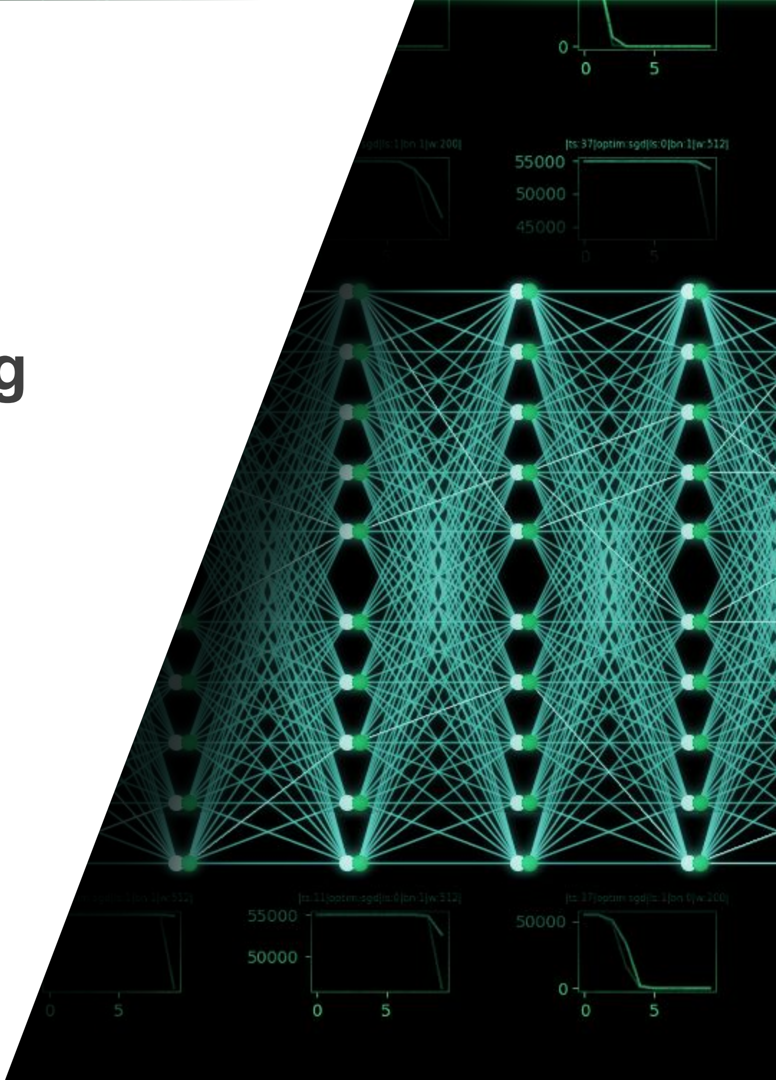
Exploring neural network training dynamics through binary node activations

Gerbrand Haasbroek & Marelie H Davel

Multilingual Speech Technologies, North-West University, South Africa
Centre for Artificial Intelligence Research



Background

- No theoretical framework for DNN training or generalisation
- Generalisation studied using
  - Geometry of loss landscape
  - Stability and robustness
  - Complexity of hypothesis space
  - Margin distributions

Background

- *Sample Set* - Samples that activate a node
- Sample sets formed during the forward pass
- Sample sets refined during the backward pass
- Evolution of sample sets has not been explored

Contributions

- Motivation of the importance of sample sets
- Demonstration of a consistent evolution of sample sets
- Interpretation of findings in terms of clustering

