ANOMALY DETECTION IN RAW AUDIO USING DEEP AUTOREGRESSIVE NETWORKS

Ellen Rushe, Brian Mac Namee

Insight Centre for Data Analytics,
University College Dublin

ABSTRACT

Anomaly detection involves the recognition of patterns outside of what is considered normal, given a certain set of input data. This presents a unique set of challenges for machine learning, particularly if we assume a semi-supervised scenario in which anomalous patterns are unavailable at training time meaning algorithms must rely on non-anomalous data alone. Anomaly detection in time series adds an additional level of complexity given the contextual nature of anomalies. For time series modelling, autoregressive deep learning architectures such as WaveNet have proven to be powerful generative models, specifically in the field of speech synthesis. In this paper, we propose to extend the use of this type of architecture to anomaly detection in raw audio. In experiments using multiple audio datasets we compare the performance of this approach to a baseline autoencoder model and show superior performance in almost all cases.

Index Terms— anomaly detection, deep learning, raw audio, WaveNet

1. INTRODUCTION

The use of semi-supervised algorithms in anomaly detection is motivated by the scarcity of anomalous events as they are either very rare or completely unseen. When detecting anomalies in time series, the task is further complicated by the temporal structure of the data [2]. In this case, the task tends to fall under the categorization of *collective* anomaly detection, where single data points tend not to hold much context or information, rather it is a collection of data points that display a novel or anomalous pattern [1]. This is especially the case when time series are collected using high sampling rates, for instance in audio which is typically sampled at several thousand samples per second.

Recent deep learning approaches to anomaly detection in this type of time series have typically used recurrent predictive models [16] [19] and autoencoders [22] [14] [18]. Although recurrent neural networks (RNN) such as Long Short Term Memory networks (LSTM) [8] are a natural choice for sequential data, an inherent drawback of these models is the difficulty to parallelize backpropagation though time, which can slow training, especially over very long sequences. This drawback has given rise to convolutional autoregressive architectures [24]. These models are highly parallelizable in the training phase, meaning that larger receptive fields can be utilised and computation made more tractable due to effective resource utilization. In this paper we adapt WaveNet [24], a robust convolutional autoregressive model originally created for raw audio generation, for anomaly detection in audio. In experiments using multiple datasets we find that we obtain significant performance gains over deep convolutional autoencoders.

The remainder of this paper proceeds as follows: Section 2 surveys recent related work on the use of deep neural networks for anomaly detection; Section 3 describes the WaveNet architecture and how it has been re-purposed for anomaly detection; Section 4 describes the setup of an experiment in which we compare the performance of the WaveNet anomaly detector to a baseline autoencoder approach across 15 datasets; Section 5 discusses the results of this experiment; and, finally, Section 6 summarises the paper and suggests directions for future work.

2. RELATED WORK

Autoencoders have been extensively utilised in anomaly detection and, more broadly, in novelty detection. In [16] a comprehensive evaluation was performed on autoencoders for acoustic novelty detection where novel sounds in various environments were detected. Although lacking evaluation on convolutional architectures, it was found that bidirectional LSTM autoencoders with denoising and non-linear prediction (prediction of current points using a previous point or points) performed best when compared to similar fully connected models and other statistical approaches—one-class support vector machine, Gaussian Mixture Models and Hidden Markov Models. A variation of adversarial autoencoders [15] was used for acoustic novelty detection of sounds in home environments by [22]. In this variation, the discriminator was used to discriminate between data reconstructed by an autoencoder and data from the training set. Convolutional autoencoders have also been used on spectrogram data to detect anomalies in machine sound [18].

This work has been supported by a research grant by Science Foundation Ireland under grant number SFI/15/CDA/3520.

Beyond audio, LSTMs and a scaled down version of WaveNet have also been utilised as predictive models for anomaly detection in radio signals [19]. A denoising autoencoder combined with RNNs was used to detect outliers in sequential data in [14]. Here the denoising autoencoder acts as a feature extractor while the RNN models the sequential structure in these features. This model first pre-trained the denoising autoencoders before training an RNN with the learned features, and finally performing finetuning on the model. An ensemble of autoencoders was used in [11] to detect unseen falls from wearable sensor data. In this work, channel-wise autoencoders were trained on raw accelerometer and gyroscope data.

