BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023

Evolutionary Optimization of Convolutional Neural Networks

Vojtěch Čoupek

Supervisor: prof. Ing. Lukáš Sekanina, Ph.D.

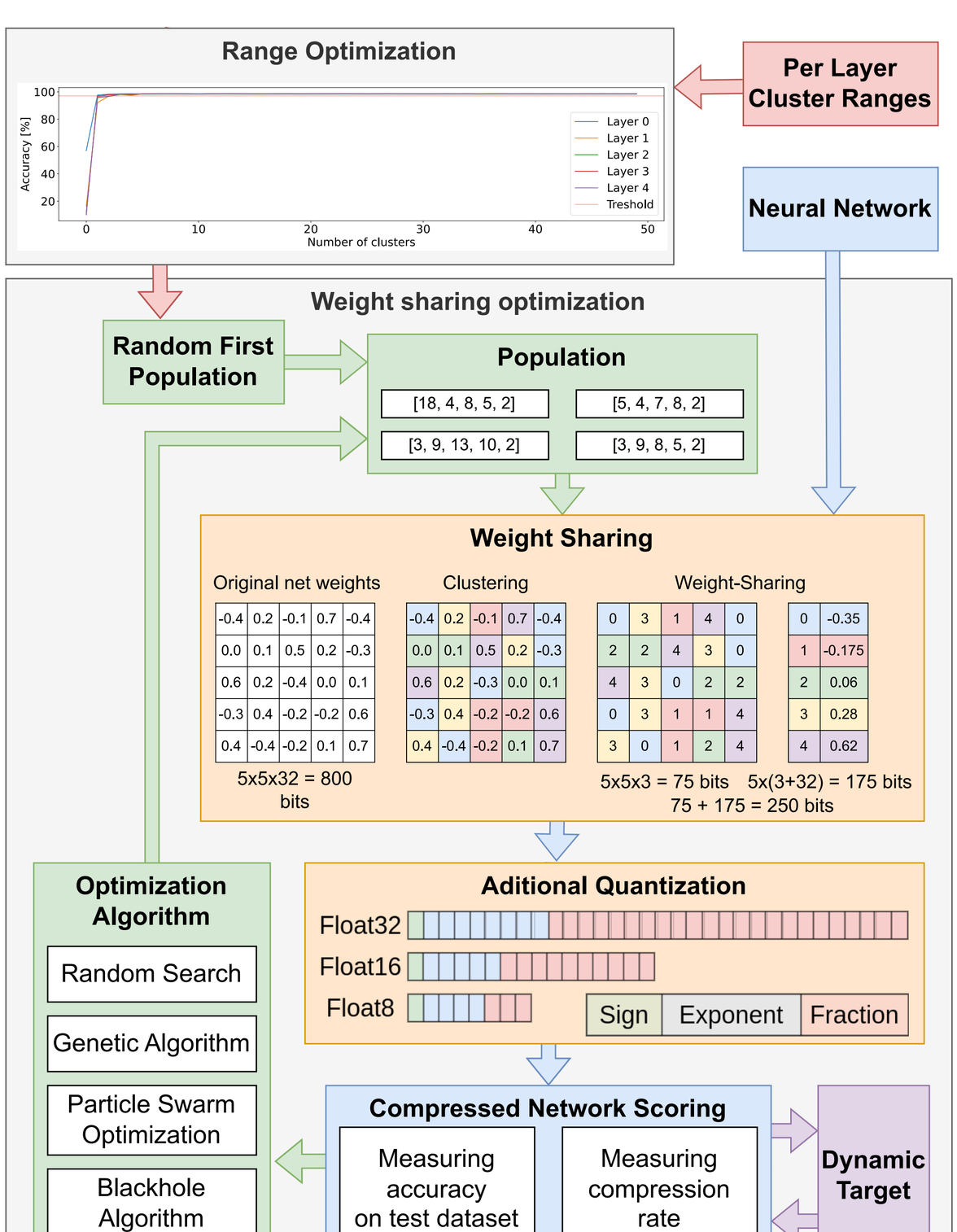

Figure 1: Weight-Sharing optimization computation flow

Weight-Sharing optimization

- Compressing neural network while retaining accuracy
- Smaller weight memory → Shorter read time → Lower energy consumption

