W UNIVERSITY of WASHINGTON

Tag, You're It!

Classifying Jets With A Transformer Architecture

Michael Rosen
Physics and Chemistry with Applied Math Minor

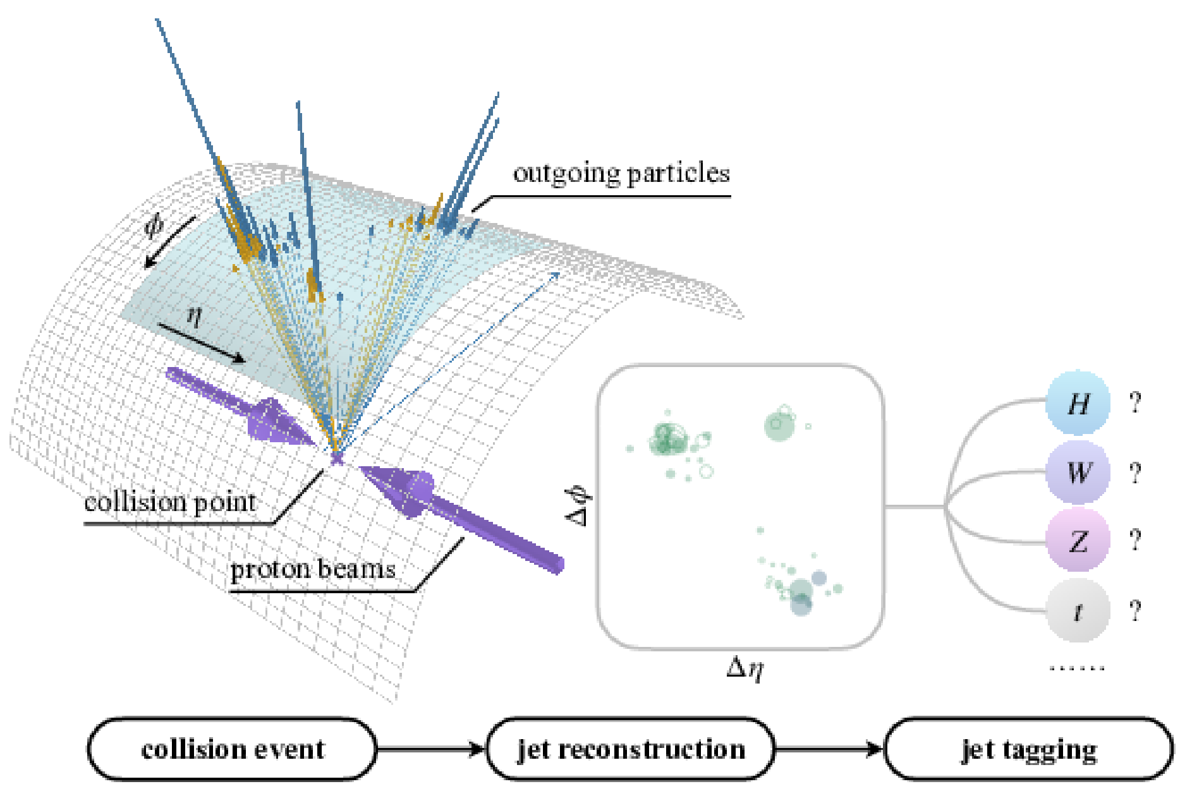
Emma Bacarra
Physics and Astronomy with Data Science Minor

Abstract

Understanding particle behavior could uncover mysteries of the universe or the discovery of new building blocks of matter. A transformer network is a powerful guide in the process.

