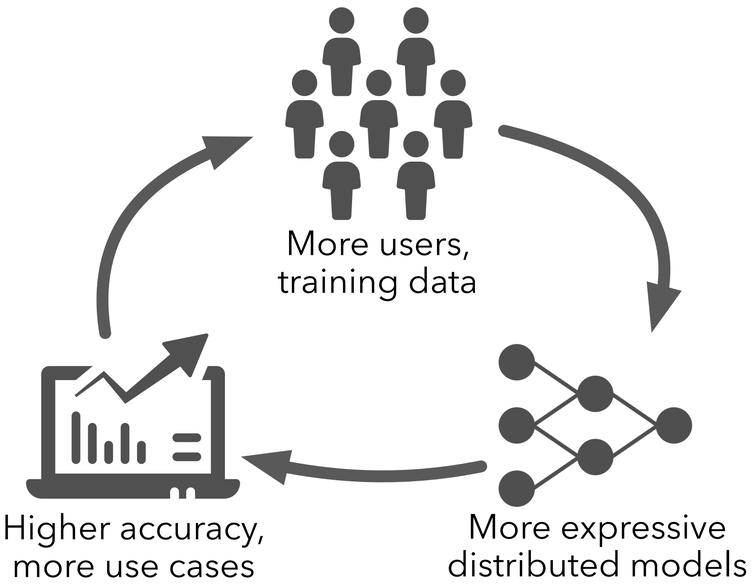
MemFlow: Memory-Aware Distributed Deep Learning

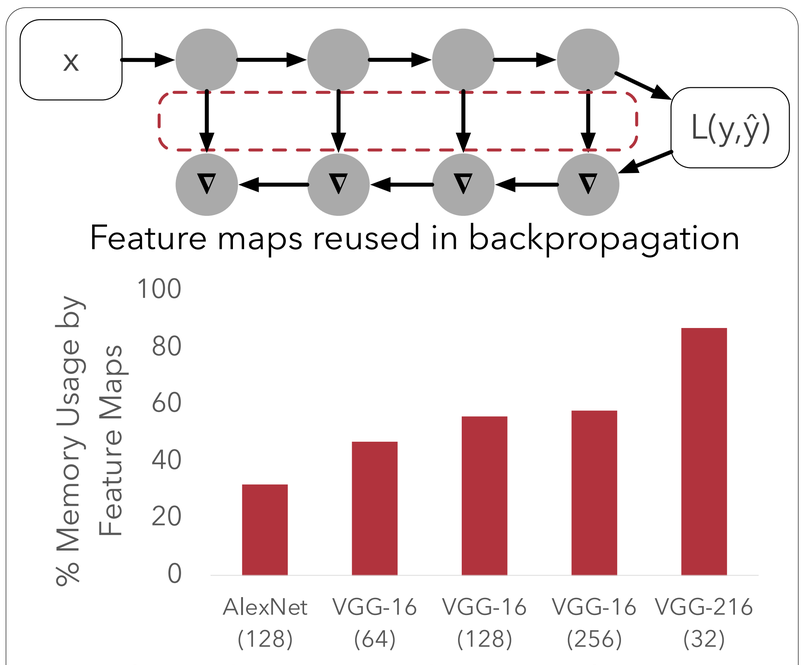
HARVARD John A. Paulson School of Engineering and Applied Sciences

