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

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

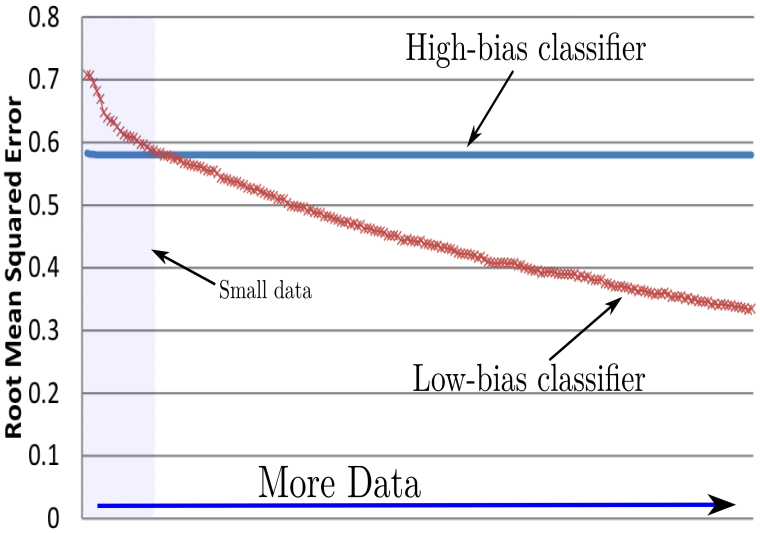
Introduction

- 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
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)

Low-Bias Models

- Bayesian Networks
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)

Low-Bias Models

- Bayesian Networks
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)

Low-Bias Models

- Bayesian Networks
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)
- Artificial Neural Networks

Low-Bias Models

- Bayesian Networks
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)
- Artificial Neural Networks
  - Why Broad? – One-hidden layer ANN are universal function-approximators

Low-Bias Models

- Bayesian Networks
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)
- 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
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)
- 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
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Buntine, W. and Hynes, M. and De Sterck, H. – Efficient Parameter Learning of Bayesian Network Classifiers, to Appear in Machine Learning (2017)
  - Martinez, A. M. Chen, S. and Webb, G. I. and Zaidi, N. A. – Scalable Learning of Bayesian Network Classifiers, Journal of Machine Learning Research, pp: 1-35, volume: 17, (2016)
- Higher-order Logistic Regression
  - Zaidi, N. A. and Webb, G. I. and Carman, M. J. and Petitjean, F. and Cerquides, J. – ALR^n: Accelerated Higher-order Logistic Regression, Machine Learning, pp: 151-194, volume: 104, (2016)
- 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

$$\mathrm{P}_{BN^k}(y|\mathbf{x}) = \frac{\mathrm{P}(y)\prod_{i=1}^{n}\mathrm{P}(x_i|\mathrm{pa}(x_i),y)}{\sum_{c=1}^{C}\mathrm{P}(c)\prod_{i=1}^{n}\mathrm{P}(x_i|\mathrm{pa}(x_i),c)}.$$

- Higher-order Logistic Regression

$$\mathrm{P}_{LR^n}(y|\mathbf{x}) = \frac{\exp\left(\beta_y + \sum_{\alpha\in\binom{\mathcal{A}}{n}}\beta_{y,\alpha,x_\alpha}\right)}{\sum_{c\in\Omega_Y}\exp\left(\beta_c + \sum_{\alpha^*\in\binom{\mathcal{A}}{n}}\beta_{c,\alpha^*,x_{\alpha^*}}\right)}.$$

- Artificial Neural Networks

$$\mathrm{P}_{ANN^{b,d}}(y|\mathbf{x}) = \frac{f_1\left[\sum_{j=1}^{n_H}\beta_{k,0} + w_{k,j}f_0\left(\beta_{j,0} + \beta_j^T\mathbf{x}\right)\right]}{\mathcal{Z}}.$$

