2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Convolutional autoencoder-based multimodal one-class classification

Firas Laakom*, Fahad Sohrab*, Jenni Raitoharju§ †, Alexandros Iosifidis‡ and Moncef Gabbouj*
*Faculty of Information Technology and Communication Sciences, Tampere University, Finland
§Quality of Information, Finnish Environment Institute, Finland
†Faculty of Information Technology, University of Jyväskylä, Finland
‡Department of Electrical and Computer Engineering, Aarhus University, Denmark

***Abstract*—One-class classification refers to approaches of learning using data from a single class only. In this paper, we propose a deep learning one-class classification method suitable for multimodal data, which relies on two convolutional autoencoders jointly trained to reconstruct the positive input data while obtaining the data representations in the latent space as compact as possible. During inference, the distance of the latent representation of an input to the origin can be used as an anomaly score. Experimental results using a multimodal macroinvertebrate image classification dataset show that the proposed multimodal method yields better results as compared to the unimodal approach. Furthermore, study the effect of different input image sizes, and we investigate how recently proposed feature diversity regularizers affect the performance of our approach. We show that such regularizers improve performance.**

***Index Terms*—Multimodal learning, one-class classification, anomaly detection, computer vision**

I. Proposed approach

In this work, we consider the problem of one-class classification in the presence of multimodal data. We propose an approach that requires multimodal data only from the positive class in the training phase. Here, we describe the formulation for two modalities, while extension to more modalities could be easily obtained. During training, the main target of our approach is to learn a compact mutual embedding of both modalities. Let $\{(x_i, x_i')\}_{i=1}^N$ be the available training data from the positive class, where x_i and x_i' are the first and the second modality of the i^{th} sample, respectively. Our model is composed of two autoencoders, one for each modality. Let $E_1(x_i) \in \mathbb{R}^{m_1 \times p_1 \times d_1}$ be the convolutional output of the encoder on the first modality x_i and $E_2(x_i') \in \mathbb{R}^{m_2 \times p_2 \times d_2}$ be the encoder output of for the second modality. Based on these two outputs, we construct the joint latent representation ϕ_i of the sample (x_i, x_i') as follows:

$$\phi_i = \phi(x_i, x_i') = \text{concat}(\text{Flat}(E_1(x_i), \text{Flat}(E_2(x_i'))), \quad (1)$$

where $\text{Flat}(\cdot)$ is the flattening operation, i.e, it flattens $E_1(x_i) \in \mathbb{R}^{m_1 \times p_1 \times d_1}$ into a $m_1 p_1 d_1$-dimensional vector. concat is the vector concatenation operation compiling $\phi(x_i, x_i') \in \mathbb{R}^{(m_1 p_1 d_1 + m_2 p_2 d_2)}$ as the final representation of the input.

Our main aim is to learn to map the input data from the positive class into a compact space. This is an objective commonly used by regression-based OCC models and it is usually expressed by minimizing the distance of the latent representations to a pre-defined point [3]–[5]. Using as the target point the origin, the objective becomes to minimize their L_2-norm: $\frac{1}{N}\sum_{i=1}^N ||\phi_i||^2$. By minimizing this loss, the model learns to map the samples from the positive class into a hypersphere centered at the origin. In the test phase, any sample that falls close to the origin is assigned to the positive class, and the rest are classified as anomalies.

Minimizing the aforementioned loss can lead to a degenerate solution, i.e., the model learns to map all inputs to the origin and thus fails to distinguish between positive and anomalous samples. To avoid obtaining such solutions, we propose to augment our model using two decoders (one for each modality), aiming to learn to reconstruct the inputs. The outputs of the encoders $E_1(x_i)$ and $E_2(x_i')$ are passed through to the decoders D_1 and D_2. To incorporate the reconstruction objective into the training, we propose to augment the loss in (**??**) using the mean squared loss. The final loss used to train the network can be expressed as follows:

$$L := \frac{1}{N}\sum_{i=1}^N \Big(||\phi_i||^2 + (||D_1(E_1(x_i)) - x_i||^2 + ||D_2(E_2(x_i')) - x_i'||^2) \Big). \quad (2)$$

The weights of E_1 and E_2 are shared as well as the weights for D_1 and D_2, and they can be trained in an end-to-end manner using gradient-based optimization by minimizing (2). The first term of the loss forces the model to learn a compact representation for both modalities of the same sample in the bottleneck of the architecture, while the other terms regularize the model to avoid degenerate and undesired solutions.

In the test phase, we discard the decoder part. Given a test sample (y, y'), we compile its feature output ϕ, as expressed in (1). The distance of the latent representation of the data point from the origin can be used as an anomaly score. Based on this distance, we can assign the sample to the corresponding class (positive class or negative class). To this end, we need to determine a threshold τ, which is used to define the hypersphere enclosing the positive class. The value of τ can be obtained using the training data. As all the training data is

TABLE I
THE RECALL, THE PRECISION AT RANK N (P@N), AND THE AREA UNDER THE CURVE (ROC) OF THE THREE DIFFERENT MODELS TESTED ON THE FOUR ONE-CLASS TASKS. WE ALSO REPORT THE AVERAGE PERFORMANCE OVER THE FOUR TASKS FOR EACH MODEL.

