Quark/Gluon Discrimination with Jet-Images and Deep Learning

BSM/LHC/DM Journal Club

Patrick T. Komiske

Center for Theoretical Physics, Massachusetts Institute of Technology

Based on arXiv:1612.01551 – PTK, Eric Metodiev, Matthew Schwartz
Ongoing work – PTK, Eric Metodiev, Ben Nachman, Francesco Rubbo, Matthew Schwartz

September 8, 2017

Overview

- Jet Images and Neural Networks
- Quark/Gluon Discrimination
- ATLAS Simulation
- Towards Learning with Data

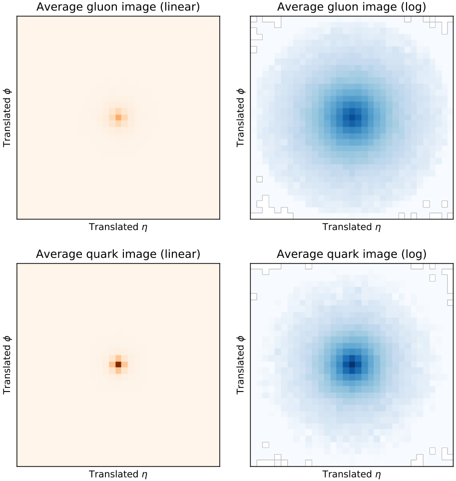
- Perspective on deep learning:
  - Deep learning is an incredible tool that HEP should explore (it's 2017!)
  - There are obvious limitations (what is it learning?) so more work is needed
  - Goal of this work is to demonstrate a use case for deep learning and inspire further studies

Jet Image Basics

- Simple idea: treat calorimeter towers as pixels in an image with intensity given by the p_T
- History:
  - Pumplin (1991): uses jet images to construct powerful single observables for q/g discrimination (e.g. $N_f =$ min $\#$ of pixels needed to account for f% of the p_T)
  - Cogan, Kagan, Strauss, Schwartzman (2015): applies Fisher Linear Discriminant (FLD) to jet images, studies W vs. QCD background
  - Almeida, Backovic, Cliche, Lee, Perelstein (2015): jet images for top vs. QCD
  - Oliveira, Kagan, Mackey, Nachman, Schwartzman (2015): W vs. QCD with jet images and Deep Neural Networks (DNN)
  - PTK, Metodiev, Schwartz (2016): light quark vs. gluon with jet images and DNNs

Jet Image Example - Average Quark and Average Gluon

- Gluons radiate proportional to $C_A = 3$, quarks radiate proportional to $C_F = 4/3$
- Gluon jets fatter than quarks for given energy bin
- Image details:
  - 33x33 pixels
  - 0.8x0.8 in (η, ϕ) space
  - Resolution of 0.024x0.024 (comparable to ECAL)



Neural Network Basics

- Neural Network (NN) = arbitrary function approximator
- Lönnblad, Peterson, Rögnvaldsson (1990): applied small NN to quark vs. gluon problem (inferior to Pumplin's N_{90} approach at the time)
- Recent advances make more sophisticated (deep) NNs possible - hardware (GPUs), architecture design (convolutions), activation functions (ReLU), accessibility (Keras)
- Two key choices:
  - Choice of representation of the jet
    - Other choices: four-vectors [Louppe, Cho, Becot, Cranmer (2017)], N-subjettiness [Datta, Larkoski (2017)], ECF(G)s, angularities
  - Choice of analysis of that representation
    - Other choices: Fisher linear discriminant, boosted decision tree (BDT), shallow/dense NNs

Neural Network Basics

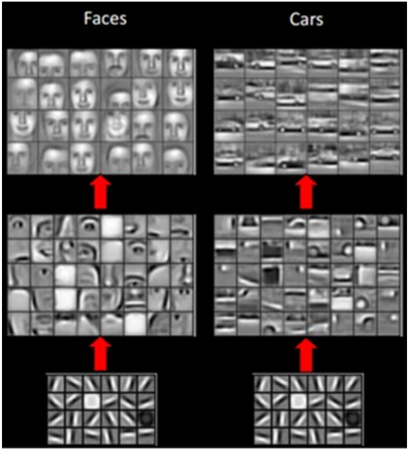
- Neural Network (NN) = arbitrary function approximator
- Lönnblad, Peterson, Rögnvaldsson (1990): applied small NN to quark vs. gluon problem (inferior to Pumplin's N_{90} approach at the time)
- Recent advances make more sophisticated (deep) NNs possible - hardware (GPUs), architecture design (convolutions), activation functions (ReLU), accessibility (Keras)
- Two key choices:
  - Choice of representation of the jet $\longrightarrow$ jet images
    - Other choices: four-vectors [Louppe, Cho, Becot, Cranmer (2017)], N-subjettiness [Datta, Larkoski (2017)], ECF(G)s, angularities
  - Choice of analysis of that representation
    - Other choices: Fisher linear discriminant, boosted decision tree (BDT), shallow/dense NNs

