Open Set Recognition using Multi-Task Autoencoder

Dov Korstag

January 21, 2026

1 Problem Formulation: Open Set Recognition

In a standard classification setting (Closed Set), the training and testing data are drawn from the same label space $\mathcal{Y}_{in} = \{1, \ldots, K\}$. However, in Open Set Recognition (OSR), the model encounters samples from unknown classes during inference.

Let $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training set, where $x_i \in \mathcal{X}$ is the input image and $y_i \in \mathcal{Y}_{in}$ is the label. During the testing phase, the model receives a sample x_{test} which may belong to a known class ($y \in \mathcal{Y}_{in}$) or an unknown/out-of-distribution (OOD) class ($y \in \mathcal{Y}_{out}$).

The goal of the OSR model is to learn a function $f(x)$ such that:

$$f(x) = \begin{cases} k \in \mathcal{Y}_{in} & \text{if } x \text{ belongs to class } k \\ \text{Unknown} & \text{if } x \in \mathcal{Y}_{out} \end{cases} \tag{1}$$

This requires the model to create a compact and well-separated representation for known classes while identifying inputs that deviate significantly from the learned distribution.

2 Proposed Method

To address the OSR challenge, we propose a multi-task architecture that integrates an Autoencoder with a Classifier, enhanced by Triplet Loss constraints. The inference mechanism relies on both reconstruction quality and latent space geometry.

2.1 Model Architecture

The model consists of three main components sharing a common feature extractor:

- **Encoder** $E_\phi : \mathcal{X} \to \mathbb{R}^d$: Maps the input x to a low-dimensional latent $z = E_\phi(x)$.
- **Decoder** $D_\psi : \mathbb{R}^d \to \mathcal{X}$: Reconstructs the input from the latent vector, yielding $\hat{x} = D_\psi(z)$.
- **Classifier** $C_\theta : \mathbb{R}^d \to \mathbb{R}^K$: Predicts the class probability distribution from z.

