Modern Neural Networks Generalize on Small Data Sets

Matthew Olson
UPenn

Abraham J. Wyner
UPenn

Richard Berk
UPenn

Motivation

Deep networks work really well on the standard image data sets: large sample sizes, very low label noise, highly structured.
Questions:

- Do they work in high noise, “small data" settings outside of image/speech recognition?
- Can they simultaneously memorize (noisy) training labels and still generalize well?
- We know "yes" to (1) and (2) for random forests - can we borrow insights?

Highlights

Empirical Insights

- 116 multi-class data sets from UCI repository ($n \approx 600$ observations on average)
- fit **unregularized** deep networks, zero training error on all data sets
- **test error comparable to a random forest!**

Ensemble Interpretation

- networks decompose into sub-networks with low bias and relatively low correlation
- deep layers serve to aid variance reduction

Ensembling

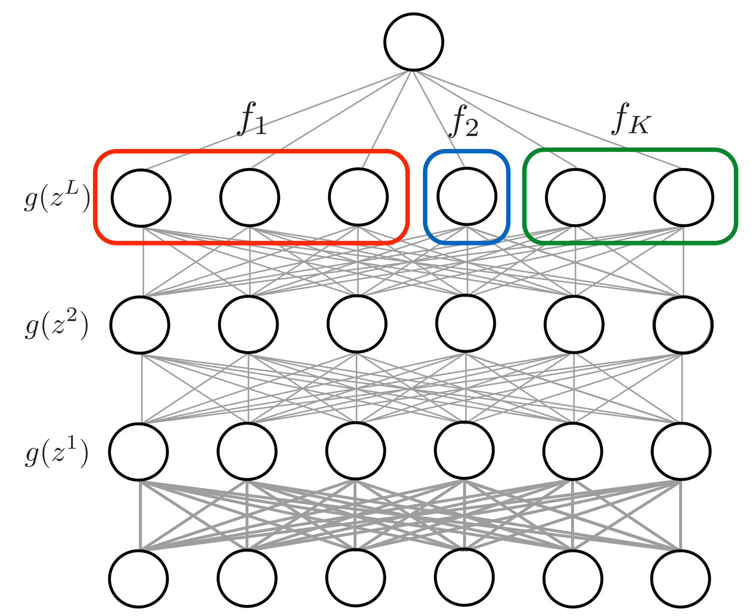
A neural network with L hidden layers and M hidden nodes can be written

$$z^{\ell+1} = W^{\ell+1} g(z^\ell) \quad \ell = 0, \dots, L$$
$$f(x) = \sigma(z^{L+1})$$

where final hidden layer is a sum of sub-networks:

$$\begin{aligned} z^{L+1}(x) &= W^{L+1} g(z^L(x)) \\ &= \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_{m,k} g(z_m^L(x)) \\ &= \sum_{k=1}^{K} f_k(x) \end{aligned}$$

Decomposing a Neural Network into an Ensemble



Ensemble Program

- decompose final layer into sub-networks $f_1, f_2, \dots, f_K$
- search for sub-networks with low bias and low pairwise error correlation
- construct $f_1, f_2, \dots, f_K$ from linear program
- low bias + low variance $\rightarrow$ good generalization

Ensemble Hunting via Linear Programming

Train a network with M hidden nodes, L hidden layers until zero training error
Find $\alpha \in \mathbb{R}^{M \times K}$ satisfying the linear system (target K sub-networks):

$$\sum_{k=1}^{K} \alpha_{m,k} = W_{1,m}^{L+1} \quad 1 \le m \le M \qquad \text{(decompose the final hidden layer)}$$

$$\alpha_{m_{j,k},k} = 0 \quad 1 \le j \le \frac{M}{2}, 1 \le k \le K \qquad \text{(diversity constraint)}$$

