Artificial Neural Networks: Deep or Broad? An Empirical Study

Nian Liu and Nayyar A. Zaidi

MONASH University
Information Technology
Introduction

- Two significant trends in machine learning in last 10 years:
  - Ever-growing quantities of training data – Advent of Big Data
  - Success of Deep Learning on many problems

Lessons learned

For big data we need low-bias models

Feature Engineering: Main reason behind the success of deep learning

Big Learning: Feature Engineering (low-bias), Minimal Pass, Minimal Tuning Parameters, Dynamic Models

Is feature engineering and low-bias models two new phenomenon?
Introduction

- Two significant trends in machine learning in last 10 years:
  - Ever-growing quantities of training data – Advent of Big Data
  - Success of Deep Learning on many problems

- Lessons learned
  - For big data we need low-bias models
  - Feature Engineering: Main reason behind the success of deep learning
Introduction

- Two significant trends in machine learning in last 10 years:
  - Ever-growing quantities of training data – Advent of Big Data
  - Success of Deep Learning on many problems
- Lessons learned
  - For big data we need low-bias models
  - Feature Engineering: Main reason behind the success of deep learning
- Big Learning: Feature Engineering (low-bias), Minimal Pass, Minimal Tuning Parameters, Dynamic Models
Introduction

Two significant trends in machine learning in last 10 years:
  ▶ Ever-growing quantities of training data – Advent of Big Data
  ▶ Success of Deep Learning on many problems

Lessons learned
  ▶ For big data we need low-bias models
  ▶ Feature Engineering: Main reason behind the success of deep learning

Big Learning: Feature Engineering (low-bias), Minimal Pass, Minimal Tuning Parameters, Dynamic Models

Is feature engineering and low-bias models two new phenomenon?
The Need for Low-Bias

- Much of machine learning has been conducted in the context of small datasets
- Variance dominates most of the error
- Low-bias models will lead to over-fitting
- Lots of emphasis on Regularization
- Big datasets requires low-bias models
Low-Bias Models

- Bayesian Networks
- Higher-order Logistic Regression
  - Generalized Linear Models
- Artificial Neural Networks
  - Deep Learning
- Random Forests
  - Other ensemble-based and tree models
- Support Vector Machines
  - Kernel Engineering $\equiv$ Feature Engineering
Low-Bias Models

- Bayesian Networks
- Higher-order Logistic Regression
  - Generalized Linear Models
- Artificial Neural Networks
  - Deep Learning
- Random Forests
- Support Vector Machines
Low-Bias Models

▶ Bayesian Networks

Low-Bias Models

- **Bayesian Networks**
Low-Bias Models

▶ Bayesian Networks


▶ Higher-order Logistic Regression

Low-Bias Models

- **Bayesian Networks**

- **Higher-order Logistic Regression**

- **Artificial Neural Networks**
Low-Bias Models

- **Bayesian Networks**

- **Higher-order Logistic Regression**

- **Artificial Neural Networks**
  - Why Broad? – One-hidden layer ANN are universal function-approximators
Low-Bias Models

- **Bayesian Networks**

- **Higher-order Logistic Regression**

- **Artificial Neural Networks**
  - Why Broad? – One-hidden layer ANN are universal function-approximators
  - Why Deep? – Constant-depth circuits are less powerful than deep circuits and Less no. of parameters
Low-Bias Models

▶ Bayesian Networks

▶ Higher-order Logistic Regression

▶ Artificial Neural Networks
  ▶ Why Broad? – One-hidden layer ANN are universal function-approximators
  ▶ Why Deep? – Constant-depth circuits are less powerful than deep circuits and Less no. of parameters
  ▶ Why not Deep?
Low-Bias Models

- **Bayesian Networks**

- **Higher-order Logistic Regression**

- **Artificial Neural Networks**
  - Why Broad? – One-hidden layer ANN are universal function-approximators
  - Why Deep? – Constant-depth circuits are less powerful than deep circuits and Less no. of parameters
  - Why not Deep?
    - Architecture Selection
    - Vanishing gradients
    - Solution: Greedy layer-wise trainings
Low-Bias Models