Neural Network Basics

- Neural Network (NN) = arbitrary function approximator
- Lönnblad, Peterson, Rögnvaldsson (1990): applied small NN to quark vs. gluon problem (inferior to Pumplin's N_{90} approach at the time)
- Recent advances make more sophisticated (deep) NNs possible - hardware (GPUs), architecture design (convolutions), activation functions (ReLU), accessibility (Keras)
- Two key choices:
  - Choice of representation of the jet $\longrightarrow$ jet images
    - Other choices: four-vectors [Louppe, Cho, Becot, Cranmer (2017)], N-subjettiness [Datta, Larkoski (2017)], ECF(G)s, angularities
  - Choice of analysis of that representation $\longrightarrow$ deep convolutional NNs
    - Other choices: Fisher linear discriminant, boosted decision tree (BDT), shallow/dense NNs

Convolutional Neural Networks

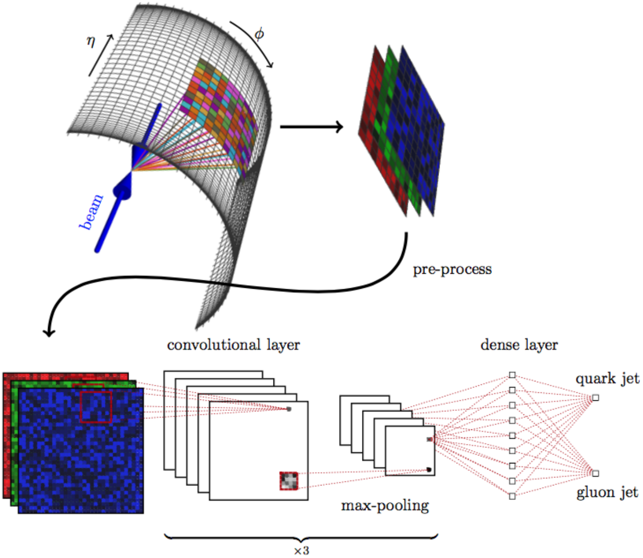
- Standard network architecture for modern image recognition
- Filters are convolved with previous layer to produce output
- Reasons for use:
  - Translation invariance
  - Efficient computation
- Different filters are used for detecting different "features"
- Deeper layers correspond to higher level features



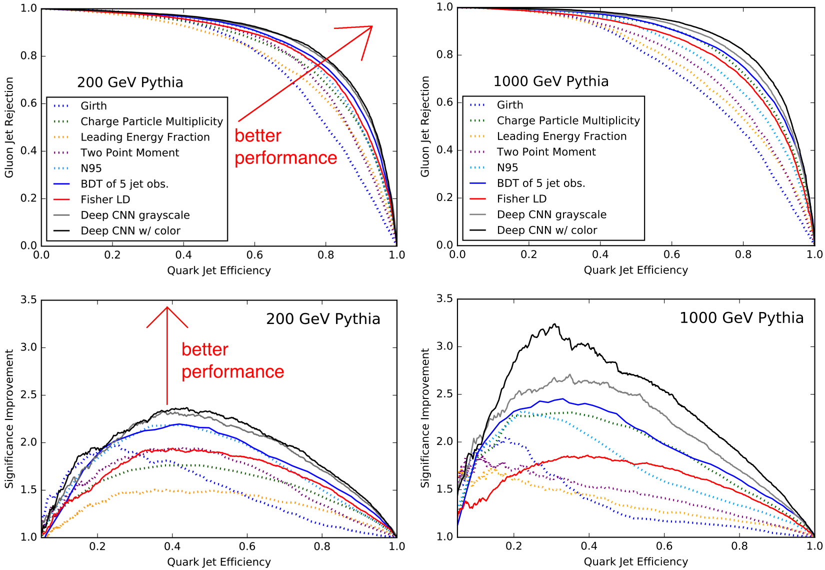
Additional Information – Multi-Channel Jet Images

