Supplementary Material

1. On the influence of feature norm to L2 reconstruction error

In order to set the theoretical ground for supporting our normalized L2 distance introduced in main paper, we discuss and prove in this section the fact that feature with smaller norm leans to produce smaller reconstruction error.

Since the input of our autoencoder is one-dimensional AV feature, it is natural to employ FC network as the architecture of the encoder and decoder. Noticeably, the convolutional layer, shortcut connection and average pooling are all in essence linear mappings, and therefore the additional use of them will not concern the following discussion. For a FC autoencoder with L layers, we denote the weight matrix of l^{th} FC layer as $W^l \in \mathbb{R}^{n^l \times n^{l-1}}$, the offset bias as $\boldsymbol{b}^l \in \mathbb{R}^{n^l}$, the activation function as $\sigma(\cdot)$ and the input as $\boldsymbol{x} \in \mathbb{R}^H$. Then the pre-activation output of l^{th} layer could be recursively written as

$$f^l(\boldsymbol{x}) = W^l \sigma(f^{l-1}(\boldsymbol{x})) + \boldsymbol{b}^l. \tag{1}$$

Although activation functions in the context of deep learning are always nonlinear in the full space, they could be approximately considered as linear in a certain polytope. For instances, given an input in $\mathbb{R}^1$, relu is linear if only the region of $[0, +\infty)$ or $(-\infty, 0)$ is considered. Sigmoid could be approximated as linear exclusively in the region of $(-\infty, -\alpha)$, $[-\alpha, \alpha)$ or $[\alpha, +\infty)$, for some $\alpha > 0$. For simplicity, in the following discussion we consider relu as the applied activation function $\sigma(\cdot)$. Similar to [2], our FC autoencoder can be expressed as a piecewise affine function

$$\begin{aligned} f^L(\boldsymbol{x}) =& W^L \sigma(W^{L-1}\sigma(\\ & ...\sigma(W^1\boldsymbol{x} + \boldsymbol{b}^1)...) + \boldsymbol{b}^{L-1}) + \boldsymbol{b}^L \\ =& W^L \Lambda^{L-1}(\boldsymbol{x})(W^{L-1}\Lambda^{L-2}(\boldsymbol{x})(\\ & ...\Lambda^1(\boldsymbol{x})(W^1\boldsymbol{x} + \boldsymbol{b}^1)....) + \boldsymbol{b}^{L-1}) + \boldsymbol{b}^L \\ =& \Gamma\boldsymbol{x} + B, \end{aligned} \tag{2}$$

where $\Lambda^l(\boldsymbol{x}) \in \mathbb{R}^{n^l \times n^l}$, for $l = 1, ..., L-1$ are diagonal matrices defined as

$$\Lambda^l(\boldsymbol{x}) = \begin{bmatrix} \mathbb{1}(f_1^l(\boldsymbol{x}) > 0) & & \\ & ... & \\ & & \mathbb{1}(f_{n^l}^l(\boldsymbol{x}) > 0) \end{bmatrix}, \tag{3}$$

and $\Gamma \in \mathbb{R}^{H \times H}$ and $B \in \mathbb{R}^H$ are matrices defined as

$$\begin{aligned} \Gamma &= W^L(\prod_{i=1}^{L-1} \Lambda^{L-i}(\boldsymbol{x}) W^{L-i}), \\ B &= \sum_{i=1}^{L-1}(\prod_{k=1}^{L-i} W^{L+1-k}\Lambda^{L-k}(\boldsymbol{x}))\boldsymbol{b}^i + \boldsymbol{b}^L. \end{aligned} \tag{4}$$

We can further have

$$\begin{aligned} \|\boldsymbol{x} - f^L(\boldsymbol{x})\| &= \|\boldsymbol{x} - \Gamma\boldsymbol{x} - B\| \\ &\leq \|\boldsymbol{x} - \Gamma\boldsymbol{x}\| + \|B\| \\ &\leq \|I - \Gamma\|\|\boldsymbol{x}\| + \|B\|. \end{aligned} \tag{5}$$

The number of total possible variants of Γ and B is $2^{\sum_{i=1}^{L-1} n^i}$, and the specific forms of Γ and B are determined by which polytope defined as intersection of $\sum_{i=1}^{L-1} n^i$ half spaces in $\mathbb{R}^H$ the input $\boldsymbol{x}$ is in. Therefore, given an input, an upper bound of its L2 reconstruction error is definite, and is approximately proportional to its norm. This result clearly supports our claim in the main paper that *feature with smaller norm tends to have smaller $L1$ or $L2$ reconstruction loss*. For instances, given a feature with norm close to zero, its reconstruction error is approximately $\|B\|$. While for a feature with norm $\to \infty$, its reconstruction error tends to be arbitrarily larger than $\|B\|$.

