BORN AGAIN NEURAL NETWORKS

TOMMASO FURLANELLO, ZACHARY C. LIPTON, MICHAEL TSCHANNEN, LAURENT ITTI, AND ANIMA ANANKDUMAR

furlanel@usc.edu, zlipton@cmu, michaelt@nari.ee.ethz.ch, itti@usc.edu, anima@amazon.com

CONTRIBUTION

We present a simple re-training procedure between teacher and students. We report improvement of the student validation error across multiple datasets and architecture classes reaching sota on Cifar-100.

To experimetally identify the source of these gains we propose two distillation objectives :

1. Confidence-Weighted by Teacher Max (CWTM)
2. Dark Knowledge with Permuted Predictions (DKPP)

BORN AGAIN MODELS

The single-sample gradient of the cross-entropy between student logits z_j and target logits t_j with respect to the ith output is given by:

$$\frac{\partial \mathcal{L}_i}{\partial z_i} = q_i - p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} - \frac{e^{t_i}}{\sum_{j=1}^{n} e^{t_j}}. \qquad (3)$$

When the target probability distribution function corresponds to the ground truth $*$ one-hot label $p_* = y_* = 1$ this reduces to:

$$\frac{\partial \mathcal{L}_*}{\partial z_*} = q_* - y_* = \frac{e^{z_*}}{\sum_{j=1}^{n} e^{z_j}} - 1 \qquad (4)$$

In Knowledge Distillation (KD) the loss is:

$$\sum_{s=1}^{b}(q_{*,s} - p_{*,s}) + \sum_{s=1}^{b}\sum_{i=1}^{n-1}(q_{i,s} - p_{i,s}), \qquad (5)$$

The second term corresponds to the information incoming from all the wrong outputs, i.e. **dark knowledge**. The first term corresponds to the gradient from the correct choice and can be written as

$$\frac{1}{b}\sum_{s=1}^{b}(q_{*,s} - p_{*,s} y_{*,s}) \qquad (6)$$

In Eq. (6) the teacher prediction $p_{*,s}$ can be interpreted as a scaling factor of the ground truth gradient of (4).
In importance weighting of samples the gradient of each sample in a mini-batch is balanced based on its importance weight w_s.]:

$$\sum_{s=1}^{b} \frac{w_s}{\sum_{u=1}^{b} w_u}(q_{*,s} - y_{*,s}) \qquad (7)$$

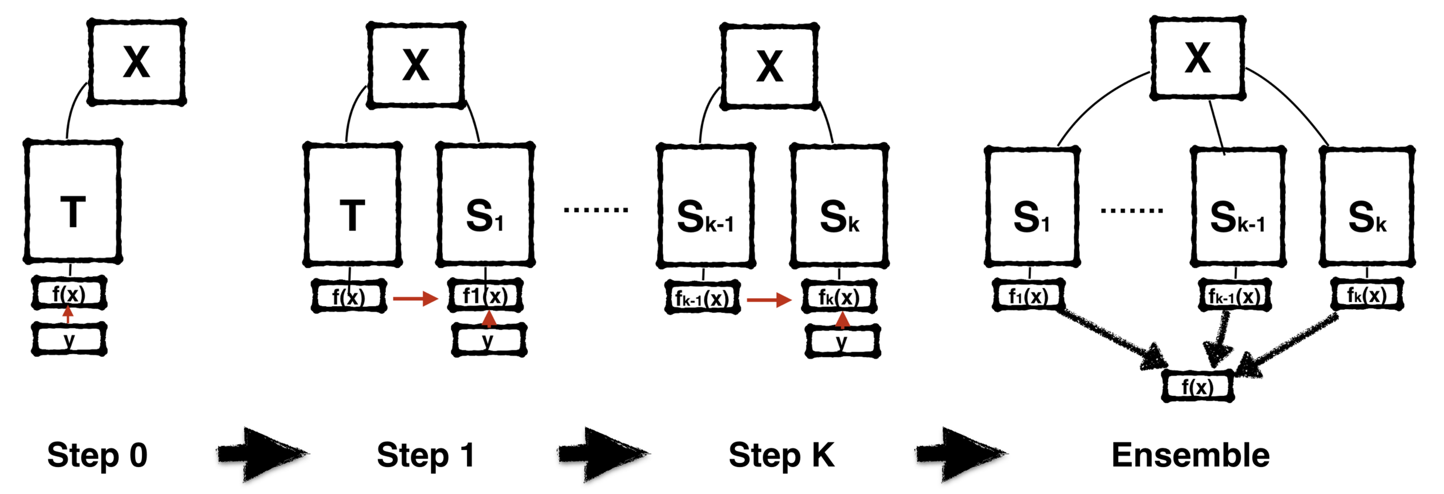
When the importance weights correspond to the output of a teacher for the correct dimension

$$\sum_{s=1}^{b} \frac{p_{*,s}}{\sum_{u=1}^{b} p_{*,u}}(q_{*,s} - y_{*,s}). \qquad (8)$$

