Machine Learning for Data Quality Monitoring

Thomas Britton
Goal

• **Goal: Replace myself with an AI**
  
  – Further goals: Create a system which can monitor output from the detectors and flag problems early. Potentially pick out correctable issues (through calibration) and perform adjustments as needed
Tools and tactics

- Chose Keras with tensorflow backend
- Start by looking at the histograms, as .pngs (just like how I do it)
  - Use convolutional NN to do image classification
  - Take each classification/confidence and use an aggregator to classify the entire run/take action on flagged issues
Early results

- Started with one image (BCAL_occupancy)
- Low sample training more okay in this case
  - **No** need for image augmentation!
- Training set: 75% (randomly selected of the 2018-08 runs)
- Validation: the other 25%
- Grow the set by testing and labeling on other run periods
The Network

- My own: achieved \(~96\%\) accuracy.
- Needed to use a more sophisticated network
  - Introducing **inceptionV3**
    - Think of it like a network trying various convolutions and figuring out which is best
The Network

5x5 in GoogleNet (Inception-v1)

Inception Module A

Grid Size Reduction (with some modifications)

5x Inception Module A

4x Inception Module B

2x Inception Module C

Auxiliary Classifier

For promoting high dimensional representations

1x7 and 7x1 replacing 7x7

n = 7 in implementation

Inception Module B

Inception Module C
The Network

- “Good” accuracy of 99.4%
  - False positive rate of 1.8%
- “NoData” accuracy of 100.0%
  - No false positives/negatives
- “bad” accuracy of 93.3%
Towards Crowd sourcing labeling

AI Classification: cosmic
Confidence%: c:87.2 b:10.8 u:0.00 g:1.94
Future

- **Expert labeling**
  - Different/more varied classifications
- **Build and train a few models**
  - Connect them together (LSTM/Recurrent NN?)
- **Integrate with data taking**
  - Incremental training pipeline