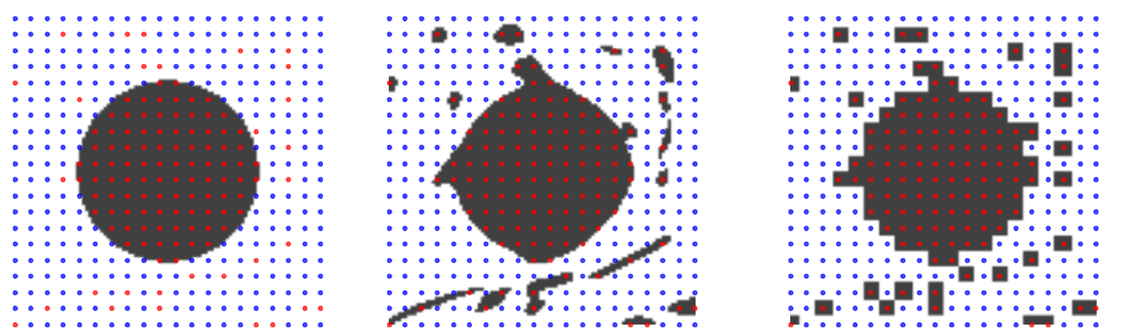
$$\left(\sum_{m=1}^{M} \alpha_{m,k} g(z_m^L(x_i)) \right) y_i \ge 0 \quad 1 \le i \le n, 1 \le k \le K \qquad \text{(each sub-network has non-negative margin)}$$

Ensemble Hunting: Simulated Example

Draw samples $(x_i, y_i) \in [-1,1]^2 \times \{-1,1\}$ from

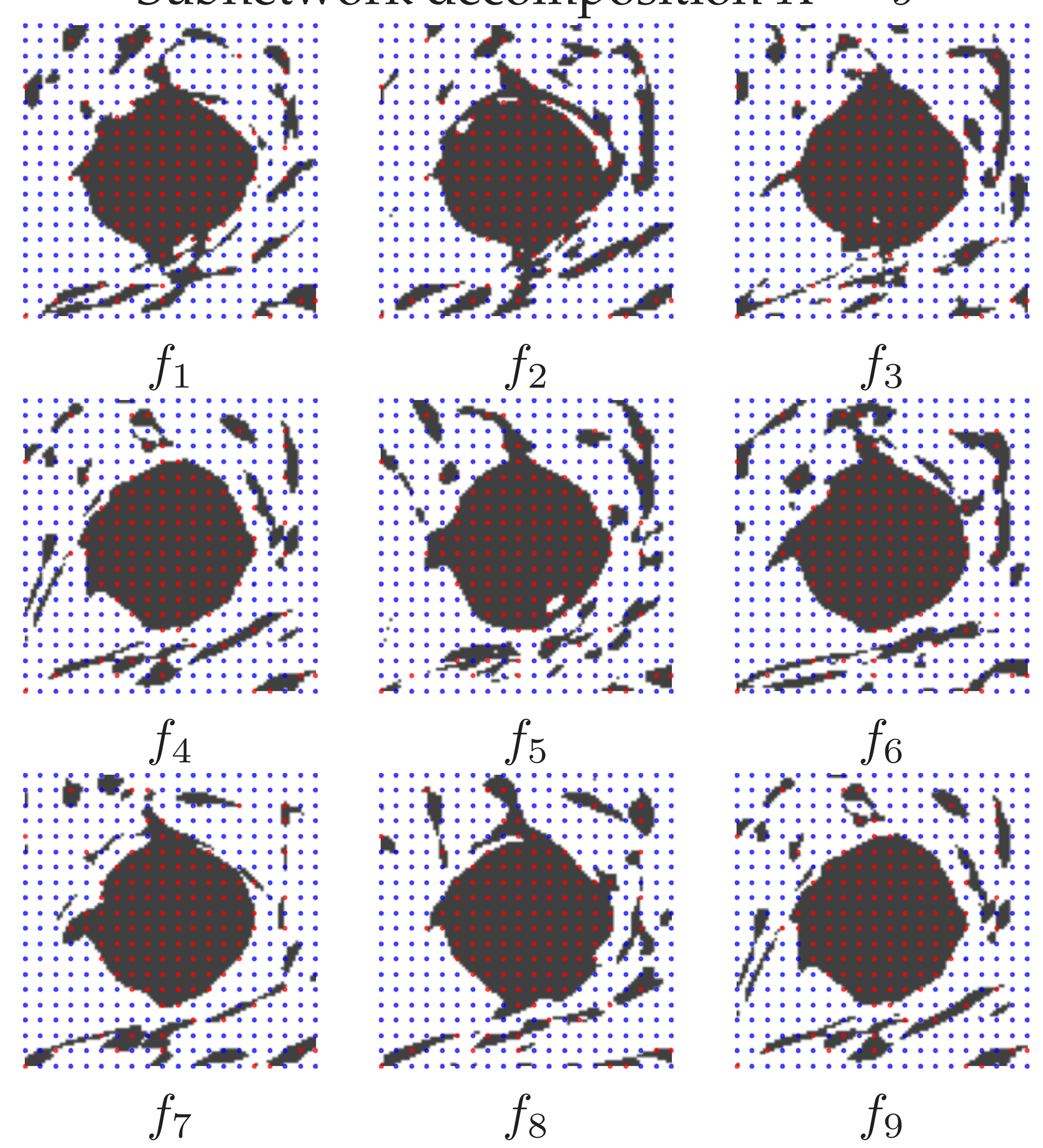
$$p(y=1|x) = \begin{cases} 1 & \text{if } \|x\|_2 \le 0.3 \\ 0.15 & \text{otherwise} \end{cases}$$