SEQUENCE OF IDENTICAL TEACHING SELVES

K-steps Born Again Neural Network

$$\min_{\theta_k} \mathcal{L}(y, f(x,\theta_k)) + \mathcal{L}(f(x, \arg\min_{\theta_{k-1}} \mathcal{L}(y, f(x,\theta_{k-1}))), f(x,\theta_k))$$



Cifar-100	Teacher	BAN	BAN+L	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNet-112-33	18.25	**16.95**	17.68	17.61	17.22	**16.59**	15.77	15.68
DenseNet-90-60	17.69	**16.69**	16.93	16.62	**16.44**	16.72	15.39	15.74
DenseNet-80-80	17.16	**16.36**	16.5	16.26	16.30	**15.5**	15.46	15.14
DenseNet-80-120	16.87	**16.00**	16.41	**16.13**	16.13	/	**15.13**	**14.9**

Penn-Tree Bank	Parameters	Teacher Val	BAN+L Val	Teacher Test	BAN+L Test
ConvLSTM	19M	83.69	80.27	80.05	**76.97**
LSTM	52M	75.11	71.19	71.87	**68.56**

CWTM & DKPP

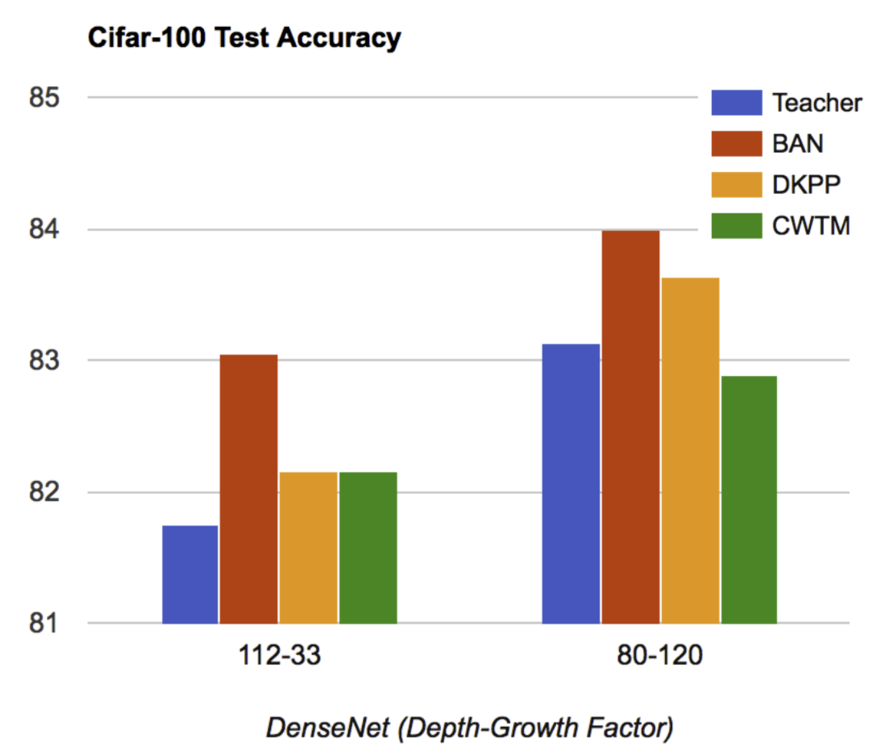
In Confidence Weighted by Teacher Max (CWTM), we weight each example in the student's loss function by the confidence of the teacher model on that example:

$$\sum_{s=1}^{b} \frac{\max p_{.,s}}{\sum_{u=1}^{b} \max p_{.,u}}(q_{*,s} - y_{*,s}). \qquad (1)$$

In dark knowledge with Permuted Predictions (DKPP), **we permute the non-argmax outputs of the teacher's predicted distribution**.

$$\sum_{s=1}^{b}\sum_{i=1}^{n} \frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \sum_{s=1}^{b}(q_{*,s} - \max p_{.,s}) + \sum_{s=1}^{b}\sum_{i=1}^{n-1} q_{i,s} - \phi(p_{j,s}), \qquad (2)$$

where $\phi(p_{j,s})$ are the permuted outputs of the teacher.

