

Optimization for inference

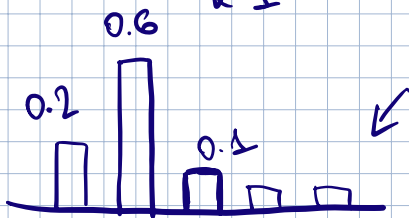
- Calibration & uncertainty estimation
- Distillation
- Pruning
- Quantization

Calibration

$$P_i = \frac{\exp(x_i)}{\sum_{k=1}^c \exp(x_k)}$$

x - logits

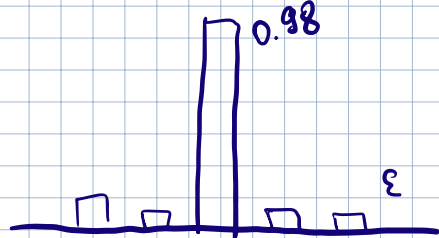
p - probabilities



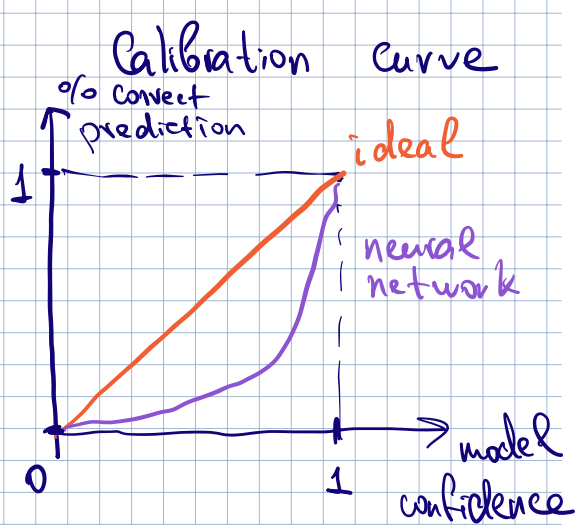
confident prediction



unconfident prediction



reality



- Temperature softmax ($\tau > 0$)

$$p_i^\tau = \frac{\exp(x_i/\tau)}{\sum_{k=1}^C \exp(x_k/\tau)}$$

$$i = \operatorname{argmax} x$$

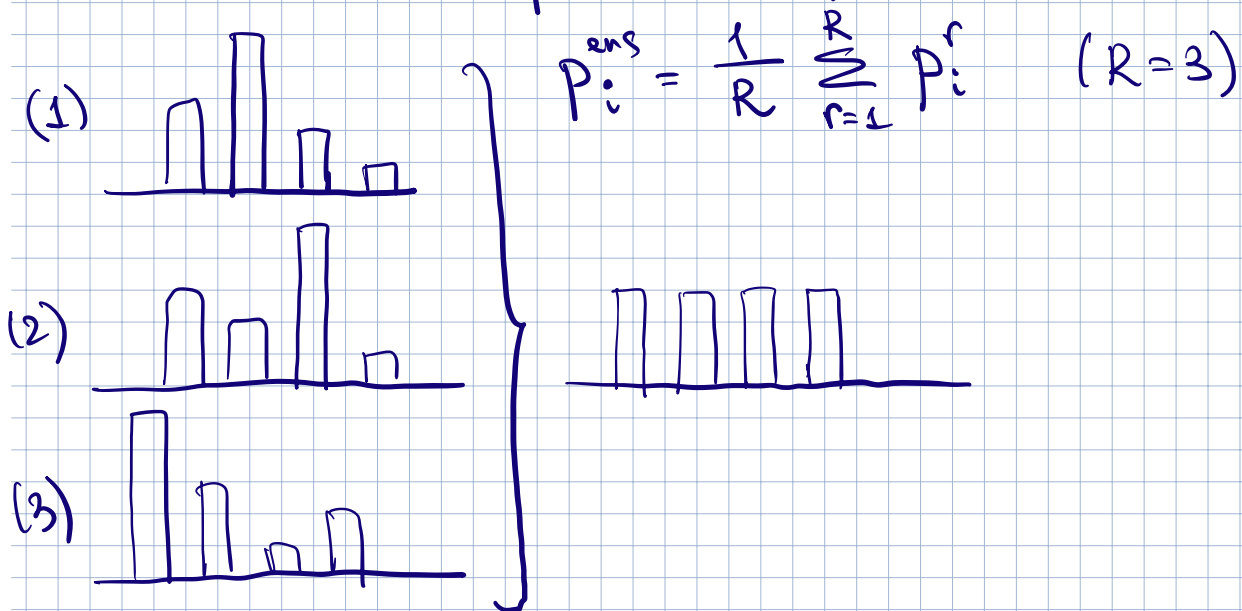
$$1) \tau \rightarrow 0 \quad p^\tau \rightarrow \delta(\operatorname{argmax} x) = (0, 0, \dots, 0, \overset{\uparrow}{1}, 0, \dots, 0)$$

$$2) \tau \rightarrow \infty \quad p^\tau \rightarrow \mathcal{U}(1, \dots, C)$$

$\tau > 1$ usually in practice

$$\tau^* = \operatorname{argmin}_{\tau} \underbrace{\left(- \sum_{n=1}^N \sum_{k=1}^C \mathbb{I}[y_n = k] \log p_{nk}^\tau \right)}_{\text{on validation set}}$$

