Autoencoder – Meeting 2/16/2021

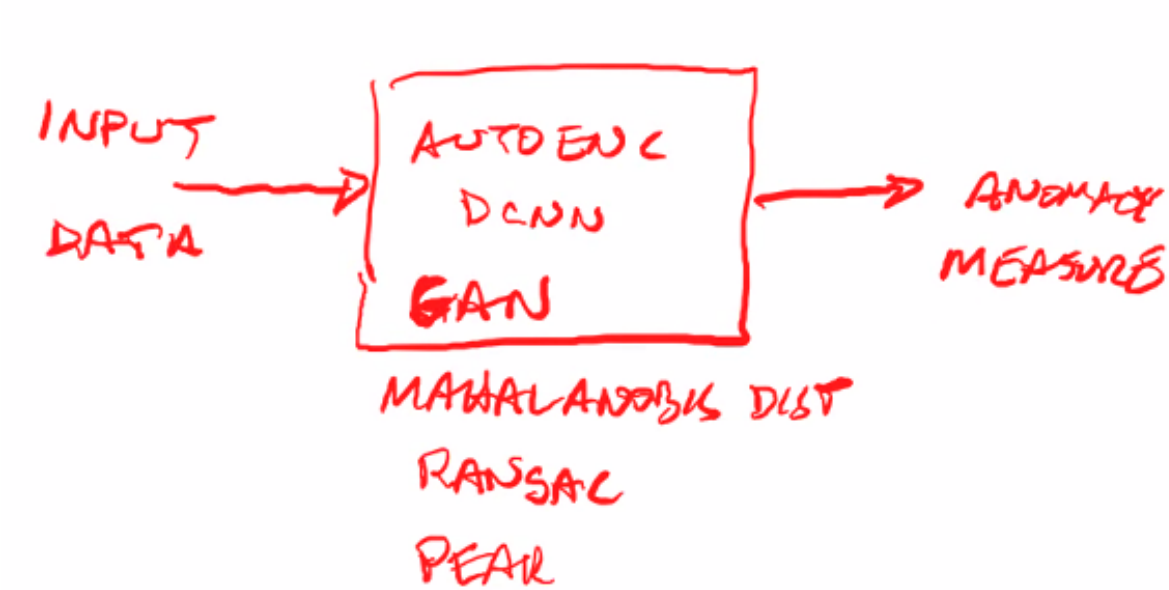
- Autoencoders are mainly used for labeling
    - Unsupervised learning, data is unlabeled, used for training purposes
- Needs a ground truth:
    - Something for the autoencoder to train on, so it has something to compare other data to
    - Ground truth data represents the normal conditions
- There is a hidden layer in this approach
- To run autoencoder on the Titan:
    1. Open Anaconda Navigator – "anaconda-navigator" on the terminal
    2. Go to Environments, select the tf environment
    3. Open Jupyter Notebook
    4. File is IoTAE
    5. Steps to ML
        a. Add CSV file for normal data
        b. Add CSV file for test data
        c. Normalize data (ln [42]) – everything is scaled between 0 and 1 so its all in the same range
        d. The model is in ln [50] – 3 layers with a hidden layer in the middle
        e. Model gets trained, the loss gets lowered (which is what we want)
        f. Plot the distribution of the loss, want a good distribution to determine a threshold
        g. To get better results, we can adjust the NUM_EPOCHS
    6. The reconstruction error/score is what we will use to separate anomalous data from normal data
    7. X_Test – first it is normal, then after you change it (because the model has been made) ONLY run lines 64 and 65

See comments – the script deals with one file that is split, we can get around this my removing that step

