Born Again Neural Networks

Published @**ICML 2018**

Furlanello, T., Lipton, Z.C.,
Tschannen, M., Itti, L.
Anandkumar, A

Outline

- Intro to Distillation
- Born Again Networks (BANs)
- Where is the novelty in the BANs paper?
- BAN improvements
- Discussion

“Model compression”

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. **(2006)**

- Pre Deep Learning era – **Ensembles** of naive classifiers

- Main idea: Model **compression** (for semi-supervised learning)

- Idea: Instead of using an ensemble, use the ensemble to created synthetic labeling and use that to train one **single Neural Network.**

- The **Teachers** will always perform **better**.

“Do deep nets really need to be deep?”

Ba, J., & Caruana, R. **(2014)**

- Idea 1: DNNs can be approximated by shallow networks

- Idea 2: Do not match the outputs, match the **logits**:

$$\mathcal{L} = \|z_S - z_T\|_2^2$$

  - Why?
    - There is **information** loss when passing from **logits** to probability space.
    - The information loss is the **relationship** learned by the teacher model across all of the **targets**.
- Insights:
  - “model compression” is a form of **regularization**
  - More powerful Teacher → Better student

“Distilling the knowledge in a neural network”

Hinton, G., Vinyals, O., & Dean, J. **(2015)**

- Idea: raise the **temperature** of the final **SoftMax** until the cumbersome model produces a suitably **soft set of targets**.

$$p_i = \frac{\exp(z_i/t)}{\sum_j \exp(z_j/t)}$$

- Matching logits is a special case of distillation
- Use both the true labels and the Teachers output
  - Captures not only the information provided by the true labels
  - Emulates the internal structure that the complex teacher has learned
  - $\mathcal{L} = H_{T=1}(\mathbf{y}, P_S) + \lambda \cdot H_{T>1}\,(P_{\mathrm{T}}, P_{\mathrm{S}})$

What have we learned so far about Distillation?

- Is **not** a compression technique
- Distillation is a **regularization** method
  - label augmentation
- **Hard labeling** is wrong – **Logits** are sub-optimal
- Better Teacher -> Better Student

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- Idea: Students parameterized **identically** to their Teachers
  - Born-Again Networks
  - The Students **outperform** the Teachers

- What is the **effect** of the teacher outputs?:
  - Confidence-Weighted by Teacher Max (CWTM)
  - Dark Knowledge with Permuted Predictions (DKPP)

- Experiments with:
  - DenseNets, ResNets
  - LSTM-based sequence models

Born-Again Networks

"Born again neural networks"

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- They distilled:
  - DenseNets to ResNets
  - ResNets to DenseNets

This is literally the concept of Distillation for compression

- Distillation in multiple **generations**:
  - the k-th model is trained, with knowledge transferred from the (k-1)-th student
    - ResNets
    - DenseNets
    - LSTMs

- Born-Again Network Ensembles (BANE)

Born-Again Networks

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- Results:
  - **Simple KD** with cross entropy is the best way to go
  - **Only KD without ground truth is better**
    - – This is not true for LSTM models
  - The Students **outperform** the Teachers in almost every experiment
    - – This should not be a surprise
      - Hint-Based Learning (see later)

Born-Again Networks

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- ResNets perform better than DenseNets