Observations and Motivations

Observations

- We know that:
  - higher k will lead to low-bias BN^k
  - higher n will lead to low-bias LR^n
- We do not know:
  - higher b or d will lead to low-bias $\mathrm{ANN}^{b,d}$
  - 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

	vs. NN0		vs. NN2		vs. NN4		vs. NN6		vs. NN8	
	W-D-L	*p*	W-D-L	*p*	W-D-L	*p*	W-D-L	*p*	W-D-L	*p*
All Datasets - Bias										
NN2	35/3/34	1								
NN4	45/4/23	**0.010**	49/7/16	**<0.001**						
NN6	47/4/21	**0.002**	47/5/20	**0.001**	37/7/28	0.321				
NN8	48/3/21	**0.002**	44/5/23	**0.014**	37/7/28	0.321	36/11/25	0.200		
NN10	52/3/17	**<0.001**	47/5/20	**0.001**	41/9/22	**0.023**	43/10/19	**0.003**	40/15/17	**0.003**
All Datasets - Variance										
NN2	20/2/50	**<0.001**								
NN4	21/2/49	**0.001**	38/6/28	0.268						
NN6	27/3/42	0.091	43/7/22	**0.013**	40/8/24	0.060				
NN8	32/2/38	0.550	42/7/23	**0.025**	44/8/20	**0.004**	36/9/27	0.314		
NN10	30/3/39	0.336	42/7/23	**0.025**	43/9/20	**0.005**	34/13/25	0.298	33/10/29	0.704

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

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

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 $p \leq 0.05$.

