HUAWEI

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

One-Shot Mixed Precision Search

Ivan Koryakovskiy, Alexandra Yakovleva, Valentin Buchnev,
Temur Isaev, Gleb Odinokikh

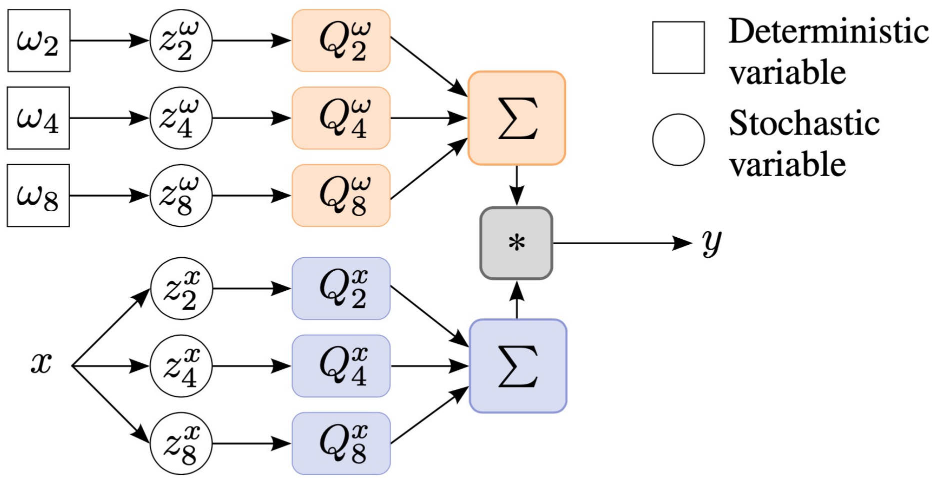
CVPR 2023
(TUE-PM-365)

Huawei Technologies

18.05.2023

One-Shot Mixed Precision Search

Contribution 1: Using Variational Inference, we theoretically derive the earlier empirically-found state-of-the-art searching methods (EdMIPS, DNAS).



- Propose a **generic approach** to model hardware constraints by a Boltzmann distribution

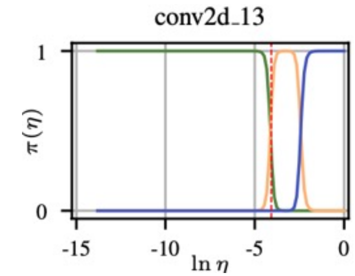
Bit width probability model $f_\theta(\eta)$

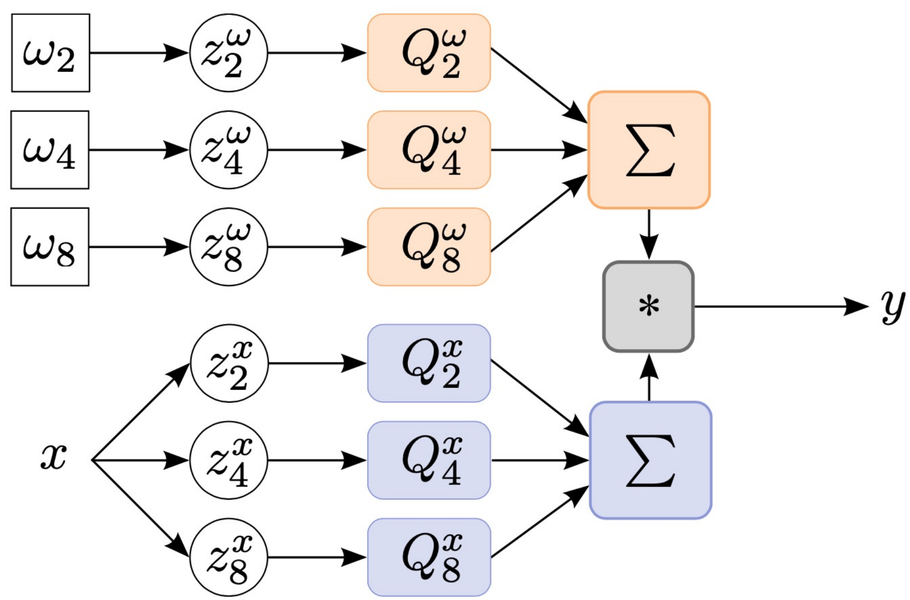
Contribution 2: We propose to augment a supernet with a bit width prediction model that allows searching for Pareto-front models in O(1) time.

$$\mathcal{L} = \mathcal{L}_{tgt} + \eta \mathcal{L}_{HW}$$

$$z \sim \text{GumbelSoftmax}(f_\theta(\eta))$$

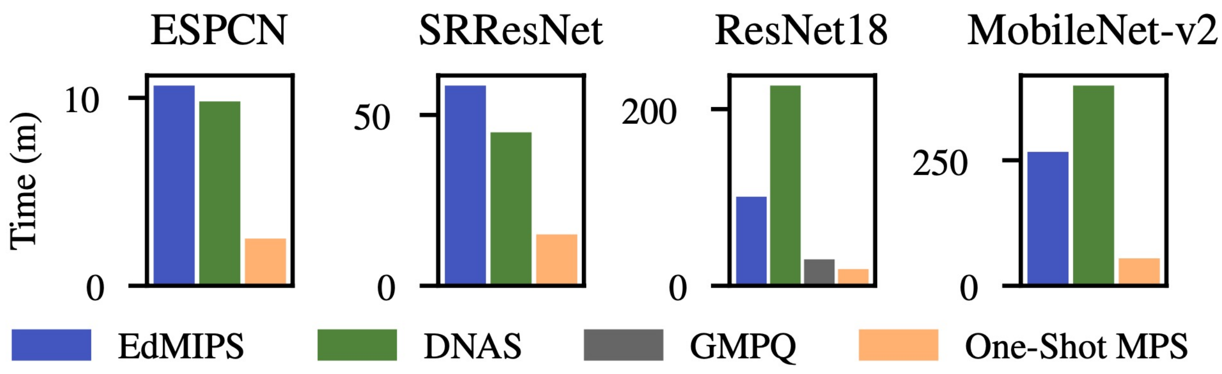
$$\pi_\theta(\eta) = \text{Softmax}(f_\theta(\eta))$$





