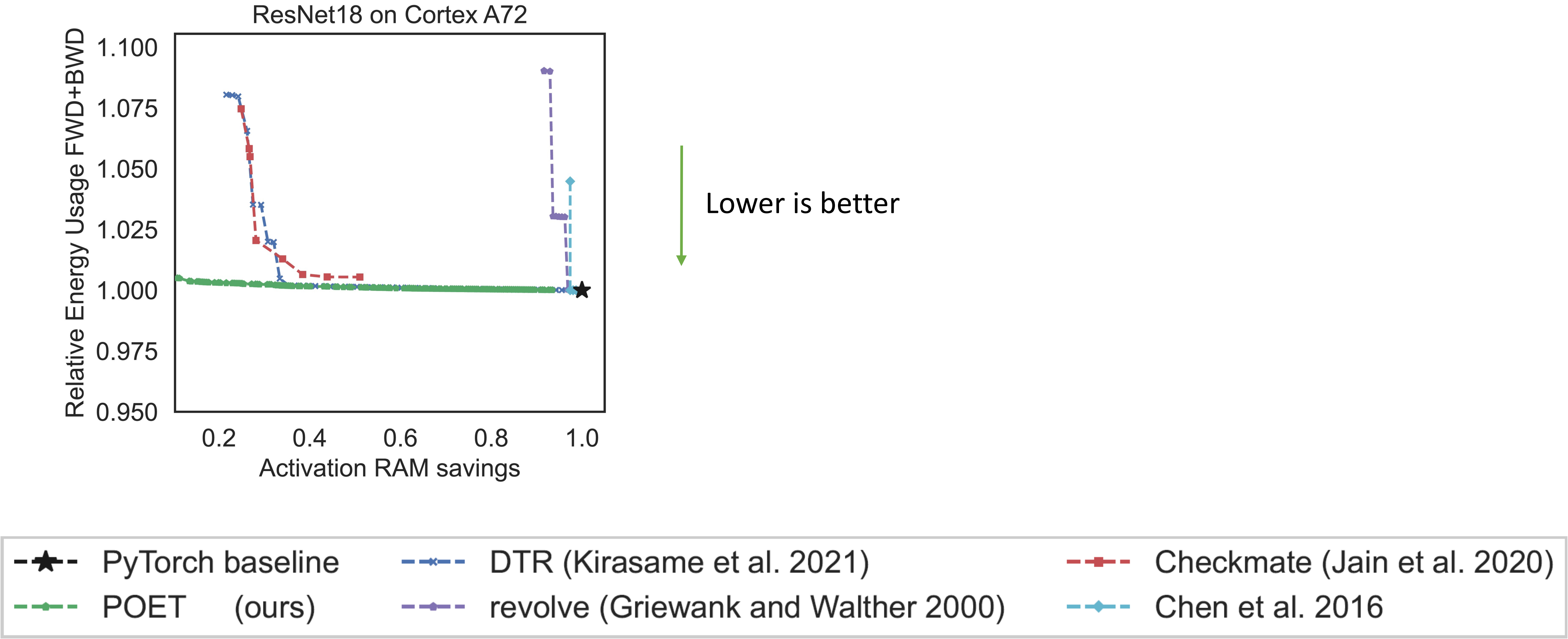
POET: Private Optimal Energy Training



Pixelated box indicates activation tensor for Layer ‘*l*’ is resident in RAM at timestep ‘*t*’

