Vision Transformers for End-to-End Quark-Gluon Jet Classification from Calorimeter Images

Md Abrar Jahin[1] Shahriar Soudeep[2] Arian Rahman Aditta[3] M. F. Mridha[2] Nafiz Fahad[4] Md. Jakir Hossen[4]

[1]University of Southern California [2]American International University-Bangladesh [3]Khulna University of Engineering & Technology [4]Multimedia University

Introduction

Why Distinguish Quark and Gluon Jets?

- Key for precision measurements & new physics at the LHC.
- **Gluon jets**: broader, higher multiplicity, softer p_T due to larger color factor [1].
- Discrimination is challenging due to pileup and detector noise.

Limitations of CNNs

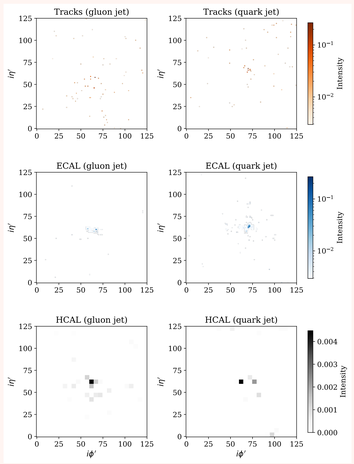
- CNNs excel at local features but struggle with long-range spatial dependencies.
- Often rely on handcrafted observables or full reconstruction pipelines.

Why Vision Transformers (ViTs)?

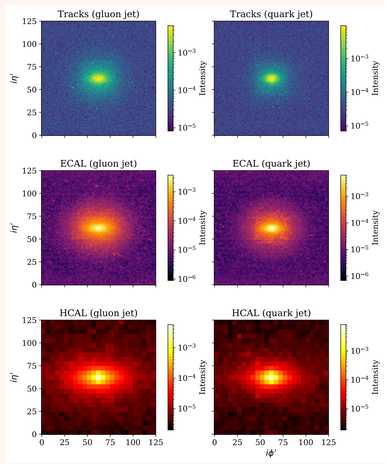
- Model global context via self-attention, capturing long-range spatial patterns.
- Naturally suited for **end-to-end learning** on multi-channel calorimeter images.
- Operate directly on detector-level energy deposits, bypassing reconstruction [2].
- **Hypothesis**: ViTs better capture subtle jet substructure differences.

Dataset and Jet Image Construction

- **Source**: Simulated 2012 CMS Open Data (QCD Dijet events, 8 TeV, 933K labeled jets).
- **Channels**: 3-channel jet images from ECAL, HCAL, and Tracks $\rightarrow$ mapped in η-ϕ space.
- **Image Size**: 125×125 pixels per jet; centered on the highest-energy HCAL tower.
- **Selection Criteria**: $|\eta| < 1.57$, $p_T > 70$ GeV, $\Delta R < 0.4$ to truth-level parton.



(a) Single event visualization



(b) Average per-pixel intensity distributions

Figure 1. **(a)** Representative gluon (left) and quark (right) jet event across Tracks, ECAL, and HCAL channels. Log scale is used for Tracks and ECAL; HCAL uses a linear scale. **(b)** Average per-pixel intensity maps over $N = 10^4$ gluon and quark jets. Log scaling highlights dynamic range; colorbars show channel-wise intensities.

Methodology

Preprocessing

- Zero suppression, Z-score normalization, clipping, sample-wise min-max scaling
- **Augmentation**: horizontal flips, random rotations, resized crops, color jitter
- **ViTs**: additional MixUp augmentation

Models Evaluated

- **CNNs**: ResNet50, EfficientNet-B0, ConvNeXt, RegNetY
- **Transformers**: ViT-Base, Swin Transformer, MaxViT, CoAtNet
- **Hybrids**: ViT+MaxViT, ViT+ConvNeXt, ViT+Swin, ViT+Triple Ensemble

Training Setup

- **Optimizer**: AdamW, **LR**: 1×10^{-4} (classifier), 1×10^{-6} (frozen layers)
- **LR scheduling**: Cosine Annealing
- **Batch size**: 32, **Epochs**: max 20, **Early stopping**: 5 epochs
- Mixed-precision enabled

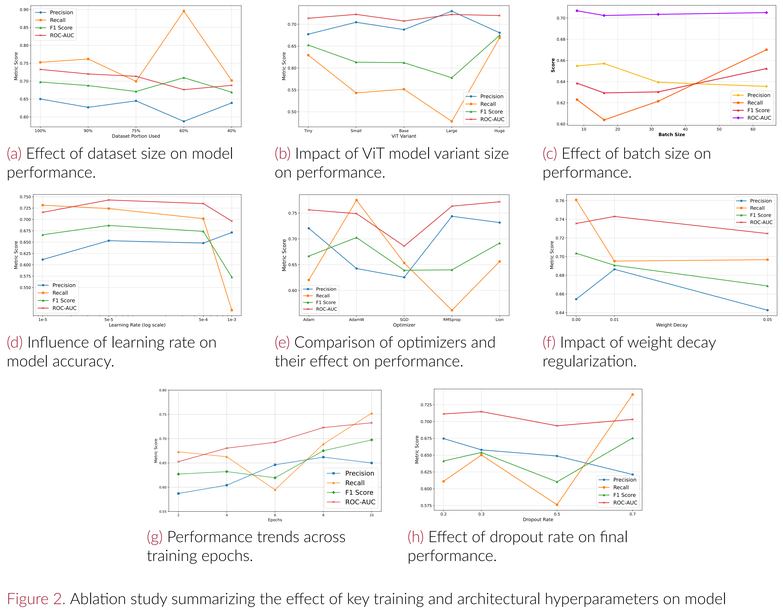
Results

Table 1. Performance comparison of benchmarked models on 7k samples from the quark-gluon dataset. Results are reported as *mean* $\pm$ *standard deviation* over three runs with different random seeds. **Bold** indicates the best performance, while underline marks the second-best result.