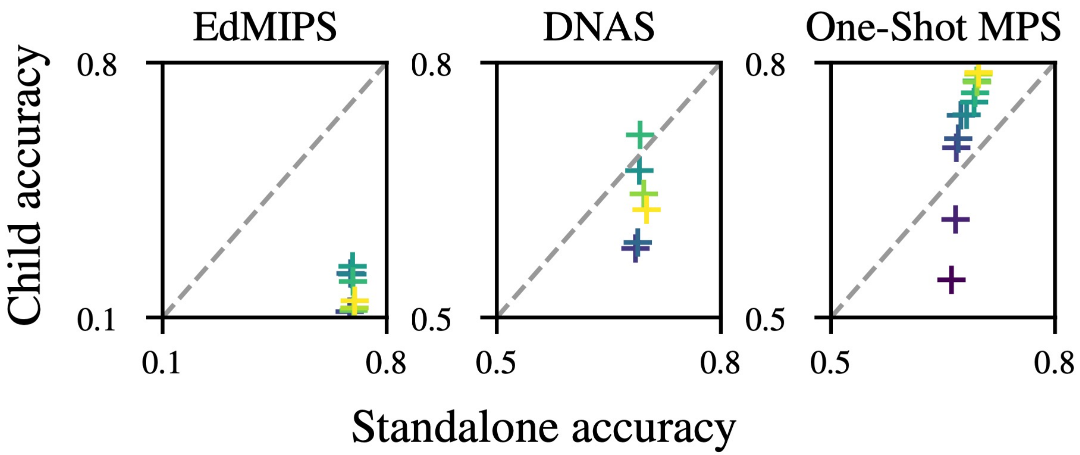
After supernet training, we can select all Pareto models by sweeping over the hardware penalty.

Pareto searching time



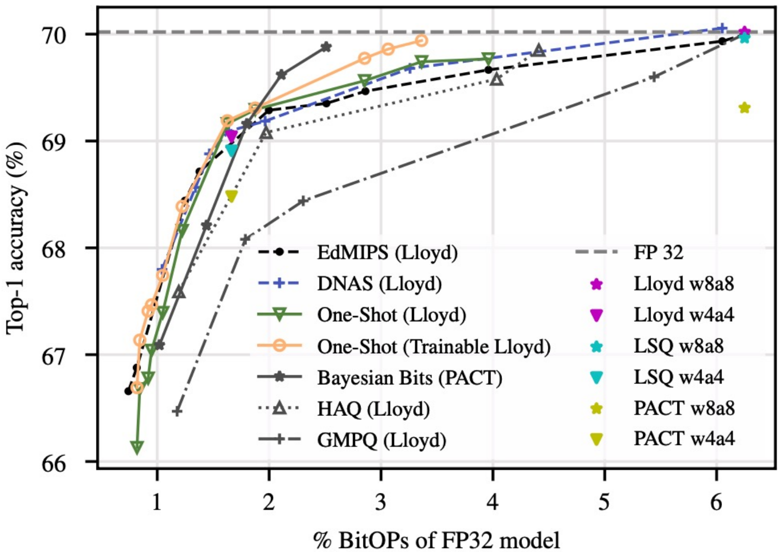
- For large models (ResNet-18 and MobileNet-v2), the proposed method is 5 times more efficient than existing methods.

Correlation between the child and standalone model performances

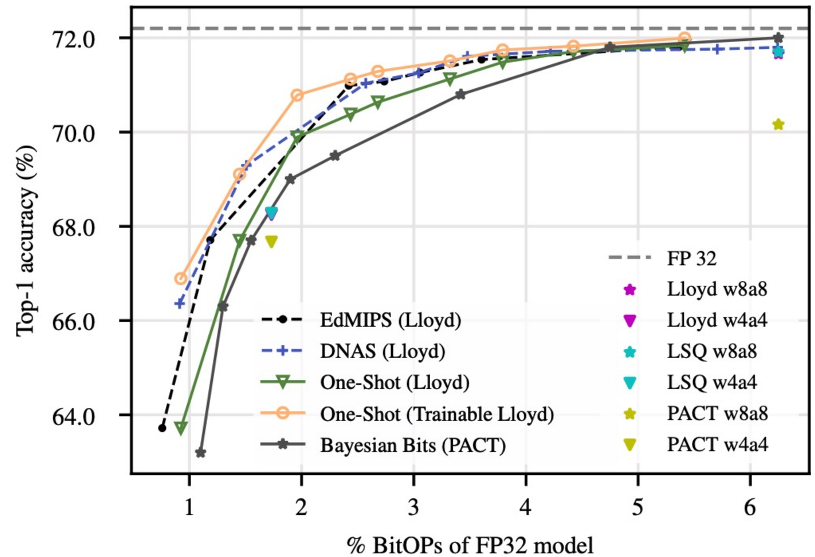


- High correlation scores (> 0.93)
- Co-adaptation of weights is avoided

Selected model quality



(c) ResNet-18 on ImageNet



(d) MobileNet-v2 on ImageNet

- Model quality is similar to other methods
- A richer set of bit width combinations is found

Quantization

Quantization is the process of converting floating-point tensors to lower precision integer tensors.

