UPSCALE:

Unconstrained Channel Pruning

Alvin Wan, Hanxiang Hao, Kaushik Patnaik, Yueyang Xu, Omer Hadad, David Guera, Zhile Ren, Qi Shan

International Conference on Machine Learning 2023 | Apple Inc.

Problem

Intuition

Method

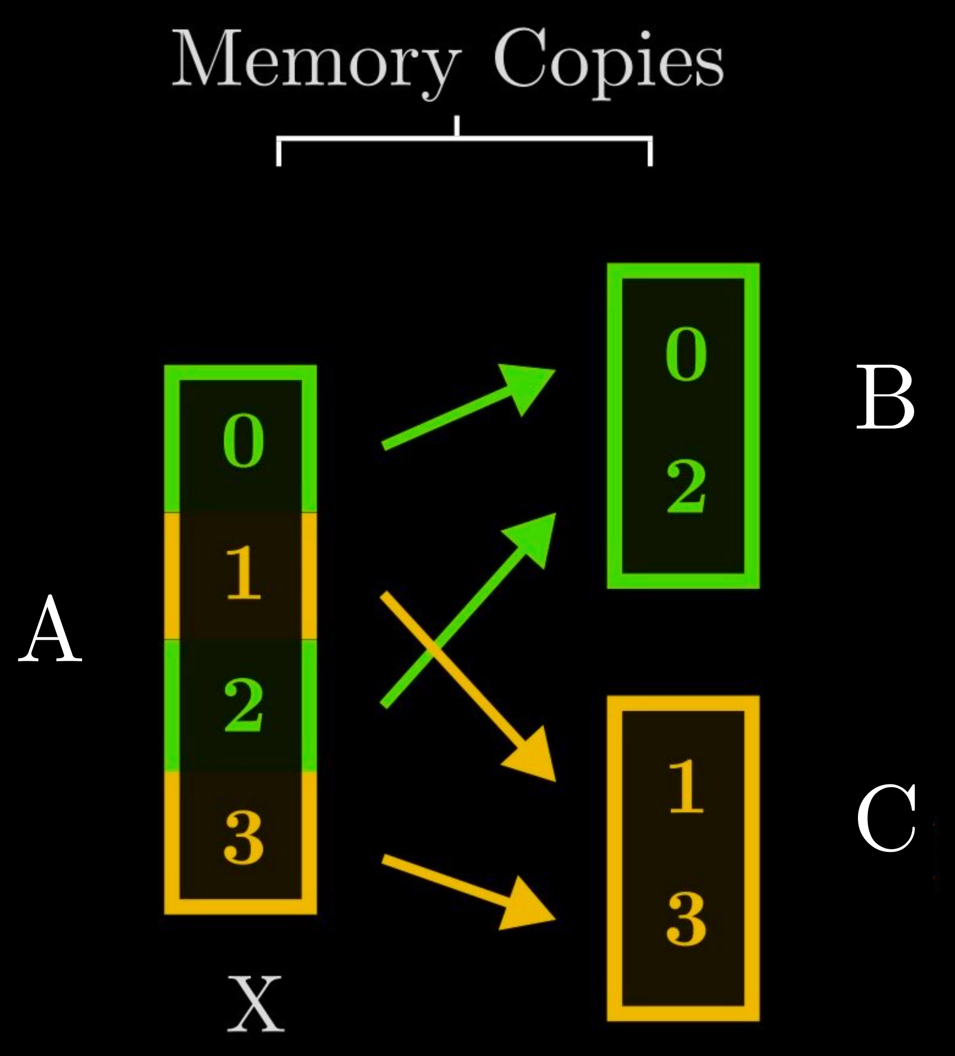
Results

Pruning is hard.

Sacrifice latency or sacrifice accuracy. Pick your poison.