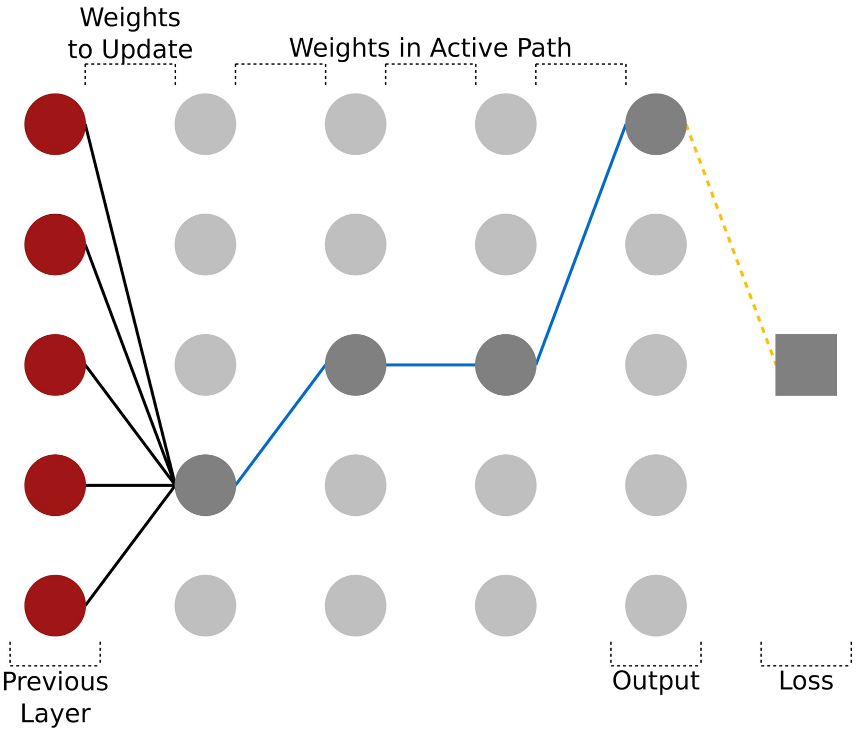
Sample Sets

- Sample added to set if

$$\mathbf{z}_{l-1}^{\mathbf{t}} \cdot \mathbf{w}_{l,j} > \eta \sum_{\mathbf{s} \in \hat{S}_{b,l,j}} \left(\mathbf{z}_{l-1}^{\mathbf{t}} \cdot \mathbf{z}_{l-1}^{\mathbf{s}} \right) \phi_{l,j}^{\mathbf{s}}$$

- Sample removed otherwise
- Node-supported cost

$$\phi_{l,j}^{\mathbf{s}} = \sum_{p \in P_j^{\mathbf{s}}} \left(\lambda_p^{\mathbf{s}} \prod_{k=1}^{N-l} w_{p_k} \right)$$



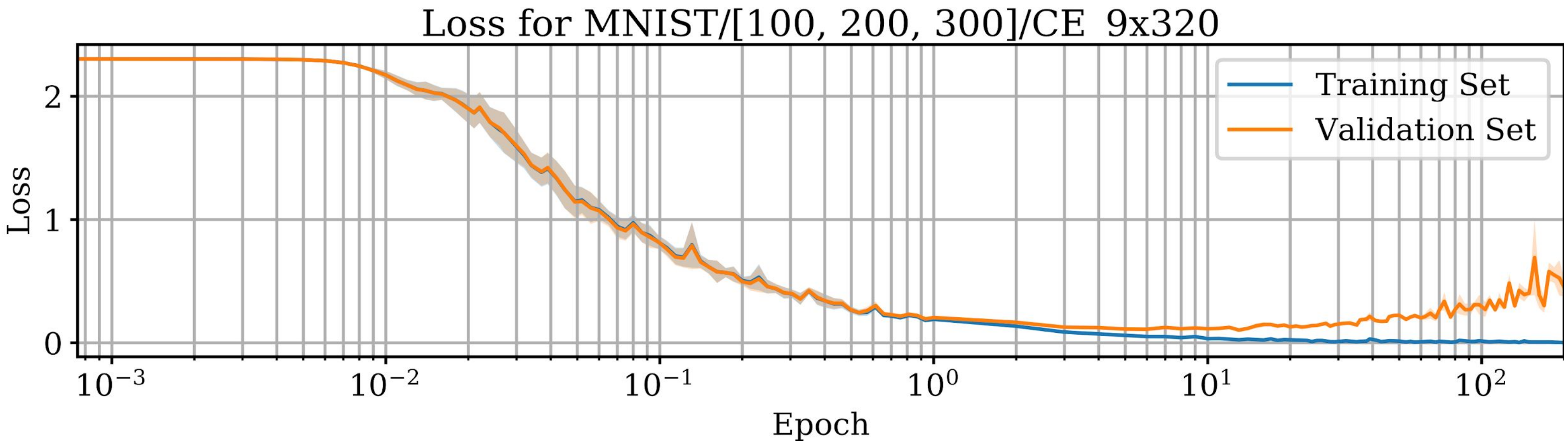
Experiments

- Train fully connected networks:
  - Number of layers: 2, 5, 9
  - Number of nodes per layer: 20, 80, 320
  - Activation functions: ReLU
  - Loss functions: mean squared error, cross entropy
  - Adam optimiser
  - MNIST dataset
  - Epochs: 200