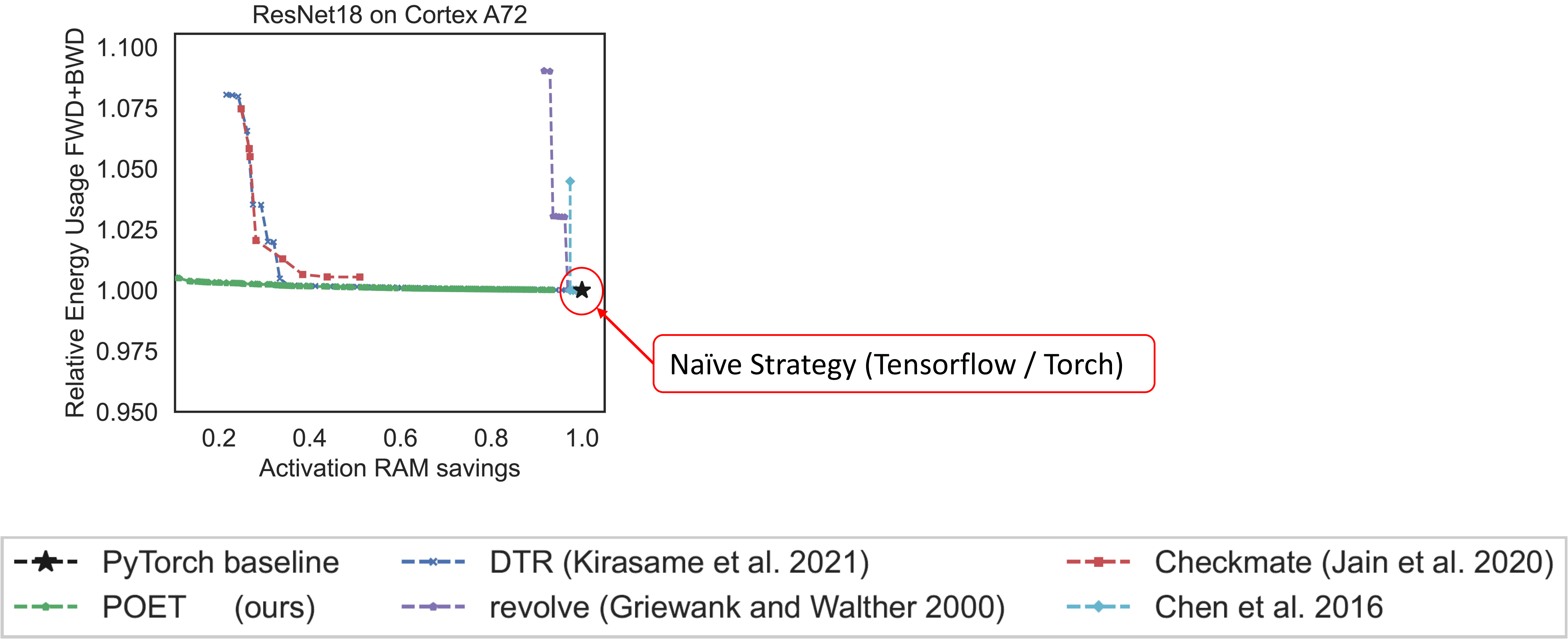
POET’s integrated rematerialization and paging search space finds advanced solutions that are not possible through simple heuristics.

Result: POET lowers energy consumption and allows training large models previously not possible!