- Ensemble (Deep Ensemble)



(Knowledge) Distillation

P_{nk}^T - teacher probability
 P_{nk}^S - student probability } with temperature

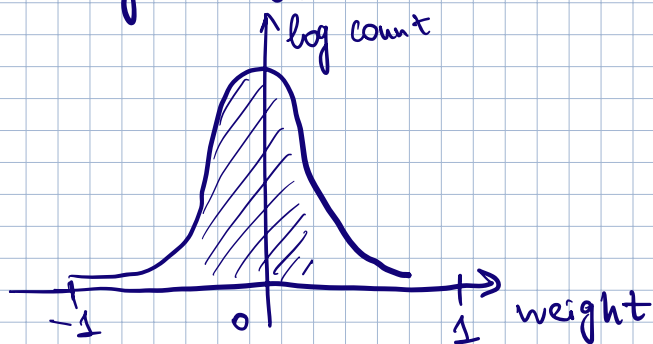
$$\text{CE loss : } \mathcal{L}_{CE} = - \sum_{n=1}^N \sum_{k=1}^C [y_n=k] \log P_{nk}^S$$

$$\text{Distill. loss : } \mathcal{L}_D = - \sum_{n=1}^N \sum_{k=1}^C P_{nk}^T \log P_{nk}^S$$

$$\mathcal{L} = \alpha \cdot \mathcal{L}_D + (1-\alpha) \mathcal{L}_{CE}, \quad \alpha \in (0,1)$$

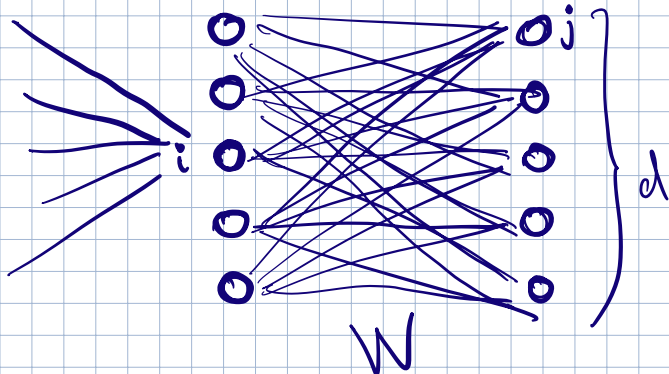
Pruning

Weights distribution



drop $\alpha\%$ weights
with lowest magnitude
- unstructured pruning

- structured pruning



$$\hat{P}_i = \sqrt[\alpha]{\sum_{j=1}^d |w_{ij}|^\alpha}$$

Several pruning stages:

1. Drop $\alpha\%$ weight
2. Fine-tune pruned network

SVD of linear layers

$$W \in \mathbb{R}^{d_1 \times d_2}$$

$W = U \Sigma V$ - SVD of weight matrix

$$U \in \mathbb{R}^{d_1 \times s} \quad \Sigma \in \mathbb{R}^{s \times s}$$

$$V \in \mathbb{R}^{s \times d_2}$$

Quantization

float 32 \rightarrow int 8

$W \in \mathbb{R}^{d_1 \times d_2}$ - weight matrix

w - one scalar weight (float 32)

$$w = s_w (w^q - z_w)$$

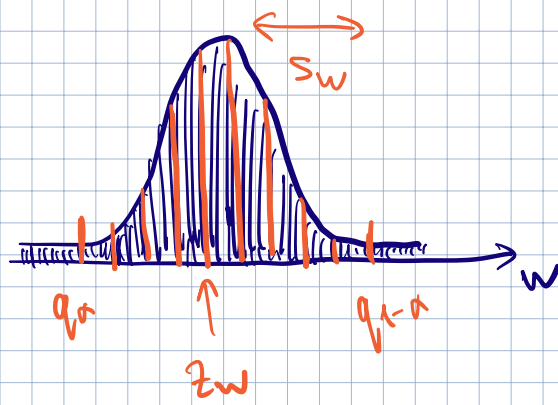
s_w - float 32 - scale parameter

z_w - int 8 - zero parameter

w^q - int 8 - quantized weight

$$w^q = z_w + \frac{w}{s_w}$$

$$w^q = \text{clip} \left(\text{round} \left(\frac{w}{s_w} \right) + z_w, -128, 127 \right)$$



$$q_\alpha \rightarrow -128$$
$$q_{1-\alpha} \rightarrow 127$$