Evaluating the Use of Reconstruction Error for Novelty Localization

Patrick Feeney [1]   Michael C. Hughes [1]

Abstract

The pixelwise reconstruction error of deep autoencoders is often utilized for image novelty detection and localization under the assumption that pixels with high error indicate which parts of the input image are unfamiliar and therefore likely to be novel. This assumed correlation between pixels with high reconstruction error and novel regions of input images has not been verified and may limit the accuracy of these methods. In this paper we utilize saliency maps to evaluate whether this correlation exists. Saliency maps reveal directly how much a change in each input pixel would affect reconstruction loss, while each pixel's reconstruction error may be attributed to many input pixels when layers are fully connected. We compare saliency maps to reconstruction error maps via qualitative visualizations as well as quantitative correspondence between the top K elements of the maps for both novel and normal images. Our results indicate that reconstruction error maps do not closely correlate with the importance of pixels in the input images, making them insufficient for novelty localization.

1 Introduction

One-class image novelty detection is a classification task where a model must determine whether an image is normal, belonging to the same distribution as the training data, or is novel, belonging to a different distribution (Japkowicz et al., 1995; Abati et al., 2019; Perera et al., 2019). Unlike typical classification tasks, one-class novelty detection requires training only on data from a single class, often under the assumption that novel data could not be collected in an applied setting. This task requires distinct methods from standard binary classification, because models must be trained without examples of novel images. Geng et al. (2020) contrasts one-class novelty detection with similar tasks such as open

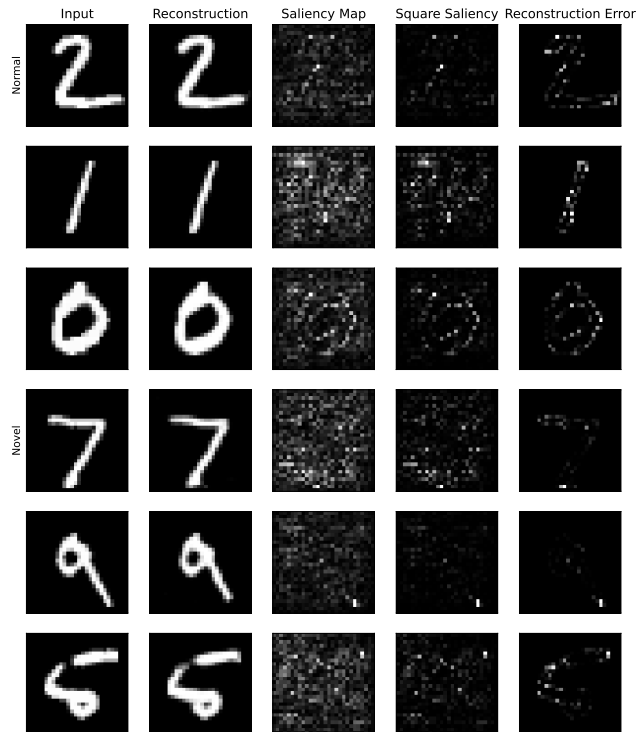[1]Department of Computer Science, Tufts University, Medford, Massachusetts, USA. Correspondence to: Patrick Feeney <patrick.feeney@tufts.edu>.

Figure 1. Comparison of normalized maps produced by a deep autoencoder given 3 examples of "normal" input (*top 3 rows*, digits 0-4) and 3 examples of "novel" input (*bottom*, digits 5-9). From left to right: model input, input reconstructions, saliency map for reconstruction error, squared saliency map, and reconstruction error map. Saliency maps represent input pixels with a high impact on the reconstruction error. While reconstruction error maps share *some* high intensity pixels with saliency maps, reconstruction error maps often fail to identify important input pixels highlighted in the saliency map that are *far from the digit boundary*. This deficiency is more pronounced on novel inputs, thus calling into question the suitability of reconstruction error maps for novelty localization.

set recognition (Oza & Patel, 2019) and out-of-distribution detection (Ren et al., 2019; Liang et al., 2017), noting that this task focuses solely on outlier detection instead of also classifying normal images.

Existing works often utilize deep autoencoders for image novelty detection. These models utilize the reconstruction error, having found that novel images will typically have a larger reconstruction error than normal images (Japkowicz et al., 1995; Richter & Roy, 2017). Various models utilize

this finding in different ways. Constructing an empirical cumulative distribution function for the reconstruction error on the training data and classifying images based on the reconstruction error percentile has been used for novelty detection in visual navigation tasks (Richter & Roy, 2017). Some methods have increased the reconstruction error percentile of novel images by using generative adversarial networks or recurrent neural networks (Perera et al., 2019; Sabokrou et al., 2018; Luo et al., 2017). The reconstruction error signal has also been combined with signals from a latent space representation learned via variational autoencoders (Abati et al., 2019) or a generative adversarial network (Perera et al., 2019).

Once a novelty has been detected, some models attempt to *localize* the novelty. The task of *novelty localization* entails determining which pixels in the input image lead to the model labeling it as novel. Localization can be useful in practice for identifying an abnormal region in a medical image or a defective part of a manufactured object (Zimmerer et al., 2018; Bergmann et al., 2020). When utilizing a deep autoencoder, novelty localization is typically done using the *reconstruction error map*, an image containing the per pixel reconstruction error produced by the autoencoder (Abati et al., 2019; Sabokrou et al., 2018; Gong et al., 2019; Bergmann et al., 2020; Zimmerer et al., 2018). Using this map assumes that pixels with high reconstruction errors correspond to novel pixels in the input image. Surprisingly, this assumption about the correlation between reconstruction error and input images has not been verified. Relying on this assumption may limit the accuracy of these methods.