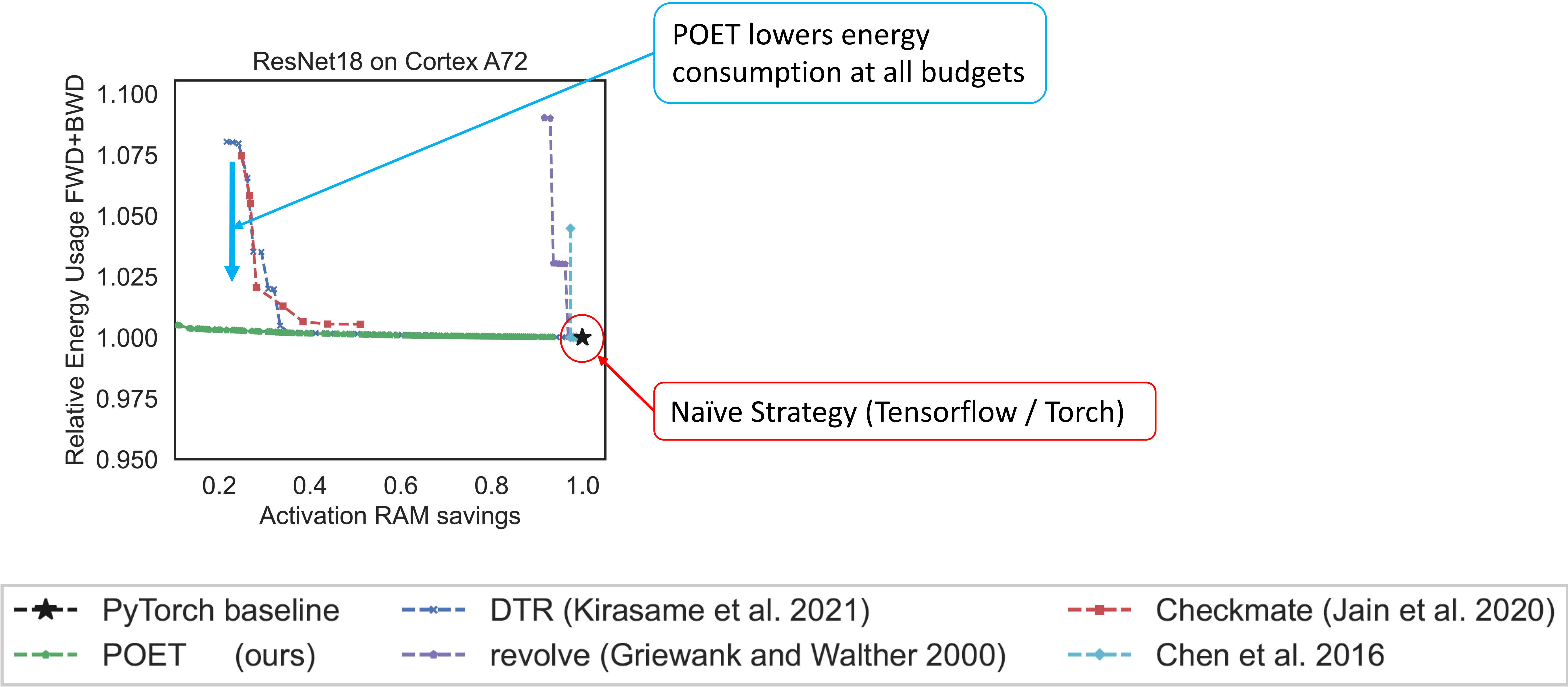
Lower is better

Result: POET lowers energy consumption and allows training large models previously not possible!



Naïve Strategy (Tensorflow / Torch)

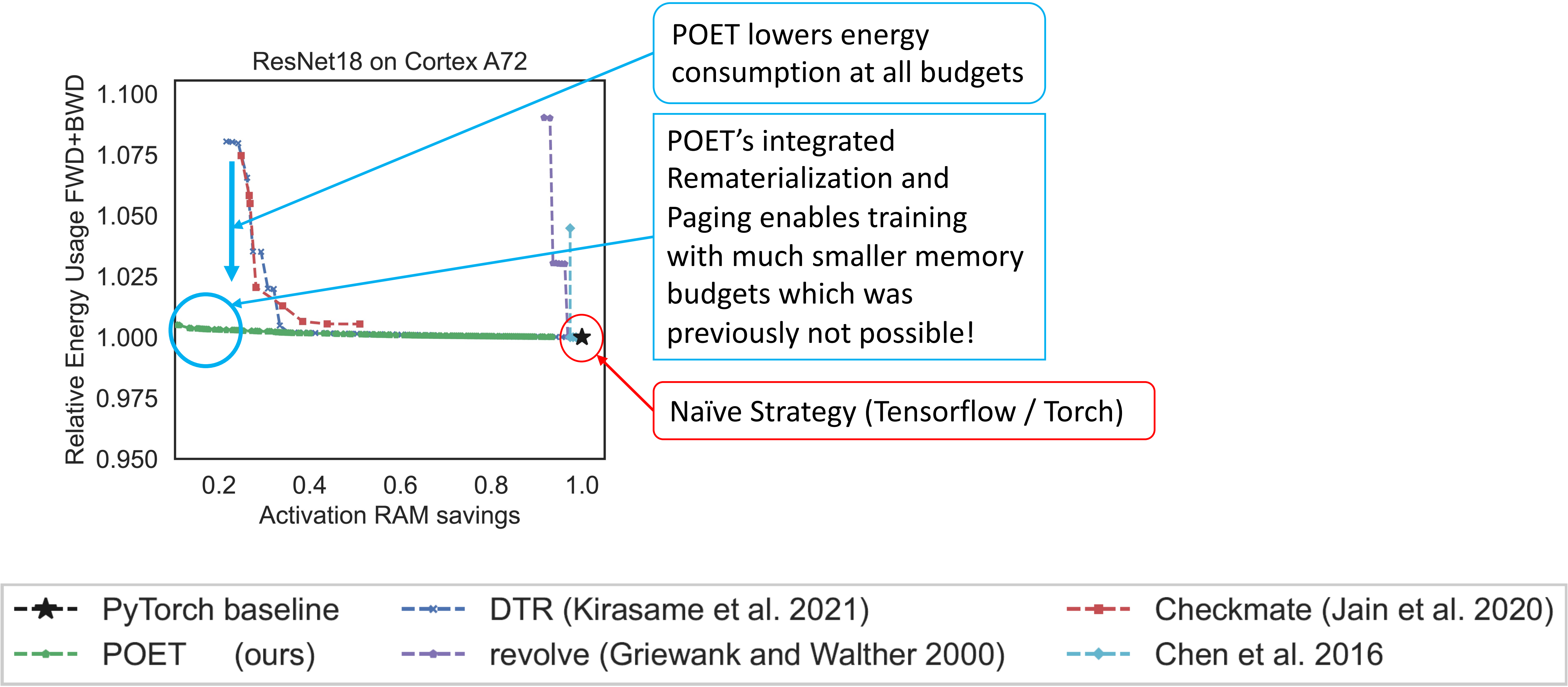
Result: POET lowers energy consumption and allows training large models previously not possible!



