UTAH ARCH

A Case for Dynamic Activation Quantization in CNNs

Karl Taht, Surya Narayanan, Rajeev Balasubramonian

University of Utah

Overview

- **Background**
- **Proposal**
- **Search Space**
- **Architecture**
- **Results**
- **Future Work**

Improving CNN Efficiency

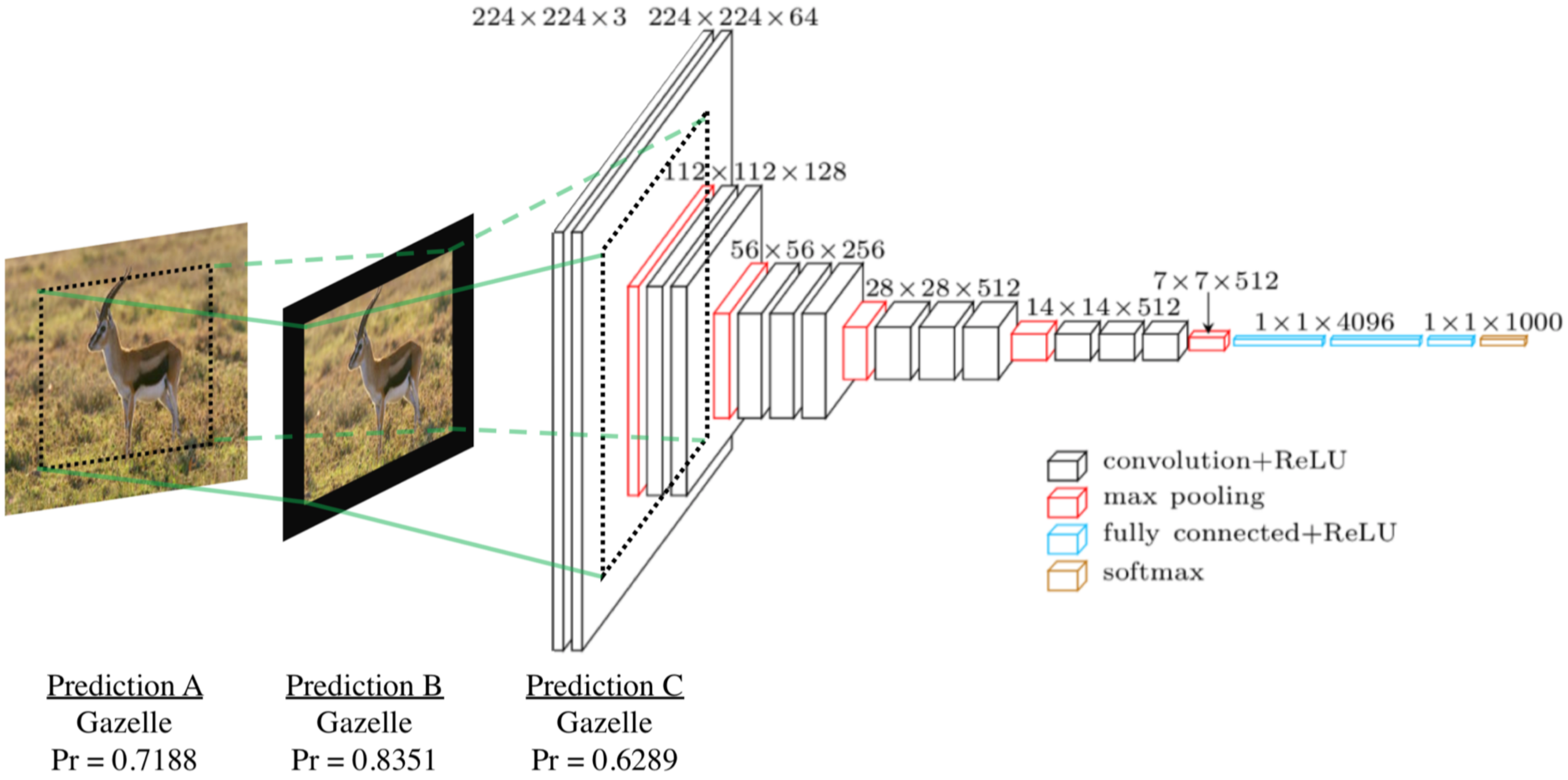
- *Stripes: Bit-Serial Deep Neural Network Computing*
  - ***Per-layer bit precisions*** net significant savings with <1% accuracy loss
  - Brute force approach to find best quantization – retraining at each step!
  - Good end result, but expensive!
- *Weight-Entropy-Based Quantization for Deep Neural Networks*
  - Quantize both weights and activations
  - ***Guided search*** to find optimal quantization (entropy and clustering)
  - Still requires retraining, still a passive approach

Can we exploit adaptive reduced precision during inference?

Proposal:
Adaptive Quantization Approach (AQuA)

- Most images contain regions of ***irrelevant information*** for the classification task
- Can avoid such computations all together?
- ***Quantize*** completely regions to ***0 bits***
  - More simply – *Crop them!*