Outside of anomaly detection, as part of an effort to explore the amount of feature engineering necessary before modelling speech signals in raw audio, [23] showed that deep networks were capable of building rough estimations of the appropriate bandpass filters that would normally be applied during pre-processing. It was also found that increased training data and deeper networks led to improved performance. Convolutional layers, however, were found to outperform fully connected layers for feature learning on the TIMIT corpus [5] by [21], and for Large Vocabulary Continuous Speech Recognition by [7]. Convolutional layers were also found to effectively learn filter banks on a large speech dataset with artificially added noise in [9]. More recently, convolutional autoregressive architectures such as WaveNet have shown promising results in raw audio generation [24] and music synthesis [4]. Motivated by these recent successes, we apply a WaveNet model to the domain of anomaly detection in raw audio.

3. ANOMALY DETECTION WITH WAVENET

WaveNet [24] is an autoregressive approach to speech synthesis based on the PixelRNN [25] and PixelCNN [26] approaches developed for high resolution image generation. In a WaveNet model, dilated causal convolutions are used in order to increase the receptive field of convolution filters, and rectified linear activations are replaced with gated units in order to provide the benefits of LSTM models without the need for a recurrent algorithm. The receptive field can also be further increased using multiple stacks of dilated convolutions. In this model samples are generated one at a time, based on a softmax distribution over all possible sample values. This is made tractable by quantizing the inputs to be 8-bit, leading to 256 possible values on the outputs for each sample.

Neural autoregressive models obtain a probability distribution over all outputs by conditioning each on all preceding outputs, assuming a sequential structure in the data [12][6]. The product rule allows the output distribution to be factorized into the product of conditionals such that

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, ..., x_{t-1}) \tag{1}$$

where x_t denotes an output at time t [24] [4]. These types of autoregressive models are a form of generative model where a given input variable is conditioned on all preceding data points. In this way, this formulation is similar to recurrent architectures, except that each output can be computed in parallel during training, which leads to significant computational efficiency, as these models are easily parallelizable.

The sum of the dilation rates R_s for a single stack s in a WaveNet model is given by

$$R_s = \sum_{i=0}^{|s|-1} 2^i \tag{2}$$

where $|s|$ is the number of layers in stack s. The sum of dilation rates for each stack can then be used to calculate the total receptive field, R_T, for a model with m stacks

$$R_T = 1 + (k-1) \sum_{s=1}^{m} R_s \tag{3}$$

where k is the kernel size. Samples at the very beginning of a sequence will not have access to past values and are therefore generally zero padded.

As causal convolution is used in the WaveNet model, each new timestep may only be computed based on previous timesteps. At the beginning of the sequence, if causality is imposed used zero padding, points at the beginning of the sequence will have less information with which to compute their outputs.

To use a WaveNet architecture for semi-supervised anomaly detection, we train the network to predict the next sample in a sequence using only normal data. This means that the network will learn a conditional distribution across normal sequences and that anomalous sequences should not follow this distribution. To recognise anomalies, sequences are presented to the trained network and the distances between the predictions generated by the network and the subsequent actual values in the sequences are calculated. A small distance is indicative of a normal sequence while a large distance is indicative of an anomaly. Mean squared error is used to calculate distance in our implementation.

One of the noted drawbacks of WaveNet is the time taken for generation, as samples are generated sequentially one at a time. There have been a number of solutions proposed to accelerate this process [27] [20]. In the context of anomaly detection, at prediction time an entire window is presented to the algorithm at once meaning that all outputs can be computed in parallel in the same way as training, leading to fast detection of anomalies.

4. EXPERIMENTAL SETUP

In this section we describe our experimental setup including the datasets used, the specific architectures implemented; and the performance measures used.

