Clustering Phenomena in Dropout

David Kewei Lin
linkewei@stanford.edu

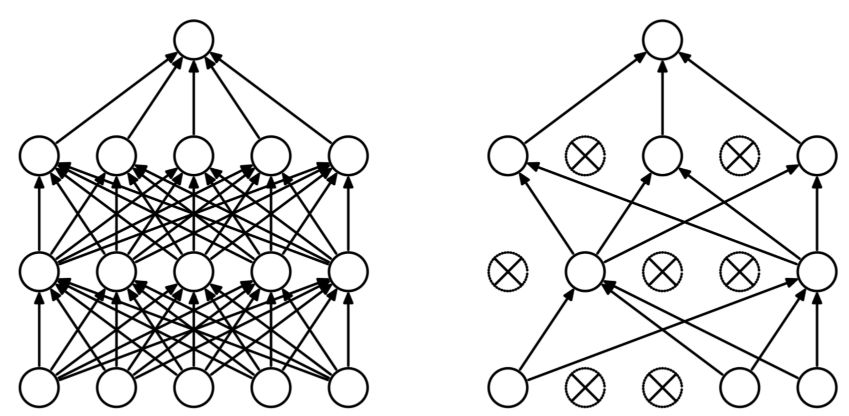
Jensen Jinhui Wang
wangjh97@stanford.edu

Phil Chen
philhc@stanford.edu

CS 229 (Spr 2019) Final Project
Mentor: Anand Avati

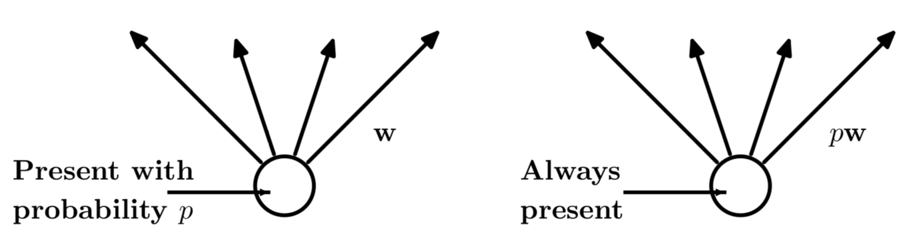
Introduction

Dropout is a regularization technique introduced in (Srivastava et al., 2014).



(a) Standard Neural Net (b) After applying dropout.

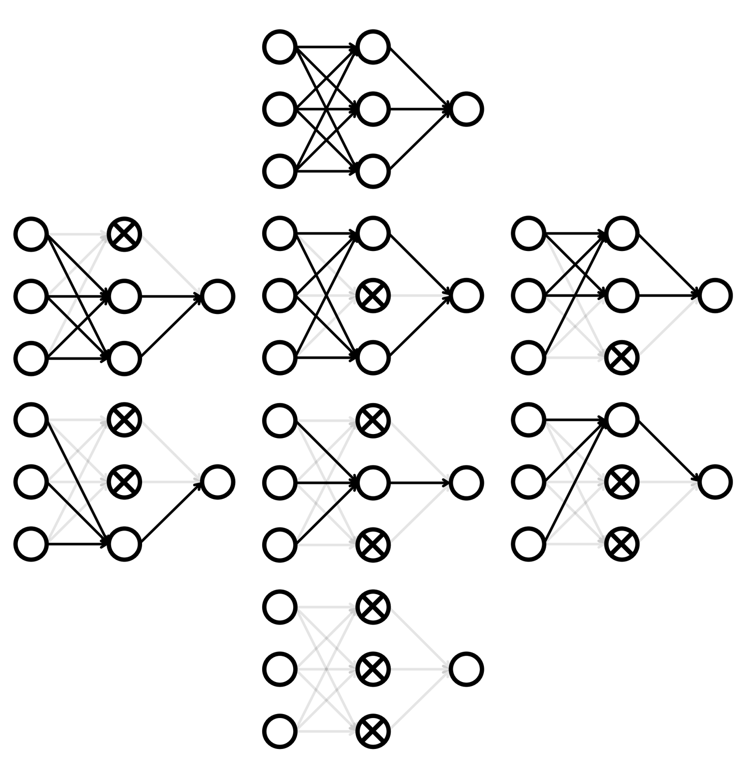
During training time, nodes will be disabled at random with a fixed probability 1-p. $(1 - p \approx 0.2 - 0.5)$



(a) At training time (b) At test time

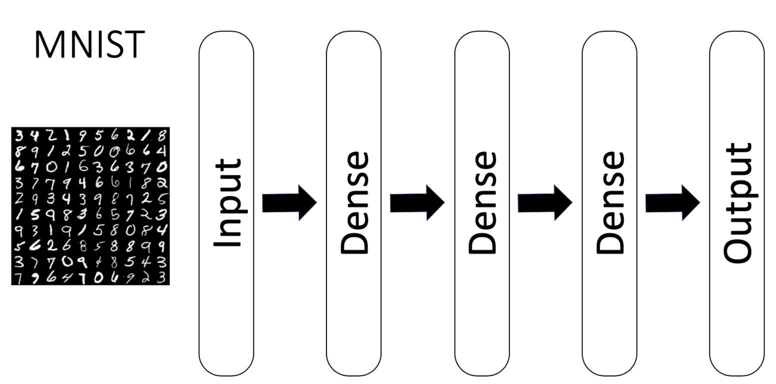
During test time, the node output is multiplied by p so that the expected outputs match up.