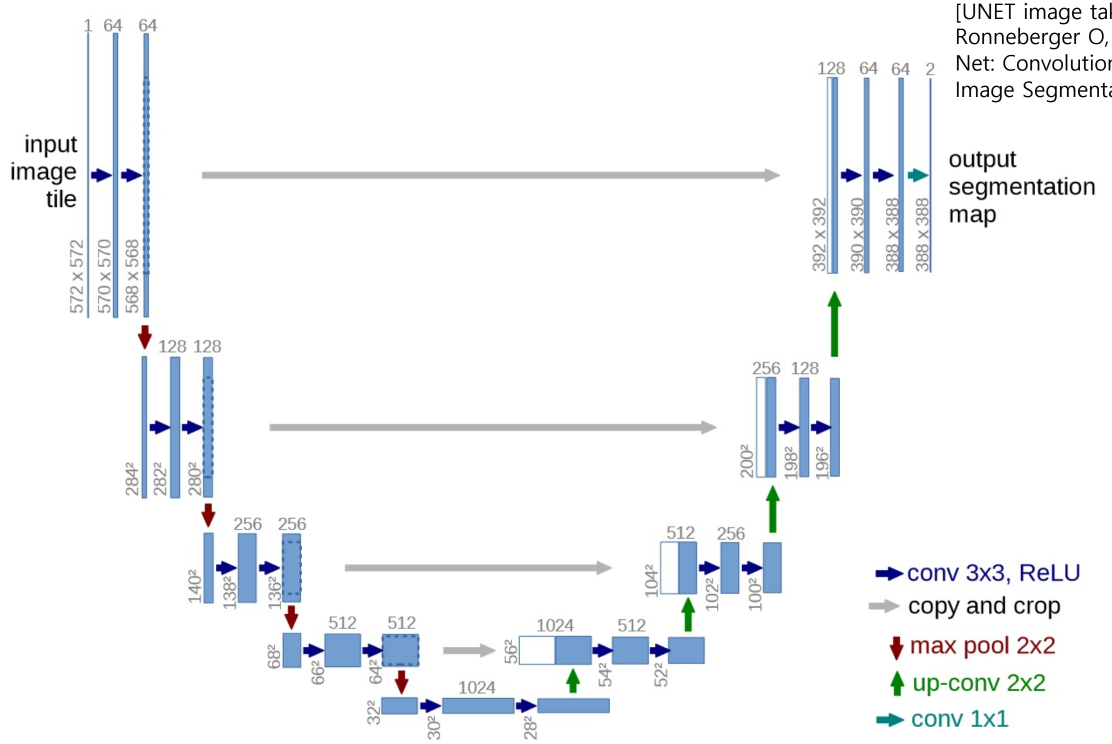
This results in a

- Reduced inference latency
- Reduced power consumption
- Potentially lower accuracy

Options:

- Quantize weights & activations
- Mixed precision quantization: {Int8, Int4, Int2}
- Uniform grid
- Quantization Aware Training (QAT)

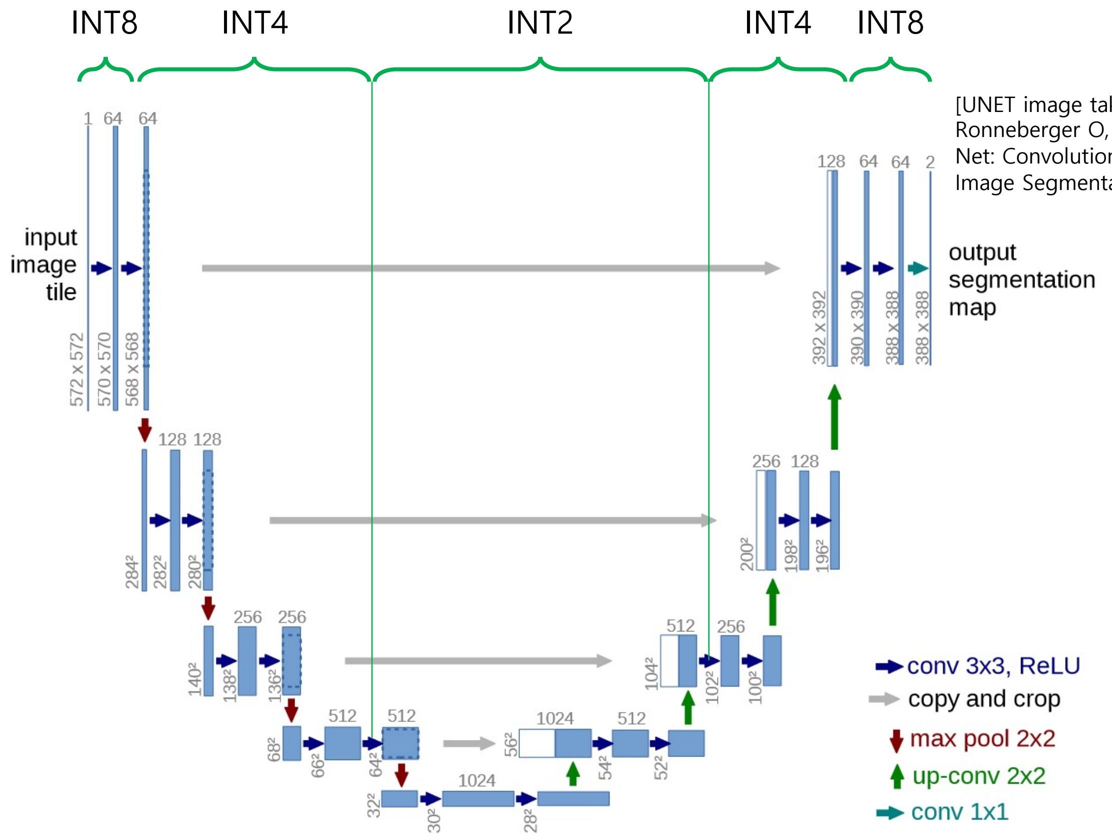
Bit Width Searching



[UNET image taken from Ronneberger O, Fischer P, Brox T (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". arXiv:1505.04597]

- Some modern hardware already supports mixed-precision operations
- Middle layers can be quantized to lower bit widths
- But search space in too large, e.g., $O(M^{2N})$, where M – is the number of bit width options, and N – is the number of layers.