- Using all available information should maximize network performance
- In analogy with an RGB image, additional jet image channels can be thought of as different "colors"
- Gallicchio, Schwartz (2012) argue there are essentially two kinds of observables for q/g discrimination, "counting" and "shape"
- Traditional jet image contains geometric information about energy flow, supplement with some count observable
- Our choice (non-canonical):
    - Channel 1: charged p_T
    - Channel 2: neutral p_T
    - Channel 3: charged particle multiplicity
- Tried an 18-channel image with p_T and charged counts for each type of particle appearing – learning too difficult to merit this approach initially

Network Architecture

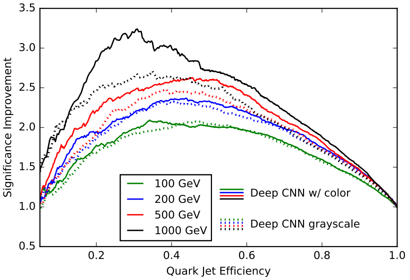


ROC Curves


200 GeV Pythia
Girth
Charge Particle Multiplicity
Leading Energy Fraction
Two Point Moment
N95
BDT of 5 jet obs.
Fisher LD
Deep CNN grayscale
Deep CNN w/ color
better performance
Gluon Jet Rejection
Quark Jet Efficiency
1000 GeV Pythia
Significance Improvement

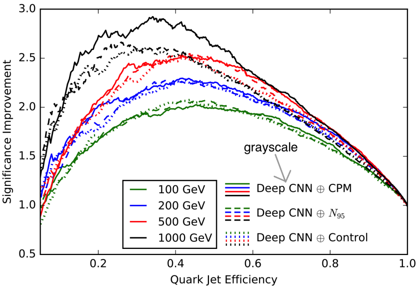
Additional Studies

- Does the multi-channel approach work?



- "Color" helpful at higher p_T

- Has the network learned common observables?

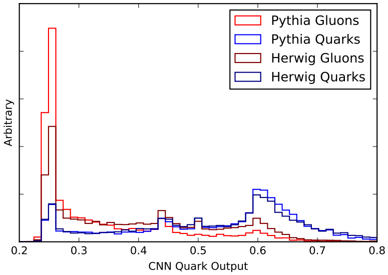


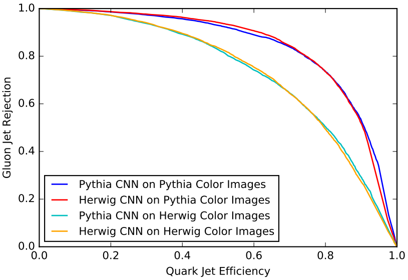
- NN knows N_{95}
- CPM boosts perf. at high p_T

Monte Carlo Comparison

- Train/test with Pythia/Herwig
- NN output defines an observable
- Output has interpretation as a confidence
- Quarks appear similar, gluons not so much

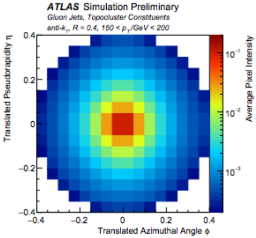
- ROC curve independent of MC used to train (working points different)
- Herwig trained model tested on Pythia images matches performance of all-Pythia setup



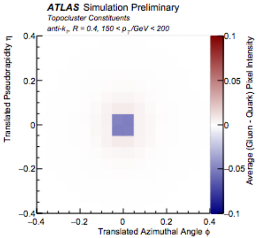


ATLAS Simulation

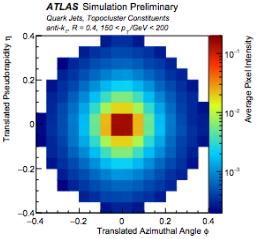
- ATLAS has investigated jet images in simulation
- Average Topocluster Images are shown below