Framework / Setup



- Interpretation as Bayesian Neural Net
  - Model drawn from Dropout distribution
  - Training equivalent to variational inference for Bayesian neural networks (Gal and Gharamani, 2015)
  - Latent variables are the usual weights
- Dropout test-time protocol is a *linearization assumption* (scaling by Dropout factor)
  - Alternative: Monte Carlo integration (sampling)

MNIST



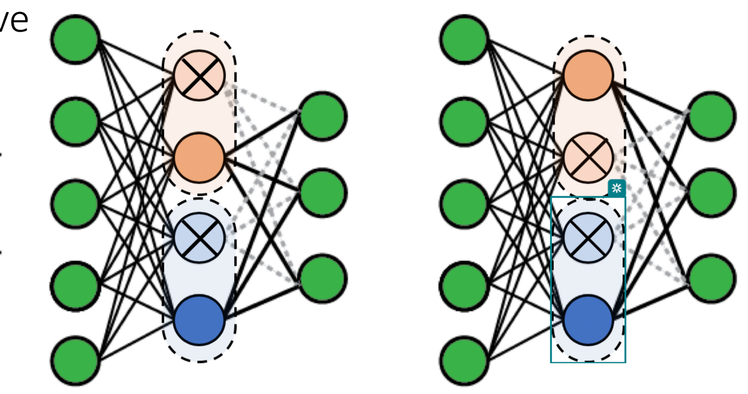
- Testing setup
  - MNIST dataset with basic 3-layer feedforward neural network.

Clustering

Idea 1 – Dropout "clusters" neurons

Result: High rand index between consecutive weight matrices

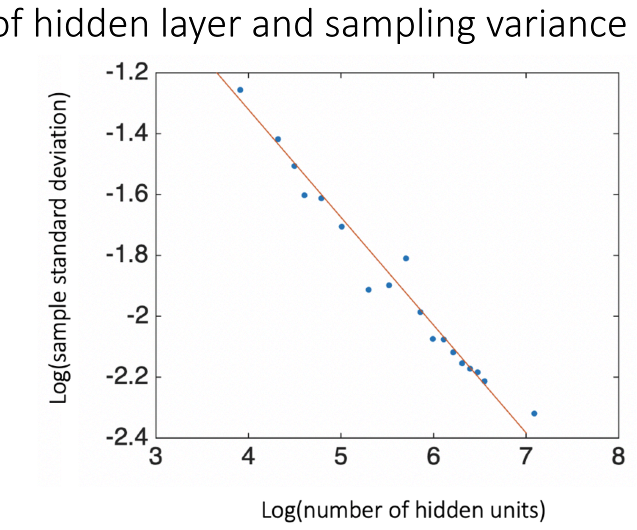
Dropout	0.0	0.2	0.4	0.6	0.8
Rand Index	0.65	0.81	0.77	0.82	0.84



Idea 2 – Clustering can predict sampling variance

Result: almost exact power relation between size of hidden layer and sampling variance

$$\mathrm{Var}(\hat{y}) \propto \mathbb{E}\left(\frac{\sum_{i=1}^{h} X_i}{h}\right)^2 = \frac{1}{h^2}\mathbb{E}\left(\sum_{i=1}^{h} X_i^2 + 2\sum_{j\neq k} X_j X_k\right)$$
$$= \frac{1}{h} + \frac{2}{h^2}\mathbb{E}\left(\sum_{j\neq k} X_j X_k\right) = \mathbb{E}_{j\neq k} X_j X_k + O\left(\frac{1}{h}\right)$$
$$\propto \frac{k}{h} + O\left(\frac{1}{h}\right) \propto h^{\kappa-1}$$



Model Compression

Idea: if dropout clusters neurons together, use "centroids" of weight matrix to compress network

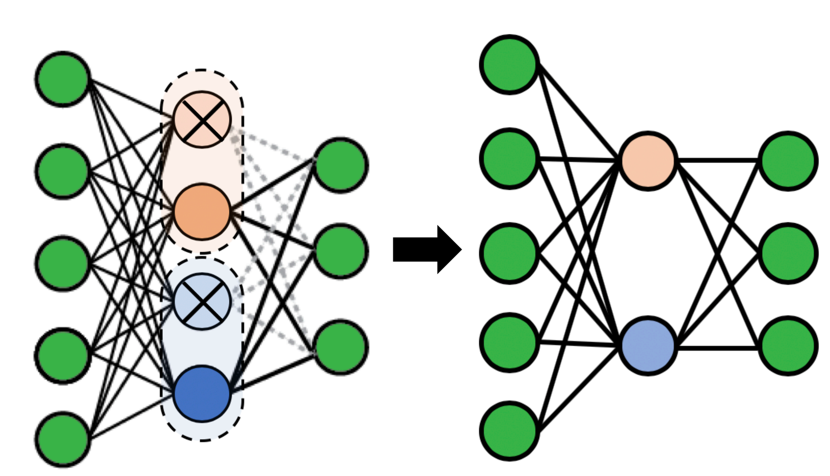
Results

Units	Test Acc	Compressed Units	Compressed Test Acc
400	0.869	162	0.855
500	0.879	162	0.847
600	0.884	167	0.833
700	0.877	177	0.839