2. Architecture of encoder and decoders

There are one encoder and two decoders employed in our OOD detection module, namely E, D_1 and D_2. The encoder has a FC layer with C filters (no additive bias term) and a following softmax function with temperature scaling, in which C is the number of ID classes.

Each decoder is a simple three-layer FC network with swish as activation function. Specifically, D_1 consists of sequential operations: FC layer with H filters and bias term $\to$ swish nonlinearity $\to$ FC layer with H filters and bias term $\to$ swish nonlinearity $\to$ FC layer with H filters and bias term, where H is the dimensionality of the AV feature

fed into the encoder. D_2 has the following operations: FC layer with 512 filters and bias term → swish nonlinearity → FC layer with 256 filters and bias term → swish nonlinearity → FC layer with C filters and bias term.

3. Model configurations in ablation study

Due to the limited pages in main paper, we did not detail the models applied in Ablation study (Table.2 in paper). Here we report them in order. For specific details and replication, please refer to the code submitted with this material.

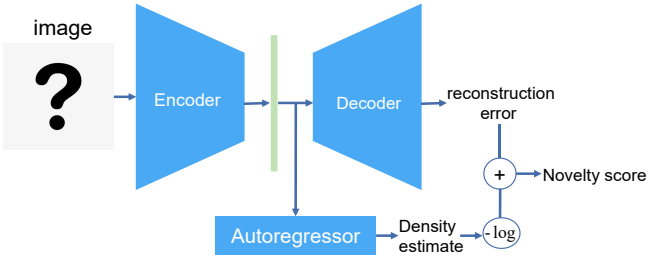

Figure 1. The overall framework of 1^{st} LSA.

1^{st} **LSA**: This is the original method of Latent Space Autoregression(LSA) [1]. Here we simply describe the framework of it, and a graphic representation is shown in Fig.1. With the training of a standard reconstruction autoencoder, this method trains an autoregressor on the latent codes output from the encoder to model the latent density while decreasing the differential entropy of the distribution of ID latent features. The input information of the autoencoder is the original image and the novelty score is computed as the summation of the L2 reconstruction error term and the negative logarithm of the density estimate produced from the autoregressor. Both terms are normalized using a validation set of ID data

$$novelty\ score = norm(L2(\boldsymbol{r}, \tilde{\boldsymbol{r}})) + norm(-log(q(\boldsymbol{r}))),$$

in which $\boldsymbol{r}$ denotes the input image, $\tilde{\boldsymbol{r}}$ the reconstruction of $\boldsymbol{r}$ and $q(\boldsymbol{r})$ the density estimate of the latent code of $\boldsymbol{r}$. The model architecture and training settings are the same as those for CIFAR-10 dataset in the original paper.

2^{nd}**-image+feature**: In this model, we change the input information of autoencoder from image to its AV feature extracted in the Wide-ResNet-28-10. Since the encoder and decoder in the original paper of LSA were designed to compress and rebuild image and therefore are parameter-redundant to be used in the case of low-dimensional AV features, we apply for this model a simpler self-designed autoencoder instead.

1. **Encoder**: FC layer with 100 filters and bias term → relu nonlinearity → FC layer with 100 filters and bias term → sigmoid nonlinearity.
2. **Decoder**: FC layer with 640 filters and bias term → swish nonlinearity → FC layer with 640 filters and bias term → swish nonlinearity → FC layer with 640 filters and bias term.
3. **Autoregressor**: Masked FC layer with 32 channels → leaky relu activation → masked FC layer with 32 channels → leaky relu activation → masked FC layer with 32 channels → leaky relu activation → masked FC layer with 32 channels → leaky relu activation → masked FC layer with 100 channels.

The applied novelty scoring function and other pipeline differ from 1^{st} **LSA** only in that AV feature is considered rather than image.

3^{rd}**-L2+NL2**: The framework of this model is identical to the above except that we use the proposed NL2 distance over the L2 distance as the $dist(\cdot)$ in the normality scoring function, which is calculated as

$$\begin{aligned} novelty\ score =& norm(NL2(\boldsymbol{v}, \tilde{\boldsymbol{v}})) \\ &+ norm(-log(q(\boldsymbol{v}))), \end{aligned}$$

where $\boldsymbol{v}$ is the extracted AV feature.

