Understanding Individual Neuron Importance Using Information Theory

Erfan Mirzaei

Information Theory and Learning Course

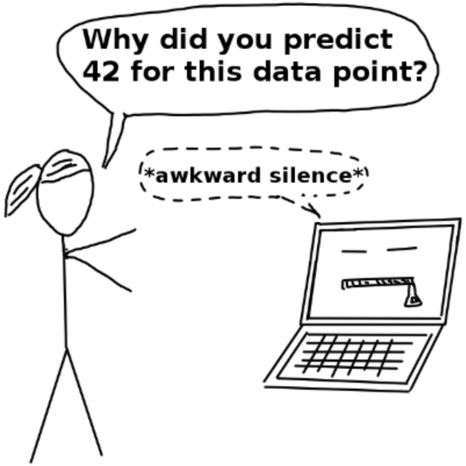
Feb 2022

Supervisors: Dr.Sabaghian, Dr.Shariatpanahi

Neural Networks Challenges

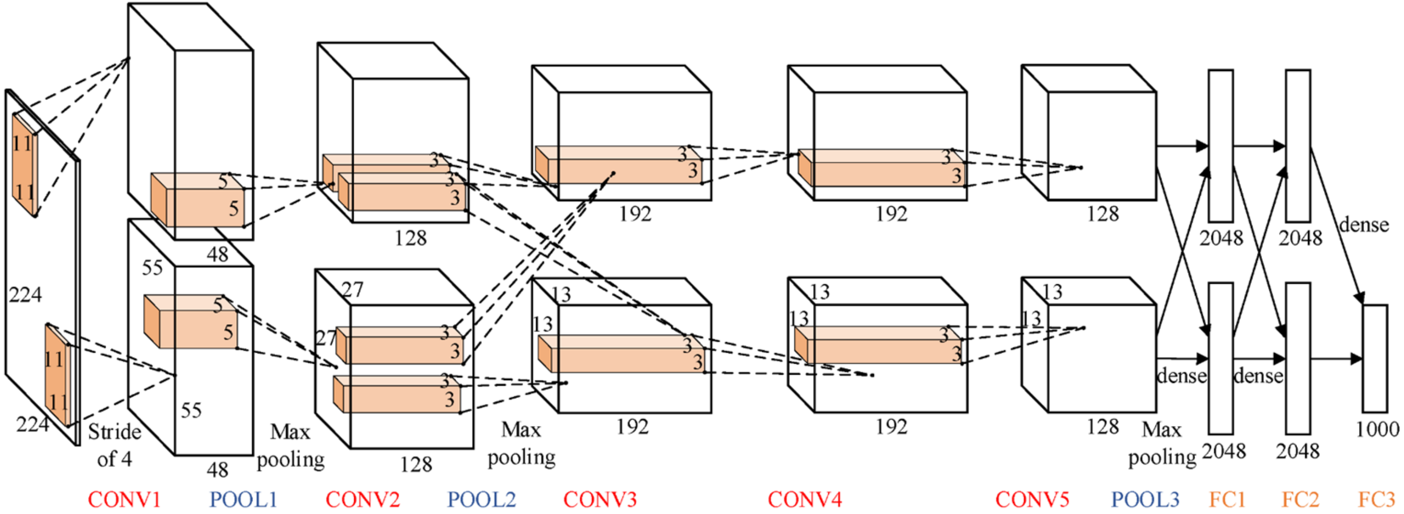
- Understand theoretically
- Understand functionality
- Interpretability of results

Interpretability of Deep Neural Nets



Neural Networks Challenges

- High Computational Complexity (Deep)



Zhang, Min, et al. "Optimized compression for implementing convolutional neural networks on FPGA." *Electronics* 8.3 (2019): 295.

Main Contributions

Pruning

+

Understanding NN

→ Cumulative Ablation

Use Information-Theoretic Measures for each neurons:

- Entropy (Variability of a neuron output)
- Mutual Information with labels(class information of a neuron output)
- KL Selectivity (class selectivity of a neuron output)

To investigate how these quantities connect with classification performance.

