A control of the false anomaly rate for Anomaly Detections problems on Images

Samy Vilhes

November 2024

Abstract

Most anomaly detection methods for images score anomalies based on projection error or reconstruction loss. These methods typically rely on setting a threshold, determined using a validation set. If the score of a test observation exceeds this threshold, it is classified as an anomaly. However, this approach lacks control over the rate of false anomalies. In some applications, it is crucial to manage this rate, ensuring that no more than $\alpha\%$ of test observations are incorrectly flagged as anomalies: we do not want to reject too many test observations by error. The code is available at: https://github.com/vilhess/AnoControl

1 Introduction to Anomaly Detections

Anomaly detection algorithms identify test observations that deviate from a conforming distribution. Consider the test samples $X_{\text{test}} = \{x_{2n+1}, x_{2n+2}, \dots, x_{2n+t}\}$ and the conforming distribution P_X representing our expected data. The goal is to identify which test samples do not belong to the conforming distribution: $\{i \mid x_{2n+i} \not\sim P_X\}$. Most methods are based on unsupervised learning, as anomalous samples are typically rare during training. A model s is trained on a training set without anomalies $X_{\text{train}} = \{x_1, x_2, \dots, x_n\}$. For a given test sample x_{2n+i}, a model s computes a score $s(x_{2n+i})$, representing the likelihood or degree to which the sample belongs to the conforming distribution. The higher the score, the more conformal the sample is according to the model. In contrast, the lower the score, the more likely the sample is to be an anomaly according to the model. In many applications, a threshold is determined based on a validation set $X_{\text{val}} = \{x_{n+1}, x_{n+2}, \dots, x_{2n}\}$. Test samples with scores lower than this threshold are classified as anomalies, as their likelihood of belonging to the conforming distribution is insufficient. This threshold ensures a balance between detecting anomalies and avoiding false positives, depending on the application's requirements.

1.1 Anomaly detections for images

Anomaly detection in images is a powerful tool for identifying irregularities, such as adversarial attacks or, more specifically, abnormal regions within an image, such as tumor areas in a brain scan. In the first case, many algorithms focus on reconstructing the

input image to identify anomalies. Models such as GANs[2] (Generative Adversarial Networks) and VAEs[4] (Variational Autoencoders) are commonly used. The conformity score is typically derived as the inverse of the reconstruction loss: a lower loss implies a higher conformity. Alternatively, methods such as One-Class SVM[8] and Deep One-Class Classification[5] directly compute a conformity score without relying on reconstruction. There is another category of methods. In an unsupervised setting, where we cannot rely on a classification loss, DROCC[3] attempts to adapt the problem to a classification framework. During training, only normal data are used, while abnormal data is artificially generated to challenge the classifier. This approach can be viewed as a form of adversarial training: the model's objective is to classify normal data correctly, while the training process actively seeks to generate abnormal data that can mislead the classifier. In the case of detecting abnormal regions, reconstruction-based algorithms, such as VAEs and GANs are employed. However, these methods focus on evaluating the reconstruction loss at a pixel level to identify regions of anomalies. Each pixel's reconstruction error is calculated, and if the error exceeds a predefined threshold, that pixel is classified as anomalous. This approach enables the precise localization of abnormal regions within the image. It is important to note that this approach differs from traditional image segmentation methods, such as those using U-Net, as it involves training exclusively on normal images (e.g., images without tumors). This makes the process inherently unsupervised, as it does not rely on labeled data for abnormal regions. Instead, the model learns to identify deviations from normal patterns during evaluation, enabling the detection of anomalies without explicit segmentation labels. In this paper, we focus exclusively on addressing the first problem: detecting abnormal images as a whole.

2 The False Anomaly Rate

The false anomaly rate is the proportion of normal test samples incorrectly classified as anomalous by the model. Controlling this metric is essential to avoid excessive false alarms. For instance, in autonomous vehicles, detecting image adversarial attacks is essential. These attacks involve subtle manipulations to the entire image, such as changes in lighting or added noise, which can deceive the system into making incorrect decisions, potentially compromising safety. However, it is equally important to control the rate at which normal images are mistakenly flagged as abnormal, as excessive false positives can disrupt operations and reduce system efficiency. Therefore, balancing sensitivity with a low false anomaly rate is crucial for maintaining both efficiency and reliability.