Gluons



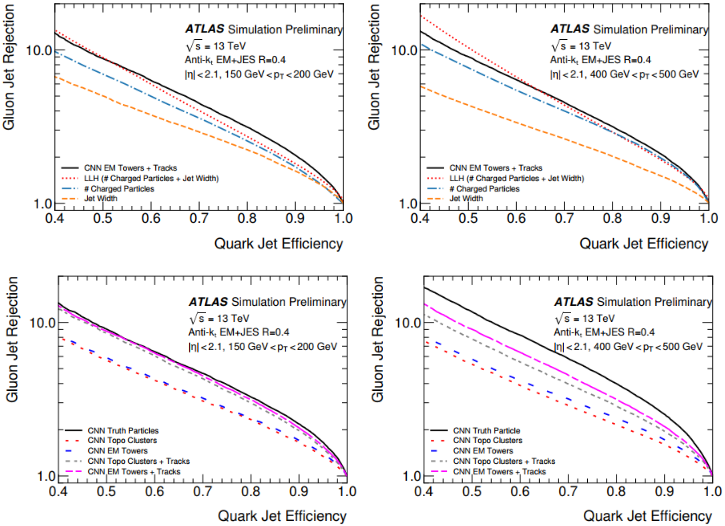
Difference



Quarks

ATL-PHYS-PUB-2017-017

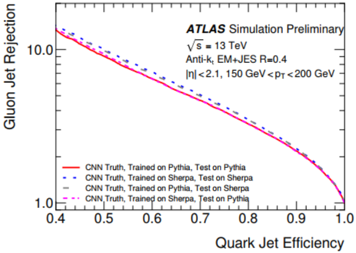
ATLAS Simulation Results

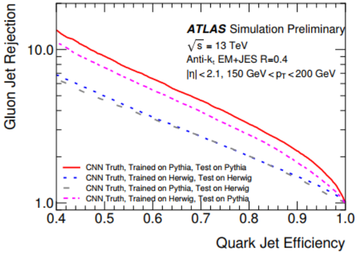


ATL-PHYS-PUB-2017-017

ATLAS Monte Carlo Comparison

- Closure property holds well for Pythia and Sherpa (left)
- Training on Pythia vs. Herwig give slightly different results but trend still holds that test sample is the dominant effect (right)





ATL-PHYS-PUB-2017-017

Weakly Supervision - Learning from Label Proportions (LLP)

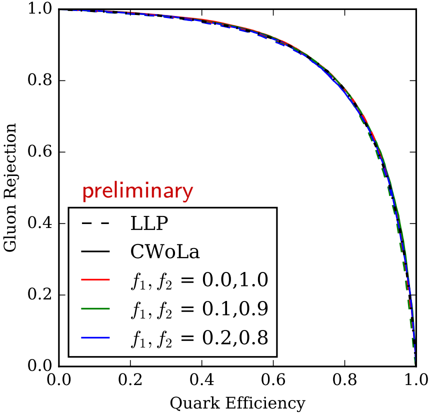
- Introduced by Dery, Nachman, Rubbo, Schwartzman (2017)
- Suppose we know only data fractions instead of sample ground truth
- Change loss function to $f_{\text{weak-loss}} \simeq \left|\sum_{i=1}^{N} \frac{\text{model}(x_i)}{N} - y\right|$, where y is batch fraction

Classification Without Labels (CWoLa)

- Metodiev, Nachman, Thaler (2017)
- Pretend that mixed samples are pure samples and train away
- Loss function is the same as in strong supervision (categorical crossentropy)
- Smoothly interpolates to strong learning in the case of pure samples

Performance

- Performance of LLP and CWoLa essentially the same as strong supervision, even for different sample purities!



Concluding Remarks

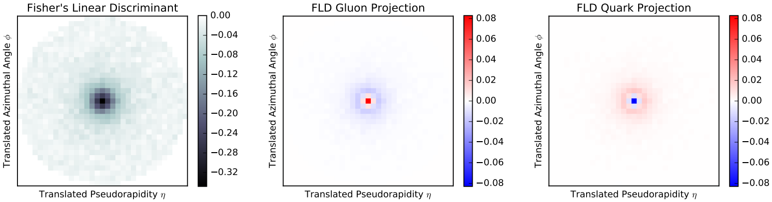
- Conclusions:
    - Jet images and deep CNNs can be successfully used to discriminate quarks and gluons
    - Multi-channel jet image approach yields additional discrimination power
    - Interesting closure test shows that training is picking up on universal features between MCs
    - ATLAS has implemented these techniques in simulation
- Further work:
    - Opening the box – need to understand what the network is learning
    - Optimizing network architecture – choices made here are reasonable but not very optimized
    - Learning directly from data – promising methods are being developed to make this a reality in the very near future

Backup Slides

Simulation Details

- Pythia 8.219, Herwig 7.0
- Train on 180k images, validate on 10k, test on 10k
- $\sqrt{s} = 13\,\text{TeV}$
- $|\eta| < 2.5$
- $R = 0.4$ anti-k_t jets
- 10% wide p_T bins

Fisher Linear Discriminant



Shallow NN Filters