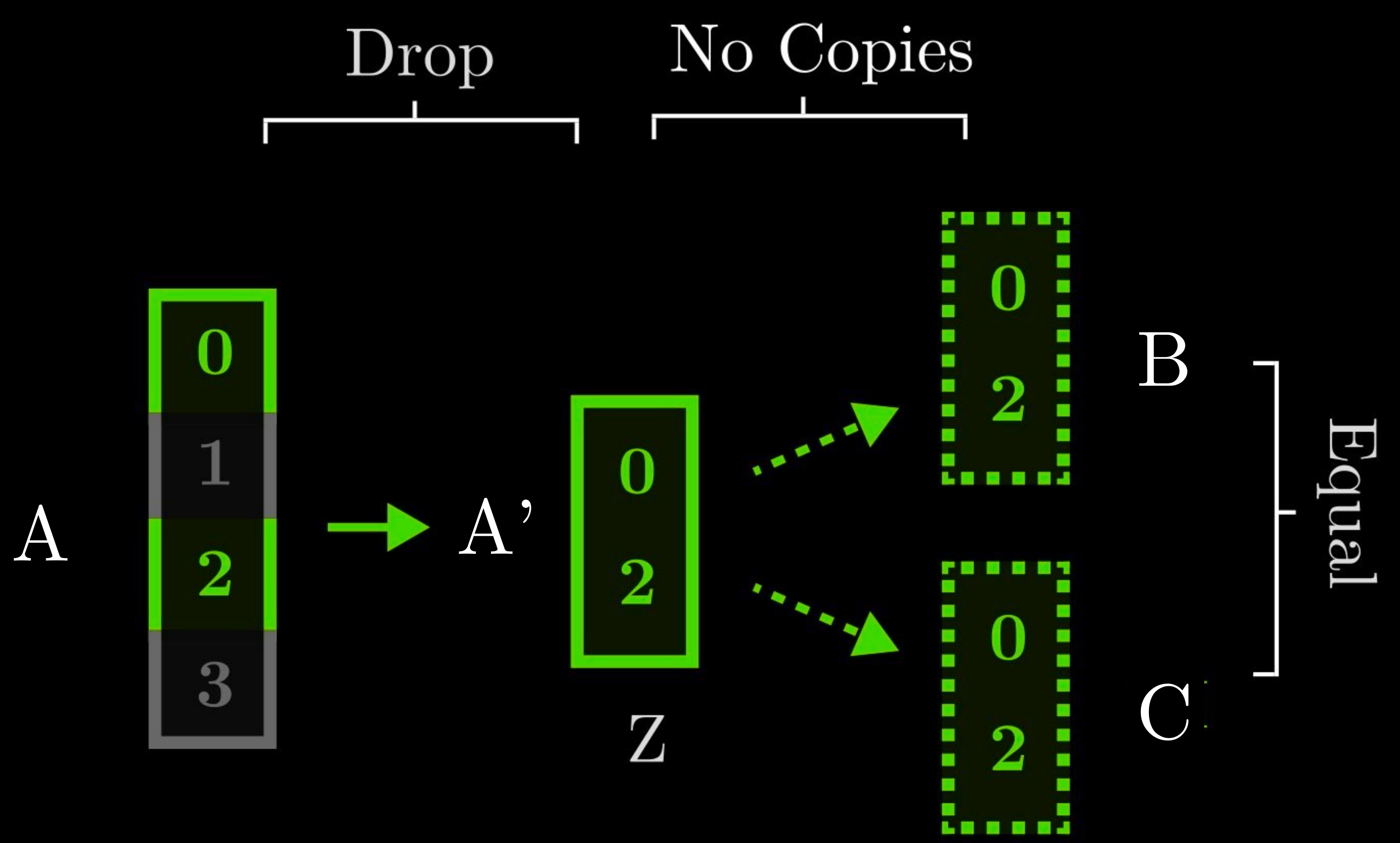
Inefficient

Prune any channel but add latency.



Constrained

Prune the same channels but lose accuracy.



UPSCALE: Unconstrained Channel Pruning by Wan et al

Add latency, or lose accuracy.

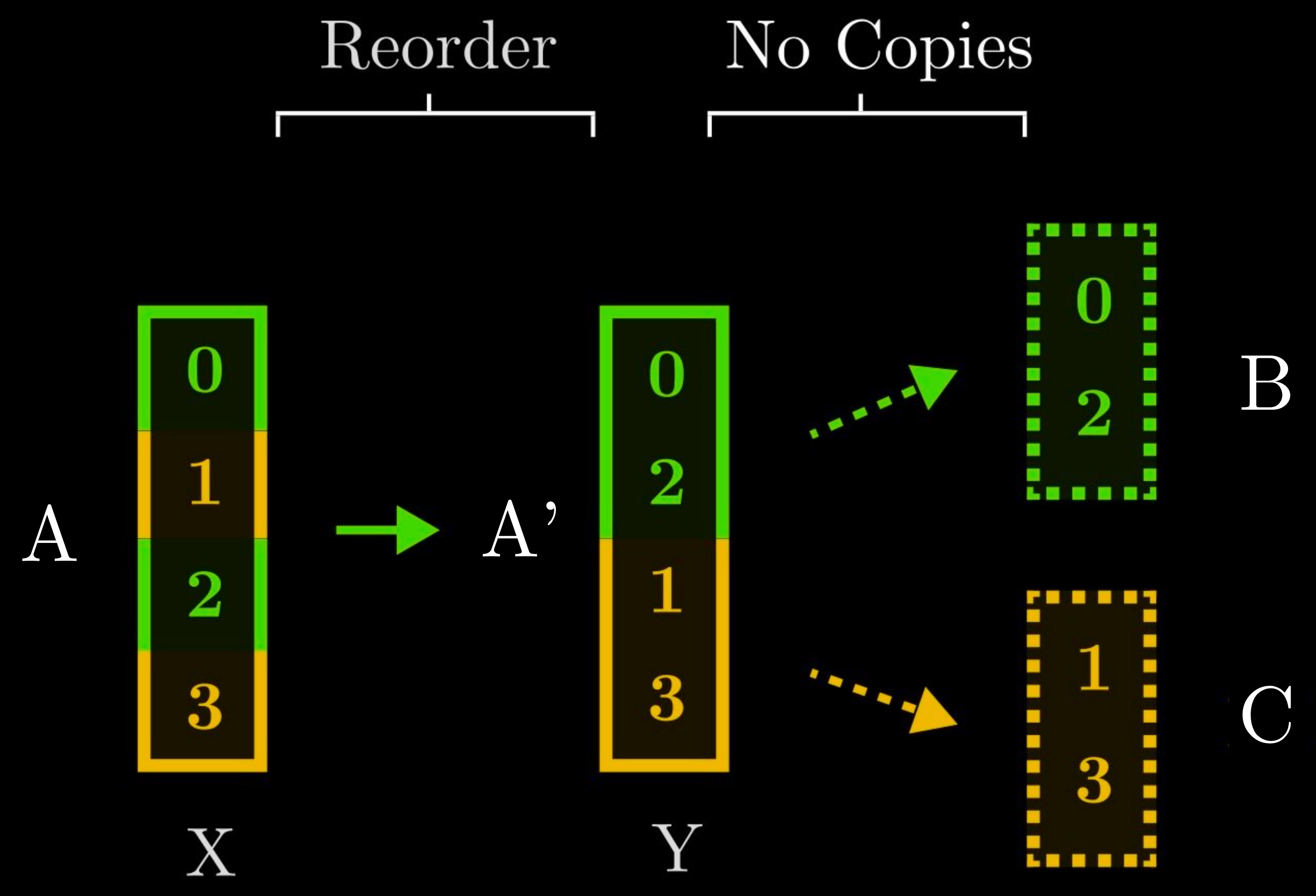
Remove constraints and add latency, or add constraints, and decrease accuracy.