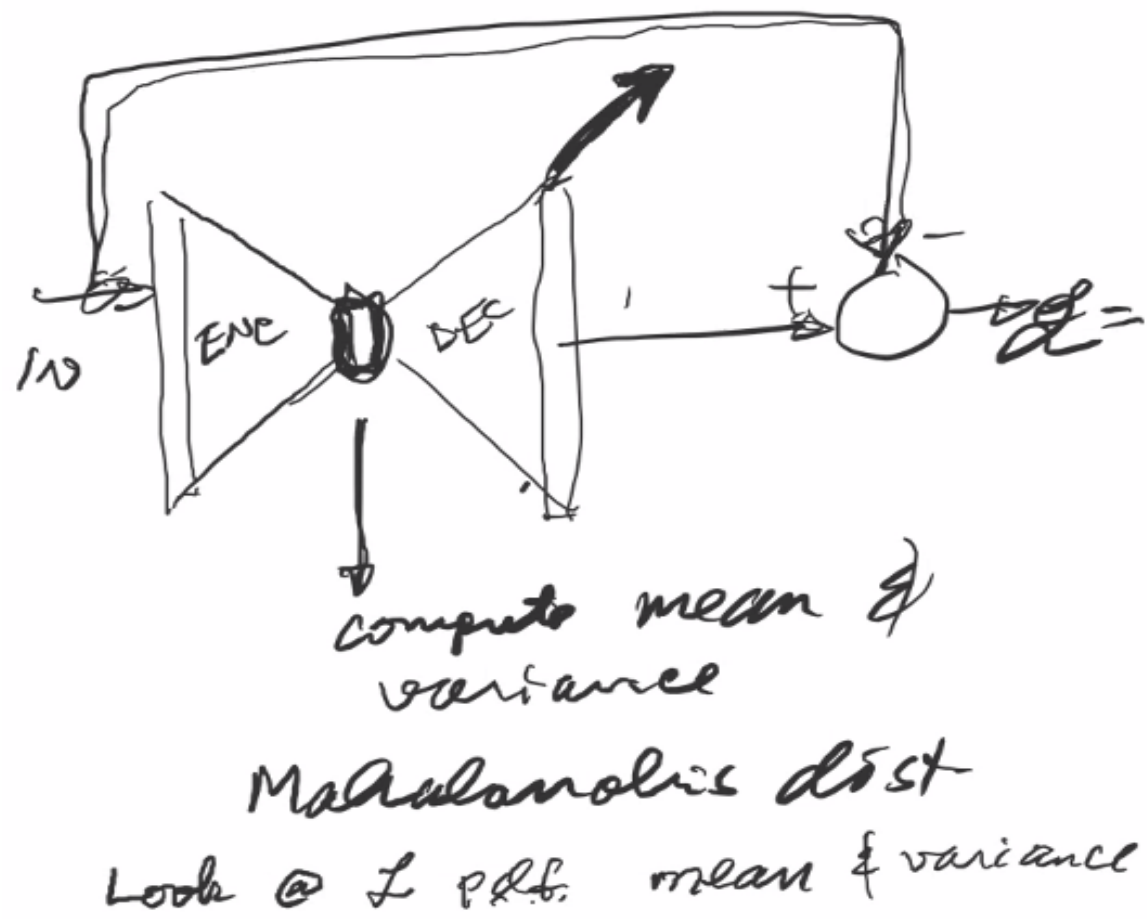
Should create our own python notebook file

Meeting 2/17/2021

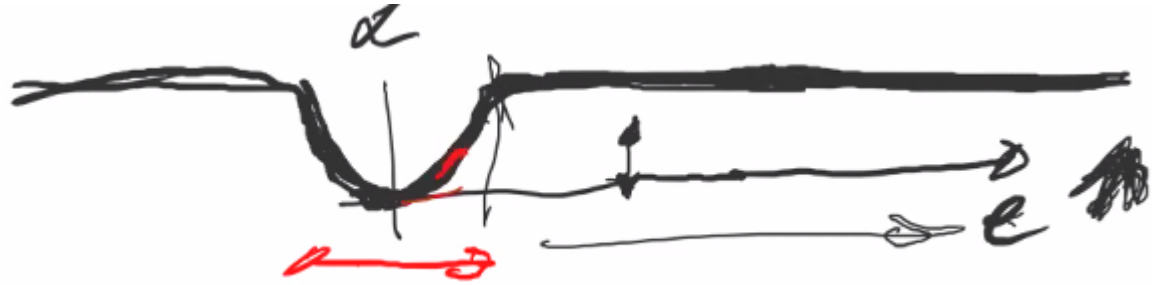
- Algorithms we will consider
    - Autoencoder
    - DCNN
    - GAN
    - Peak finding algorithm
    - RANSAC
- Algorithms go into a black box which takes in anomalous data and output flagged anomalies



-
- Autoencoder
    - Compute mean and variance
    - Mahalanobis distance between vectors
    - Our autoencoder looks at loss mean and variance



- Limit the error, anomalous data is very infrequent – can allow for some anomalies in training



-
- If the error is very far out, reject it
- Use a clipped MSE
- https://stats.stackexchange.com/questions/350211/loss-function-autoencoder-vs-variational-autoencoder-or-mse-loss-vs-binary-cross ???

- Peak finding
    - Medium filters – used to try to reject peaks
- We also want to be able to compare the results of each algorithm
- Dr. Pearlstein can provide us with more energy usage data
    - It is unlabeled though

TODO:

- Clean up COMNETS dataset
    - Remove bad columns
    - Convert everything to decimal – remove/parse out colons and extra commas between values in the same field
- Modify Adam's script
    - Take in two files separately instead of splitting one file into two (as discussed above and in code comments)
- Run the script with the COMNETS data, get output
- Meet with Raj, Adam, and Jitu to discuss results