RISE

Attracting and Developing Top Level IT and R&D CS/CE/Cyber Professionals

SAND2017-7457D

SPIRE

Clustering and Deep Learning of SAND Reports

Mark Louie, University of New Mexico

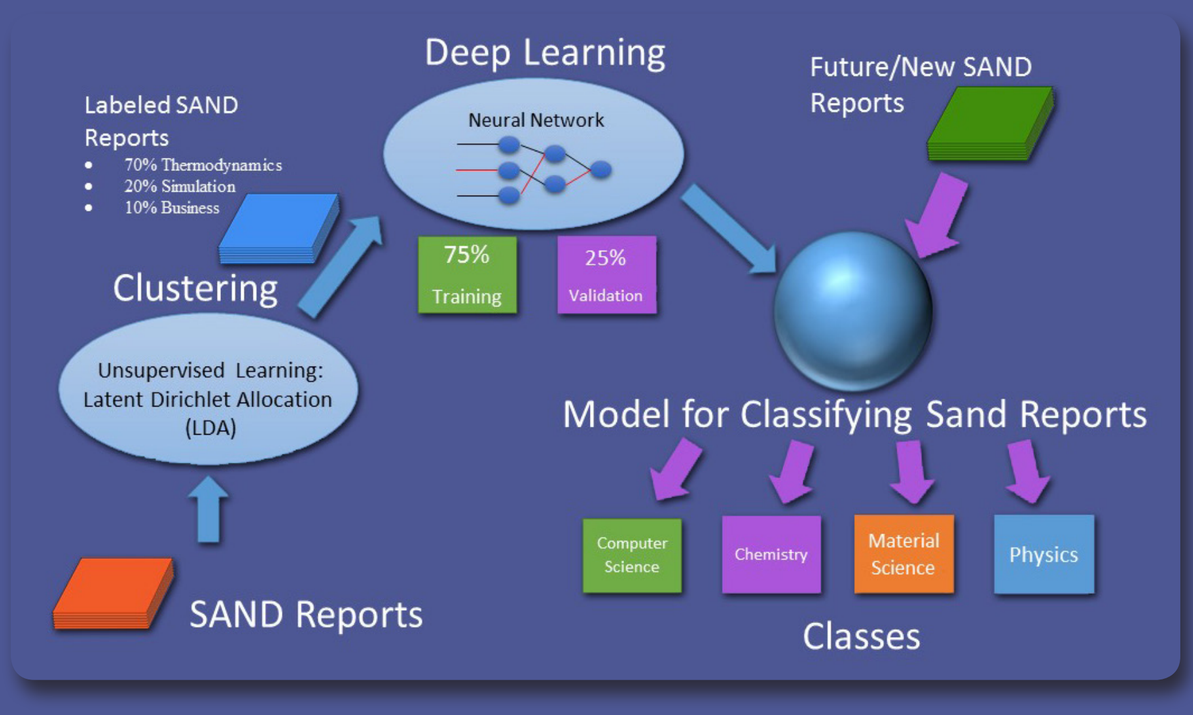
Project Mentors: Pengchu Zhang and John Herzer, Org. 10737

Abstract:

The objective of this project is to optimize a combination of machine learning and deep learning to automate the classification of SAND reports in the parent project SPIRE (Sandia Personalized Information Retrieval Environment). The scope of this project encapsulates the clustering and the creation of a model to classify SAND reports. SAND reports were cleaned and clustered to create a dataset for model training and validation. The parameters in the neural network are adjusted to find the optimal accuracy with minimizing the faults of overfitting and underfitting.

Introduction:

- **Problem:**
  - Sandia has over 140,000 SAND documents spanning from the 1950s to the present.
  - Thousands of SAND documents have not been cataloged yet and doing so is labor intensive.
- **Solution:**
  - Use Topic Clustering on SAND reports to create a dataset.
  - Use clustered dataset to train deep learning model to automatically classify SAND reports.
  - High level topology illustrated below:



Methodology:

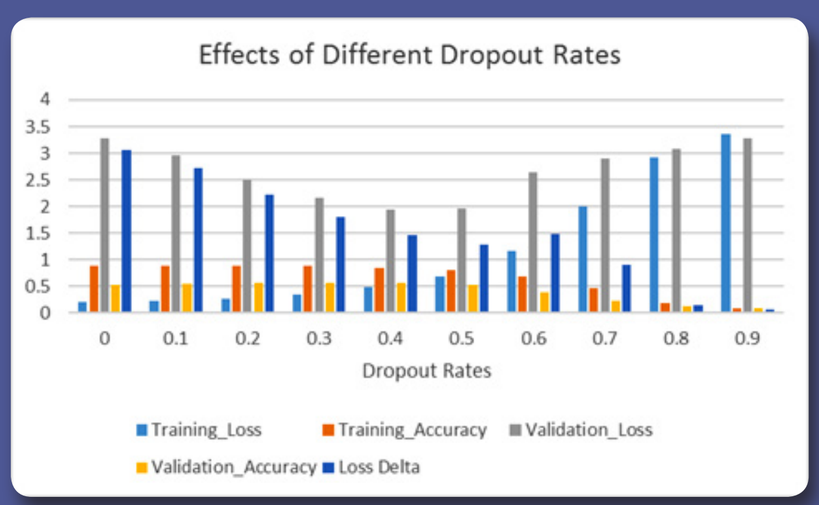
1. Clean 90,000 SAND reports to prepare for clustering.
2. Use Latent Dirichlet Allocation (LDA), a machine learning clustering algorithm, to cluster reports in topics relevant to the contents of each report.
3. Utilize Sandia's term taxonomy to label topics based on the words associated within a topic.
4. Split dataset into 75% training data and 25% validation data.
5. Create model by using deep learning libraries and tweak parameters to optimize accuracy and minimize overfit and underfit.

