New York University (Shanghai) Machine Learning Competition

Project Report

Competition Introduction:

The competition focused on representation learning through convolutional autoencoders. Participants were provided with Pokémon Fusion images and tasked with training a model capable of compressing and reconstructing these images while retaining meaningful latent representations. The final score was evaluated by reconstruction accuracy measured by mean squared error, linear probing using logistic regression to test the informativeness of latent features for predicting 170 Pokémon types, and image sampling from the latent space for qualitative assessment.

Model Introduction:

I built a convolutional autoencoder in PyTorch with a symmetric encoder–decoder design, incorporating residual blocks to stabilize training and enrich feature representation. The encoder progressively reduces the 128×128 input image through strided convolutions while increasing channel depth, with a residual block in the deeper layers to maintain gradient flow and capture complex visual patterns. A 1×1 convolution compresses the feature map into an 8192-dimensional bottleneck, preserving essential spatial structure. The decoder mirrors this process with transposed convolutions and output padding to recover the original resolution, ending with a sigmoid layer for pixel normalization. On top of the latent representation, I added a simple classification head—global average pooling followed by a linear layer—to predict among 170 Pokémon types. The model was trained for 200 epochs using AdamW with a cosine learning rate schedule and mixed-precision training for efficiency. Specifically, for the loss function, instead of the original VAE-Style object, I adopted a combined loss that takes the logarithm of both reconstruction (MSE) and classification (cross-entropy) terms. This formulation keeps the two components on a similar scale without requiring manual weighting, preventing one from dominating the optimization. The logarithmic form also smooths large gradients and amplifies small ones, which helps maintain stable convergence and encourages the model to keep improving even when one objective saturates. In short, this design and loss formulation allowed the model to learn compact yet expressive representations that balanced image reconstruction quality with strong class discrimination.