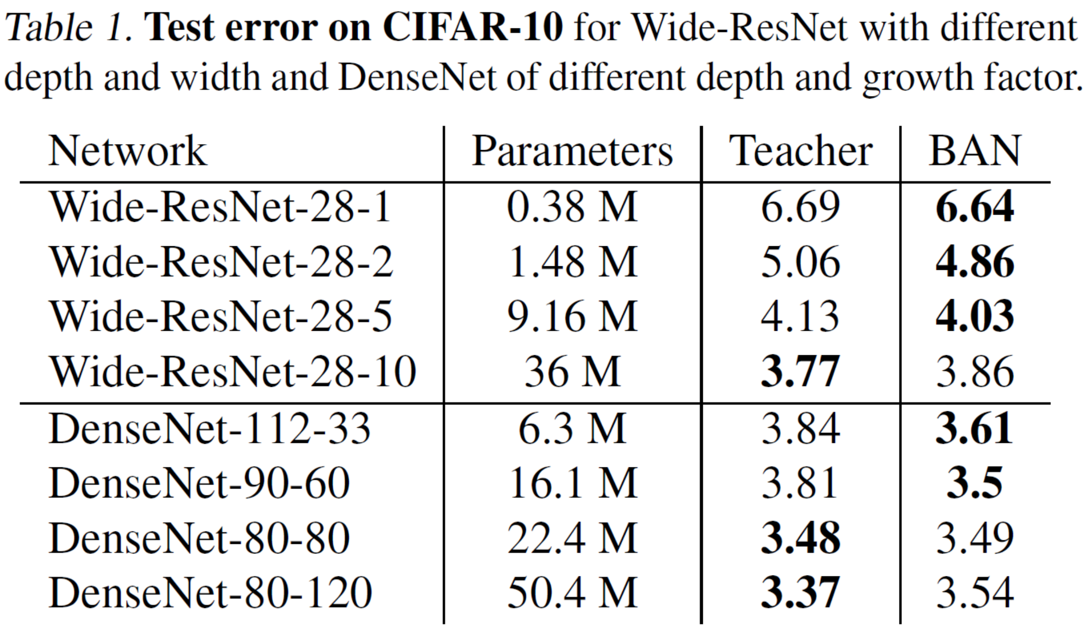


Table 1. **Test error on CIFAR-10** for Wide-ResNet with different depth and width and DenseNet of different depth and growth factor.

Network	Parameters	Teacher	BAN
Wide-ResNet-28-1	0.38 M	6.69	**6.64**
Wide-ResNet-28-2	1.48 M	5.06	**4.86**
Wide-ResNet-28-5	9.16 M	4.13	**4.03**
Wide-ResNet-28-10	36 M	**3.77**	3.86
DenseNet-112-33	6.3 M	3.84	**3.61**
DenseNet-90-60	16.1 M	3.81	**3.5**
DenseNet-80-80	22.4 M	**3.48**	3.49
DenseNet-80-120	50.4 M	**3.37**	3.54

Born-Again Networks

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- DenseNets perform better than ResNets