2.1 Control of the False Anomaly Rate for Anomaly Detections

Let's define our null hypothesis as follows:

$$\mathcal{H}_{0,i} : X_{2n+i} \sim P_X$$

where P_X is the distribution of normal data. The associated p-value is calculated as:

$$p_i = P_X(x \leq X_{2n+i}),$$

representing the probability of observing a value as extreme or more extreme than X_{2n+i} under the null hypothesis.
To identify anomalies, we compare p_i to a predefined threshold α. If $p_i \leq \alpha$, we reject $\mathcal{H}_{0,i}$ and classify X_{2n+i} as anomalous. For instance, if $\alpha = 5\%$, we consider X_{2n+i} anomalous if the probability of observing a more extreme value under P_X is less than 5%. This approach ensures that only values significantly deviating from the normal distribution are flagged as anomalies. The problem is that, most of the time, we do not know how to compute such p-values, because the distribution of our normal samples is unknown or cannot be explicitly computed. To address this issue, we will leverage conformal inference to transform the outputs of our one-class classifier into p-values, enabling us to test the null hypothesis effectively.
Consider that we have trained our scoring method $\hat{s}$ on a training dataset $X_{\text{train}} = \{x_1, x_2, \ldots, x_n\}$. Let $X_{\text{cal}} = \{x_{n+1}, x_{n+2}, \ldots, x_{2n}\}$ denote our calibration set. For a test sample X_{2n+i}, we compute the marginal p-value[1] as follows:

$$\hat{p}_i = \frac{1 + |\{X_{n+j} \in X_{\text{cal}} : \hat{s}(X_{n+j}) \leq \hat{s}(X_{2n+i})\}|}{n+1},$$

where the numerator adds one to account for the test sample itself, ensuring robustness in finite sample settings.

3 Experiments

We will experiment with various one-class classifiers on problems involving images. To ensure a fair comparison between these methods, we will construct marginal test p-values. This approach eliminates the need to set arbitrary thresholds, which can vary across models, and provides a consistent evaluation framework. We will implement one-class methods on the MNIST dataset, which consists of grayscale images of hand-written digits.

3.1 Problems:

We detail the two settings we consider:

- **1) One VS All**: One digit is treated as the normal class, while images of other digits are considered anomalies. This setup is commonly addressed in numerous papers on anomaly detection for images.
- **2) All VS One**: One digit is treated as the anomalous class, while images of other digits are considered normal. This is the setting we propose. Unlike learning the distribution of a single digit, the model must learn the distribution of multiple digits, making the task more complex.

3.2 Models

For problems **1** and **2**, the same algorithms can be applied. For instance, Deep One-Class Classification[5] attempts to map the normal data distribution into a hypersphere

of lower dimensionality using neural networks, effectively isolating anomalies as points lying outside this hypersphere. They also propose mapping normal data as close as possible to a single point in a latent space. The anomaly score is then defined as the distance of the data point's mapping to this center. To classify anomalies, a threshold must be set: any test observation with a distance exceeding this threshold is considered an anomaly. In the following, when referring to the Deep One-Class Classifier, we will focus on this specific last method. DROCC[3] can also be applied. During training, a classifier is trained to correctly identify normal data as normal. Simultaneously, gradient ascent is performed on generated samples to create adversarial examples that aim to mislead the classifier. The overall loss combines two objectives: correctly classifying normal data as normal and generated abnormal data as abnormal. These models provides a conformity score for the entire image. Generative models also can be employed. The approach for using a VAE as an anomaly detector is as follows: the VAE is trained exclusively on normal samples. For a test sample, the model generates a reconstructed version of the input. If the reconstruction loss for the entire image is significantly high, it indicates that the model has not encountered this type of image during training and is unable to reconstruct it effectively. Such a sample can then be considered anomalous. For GANs, the methodology is similar but with some differences. In a basic GAN, we lack direct control over the generated samples and we still rely on evaluating the reconstruction loss. Using models like AnoGAN [7], we first train a basic GAN on normal samples. Then, for a test sample, gradient descent is performed in the latent space to identify the latent vector that best generates the given test sample. This process aims to minimize two key differences: the difference between the test sample and its generated counterpart, and the difference between the outputs of an intermediate layer of the discriminator for the generated image and the test image. This dual objective helps ensure a more accurate representation of the test sample in the latent space. The challenge with this approach is that for N test samples, N separate optimizations are required to find the best latent vector for each sample. As N grows large, this process can become computationally prohibitive due to the extensive optimization workload. To address this limitation, the authors of AnoGAN introduced F-AnoGAN [6], a more efficient version of AnoGAN. In F-AnoGAN, an additional neural network is trained after the basic GAN is trained. This network, which acts as an encoder, maps an input image directly to its latent vector. The encoder is designed to minimize two objectives: the difference between the input sample and its generated counterpart, and the difference between the outputs of an intermediate layer of the discriminator for the input and generated images. This significantly reduces the computational cost compared to performing gradient descent for each test sample.

