Self-Distillation as Instance-Specific Label Smoothing

Zhilu Zhang (NeurIPS 2020)

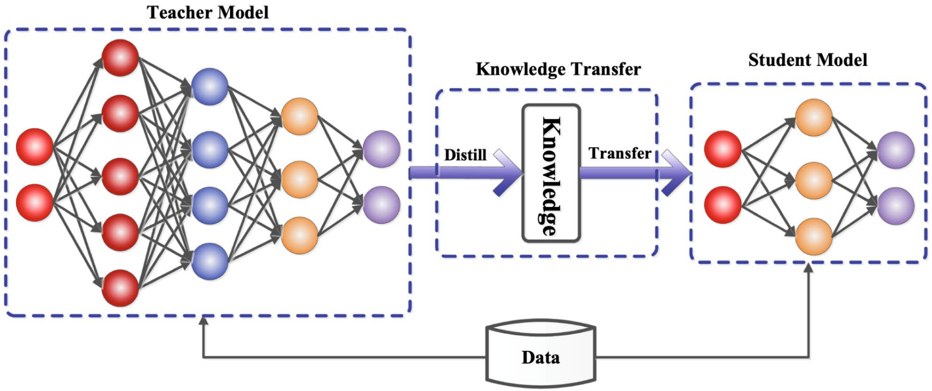
Mauricio Orbes
Research Scientist at OMHU (LEOiLAB)

Outline

- Introduction to KD
- Overview of methods
- Self Distillation.
- Why does it works?

Knowledge distillation Origin.

- Model Compression and acceleration (Hinton et al 2015),
- Learning effectively a small model (student) from a large model (teacher)
- Vanilla knowledge distillation learn the logits
- "Dark Knowledge" transfer.



Current methods.

1. Which Knowledge is being transferred?.

2. Which Training strategy?

Which Knowledge is being transferred?.

- Response-based Knowledge
- Feature-based Knowledge
- Relation-Based Knowledge.

Which Knowledge is being transferred.

- **Response based Knowledge**
- Feature-based Knowledge
- Relation-Based Knowledge.

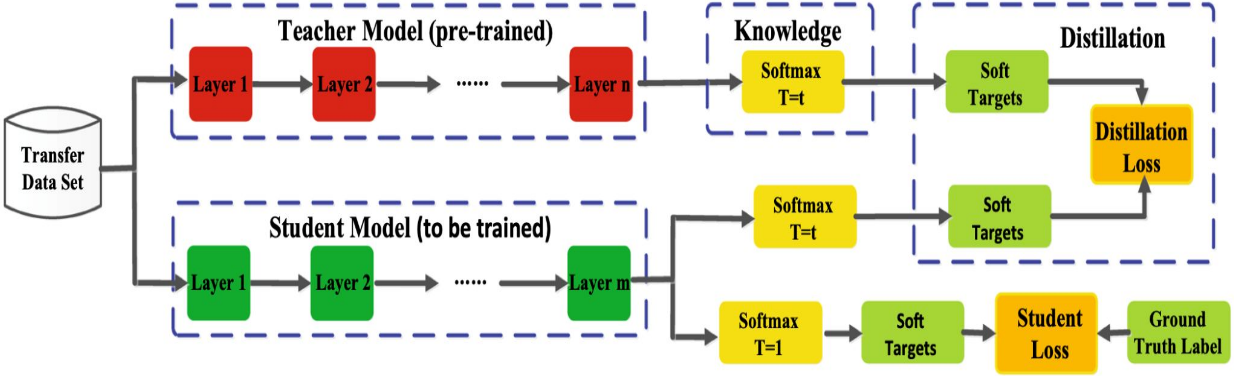


Figure from Gou et al 2020

Which Knowledge is being transferred.

- Response based Knowledge
- **Feature-based Knowledge**
- Relation-Based Knowledge.