Problem

Intuition

Method

Results



- **unconstrained**

  Each layer can prune any channel.

- **reorder**

  Move channels so that downstream inputs are contiguous.

- **contiguous**

  Contiguous slices are “free”, unlike memory copies.

UPSCALE: Unconstrained Channel Pruning by Wan et al

Don’t copy.

Instead, reorder and slice channels to reduce latency and retain accuracy.

Problem

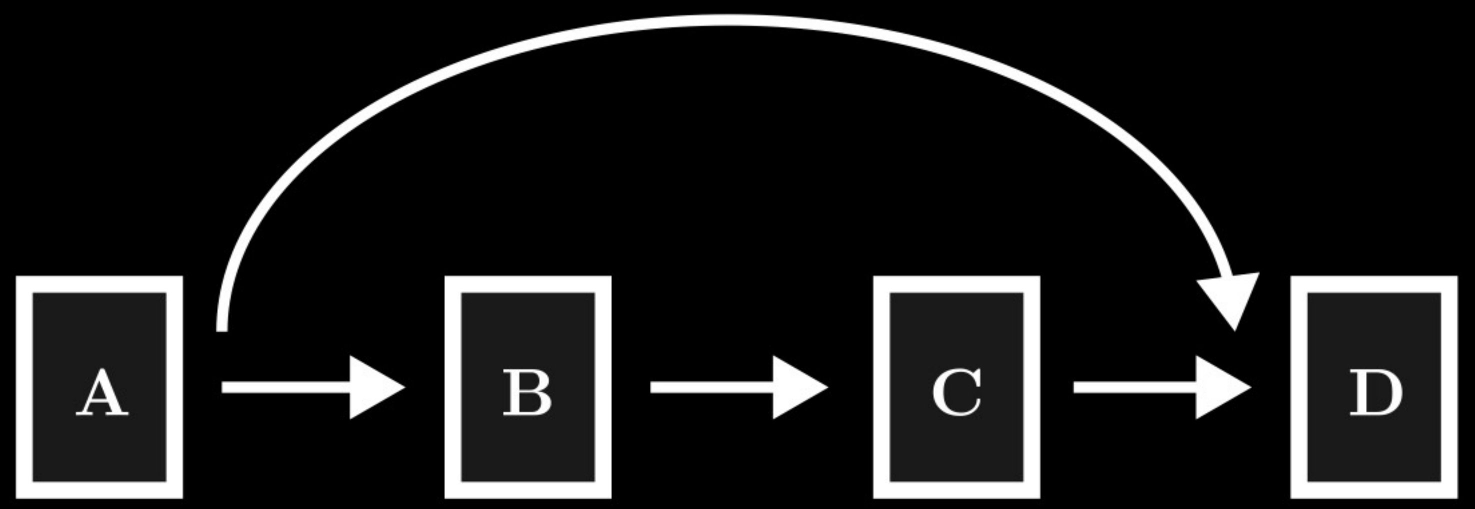
Intuition

Method

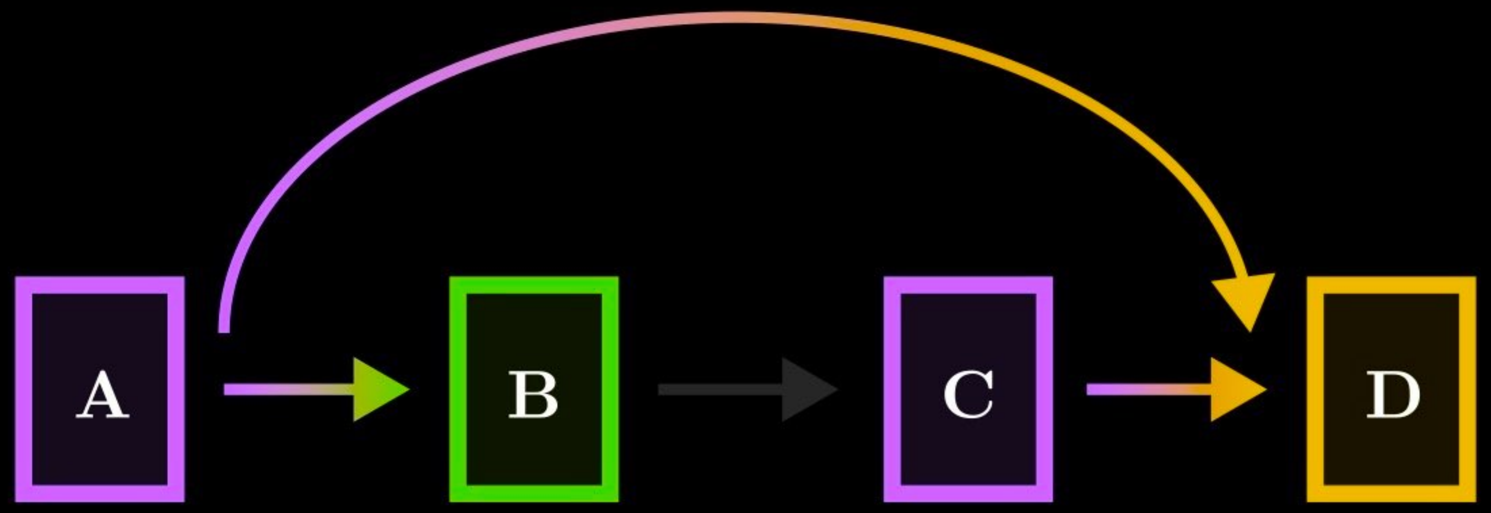
Results

1. Segment

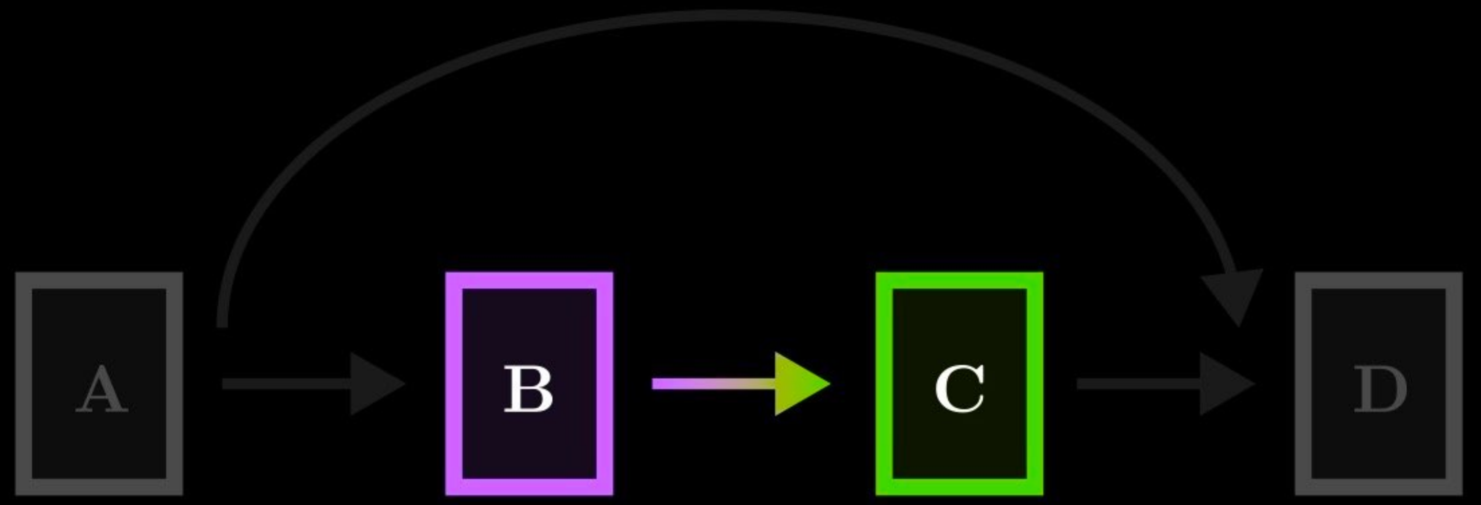
Find chunks of the network that can be pruned independently.