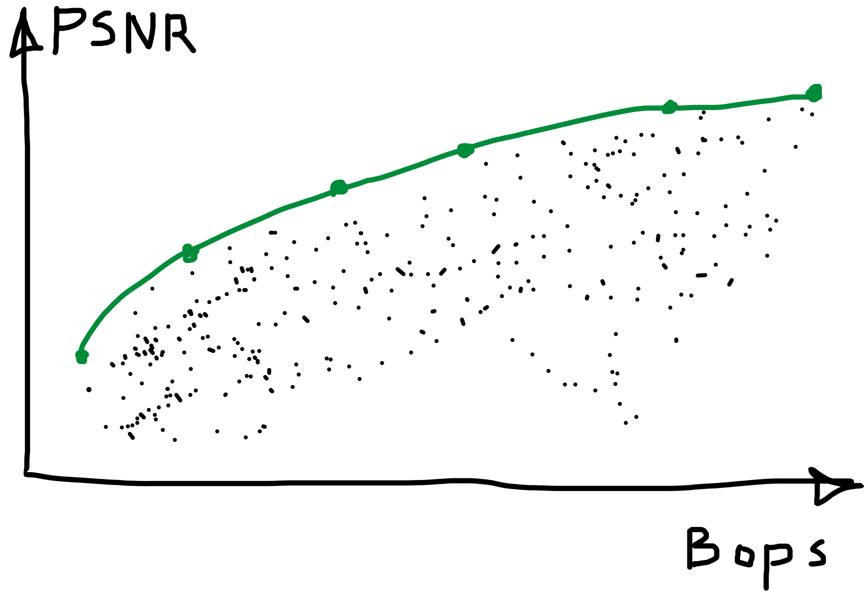
Bit Width Searching



[UNET image taken from Ronneberger O, Fischer P, Brox T (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". arXiv:1505.04597]

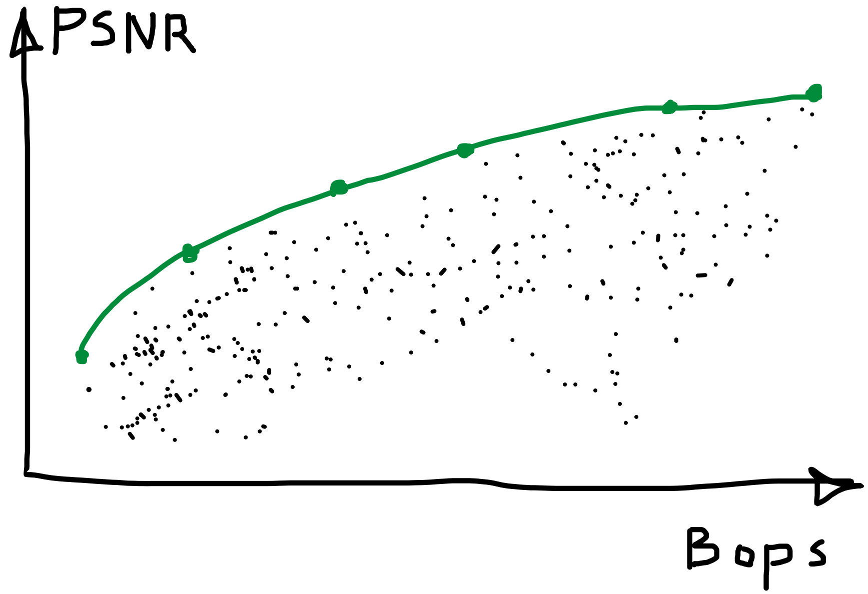
- Some modern hardware already supports mixed-precision operations
- Middle layers can be quantized to lower bit widths
- But search space in too large, e.g., $O(M^{2N})$, where M – is the number of bit width options, and N – is the number of layers.

Pareto front



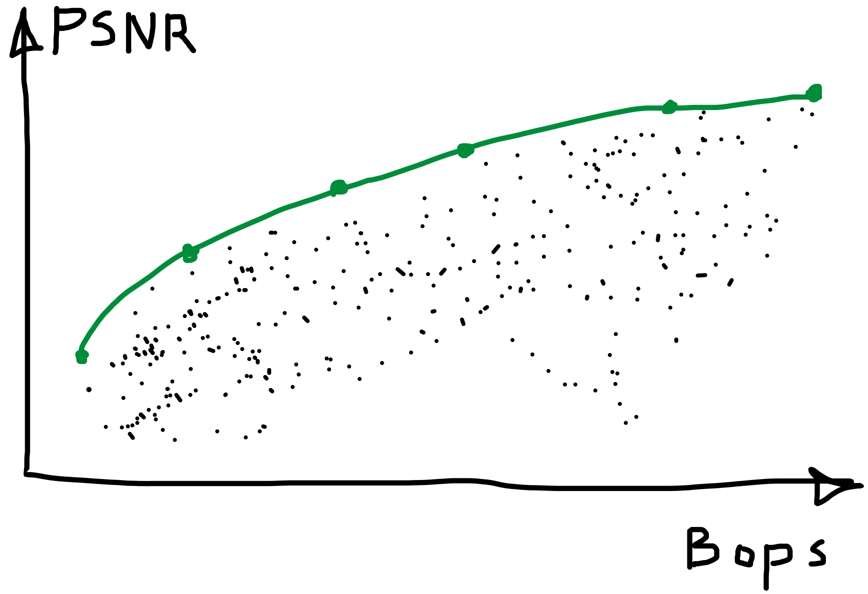
- For practical reasons, it is convenient to find multiple Pareto models
- One-stage algorithms require multiple restarts:
  - EdMIPS [1]
  - DNAS [2]
  - Bayesian Bits [3]
  - HAQ [4]
  - etc