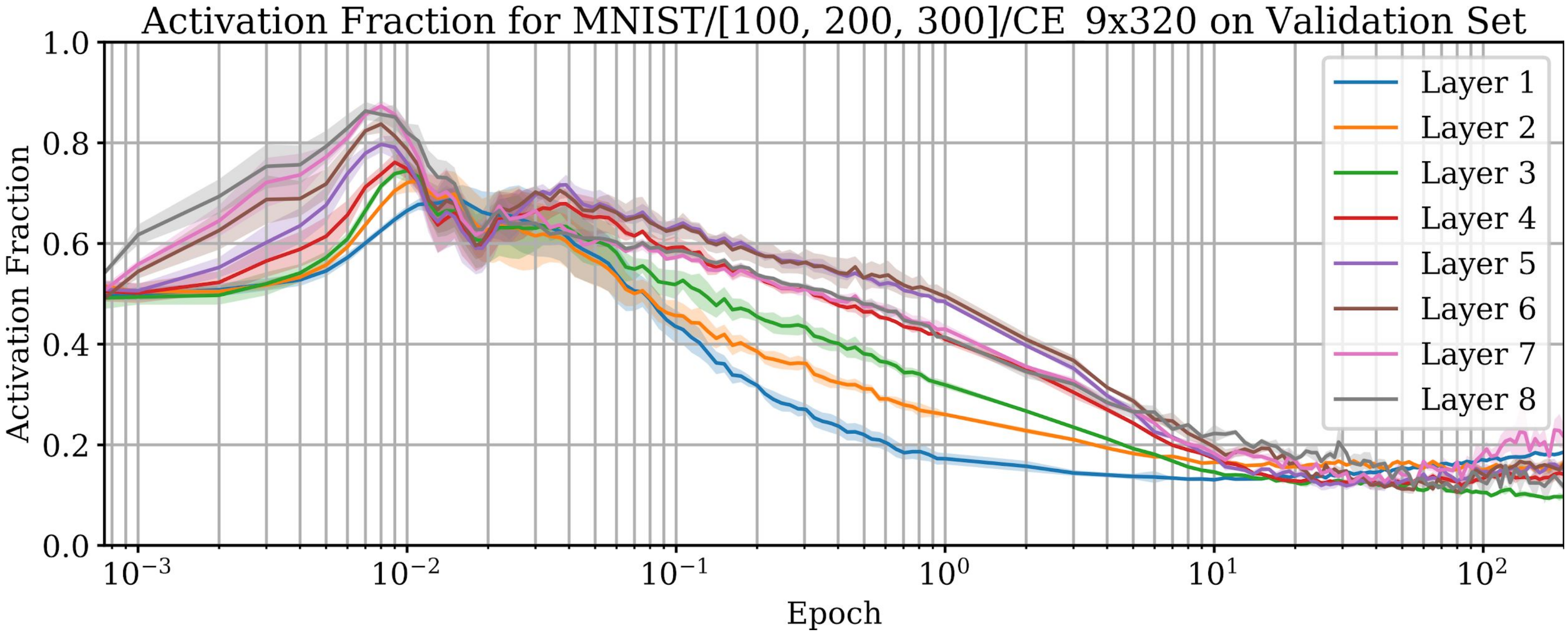
Measurements

- *Activation fraction* - Fraction of all samples in sample set
- *Predictability* - Average amount of information about binary activation given class
- *Informativeness* - Average amount of information about class given binary activation