ResNet Block



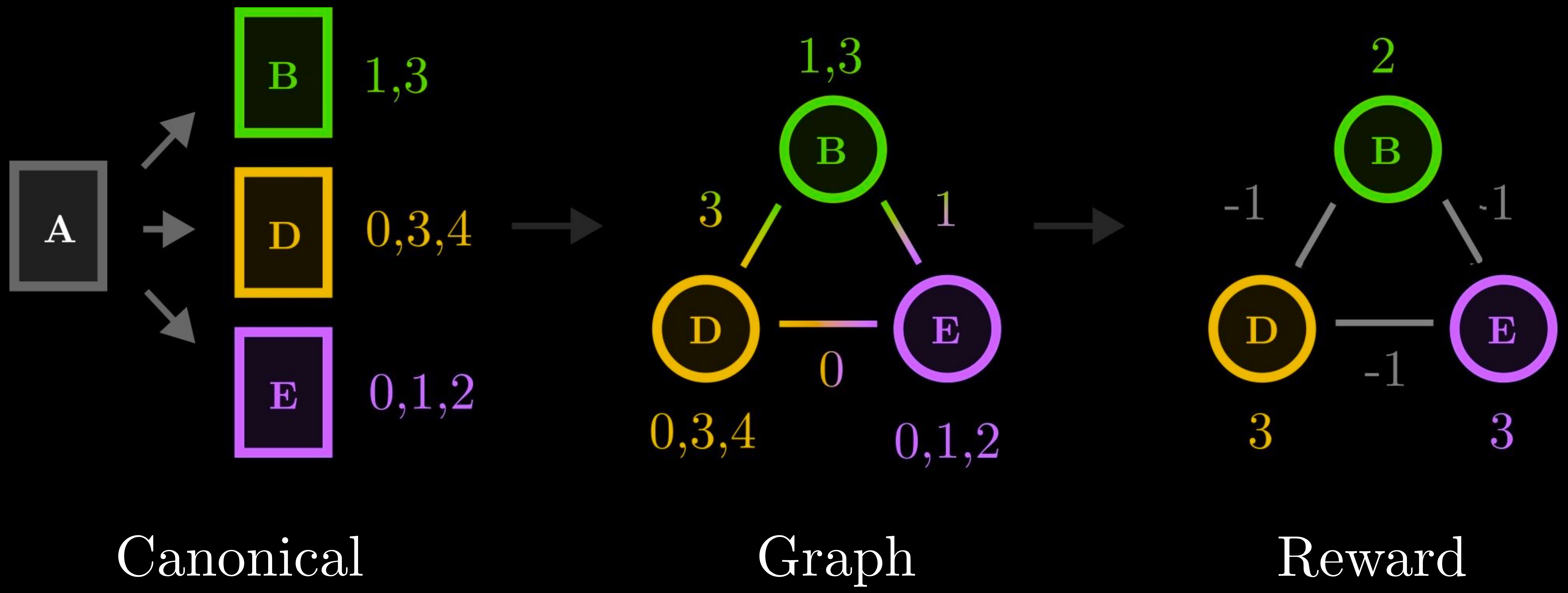
Segment 1



Segment 2

2. Define graph

Convert architecture into a graph that represents constraints.

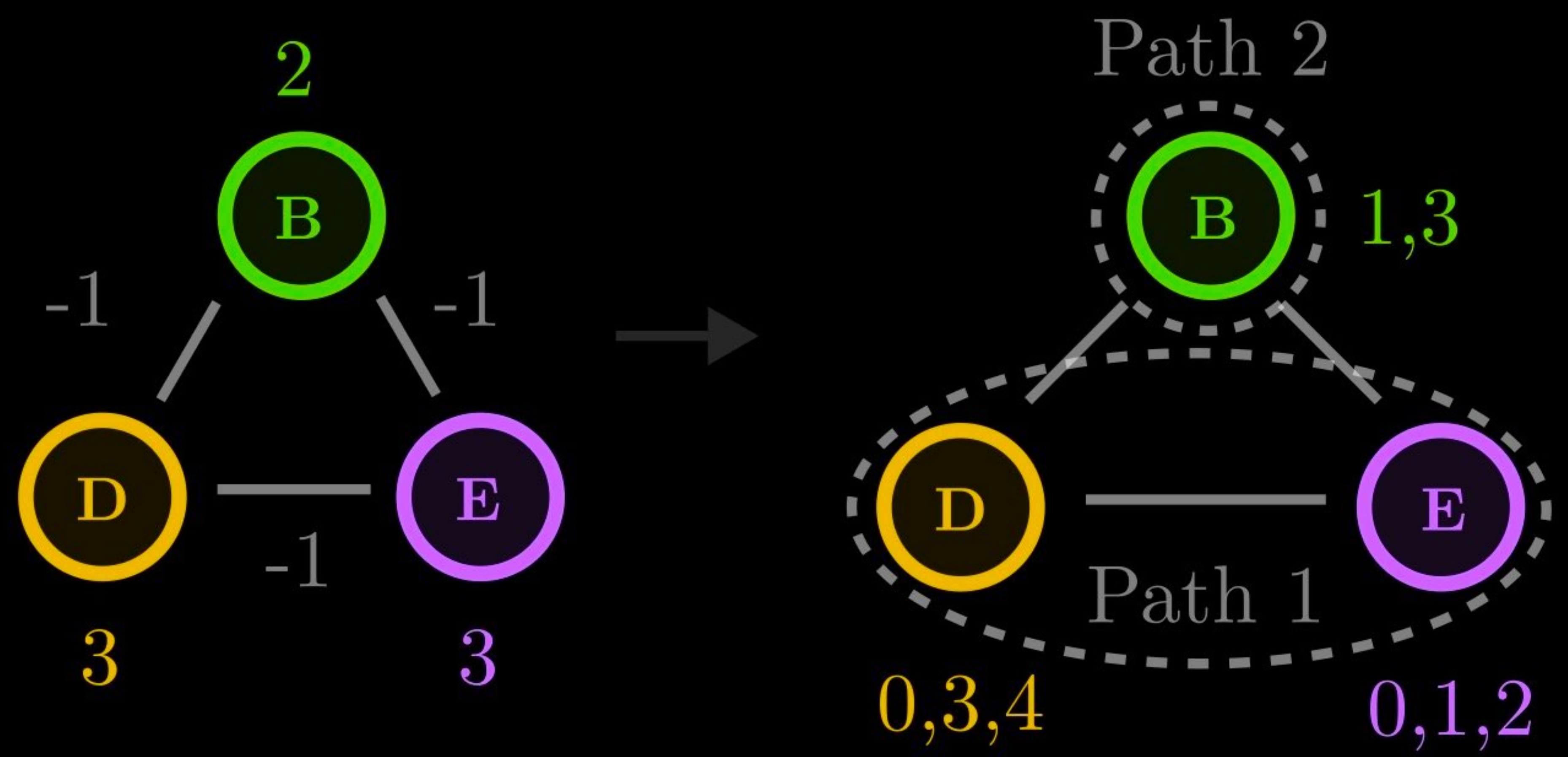


UPSCALE: Unconstrained Channel Pruning by Wan et al

3. Find path.

Find a path that maximizes reward.

Find maximum-reward acyclic path.