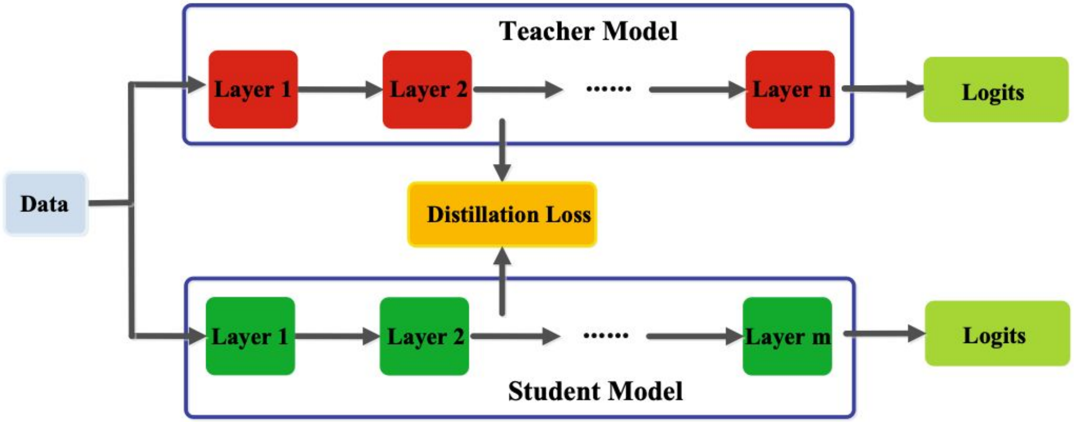


Figure from Gou et al 2020

Which Knowledge is being transferred.

- Response based Knowledge
- Feature-based Knowledge
- **Relation-Based Knowledge.**

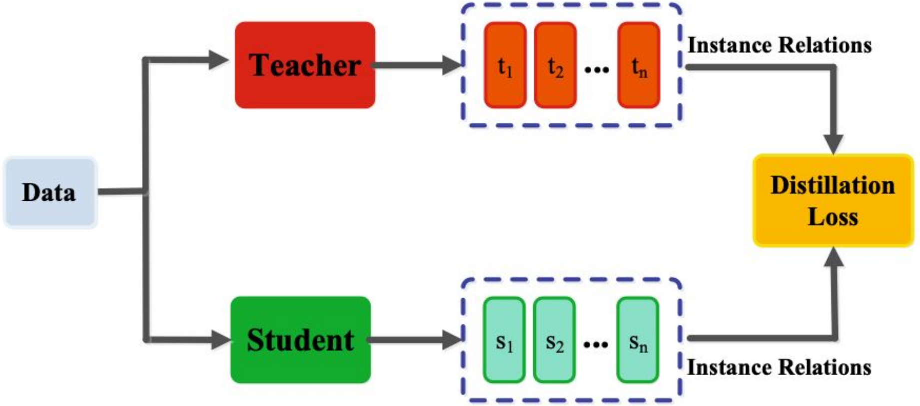


Figure from Gou et al 2020

Which training Strategy?

- Offline distillation
- Online distillation
- Self-distillation

Which training Strategy?

- **Offline distillation**
- Online distillation
- Self-distillation

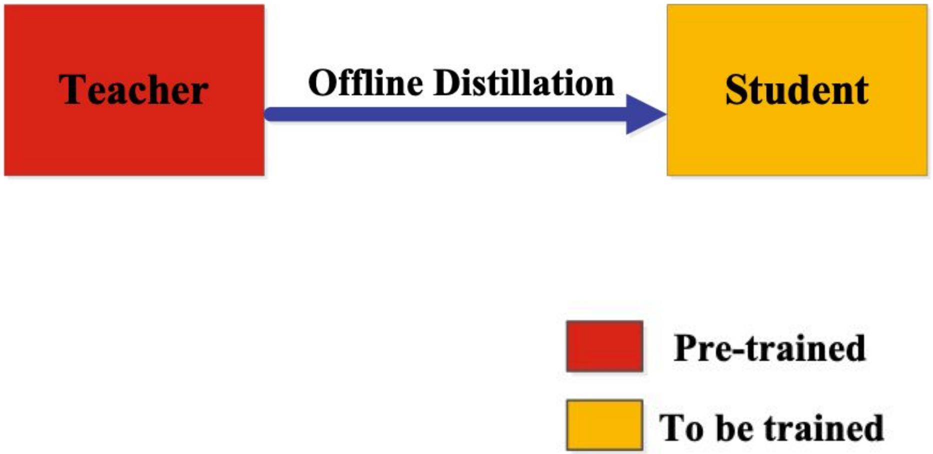


Figure from Gou et al 2020

Which training Strategy?