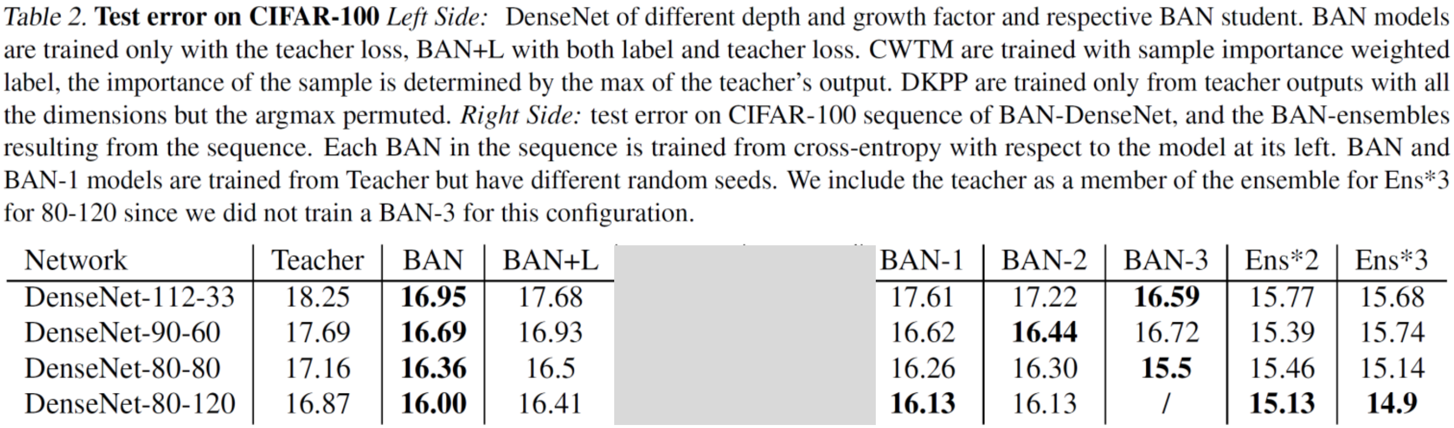


Table 2. **Test error on CIFAR-100** *Left Side:* DenseNet of different depth and growth factor and respective BAN student. BAN models are trained only with the teacher loss, BAN+L with both label and teacher loss. CWTM are trained with sample importance weighted label, the importance of the sample is determined by the max of the teacher's output. DKPP are trained only from teacher outputs with all the dimensions but the argmax permuted. *Right Side:* test error on CIFAR-100 sequence of BAN-DenseNet, and the BAN-ensembles resulting from the sequence. Each BAN in the sequence is trained from cross-entropy with respect to the model at its left. BAN and BAN-1 models are trained from Teacher but have different random seeds. We include the teacher as a member of the ensemble for Ens*3 for 80-120 since we did not train a BAN-3 for this configuration.

Network	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**

Born-Again Networks

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- ConvLSTM perform better than LSTM
  - LSTM = 1 single layer LSTM with 1500 units
  - ConvLSTM = convolutional layers, highway layers, and a 2-layer LSTM

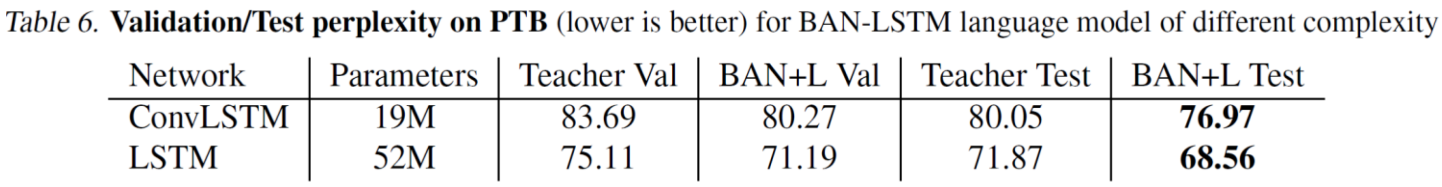


Table 6. **Validation/Test perplexity on PTB** (lower is better) for BAN-LSTM language model of different complexity

Network	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**

WOW! BANs!!!! Oh wait….

In the BAN paper they mention:
“A gift from knowledge distillation: Fast optimization, network minimization and transfer learning” (2017)

Yim, J., Joo, D., Bae, J., & Kim, J**.**

Lets see some other works:

1. “Fitnets: Hints for thin deep nets” (2014)
   Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y.

2. “Temporal ensembling for semi-supervised learning” (2016)
   Laine, S., & Aila, T.

3. “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results” (2018)
   Tarvainen, A., & Valpola, H.

4. “Deep mutual learning” (2017)
   Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H.

“Fitnets: Hints for thin deep nets”

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. **(2014)**

- Idea: **Deep** networks generalize **better**
- Distillation from a big network to a thin and deep
- **Thin** and **deep** nets are **hard** to train
  - Solution:
    - **pretrain** half of the Student with hints from the Teacher’s middle layers
    - Annealing λ

- **1st time that the Student outperformed the Teacher!!!**

“Temporal ensembling for semi-supervised learning”

Laine, S., & Aila, T. **(2016)**

- Idea: Use self-ensembling for semi-supervised tasks
- The **Π-model**:
  - Different augmentations and dropout for each epoch
  - The loss has 2 components:
    - Standard cross entropy (only labeled examples)
    - L2 for distillation for all the samples
  - The unsupervised loss weighting function ramps up along a Gaussian curve
- **Temporal Ensembling**:
  - Evaluate the network only once and keep a **moving average of the labels** $Z_i \leftarrow \alpha Z_i + (1-\alpha) z_i$

AGAIN the Student outperforms the Teacher!!!

“Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”

Laine, S., & Aila, T. **(2018)**

- Idea:
  Temporal Ensembling becomes memory demanding when learning large datasets

- Solution:
- Don’t average the labels, **average the weights**

- Results:
  Better test accuracy

AGAIN the Student outperforms the Teacher!!!

“Deep mutual learning”

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. **(2017)**

- Idea: Train a **pool** of Students that act as Teachers
  - Similarities to Temporal Ensembling
- Implementation:
  - Train all the models **from scratch** and let them distilled to each other
- The models of the pool **outperform** the powerful static **Teachers**!

Still trying to find the novelty in BANs…

- Compression ? ✗
- Distillation with same networks? ✗
- Why distillation works? ✓

CWTM - DKPP

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- What distillation provides to the Student?
  - **Hinton** et al. (2006) information on the **wrong outputs**
  - **Importance-weighting** of the real labels (teacher’s confidence in the correct prediction)

- KD gradient decomposition
  - Aim: Quantifying the contribution of each term to KD

KD gradient decomposition

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

The **single-sample gradient** of the cross-entropy

Between the Student and the Teacher is:

$$\frac{\partial \mathcal{L}_i}{\partial z_i} = q_i - p_i = \frac{e^{z_i}}{\sum\limits_{j=1}^{n} e^{z_j}} - \frac{e^{t_i}}{\sum\limits_{j=1}^{n} e^{t_j}}$$

Student Teacher

KD gradient decomposition

"Born again neural networks"

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

Across all the b samples s of the mini-batch:

Information incoming from all the **wrong outputs** (Hinton et al. (2015) hypothesis)

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

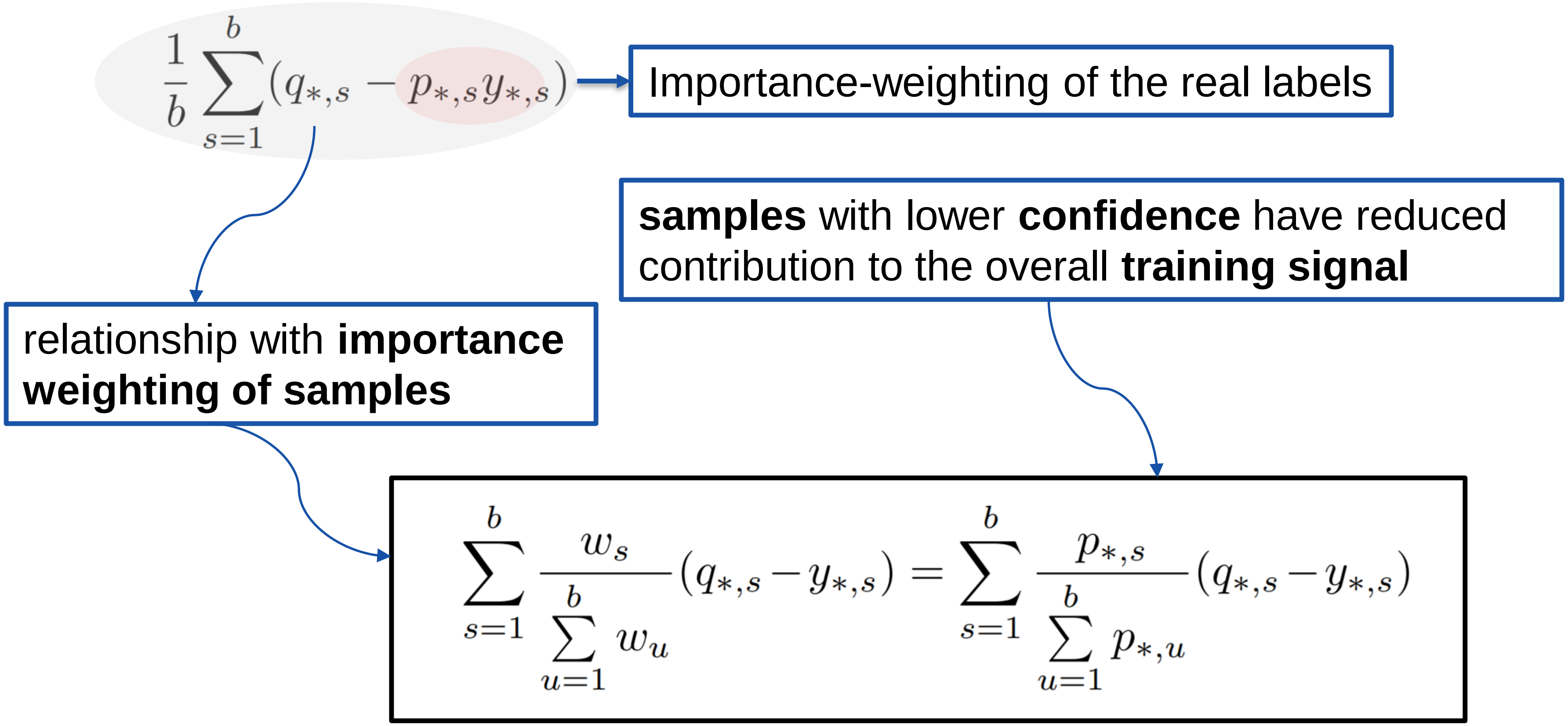
$$\frac{1}{b}\sum_{s=1}^{b}(q_{*,s}-p_{*,s}y_{*,s})$$