- Whole-Network Ablation [Remove neurons based on importance measures in whole-network]
- Layer-wise Ablation [Remove neurons based on importance measures in each layer]
- Bias Balancing [Instead of retraining]

General Background

- Entropy of a neuron:

$$H(T_j^{(i)}) = -\sum_{t \in \mathcal{T}} P_{T_j^{(i)}}(t) \log P_{T_j^{(i)}}(t)$$

- Mutual Information of a neuron and target:

$$I(T_j^{(i)}; Y) = H(T_j^{(i)}) - H(T_j^{(i)}|Y).$$

- KL Selectivity of a neuron:

$$\max_{y \in \mathcal{C}} \quad D_{\mathrm{KL}}(P_{T_j^{(i)}|Y=y} \| P_{T_j^{(i)}})$$

Problem Setting

- Classification via a feed-forward NN(|C| different classes).
- Labeled validation dataset D as follows:

$$\mathcal{D} := \{(x_1, y_1), \ldots, (x_N, y_N)\}$$ Where [N >> C]

- Neurons in each hidden layer use sigmoid as activation function:

$$t_j^{(i)}(x_\ell) = \sigma\left(b_j^{(i)} + \sum_p w_{p,j}^{(i-1)} t_p^{(i-1)}(x_\ell)\right)$$

- Quantize the output of each neuron to a finite set |T|.

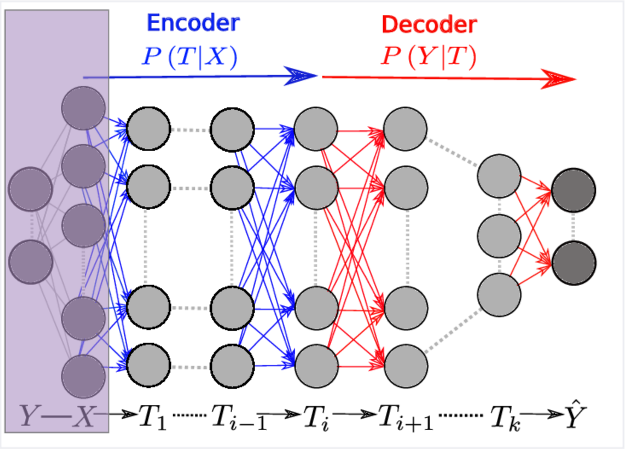
Related Works

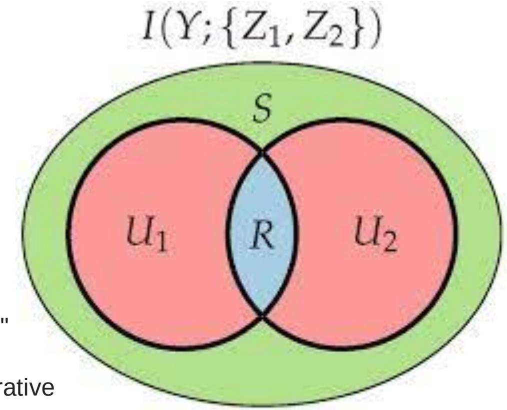
Other Methods

- Deconvolution
- Network Dissection
- Sensitivity Analysis
- Layer-wise relevance Propagation

Using Information Theory

- IB (Information Bottleneck)[1]
- PID(Partial information decomposition) [2]

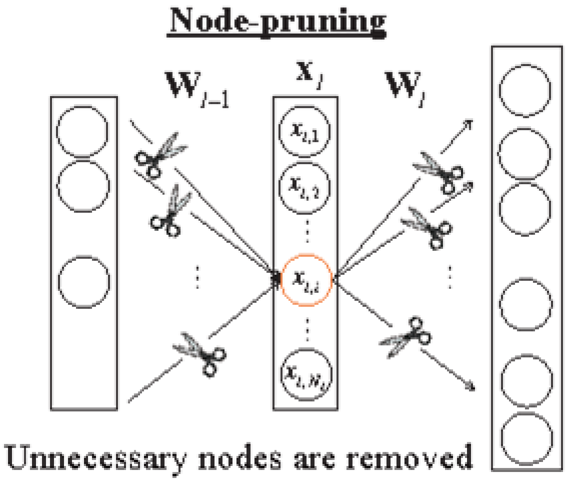




[1] Shwartz-Ziv, Ravid, and Naftali Tishby. "Opening the black box of deep neural networks via information." arXiv preprint arXiv:1703.00810 (2017).
[2] Tax, Tycho, Pedro AM Mediano, and Murray Shanahan. "The partial information decomposition of generative neural network models." *Entropy* 19.9 (2017): 474.