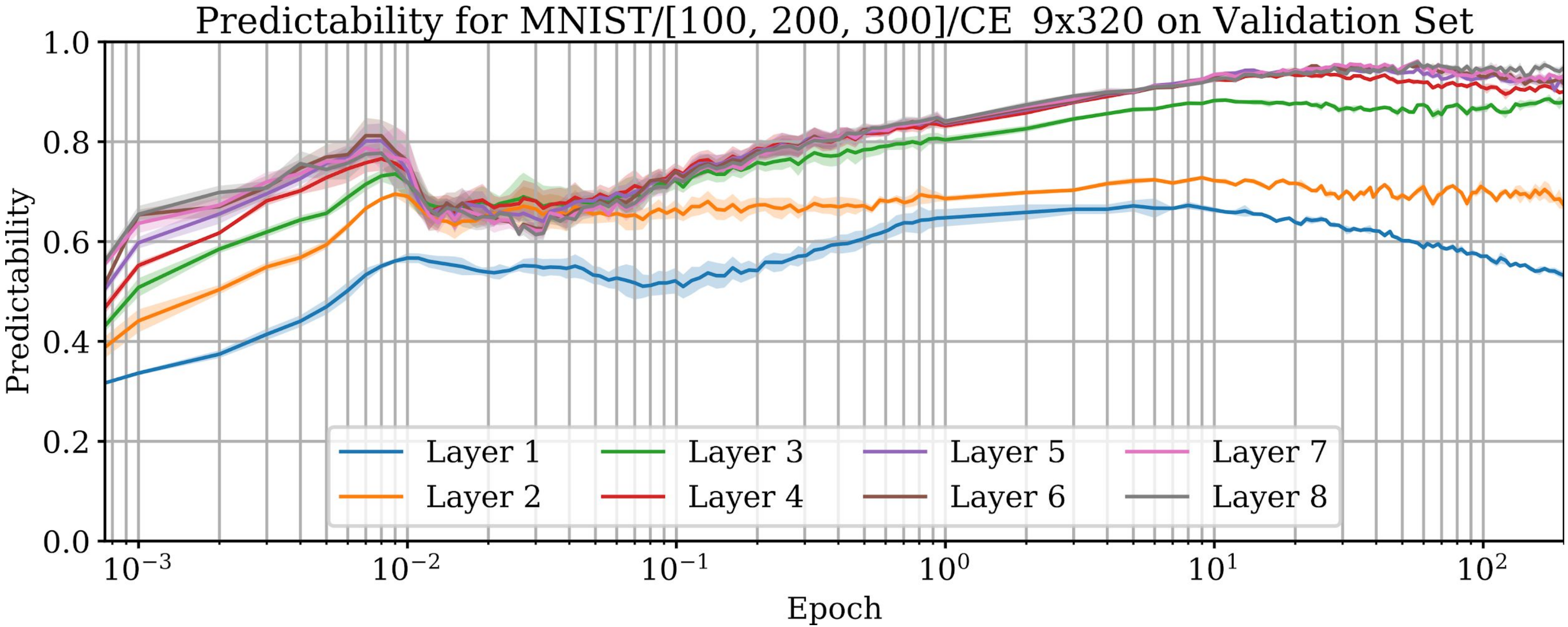
Results


Loss for MNIST/[100, 200, 300]/CE_9x320
Loss
2
1
0
Epoch
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Training Set
Validation Set