Units	Test Acc	Compressed Units	Compressed Test Acc
400	0.882	162	0.740
500	0.894	162	0.685
600	0.892	167	0.534
700	0.891	177	0.657

Advantages
- Retains original model features
- Easy computation
- Low variance

Disadvantages
- Slightly lower accuracy



Bias-Variance Tradeoff

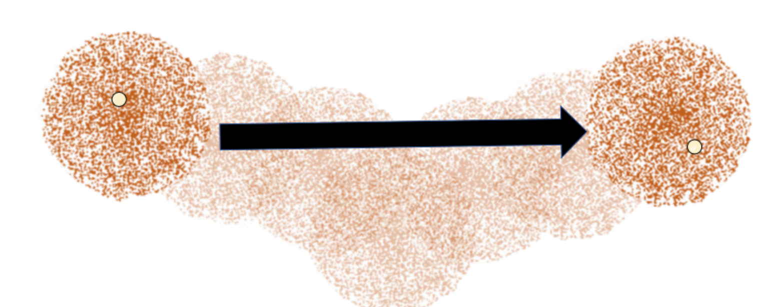
Findings
- Dropout decreases model variance
- Sampling at test time further decreases model variance

Why does sampling at test time decrease variance?

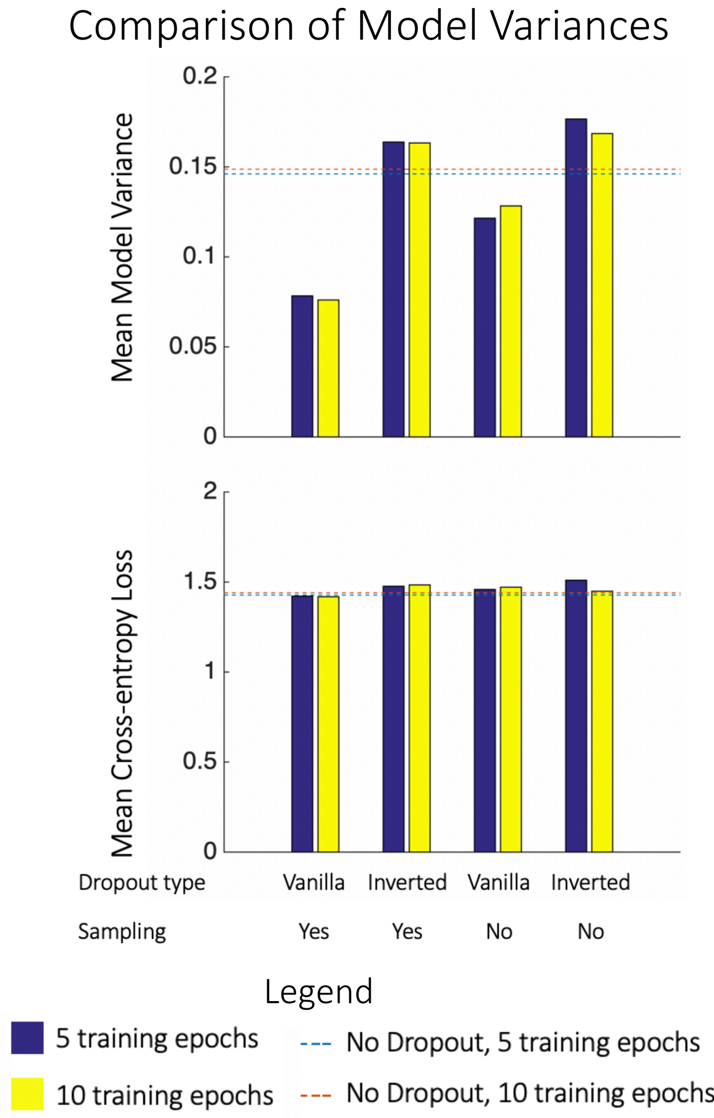
Electron cloud model

Model family	*Electron cloud*
Sampling	*Superposition of states*
Inverting at test time	*Collapsing wavefunction*

On average, the electron cloud moves less than any one "component" of the cloud



Comparison of Model Variances



Summary/Future Work

Summary
- Dropout clusters weights of hidden layers
- These clusters can predict trends for both sample variance and model variance
- Using the centroids of these clusters, we presented a model compression algorithm with strong results

Future work
- Unified model to explain both sample and model variance through clustering
- Considering both sets of weights simultaneously for model compression
- Theoretically-motivated hyperparameters for model compression

References:
[1] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.
[2] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158, 2015.
[3] William M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336):846–850, 1971. doi: 10.1080/01621459.1971.10482356

Both graphics used in the "Introduction to Dropout" section are from [1].