Binarization of Neural Networks as an Optimization Problem

Danila Doroshin

Principal Engineer, PhD

Mathematical Modeling & Optimization Algorithm Competence Center

www.huawei.com

Outline

1. Sparsification – from millions of parameters to thousands

2. Compression as a Quadratic Optimization

3. From floating point to fixed point

Sparsification of DL models
Variational Dropout

Variational Dropout Sparsifies Deep Neural Networks *(Dmitry Molchanov et al.) ICML 2017*

1. Variational Lower Bound:

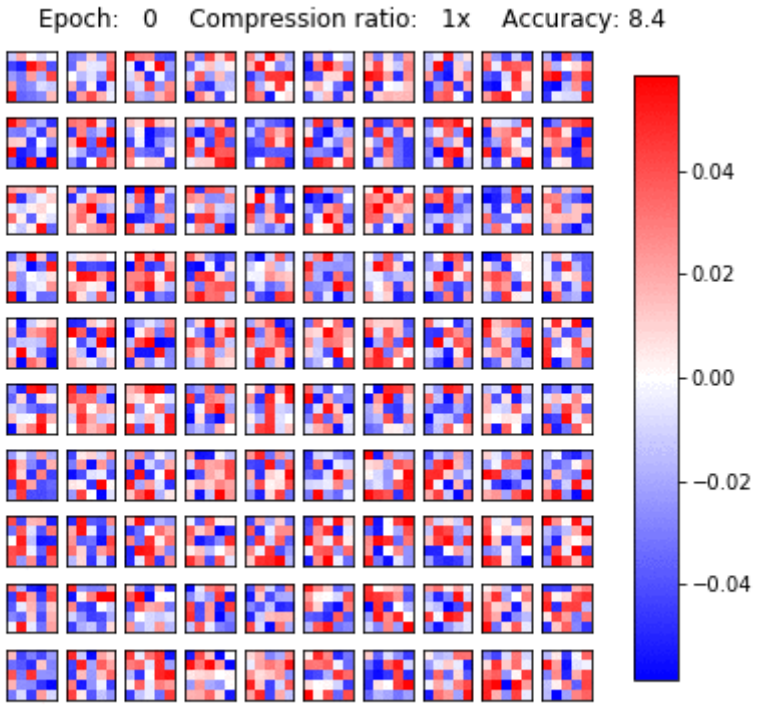
$$\mathbb{E}_{q(\widetilde{W}\,|\,\phi)} \log p(y\,|\,x, \widetilde{W}) - D_{KL}(q(\widetilde{W}\,|\,\phi)\,||\,p(\widetilde{W})) \to \max_{\phi}$$

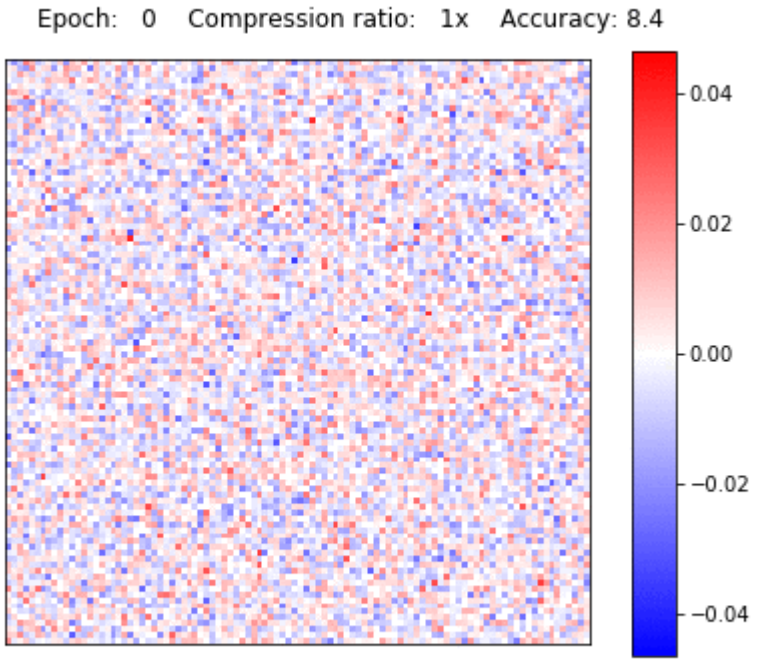
2. Prior distribution: $p(|w_{ij}|) \propto \dfrac{1}{|w_{ij}|}$