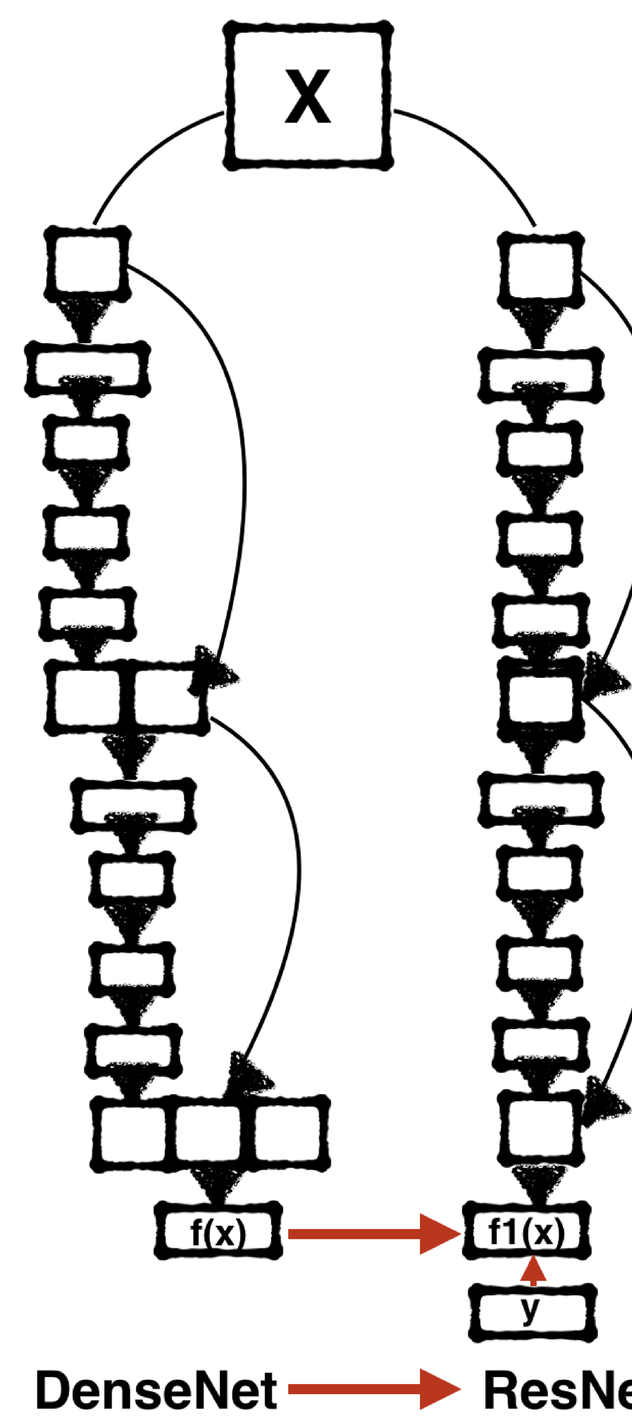


Network	CWTM	DKPP
DenseNet-112-33	**17.84**	**17.84**
DenseNet-90-60	**17.42**	**17.43**
DenseNet-80-80	17.16	**16.84**
DenseNet-80-120	17.12	**16.34**

BAN-RESNET

DenseNet $\mapsto$ Resnet

DenseNet 90-60	Parameters	Baseline	BAN
ResNet-1001	10.2 M	22.71	/
ResNet-14-0.5	7.3 M	20.28	18.8
ResNet-14-1	17.7 M	18.84	**17.39**
Wide-ResNet-1-1	20.9 M	20.4	19.12
Match-Wide-ResNet-2-1	43.1 M	18.83	17.42
Wide-ResNet-4-0.5	24.3 M	19.63	**17.13**
Wide-ResNet-4-1	87.3 M	18.77	17.18



Resnet $\mapsto$ DenseNet

Cifar100	Teacher	BAN	Dense-90-60
Wide-ResNet-28-1	30.05	29.43	24.93
Wide-ResNet-28-2	25.32	24.38	18.49
Wide-ResNet-28-5	20.88	20.93	17.52
Wide-ResNet-28-10	19.08	18.25	16.79

REFERENCES

[1] Breiman, Leo, and Nong Shang. "Born again trees."