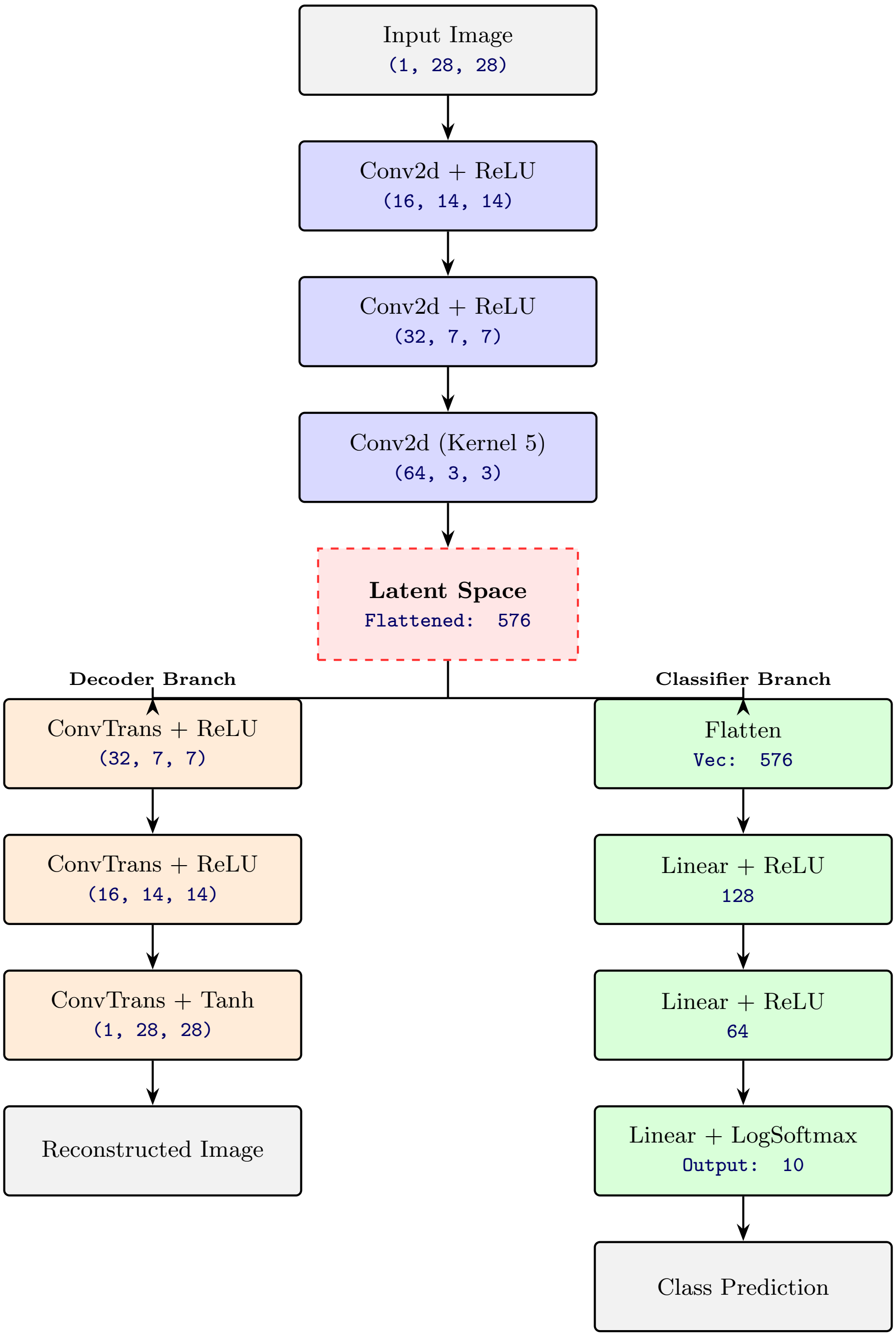

Figure 1: Proposed Multi-Task OSR Architecture with shared feature extractor.

2.2 Training Objective

The network is trained end-to-end by minimizing a combined loss function:

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{trip}\mathcal{L}_{trip} \tag{2}$$

2.2.1 Reconstruction Loss

Ensures the latent representation captures the essential features of the input:

$$\mathcal{L}_{rec}(x) = \|x - D_{\psi}(E_{\phi}(x))\|_2^2 \tag{3}$$

2.2.2 Classification Loss

Standard Cross-Entropy loss to ensure class separability:

$$\mathcal{L}_{cls}(x, y) = -\sum_{k=1}^{K} \mathbb{I}(y = k) \log(C_{\theta}(E_{\phi}(x))_k) \tag{4}$$

2.2.3 Triplet Loss

To enforce a metric structure in the latent space where inputs of the same class are clustered together:

$$\mathcal{L}_{trip} = \max(0, \|z_a - z_p\|_2^2 - \|z_a - z_n\|_2^2 + \alpha) \tag{5}$$

where z_a is the anchor, z_p is a positive sample (same class), z_n is a negative sample (different class), and α is the margin.

2.3 Inference and OOD Detection

During inference, a sample x is classified as "Unknown" based on two criteria derived from the training set statistics:

1. **Latent Distance:** We compute the centroids μ_k and maximum radii R_k for each known class k in the latent space. A sample must fall within the hypersphere of its predicted class:

$$d(z, \mu_{\hat{y}}) \leq R_{\hat{y}} \tag{6}$$

2. **Reconstruction Error:** We compute a global threshold τ (e.g., 99th percentile of training errors). A sample must be well-reconstructed:

$$\|x - \hat{x}\|_2^2 \leq \tau \tag{7}$$

If either condition fails, the sample is rejected as OOD.

3 Experimental Results

We evaluated the proposed OSR framework using MNIST as the in-distribution (ID) dataset, while FashionMNIST and CIFAR-10 served as out-of-distribution (OOD) datasets.

3.1 Quantitative Analysis

The model demonstrated robust performance in distinguishing between known digits and unknown inputs. On the combined test set (comprising MNIST and FashionMNIST), the system achieved an overall accuracy of **97.72%**.

Specifically, the breakdown of the performance is as follows:

- **In-Distribution Accuracy:** The model maintained a high classification accuracy of **96.30%** on the known MNIST classes.
- **OOD Detection Rate:** The system successfully detected and rejected **99.15%** of the unknown samples (FashionMNIST), classifying them correctly as OOD.

These results indicate that the combination of reconstruction error and latent distance thresholds effectively filters out anomalies without significantly compromising the classification performance on known data.

3.2 Qualitative Analysis

To further validate the model's behavior, we visualized the latent space using t-SNE projections. The visualizations confirm that the model learns compact, well-separated clusters for the known classes (digits 0-9). Crucially, the OOD samples are consistently mapped to low-density regions or form separate clusters far from the centroids of the ID classes.

Additionally, the reconstruction visualization demonstrates that while MNIST digits are reconstructed with high fidelity, OOD images result in significant artifacts and high Mean Squared Error (MSE), justifying the use of the reconstruction threshold as a filtering criterion.