- **Bayesian Networks**

\[
P_{BN^k}(y|x) = \frac{P(y) \prod_{i=1}^{n} P(x_i|pa(x_i), y)}{\sum_{c=1}^{C} P(c) \prod_{i=1}^{n} P(x_i|pa(x_i), c)}.
\]

- **Higher-order Logistic Regression**

\[
P_{LR^n}(y|x) = \frac{\exp \left( \beta_y + \sum_{\alpha \in \binom{A}{n}} \beta_{y,\alpha,x_{\alpha}} \right)}{\sum_{c \in \Omega_y} \exp \left( \beta_c + \sum_{\alpha^* \in \binom{A}{n}} \beta_{c,\alpha^*,x_{\alpha^*}} \right)}.
\]

- **Artificial Neural Networks**

\[
P_{ANN^b,d}(y|x) = \frac{f_1 \left[ \sum_{j=1}^{nH} \beta_{k,0} + w_{k,j} f_0 \left( \beta_{j,0} + \beta_j^T x \right) \right]}{Z}.
\]
Observations and Motivations

Observations

- We know that:
  - higher $k$ will lead to low-bias $\text{BN}^k$
  - higher $n$ will lead to low-bias $\text{LR}^n$
- We do not know:
  - higher $b$ or $d$ will lead to low-bias $\text{ANN}^{b,d}$
  - should $b$ be preferred over $d$ or vice-versa
  - what is the effect on the convergence?

Motivations

- A comparative analysis of low-bias models warrants further investigation
- Efficient, low-bias and dynamic models are the key to solving big data enigma
Experimental Design: Broad vs. Deep ANN

- 73 datasets from UCI repository
- 2-fold cross-validation
- 0-1 Loss, RMSE, Bias, Variance and Convergence performance
- Bias and Variance definition of Kohavi and Wolpart
- Win-Draw-Loss results are reported
- Separate analysis on Big Datasets
- 12 datasets with more than 10000 instances
Experimental Design: Broad vs. Deep ANN

- Deep Models denoted as: NN2, NN22, NN222, NN2222, NN2222, representing 1, 2, 3, 4, and 5 hidden layers each with two nodes each.

- Broad Models denoted as: NN2, NN4, NN6, NN8, NN10, representing 1 hidden layer with 2, 4, 6, 8 and 10 nodes.

- For sake of comparison, we also include NN0, this zero-hidden layer ANN and is equivalent to linear Logistic Regression.
Broad ANN – Bias, Variance Comparison

<table>
<thead>
<tr>
<th></th>
<th>vs. NN0</th>
<th>vs. NN2</th>
<th>vs. NN4</th>
<th>vs. NN6</th>
<th>vs. NN8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W-D-L</td>
<td>p</td>
<td>W-D-L</td>
<td>p</td>
<td>W-D-L</td>
</tr>
<tr>
<td>All Datasets - Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>35/3/34</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN4</td>
<td>45/4/23</td>
<td>0.010</td>
<td>49/7/16</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>NN6</td>
<td>47/4/21</td>
<td>0.002</td>
<td>47/5/20</td>
<td>0.001</td>
<td>37/7/28</td>
</tr>
<tr>
<td>NN8</td>
<td>48/3/21</td>
<td>0.002</td>
<td>44/5/23</td>
<td>0.014</td>
<td>37/7/28</td>
</tr>
<tr>
<td>NN10</td>
<td>52/3/17</td>
<td>&lt;0.001</td>
<td>47/5/20</td>
<td>0.001</td>
<td>41/9/22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43/10/19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40/15/17</td>
</tr>
<tr>
<td>All Datasets - Variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>20/2/50</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN4</td>
<td>21/2/49</td>
<td>0.001</td>
<td>38/6/28</td>
<td>0.268</td>
<td></td>
</tr>
<tr>
<td>NN6</td>
<td>27/3/42</td>
<td>0.091</td>
<td>43/7/22</td>
<td>0.013</td>
<td>40/8/24</td>
</tr>
<tr>
<td>NN8</td>
<td>32/2/38</td>
<td>0.550</td>
<td>42/7/23</td>
<td>0.025</td>
<td>44/8/20</td>
</tr>
<tr>
<td>NN10</td>
<td>30/3/39</td>
<td>0.336</td>
<td>42/7/23</td>
<td>0.025</td>
<td>43/9/20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34/13/25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33/10/29</td>
</tr>
</tbody>
</table>

