Do Deep Neural Networks Learn Shallow Learnable Examples First ?

Karttikeya Mangalam, Vinay Uday Prabhu
Email: mangalam@stanford.edu

Motivation

What characterizes the generalization process of a deep learning network as training progresses?

- Generalization error decreases first then overfitting sets in
- U-shaped test error curve explained by Bias-Variance tradeoff [1]
- DNNs learn simple patterns first before memorizing [2]
- Input domains consist of a subsets of both task relevant and task irrelevant information and representations first learn to effectively compress the task irrelevant information [3]

Core Questions Investigated

- Is the notion of *easiness* for classification same for models with as different parameterizations and architectures as classical machine learning models and deep networks the same? And hence is largely related to the example independently of model?
- As training progresses, is there a shallow learnable to deep learnable *regime change* viewed through the test set?
- Are there examples that are shallow learnable but for some reason a DNN with a far better overall accuracy fails to classify?

Datasets & Models

❖ Datasets:
To study the phenomenon on a wide range of examples we perform experiments on:
● MNIST ● CIFAR10 ● CIFAR100

❖ Classical Machine Learning Models:
To compare the learning process against different classical machine learning models we use the following models:
● Support Vector Machine (RBF Kernel) ● Random Forests

❖ Deep Learning Models:
We choose diverse network architectures to account for different inductive biases like skip connections, dense networks etc. and also according to the dataset simplicity and size. With these considerations, we study the generalization process of the following three deep learning networks:

● 2 layer Convolution Neural Network (MNIST)
● DenseNet 121 (CIFAR10)
● ResNet 101 (CIFAR100)

Note that each DNN is compared against both the ML models.

Experimental Procedure

Tracking the Learning Process

Traditionally, generalization performance on a held out set is tracked.

Given models **M** and **D** we propose to keep track of the contingency matrix **T** as training of **D** progresses. Several other interesting metrics are obtained from **T**

	M incorrect	**M correct**
D incorrect	T_{00}	T_{01}
D correct	T_{10}	T_{11}

❖ Accuracy
Accuracy of models **D** and **M** can be found simply as:

Accuracy (M) = $\frac{T_{01}+T_{11}}{T_{11}+T_{00}+T_{10}+T_{01}}$ Accuracy (D) = $\frac{T_{10}+T_{11}}{T_{01}+T_{11}+T_{10}+T_{00}}$

❖ Marginal Accuracy
Accuracy of **D** on subsets that M classifies correct (R_+) & incorrect (R_-)

$$R_+ = \frac{T_{11}}{T_{11}+T_{01}} \qquad R_- = \frac{T_{10}}{T_{10}+T_{00}} \qquad R_{\pm} = \frac{R_+}{R_-}$$

❖ Ratio of Accuracies
Ratio of marginal accuracies $R_{\pm}$ is also obtained which serves as a measure of how the correctly classified by **D** overlap with the those classified by **M**.