Broad ANN – Geometric Averages

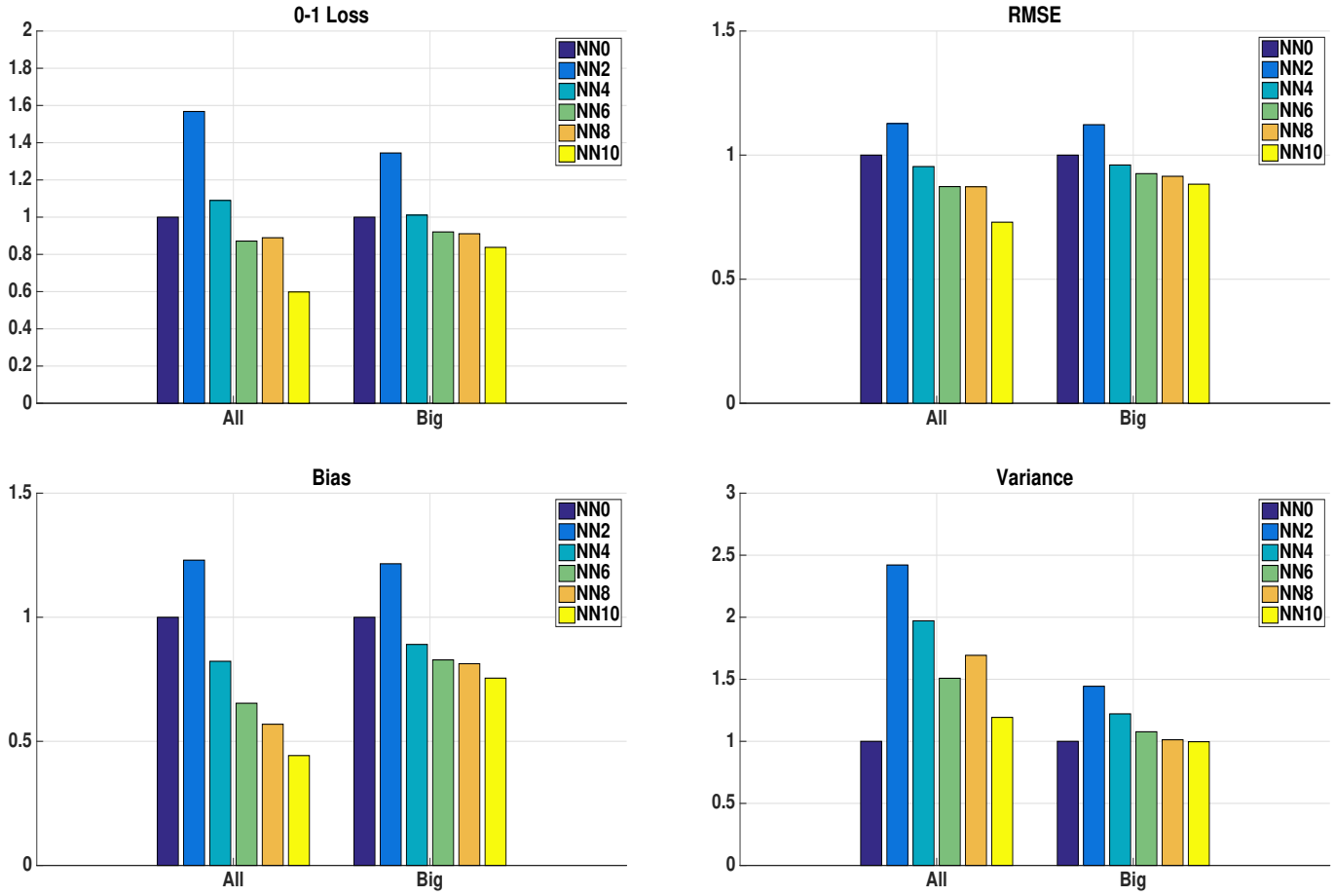


Figure: Comparison (geometric average) of 0-1 Loss, RMSE, Bias and Variance for broad models on All and Big datasets. Results are normalized w.r.t NN0.

Deep ANN – Bias, Variance Comparison

	vs. NN0		vs. NN2		vs. NN22		vs. NN222		vs. NN2222	
	W-D-L	p	W-D-L	p	W-D-L	p	W-D-L	p	W-D-L	p
All Datasets – Bias										
NN2	35/3/34	1								
NN22	30/3/39	0.336	28/4/40	0.182						
NN222	26/1/45	**0.032**	21/3/48	**0.002**	24/4/44	**0.021**				
NN2222	5/0/67	**<0.001**	3/1/68	**<0.001**	3/2/67	**<0.001**	4/9/59	**<0.001**		
NN22222	0/1/71	**<0.001**	0/1/71	**<0.001**	1/2/69	**<0.001**	1/9/62	**<0.001**	0/61/11	**<0.001**
All Datasets – Variance										
NN2	20/2/50	**<0.001**								
NN22	20/1/51	**<0.001**	27/6/39	0.175						
NN222	24/1/47	**0.009**	34/3/35	1	32/4/36	0.905				
NN2222	34/1/37	0.813	34/1/37	0.813	36/2/34	0.905	32/9/31	1		
NN22222	40/2/30	0.282	38/1/33	0.6353	39/2/31	0.403	35/9/28	0.450	8/61/3	0.227

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

Deep ANN – Error Comparison

	vs. NN0		vs. NN2		vs. NN22		vs. NN222		vs. NN2222	
	W-D-L	p	W-D-L	p	W-D-L	p	W-D-L	p	W-D-L	p
All Datasets – 0-1 Loss										
NN2	27/2/43	0.072								
NN22	28/1/43	0.096	24/5/43	**0.027**						
NN222	24/1/47	**0.009**	25/5/42	**0.050**	28/3/41	0.148				
NN2222	7/0/65	**<0.001**	4/2/66	**<0.001**	4/2/66	**<0.001**	3/9/60	**<0.001**		
NN22222	7/1/64	**<0.001**	5/1/66	**<0.001**	4/2/66	**<0.001**	3/9/60	**<0.001**	1/61/10	**0.012**
Big Datasets – 0-1 Loss										
NN2	6/0/6	1.226								
NN22	5/0/7	0.774	4/0/8	0.388						
NN222	4/0/8	0.388	2/0/10	**0.039**	4/0/8	0.388				
NN2222	2/0/10	**0.039**	0/0/12	**<0.001**	1/0/11	**0.006**	1/1/10	**0.012**		
NN22222	1/1/10	**0.012**	0/0/12	**<0.001**	0/0/12	**<0.001**	0/1/11	**<0.001**	0/6/6	**0.031**