**Table:** A comparison of Bias and Variance of broad models in terms of W-D-L on All datasets. *p* is two-tail binomial sign test. Results are significant if $p \leq 0.05$. 
## Broad ANN – Error Comparison

Table: A comparison of 0-1 Loss and RMSE of broad models in terms of W-D-L on All and Big datasets. *p* is two-tail binomial sign test. Results are significant if *p* ≤ 0.05.

<table>
<thead>
<tr>
<th></th>
<th>vs. NN0</th>
<th></th>
<th>vs. NN2</th>
<th></th>
<th>vs. NN4</th>
<th></th>
<th>vs. NN6</th>
<th></th>
<th>vs. NN8</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Datasets – 0-1 Loss</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>27/2/43</td>
<td>0.072</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN4</td>
<td>31/6/35</td>
<td>0.712</td>
<td>50/9/13</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN6</td>
<td>33/3/36</td>
<td>0.801</td>
<td>49/3/20</td>
<td>&lt;0.001</td>
<td>45/7/20</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN8</td>
<td>37/1/34</td>
<td>0.813</td>
<td>50/5/17</td>
<td>&lt;0.001</td>
<td>44/8/20</td>
<td>0.004</td>
<td>31/14/27</td>
<td>0.694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN10</td>
<td>40/2/30</td>
<td>0.282</td>
<td>51/4/17</td>
<td>&lt;0.001</td>
<td>49/5/18</td>
<td>&lt;0.001</td>
<td>38/9/25</td>
<td>0.130</td>
<td>40/8/24</td>
<td>0.060</td>
</tr>
<tr>
<td><strong>Big Datasets – 0-1 Loss</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>6/0/6</td>
<td>1.226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN4</td>
<td>7/0/5</td>
<td>0.774</td>
<td>12/0/0</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN6</td>
<td>7/0/5</td>
<td>0.774</td>
<td>12/0/0</td>
<td>0.001</td>
<td>11/0/1</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN8</td>
<td>8/0/4</td>
<td>0.388</td>
<td>12/0/0</td>
<td>&lt;0.001</td>
<td>9/0/3</td>
<td>0.146</td>
<td>8/0/4</td>
<td>0.388</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN10</td>
<td>8/0/4</td>
<td>0.388</td>
<td>12/0/0</td>
<td>&lt;0.001</td>
<td>10/0/2</td>
<td>0.039</td>
<td>9/0/3</td>
<td>0.146</td>
<td>9/0/3</td>
<td>0.146</td>
</tr>
</tbody>
</table>

---

2016: The 29th Australasian Joint Conference on Artificial Intelligence  Nian Liu and Nayar A. Zaidi
Figure: Comparison (geometric average) of 0-1 Loss, RMSE, Bias and Variance for broad models on All and Big datasets. Results are normalized w.r.t NN0.
Deep ANN – Bias, Variance Comparison

<table>
<thead>
<tr>
<th></th>
<th>vs. NN0</th>
<th>vs. NN2</th>
<th>vs. NN22</th>
<th>vs. NN222</th>
<th>vs. NN2222</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>NN2</td>
<td>35/3/34</td>
<td>28/4/40</td>
<td>24/4/44</td>
<td>21/3/48</td>
<td>20/2/50</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.182</td>
<td>0.021</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NN22</td>
<td>30/3/39</td>
<td>21/3/48</td>
<td>32/4/36</td>
<td>32/9/31</td>
<td>38/1/33</td>
</tr>
<tr>
<td></td>
<td>0.336</td>
<td>0.002</td>
<td>0.905</td>
<td>1</td>
<td>0.6353</td>
</tr>
<tr>
<td>NN222</td>
<td>26/1/45</td>
<td>24/4/44</td>
<td>36/2/34</td>
<td>35/9/28</td>
<td>39/2/31</td>
</tr>
<tr>
<td></td>
<td>0.032</td>
<td>0.001</td>
<td>0.905</td>
<td>0.450</td>
<td>0.403</td>
</tr>
<tr>
<td>NN2222</td>
<td>5/0/67</td>
<td>3/2/67</td>
<td>4/9/59</td>
<td>1</td>
<td>8/61/3</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NN22222</td>
<td>0/1/71</td>
<td>1/2/69</td>
<td>1/9/62</td>
<td>0/61/11</td>
<td>0.282</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Table:** Bias W-D-L on **All** and **Big** datasets. *p* is two-tail binomial sign test. Results are significant if *p* ≤ 0.05.
Deep ANN – Error Comparison

