Improving Discriminative Capability of Anomaly Detector using Pseudo Anomalies and Triplet Loss

Marcella Astrid[1, 2], Minsu Jang[1, 2], Seung-Ik Lee[1, 2, *]

[1]University of Science and Technology, [2]Electronics and Telecommunications Research Institute

marcella.astrid@ust.ac.kr, minsu@etri.re.kr, the_silee@etri.re.kr

[*]corresponding author

의사 이상 및 트리플렛 손실을 이용한 이상 탐지기의 분별 능력 향상

Marcella Astrid[1, 2], 장민수[1,2], 이승익[1, 2, *]

[1] 과학기술연합대학원대학교, [2] 한국전자통신연구원

Abstract

Anomaly detector models are often trained using only normal data, due to the difficulties in collecting anomalous events. Several works utilize pseudo anomalies, made from the normal data, to enforce autoencoder based model to poorly-reconstruct anomalies while well-reconstructing normal data. In this work, on top of the previously proposed model, we add triplet loss to enforce the separability between normal features and pseudo anomalous features. Experiment on Ped2 dataset shows the effectiveness of triplet loss to train a more discriminative model.

Ⅰ. Introduction

Anomaly detection is problem to separate anomalous data from the normal data. One of the applications is for detecting anomalous behaviors in surveillance videos. However, due to the difficulties in collecting anomalous data, the training of anomaly detector models is often conducted using only normal data.

One way to build anomaly detector trained using only normal data is to train an autoencoder (AE) to well-reconstruct its normal input [1-3]. Despite trained using only normal data, AE can unfortunately also well-reconstruct anomalous data, which results in the low capability to discriminate between normal and anomalous data.

Several works synthesize pseudo anomalies by augmenting the available normal training data [2,3]. The pseudo anomalies are then used to assist the training the AE so that the AE can poorly-reconstruct out-of-normal data, in addition to well-reconstruct the normal data. In this work, we extend the work from [3] and further separate the normal and abnormal features

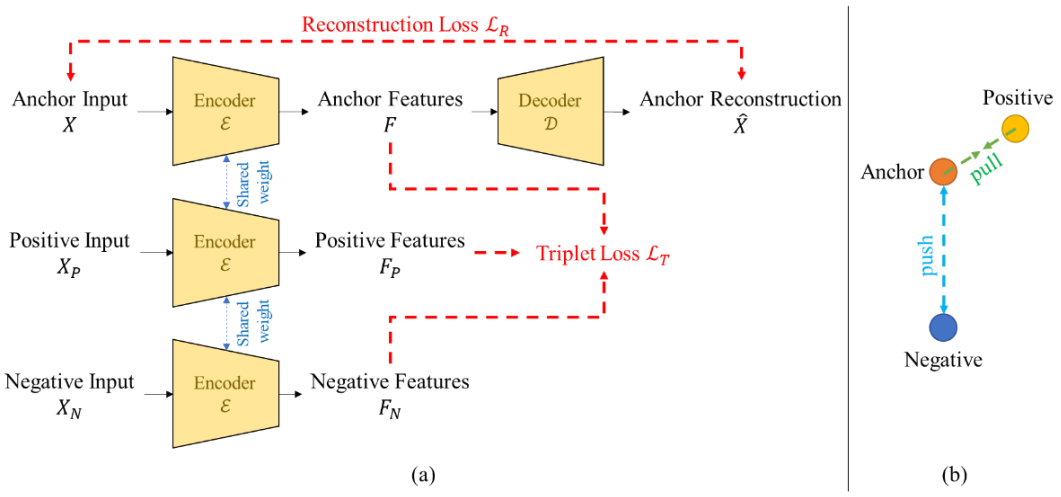

Figure 1. (a) Overall training configuration of our method. (b) Illustration of triplet loss.

using triplet loss [4]. Our experiment on Ped2 dataset [5] demonstrates that triplet loss further improves the discriminative capability of AE in detecting anomalies.