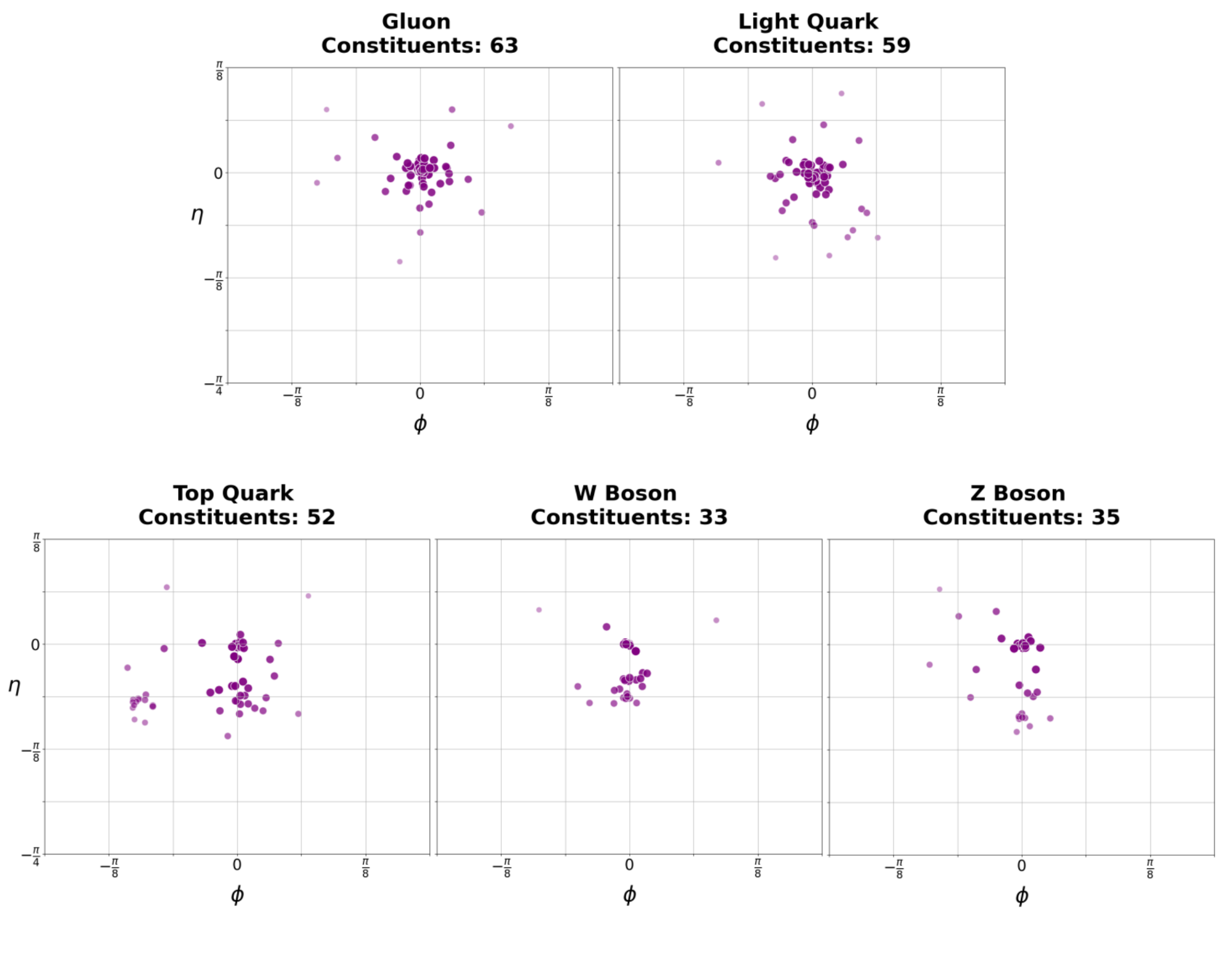
A particle beam collision sprays out debris of various, unstable particles that decay into a cluster, or **jet**. With hundreds of particles, or **constituents**, in a single jet, it can be difficult to identify types and reconstruct the event with static parameters.

A Jet Event at the ATLAS Detector [1]



The flexibility of a many-to-one transformer is advantageous in finding patterns to categorize, or **tag**, these particle jets. Source particles can be accurately identified while simultaneously lowering the complexity of classification.

DECAY PATTERNS



The Top Tagging Dataset

This dataset is sourced from the **ATLAS Detector** in the Large Hadron Collider at CERN. Comprised of ~2M event files, each jet is categorized into a **class**: Gluon, Light Quark, W Boson, Z Boson, or Top Quark.

Recorded Features of Each Constituent

p_T - Momentum as a Fraction of the Jet Total

η – Angular Coordinate (Pseudorapidity)

ϕ – Angular Coordinate (Azimuthal Angle)

