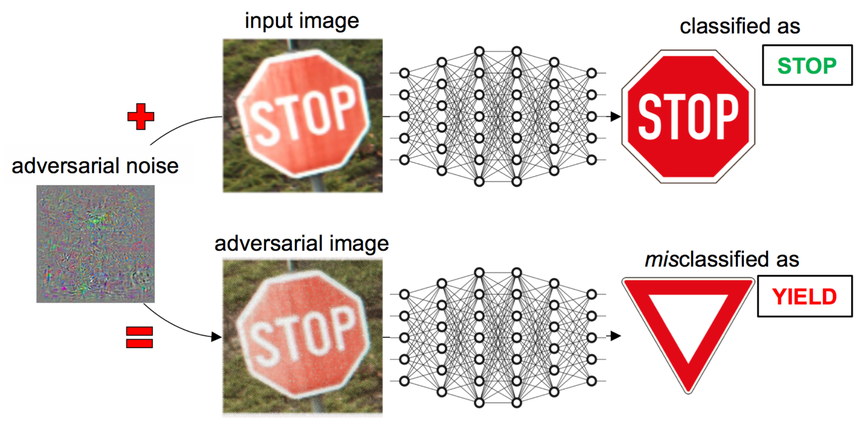
Uncertainty Aware Small Data Neural Networks

Ankur Mallick[1],
joint work with Bhavya Kailkhura[2], Brian Gallagher[2], Yong Han[2]
[1] Carnegie Mellon University, [2] Lawrence Livermore National Laboratory

Abstract

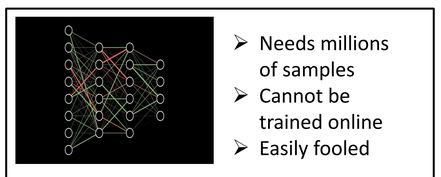
This work deals with learning in high dimensional settings with very few labelled examples where conventional neural networks have a high tendency to overfit and predict incorrect outputs with a high confidence. It is preferable to make predictions with a lower degree of confidence i.e. higher uncertainty, in regions where we do not have enough labeled points. We propose 'Small Data Neural Networks' which compress the high dimensional data to a low dimensional latent space and then performs prediction on the latent space outputting predicted values along with the corresponding uncertainty estimates.



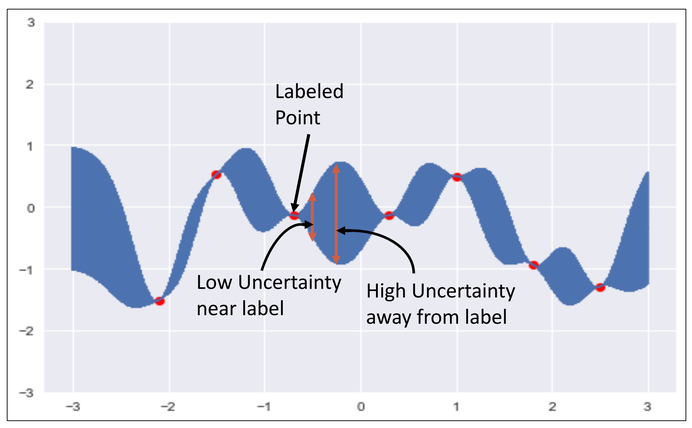
"I don't think it's how the brain works, We clearly don't need all the labeled data."
– Geoffrey Hinton

Motivation

- Conventional Neural Networks need a huge amount of labeled data (which is expensive to obtain) to learn an accurate prediction model
- In the absence of enough labeled data they make incorrect predictions with a high degree of confidence
- **Sample Efficiency** and **Uncertainty Quantification** are two of the biggest challenges in making machine learning models more reliable



Brain v/s AI



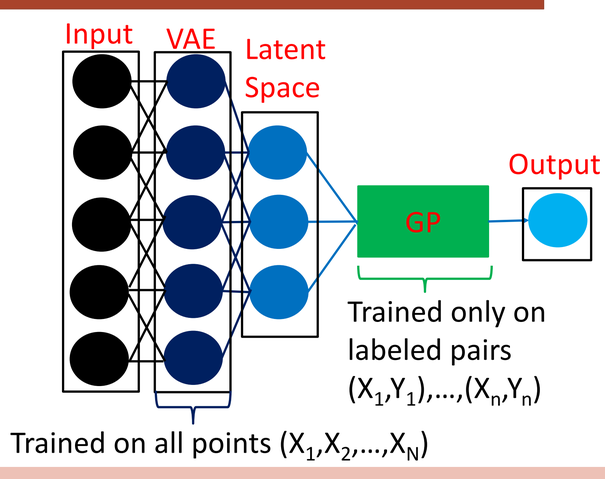
Desired Uncertainty Quantification

What we want: The Holy Grail

- A predictive model that can generalize well from small amounts of labeled data
- Mathematical expressions for the uncertainty of the model in regions of the data space that do not have adequate labels
- The ability to explore regions of the data space with high uncertainty to improve the performance of the model in these regions

