Born Again Neural Networks

Radek Bartyzal

Let's talk ML in Prague

21. 2. 2018

Prior work

Ensembles

Diverse models with similar validation performances can be often be combined to achieve predictive power superior to each of the constituent models. [3]

Born again trees

Learn a single tree that is able to recover the performance of a multiple-tree predictor. [4]

Knowledge distillation = model compression

Transfer knowledge acquired by a learned teacher model to a new simpler student model. [5]

Knowledge distillation

Teacher

- high-capacity model
- good performance

Student

- more compact model
- not as good performance as the teacher but better than if it was trained without it

By transferring knowledge, one hopes to benefit from the student's compactness while suffering only minimal degradation in performance.

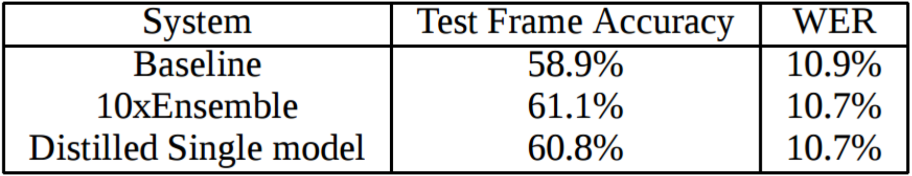
Knowledge distillation

Teacher produces soft targets = probabilities of incorrect classes = the key to generalization outside of the training dataset.

Training student = minimize weighted average of:

- cross entropy with the soft targets
- cross entropy with the hard targets = labels

Knowledge distillation results



System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Figure: DNN acoustic models used in Automatic Speech Recognition. Distilled model trained by an ensemble of models performs better than the baseline. [5]

Born Again Networks (BANs)

- not compressing models
- students are parameterized identically to their parents
- students outperform teachers
- knowledge transfer between dense networks and residual networks of similar capacity

$$\min_{\theta_2} \mathcal{L}(y, f(x, \theta_2)) + \mathcal{L}(f(x, \arg\min_{\theta_1} \mathcal{L}(y, f(x, \theta_1))), f(x, \theta_2))$$

Figure: BAN loss function adding Kullback–Leibler divergence between the new model's outputs and the outputs of the original model. [1]

BAN Ensembles

Apply BANs sequentially with multiple generations of knowledge transfer. In each case, the k-th model is trained, with knowledge transferred from the $k-1$-th student:

$$\min_{\theta_k} \mathcal{L}(y, f(x, \theta_k)) + \mathcal{L}(f(x, \arg\min_{\theta_{k-1}} \mathcal{L}(y, f(x, \theta_{k\_1}))), f(x, \theta_k))$$

Born Again Network Ensemble (BANE)

Averaging the prediction of multiple generations of BANs.

Dense Convolutional Network (DenseNet)

- For each layer, the feature-maps of all preceding layers are used as inputs.
- The received feature maps are concatenated not added like in ResNet.
- Growth rate (k) = number of feature maps per layer in the dense block.
- Last layer of block has $k_0 + k \times (L - 1)$ feature maps.

Advantages:

- alleviate the vanishing-gradient problem
- reduce the number of parameters - hmm how?
- dense connections have a regularizing effect = reduce overfitting
- promote feature reuse

DenseNet: Dense block

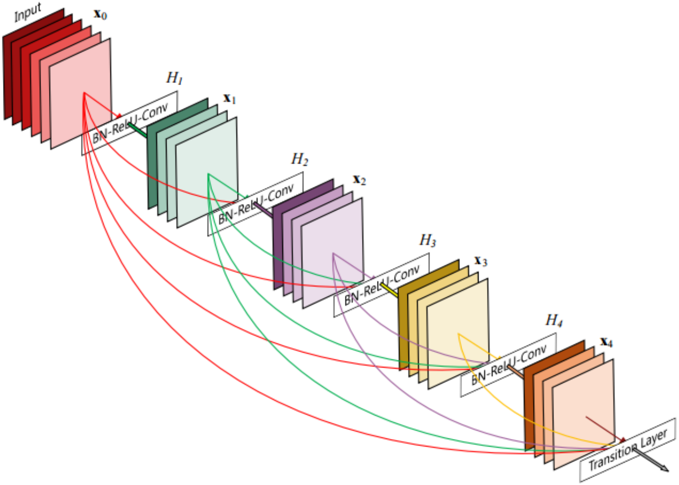


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

DenseNets reducing number of parameters

In classic Feed Forward CNN each layer:

- receives information (= state) from the previous layer
- modifies the state
- keeps the information that needs to be passed on
- sends the whole state to the next layer

In DenseNet each layer:

- explicitly differentiates between information that is added to the network and information that is preserved
- is very narrow (e.g. 12 filters)
- receives all the previous feature-maps unchanged
- adds small set of feature-maps to the "collective knowledge"

Deep DenseNet

- Feature map dimensions must stay the same in one dense block.
- Pooling is an important part of CNNs.
- Put pooling between the dense blocks.