E – Energy of the Constituent Particle

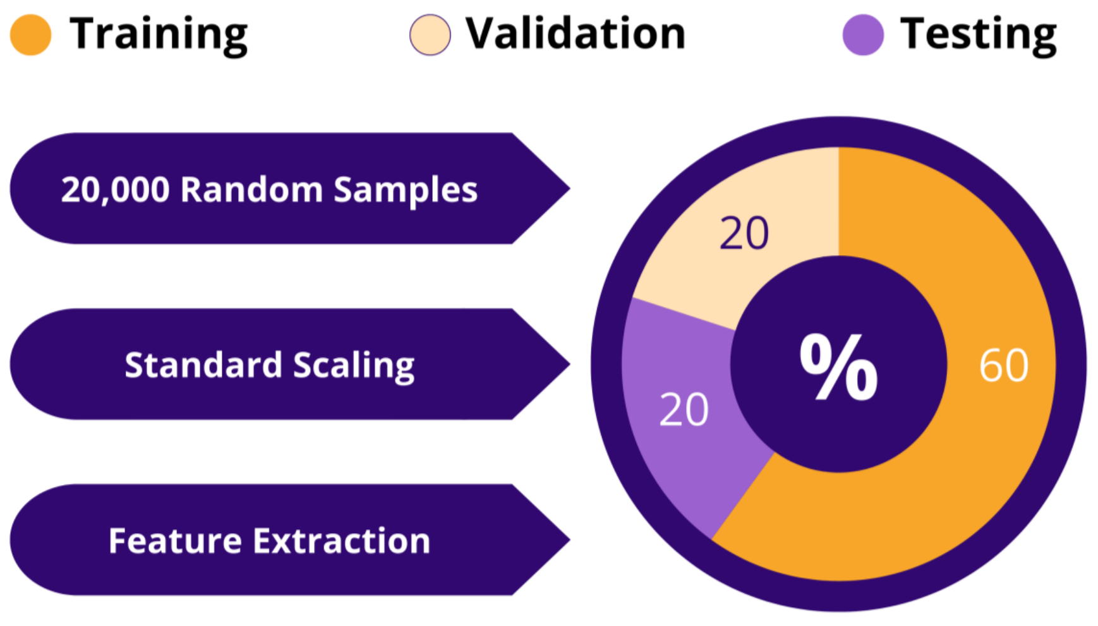
$\Delta R \; - \sqrt{\eta^2 + \phi^2}$

Total Jet Class Characteristics

$$I = \begin{bmatrix} p_{t,1} & \eta_1 & \phi_1 & E_1 & \Delta R_1 \\ p_{t,2} & \eta_2 & \phi_2 & E_2 & \Delta R_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{t,n} & \eta_n & \phi_n & E_n & \Delta R_n \end{bmatrix}$$

Data Preprocessing

A **60 / 20 / 20 split** was used to allocate the proper training, testing, and validation data, respectively.



MODEL ARCHITECTURE

Classification of All Five Jets in Parallel

Encoder Layer

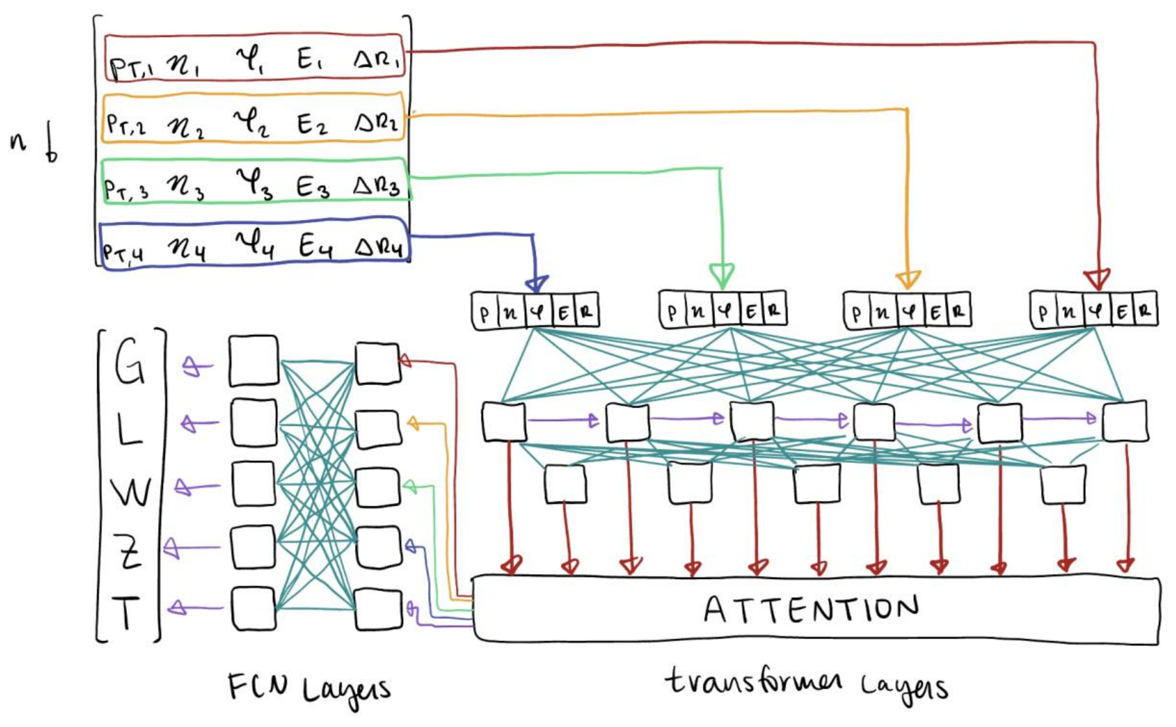
- 5 Attention Heads
- 2 Layers
- 256 Neurons per Layer
- 10% Dropout Rate