3. Gaussian approximation for prior: $w_{ij} \sim \mathcal{N}(\mu_{ij}, \alpha_{ij}\mu_{ij}^2)$

Sparsification of DL models

Variational Dropout Sparsifies Deep Neural Networks (Dmitry Molchanov et al.) ICML 2017





https://github.com/ars-ashuha/variational-dropout-sparsifies-dnn

Sparsification of DL models
Pros and Cons of Variational Dropout

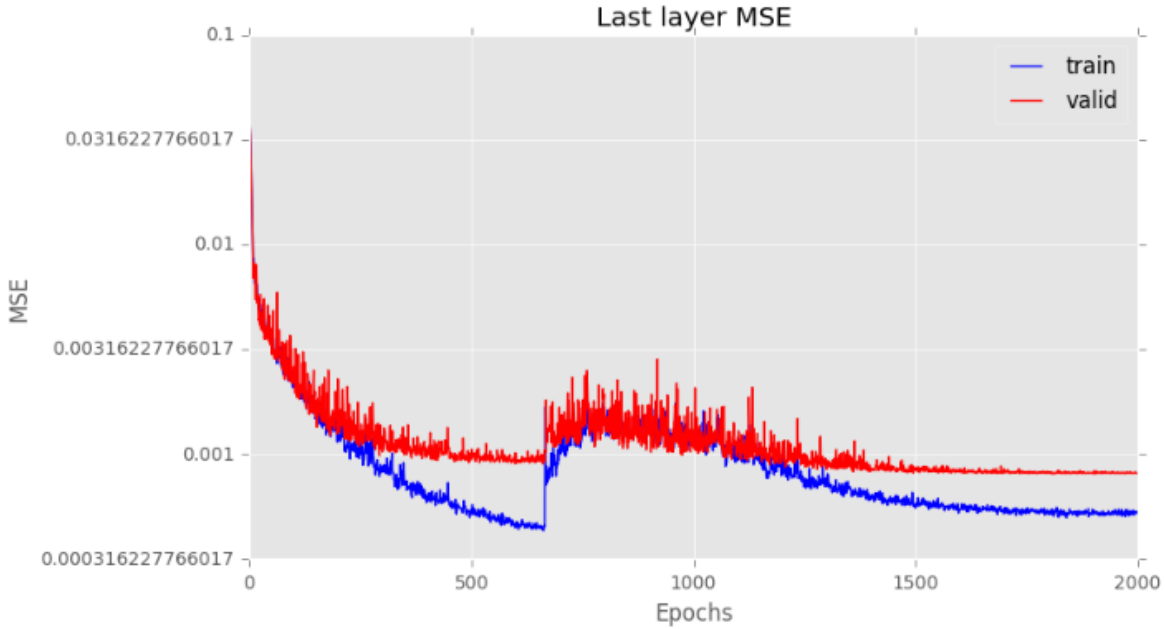
Pros: reduce from millions of parameters to tens of thousands.

Cons:
1. Unpredictable final performance. Fine tuning with turned off Dropout is required.

2. Sensitive to regularizer weight. High risk of under/over compression.

3. Has to be adapted to nonstandard layers.

4. Prior is very strong, hence
a) Good for compression;
b) Bad for performance.



Sparsification of DL models

What else?

Brain Surgeon

Sparsification of DL models
Optimal Brain Surgeon

Optimal Brain Surgeon: Extensions and performance comparisons
(Babak Hassibi et al.) *Advances in neural information processing systems*. 1994

$$\delta E = \underbrace{\left(\frac{\partial E}{\partial \mathbf{w}}\right)^T}_{\approx 0} \cdot \delta \mathbf{w} + \frac{1}{2}\, \delta \mathbf{w}^T \cdot \underbrace{\frac{\partial^2 E}{\partial \mathbf{w}^2}}_{\equiv \mathbf{H}} \cdot \delta \mathbf{w} + \underbrace{O(\|\delta \mathbf{w}\|^3)}_{\approx 0},$$

The Hessian can be estimated as a covariance matrix of the stochastic gradients (it is due to the Fisher Information properties of the Maximum Likelihood estimation).

Sparsification of DL models

Optimal Brain Surgeon as a Quadratic Programming

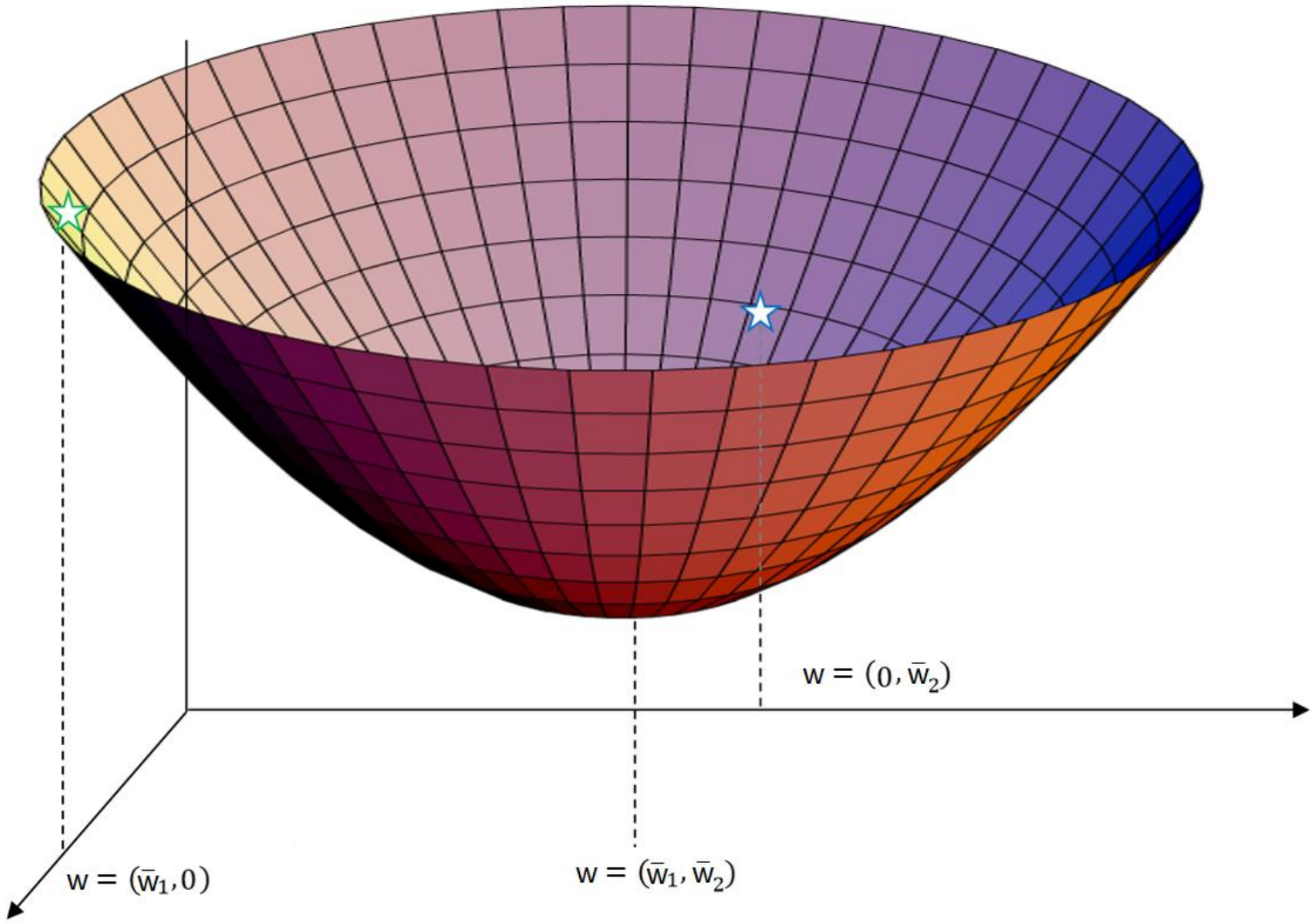
$$\frac{1}{2}(\mathrm{w}-\bar{\mathrm{w}})^{\mathrm{T}}\boldsymbol{H}\,(\mathrm{w}-\bar{\mathrm{w}}) \underset{\theta}{\rightarrow} min$$

$$\text{wrt } \mathbf{constraints}(\mathrm{w}) = 0$$

a) Sparsification constraints: keep predefined percentage of nonzero parameters. Here we reduce from tens of thousands of parameters to thousands.

b) Switch from floating point to the fixed point arithmetic *(p, k)*.

