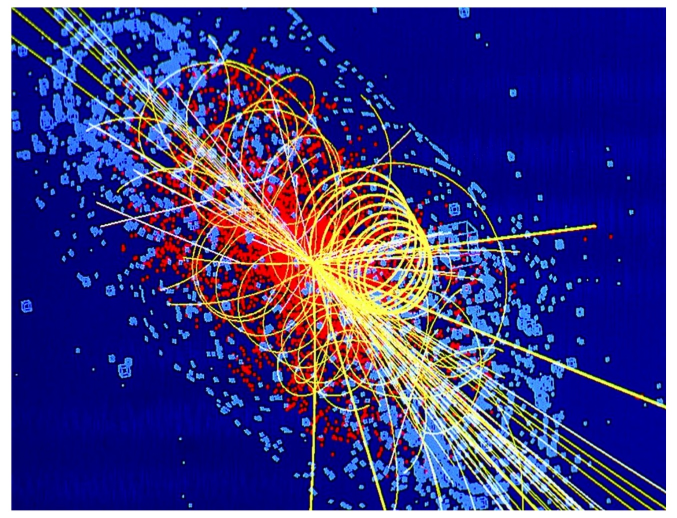
Identifying the Higgs Boson

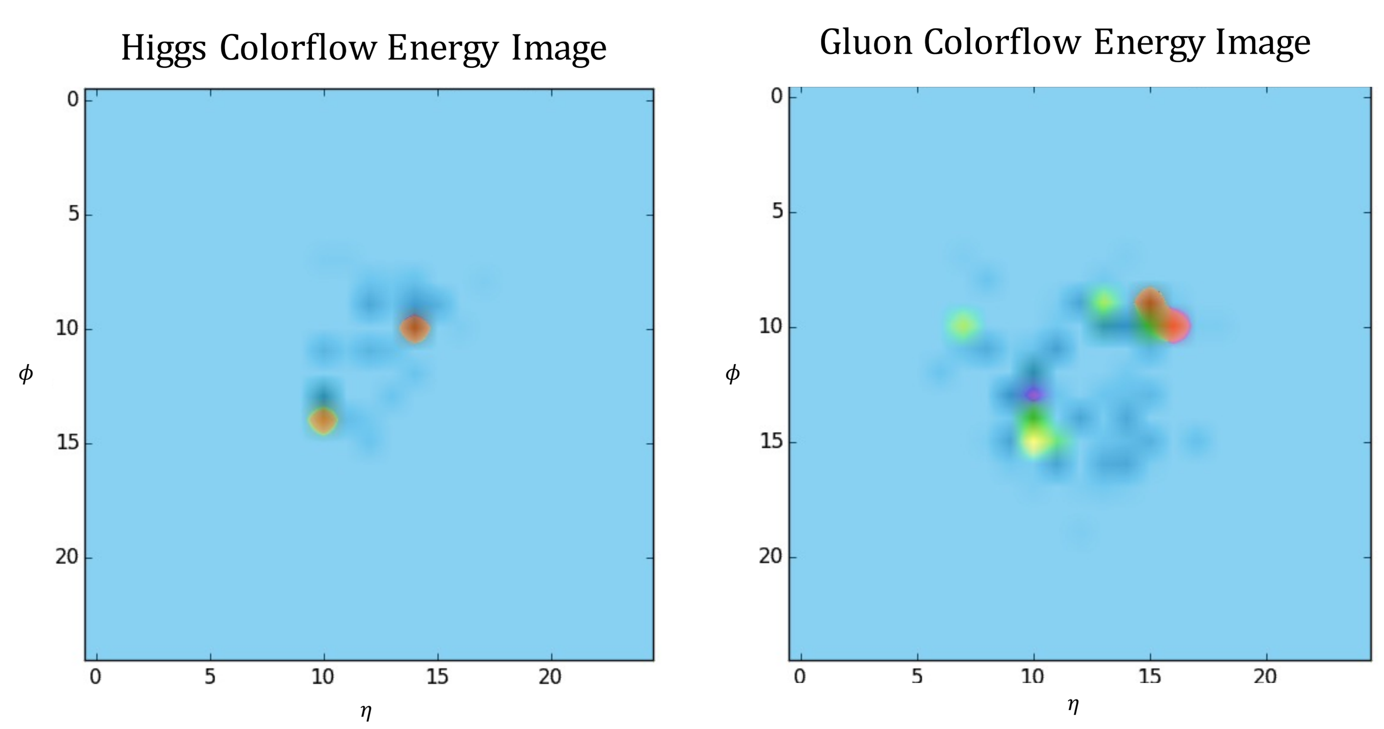
Anton Apostolatos, Leonard Bronner

LHS and the Higgs Boson



The ATLAS detector at the Large Hadron Collider, recording 40 million energy snapshots of proton-proton collisions a second, is tasked to finding the elusive Higgs boson subatomic particle. In a super-charged state, the particle decays in observable ways. By measuring and analyzing the energy patterns of this decay we can identify the subatomic particle itself. Human curation of these events is unfeasible, making an accurate classification system that would label the most promising observations as Higgs boson particle decays required. Discovering the Higgs boson would validate a generation of scientific thought in particle physics and alter our understanding of mass as a physical property.