Proposal: Activation Cropping

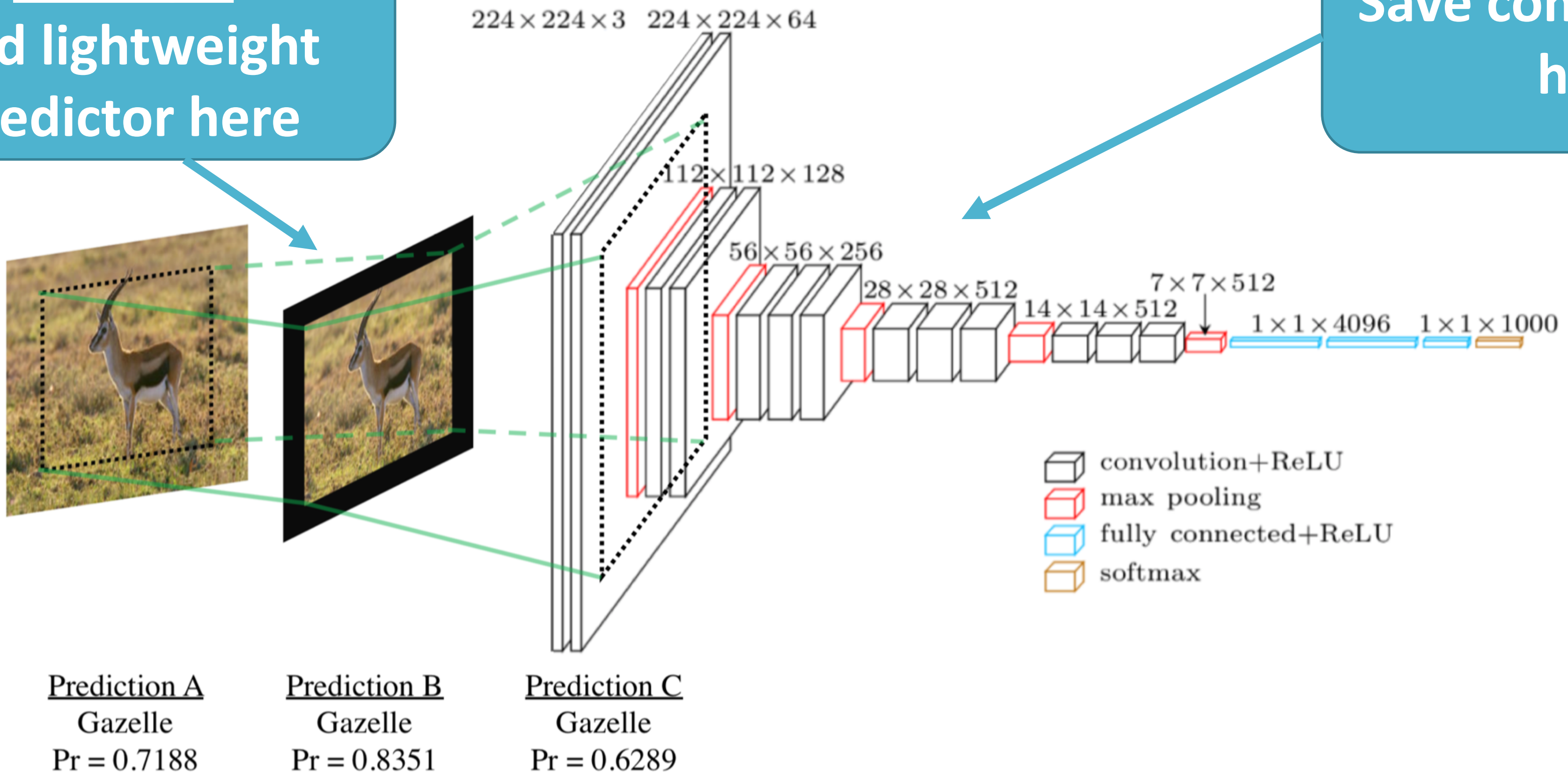


Proposal: Activation Cropping

Concept:

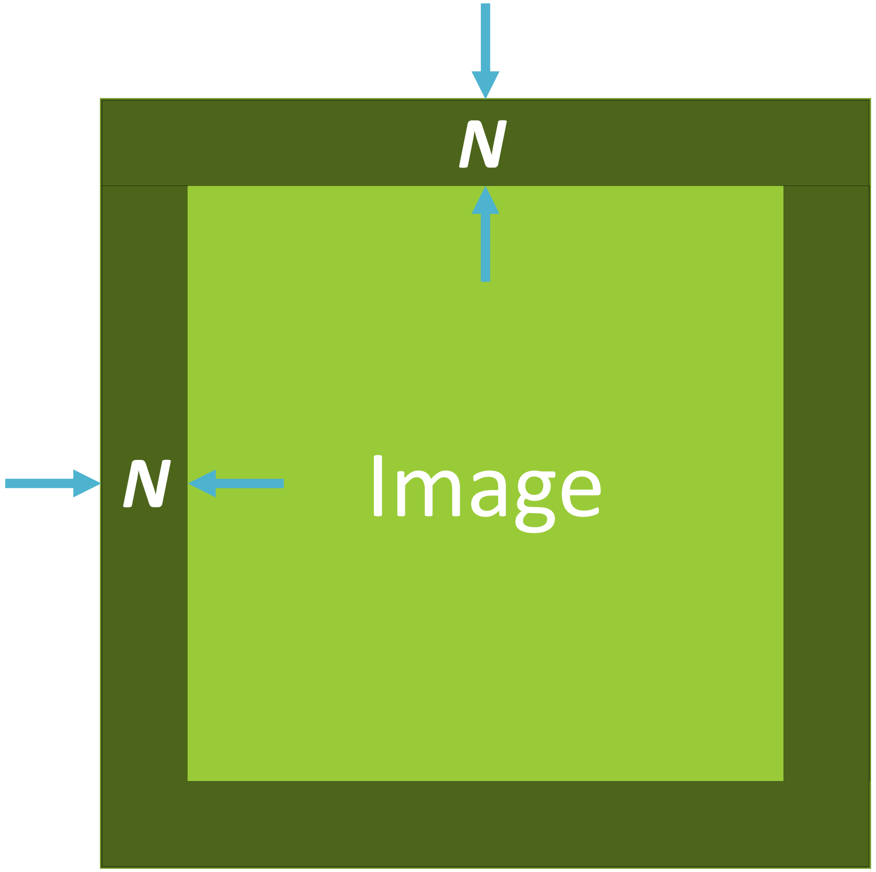
Add lightweight predictor here

Save computations here

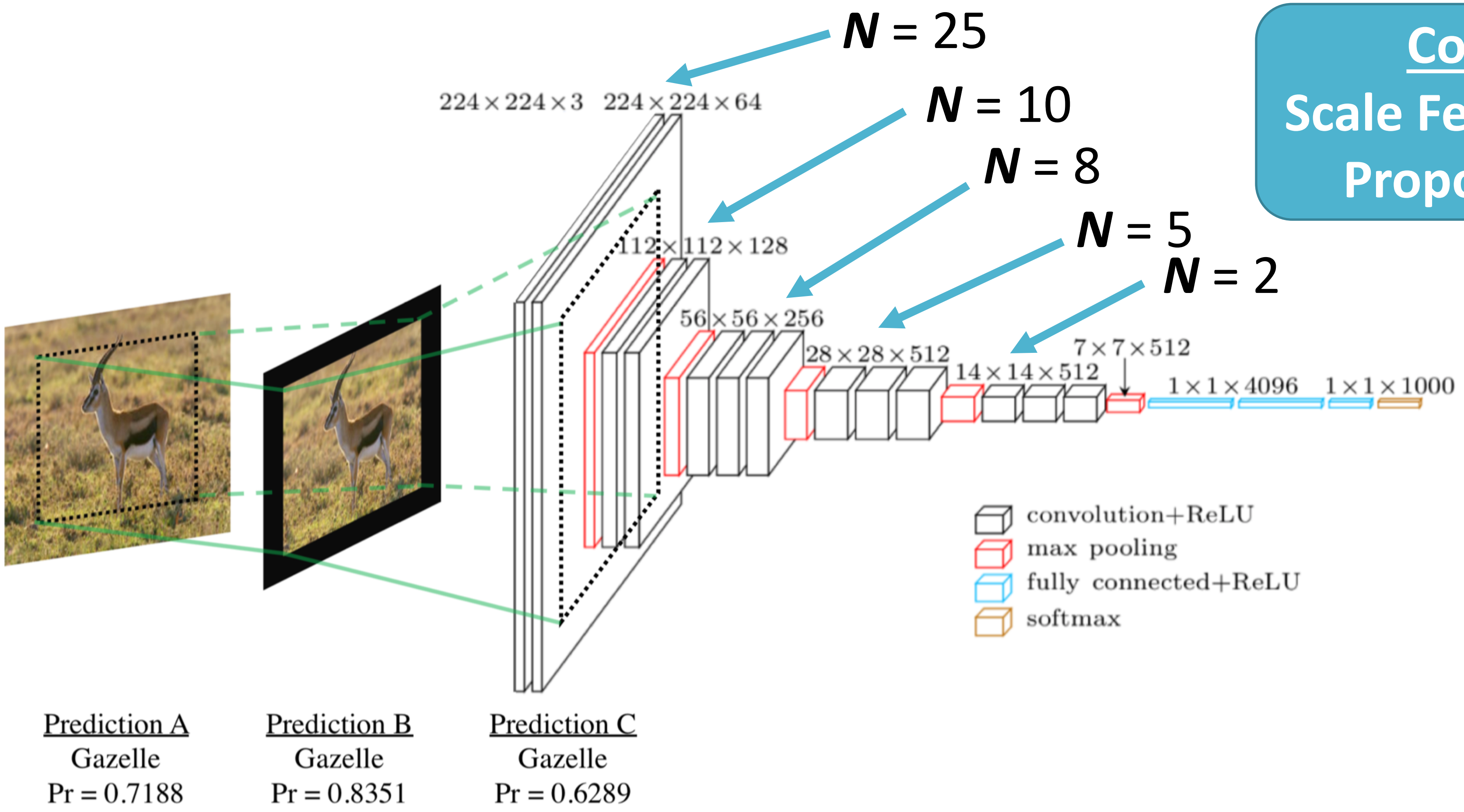


Search Space – How to Crop

- Exploit domain knowledge
  - Information is typically centered within the image (>55% in our tests)
- Utilize a regular pattern
  - Less control logic required
  - Maps easier to different hardware
- Added bonus:
  - While objects are centered, majority of area (and thus computation) is on the outside!