Fully Connected Network Layer

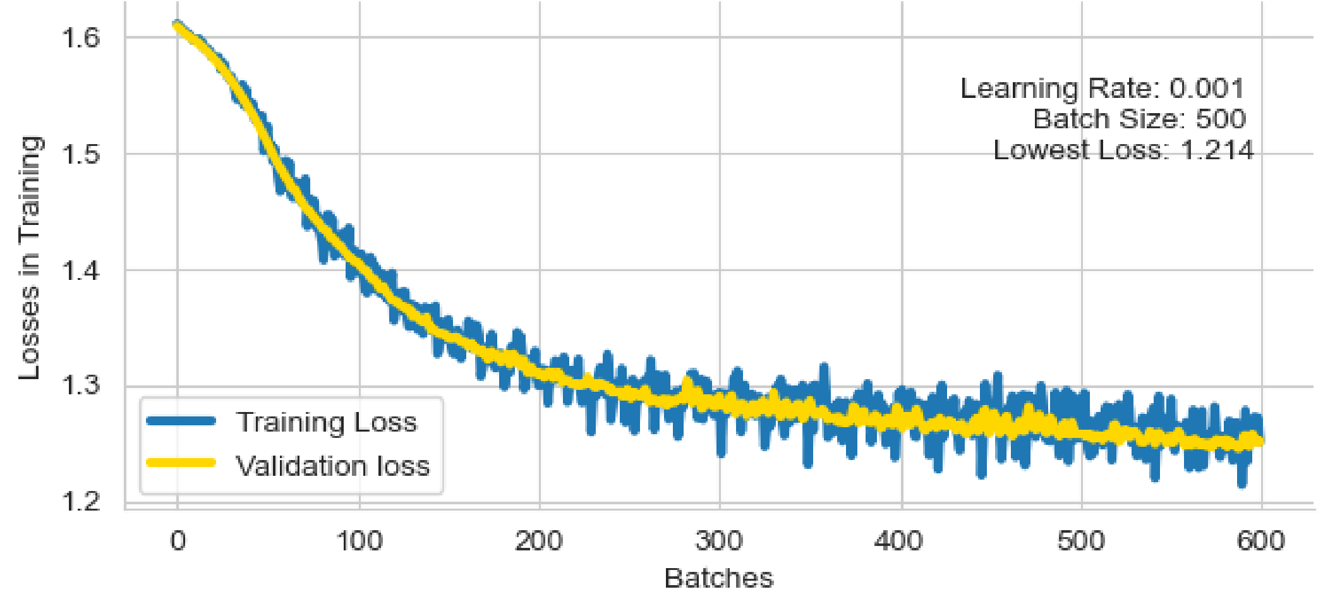
- Hidden Layer with 5→100 Neurons and ReLU Activation
- Output Layer with 100→5 Neurons and Softmax Activation

Training Performance

Following the architecture design, the transformer trained to tag the jet classes **in parallel**.


$p_{T,1}$ η_1 ϕ_1 E_1 ΔR_1
$p_{T,2}$ η_2 ϕ_2 E_2 ΔR_2
$p_{T,3}$ η_3 ϕ_3 E_3 ΔR_3
$p_{T,4}$ η_4 ϕ_4 E_4 ΔR_4
n
G
L
W
Z
T
ATTENTION
FCN Layers
transformer Layers

With 25 epochs, it learned to recognize patterns in 24 batches of 500 events. The **training loss** and **validating accuracy** were recorded throughout the process to monitor a stable convergence towards high accuracies.