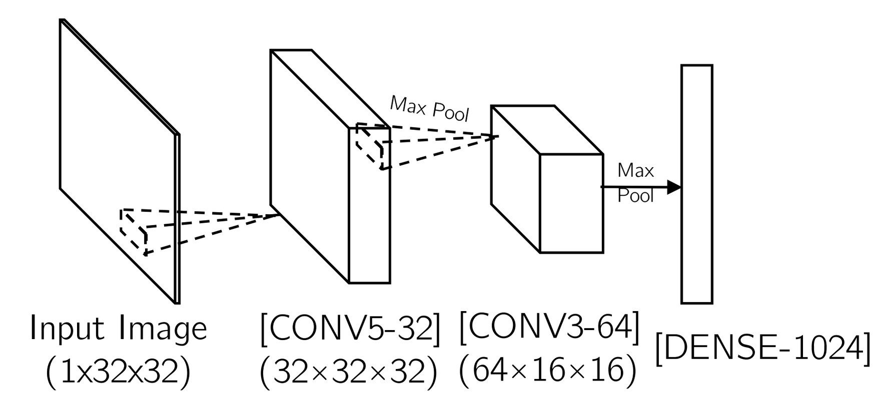
Data



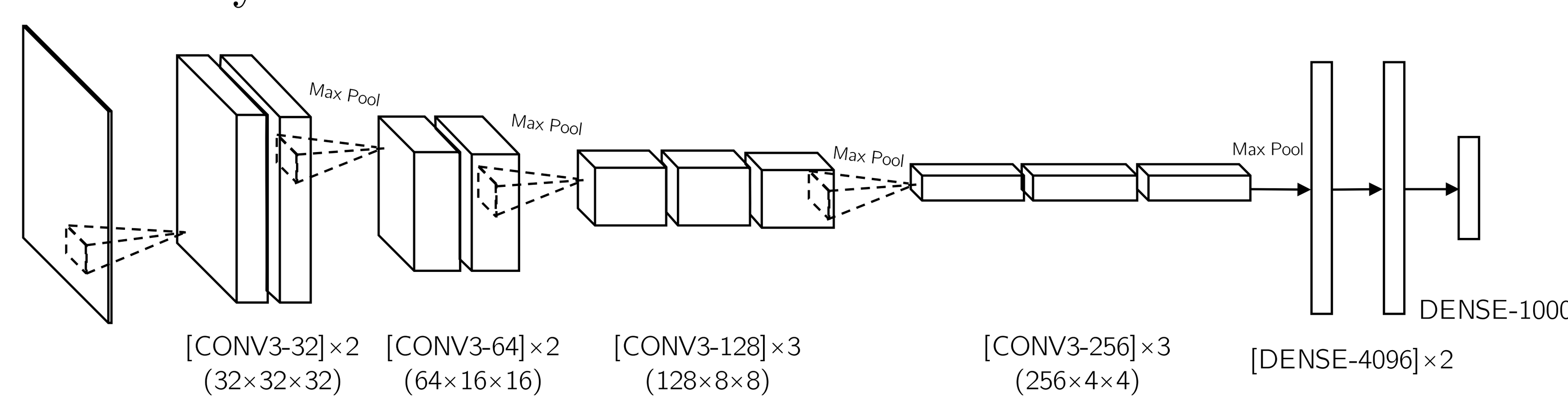
In collaboration with ATLAS we were given access to energy images for both Higgs boson and gluon decay. We had 100,000 25x25 pixel images in η and ϕ cylindrical coordinates, where η is the angle in the x-z and ϕ is the azimuthal angle in the x-y plane perpendicular to the beam direction. These were pre-processed so that the jet center is at the center of an image, with resonance being kept constant in every sampled data point. We also zero-centered the images with the training mean and normalized all values to be between [-1, 1]. We also zero-padded the images to 32x32 to allow for Max-Pooling.