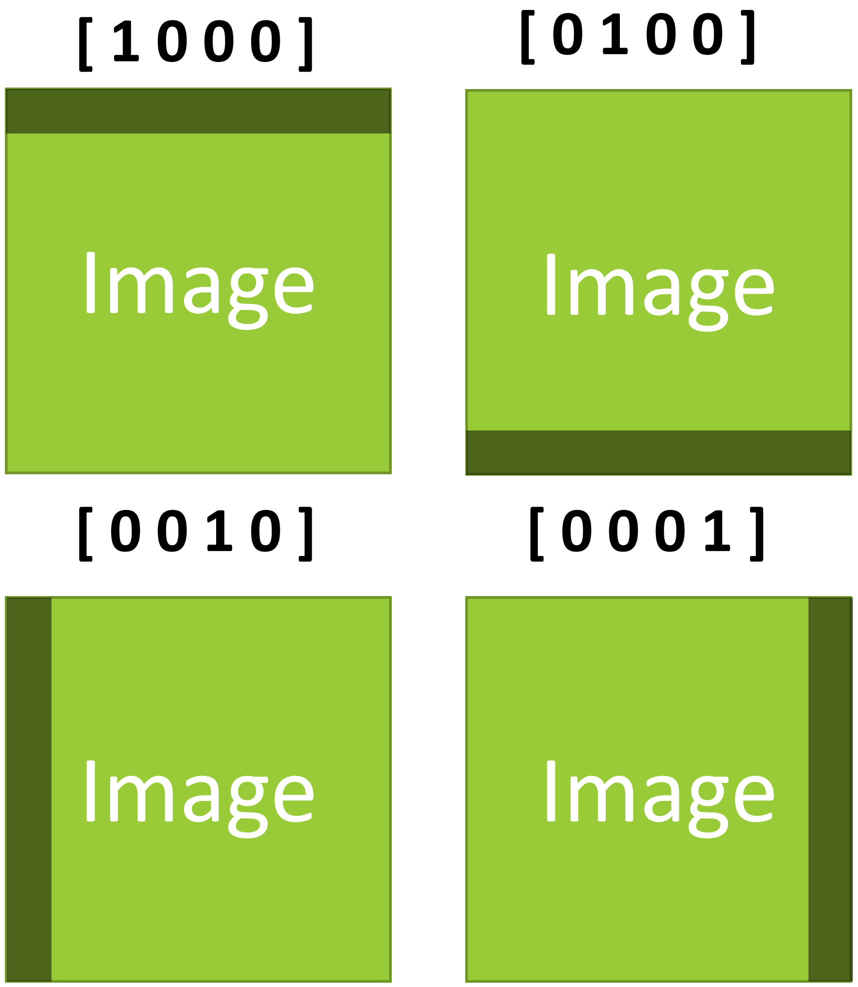
Proposal: Activation Cropping



Concept:

Scale Feature Maps Proportionally

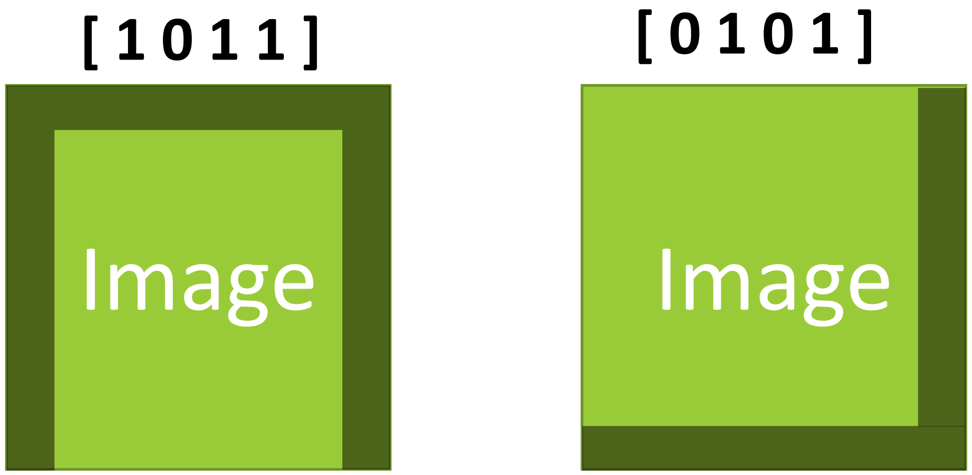
Search Space – Crop Directions



- We consider ***16 possible crops*** as permutations of top, bottom, left, and right crops encoded as a vector:

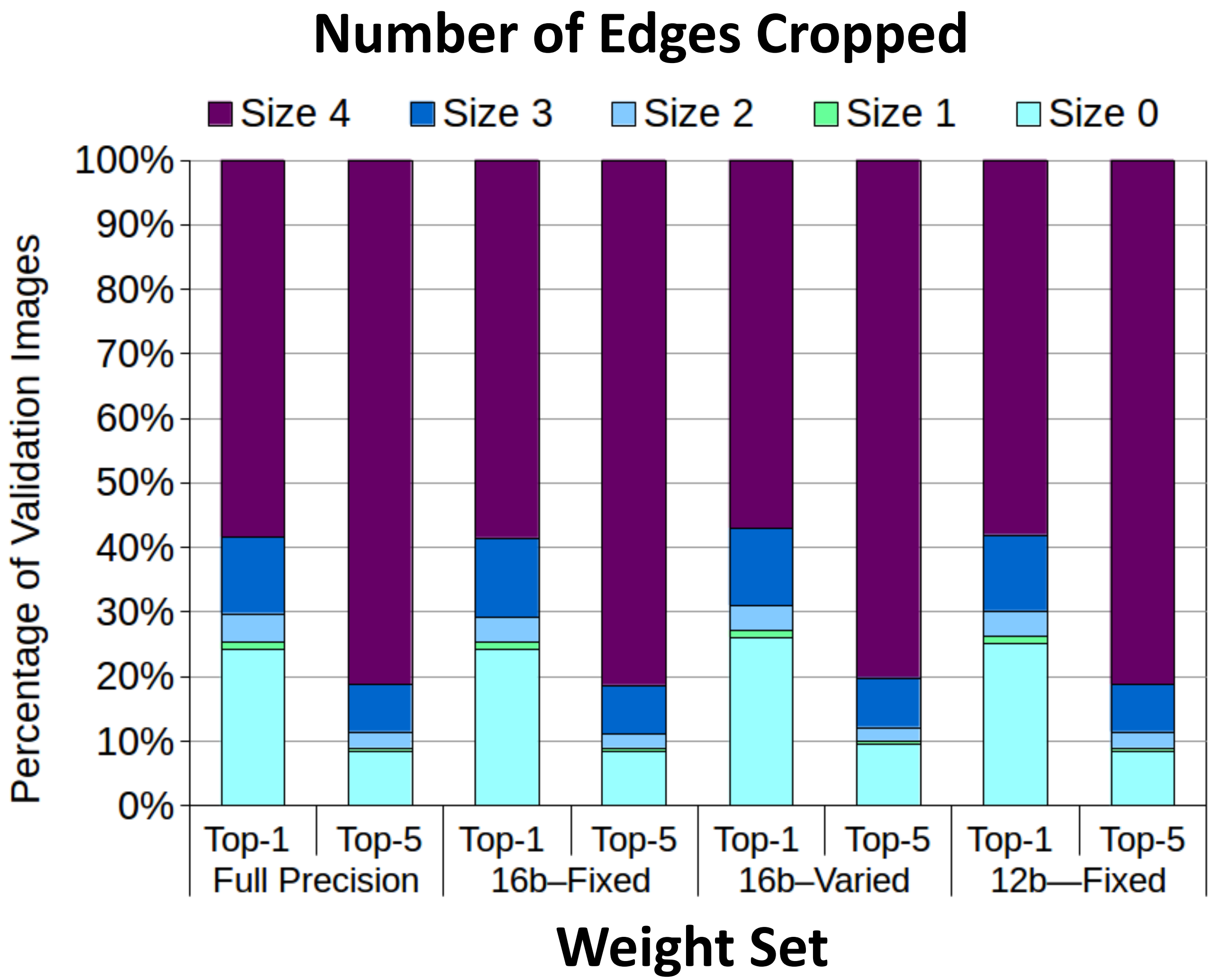
  [TOP , BOTTOM , LEFT , RIGHT]

- Unlike traditional pruning, AQuA can exploit ***image-based information*** to enhance pruning options.



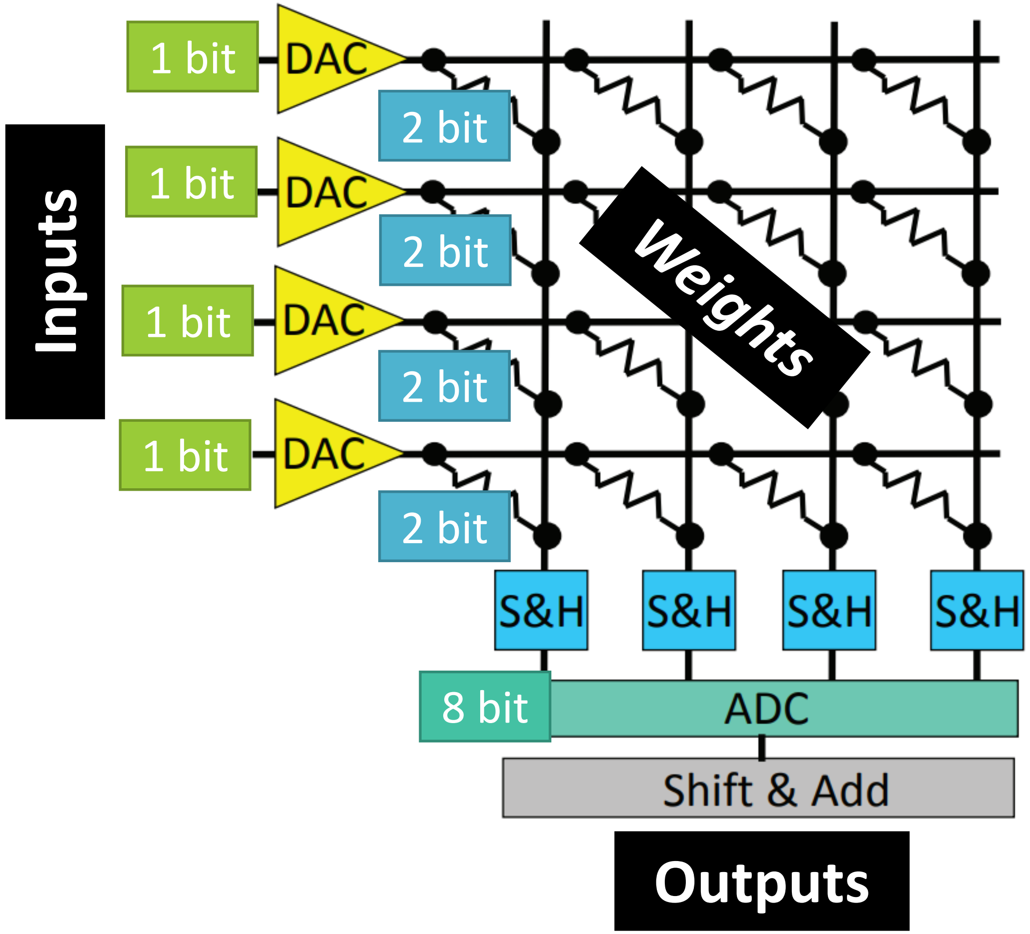
Quantifying Potentials

- For maintaining original Top-1 accuracy, ***75% images can tolerate some type of crop!***
- Greater savings with top-5 predictions
- Technique ***invariant*** to weight quantization