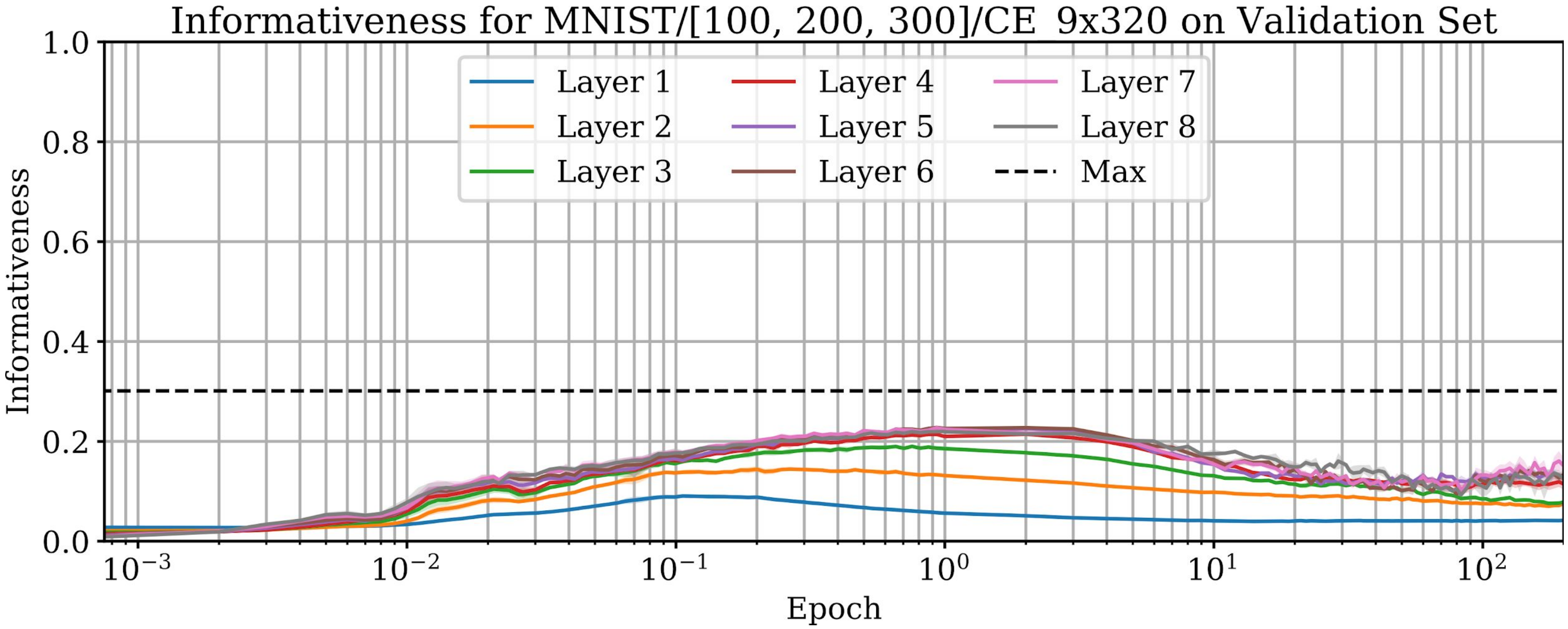
Results


Activation Fraction for MNIST/[100, 200, 300]/CE 9x320 on Validation Set
Activation Fraction
Epoch
0.0
0.2
0.4
0.6
0.8
1.0
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8

Results


Predictability for MNIST/[100, 200, 300]/CE 9x320 on Validation Set
Predictability
1.0
0.8
0.6
0.4
0.2
0.0
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Epoch
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8

Results

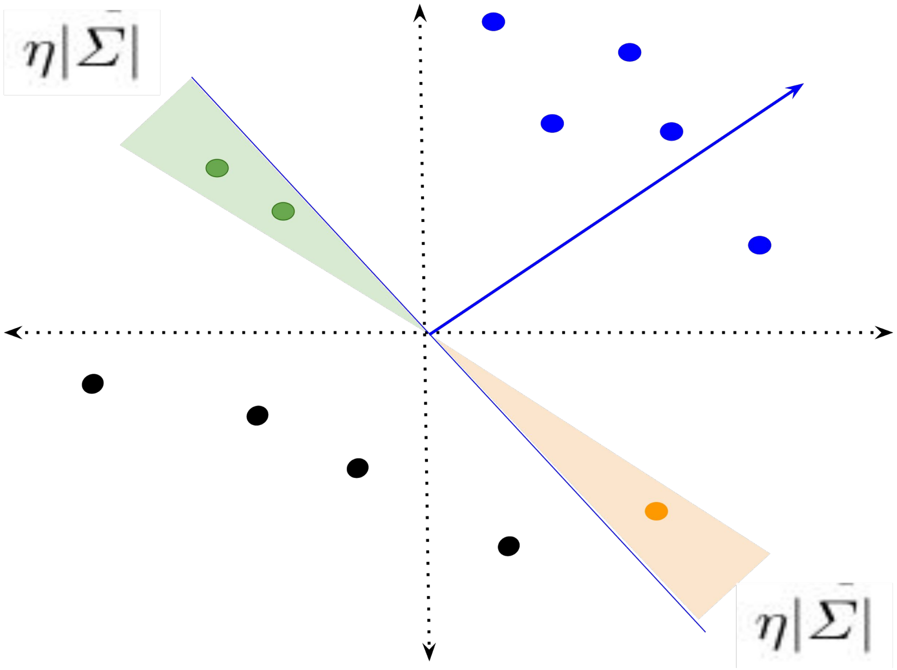

Informativeness for MNIST/[100, 200, 300]/CE_9x320 on Validation Set
Informativeness
Epoch
0.0
0.2
0.4
0.6
0.8
1.0
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Max

Interpretation

- Initial loss magnitude is large
- Sample sets grow until loss starts decreasing
- Sample sets shrink as loss decreases
- Samples with zero loss do not contribute to weight updates
- Training finds clusters of samples that cause coherent behaviour

Questions

Appendix



Appendix

- Predictability

$$p_{l,i} \approx 1 - \frac{1}{\log(2)} H\left(\hat{Z}_{l,i} \,|\, Y\right)$$

- Informativeness

$$u_{l,i} \approx 1 - \frac{1}{\log(|C|)} H\left(Y \,|\, \hat{Z}_{l,i}\right)$$