Models

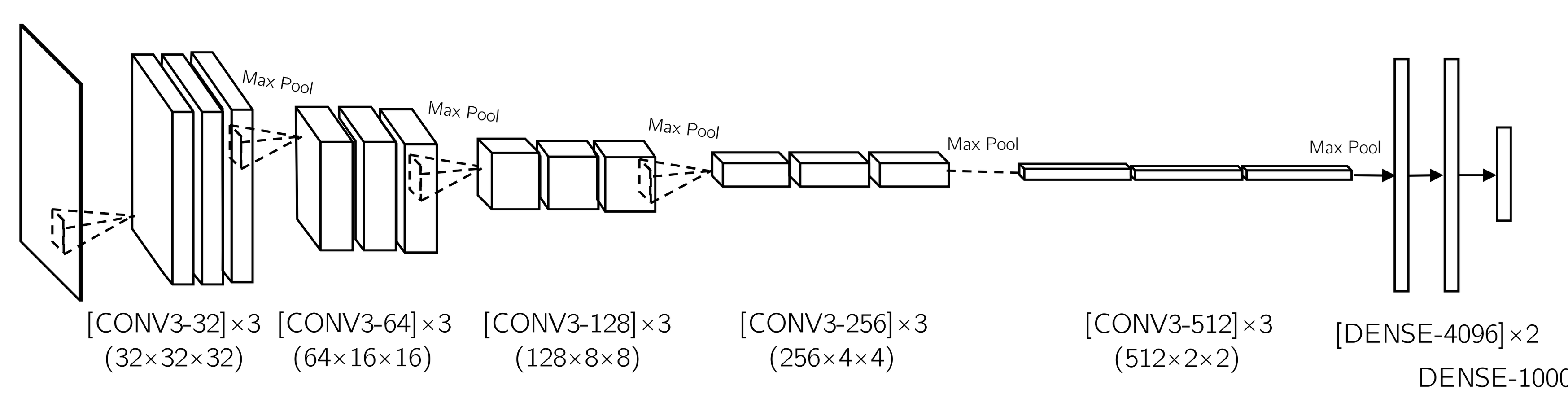
We had initial doubts about the effectiveness of CNNs for this task as these are not traditional visual recognition problems. This also meant that we had to design, build and learn a CNN from scratch – given the lack of similarities between these colorflow energy representations and common image datasets, transfer learning was useless.



To establish a baseline in this new framework we ran a simple model. We did this in order to make sure that this problem was tractable for CNNs. The simple model includes two convolutional layers, followed by a Max-Pool layer each, and a dense layer at the end.