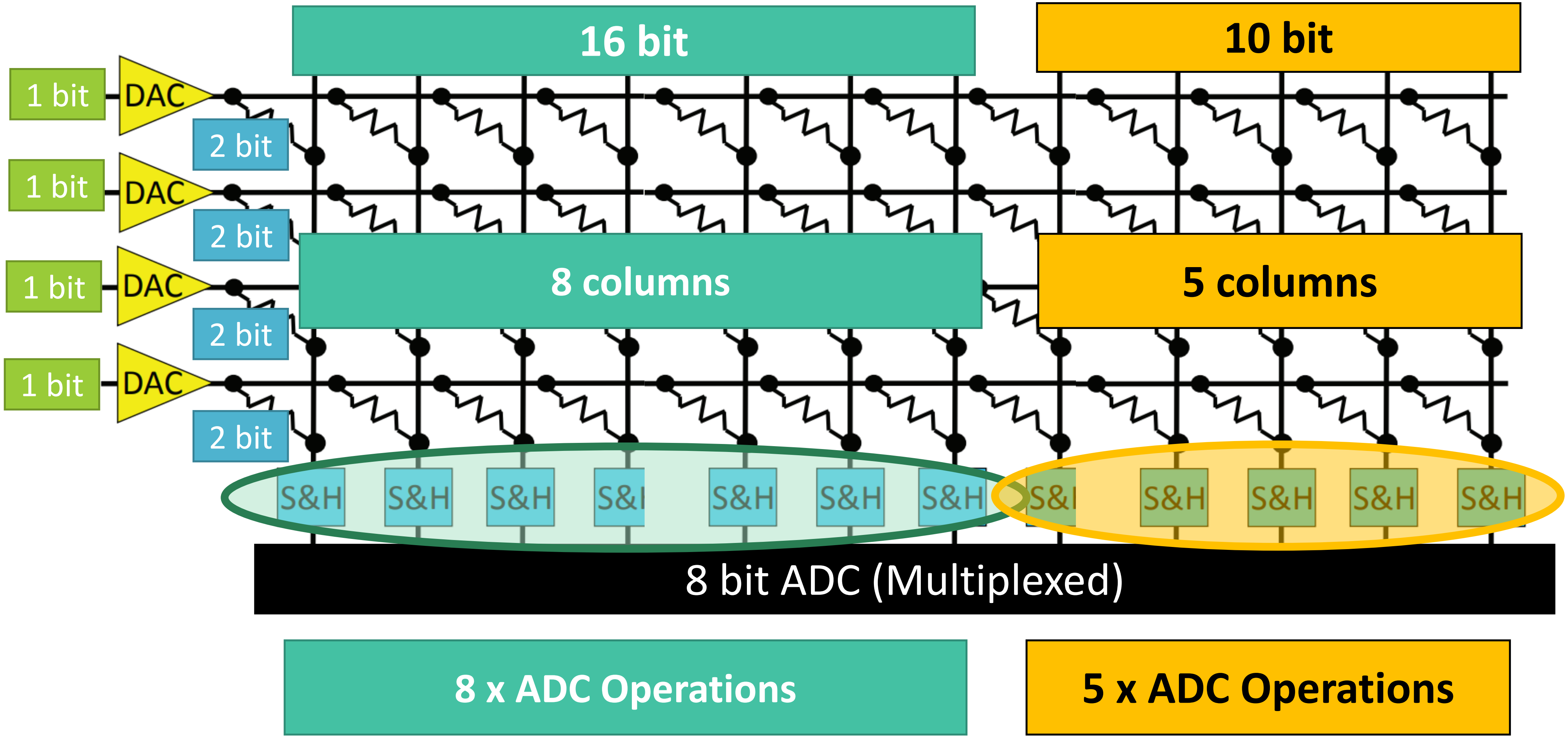


Exploiting Energy Savings with ISAAC

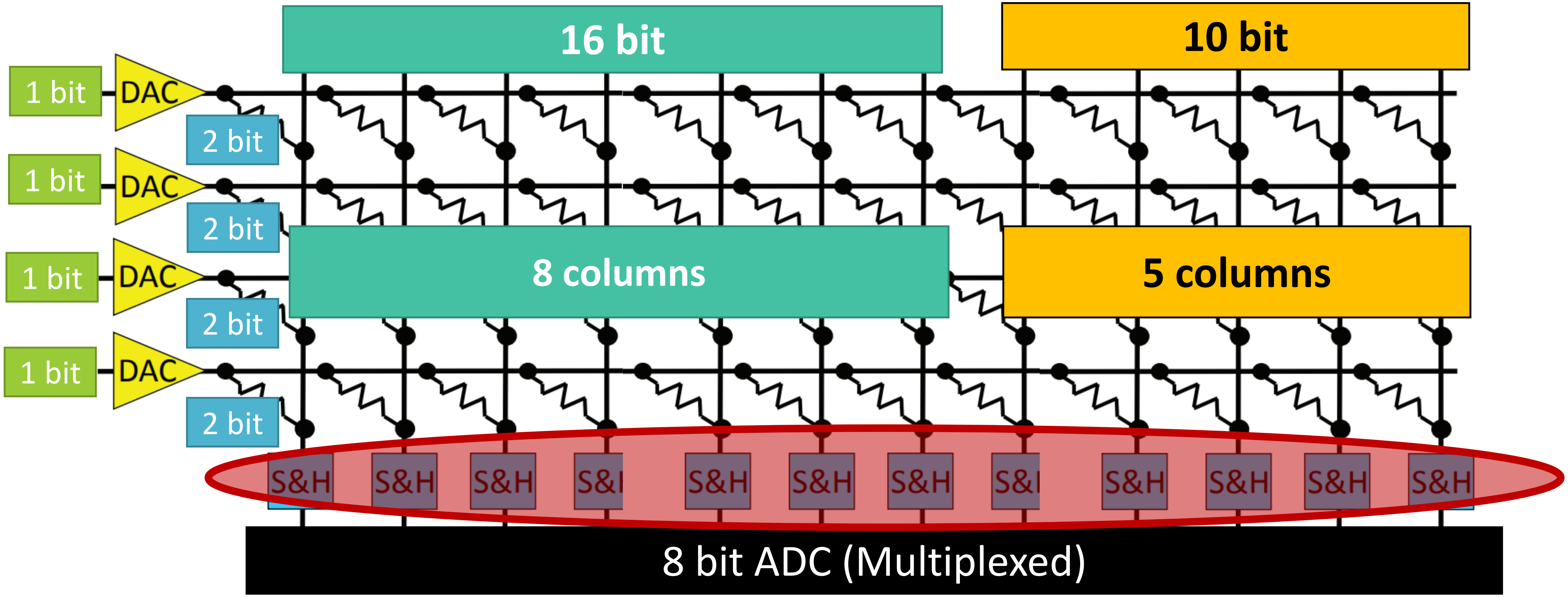
- Activation cropping technique can be applied to any architecture
- We use the ISAAC accelerator due to its flexibility
- Future work includes leveraging additional variable precision techniques



Weight Precision Savings

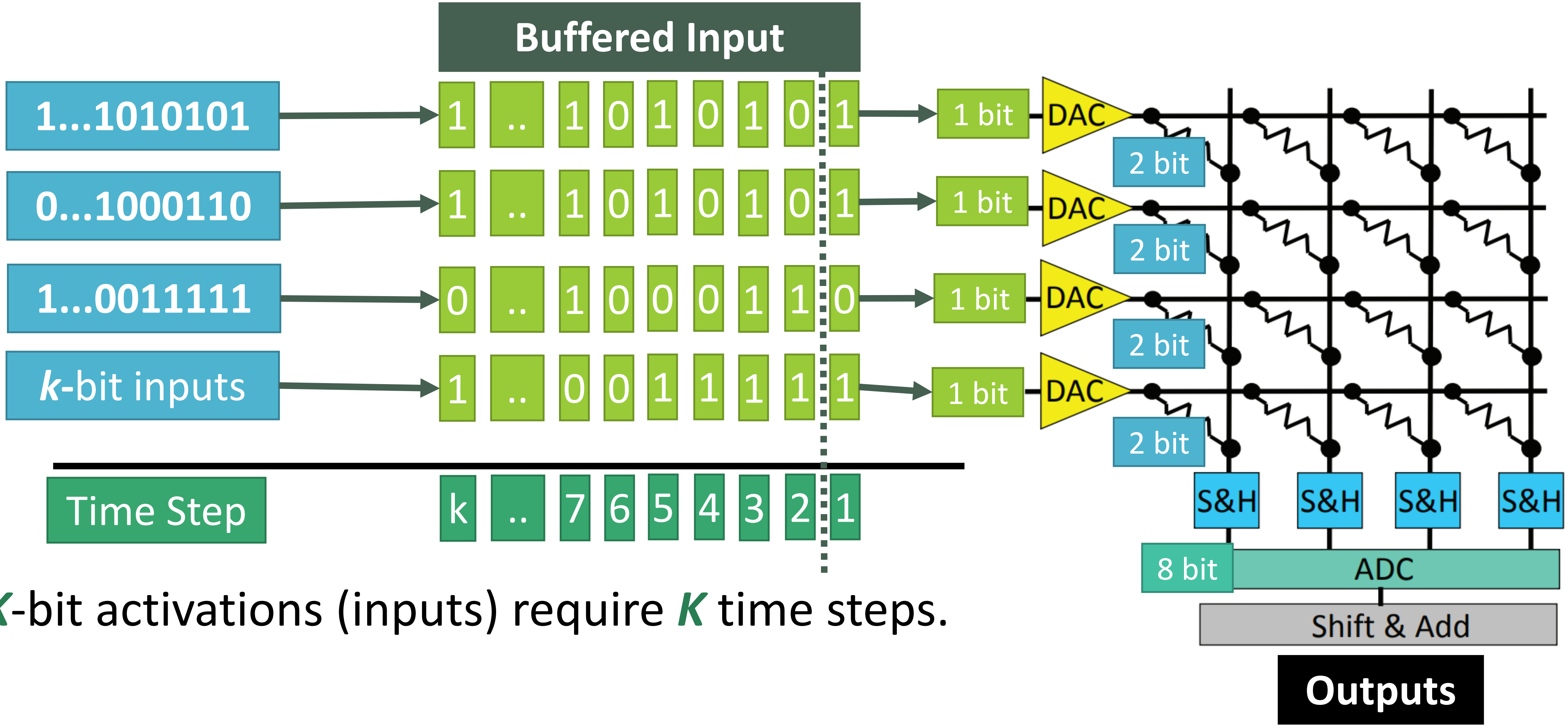


“FlexPoint” Support



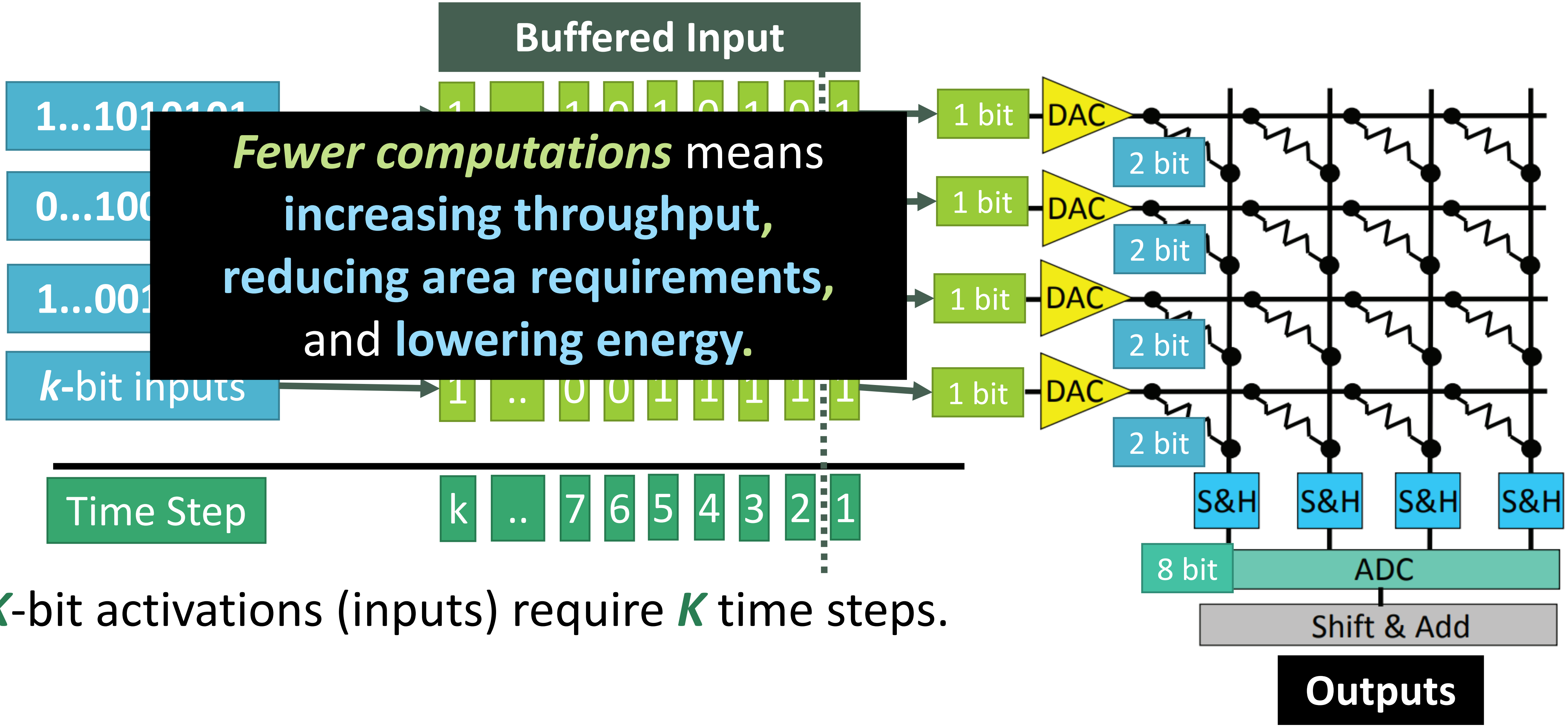
Can vary shift amount to compute fixed point computations with different exponents