4.1. Datasets

We extract 15 datasets from the 2017 DCASE Challenge Task 2 dataset [17]. This dataset consists of three different classes of rare event—*babycry*, *glassbreak* and *gunshot*—that have been artificially mixed with background audio from 15 different environmental settings (e.g. *beach*, *restaurant*, and *train*). Rare events were mixed at different "event-to-background" ratios which is defined as the ratio between the root mean square energy of the rare event and the background audio [17]. The original audio mixtures were sampled at 44.1KHz with 24-bit bit depth. We use the original mixture audio files from the challenge, which contains 491, 496 and 500 audio files of roughly 30 seconds each for training, validation and testing respectively.

Assuming that events may be more difficult to detect in some environments than others, we examine the performance of these models across multiple contexts, leading to 15 datasets, one for each type of environment, see Table 1. As we take a semi-supervised approach, we train a single model on only data from a single environment that contains no anomalous examples, and consider an anomaly to be an instance of one of the rare event sounds. For each audio file in a given scene, we window the training set with a window size of 4096 samples with no overlap. We discard any windows containing anomalies in the training phase. A window is considered anomalous if most of the samples from that window are from any of the three rare event classes.

4.2. Baseline Convolutional Autoencoder

We use a convolutional autoencoder (CAE) as a baseline model against which to compare the performance of the WaveNet model. The use of a convolutional autoencoder as a robust baseline is motivated by previous research showing the ability of convolutional architectures to model raw audio [9][7][21]. The CAE is composed of 20 layers (10 layer encoder, 10 layer decoder), matching the number of layers in the WaveNet model, with a kernel size of 3 throughout. The data is scaled between -1 and 1. Strided convolution is used for downsampling, with a stride of 2 being applied at every second convolutional layer in the encoder. For upsampling transposed convolution otherwise known as *fractionally strided convolution* [3], is used—see Figure 1. For each audio window, the mean squared error between the output of the network and the example originally input into the network is computed, i.e. the reconstruction error.

Table 1. Number of examples (windows) in each subset of data. Data is highly imbalanced in favor of the normal class.

Scene	# train	# test	% anomalies (test)
beach	34 768	37 352	2.11
bus	26 482	33 810	2.27
cafe/restaurant	36 097	30 590	2.29
car	30 303	36 386	1.70
city center	34 343	33 488	2.16
forest path	27 132	30 590	1.99
grocery store	35 336	28 658	1.79
home	29 774	30 590	2.46
library	34 054	32 200	2.07
metro_station	33 153	28 980	2.18
office	32 985	34 454	2.88
park	28 245	30 590	2.29
residential_area	29 524	30 268	1.87
train	28 107	35 742	1.80
tram	24 894	29 302	2.19

Fig. 1. **Baseline Convolutional autoencoder.** Input is downsampled into an encoding of size 1x64 along the time dimension using strided convolution in every second layer. Encoding is upsampled using fractionally strided transposed convolution.

4.3. WaveNet model

The WaveNet model uses two stacks of 10 layer causal dilated convolutions, leading to a total of 20 layers. Residual and skip connections are used along with an exponentially growing dilation rate in each stack, as per the original WaveNet paper [24]. Data is quantized using μ-law companding and scaled between -1 and 1 for the inputs. The softmax distribution of the quantized integer range (256 values) for each sample is then generated. The number of filters in each layer is based on those found to be optimal in previous literature [4] with 512 filters in skip connections, and a further 256 filters for residuals. Cross-entropy is minimized during training however the reconstruction error is measured at testing time using mean-squared error. The implementation of this model

is based on code from [4][1] and also with reference to other open source implementations [2] [3]. The code for our particular implementation can be found on GitHub [4].

4.4. Evaluating Model Performance

To evaluate the model we train both the CAE and WaveNet model using training sets containing normal examples only, and evaluate their performance on hold-out test sets (as described in Table 1). In the testing phase the reconstruction error, in the form of mean squared error, is computed for each example. Using these reconstruction errors as an anomaly detection signal we can compute the Area Under the ROC Curve (AUC) [10] to show the ability of the models to recognize anomalies.

5. RESULTS

The performance of the two models across the 15 datasets is shown in Table 2. We find that the WaveNet models consistently outperform the baseline CAE models in almost all datasets, with a tie in the *home* and *office* scenarios. It is also noteworthy that performance of both models noticeably varies across the different acoustic scenarios, indicating that the ability of the models to detect anomalies can be significantly affected by different acoustic environments.

Table 2. AUC scores for both models on each dataset.

Scene	CAE	WaveNet
beach	0.69	**0.72**
bus	0.79	**0.83**
cafe/restaurant	0.69	**0.76**
car	0.79	**0.82**
city center	0.75	**0.82**
forest path	0.65	**0.72**
grocery store	0.71	**0.77**
home	**0.69**	**0.69**
library	0.59	**0.67**
metro station	0.74	**0.79**
office	**0.78**	**0.78**
park	0.70	**0.80**
residential area	0.73	**0.78**
train	0.82	**0.84**
tram	0.80	**0.87**

6. CONCLUSIONS AND FUTURE WORK

In this paper we have adapted the WaveNet architecture to the domain of acoustic anomaly detection. We find that we obtain significant performance gains over standard convolutional autoencoders across multiple datasets.

In the future we intend to explore building a single model that will operate across all of the scenarios in the DCASE Challenge Task 2 datasets by using conditioning [24] to adapt models to different contextual environments. This will be a first step towards exploring the use of conditioning to build models that can address the issue of concept drift [13].

[1] https://github.com/tensorflow/magenta/tree/master/magenta/models/nsynth

[2] https://github.com/ibab/tensorflow-wavenet

[3] https://github.com/r9y9/wavenet_vocoder

[4] https://github.com/EllenRushe/AudioAnomalyDetectionWaveNet

7. REFERENCES

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):823–839, 2012.

[3] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[4] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. *Proceedings of the 34th International Conference on Machine Learning*, 70:1068–1077, 2017.

[5] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*, 1993.

[6] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 881–889, 2015.

[7] Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney. Convolutional neural networks for acoustic modeling of raw time signal in lvcsr. In *Interspeech*, pages 26–30, 2015.

[8] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson. Speech acoustic modeling from raw multichannel waveforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4624–4628, 2015.

[10] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[11] Shehroz S Khan and Babak Taati. Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders. *Expert Systems with Applications*, 87:280–290, 2017.

[12] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011.

[13] Patrick Lindstrom, Sarah Jane Delany, and Brian Mac Namee. Handling concept drift in a text data stream constrained by high labelling cost. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 2010.

[14] Weining Lu, Yu Cheng, Cao Xiao, Shiyu Chang, Shuai Huang, Bin Liang, and Thomas Huang. Unsupervised sequential outlier detection with deep architectures. *IEEE Transactions on Image Processing*, 26(9):4321–4330, 2017.

[15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[16] Erik Marchi, Fabio Vesperini, Stefano Squartini, and Björn Schuller. Deep recurrent neural network-based autoencoders for acoustic novelty detection. *Computational Intelligence and Neuroscience*, 2017.

[17] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[18] Dong Yul Oh and Il Dong Yun. Residual error based anomaly detection using auto-encoder in smd machine sound. *Sensors (Basel, Switzerland)*, 18(5), 2018.

[19] Timothy J O'Shea, T Charles Clancy, and Robert W McGwier. Recurrent neural radio anomaly detection. *arXiv preprint arXiv:1611.00301*, 2016.

[20] Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A Hasegawa-Johnson, and Thomas S Huang. Fast wavenet generation algorithm. *arXiv preprint arXiv:1611.09482*, 2016.

[21] Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *Interspeech*, 2013.

[22] Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. Acoustic novelty detection with adversarial autoencoders. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3324–3330, 2017.

[23] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney. Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *Interspeech*, 2014.

[24] Aäron 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.

[25] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756, 2016.

[26] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and koray kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[27] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926, 2018.