Finding Pareto front



- Two-stage algorithms: first train a Supernet (done once), then search for suitable bit widths:
  - SPOS [5] uses Evolutionary algorithm (very slow)
  - Bit-Mixer [6] and FN^3[7] use heuristics
    (largest eigenvalue of a Hessian or pruning)

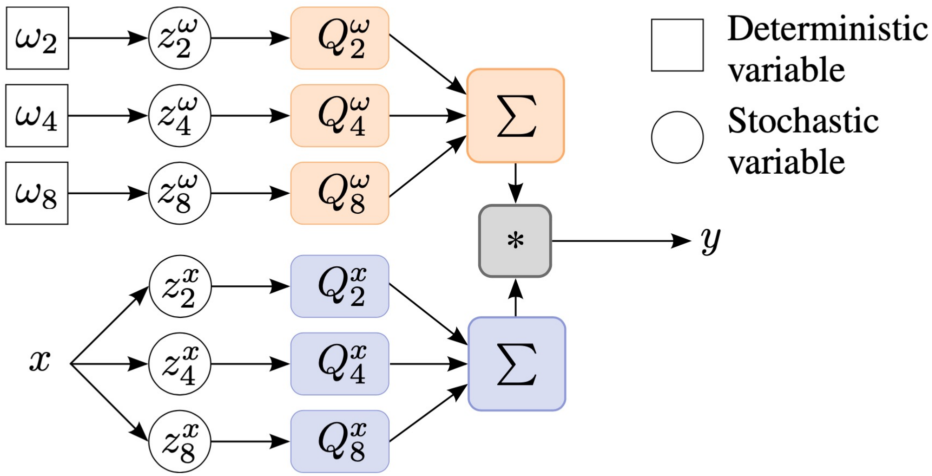
Finding Pareto front



- Two-stage algorithms: first train a Supernet (done once), then search for suitable bit widths:
  - SPOS [5] uses Evolutionary algorithm (very slow)
  - Bit-Mixer [6] and FN^3[7] use heuristics
    (largest eigenvalue of a Hessian or pruning)

Can we do better? Yes!

Variational Inference approach (VIMPS)

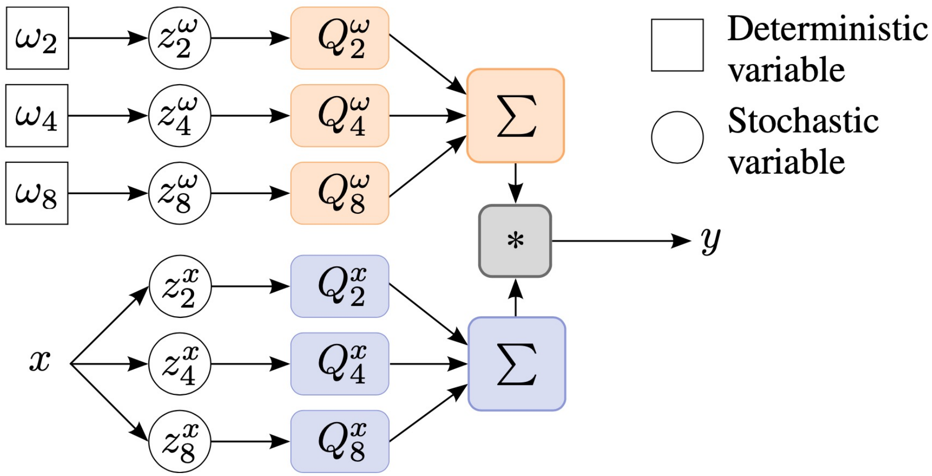


w – FP weights
x – FP activation
z – Stochastic gate
Q – Quantizer
y – Output
$\mathcal{D}$ – Dataset

The task is to find the posterior probability $p_w(z|\mathcal{D})$
We cannot calculate it using Bayes Rule, but we can approximate it using some variational distribution $q_\pi(z)$ by minimizing:

$$\mathrm{KL}\big(q_\pi(z)||p_w(z|\mathcal{D})\big) = \underbrace{-\mathcal{F}(w,\pi)}_{\text{ELBO}} + \underbrace{\log p_w(\mathcal{D})}_{\text{const w.r.t. } q}$$

Variational Inference approach (VIMPS)