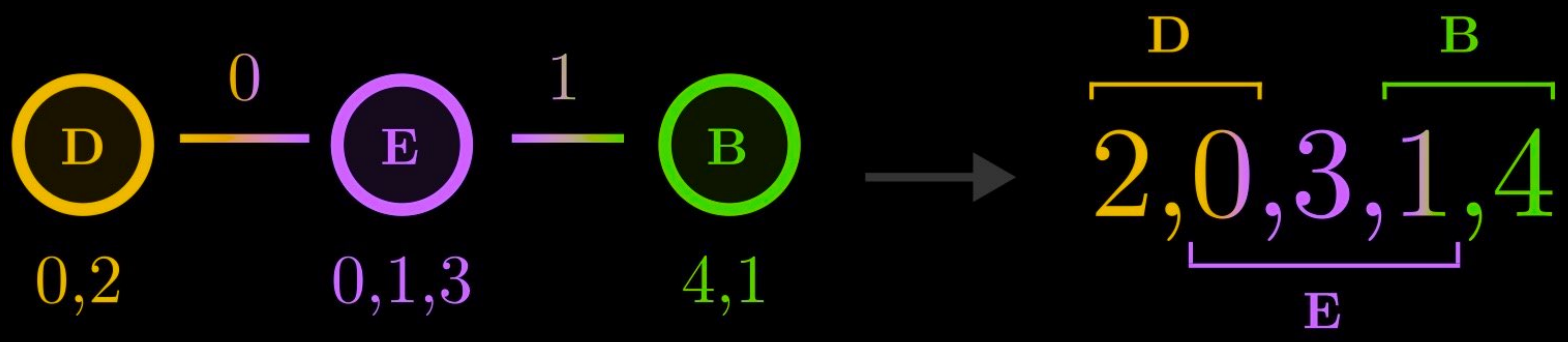


UPSCALE: Unconstrained Channel Pruning by Wan et al

4. Determine order.

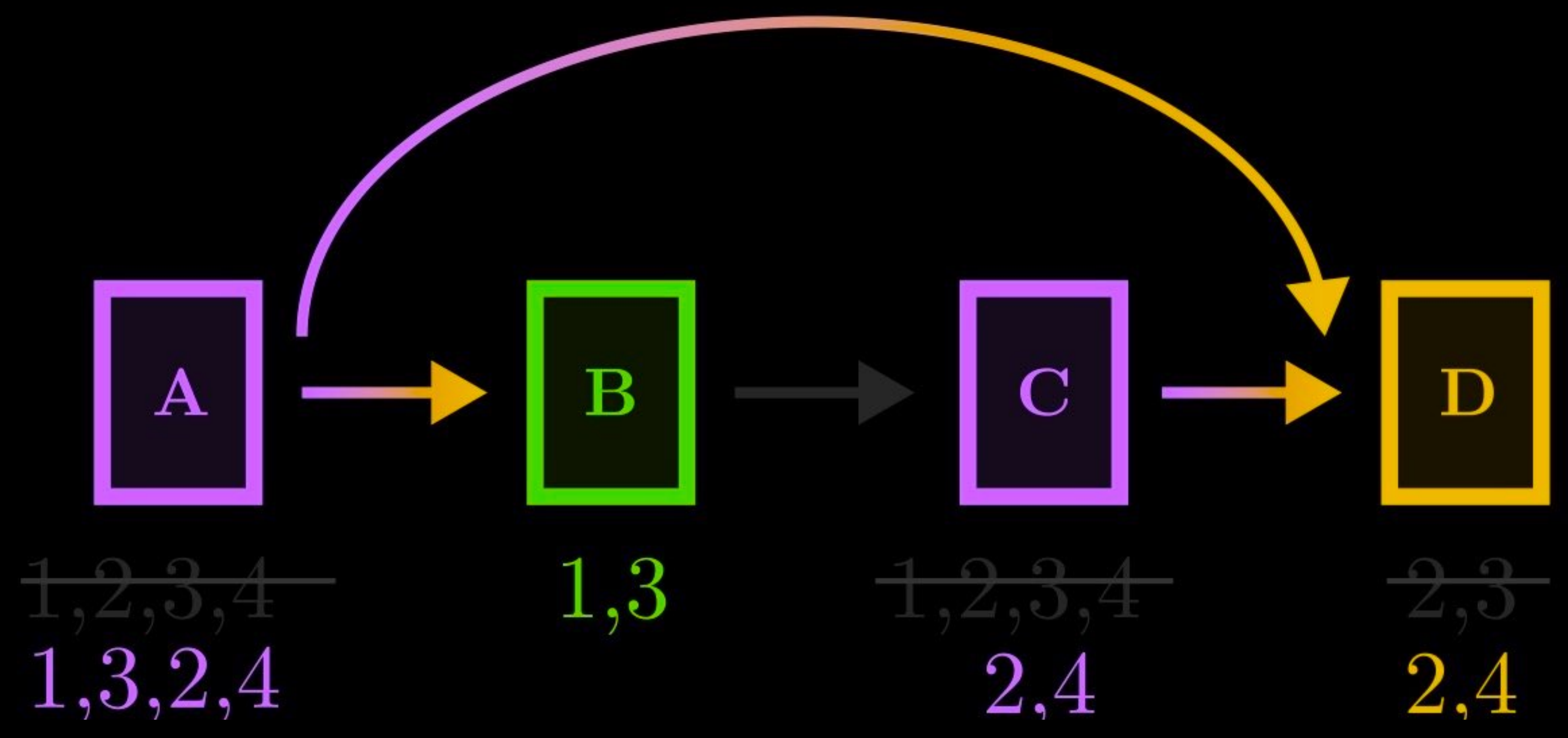
Convert into channel order.