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

Deep ANN – Geometric Averages

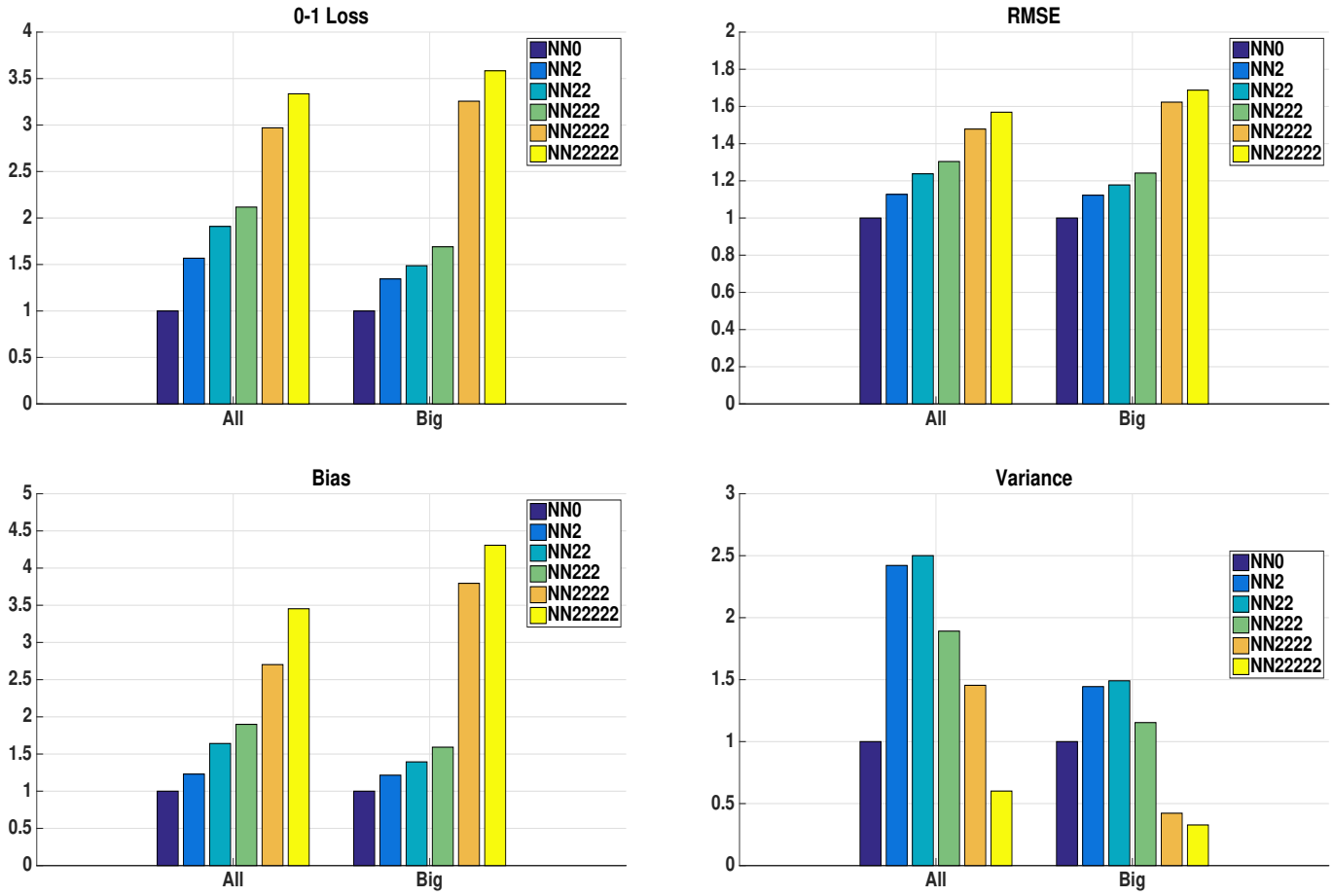


Figure: Comparison (geometric average) of 0-1 Loss, RMSE, Bias and Variance for deep models on Little and Big datasets. Results are normalized w.r.t NN0.