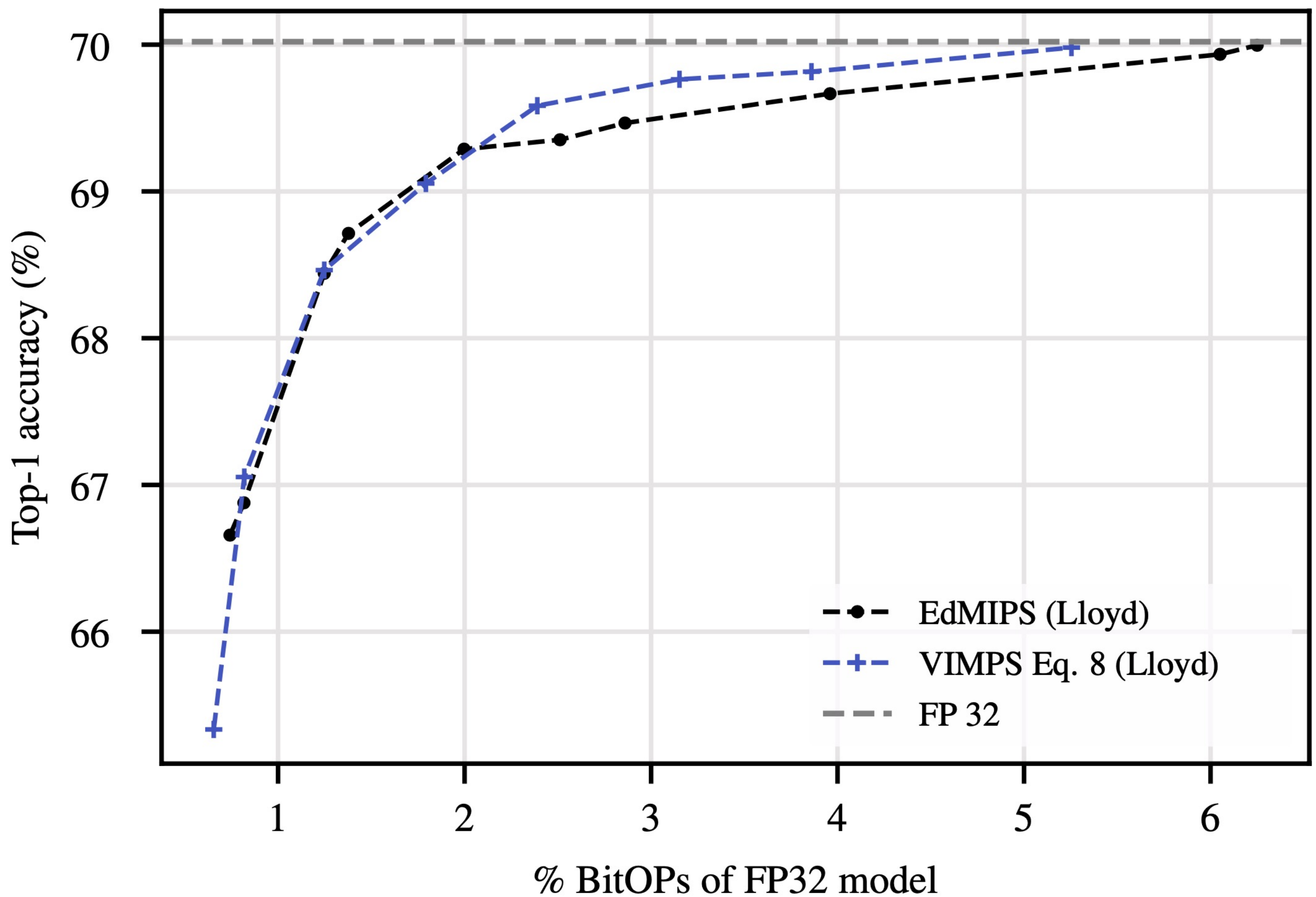


K – Number of layers
$\mathcal{B}$ – a set of all bit width configuration in a layer
η – hardware penalty
$h_{k,b}$ – is the amount of resources (BOPs) required when layer k is quantized in bit width b

$$p_{k,b} = \frac{e^{-\eta h_{k,b}}}{\sum_{b \in \mathcal{B}} e^{-\eta h_{k,b}}}$$

$$\mathcal{F}(w,\pi) = \underbrace{\mathbb{E}_{z\sim q_\pi(z)}[\log p_w(\mathcal{D}|z)]}_{\text{Task loss}} + \underbrace{\sum_{k=1}^{K}\sum_{b\in\mathcal{B}} \pi_{k,b} \log p_{k,b}}_{\text{Hardware penalty term per layer}} + \underbrace{H(\pi)}_{\text{Entropy}}$$

Variational Inference approach (VIMPS)



Generalization to EdMIPS and DNAS

$$\mathcal{F}(w,\pi) = \mathbb{E}_{z\sim q_\pi(z)}[\log p_w\,(\mathcal{D}|z)] + \eta\sum_{k=1}^{K}\mathbb{E}_{\pi^w}[b^w]\mathbb{E}_{\pi^x}[b^x]\mathrm{MACs}(k) + H(\pi)$$

1. Approximate the variational $q_\pi(z)$ using a differentiable Concrete distribution (DNAS):

$$\mathbb{E}_{z\sim q_\pi(z)}[\log p_w\,(\mathcal{D}|z)] = \mathbb{E}_{g\sim \mathrm{Gumbel}(0,1)}[\log p_w\,(\mathcal{D}|z^g)]$$

2. Use a Softmax function as a proxy for the gate probabilities (EdMIPS):

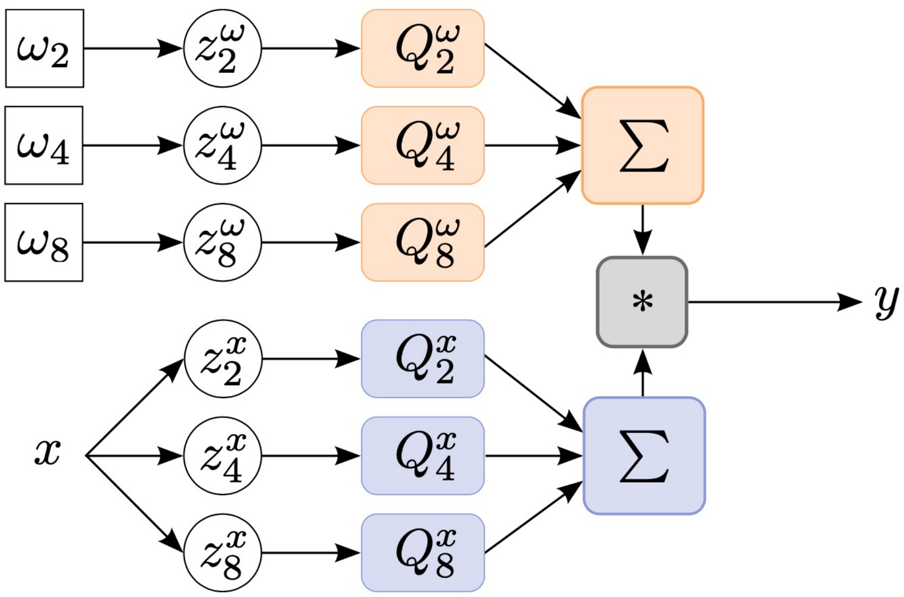
$$\mathbb{E}_{z\sim q_\pi(z)}[\log p_w\,(\mathcal{D}|z)] = \log p_w\,(\mathcal{D}|\mathrm{Softmax}(l))$$

Other differences:

1. Both DNAS and EdMIPS do not use entropy.
2. DNAS uses a multiplicative hardware loss.
3. EdMIPS uses a crude approximation of the expected conditional distribution.