Convert to ordering.

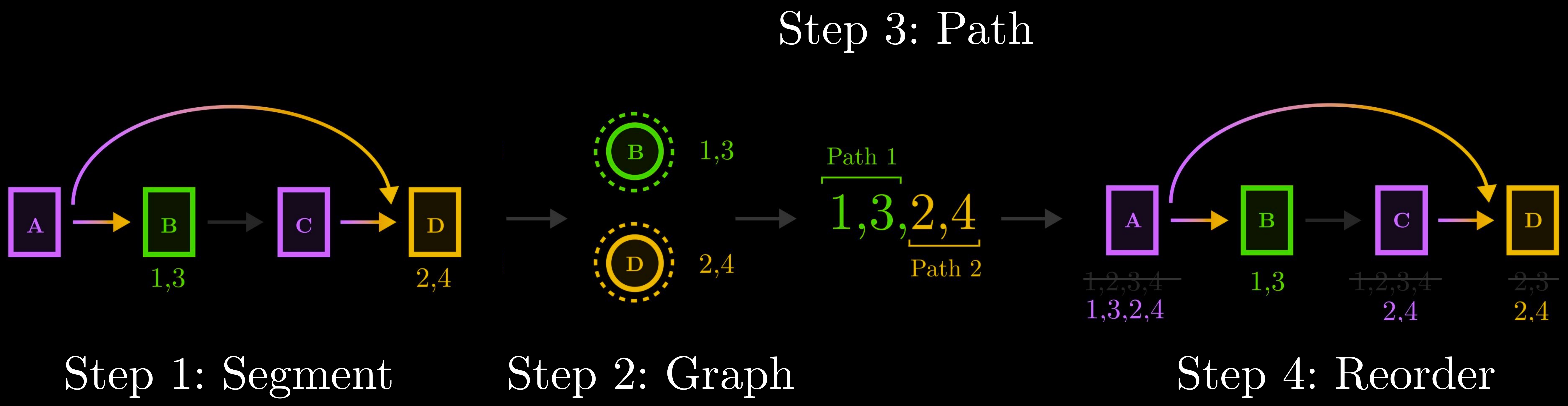


UPSCALE: Unconstrained Channel Pruning by Wan et al

Reorder channel weights.



