Data2Vec

Teacher parameterization:

EMA update

$$\Delta \leftarrow \tau \Delta + (1-\tau)\,\theta$$

τ → scheduled hyper parameter.

Target

Top k block in teacher network

a_t^l → output of block l at t time

target: $y_t = \frac{1}{K} \sum_{l=L-K+1}^{L} \hat{a}_t^l$ → normalization (prevent collapse) BN

(for student)

Objective:

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & ; |y_t - f_t(x)| \le \beta \\ |y_t - f_t(x)| - \frac{1}{2} & ; \text{otherwise} \end{cases}$$

squared → L_1 loss

[β controls transition from squared loss to L_1 loss]