Ⅱ. Method

The overall training configuration of our method is illustrated in Fig. 1(a). The overall training loss consists of a reconstruction loss $\mathcal{L}_R$ and a triplet loss $\mathcal{L}_T$ as follows:

$$\mathcal{L} = \mathcal{L}_R + \mu\,\mathcal{L}_T,$$

where μ is weighting hyperparameter and $\mathcal{L}_R$ is same as [2], i.e. mean squared error between X and $\hat{X}$ in case of normal anchor data and negative error in case of pseudo anomalous anchor data.

To increase the discriminative capability, we add $\mathcal{L}_T$ that processes latent features of anchor F, positive F_P, and negative F_N data taken after the encoder. The features are then global average pooled $\mathcal{P}$ before the final triplet loss is calculated:

$$\mathcal{L}_T = \max\Big(0, m + d\big(\mathcal{P}(F), \mathcal{P}(F_P)\big) - d\big(\mathcal{P}(F), \mathcal{P}(F_N)\big)\Big),$$

where $d(.,.)$ is 2-norm distance and m is margin hyperparameter. Simply, as illustrated in Fig. 1(b), this loss encourages the distance between same class (anchor and positive pair) to be minimized while the distance between different classes (anchor and negative pair) to be maximized. This loss is originally used for face recognition to make the face features from same person to be close and from different person to be far [4]. The margin can stabilize the training so that it is enough when $d\big(\mathcal{P}(F), \mathcal{P}(F_P)\big)$ is already smaller than $d\big(\mathcal{P}(F), \mathcal{P}(F_N)\big)$ by m. In our method, if X comes from normal data, X_P is another normal data while X_N is the pseudo anomaly generated from the anchor. Whereas, if X is pseudo anomalous data, X_N is another pseudo anomalous data while X_P is the normal data used to synthesize X.

Ⅲ. Experiments

We utilize the same 3D-AE architecture used in [2] that takes input of size $T \times C \times H \times W = 16 \times 1 \times 256 \times 256$ and output reconstruction of the same size, where T, C, H, W are time, channel, height, and width dimensions, respectively. The latent features between the encoder and decoder (F, F_P, F_N) have size of $2 \times 256 \times 16 \times 16$. Global average pooling averages the feature in the spatial dimension and combines the time and channel dimensions, which leads to a feature of size 512. This feature is then used to calculate $d(.,.)$. We also follow [2] for the pseudo anomalies, i.e. skipping-frame with probability $p = 0.01$ and number of skipped frames $s = [2,3,4,5]$. Anomaly scores during test time is calculated using min-max normalization of the reconstruction loss, same as [2]. μ and m are set to 0.005 and 10, respectively.

We conduct the experiment in Ped2 [5], a video anomaly detection dataset. AUC (Area Under ROC Curve) comparisons with other methods can be seen in Table 1, which shows the superiority of our method against several state-of-the-art approaches. Specifically, without triplet loss, the AE yields performance of 98.4% AUC [2]. Adding triplet loss to the training boosts the performance to 98.9%, which demonstrates the effectiveness of triplet loss in improving the discriminative capability of anomaly detector.

Table 1. AUC comparisons on Ped2 dataset.

Method	AUC
AE-Conv2D [1]	90.0%
LNTRA-skip frame [3]	96.5%
STEAL [2] (ours without triplet loss)	98.4%
Ours	**98.9%**

Ⅳ. Conclusion

In this work, we propose adding triplet loss to enforce separation between normal and anomaly (represented by pseudo anomaly during training) of AE-based anomaly detector model. Experiments on Ped2 dataset shows the performance improvement of using triplet loss compared to not using it.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions).

REFERENCES

[1] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K. and Davis, L.S., 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 733-742).

[2] Astrid, M., Zaheer, M.Z. and Lee, S.I., 2021. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 207-214).

[3] Astrid, M., Zaheer, M.Z., Lee, J.Y. and Lee, S.I., 2021. Learning not to reconstruct anomalies. In *British Machine Vision Conference*.

[4] Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

[5] Li, W., Mahadevan, V. and Vasconcelos, N., 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1), pp.18-32.