- Offline distillation
- **Online distillation**
- Self-distillation

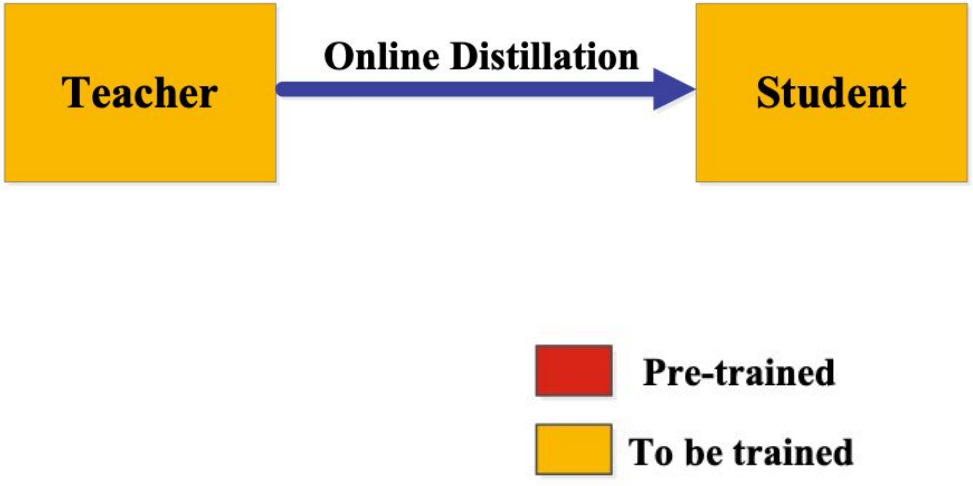


Figure from Gou et al 2020

Which training Strategy?

- Offline distillation
- Online distillation
- **Self-distillation**

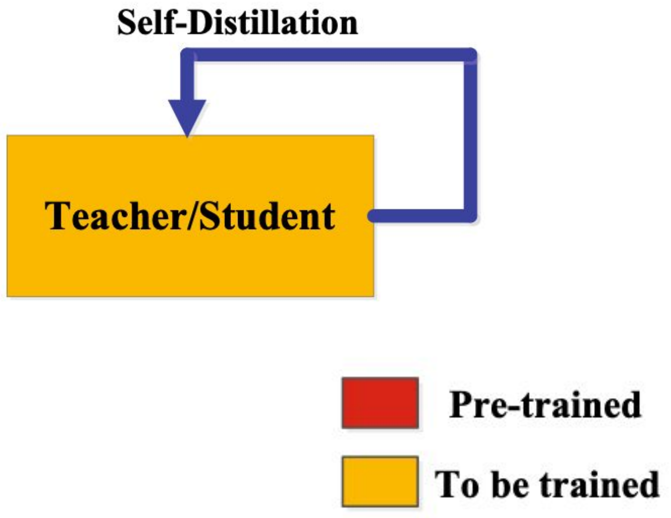
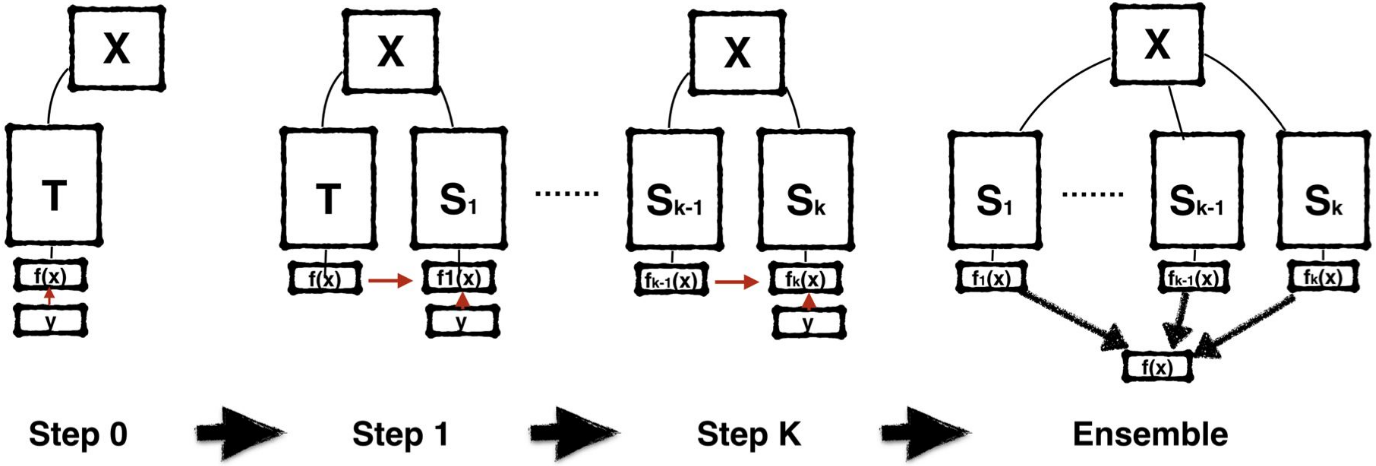