3.3 Results

Initially, we will assess the models' performance using the ROC-AUC score to evaluate their accuracy and effectiveness. The ROC-AUC score ranges from 0 to 1, taking into account both the False Positive Rate and the True Positive Rate. A higher ROC-AUC score implies better model performance, indicating a greater ability to correctly distinguish between positive and negative instances. Following this, to conduct a more detailed analysis of model rejections, we will use p-values with a significance threshold

of 5%. Any image with a p-value below this threshold will be classified as an anomaly. We will present tables summarizing results for selected digits classified as either normal or anomalous and analyze the overall rejection rate, including both anomalies and normal samples.
We will present results specifically for digits 0 and 7, considered as normal in the first setting and anomalous in the second setting. Complete results are available on our GitHub repository, where we have developed a web interface using Streamlit to facilitate result visualization.

3.4 One vs All framework

3.4.1 ROC-AUC scores:

Normal digit	VAE	CVAE	DeepSVDD	DROCC	f-ANOGAN
0	0.995	0.995	0.989	0.944	0.967
1	0.999	1	0.995	0.927	0.999
2	0.927	0.930	0.913	0.772	0.834
3	0.951	0.943	0.866	0.558	0.880
4	0.961	0.946	0.931	0.823	0.865
5	0.967	0.954	0.879	0.568	0.856
6	0.989	0.994	0.975	0.849	0.933
7	0.969	0.962	0.949	0.911	0.897
8	0.897	0.891	0.917	0.881	0.861
9	0.970	0.974	0.958	0.814	0.920

In this initial framework, the ROC-AUC score indicates that the best-performing models are the VAEs using the reconstruction loss. Their results are quite impressive. The Deep SVDD model comes after, then DROCC and f-ANOGAN that performed the worst. Limited time was spent on fine-tuning the models, so the results may vary with further optimization. We ensured that the training time for each model was kept comparable to maintain fairness, which could explain why f-ANOGAN underperformed. This model is particularly resource-intensive, as it requires the training of three neural networks.

3.4.2 Study of the rejection rates with a threshold of 5%

Rejection Rate for Normal digit 0

Digit	Normal	Deep SVDD	VAE	CVAE	f-ANOGAN	DROCC
0	Yes	4.6%	4.7%	4.6%	3.6%	4.5%
1	No	99.8%	99.3%	97.2%	100.0%	98.8%
2	No	98.0%	99.4%	100.0%	91.3%	77.6%
3	No	97.0%	99.4%	100.0%	84.6%	46.3%
4	No	99.6%	99.7%	99.8%	98.8%	66.2%
5	No	95.4%	98.5%	99.6%	82.7%	62.0%
6	No	86.0%	93.2%	96.8%	69.9%	60.1%
7	No	100.0%	99.0%	97.5%	98.8%	94.8%
8	No	82.4%	98.7%	99.2%	85.7%	25.6%
9	No	98.7%	99.3%	98.5%	91.2%	59.3%

Our strategy effectively controls the False Anomaly Rate at a target of 5% while successfully rejecting other digits.