Related Works

- Weight matrices [weight pruning or low-rank approx.]
- Binary or ternary weights
- Pruning (neurons or filters)
- Merging



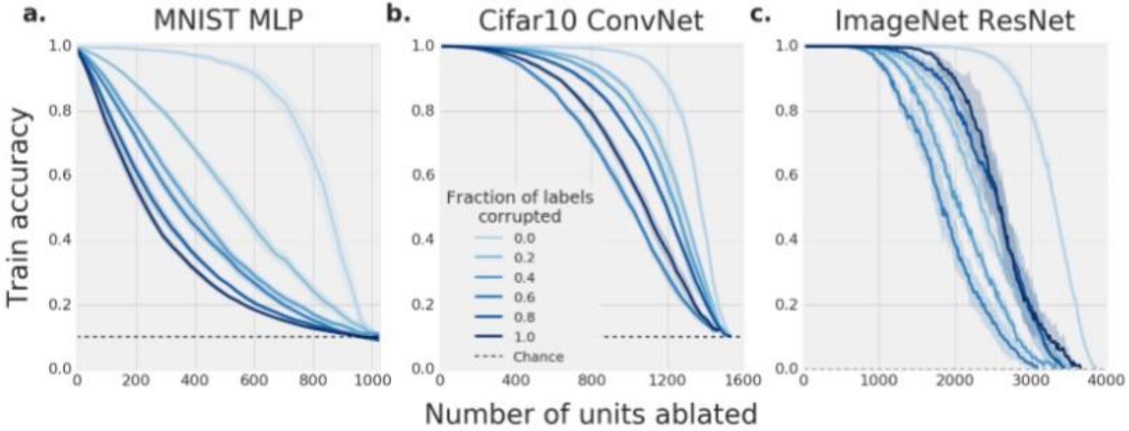
He, Tianxing, et al. "Reshaping deep neural network for fast decoding by node-pruning." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

Related Works

Cumulative Ablation across layers

Based on *Selectivity* of each neuron to different classes

Is **NOT a** good Indicator



Also Mutual Information is **Not** a good indicator

Morcos, Ari S., et al. "On the importance of single directions for generalization." *arXiv preprint arXiv:1803.06959* (2018).

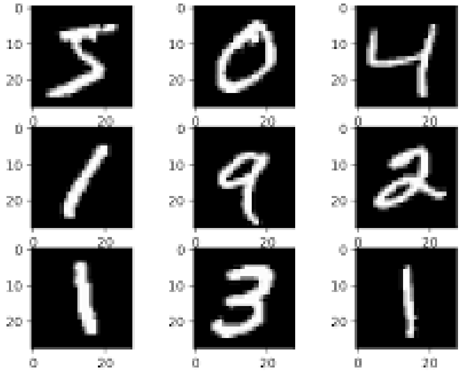
Experiments Setup

- Using a trained NN with 2 hidden layers with 200 neurons
- They apply one-bit quantization,i.e., $|T| = 2$ [sigmoid thresh = 0.5]
- MNIST Dataset [28 * 28 gray-scale images]:

  50,000 Train samples

  10,000 Validation samples

  10,000 Test samples

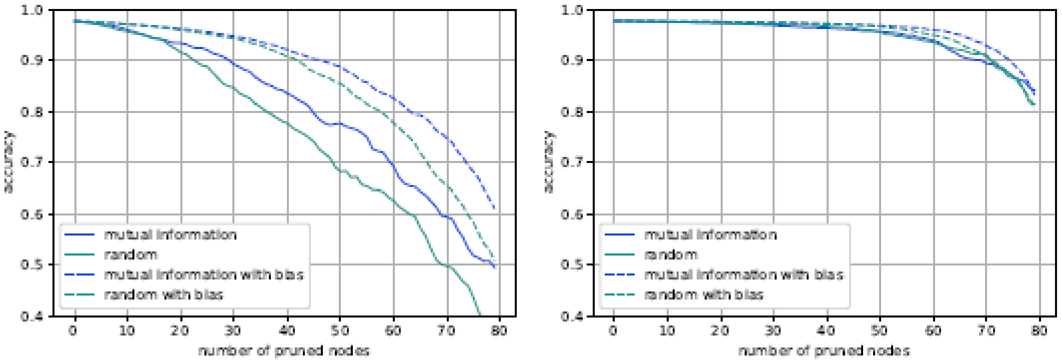


- Cross-entropy loss + L2 regularization
- Adam optimizer(lr =0.001, batch_size = 32)
- Bias balancing:

$$b_k^{(i+1)} + w_{j,k}^{(i)} \sum_{\ell} \frac{t_j^{(i)}(x_\ell)}{N}.$$

Experiments Results

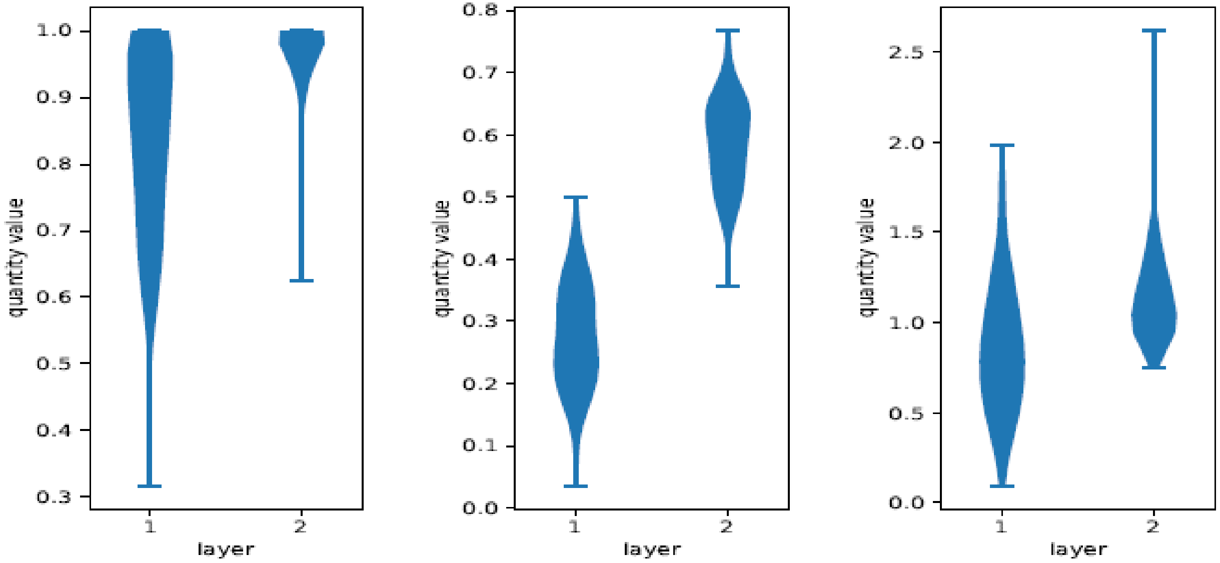
Classification Performance with and without bias balancing:



(a) First hidden layer (100 neurons) (b) Second hidden layer (100 neurons)

Experiments Results

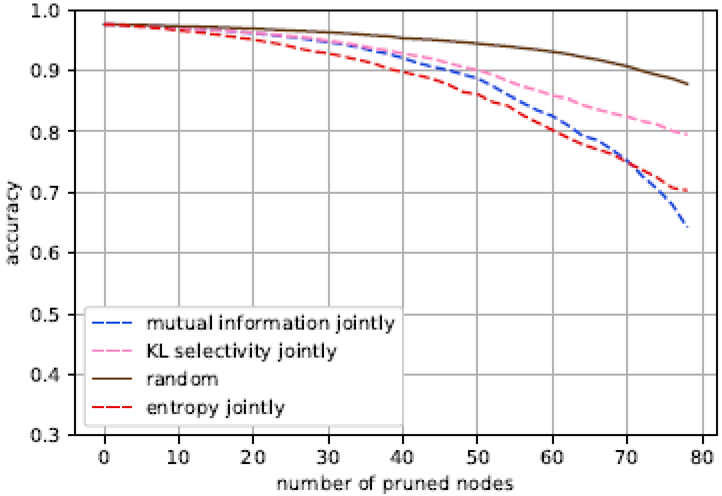
Dependence of Importance Measures on Layer Number