Bit width probability model $f_\theta(\eta)$

How to find Pareto architectures at once?



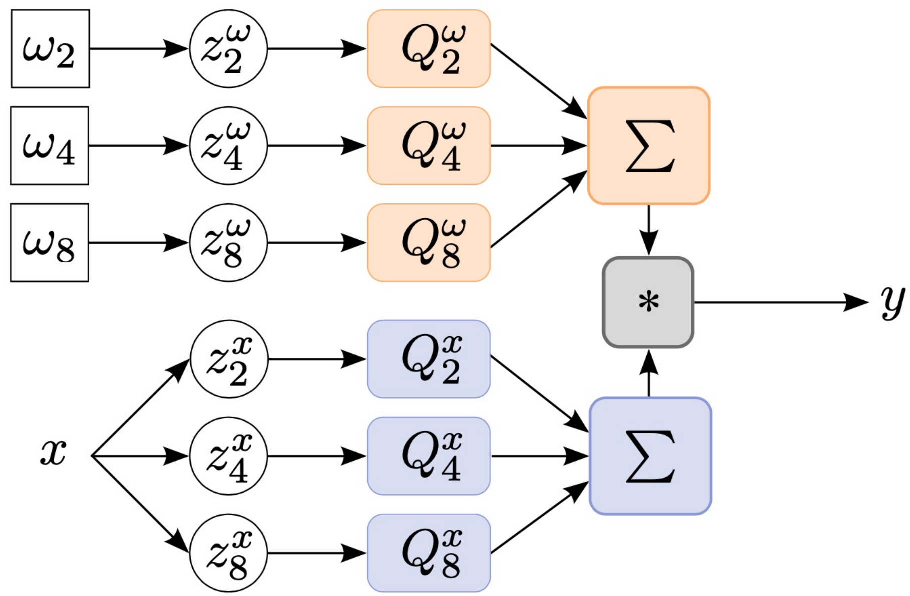
η – hardware penalty
$\pi_\theta(\eta)$ – bit width probability model
z – stochastic gate
$\mathcal{D}$ – dataset

Bit width probability model $f_\theta(\eta)$

How to find Pareto architectures at once?

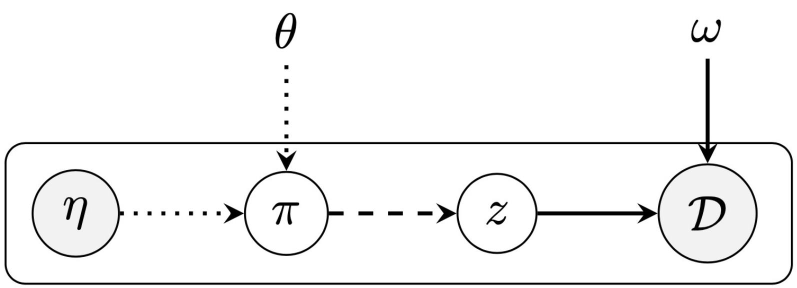
$$\mathcal{L} = \mathcal{L}_{tgt} + \eta \mathcal{L}_{HW} \qquad z \sim \mathrm{GumbelSoftmax}(f_\theta(\eta))$$

$$\pi_\theta(\eta) = \mathrm{Softmax}(f_\theta(\eta))$$



η – hardware penalty
$\pi_\theta(\eta)$ – bit width probability model
z – stochastic gate
$\mathcal{D}$ – dataset