- 10% label noise (red points)
- train 10 layer network until zero training error
- influence of noise points localized
- **better test error than random forest**



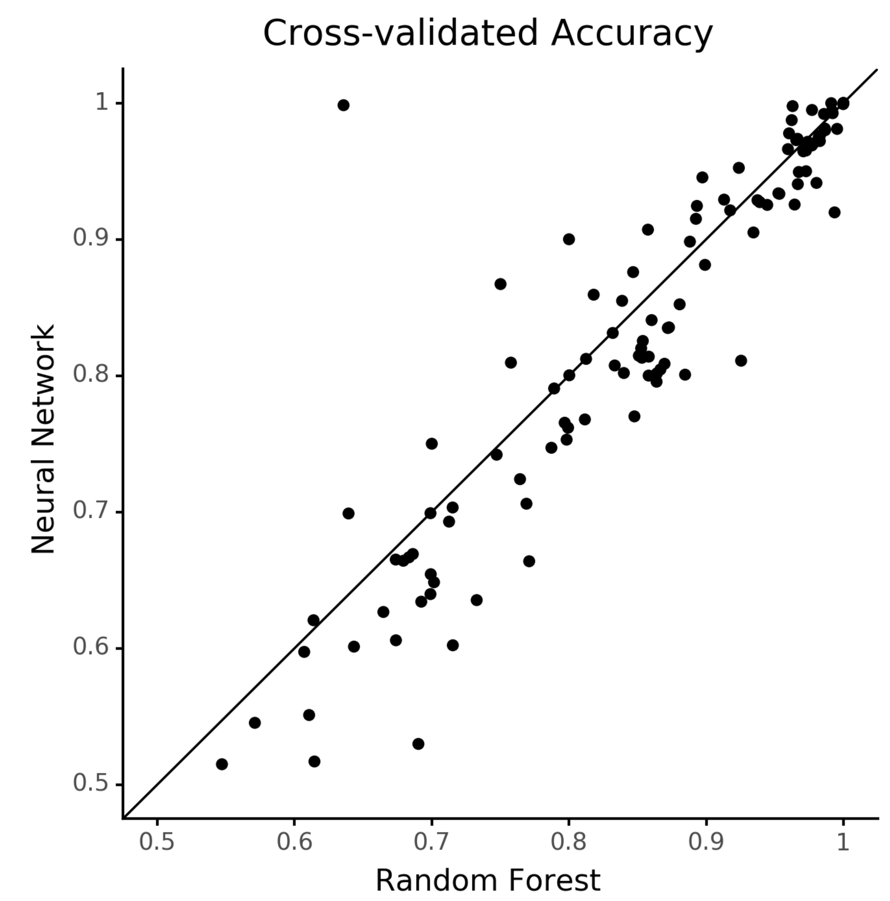
Bayes rule | DNN | Random Forest

Subnetwork decomposition $K = 9$



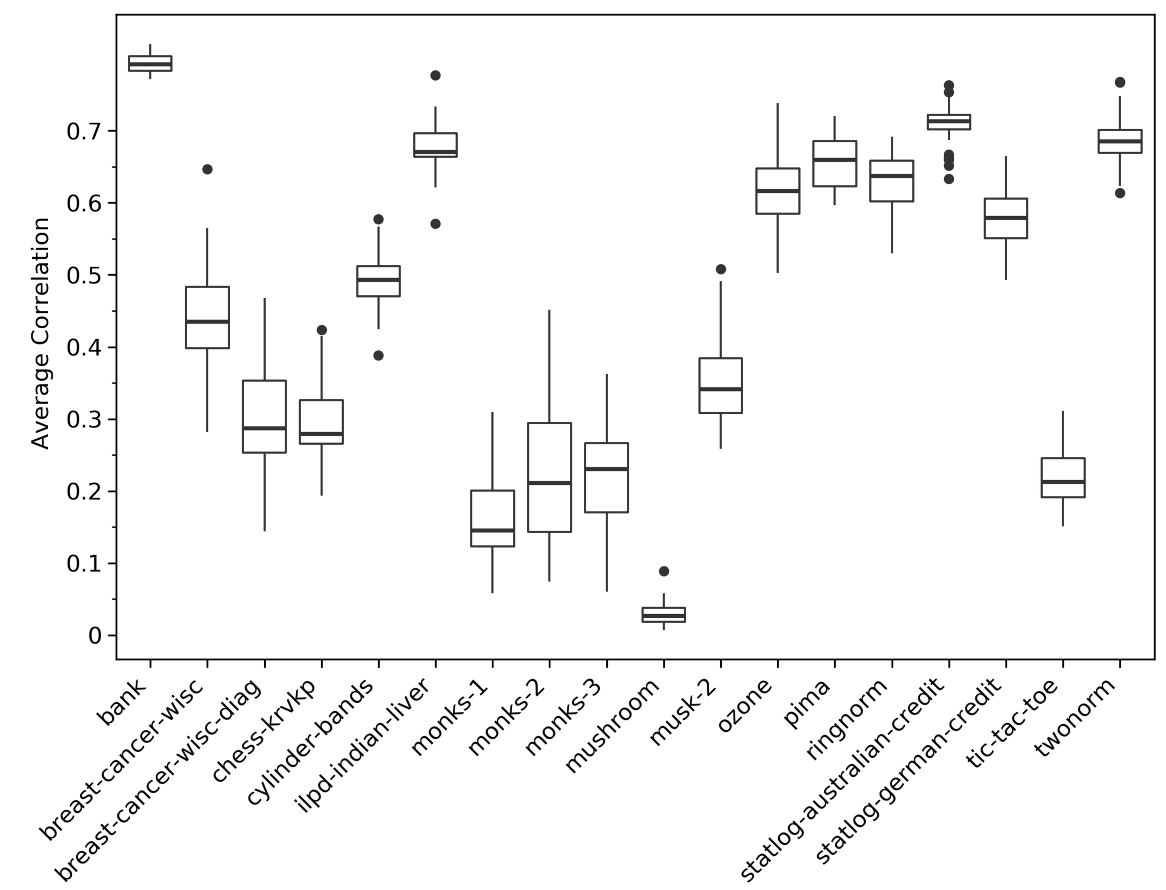
Empirical Evaluation

- 116 UCI repository data sets (classification)
- depth 10 networks trained to 100% training accuracy (no regularization)
- hundreds of parameters per observation



Decorrelation

Average pairwise error correlation between ensemble components $f_1, \dots, f_9$



Takeaways

- high capacity networks can still generalize well on small data sets with non-trivial noise
- ensemble interpretation of deep networks, deeper layers offer variance reduction