1. Range optimization – iterative sharing of layers by a given number of clusters. If the accuracy drops below the threshold, it is removed from the search space.
2. Generate first population. Each member wraps a list of cluster numbers – position in the list refers to the layer in the model.
3. Weight-Sharing – for each member it performs K-means clustering layer by layer. Weights are replaced by the cluster means. Additional Quantization to the means can be applied.
4. Compressed network scoring – For each member, compression rate (CR) and accuracy (ACC) is computed. If CR or ACC of any member surpasses the $\mathrm{ACC}_{\mathrm{target}}$ or $\mathrm{CR}_{\mathrm{target}}$ values, they are updated, then the fitness is computed by equation 1.
5. A new population is generated by the chosen optimization algorithm and the process is repeated from step 2 for the given number of iterations.
6. The best found solution is fine-tuned by modulating K-means clustering space adding the second dimension using equation 3 – Searching for the best spread and focus values for each layer is the last step.

$$\mathrm{Fit} = \frac{1}{\sqrt{(1 - \frac{\mathrm{ACC}_{\mathrm{current}}}{\mathrm{ACC}_{\mathrm{target}}})^2 + (1 - \frac{\mathrm{CR}_{\mathrm{current}}}{\mathrm{CR}_{\mathrm{target}}})^2}} \tag{1}$$

$$\mathrm{CR} = \frac{\mathrm{n\_w} \cdot \mathrm{bits}_{\mathrm{weight}}}{\mathrm{n\_w} \cdot \mathrm{bits}_{\mathrm{key}} + \mathrm{n\_k} \cdot (\mathrm{bits}_{\mathrm{red\_weight}} + \mathrm{bits}_{\mathrm{key}})} \tag{2}$$

$$y = \mathrm{spread} \cdot (\max(\mathbf{w}) - \min(\mathbf{w})) \cdot \mathrm{Tanh}(\mathrm{focus} \cdot (x - \mathrm{mean}(\mathbf{w})) \tag{3}$$

Results – Le-Net-5

Activation	Quantization	Compress.	Acc. [%]	Acc. ft [%]	Acc. loss [%]
ReLu	float32	12.7529	97.56	97.80	0.66
	float16	14.1083	97.54	97.66	0.80
	float8	12.7291	97.48	97.78	0.68
Tanh	float32	19.9433	97.66	98.04	0.60
	float16	20.5451	97.68	98.02	0.62
	float8	19.7939	97.86	98.04	0.60

Table 1: Results of weight compression on Le-Net-5 with MNIST dataset

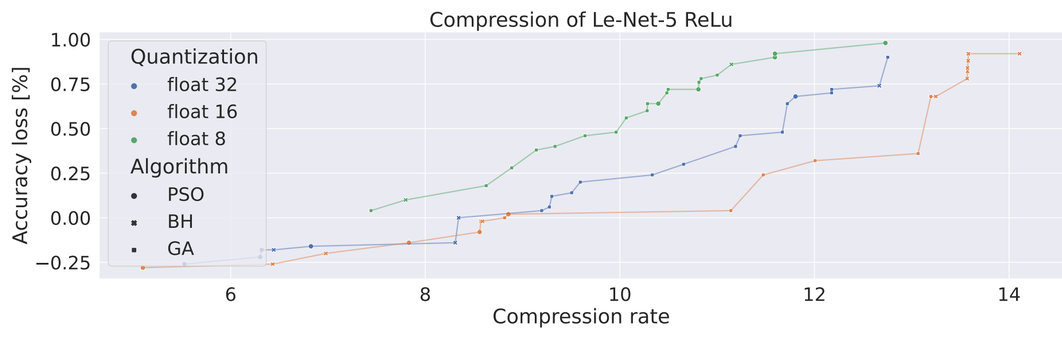

Figure 2: Different additional Quantizations on Weight-Sharing – Le-Net-5 ReLu

Results – Mobilenet_v2

Quantization	Compression	Top-1 Acc. [%]	Top-1 Acc. loss [%]
float32	4,8605	80,75	2,26
float16	4,7617	80,53	2,47
float8	–	–	–

Table 2: Results of weight compression on Mobilenet_v2 with Imagenette dataset