Results & Observations

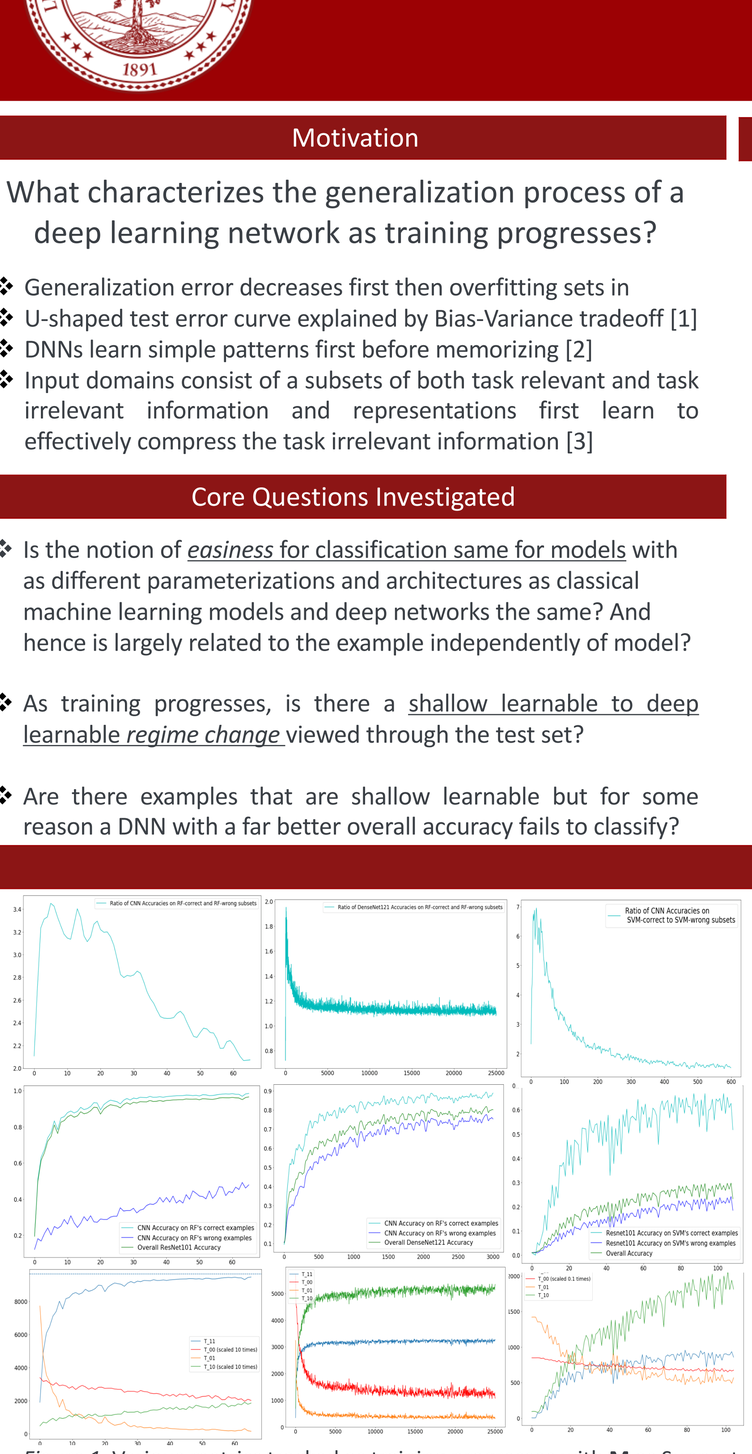


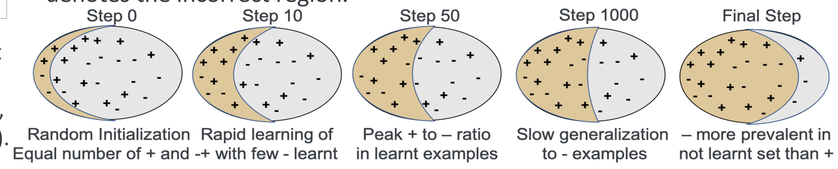
Figure 1. Various metrics tracked as training progresses with **M** as Support Vector Machine . Plots of $\boldsymbol{R_{\pm}}$ (Top Row), Marginal Accuracies ($\boldsymbol{R_+}$, $\boldsymbol{R_-}$) (Middle Row) and **T** (Bottom Row) on the pairs of {MNIST , CNN} (Left Col), {CIFAR10, DenseNet121} (Middle Col) & {CIFAR100, ResNet101} (Right Col).

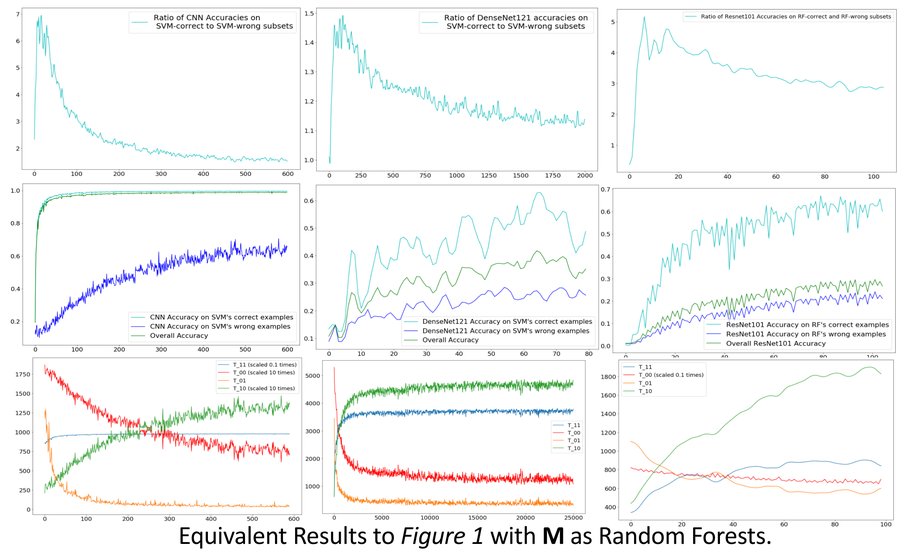
❖ Key Observations:

- $R_{\pm}$ has a right skewed unimodal shape. Of the two subsets of testing data, **M**-correct and **M**-incorrect were completely irrelevant for generalization process of **D**, $R_{\pm}$ would stay identically at 1.
- Instead, the observed peak indicates that **D** learns **M**-correct examples much earlier in the training than **M**-incorrect. Then slowly over the epochs generalized to *harder* **M**-incorrect set.
- Plots of R_+, R_- (middle row) validate this observation where R_+ can sometimes be sometimes be as high as **60%** where the overall accuracy is still only **20%** and R_- is still around **15%.**

Conclusion

The following infographic succinctly expresses our findings. The Oval denotes the entire test set littered with + and – which denote **M** correct and incorrect examples. Finally, golden color denotes the region **D** classifies correctly and gray denotes the incorrect region.





Equivalent Results to *Figure 1* with **M** as Random Forests.

Relevant Previous Work

[1] Vapnik, V. N. Statistical learning theory. Adaptive and learning systems for signal processing, communications and control series, 1998.
[2] Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 , pp. 233– 242. JMLR
[3] Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A.,Tracey, B. D.&Cox, D. D. On the information bottleneck theory of deep learning.