POET lowers energy consumption at all budgets

Naïve Strategy (Tensorflow / Torch)

Result: POET lowers energy consumption and allows training large models previously not possible!

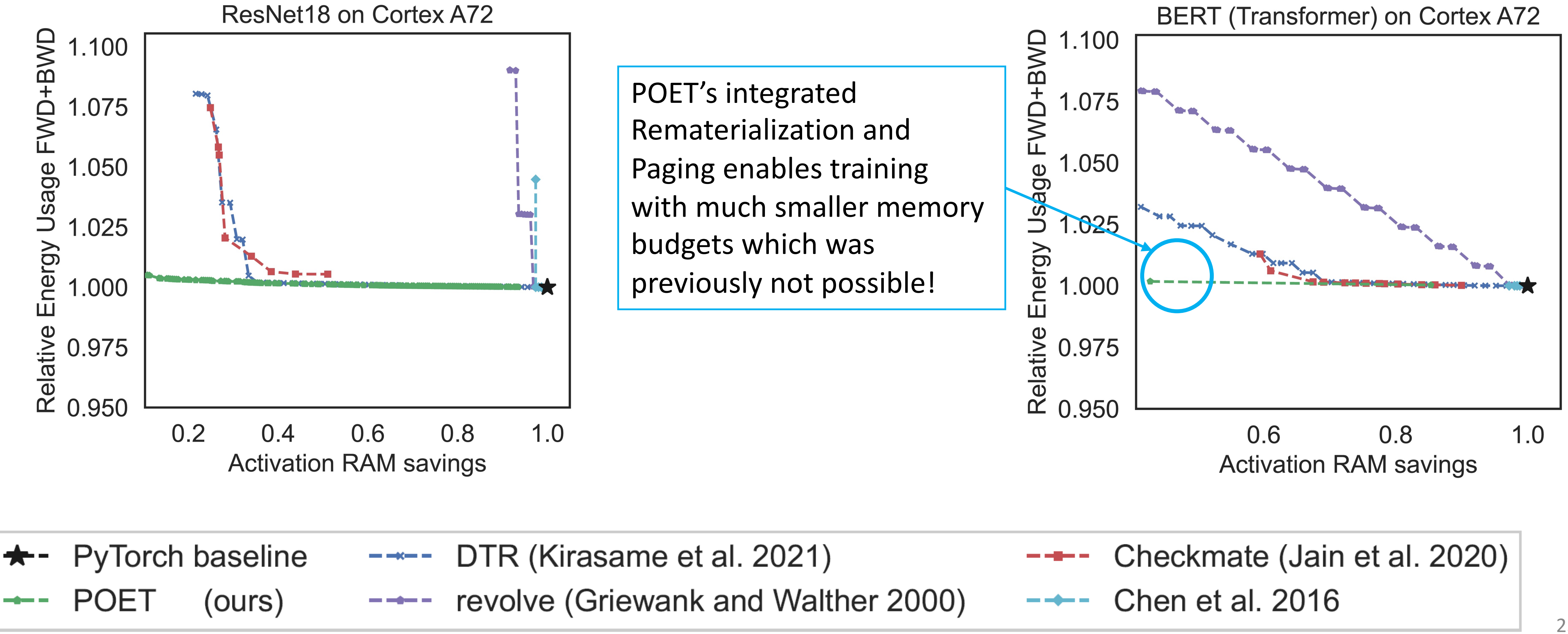


POET lowers energy consumption at all budgets

POET’s integrated Rematerialization and Paging enables training with much smaller memory budgets which was previously not possible!

Naïve Strategy (Tensorflow / Torch)

Result: POET lowers energy consumption and allows training large models previously not possible!



POET's integrated Rematerialization and Paging enables training with much smaller memory budgets which was previously not possible!

POET – Private Optimal Edge Training

Conclusion

- POET enables training SOTA DNN models locally on memory-constrained edge devices.
- POET's fine grained profiling results in accurate cost profiles.
- POET's MILP formulation finds the optimal training schedule through integrated **rematerialization** and **paging.**

https://poet.cs.berkeley.edu

https://github.com/ShishirPatil/poet

shishirpatil@berkeley.edu