Rejection Rate for Normal digit 7

Digit	Normal	Deep SVDD	VAE	CVAE	f-ANOGAN	DROCC
0	No	86.9%	99.9%	100.0%	98.4%	94.7%
1	No	84.5%	18.4%	18.0%	88.7%	86.6%
2	No	93.8%	99.3%	99.1%	94.1%	95.4%
3	No	87.1%	99.2%	99.7%	95.4%	78.7%
4	No	79.4%	87.6%	96.6%	66.5%	52.0%
5	No	93.3%	99.8%	99.9%	96.4%	77.7%
6	No	99.5%	100.0%	100.0%	99.9%	99.2%
7	Yes	5.7%	5.7%	5.6%	6.2%	6.2%
8	No	61.8%	99.1%	99.8%	86.3%	49.0%
9	No	16.7%	55.1%	75.4%	39.3%	9.5%

The results are similar to the previous ones. However, the False Anomaly Rate slightly exceeds our threshold. This suggests the need for a larger validation set to capture a more comprehensive range of normal scores.

3.5 All vs One framework

3.5.1 ROC-AUC scores:

Anormal digit	VAE	CVAE	DeepSVDD	f-ANOGAN	DROCC
0	0.955	0.977	0.678	0.823	0.824
1	0.386	0.209	0.909	0.359	0.908
2	0.962	0.967	0.702	0.966	0.669
3	0.916	0.968	0.579	0.870	0.740
4	0.911	0.940	0.672	0.858	0.425
5	0.945	0.969	0.565	0.899	0.567
6	0.946	0.945	0.570	0.914	0.662
7	0.664	0.791	0.663	0.676	0.531
8	0.938	0.966	0.419	0.907	0.431
9	0.729	0.742	0.460	0.597	0.333

In this second framework, the ROC-AUC scores are less impressive compared to the previous framework. This is expected, as the problem is more complex due to the inclusion of multiple digits as normal classes during training. Nevertheless, the best-performing models remain the VAEs based on reconstruction loss. These models outperform all others across most digits, except for one specific digit, the 1, where they struggle significantly. For this particular digit, Deep SVDD and DROCC perform relatively well but show weaker performance on the other digits.

3.5.2 Study of the rejection rates with a threshold of 5%

Rejection Rate for Anormal digit 0

Digit	Anormal	Deep SVDD	VAE	CVAE	f-ANOGAN	DROCC
0	Yes	39.8%	64.2%	86.5%	23.0%	36.0%
1	No	66.7%	0.4%	0.5%	0.3%	0.0%
2	No	7.6%	7.0%	6.5%	9.3%	11.6%
3	No	5.9%	5.6%	6.6%	5.1%	6.8%
4	No	10.6%	1.4%	2.9%	2.7%	3.7%
5	No	13.5%	6.8%	7.3%	6.4%	5.7%
6	No	9.2%	4.2%	2.5%	4.6%	8.8%
7	No	13.6%	1.2%	3.1%	2.0%	0.2%
8	No	3.0%	11.6%	11.8%	9.4%	3.8%
9	No	3.3%	2.2%	1.9%	1.0%	0.8%

The best-performing models remain the VAEs, particularly the Convolutional VAE. It effectively rejects anomalous digits while correctly preserving normal ones. Although the threshold is not strictly adhered to, the rate is close to the target. Other models perform less effectively compared to the VAEs.

Rejection Rate for Anormal digit 7

Digit	Normal	Deep SVDD	VAE	CVAE	f-ANOGAN	DROCC
0	No	25.0%	2.2%	2.4%	2.4%	16.6%
1	No	72.6%	0.2%	0.2%	0.2%	0.1%
2	No	15.1%	7.9%	7.8%	7.9%	4.3%
3	No	14.3%	4.9%	6.5%	3.9%	9.7%
4	No	20.4%	2.2%	3.2%	3.0%	0.6%
5	No	17.2%	6.7%	7.7%	5.6%	2.7%
6	No	13.4%	3.3%	1.8%	3.8%	0.6%
7	Yes	40.3%	12.1%	27.8%	8.6%	3.9%
8	No	6.5%	10.6%	11.8%	7.8%	0.7%
9	No	11.8%	1.9%	2.1%	1.5%	0.3%

As expected, the VAEs once again emerge as the best-performing models. They effectively preserve normal digits while accurately detecting the anomalous digit. In this case, where the abnormal class is digit 7, the other models also perform well, showing better results compared to the scenario where the abnormal class was digit 0.