One-shot Mixed Precision Search

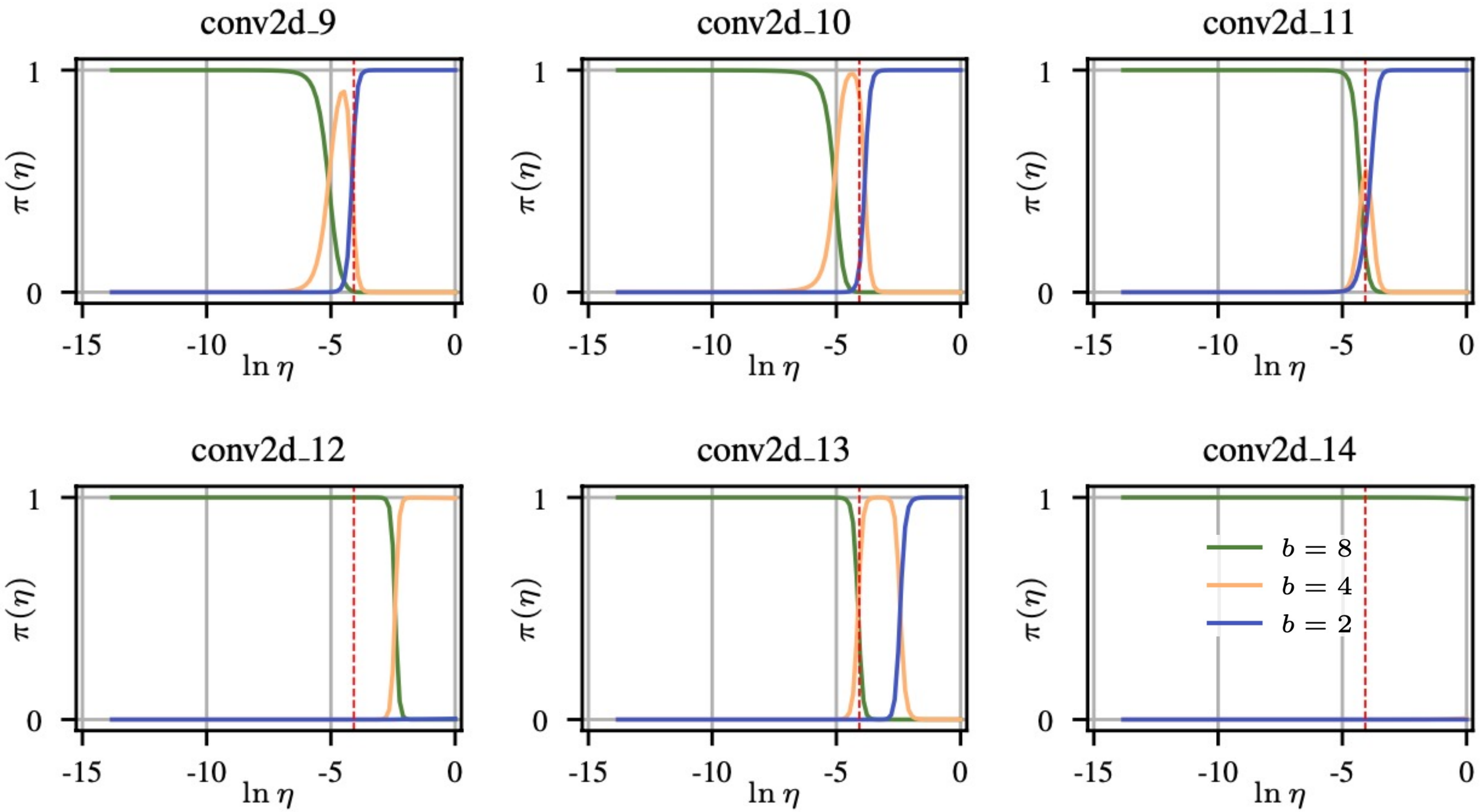


η – hardware penalty
$\pi_\theta(\eta)$ – bit width probability model
z – stochastic gate
$\mathcal{D}$ – dataset

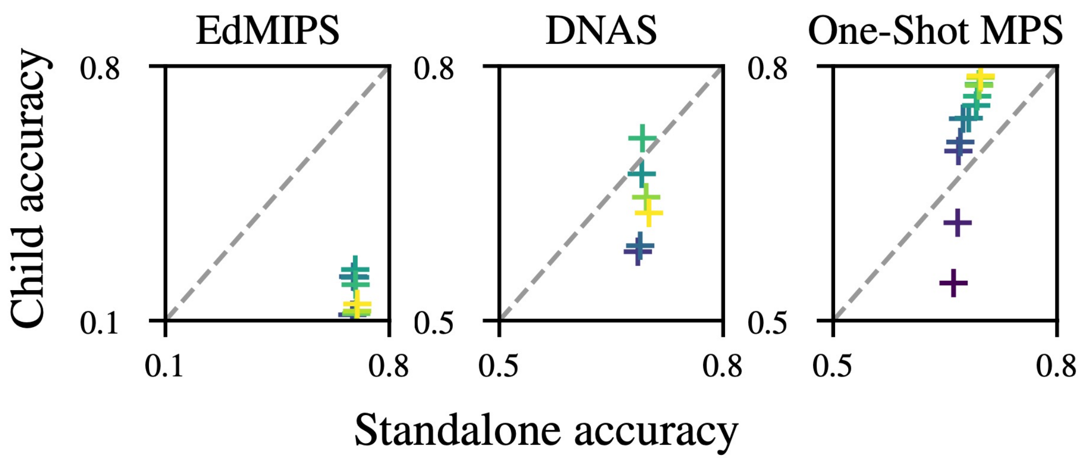
$$\mathcal{L}(\mathcal{D}; w, \theta) = -\mathbb{E}_{\eta \sim p(\eta)} \left[\mathbb{E}_{z \sim q_{\pi(\eta)}(z)} [\log p_w (\mathcal{D}|z)] + \sum_{k=1}^{K} \sum_{b \in \mathcal{B}} \pi_{k,b}(\eta) \log p_{k,b}(\eta) + \lambda H(\pi(\eta)) \right]$$

$$\underbrace{+\mu \sum_{i \in \mathcal{B}^w} \sum_{\substack{j \in \mathcal{B}^w \\ j > i}} \left\| w_i - w_j \right\|_2}_{\text{Kernel similarity loss}}$$

Intuition behind the bit width probability model



Correlation between the child and standalone model performances



Model	Method	Kendall's Tau correlation score
ResNet-18	One-Shot	**0.97**
	EdMIPS	0.29
	DNAS	0.52

- High correlation scores (> 0.93)
- Co-adaptation of weights is avoided

Conclusion

1. We theoretically derived two new searching methods: VIMPS and One-Shot MPS.

2. We showed that the bit width probability model allows for a straight-forward Pareto-front architecture selection.

3. The bit width probability model imposes structure on the selected architectures due to which we can
    1. find a richer set of bit width combinations, and
    2. improve a Kendall's tau correlation which is useful for predicting the fine-tuned model performance.

For questions, feel free to email me: i.koryakovskiy@gmail.com

References

[1] Zhaowei Cai and Nuno Vasconcelos. Rethinking Differentiable Search for Mixed-Precision Neural Networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2020.

[2] Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed Precision Quantization of Convnets via Differentiable Neural Architecture Search. *arXiv:1812.00090*, 2018.

[3] Mart Van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. Bayesian Bits: Unifying Quantization and Pruning. *Advances in Neural Information Processing Systems*, 33:5741– 5752, 2020.

[4] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. HAQ: Hardware-Aware Automated Quantization with Mixed Precision. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8612–8620, 2019.

[5] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. *In Proceedings of the European Conference on Computer Vision*, pages 544–560, 2020.

[6] Adrian Bulat and Georgios Tzimiropoulos. Bit-Mixer: Mixed-Precision Networks with Runtime Bit-Width Selection. *In Proceedings of IEEE International Conference on Computer Vision*, 2021.

[7] Yufei Cui, Ziquan Liu, Wuguannan Yao, Qiao Li, Antoni B. Chan, Tei-wei Kuo, and Chun Jason Xue. Fully Nested Neural Network for Adaptive Compression and Quantization. *In Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.