Parallel Classifier Accuracy of 64.80%

Advanced models have demonstrated accuracies of **up to 84%**, prompting for a further investigation on the dataset and the parallel model. Subsets of the training data were created by filtering pairs of classes for a **binary classifier** for better context of the model's limitations.

Testing Statistics

For statistical analysis, these equations are derived:

Overall Accuracy

$$A_T = \frac{\sum_m C_{mm}}{\sum_n \sum_l C_{nl}}$$

Binary Subsets

$$c_{i \rightleftharpoons j} = \begin{bmatrix} C_{ii} & C_{ij} \\ C_{ji} & C_{jj} \end{bmatrix}$$

Individual Accuracy

$$A_i = \frac{C_{ii}}{\sum_m C_{mi}}$$

Binary Accuracy

$$a_{i \rightleftharpoons j} = \frac{\sum_m c_{mm}}{\sum_n \sum_l c_{nl}}$$

The Confusion Matrix

The **confusion matrix** is a powerful metric for analyzing performance through the distributions of correct and incorrect classifications. Here, the matrices have been normalized to show the **ratio of each outcome to the total outputs**:

Parallel Classifier

$$C_{5,net} = \begin{bmatrix} 14.62 & 3.42 & 1.50 & 1.60 & 2.67 \\ 2.25 & 12.12 & 3.20 & 2.00 & 0.35 \\ 0.30 & 0.90 & 14.45 & 6.92 & 0.85 \\ 0.27 & 0.47 & 2.30 & 8.90 & 1.52 \\ 2.40 & 1.42 & 0.15 & 0.68 & 14.70 \end{bmatrix} \qquad A_T = 64.80\%$$

Binary Classifiers

$$C_{5,binary} = \begin{bmatrix} 11.25 & 3.08 & 1.34 & 1.67 & 1.21 \\ 2.61 & 9.88 & 1.14 & 2.81 & 1.07 \\ 0.94 & 3.62 & 11.94 & 5.76 & 1.81 \\ 0.61 & 2.81 & 4.35 & 10.55 & 1.41 \\ 3.08 & 1.41 & 1.34 & 1.94 & 12.37 \end{bmatrix} \qquad A_T = 56\%$$

Success Rate Table:
The Models' Abilities to Distinguish Between Classes

Parallel Classifier

Gluon	83 Light Quark Gluon	94 W Boson Gluon	93 Z Boson Gluon	85 Top Quark Gluon
80 Light Quark Gluon	Light Quark	87 Light Quark W Boson	89 Light Quark Z Boson	94 Light Quark Top Quark
92 W Boson Gluon	82 W Boson Light Quark	W Boson	72 W Boson Z Boson	97 W Boson Top Quark
91 Z Boson Gluon	78 Z Boson Light Quark	63 Z Boson W Boson	Z Boson	92 Z Boson Top Quark
84 Top Quark Gluon	90 Top Quark Light Quark	88 Top Quark W Boson	87 Top Quark Z Boson	Top Quark

Binary Classifier

Discussion

To determine whether our model met or exceeded the performance of binary classifiers, we can evaluate the following (matrix dot and magnitude)

Theory	Calculation
$\hat{C} = \frac{C}{\sqrt{\sum_{m,n} \lvert C_{mn} \rvert^2}}$	$\lvert C_{5,net} \rvert_F = 31.21$
$S_{dir} = \sum_m \sum_n \hat{C}_{1,mn} \hat{C}_{2,mn}$	$\lvert C_{5,binary} \rvert_F = 27.58$
$D_{mag} = \sqrt{\sum_{m,n} \lvert C_{1,mn} - C_{2,mn} \rvert^2}$	$S_{dir} = 0.98$
	$D_{mag} = 7.27$

From this, it is evident that there is no significant lapse in performance between the binary and parallel classifiers.

CONCLUSIONS

The application of this novel transformer model to CERN's Top Tagging dataset paves the way for further research in this area. Computing power and the architectural design itself may soon limits constrained the potential of this model, but the rapid development in both consumer computing power and the architectural design itself may soon enable more accurate networks to be created.

A big thank you to **Professor Shih-Chieh Hsu** and **Daniela Koch** for their valuable help and support.

[1] Huilin Qu, Congqiao Li, & Sitian Qian. Particle Transformer for Jet Tagging. 39th International Conference on Machine Learning (ICML), 2022