Convergence Analysis (Broad)

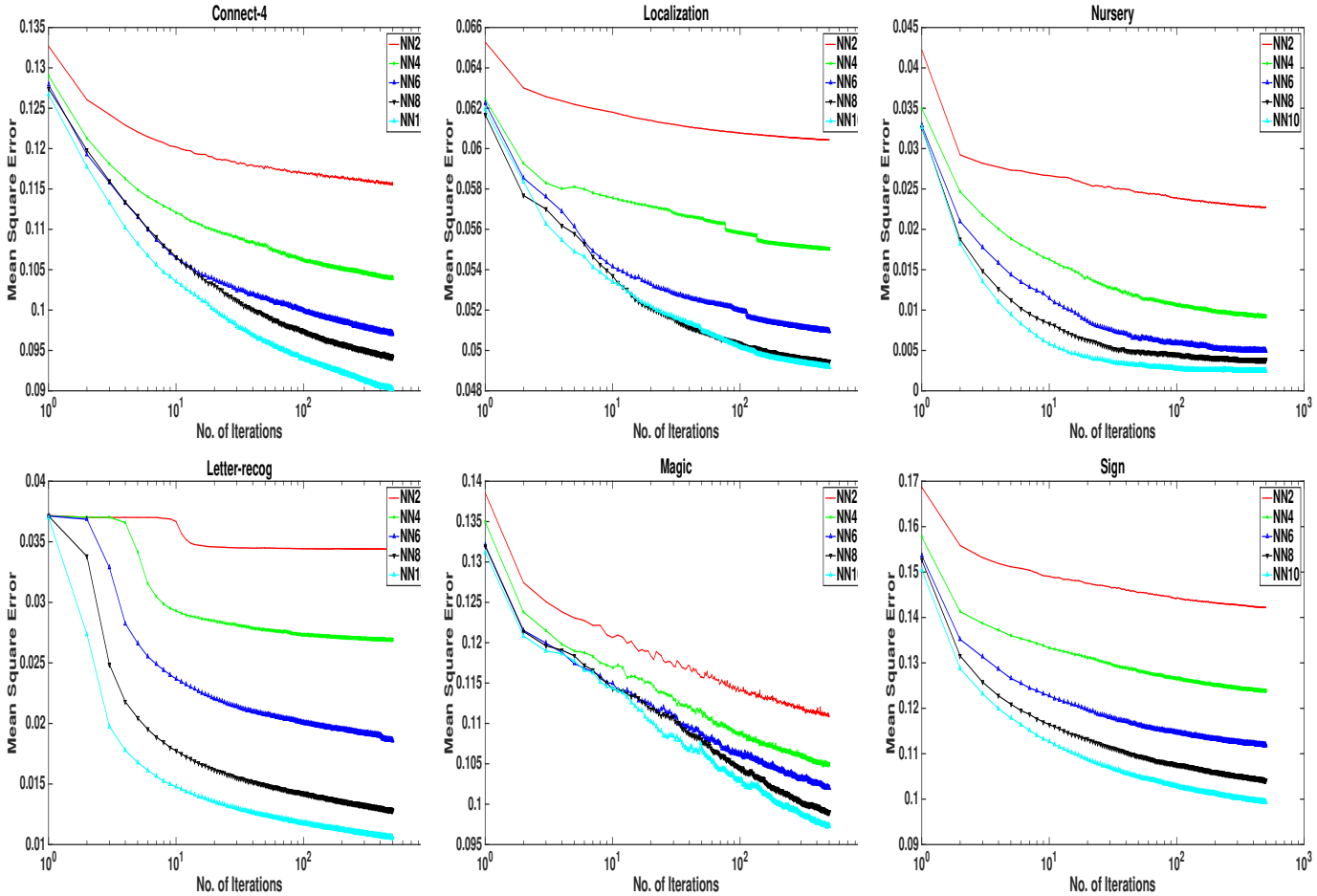


Figure: Variation in Mean Square Error of NN2, NN4, NN6, NN8 and NN10 with increasing number of (optimization) iterations on sample datasets.