Model	Accuracy (%) (↑)	Precision (%) (↑)	Recall (%) (↑)	F1 Score (%) (↑)	ROC-AUC (%) (↑)	# Params (↓)	Train Time (↓)	Inference Time (ms) (↓)
ViT + MaxViT	70.29 ± 0.0224	**77.35 ± 0.0397**	76.45 ± 0.0613	**72.02 ± 0.0392**	**76.65 ± 0.0287**	236M	54m 41s	276.11
ViT + ConvNeXt	**70.57 ± 0.0354**	72.67 ± 0.0477	75.47 ± 0.0914	71.33 ± 0.0308	76.25 ± 0.0304	287M	34m 27s	354.33
ViT + EfficientNet	70.00 ± 0.0186	71.26 ± 0.0320	76.36 ± 0.0584	70.75 ± 0.0241	76.14 ± 0.0229	190.8M	17m 8s	67.96
RegNetY	69.43 ± 0.0164	71.30 ± 0.0310	66.05 ± 0.0508	68.58 ± 0.0200	75.89 ± 0.0224	2.98M	10.41m	10.89
ViT + Swin	69.86 ± 0.0235	74.43 ± 0.0417	80.93 ± 0.0752	71.27 ± 0.0380	75.62 ± 0.0213	183M	27m 32s	44.74
ViT + RegNetY	69.79 ± 0.0148	70.54 ± 0.0343	69.85 ± 0.0616	70.19 ± 0.0227	74.92 ± 0.0148	89.77M	24.07m	169.11
ConvNeXt	67.57 ± 0.0241	72.93 ± 0.0308	75.24 ± 0.0597	70.33 ± 0.0401	72.91 ± 0.0276	89M	8m 50s	54.20
ViT + CoAtNet	66.79 ± 0.0113	67.59 ± 0.0130	67.87 ± 0.0402	67.73 ± 0.0194	71.79 ± 0.0108	0.79M	7m 52s	20.21
ViT + ConvNeXt + Swin	66.64 ± 0.0280	68.01 ± 0.0275	78.32 ± 0.0492	69.09 ± 0.0314	71.11 ± 0.0159	312M	17m 45s	92.66
ViT	69.29 ± 0.0416	69.34 ± 0.0474	73.68 ± 0.0448	70.09 ± 0.0150	69.28 ± 0.0419	85M	11.1m	55.52
Swin	69.29 ± 0.0555	69.36 ± 0.0547	84.92 ± 0.0559	70.93 ± 0.0390	69.28 ± 0.0557	87M	23.3m	36.58
CoAtNet	61.29 ± 0.0238	66.83 ± 0.0416	**88.00 ± 0.1214**	67.26 ± 0.0619	66.65 ± 0.0285	82M	15m 30s	68.50
MaxViT	66.36 ± 0.0152	65.69 ± 0.0186	78.95 ± 0.0549	69.29 ± 0.0210	66.34 ± 0.0153	119M	59.3m	141.63
ResNet	63.79 ± 0.0146	63.13 ± 0.0134	74.82 ± 0.0723	66.97 ± 0.0365	63.77 ± 0.0145	15M	5m 30s	103.15
EfficientNet	59.57 ± 0.0205	60.33 ± 0.0225	57.33 ± 0.0361	58.57 ± 0.0252	59.58 ± 0.0205	29M	6.1m	58.29

Sensitivity Analysis

1. **Dataset Size**: ViT retains 70.96% F1 score with only 60% training data
2. **Model Size**: ViT-Huge improves F1/Recall but with more compute
3. **Batch Size**: 64 yields best F1/Recall
4. **Learning Rate**: 5×10^{-5} optimal across all metrics
5. **Optimizer**: Lion > AdamW > RMSprop > SGD
6. **Weight Decay**: 0.01 achieves the best generalization
7. **Epochs**: F1 peaks after gradual unfreezing of ViT and MaxViT blocks
8. **Dropout**: 0.3 balances recall and precision well



(a) Effect of dataset size on model performance.
(b) Impact of ViT model variant size on performance.
(c) Effect of batch size on performance.
(d) Influence of learning rate on model accuracy.
(e) Comparison of optimizers and their effect on performance.
(f) Impact of weight decay regularization.
(g) Performance trends across training epochs.
(h) Effect of dropout rate on final performance.

Figure 2. Ablation study summarizing the effect of key training and architectural hyperparameters on model performance. Each subfigure isolates a single factor while holding others constant, illustrating its individual impact.

Conclusion & Future Work

Conclusion

ViT and ViT-CNN hybrid models establish new baselines for quark-gluon jet classification, outperforming traditional CNNs. Their ability to model global spatial dependencies through attention mechanisms enables more effective exploitation of jet substructure. This work presents the first public benchmark of ViTs on CMS calorimeter images in an end-to-end learning setting.

Future Directions

- Test on real experimental data.
- Optimize models for real-time use.
- Understand learned features physically.

Acknowledgment

This work was supported by Multimedia University (MMU), Malaysia.

References

[1] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. Jet substructure at the Large Hadron Collider: a review of recent advances in theory and machine learning. *Physics Reports*, 841:1–63, 2020. Publisher: Elsevier.

[2] M. Andrews, J. Alison, S. An, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczos, and E. Usai. End-to-end jet classification of quarks and gluons with the CMS Open Data. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 977:164304, October 2020.