Figure from Gou et al 2020

Self - Distillation, Born Again Neural networks (BAN)



Furlanello et al 2018

Distillation Loss.

1. Traditional Supervised Training.

$$p(y|\boldsymbol{x};\boldsymbol{w}) = \mathrm{Cat}\left(\mathrm{softmax}\left(f_{\boldsymbol{w}}(\boldsymbol{x})\right)\right)$$

$$\mathcal{L}_{cce}(\boldsymbol{w}) = -\sum_{i=1}^{n}\sum_{j=1}^{k} \boldsymbol{y}_{ij} \log p(y = j|\boldsymbol{x}_i;\boldsymbol{w})$$

GT -Label

Sample

NN parameters

2. Distillation loss.

$$\mathcal{L}_{dist}(\boldsymbol{w}) = -\sum_{i=1}^{n}\sum_{j=1}^{k} \left[\mathrm{softmax}\left(f_{\boldsymbol{w}_t}(\boldsymbol{x})/T\right)\right]_j \log p(y = j|\boldsymbol{x}_i;\boldsymbol{w})$$

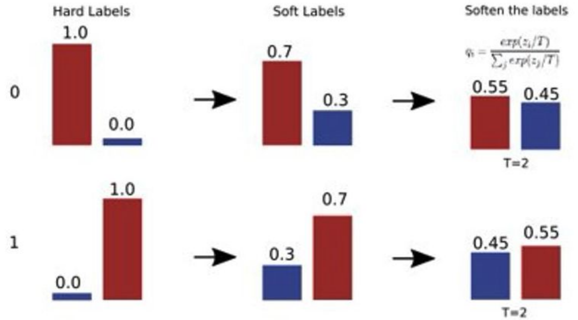
Teacher model (same architecture)

3. Total loss.

$$\mathcal{L}(\boldsymbol{w}) = \alpha\mathcal{L}_{cce}(\boldsymbol{w}) + (1-\alpha)\mathcal{L}_{dist}(\boldsymbol{w})$$

Why it works?

"Dark Knowledge"



1. Empirical Findings.
   a. Predictive uncertainty
   b. Confidence diversity
2. Theoretical Interpretation of teacher/student training.
   a. Instance-specific regularization (Via MAP formulation)
   b. Teacher predictions instance-specific priors conditioned in the inputs.
   c. Regularize predictive uncertainty + regularization on outputs lead to a better generalization.
3. A new method for label smoothing - Beta smoothing.

Empirical Findings

1. Predictive Uncertainty

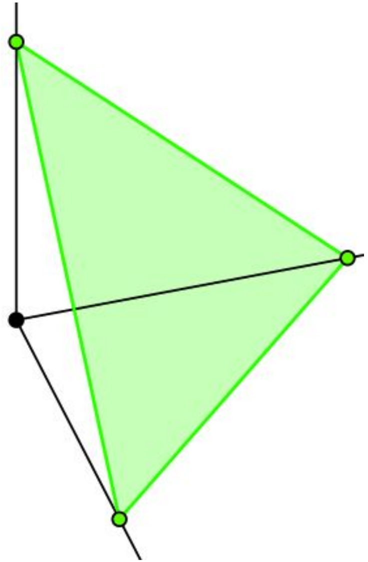
$$\mathbb{E}_{\boldsymbol{x}}\left[H\left(p(\cdot|\boldsymbol{x};\boldsymbol{w}_i)\right)\right] \approx \frac{1}{n}\sum_{j=1}^{n} H\left(p(\cdot|\boldsymbol{x}_j;\boldsymbol{w}_i)\right) = \frac{1}{n}\sum_{j=1}^{n}\sum_{c=1}^{k} -p(y_c|\boldsymbol{x}_j;\boldsymbol{w}_i)\log p(y_c|\boldsymbol{x}_j;\boldsymbol{w}_i).$$

2. Confidence Diversity (Amount of spread over the probability simplex)

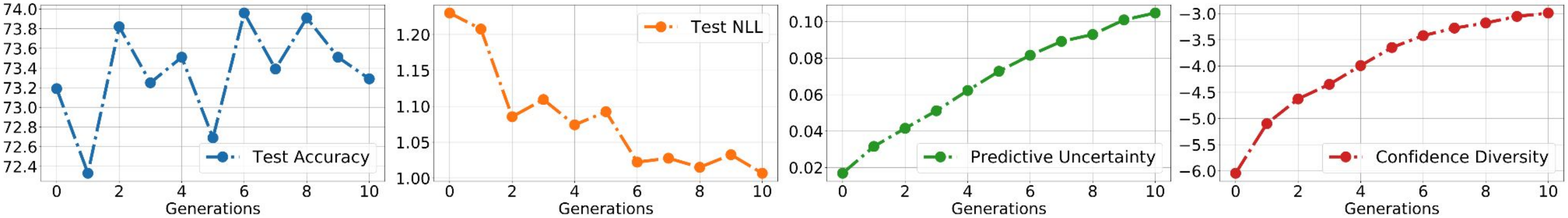
$$c = \phi(\boldsymbol{x}, y) := \left[\mathrm{softmax}\left(f_{\boldsymbol{w}}(\boldsymbol{x})\right)\right]_y$$

$$h(C) = -\int p_C(c)\log p_C(c)\, dc.$$

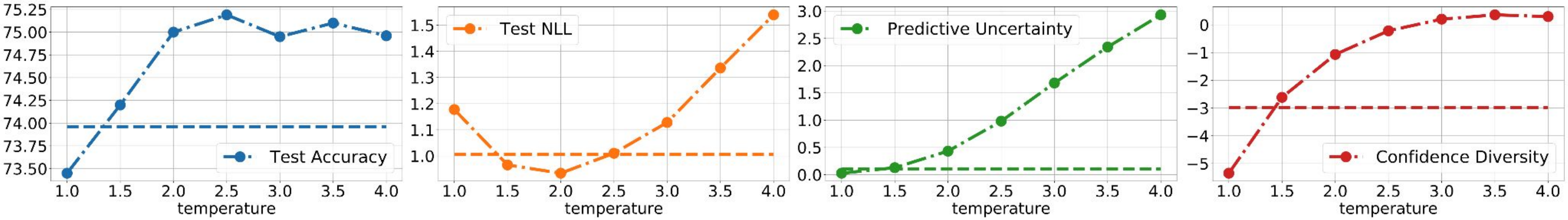
$$C := \phi(\boldsymbol{X}, Y) \text{ where } (\boldsymbol{X}, Y) \sim p(\boldsymbol{x}, y)$$



Born Again experiment.



Varying Temperature.



MAP Perspective of Self-Distillation

- Using a student network to amortize the MAP estimatior $\hat{\boldsymbol{z}}_i \approx \operatorname{softmax}(f_{\boldsymbol{w}}(\boldsymbol{x}_i))$

- MAP estimation on the softmax probability Vector (z).
- P(y|x) = cat(z), z ∈ Δ (L)
- P(z|x) = Dir(αx) instance specific parameter αx

- Close form solution:

$$\hat{\boldsymbol{z}}_i = \frac{\boldsymbol{c}_i + \boldsymbol{\alpha}_{\boldsymbol{x}\,i} - 1}{\sum_j \boldsymbol{c}_j + \boldsymbol{\alpha}_{\boldsymbol{x}\,j} - 1}$$

  - Training a student network

$$\max_{\boldsymbol{w}} \sum_{i=1}^{n} \log p(\boldsymbol{z}|\boldsymbol{x}_i, y_i; \boldsymbol{w}, \boldsymbol{\alpha}_{\boldsymbol{x}}) = \max_{\boldsymbol{w}} \sum_{i=1}^{n} \log p(y = y_i|\boldsymbol{z}, \boldsymbol{x}_i; \boldsymbol{w}) + \log p(\boldsymbol{z}|\boldsymbol{x}_i; \boldsymbol{w}, \boldsymbol{\alpha}_{\boldsymbol{x}})$$

$$= \max_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{n} \log[\operatorname{softmax}(f_{\boldsymbol{w}}(\boldsymbol{x}_i))]_{y_i}}_{\text{Cross entropy}} + \underbrace{\sum_{i=1}^{n} \sum_{c=1}^{k} ([\boldsymbol{\alpha}_{\boldsymbol{x}_i}]_c - 1) \log[\boldsymbol{z}]_c}_{\text{Instance-specific regularization}}$$

Label Smoothing as MAP.

- Assuming p(z|x) = P(z) and a uniform distribution across all possible labels.
- Choosing

$$[\boldsymbol{\alpha}_{\boldsymbol{x}}]_c = [\boldsymbol{\alpha}]_c = \frac{\beta}{k} + 1$$

$$\mathcal{L}_{LS} = \sum_{i=1}^{n} -\log[\boldsymbol{z}]_{y_i} + \beta \sum_{i=1}^{n} \sum_{c=1}^{k} -\frac{1}{k} \log[\boldsymbol{z}]_c$$

$$\begin{aligned}\max_{\boldsymbol{w}} \sum_{i=1}^{n} \log p(\boldsymbol{z}|\boldsymbol{x}_i, y_i; \boldsymbol{w}, \boldsymbol{\alpha}_{\boldsymbol{x}}) &= \max_{\boldsymbol{w}} \sum_{i=1}^{n} \log p(y = y_i|\boldsymbol{z}, \boldsymbol{x}_i; \boldsymbol{w}) + \log p(\boldsymbol{z}|\boldsymbol{x}_i; \boldsymbol{w}, \boldsymbol{\alpha}_{\boldsymbol{x}}) \\ &= \max_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{n} \log[\mathrm{softmax}\,(f_{\boldsymbol{w}}(\boldsymbol{x}_i))]_{y_i}}_{\text{Cross entropy}} + \underbrace{\sum_{i=1}^{n} \sum_{c=1}^{k} ([\boldsymbol{\alpha}_{\boldsymbol{x}_i}]_c - 1) \log[\boldsymbol{z}]_c}_{\text{Instance-specific regularization}}\end{aligned}$$

Self-Distillation as MAP

- Considering a teacher fwt trained by maximizing: $p(y|\boldsymbol{x}; \boldsymbol{w}_t) = \mathrm{Cat}(\mathrm{softmax}(f_{\boldsymbol{w}_t}(\boldsymbol{x})))$
- With $[\mathrm{softmax}\,(f_{\boldsymbol{w}_t}(\boldsymbol{x}))]_i = \frac{[\exp(f_{\boldsymbol{w}_t}(\boldsymbol{x}))]_i}{\sum_j [\exp(f_{\boldsymbol{w}_t}(\boldsymbol{x}))]_j}$

- $p(y|\boldsymbol{x}; \boldsymbol{\alpha}_{\boldsymbol{x}})$ Is a dirichlet-multimodal distribution (conjugacy of Dirichlet prior).
- Marginal likelihood reduces to categorical distribution: $p(y|\boldsymbol{x}; \boldsymbol{\alpha}_{\boldsymbol{x}}) = \mathrm{Cat}(\overline{\boldsymbol{\alpha}_{\boldsymbol{x}}})$
- $[\overline{\boldsymbol{\alpha}_{\boldsymbol{x}}}]_i = \frac{[\boldsymbol{\alpha}_{\boldsymbol{x}}]_i}{\sum_j [\boldsymbol{\alpha}_{\boldsymbol{x}}]_j}$
-

$$\boldsymbol{\alpha}_{\boldsymbol{x}} = \beta \exp(f_{\boldsymbol{w}_t}(\boldsymbol{x})/T) + \gamma$$