Normal Class	unimodal (left)			unimodal (right)			ours (multimodal)		
	Recall	P@n	ROC	Recall	P@n	ROC	Recall	P@n	ROC
Leptophlebia sp.	0.963	0.921	0.806	0.888	0.897	0.535	0.938	0.921	0.742
Baetis rhodani	0.961	0.893	0.428	0.882	0.892	0.330	0.961	0.907	0.582
Elmis aenea larva	0.844	0.900	0.367	0.896	0.900	0.456	0.935	0.899	0.427
Oulimnius tuberculatus larva	0.896	0.897	0.444	0.909	0.919	0.646	0.896	0.897	0.560
Average	0.916	0.903	0.511	0.894	0.902	0.49	**0.932**	**0.906**	**0.578**

from the positive class, τ can be set to 95^{th} percentile of the feature norms of the training data $\{\phi_i\}_{i=1}^{N}$. Then, given the test sample (y, y'), if $\phi(y, y') \leq \tau$, it is considered to be from the target class. Otherwise, it is considered an anomaly.

It should also be noted that, although in this work, we use shared weights for E_1 and E_2, it is possible to use a different model for every modality. However, one-class classification tasks usually have scarce data [1], [6]. Using shared weights reduces the total number of parameters and acts as a regularization, which makes our model suitable for learning from a limited amount of data.

II. EXPERIMENTS AND DISCUSSION

In this section, we evaluate the performance of the proposed multimodal one-class classification method.

We used a subset of the multimodal image classification dataset of benthic macroinvertebrates, FIN-Benthic [2]. In particular, we used data from 4 classes. Each sample point is presented with two RGB images (which act as the two modalities) from two perpendicular viewpoints. Using this dataset, we constructed four different one-class classification tasks. In each task, data from a single class (out of the four) is considered the normal class, and the remaining three classes are combined to form the anomaly class. In each of the four experiments, we used 66% of normal class data as training data, and we held the rest along with the anomaly data (the remaining three classes) as our test data. All the images were resized to 32×32 pixels.

Our implementation is based on [7]. To train our models, we use Adam optimizer with a learning rate of 0.001 and weight decay of 10^{-3}. The training is conducted with 4 epochs and a batch size of 32. The input image size is 32×32 pixels. For the encoders E_1 and E_2, we used a fully convolutional model which consists of three blocks of convolution, batch normalization, *maxpooling*, and *dropout* layers. All the convolution filters have a size of 3×3 and were selected to be 64, 32, and 16 in the first, second, and third layers, respectively. For the decoder part, i.e., D_1 and D_2, we used the corresponding symmetric layers.

To test the hypothesis that multimodal learning helps in the context of one-class classification, we also experimented with the unimodal variant of the method, i.e., using only one branch of the model and using images from one modality. This yields two competing methods, namely unimodal (left) and unimodal (right), for the left and right modalities, respectively.

In Table I, we report the results for the multimodal model along with the two unimodal models on the four one-class classification tasks. We also report the average results over the four tasks. For each method, we report the Recall scores, the Precision at rank n (P@n), and the Area Under the Curve (ROC) [7].

As shown in Table I, multimodal learning, indeed, yields better performance compared to both unimodal cases in all three metrics. For instance, in the average performance, the proposed multimodal model yields 0.067 and 0.088 improvement in ROC compared to the unimodal models using the left right images, respectively. We also note that on the four tasks, the worst ROC achieved by unimodal (left) and unimodal (right) models are 0.367 and 0.330, respectively, whereas for the multimodal model, the lowest ROC corresponds to *Elmis aenea larva* and is equal to 0.427.

ACKNOWLEDGEMENT

This work was supported by the NSF-Business Finland Center for Visual and Decision Informatics (CVDI) project AMALIA. Foundation for Economic Education (Grant number: 220363) funded the work of Fahad Sohrab at Haltian. The work of Jenni Raitoharju was funded by the Academy of Finland project TIMED (project 333497).

REFERENCES

[1] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis, "Subspace support vector data description," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 722–727.

[2] Jenni Raitoharju, Ekaterina Riabchenko, Iftikhar Ahmad, Alexandros Iosifidis, Moncef Gabbouj, Serkan Kiranyaz, Ville Tirronen, Johanna Ärje, Salme Kärkkäinen, and Kristian Meissner, "Benchmark database for fine-grained image classification of benthic macroinvertebrates," *Image and Vision Computing*, vol. 78, 2018.

[3] Qian Leng, Honggang Qi, Jun Miao, Wentao Zhu, and Guiping Su, "One-class classifcation with extreme learning machine," *Mathematical Problems in Engineering*, vol. 2015, Article ID 412957, 2015.

[4] Alexandros Iosifidis, Vasileios Mygdalis, Anastasios Tefas and Ioannis Pitas, "One-class classification based on extreme learning and geometric class information," *Neural Processing Letters*, vol. 45, no. 2, pp. 577–592, 2017.

[5] Haozhen Dai, Jiuwen Cao, Tianlei Wang, Muqing Deng and Zhixin Yang, "Multilayer one-class extreme learning machine," *Neural Networks*, vol. 115, pp. 11–22, 2009.

[6] Naeem Seliya, Azadeh Abdollah Zadeh, and Taghi M Khoshgoftaar, "A literature review on one-class classification and its potential applications in big data," *Journal of Big Data*, vol. 8, no. 1, pp. 1–31, 2021.

[7] Yue Zhao, Zain Nasrullah, and Zheng Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, 2019.