Neil Band | Advised by Abdul Wasay, Stratos Idreos

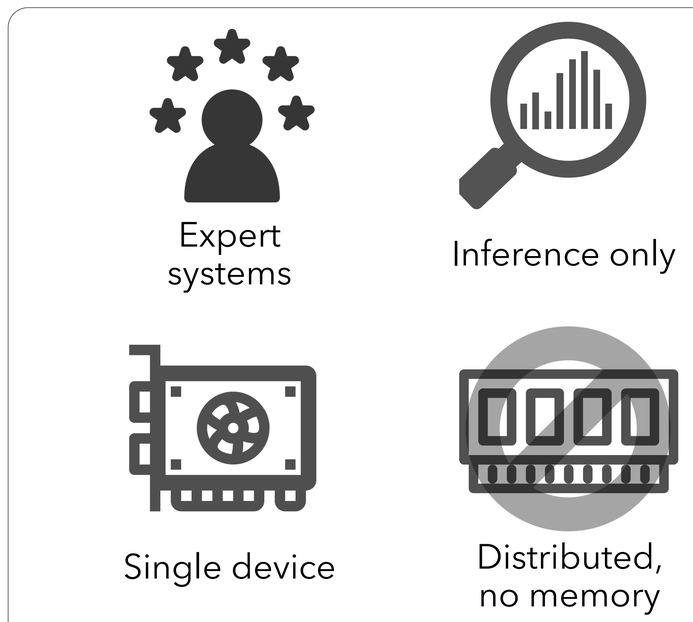
Motivation



Increasing memory requirement of deep learning

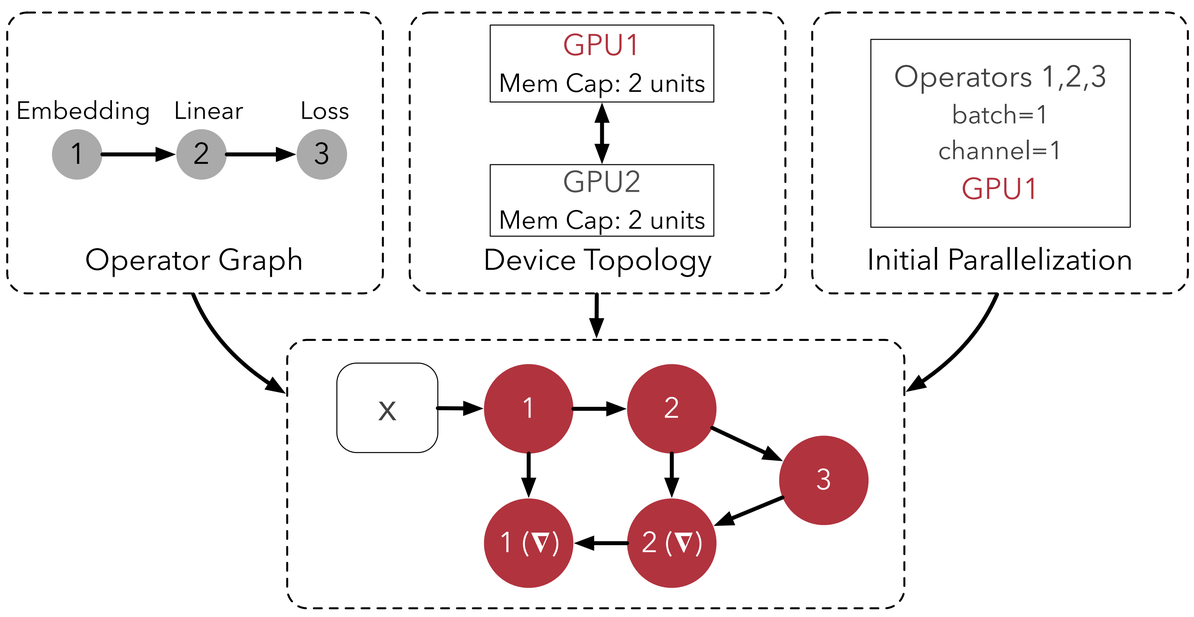


Bottleneck: intermediate result storage

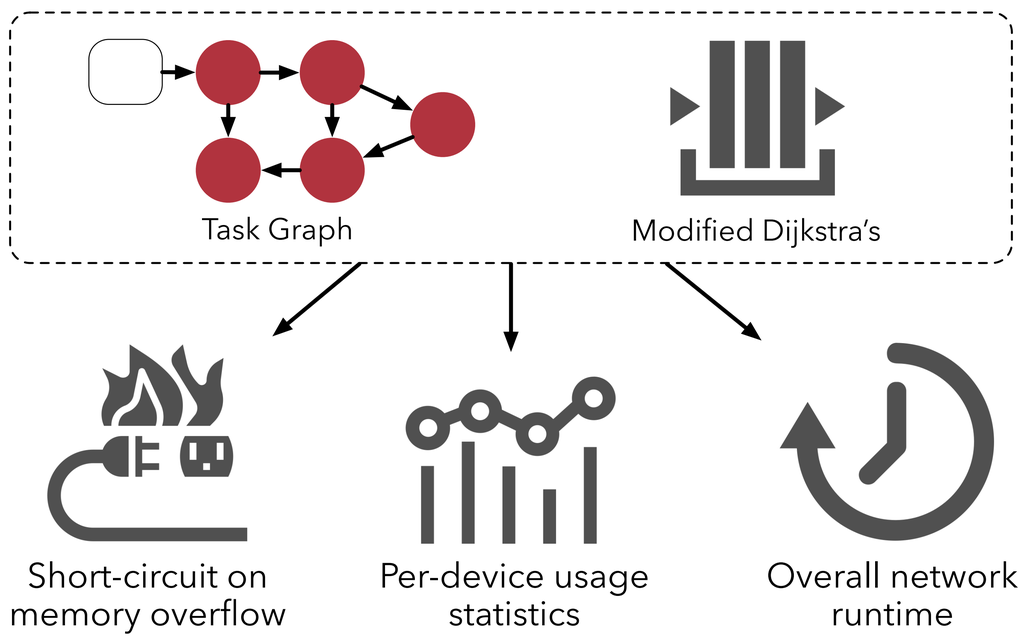


Problem: no general memory-aware framework

MemFlow