Testing Methodology:

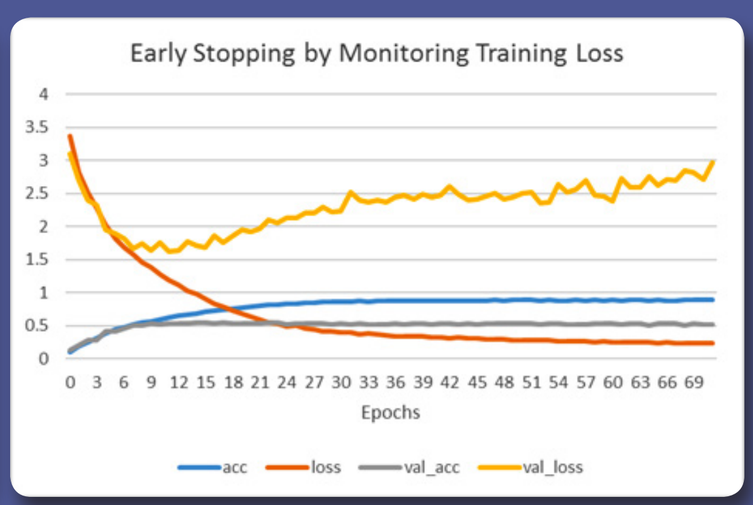
1. To test the effects of dropout, a test was conducted with a fixed number of epochs of 100 and five points of dropout. The dropout rates varied from 0 to 0.9 with increments of 0.1.
2. The testing methodology for early stopping uses a built-in function to monitor the loss parameter and terminate training if loss does not decrease after successive iterations.
3. The final test is to repeat early stopping with monitoring the validation accuracy parameter except the training terminates when validation accuracy no longer increases.

Results:

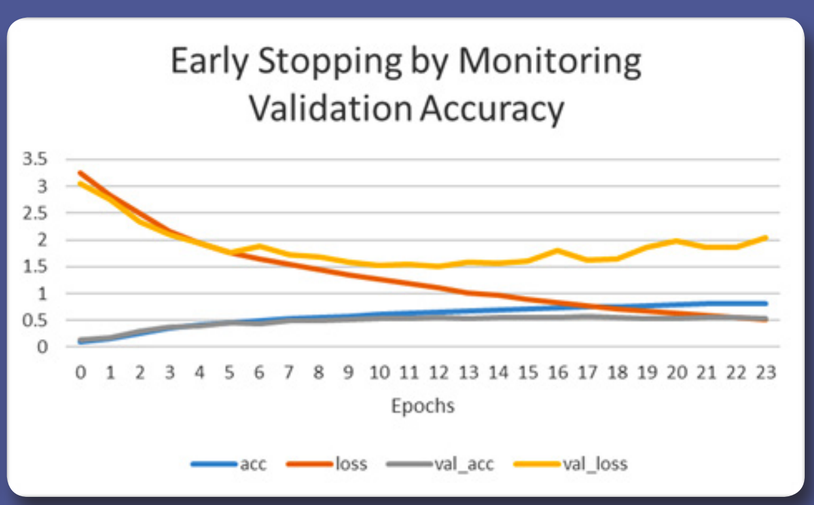
Test 1: The results of the test involving dropout rates for 5 points of dropout is described in the figure below.



Test 2: The figure below describes the results for early stopping of training by terminating training if Training Loss no longer decreases.



Test 3: The figure in below illustrates the results of the third test involving early stopping on the parameter Validation Accuracy.



Discussion:

From all three tests, overfitting occurs prevalently despite the fact that Dropout and Early Stopping are ways to mitigate overfitting. No conclusions can be made with this data. More testing has to be done with tweaking Dropout and Early Stopping. This may include combining Early Stopping with Dropout.

Future Steps:

- Repeat testing of Dropout and Early Stopping with other Neural Network architectures such as: Recurrent Neural Network (RNN) and combining Convolution Neural Network (CNN) with RNN.
- Increase the size of dataset because the dataset may have too few data points.
- Explore different clustering algorithms other than LDA, perhaps K-Means.

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2017-XXXXC.

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration