Deep Context-Aware Novelty Detection

Ellen Rushe
School of Computer Science
University College Dublin
ellen.rushe@insight-centre.org

Brian Mac Namee
School of Computer Science
University College Dublin
brian.macnamee@ucd.ie

Abstract

A common assumption of novelty detection is that the distribution of both “normal" and “novel" data are static. This, however, is often not the case—for example scenarios where data evolves over time or where the definition of normal and novel depends on contextual information both lead to changes in these distributions. This can lead to significant difficulties when attempting to train a model on datasets where the distribution of normal data in one scenario is similar to that of novel data in another scenario. In this paper we propose a context-aware approach to novelty detection for deep autoencoders to address these difficulties. We create a semi-supervised network architecture that utilises auxiliary labels to reveal contextual information and allow the model to adapt to a variety of contexts in which the definitions of normal and novel change. We evaluate our approach on both image data and real world audio data displaying these characteristics and show that the performance of individually trained models can be achieved in a single model.

1 Introduction

Novelty, or anomaly, detection is often framed as a task where the definition of “normal", or data that has been “seen" before, is static. Furthermore, the nature of what is “novel" is often also assumed to be fixed. In real world scenarios, however, these assumptions are often invalid: the nature of normality and novelty may evolve as a function of time or depend on the *context* in which data is observed [2]. This paper addresses the latter scenario. There are many situations where an event that is normal in one context, may be considered novel in another. For example the sound of an ice-cream truck on a sunny day versus the same sound in the middle of the night. The notion of context is particularly useful when one considers the lack of labelled data inherent in novelty detection. In many novelty detection scenarios it is not practical to build individualized models for each context and so it would be useful to have an effective way to build novelty detection models that can effectively utilize data from multiple contexts to build a single, accurate model. Such a model would exploit two types of features: *behavioral features* that are ubiqitous across all contexts and *contextual features* that depend on context [2]. This motivates a semi-supervised approach where the detection algorithm is *conditioned* on context, and therefore context-aware.

In this paper we propose the *Context-Aware Novelty Detection autoEncoder* (CANDE), a novel approach to detecting contextual anomalies based on conditioned autoencoder neural networks. In experiments using two datasets, including a large real world dataset, we show that shifts in context degrade detection but that a single CANDE model can alleviate this degradation, removing the necessity for individual models to be trained for each context.

2 Related Work

There is a significant amount of existing work using deep learning for novelty detection. Golan et al. [4] use a discriminative learning strategy for detecting anomalies in images by training a network

1st NeurIPS workshop on Interpretable Inductive Biases and Physically Structured Learning (2020), virtual.

to recognise transformations applied to images, and flagging anomalies when these transformations cannot be recognised. Pidhorskyi et al. [11] combined adversarial losses with reconstruction error in order to compute the likelihood of samples being generated from an inlier distribution. Gong et al. [5] use a memory augmented network to learn a fixed number of sparse representations of the normal data via a content addressable memory and retrieve these representations at test time via an attention mechanism. This way, only normal representations learned at training time can be retrieved, resulting in a poor reconstruction error for anomalous queries. Nguyen et al. [8] addressed the issue of high likelihood being assigned to anomalous regions using a variational autoencoder that learned multi-model distributions using multiple hypothesis networks [14] along with a discriminator network that prevented the network from assigning high likelihood to non-existent input regions.

There is little research on deep contextual novelty detection. Shulman [16] used variational autoencoders to reconstruct contextual and behavioural features separately. A continual learning approach using generative replay [15] as a data augmentation strategy was taken by Wiewel et al. [17] for streaming data scenarios where past data is unavailable when retraining models on new incoming data. Araya et al. [1] used historical sensor data alongside contextual features with autoencoders for novelty detection in smart buildings. Komatsu et al. [6] proposed a scene-dependent acoustic event detector. This used *I-vectors*, a low-dimensional embedding based on factor analysis of the difference between a Universal Background Model and a short audio segment-specific model, as an additional input to a WaveNet [9] model. This work is perhaps closest the proposal in this paper. Our proposed approach however, is not domain-specific and differs in both the method of embedding and the conditioning mechanism used.

3 Context-Aware Novelty Detection Autoencoder

We propose the Context-Aware Novelty Detection autoEncoder (CANDE), an approach that dynamically adapts the output of a standard deep novelty detector to contextual features. The advantage of this architecture is that it allows a single model to be trained for disparate contexts. The architecture includes two components: a deep autoencoder and a contextual encoding function. We define a deep autoencoder with an encoder network $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$, with network weights θ^e where m is the dimension of the encoding $\boldsymbol{z}$; and a decoder network $g : \mathbb{R}^m \rightarrow \mathbb{R}^d$ with network weights θ^d. To encourage this autoencoder to adapt to contextual information, and thereby modulate its output depending on context, the network is conditioned using an auxiliary context label. The auxiliary context label denotes some form of partitioning of the dataset which indicates the context from which the data was derived. Context is recorded in a training dataset through a set of auxiliary labels, $\langle C_i \rangle_{i \in I}$. Crucially, these labels do not contain any information about the nature of anomalies but only serve as a more fine-grained representation of each normal example used to train the network. To condition the deep autoencoder in CANDE, we use the Feature-wise Linear Modulation (FiLM) [10] network conditioning strategy.[1]. This is illustrated in Figure 1. An affine transformation is applied the kth layer of the network, denoted by $\boldsymbol{z}_k \in \mathbb{R}^q$, using scaling and shifting factors $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$. These factors are derived using the context vector $\boldsymbol{h}_c \in \mathbb{R}^p$ as follows:

$$\boldsymbol{\gamma}_k = \boldsymbol{h}_c W_{\gamma k} + \boldsymbol{b}_{\gamma k} \quad (1) \qquad \boldsymbol{\beta}_k = \boldsymbol{h}_c W_{\beta k} + \boldsymbol{b}_{\beta k} \quad (2)$$

where $W_{\gamma k}$ and $W_{\beta k}$ are weight matrices corresponding to $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$, and $\boldsymbol{b}_{\gamma k}$ and $\boldsymbol{b}_{\beta k}$ represent their respective bias vectors. The context vector $\boldsymbol{h}_c$ can be a one-hot-encoding of the context label or a more complex representation, such as an embedding. The transformations in equations 1 and 2 are performed such that their output will be of dimension q, thus aligning with that of the layer to be conditioned, $\boldsymbol{z}_k$. Conditioning is performed using an element-wise affine transformation [10]:

$$\boldsymbol{z}'_k = \boldsymbol{\gamma}_k \odot \boldsymbol{z}_k + \boldsymbol{\beta}_k \quad (3)$$

After this transformation, the output of this layer, $\boldsymbol{z}'_k$, is passed through a ReLU [3] non-linearity. With conditioning, the encoder and decoder network of the autoencoder, f and g, are now not only functions of their inputs, $\boldsymbol{x}$ and $\boldsymbol{z}'$ respectively, and their learned parameters, but also $\boldsymbol{h}_c$:

$$\boldsymbol{z}' = f(\boldsymbol{x}, \boldsymbol{h}_c; \theta^e) \quad (4) \qquad \boldsymbol{x}' = g(\boldsymbol{z}', \boldsymbol{h}_c; \theta^d) \quad (5)$$

[1]This layer-by-layer conditioning is also in contrast with a previous approach to autoencoder conditioning proposed by [13] where simply the layer before the "bottleneck" is conditioned.

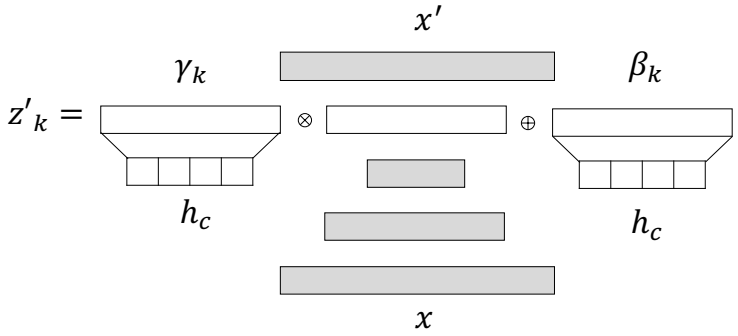

Figure 1: For a layer k, a label corresponding to the context of a particular example is transformed according to Equations 1 and 2. The conditioned layer corresponds to the layer after being scaled by γ_k and shifted by β_k. These operations occur at each layer in the network.

4 Experiments

We use two datasets in our experiments: a novelty detection dataset based on the MNIST dataset and the MIMII dataset, a large dataset capturing a real-world novelty detection scenario using audio recorded from industrial machines.

MNIST: We replicate contextual novelties artificially using the MNIST dataset [7]. This is done in order to mimic the phenomenon of different contexts in novelty detection in a controlled setting where we can observe the effects of conditioning more clearly. The MNIST dataset is partitioned into different "contexts". A context denotes a set of classes from the dataset. This set of classes is considered "normal" for that context. Let $\langle C_i \rangle_{i \in I}$ denote a set of different contexts in dataset D, where $\forall i, j \in I : i \neq j \Rightarrow C_i \cap C_j = \emptyset$. To simulate contextual anomalies, we then relabel data as being 'normal' or 'novel' and create scenarios where in one context images of a specific digit are 'novel', whereas in another context images of the same digit are considered 'normal'. This is illustrated in Figure 2.

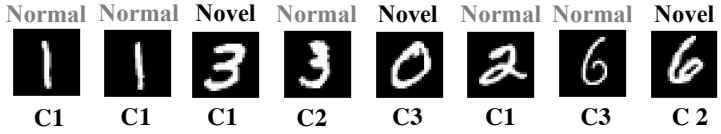

Figure 2: Example of MNIST dataset configuration. C1, C2 and C2 refer to Context 1, 2 and 3.

MIMII: As an example of a real-world novelty detection problem, we use the recently proposed MIMII public dataset [12] which consists of industrial machine sounds recorded from a number of different machine types and models. The task requires novel audio to be detected in 10-second audio segments, where a novel audio segment might indicate a machine about to fail. As in the original benchmarks in [12], in our experiments we use a representation of audio segments that consists of five frames of 64 log-mel spectrogram filters derived from the first channel of audio with a frame size of 1,024 and a hop-length of 512. In the public MIMII dataset four different types of machine were recorded: 'valve', 'pump', 'fan' and 'slide rail'. The machine types and models define the different contexts in this scenario. Recordings from four different models of each machine are included.

For all autoencoder models mean squared error (MSE) is used as the cost function. To calculate a novelty score, the MSE is calculated between a test example, $\boldsymbol{x}_i$, and its reconstruction, $\hat{\boldsymbol{x}}_i$. We base our evaluation on the Area Under the Receiver Operator Characteristic (ROC) Curve (AUC). This measures the ability of the models to generate accurate novelty scores without requiring a novelty classification threshold to be set. Three different types of models are compared in our experiments:

Individual Models: For each context, an individual autoencoder model was trained in order to ascertain the best performance possible given a model trained on a particular context without intrusion of other normal contexts.

Unconditioned Combined Models: An Unconditioned Combined Model takes all data from all contexts and combines them to train a single novelty detection model. In the case of MNIST this means that, in the test set, there will be instances where two examples that are drawn from the same distribution (for instance two "0" digits) will have opposite labels. This represents an extreme overlap between normal and novel distributions. In the case of the MIMII data, this effect is more subtle, as the combined dataset simply includes examples from all machine models. We use unconditioned combined models as a benchmark against which to compare conditioned models.

One-hot-encoded CANDE Combined Models: A one-hot-encoded vector containing the ID of each context was used for conditioning in the CANDE models used in this experiment.

5 Results & Conclusion

Table 1: Each model was trained and evaluated with ten different random weight initializations for MNIST and three for the MIMII dataset. (a) AUC for MNIST dataset with three artificially created contexts. (b) AUC shows results for the MIMII dataset. AUCs are calculated for each machine type and ID and averaged over three different signal-to-noise ratios.

(a)

Normal digits	Novel digit	AE separate (reference)	AE uncond.	CANDE one hot
0, 1, 2	3	0.945	0.609	0.921
3, 4, 5	6	0.944	0.491	0.916
6, 7, 8	0	0.935	0.518	0.893

(b)

Model	ID	AE separate (reference)	AE uncond.	CANDE one hot
fan	00	0.663	0.614	0.647
	02	0.850	0.747	0.845
	04	0.748	0.714	0.764
	06	0.930	0.773	0.904
pump	00	0.609	0.466	0.626
	02	0.519	0.426	0.430
	04	0.950	0.737	0.938
	06	0.805	0.647	0.775
slider	00	0.973	0.968	0.964
	02	0.860	0.796	0.847
	04	0.765	0.755	0.840
	06	0.625	0.668	0.636
valve	00	0.540	0.362	0.507
	02	0.619	0.647	0.635
	04	0.625	0.504	0.594
	06	0.651	0.565	0.590

Table 1 (a) shows the performance of the different models trained on MNIST. The separate autoencoder models trained for each context perform very well, however, there is a huge degradation in performance when the combined model is trained without conditioning. This is not surprising given the degree of overlap between the novel and normal classes in the three different contexts. Comparing the results of CANDE to the unconditioned combined model shows the clear improvement provided by contextual conditioning, and demonstrates that performance similar to that achieved with three separate models is possible using a single model with conditioning.

The MIMII dataset provides a more realistic view of novelty detection in real world scenarios with data arising from a number of sources. Table 1 (b) shows the results on the MIMII dataset. The need for conditioning is clearly illustrated by the degradation between the performance of the separate models and the performance of the combined models without conditioning, though not as extreme as in MNIST. In all but a few cases, CANDE models outperform unconditioned models. We have therefore been able to show that context-aware architectures clearly outperform their unconditioned counterparts in nearly all cases, especially where there is a high degree of overlap between normal and novel labels. The results from both datasets demonstrate that this architecture can recover much of the performance lost by training a single model on all data combined and in some cases can even out-perform individually trained models.

Acknowledgments This work was supported by Science Foundation Ireland under Grant No. 15/CDA/3520 and Grant No. 12/RC/2289_P

References

[1] Daniel B Araya, Katarina Grolinger, Hany F ElYamany, Miriam AM Capretz, and G Bitsuamlak. Collective contextual anomaly detection framework for smart buildings. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 511–518. IEEE, 2016.

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[4] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9781–9791, 2018.

[5] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.

[6] Tatsuya Komatsu, Tomoki Hayashiy, Reishi Kondo, Tomoki Todaz, and Kazuya Takeday. Scene-dependent anomalous acoustic-event detection based on conditional wavenet and i-vector. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 870–874. IEEE, 2019.

[7] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[8] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, pages 4800–4809. PMLR, 2019.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[10] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems*, pages 6822–6833, 2018.

[12] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv:1909.09347*, 2019.

[13] Jan Rudy and Graham Taylor. Generative class-conditional autoencoders. *arXiv preprint arXiv:1412.7009*, 2014.

[14] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017.

[15] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2017.

[16] Yaniv Shulman. Unsupervised contextual anomaly detection using joint deep variational generative models. *arXiv preprint arXiv:1904.00548*, 2019.

[17] Felix Wiewel and Bin Yang. Continual learning for anomaly detection with variational autoencoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3837–3841. IEEE, 2019.