Importance-weighting of the real labels

KD gradient decomposition

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**



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

Importance-weighting of the real labels

samples with lower **confidence** have reduced contribution to the overall **training signal**

relationship with **importance weighting of samples**

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

Confidence-Weighted by Teacher Max (CWTM)

"Born again neural networks"

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- Is dark knowledge performing a kind of **importance weighting**?
  - Experimental procedure:

    Weight each example in the student's loss function
    by the **confidence of the teacher** model on that example

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

Dark Knowledge with Permuted Predictions (DKPP)

"Born again neural networks"

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- Does the success of dark knowledge owe to the information contained in the non argmax outputs of the teacher?
  - i.e. Was **Hinton** et al. (2014) correct?

- Experimental procedure:

  Permute the non-argmax outputs of the teacher's predicted distribution to **destroy** the pairwise similarities of the original output covariance matrix

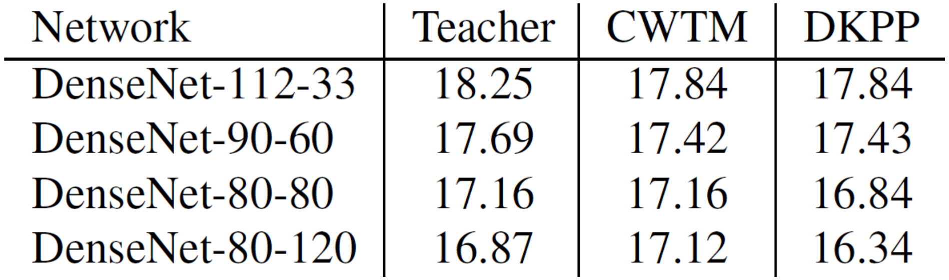
$$\sum_{s=1}^{b}\sum_{i=1}^{n}\frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \sum_{s=1}^{b}(q_{*,s} - \max p_{.,s}) + \sum_{s=1}^{b}\sum_{i=1}^{n-1} q_{i,s} - \phi(p_{j,s})$$

CWTM - DKPP

“Born again neural networks”

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. **(2018)**

- Results:
  - **CWTM** leads to **weak** improvements
  - **DKPP** leads to systematic **improvement**



Network	Teacher	CWTM	DKPP
DenseNet-112-33	18.25	17.84	17.84
DenseNet-90-60	17.69	17.42	17.43
DenseNet-80-80	17.16	17.16	16.84
DenseNet-80-120	16.87	17.12	16.34

BAN improvements

- "Knowledge Distillation in Generations: More Tolerant Teachers Educate Better Students" (2018)

  Yang, C., Xie, L., Qiao, S., & Yuille, A.

  - Models should preserve secondary information so that the students become stronger
  - Compute the gap between the confidence scores of the primary class and other K-1
  - Control the secondary information

We went so far with KD… Did we missed something in the process?

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Now let's talk about Distillation...

Thank you