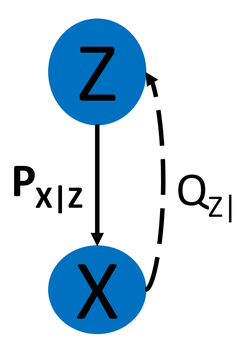
What we have: Small Data Neural Nets

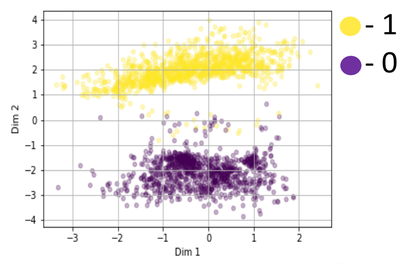
- **Input:** $(X_1,X_2,\ldots,X_N) \in \mathbb{R}^D$
- **Labels:** $f(X_1) = y_1,\ldots,f(X_n) = y_n$ (n<<N)
- **Output:** Probabilistic estimate of labels of 'N-n' unlabeled points i.e. $P(f(X_i) = y)$, i=n+1,…,N
- **Algorithm:**
  Step 1: Learn a low dimensional latent space i.e. transform $X_1 \rightarrow Z_1$, $X_2 \rightarrow Z_2$,…, $X_N \rightarrow Z_N$ where $(Z_1,Z_2,\ldots,Z_N) \in \mathbb{R}^d$ (d<<D)
  Step 2: Train a Gaussian Process (GP) [1] model on the low-dimensional representation of the 'n' labeled points and use it to predict labels for the 'N-n' unlabeled points



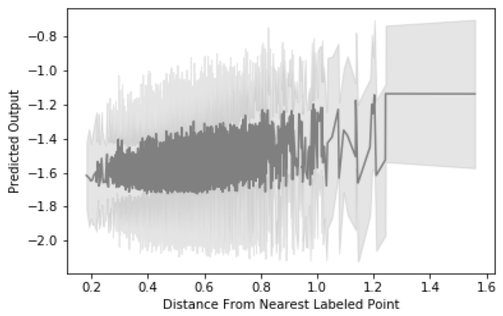
Step 1: Deep Unsupervised Learning

- Most of the information in real data can be expressed using a few variables and the actual data can be viewed as a (non-linear) transformation of these variables [3]. So if we can learn this transformation we can reduce the problem of labeling the D-dimensional data to the (much) easier problem of labeling the d-dimensional data (d<<D)
- We use Variational Autoencoders (VAEs) [2] to model the low dimensional representation of high dimensional data in a probabilistic fashion i.e. for each data point 'X' they output the probability of 'X' having been generated from every latent point 'Z'





Latent Space for MNIST

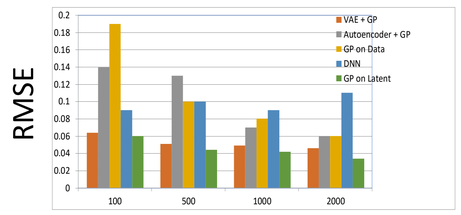


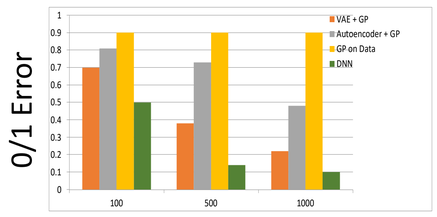
Step 2: Uncertainty Aware Prediction

- GPs have much fewer parameters than DNNs and also output probabilistic estimates of labels for unlabeled points thus providing a direct quantification of the uncertainty in unlabeled regions
- Since the output of the VAE is a random variable and GPs are conventionally trained on deterministic inputs we designed a GP architecture with flexible kernels that can accept random inputs following any (unknown) distribution

Experiments

Type of Expt.	Regression	Classification
Input Data	Random Points in 10D Space	28X28 MNIST Handwritten Digits
Latent Size	6	10
Target	Hartmann 6 Function on Latent Space	Digit Label (0-9)
Total Points	60000	60000
Error Metric	RMSE	0/1 Error
Test Points	5000	10000

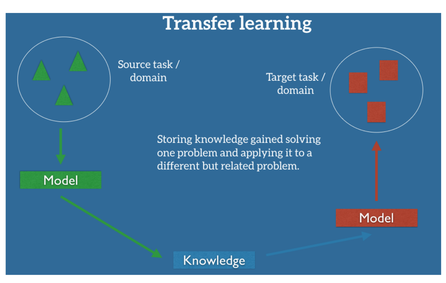


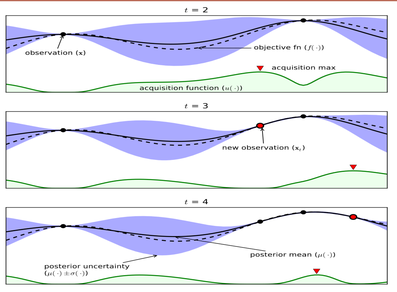


Our scheme requires 20x fewer samples for the same RMSE (6%)

Observations and Conclusion

- The proposed VAE + GP model for learning outperforms conventional DNNs on a regression task over a 10-Dimensional dataset
- In addition to the superior accuracy our model can provide fully probabilistic estimates of both the latent variables and the target variables
- This can be extended to tasks like transfer learning/domain and Bayesian Optimization (Future Work)





IM Release Number: LLNL-POST-756056

References:

[1] Rasmussen, Carl Edward. "Gaussian processes in machine learning." *Advanced lectures on machine learning*. Springer, Berlin, Heidelberg, 2004. 63-71.
[2] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).
[3] Kailkhura, Bhavya, et al. "A Spectral Approach for the Design of Experiments: Design, Analysis and Algorithms." *arXiv preprint arXiv:1712.06028* (2017).