Energy-based Vision Transformers



Project Manager
Simon Kiefhaber

Principal Investigator
Prof. Dr. Sc. Simone Schaub-Meyer

Project Term
2026 - 2026

Clusters
Lichtenberg II Cluster Darmstadt

Software
Python, PyTorch

Institute
Department of Computer Science

University
Technische Universität Darmstadt

Introduction

Deep learning architectures in computer vision have traditionally relied on feed-forward models, where outputs are directly predicted from a given set of inputs. However, energy-based models (EBMs) present a compelling paradigm shift. Rather than directly predicting an output, EBMs learn an energy function that evaluates the compatibility between inputs and possible outputs. Predictions are then made by optimizing this function to find the output with the lowest energy, a process that effectively verifies rather than generates predictions. This process also allows for quantifying uncertainty, which is not directly possible in feed-forward ViTs. This project aims to combine Vision Transformers (ViTs) with EBMs, resulting in very resource-intensive training, as EBM training also requires additional sampling beyond the usual ViT training.

Methods

In this project, we explore a combination of DeiT3-ViTs [1] with energy-based learning. We use different sampling-based loss formulations for training EBMs [2,3,4], pre-train our combination of ViTs and EBMs on ImageNet1K, and verify its generalization capabilities by finetuning it on other classification datasets such as Flower101 and StanfordCars.

[1] H. Touvron et al., "DeiT III: Revenge of the ViT," ECCV 2022

[2] A. Gladstone et al., "Energy-Based Transformers are Scalable Learners and Thinkers," CoRR abs/2507.02092, 2025

[3] Y. Song et al., "How to Train Your Energy-Based Models," CoRR, vol. abs/2101.03288, 2021 [4] F. Gustafsson et al., "How

to Train Your Energy-Based Model for Regression," BMVC 2020

Results

We run very explorative and preliminary experiments to analyze the potential of energy-based vision transformers for classification and transfer learning in comparison to traditional vision transformers with the following results:

- For the pre-trained classification task (ImageNet), we find no notable difference in accuracy between ViTs and energy-based ViTs (80.3% vs 80.2% Top-1 Accuracy).

- In transfer-learning tasks to other classification datasets, we find that the energy-based ViT performs significantly worse than the DeiT3-baseline.

- The overall computational cost of energy-based models is much higher than that of other methods, as multiple forward passes have to be done for each prediction.

Discussion

While we hypothesized that energy-based Vision Transformers (ViTs) would improve both generalization and overall performance compared to conventional feed-forward Vision Transformers, our empirical evaluation reveals substantial limitations of the current approaches in discriminative vision tasks. Across a range of model variants, training objectives, and loss formulations, we were unable to identify consistent benefits that would justify the additional complexity of the energy-based framework. On the contrary, our experiments highlight several practical drawbacks. In particular, the tested methods introduce significant computational overhead due to the iterative inference process, fail to provide measurable accuracy improvements during large-scale pre-training, and consistently underperform their feed-forward counterparts in transfer learning scenarios. Taken together, these findings suggest that, despite their theoretical appeal, current energy-based ViT formulations do not yet offer a compelling alternative for discriminative computer vision applications and require further methodological advances before their potential advantages can be realized in practice.

Last Update: 2026-09-16 11:27