4 Conclusion

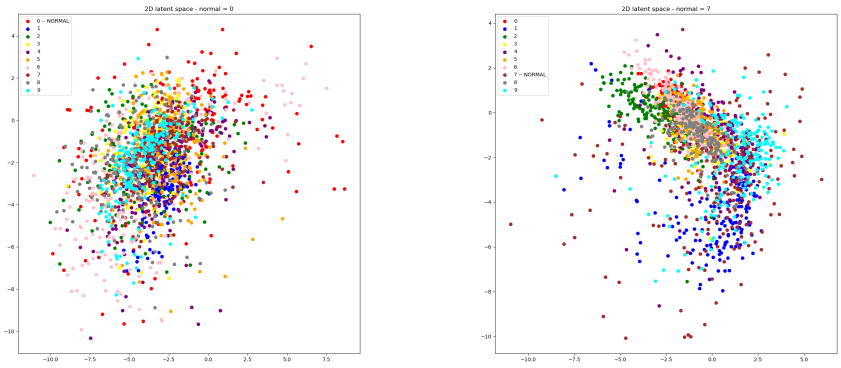
While models strongly perform in One-vs-All Setting, they struggle a bit more with the All-vs-One Setting. VAEs reconstruction-based perform better than other models across all settings. The introduction of test p-values allows us to standardize the thresholding process, eliminating the need to define separate thresholds for each model. This approach enables the use of a unified threshold across all models while controlling the false anomaly rate.

5 Annexe: study of latent spaces

Our models considered latent spaces, such as those provided by Variational Autoencoders (VAEs), and projection spaces, exemplified by Deep SVDD. For VAEs, we aim to observe that the anomalous class (or classes, depending on the framework) resides significantly distant from the normal class (or classes) in the latent space. In the case of Deep SVDD, our objective is for the center of the projection space to align closely with the true normal class (or classes).

5.1 VAE

For the VAEs, we utilized a 2-dimensional latent space to facilitate easy visualization. We only consider Linear VAEs for this experiment. In both frameworks (All vs. One and One vs. All), we conducted 10 experiments for each class, considering it alternately as normal and anomalous. For simplicity and brevity, we will present two results that are representative of the majority.

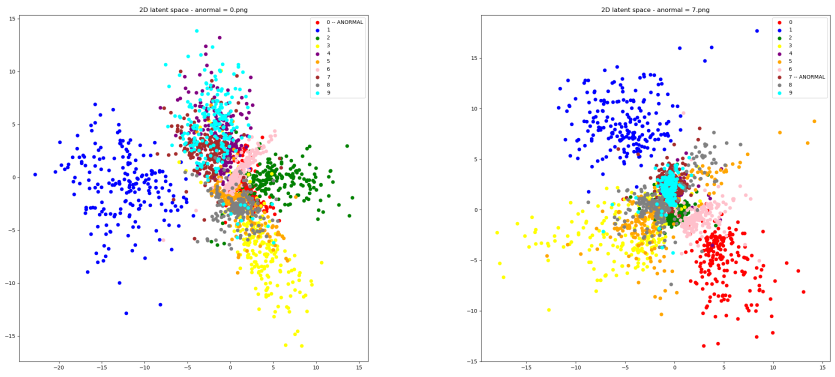


(a) Normal class 0 (b) Normal class 7

Figure 1: Visualization for the One vs All framework.

For this given problem, our VAE works better than the F-ANOGAN model. H

In the One vs All framework, results shown in Fig.1 reveal that all digits, including the anomalies, are mapped close to the origin, with little distinction between them. This is not the desired outcome. The goal is to have one cluster for the normal digit and another separate cluster for all other digits. This behavior can be attributed to the assumptions made about the latent space during the training of the VAE, where we assume the latent space follows a Gaussian distribution. As a result, the encoder is trained to map all data points to this distribution. However, for our problem, it is the decoder that plays a crucial role in this process. The decoder is specifically trained to reconstruct the normal digit, and thus, for any position in the latent space, it will generate a reconstruction of the normal digit. It does not have the capability to accurately reconstruct anomalous digits.