Convergence Analysis (Deep)

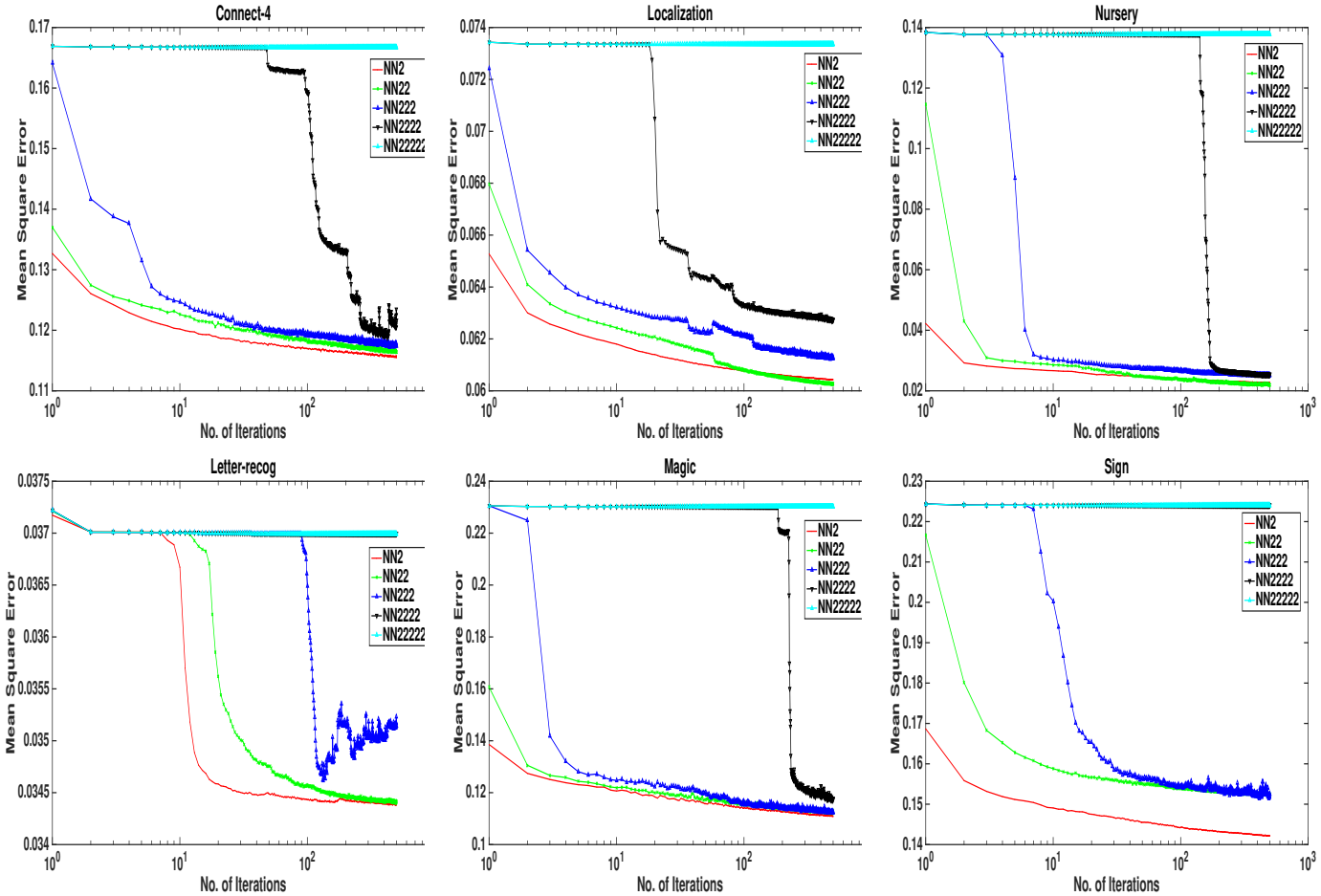


Figure: Variation in Mean Square Error of NN2, NN22, NN222, NN2222 and NN22222 with increasing number of (optimization) iterations on sample datasets.

Conclusion

- Results warrants further investigation
- Deep versus Broad
- Deep versus Shallow
- Q & A
- For Further Discussions
  - @nayyar_zaidi
  - nayyar.zaidi@monash.edu
  - nayyar_zaidi
  - `http://users.monash.edu.au/~nzaidi`