$$\boldsymbol{\alpha}_{\boldsymbol{x}} = \beta \exp(f_{\boldsymbol{w}_t}(\boldsymbol{x})/T) + 1 = \beta \sum_j [\exp(f_{\boldsymbol{w}_t}(\boldsymbol{x})/T)]_j \,\mathrm{softmax}(f_{\boldsymbol{w}_t}(\boldsymbol{x})/T) + 1.$$

$$\mathcal{L}_{SD} = \sum_{i=1}^{n} -\log[\boldsymbol{z}]_{y_i} + \beta \sum_{i=1}^{n} \omega_{\boldsymbol{x}_i} \sum_{c=1}^{k} -[\mathrm{softmax}(f_{\boldsymbol{w}_t}(\boldsymbol{x}_i)/T)]_c \log[\boldsymbol{z}]_c,$$

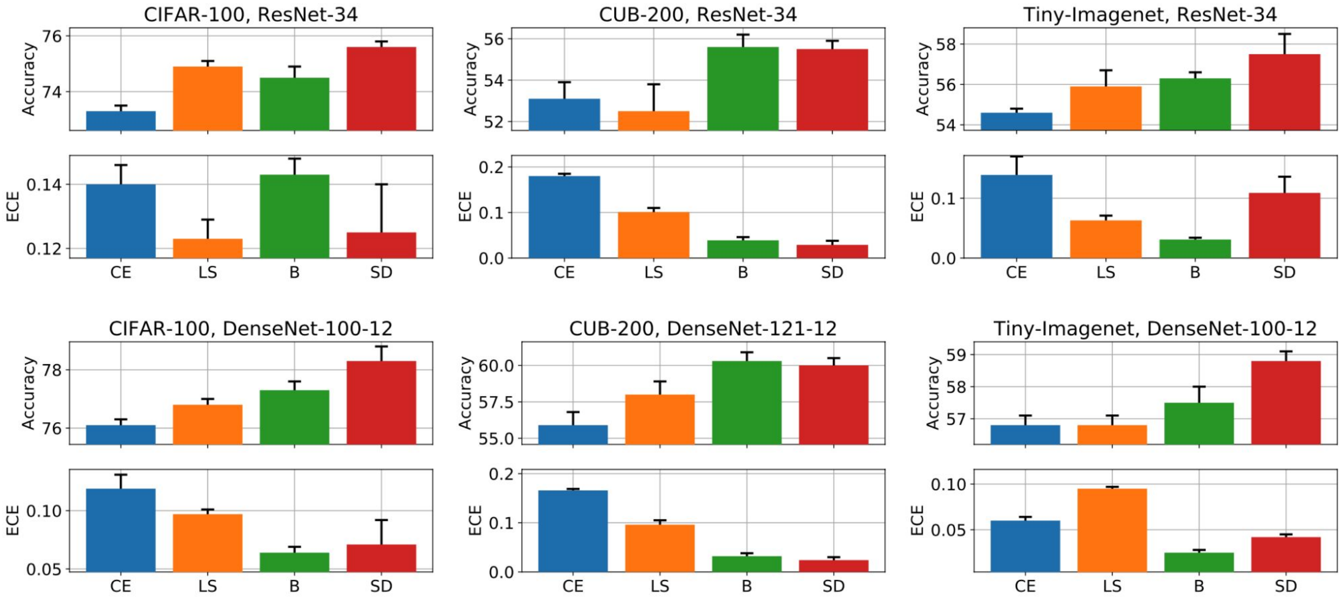
$$\omega_{\boldsymbol{x}_i} = \sum_j [\exp(f_{\boldsymbol{w}_t}(\boldsymbol{x}_i)/T)]_j\,!$$

Beta Smoothing labels.

- Amount of smoothing proportional to the uncertainty of predictions (Exponential Moving average -EMA)
- Ranking $\{b_1 \leq ... \leq b_m\}$ from Beta(a,1), m is the batch size, a is an hyperparameter of beta distribution.
-

$$[\boldsymbol{\alpha}_{\boldsymbol{x}_i}]_{y_i} = \beta b_i + 1 \text{ and } [\boldsymbol{\alpha}_{\boldsymbol{x}_i}]_c = \beta \tfrac{1-b_i}{k-1} + 1 \text{ for all } c \neq y_i$$

Experimental results.



Questions.