<table>
<thead>
<tr>
<th></th>
<th>vs. NN0</th>
<th>vs. NN2</th>
<th>vs. NN22</th>
<th>vs. NN222</th>
<th>vs. NN2222</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Datasets – 0-1 Loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>27/2/43</td>
<td>0.072</td>
<td>24/5/43</td>
<td>0.027</td>
<td>28/3/41</td>
</tr>
<tr>
<td>NN22</td>
<td>28/1/43</td>
<td>0.096</td>
<td>25/5/42</td>
<td>0.050</td>
<td>28/3/41</td>
</tr>
<tr>
<td>NN222</td>
<td>24/1/47</td>
<td>0.009</td>
<td>25/5/42</td>
<td>0.050</td>
<td>28/3/41</td>
</tr>
<tr>
<td>NN2222</td>
<td>7/0/65</td>
<td>&lt;0.001</td>
<td>4/2/66</td>
<td>&lt;0.001</td>
<td>4/2/66</td>
</tr>
<tr>
<td>NN22222</td>
<td>7/1/64</td>
<td>&lt;0.001</td>
<td>5/1/66</td>
<td>&lt;0.001</td>
<td>4/2/66</td>
</tr>
<tr>
<td>Big Datasets – 0-1 Loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN2</td>
<td>6/0/6</td>
<td>1.226</td>
<td>4/0/8</td>
<td>0.388</td>
<td>4/0/8</td>
</tr>
<tr>
<td>NN22</td>
<td>5/0/7</td>
<td>0.774</td>
<td>2/0/10</td>
<td>0.039</td>
<td>4/0/8</td>
</tr>
<tr>
<td>NN222</td>
<td>4/0/8</td>
<td>0.388</td>
<td>2/0/10</td>
<td>0.039</td>
<td>4/0/8</td>
</tr>
<tr>
<td>NN2222</td>
<td>2/0/10</td>
<td>0.039</td>
<td>0/0/12</td>
<td>&lt;0.001</td>
<td>1/0/11</td>
</tr>
<tr>
<td>NN22222</td>
<td>1/1/10</td>
<td>0.012</td>
<td>0/0/12</td>
<td>&lt;0.001</td>
<td>0/0/12</td>
</tr>
</tbody>
</table>

Table: 0-1 Loss W-D-L on All and Big datasets. $p$ is two-tail binomial sign test. Results are significant if $p \leq 0.05$. 

AI 2016: The 29th Australasian Joint Conference on Artificial Intelligence
Nian Liu and Nayyar A. Zaidi
Figure: Comparison (geometric average) of 0-1 Loss, RMSE, Bias and Variance for deep models on Little and Big datasets. Results are normalized w.r.t NN0.
Convergence Analysis (Broad)

Figure: Variation in Mean Square Error of NN2, NN4, NN6, NN8 and NN10 with increasing number of (optimization) iterations on sample datasets.
Figure: Variation in Mean Square Error of NN2, NN22, NN222, NN2222 and NN22222 with increasing number of (optimization) iterations on sample datasets.
Conclusion

- Results warrants further investigation
- Deep versus Broad
- Deep versus Shallow
- Q & A

- For Further Discussions
  - @nayyar_zaidi
  - nayyar.zaidi@monash.edu
  - nayyar_zaidi