Activation Quantization Savings



***K*-bit activations (inputs) require *K* time steps.**

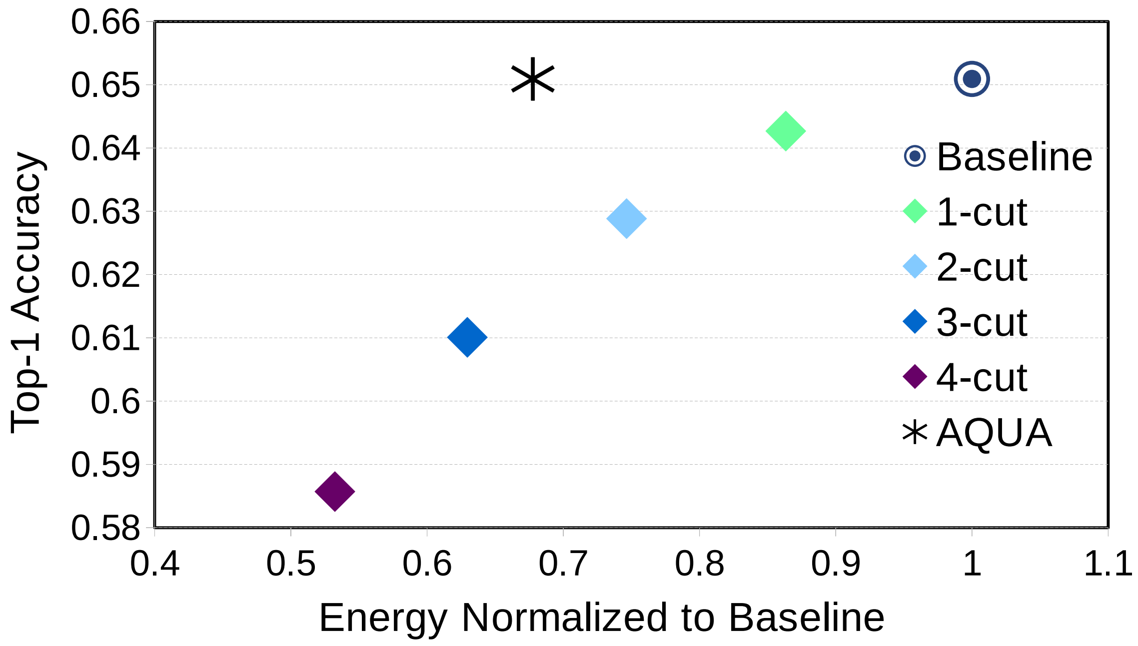
Activation Quantization Savings

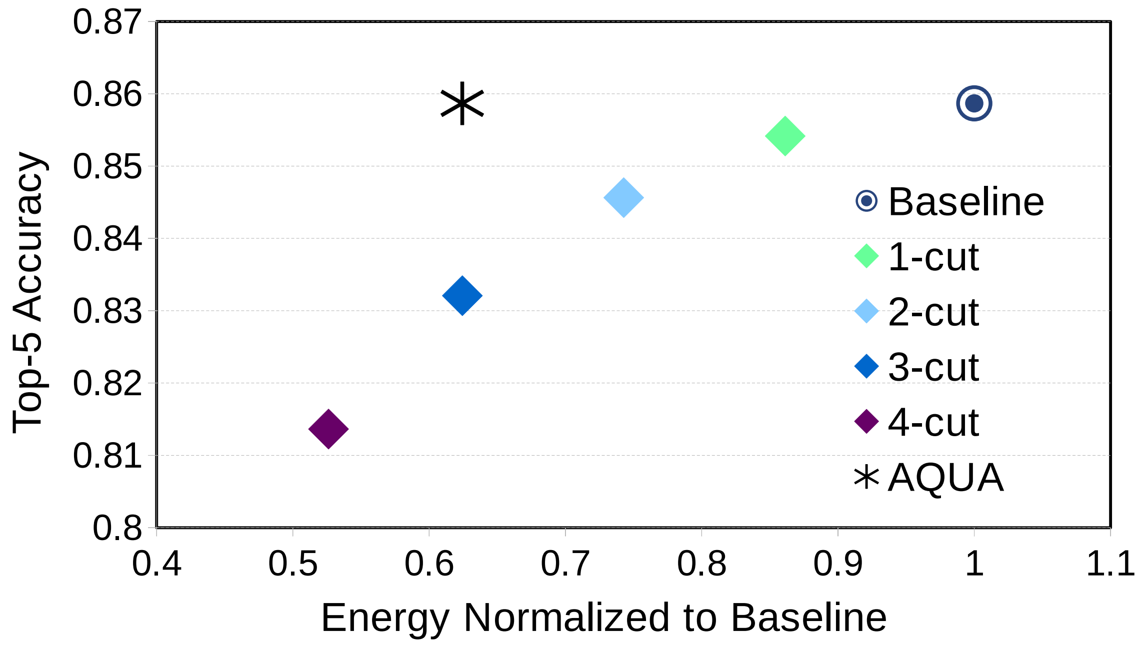


Fewer computations means **increasing throughput, reducing area requirements,** and **lowering energy.**

K-bit activations (inputs) require ***K*** time steps.

Naive Approach – Crop Everything

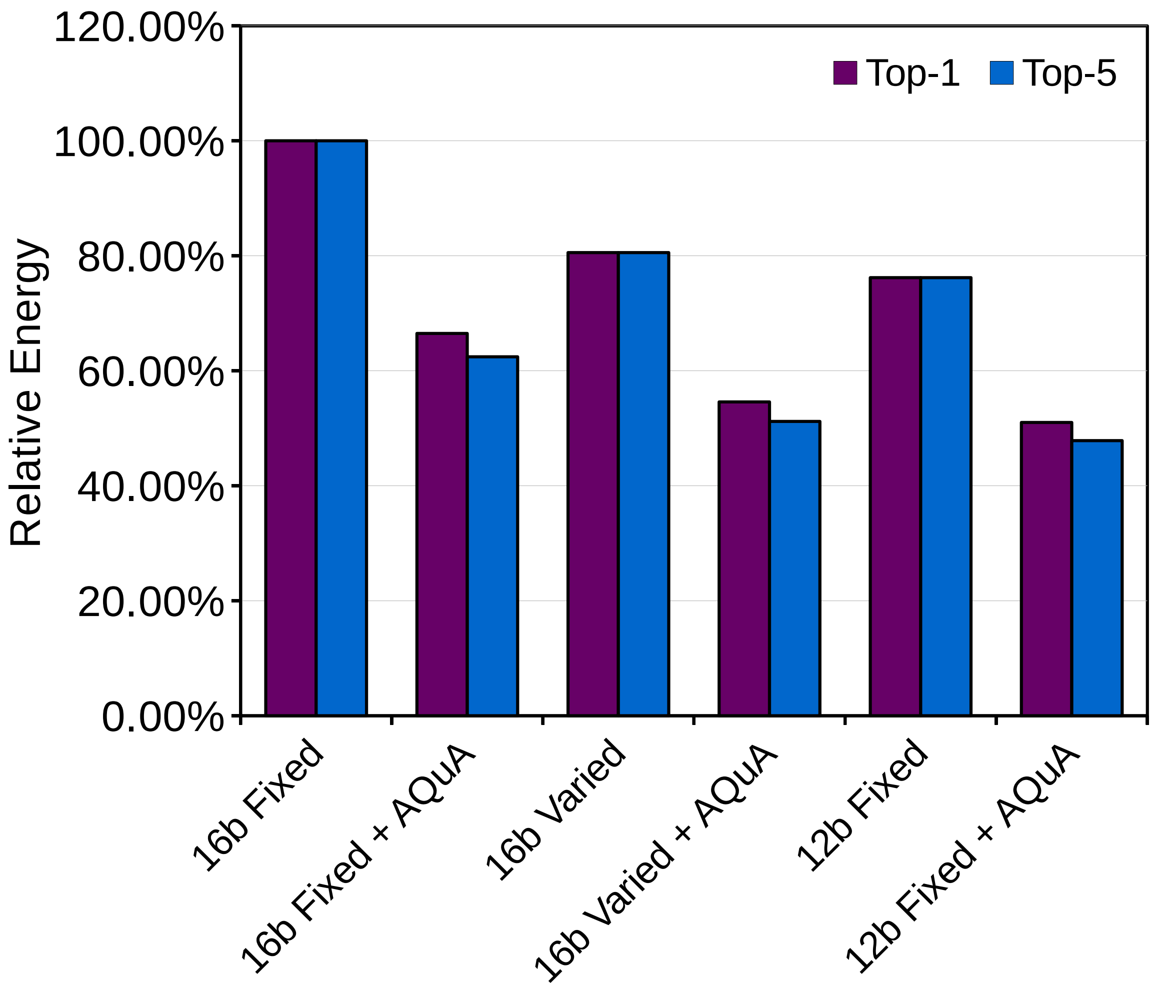




- Substantial energy savings at a cost to accuracy
- Theoretically, can save over 33% energy and maintain original accuracy!