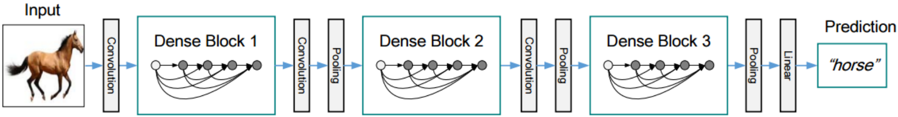
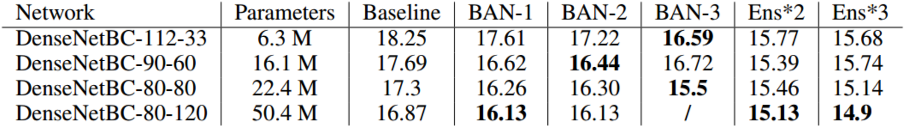


Figure: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling. [6]

BAN DenseNets Experiments

All experiments on: CIFAR-100 (100 classes each containing 600 32x32 colour images).

- DenseNet-BC-(depth)-(growth rate)
- BAN-1/2/3 = sequential training by previous BAN-(k-1)
- Ens*2/3 = ensembles of 2/3 BAN-x



Network	Parameters	Baseline	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNetBC-112-33	6.3 M	18.25	17.61	17.22	**16.59**	15.77	15.68
DenseNetBC-90-60	16.1 M	17.69	16.62	**16.44**	16.72	15.39	15.74
DenseNetBC-80-80	22.4 M	17.3	16.26	16.30	**15.5**	15.46	15.14
DenseNetBC-80-120	50.4 M	16.87	**16.13**	16.13	/	**15.13**	**14.9**

Figure: BAN training is clearly beneficial for DenseNets on CIFAR. [1]

Residual Networks (ResNet)

- Add skip-connection that bypasses the non-linear transformations with an identity function.
- Identity function and the output of previous layer are combined by summation, which may impede the information flow in the network.
- "Thin and deep" = small number of filters, 1000+ layers

Diminishing feature reuse

Gradient flowing through skip connections is not forced to go through residual block weights $\Longrightarrow$

- few blocks learn useful representations
- many blocks share very little information with small contribution to the final goal

ResNet: Residual block

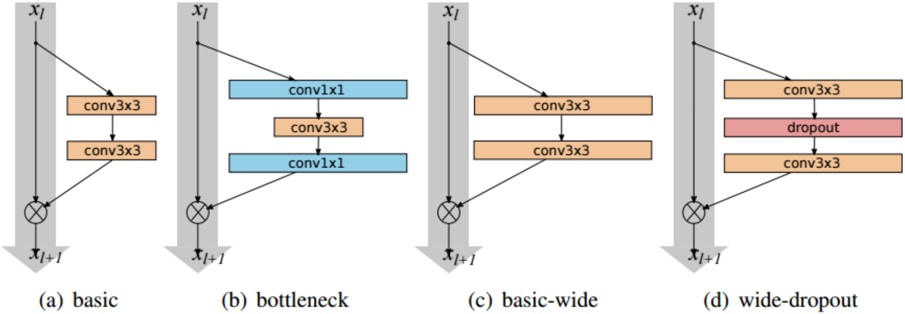


Figure: Various residual blocks used in the paper. Batch normalization and ReLU precede each convolution (omitted for clarity). [7]

Wide ResNet

Multiply the number of filters in each convolution layer:

- basic $= 3 \times 3$, 16 filters
- wide $= 3 \times 3$, 16 filters $\times$ k

And reduce the number of layers to keep the number of parameters the same.

- wide 16-layer deep network has the same accuracy as a 1000-layer thin deep network and a comparable number of parameters, although being several times faster to train
- wide networks are trained faster because the added convolution filters are easily parallelized on GPUs

ResNet vs DenseNet

BAN ResNets

BAN-ResNets:

- trained by DenseNet 90-60 teacher
- student always shares the first and last layer with its teacher
- dense blocks replaced by residual blocks
- baseline = wide-ResNet28 [7]
- tested multiple nets with different number of units per block
- all benefit from BAN training

BAN-ResNets outperform:

- traditional counterparts
- equivalent ResNets trained without DenseNet teacher
- their DenseNet teacher

BAN Results

Single model non-ensemble SOTA on CIFAR 100 trained with SGD without any sort of shake-shake regularization:

- BAN-3-DenseNet-80-80
- 22M parameters
- 15.5% error

Ensemble SOTA under the same conditions:

- BAN-3-DenseNet-BC-80-120
- 150M parameters
- 14.9% error

Sources

1. Tommaso Furlanello et al. ”Born Again Neural Networks.” Workshop on Meta-Learning (MetaLearn 2017) at NIPS. Accessible from: http://metalearning.ml/papers/metalearn17_furlanello.pdf

2. Breiman, Leo. ”Statistical modeling: The two cultures (with comments and a rejoinder by the author).” Statistical science 16.3 (2001): 199-231. Accessible from: https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726%20

3. Hansen, Lars Kai, and Peter Salamon. ”Neural network ensembles.” IEEE transactions on pattern analysis and machine intelligence 12.10 (1990): 993-1001.

Sources

4. Breiman, Leo, and Nong Shang. "Born again trees." ps (1996). Accessible from: https://www.stat.berkeley.edu/~breiman/BAtrees.pdf

5. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015). Accessible from: https://arxiv.org/pdf/1503.02531.pdf

6. Huang, Gao, et al. "Densely connected convolutional networks." arXiv preprint arXiv:1608.06993 (2016). Accessible from: https://arxiv.org/abs/1608.06993

7. Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016). Accessible from: https://arxiv.org/abs/1605.07146