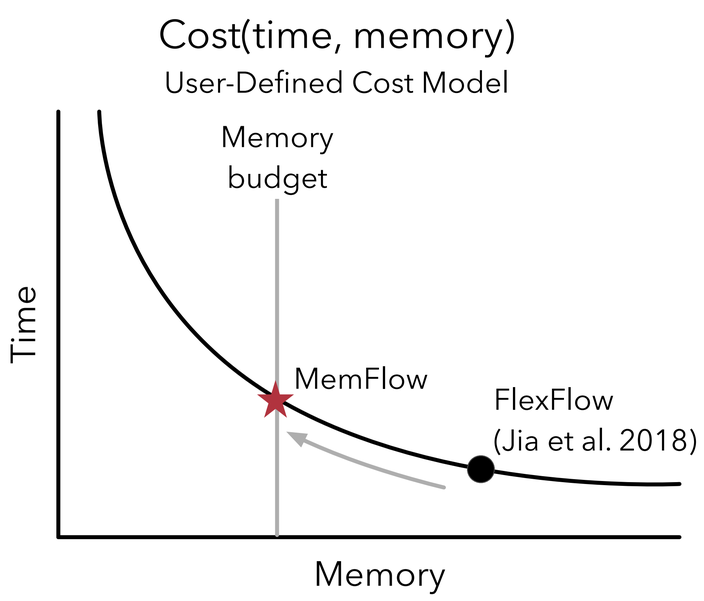


Step 1: Construct a task graph given a deep learning model and hardware

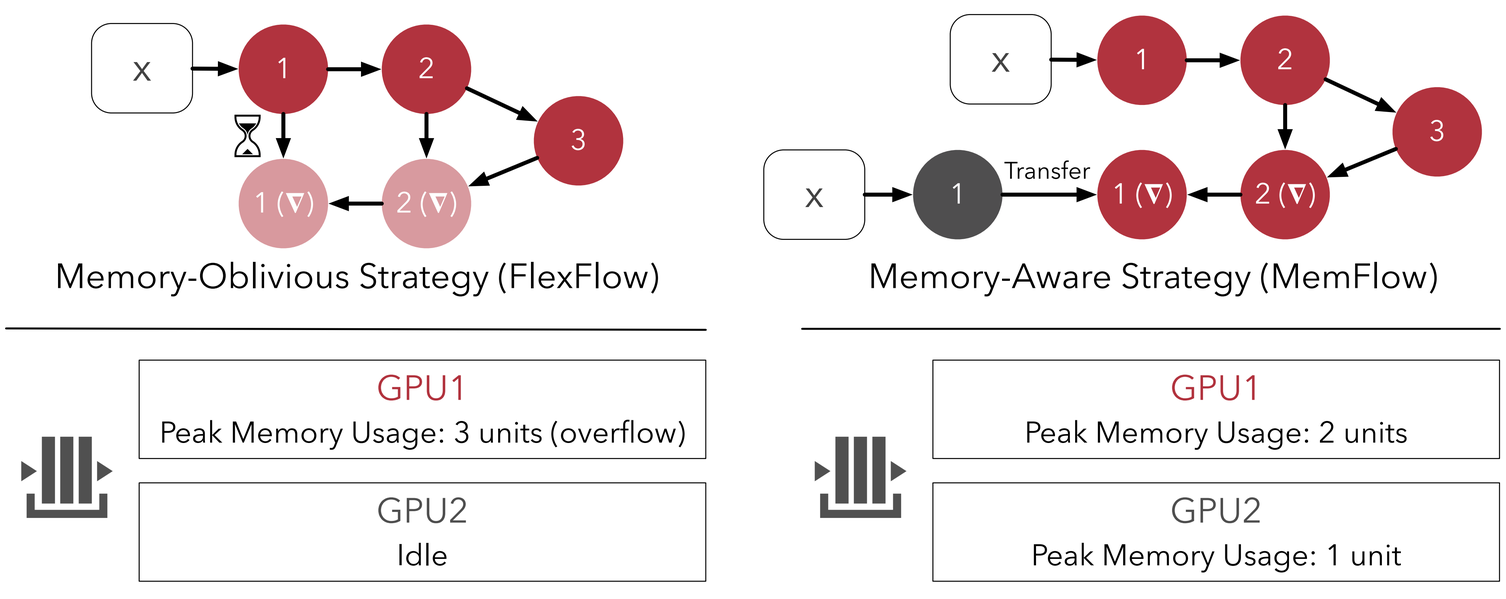


Step 2: Evaluate the task graph using a memory-aware simulator

MCMC Search



Step 3: Perform MCMC search with a user-defined cost model

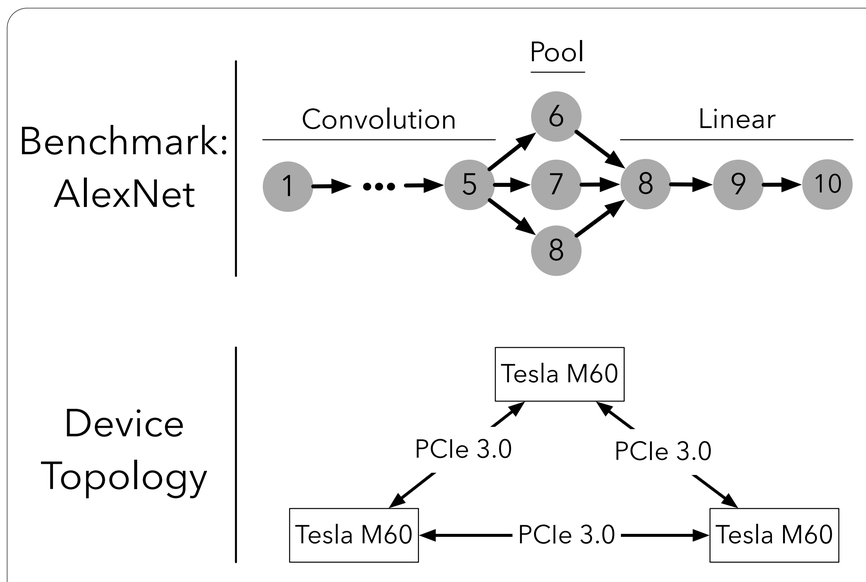