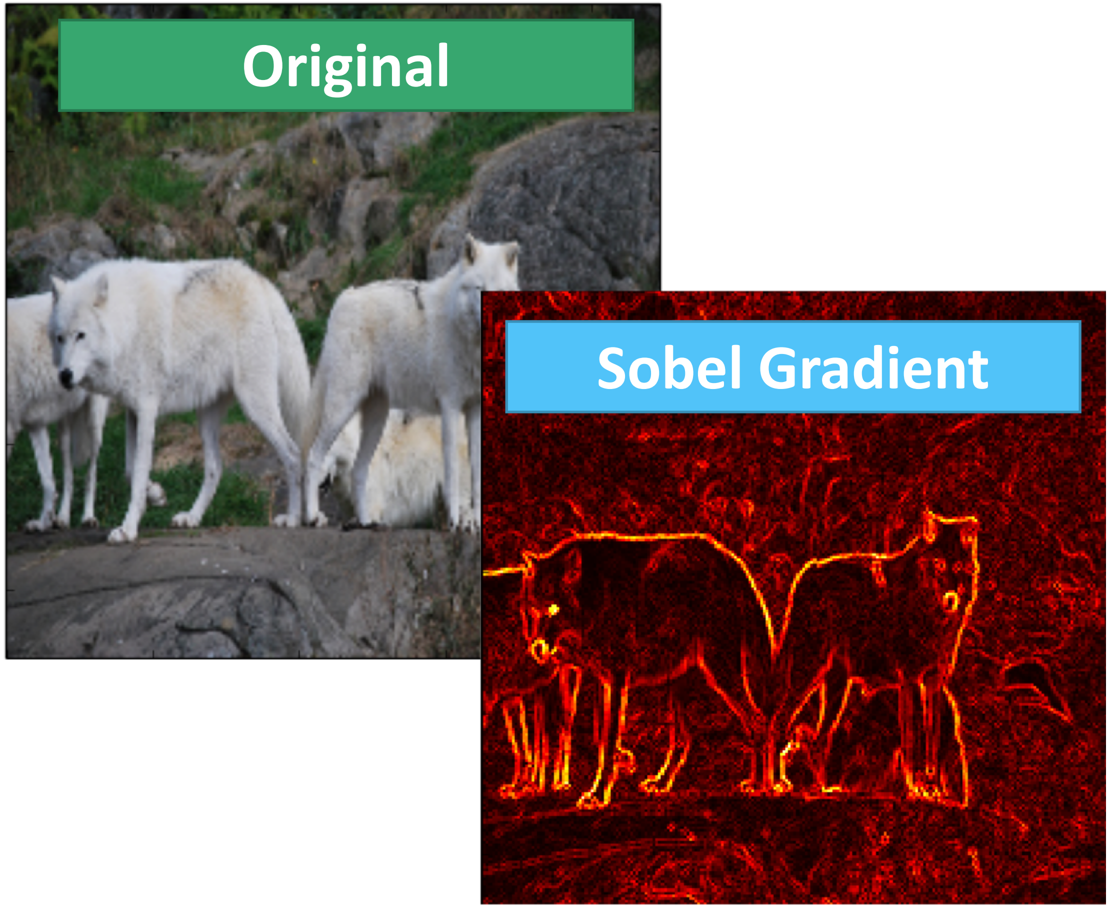
Overall Energy Savings

- Adaptive quantization saves 33% on average compared to an uncropped baseline.
- Technique can be applied in conjunction with weight quantization techniques with nearly identical relative savings



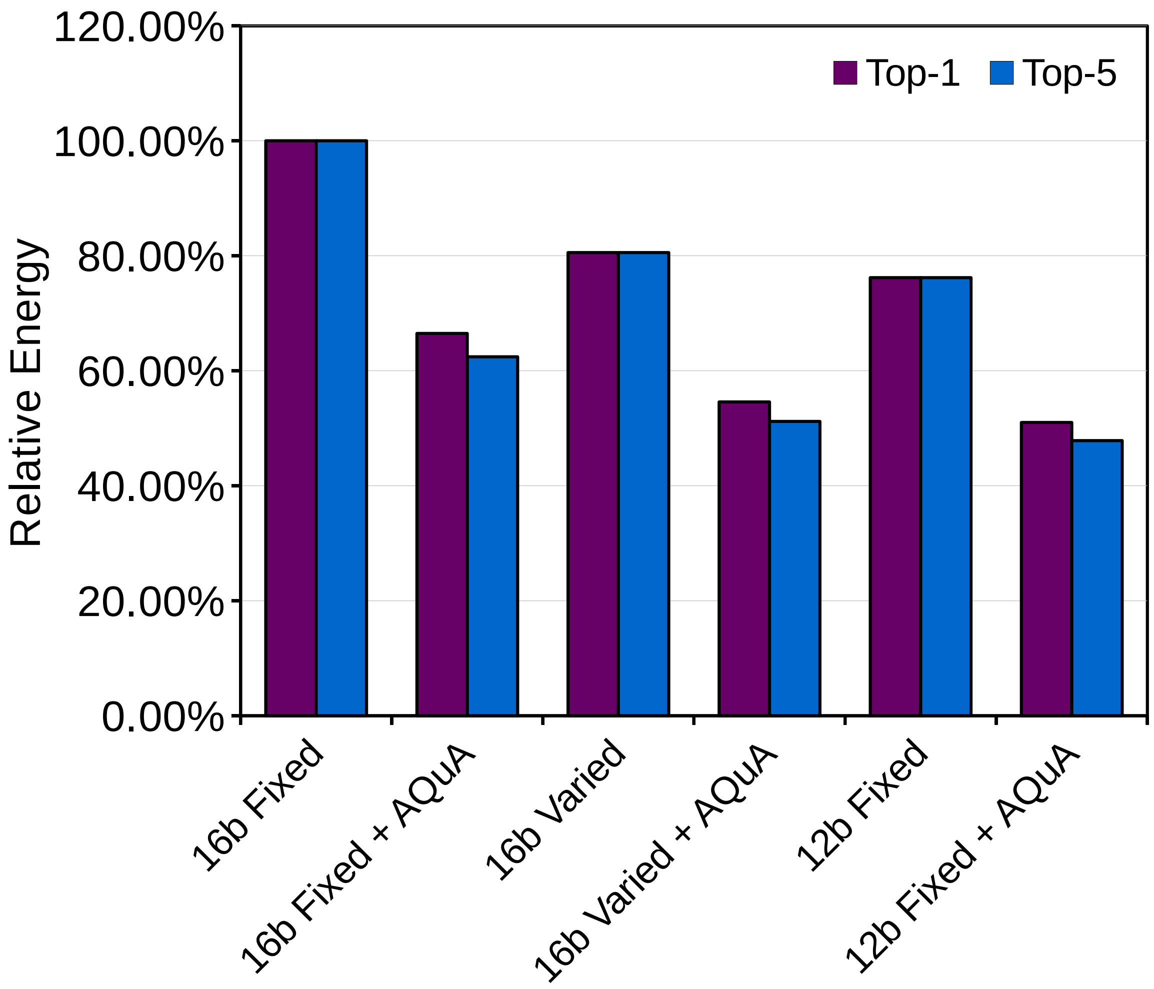
Future Work

- ***Predict*** unimportant regions
  - Using a “0^{th}” layer with a just a few gradient-based kernels
- Use ***variable low precision*** computations unimportant regions (not just cropping)
- ***Quantify energy and latency*** changes due to additional prediction step, but fewer overall computations



Conclusion

- Adaptive quantization saves 33% on average compared to an uncropped baseline.
- Technique can be applied in conjunction with weight quantization techniques with nearly identical relative savings



Thank you!

Questions?