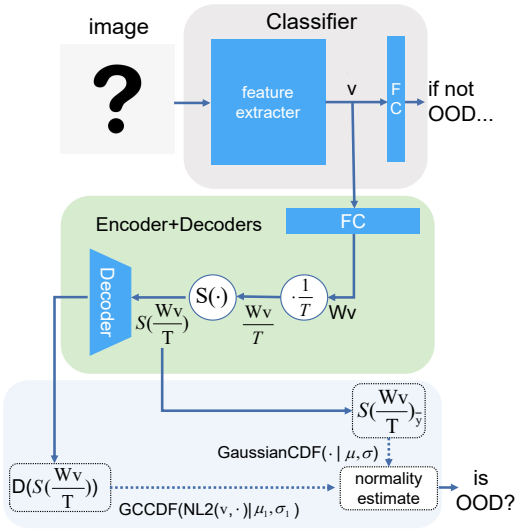

Figure 2. The overall framework of 4^{th} -AutoReg+CE.

4^{th}**-AutoReg+CE**: To restrict the span of latent representations, the models introduced above all apply a latent space autoregressor. In this model, we keep the decoder and change the encoder from that in 2^{nd}**-image+feature** to the proposed one in order to use a simple cross entropy loss to compress the latent space. Thus, in test time the normality scoring function becomes

$$\textit{normality score} = \Phi(S(\frac{W\boldsymbol{v}}{T})_{\overline{y}}|\mu, \sigma + 10\sigma) \cdot \Psi(\|\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{D(S(\frac{W\boldsymbol{v}}{T}))}{\|\boldsymbol{v}\|}\||\mu_1, \sigma_1 + 10\sigma_1).$$

The overall framework can be seen in Fig.2.

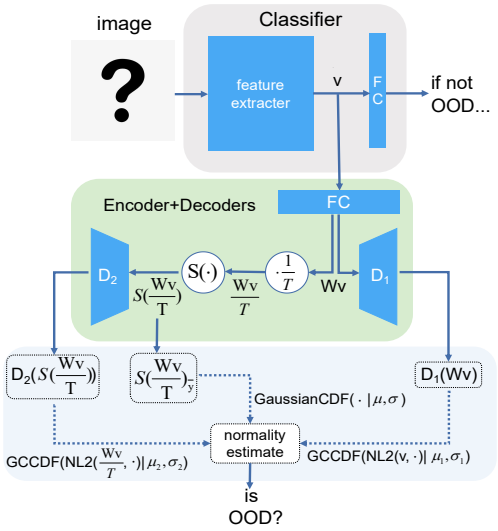

Figure 3. The overall framework of 5^{th}-basic+layerwise.

5^{th}**-basic+layerwise**: We apply the framework of layer-wise reconstruction for this model. Thus it no longer recovers the input directly from the latent space and an additional decoder D_2 is involved. In essence, this model is identical to the one applied for benchmark experiments except that we do not use ϵ terms in the normality scoring function

$$\begin{aligned}\textit{normality score} = \Phi(S(\frac{W\boldsymbol{v}}{T})_{\overline{y}}|\mu_0, \sigma_0) \\ \cdot \Psi(\|\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{D_1(W\boldsymbol{v})}{\|\boldsymbol{v}\|}\||\mu_1, \sigma_1) \\ \cdot \Psi(\|\frac{W\boldsymbol{v}}{\|W\boldsymbol{v}\|} - \frac{D_2(S(\frac{W\boldsymbol{v}}{T}))}{\|\frac{W\boldsymbol{v}}{T}\|}\||\mu_2, \sigma_2).\end{aligned}$$

The overall framework is presented in Fig.3.

6^{th}**+epsilon**: This model is the one introduced and tested in benchmark experiment. Difference with respect to the model introduced above, we add the ϵ terms into the normality scoring function to prevent it from collapsing

$$\begin{aligned}P(\boldsymbol{v} \in V) = \Phi(S(\frac{W\boldsymbol{v}}{T})_{\overline{y}}|\mu_0, \sigma_0 + \epsilon_0) \\ \cdot \Psi(\|\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{D_1(W\boldsymbol{v})}{\|\boldsymbol{v}\|}\||\mu_1, \sigma_1 + \epsilon_1) \\ \cdot \Psi(\|\frac{W\boldsymbol{v}}{\|W\boldsymbol{v}\|} - \frac{D_2(S(\frac{W\boldsymbol{v}}{T}))}{\|\frac{W\boldsymbol{v}}{T}\|}\||\mu_2, \sigma_2 + \epsilon_2).\end{aligned}$$

Other configurations keep unchanged.

References

[1] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara. Latent space autoregression for novelty detection. 2018. 2

[2] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1