Born Again Neural Networks

Radek Bartyzal

Let's talk ML in Prague

Date TBA

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]

Knolewdge 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.

Knolewdge 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

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. [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.

Sources

- 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
- 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
- Hansen, Lars Kai, and Peter Salamon. "Neural network ensembles." IEEE transactions on pattern analysis and machine intelligence 12.10 (1990): 993-1001.
- Breiman, Leo, and Nong Shang. "Born again trees." Accessible from: https://www.stat.berkeley.edu/~breiman/BAtrees.pdf ps (1996).