Pipeline



Step 1: Segment

Step 2: Graph

Step 3: Path

Step 4: Reorder

UPSCALE: Unconstrained Channel Pruning by Wan et al

Problem

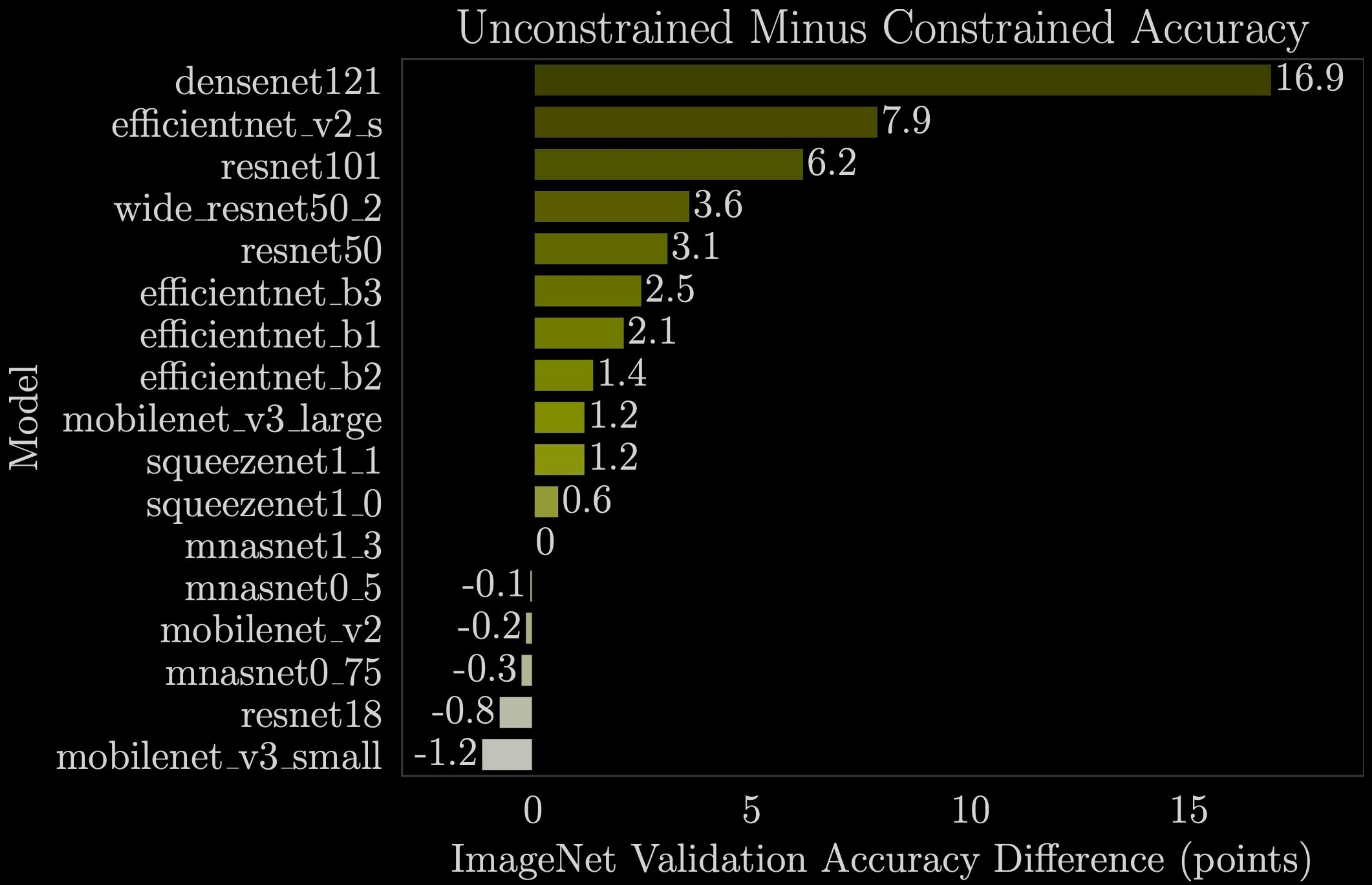
Intuition

Method

Results

Removing constraints raises accuracy.

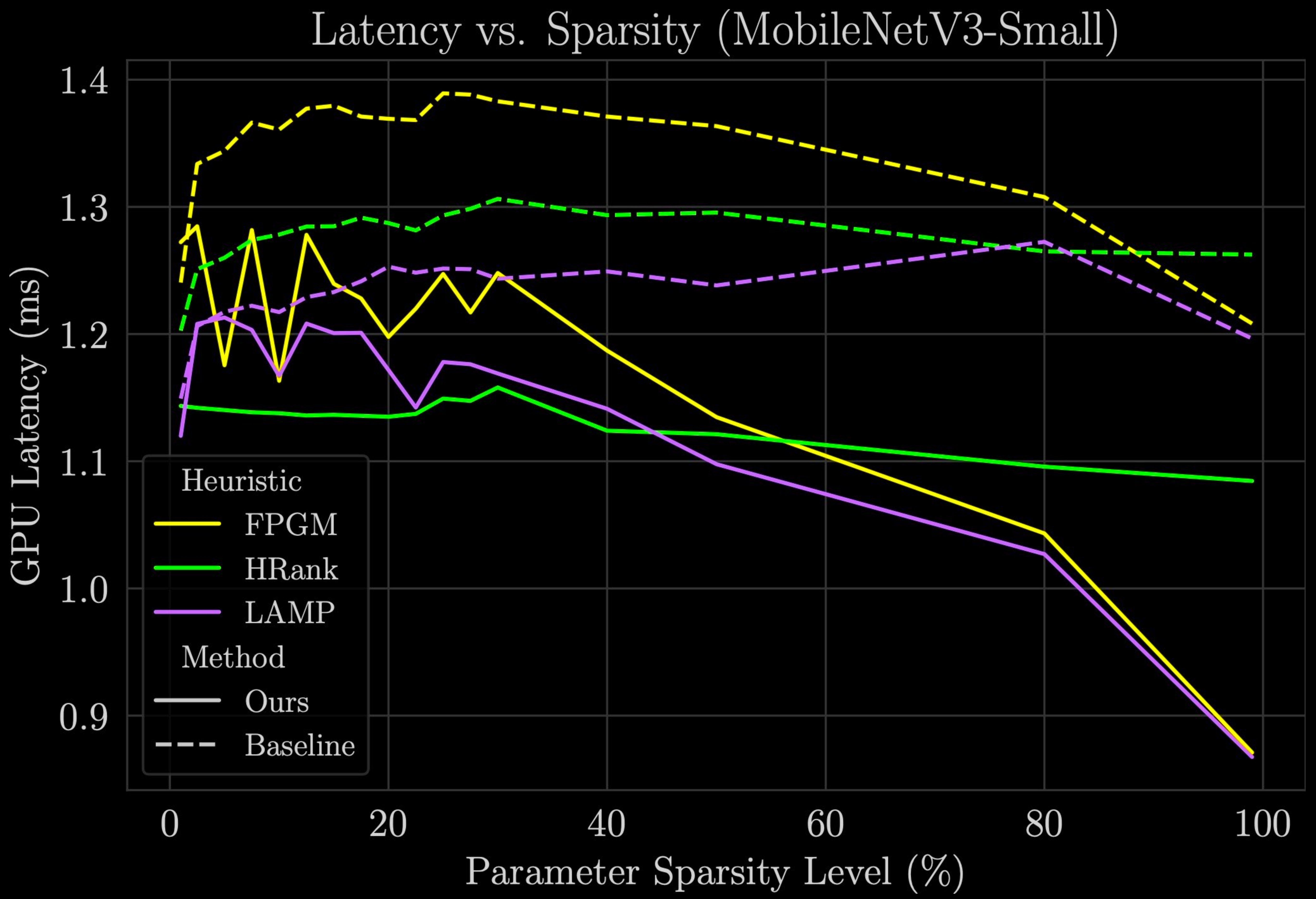
Across architectures, heuristics, and sparsity levels.



UPSCALE: Unconstrained Channel Pruning by Wan et al

Reordering lowers latency.

Across architectures, heuristics, and sparsity levels.



UPSCALE: Unconstrained Channel Pruning by Wan et al

```
pip install apple-upscale
```

UPSCALE:

Unconstrained Channel Pruning

alvinwan@apple.com | @lvinwan | github.com/apple/ml-upscale

Alvin Wan, Hanxiang Hao, Kaushik Patnaik, Yueyang Xu, Omer Hadad, David Guera, Zhile Ren, Qi Shan

International Conference on Machine Learning 2023 | Apple Inc.