(a) Entropy (b) Mutual Information (c) KL Selectivity

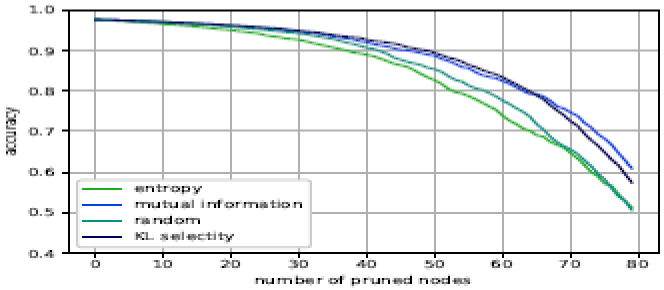
Experiments Results

Effect of cumulative ablation across all layers on classification performance:

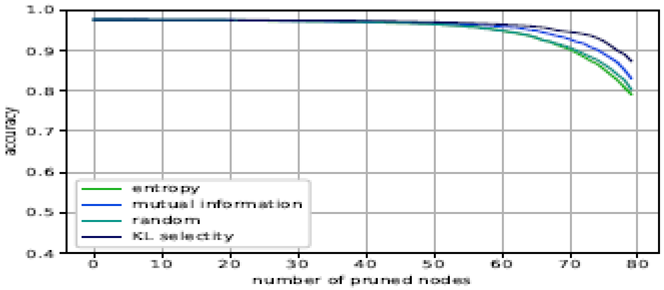


Experiments Results

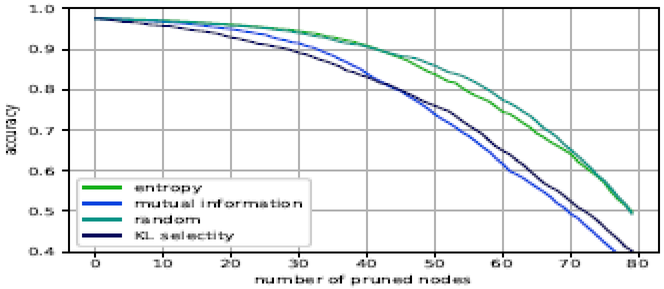
Effect of cumulative ablation across all layers on classification performance:



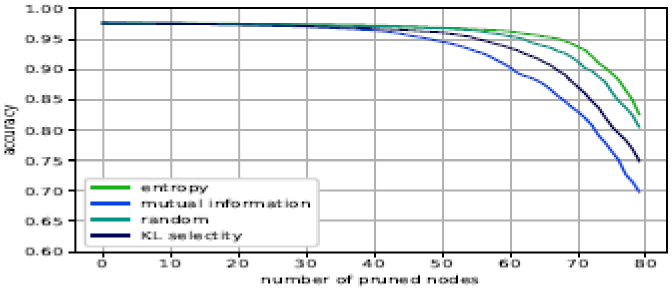
(a) First hidden layer, low values first



(b) Second hidden layer, low values first



(c) First hidden layer, high values first



(d) Second hidden layer, high values first

Discussion of results

- The distribution of importance measures changes from layer to layer.
- Therefore seems ill-advised to compare the importance of neurons of different layers.
- In deeper layers, ablation has smaller effects on classification performance.
- Deeper layers, KL selectivity seems to be the most adequate importance measure.
- Class-dependent importance measures, such as MI or KL selectivity, are connected more strongly to classification performance than class-independent ones, such as entropy.
- The connection between importance measures and classification performance depends on the activation function, as does the benefit of bias balancing.

Critique / Limitations / Open Issues

- Small and limited dataset was examined.
- They experimented just on shallow feed-forward NN, not CNN, or RNN.
- They didn’t examine different activation functions in the main results.
- The number of repetitions was low.
- Information-theoretic importance functions depending on the distribution of an individual neuron output are not sufficient. [Counter examples]
- Partial information decomposition may be used to shed more light on the behavior of neural networks.

Thanks for your attention

Erfan Mirzaei
erfunmirzaei@gmail.com