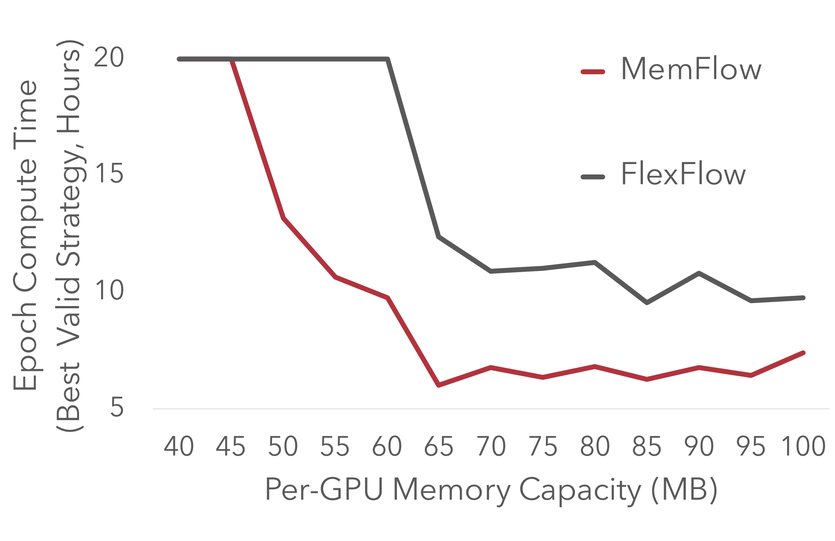
Example: Duplicate nodes across devices to exploit favorable time-memory tradeoff

Use GPU2 space to avoid cold storage of Task 1 intermediate result on GPU1

Recomputation

Evaluation





MemFlow locates valid and superior parallelization strategies for a wide range of memory limits

Application