$$\mathrm{w}_i = \sum_{m=0}^{k} \alpha_m^i 2^{p-m}$$

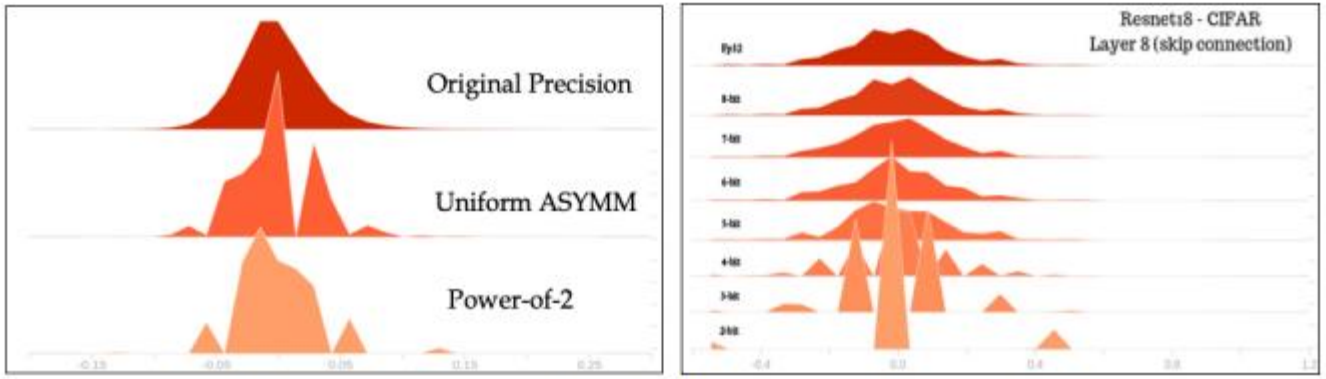


Sparsification of DL models
From floating point to fixed

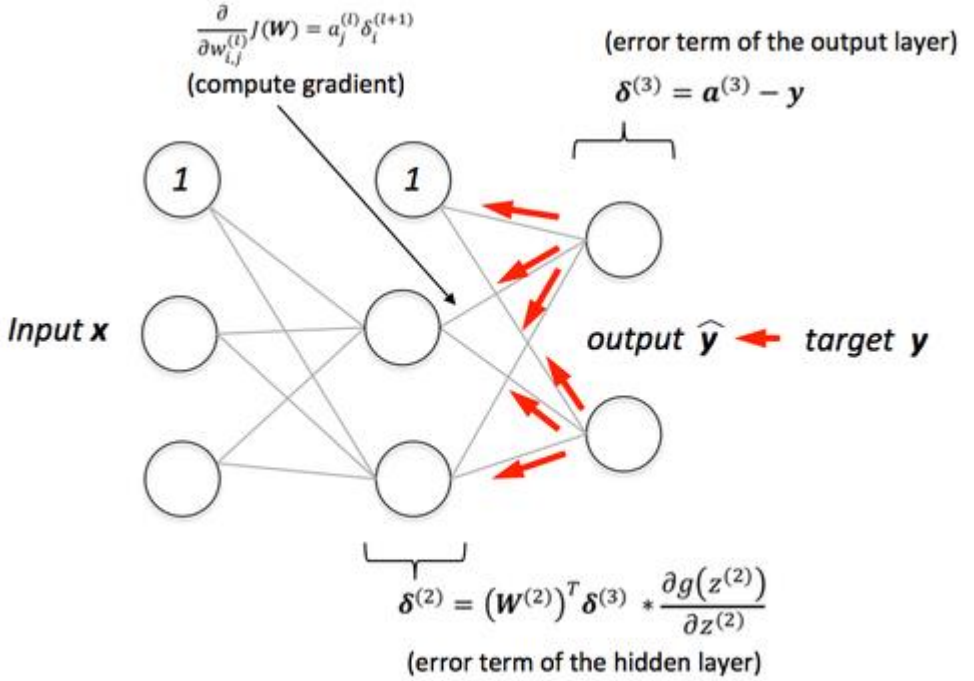
Can we reduce a fixed point approximation to one of the standard optimization problem?

$$w_i = \sum_{m=0}^{k} \alpha_m^i 2^{p-m}$$

One possible way is to consider the rounding error accumulation process as a stochastic process on the graph.



Histogram of weight distribution of ResNet Layers for bit precisions and approaches.



Bit Efficient Quantization for Deep Neural Networks (P Nayak et. al.) 2019

We reduce from millions
of parameters to thousands

Our goal is hundreds

For any questions:

doroshin.danila@huawei.com