Deep Semi-Supervised Anomaly Detection

August 19, 2022

1 Information Theory view on Deep Anomaly Detection

In the supervised classification where we have input variable X, latent variable Z (e.g., the final layer of a deep network), and output variable Y (i.e., the label), we aim to find a minimal compression Z of input X, while retaining informativeness of Z for predicting label Y, i.e.

$$\min_{p(z|x)} \mathcal{I}(X;Z) - \alpha \mathcal{I}(Z;Y) \tag{1}$$

where α is hyperparameter for trade-off between compression and classification accuracy.

In unsupervised learning, one of most widely used principals is *Infomax Principle*, which maximizes the mutual information $\mathcal{I}(X;Z)$ between data and its latent representation under some constraint or regularization $\mathcal{R}(Z)$:

$$\max_{p(z|x)} \mathcal{I}(X;Z) + \beta \mathcal{R}(Z). \tag{2}$$

The constraint term, $\mathcal{R}(Z)$, can be a distance to a prior distribution, an adverserial loss or bottleneck in dimensionality (e.g. autoencoders).

2 Deep Semi-Supervised Anomaly Detection

Supervised or semi-supervised methods for anomaly detection only learn to detect anomalies similar to those seen during training and for this reason, previous semi-supervised anomaly detection methods only used datapoints labeled as normal in their learning process, ignoring anomaly samples in labeled data. Tis paper proposes a way based on **Depp SVVD** (2) to effectively use aomaly samples as well.

With n unlabeld samples $x_1, \dots x_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and m labeled samples $(\hat{x}_1, \hat{y}_1), \dots (\hat{x}_m, \hat{y}_m) \in \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = -1, +1$ with -1 for anomaly and $+1$ for normal samples, *Deep SAD* objective is defined as:

$$\min_{\mathcal{W}} \frac{1}{n+m} \sum_{i=1}^{n} \|\phi(x_i) - c\|^2 + \frac{\eta}{n+m} \sum_{j=1}^{m} \left(\|\phi(\hat{x}_j) - c\|^2\right)^{\hat{y}_j} + \frac{\lambda}{2} \sum_{l=1}^{L} \|\mathbf{W}^l\|_F^2 \tag{3}$$

They same term as Deep SVVD is employed for unlabeled data and the second term, seeks to minimize variance of normal labeled data (in the same way as unlabeled) and maximize variance of anomaly labeled data, with punishing the inverse of the distances such that anomalies to our selected center.

In terms of information theory, this objective function can be viewed as:

$$\max_{p(z|x)} \mathcal{I}(X;Z) + \beta(\mathcal{H}(Z^{-} - Z^{+})). \tag{4}$$

Where maximizing mutual information of X and Z is induced by using pretrained endcoder network of an autoencoder. Also $\mathcal{H}(Z)$ is entropy of latent variable, and as we know, if we assume Z follows a gaussian distribution, its entropy is proportional to its log-variance which we maximize and mimize for anomaly and normal samples respectively with penalizing their distance to c.