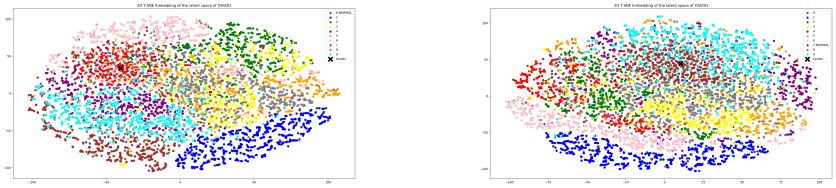
(a) Normal class 0 (b) Normal class 7

Figure 2: Visualization for the All vs One framework.

For the second framework, All vs One, some results are shown in Fig.2. In this case, we observe distinct clusters of normal digits in the latent space, generally distributed around the origin. This behavior aligns with the hypotheses underlying the training of the VAE: the entire latent space is expected to follow a Gaussian distribution, and each normal digit corresponds to its own Gaussian distribution with specific parameters for mean and variance. This explains the formation of distinct clusters for the same digits. Conversely, anomalous digits are consistently mapped near the center of the latent space. Since the decoder is not trained to reconstruct these anomalous digits, it instead generates reconstructions of normal digits. This observation reinforces the role of the decoder as the principal component for anomaly detection using VAEs as we know that these models work well for the problems we considered.

5.2 Deep One-Class Classifier

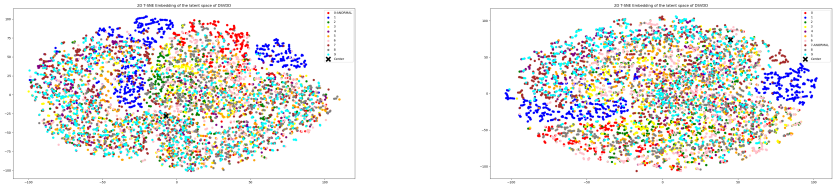
In this model, we project our images into a 32-dimensional space, which makes it challenging to compare the sample projections to the center directly. To address this, we employ T-SNE to map the projections into a 2-dimensional latent space, enabling visualizations similar to those used for the VAEs.



(a) Normal class 0 (b) Normal class 7

Figure 3: Visualization for the One vs All framework.

The results in Fig.3 for the One vs All show that both normal and anomalous digits are well separated in the T-SNE space. Furthermore, we observe that the center lies within and approximately at the midpoint of each normal digit cluster. These results are satisfactory overall. However, it would be more desirable if the anomalies were not grouped into distinct clusters, as this could further enhance the separation between normal and anomalous digits.



(a) Normal class 0 (b) Normal class 7

Figure 4: Visualization for the All vs One framework.

The results in Fig.4 for the All vs One framework show that each digit resides in the same overall cluster within the latent space. The center is not particularly far from the anomalous class, which may explain why the Deep SVDD performs poorly in this framework. All digits are mapped to overlapping regions in the latent space, with the exception of digit ”one,” which forms a distinct sub-cluster within the larger cluster.

References

[1] Stephen Bates et al. “Testing for outliers with conformal p-values”. In: *The Annals of Statistics* 51.1 (Feb. 2023). ISSN: 0090-5364. DOI: 10.1214/22-aos2244. URL: http://dx.doi.org/10.1214/22-AOS2244.

[2] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[3] Sachin Goyal et al. “DROCC: Deep Robust One-Class Classification”. In: *CoRR* abs/2002.12718 (2020). arXiv: 2002.12718. URL: https://arxiv.org/abs/2002.12718.

[4] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: https://arxiv.org/abs/1312.6114.

[5] Lukas Ruff et al. “Deep One-Class Classification”. en. In: *Proceedings of the 35th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2018, pp. 4393–4402. URL: https://proceedings.mlr.press/v80/ruff18a.html (visited on 11/12/2024).

[6] Thomas Schlegl et al. “f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks”. In: *Medical Image Analysis* 54 (Jan. 2019). DOI: 10.1016/j.media.2019.01.010.

[7] Thomas Schlegl et al. *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*. arXiv:1703.05921. Mar. 2017. URL: http://arxiv.org/abs/1703.05921 (visited on 11/18/2024).

[8] Bernhard Schölkopf et al. “Estimating Support of a High-Dimensional Distribution”. In: *Neural Computation* 13 (July 2001), pp. 1443–1471. DOI: 10.1162/089976601750264965.