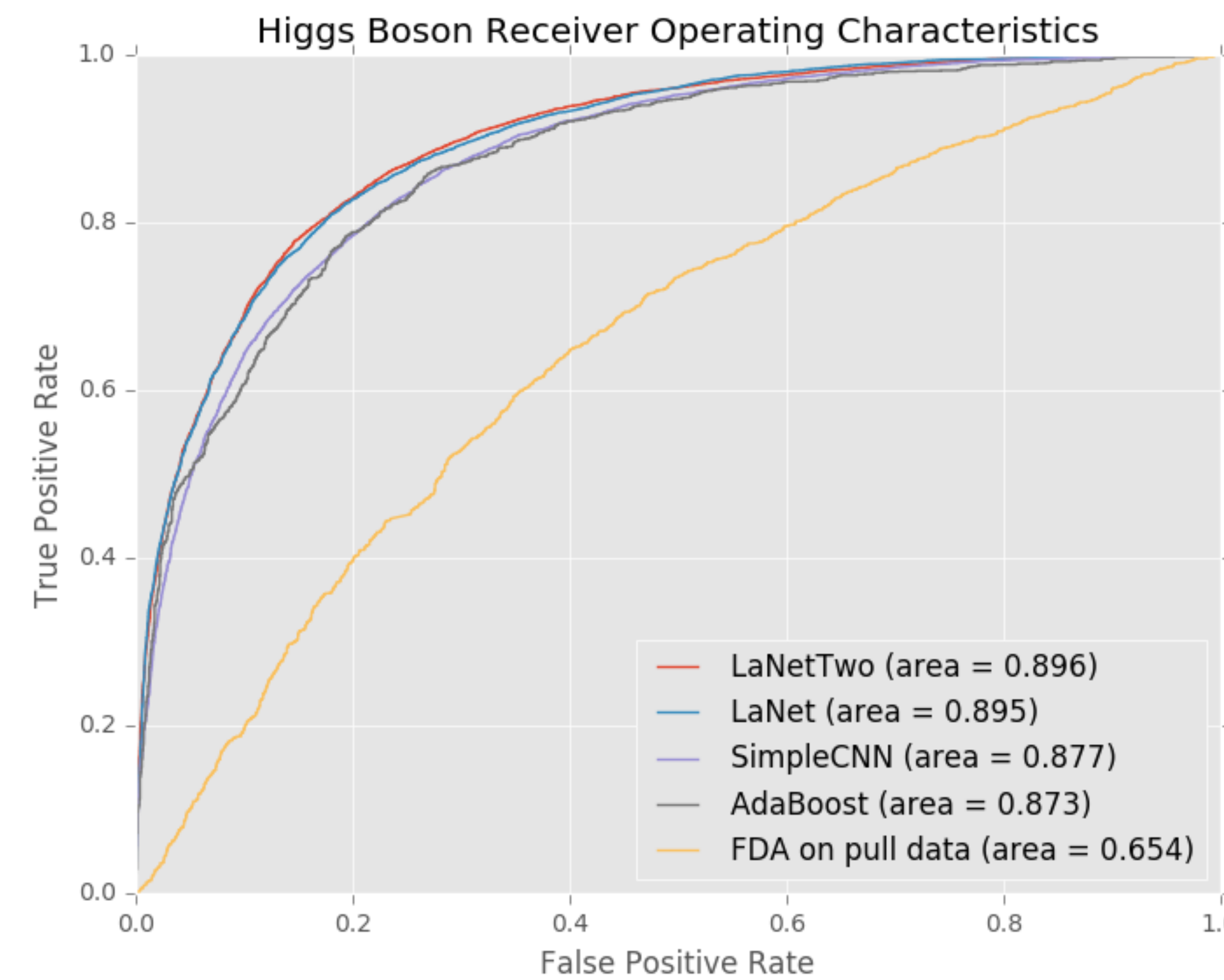
We proceeded by building a larger model which we refer to as LaNet. This model includes 10 convolutional layers, four Max-Pools, and two final dense layers. The size of the network was constrained by the low dimensionality of the image.



We also decided to maximize our network size, designing an even larger network, which we will refer to as LaNetTwo. given the dimensionality problem explained above,we could only add one more set of convolutional and Max-Pool layers. We also increased all layer convolution sets from two to three.

Results

Current state-of-the-art runs Fisher Discriminant Analysis on pull data, a feature class which characterizes the superstructure of the particle collision events. This classification method has a AUC of 0.654.



Our first baseline after the state-of-the-art Fisher Discriminant Analysis was an AdaBoost tree algorithm with Random Forest base estimators on the vectorized images. The AdaBoost classifier is an adaptive model that runs and tweaks weaker learning algorithms. With this algorithm we achieved an AUC of 0.873.

Our SimpleCNN achieved an AUC of 0.877, showing us that CNNs were a promising framework for this task. Our larger models, LaNet and LaNetTwo, both achieved AUC scores of 0.895 and 0.896, respectively. It's interesting to note that the difference between these two models is negligible, indicating that to achieve greater classification power a substantially more complex convolutional neural network is required.