We note the similarity between the novelty localization and classification model explanation problems. Both problems require the identification of input pixels with the largest influence on some image-level output function, novelty detection and classification respectively. In this paper we utilize saliency maps to evaluate whether reconstruction error maps correlate with the importance of pixels in the input images for deep autoencoder architectures. To assess this correlation between saliency maps and reconstruction error maps, we use both qualitative visualizations (see Fig. 1) and quantitative metrics. We introduce two metrics to better summarize the correspondence between two maps. Thorough experiments on MNIST digit data show the utility of our saliency map approach and metrics. Our results indicate that reconstruction error maps do not closely correlate with the importance of pixels in the input images, making them insufficient for novelty localization.

2 Related Work

2.1 Deep Autoencoders for Novelty Detection

The simplest deep autoencoder approaches to novelty detection consist of two parts: a deep autoencoder (Bourlard & Kamp, 1988; Vincent et al., 2008) that produces a reconstruction map, and a method that consumes the reconstruction map and produces a decision about novelty. In the case of single image inputs, a threshold on the reconstruction error is often used to determine novelty (Japkowicz et al., 1995; Richter & Roy, 2017). This threshold may be set using any method for hyperparameter selection, but thresholding at a high percentile of the empirical cumulative distribution function for the reconstruction error on the training data provides a low-cost solution (Richter & Roy, 2017).

Recent advances tend to focus on introducing new methods for one part of the larger model. Autoencoder focused methods have created models that increase the gap between the reconstruction error of normal and novel images by using generative adversarial networks (GANs) during training (Perera et al., 2019; Sabokrou et al., 2018). Adversarial training forces the network to decode more latent vectors into plausible, normal images. Other efforts use sparse coding representations within neural networks to similar effect (Luo et al., 2017). Furthermore, latent representations can be constrained to a subset of the embeddings of training inputs (Gong et al., 2019).

Approaches focused on improving the novelty detection part introduce ways to utilize the latent space for detection. The probability of an input image being mapped to a latent vector can be probabilistically modeled via a Gaussian mixture model (Zong et al., 2018), a Gaussian classifier (Sabokrou et al., 2018), a variational autoencoder (Zimmerer et al., 2018), or an autoregressive process (Abati et al., 2019). This probability can be utilized independently or combined with reconstruction error to form a novelty score.

2.2 Saliency Maps

Saliency maps provide a method for visualizing the representations learned by deep neural networks. The resulting map reveals which pixels will most impact the loss function when changed. Various methods for image-specific saliency maps have been proposed (Springenberg et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017). In this paper we utilize Simonyan et al. (2014)'s gradient based method. Recent independent evaluations (Adebayo et al., 2020; Arun et al., 2020) suggest this method is better than alternatives at exposing differences between a properly trained neural network and one with some randomized internal weights.

2.3 Similarity Metrics

Previous works have used several metrics for quantitative comparison between saliency maps, such as Spearman rank correlation with and without absolute value, the structural similarity index, and the Pearson correlation of the histogram of gradients (Adebayo et al., 2020; Arun et al., 2020). These metrics were found to be ill-suited to the compari-

son between saliency maps and reconstruction error maps. Although intense values are most important in both maps, directly comparing values across maps is impractical.

3 Method

In our formal setup, an input is a vector $x \in \mathbb{R}^n$. A deep autoencoder is a neural network model with an encoder $f(x) : \mathbb{R}^n \to \mathbb{R}^m$ and decoder $g(z) : \mathbb{R}^m \to \mathbb{R}^n$.

The *reconstruction error map* $r(x) \in \mathbb{R}^n$ for image x is defined as

$$r(x) = (x - g(f(x)))^2, \tag{1}$$

producing a per pixel squared error map. The error at pixel index i is denoted $r_i(x)$ The reconstruction loss $\bar{r}(x)$ is

$$\bar{r}(x) = \frac{1}{n} \sum_{i=1}^{n} r_i(x), \tag{2}$$

the mean of $r(x)$ across pixel values.

Following Simonyan et al. (2014), a *saliency map* $s(x) \in \mathbb{R}^n$ is defined as

$$s(x) = \left| \frac{\delta \bar{r}}{\delta x} \right|, \tag{3}$$

which is found by backpropagating $\bar{r}(x)$. In the case of multichannel images, the maximum of the channel values is taken for each pixel (Simonyan et al., 2014).

To evaluate whether reconstruction error maps correlate with important pixels in the input image, we compare $r(x)$ and $s(x)$. A saliency map $s(x)$ reveals which input pixels most impact $\bar{r}(x)$ when their values are changed. This differs from the $g(f(x))$ term of $r(x)$, where each output pixel no longer has a clear correspondence to an input pixel due to the transformations applied. Via convolutions and compression to an m-dimensional encoding vector, it is likely each input pixel has some impact on each pixel of $r(x)$.

If high values in $r(x)$ do indicate which parts of the input image are novel, a correspondence should exist between high values in $r(x)$ and high values in $s(x)$, where changes to the input x have the greatest effect.

3.1 Evaluation Metrics

When comparing the reconstruction map $r(x)$ to the saliency map $s(x)$, it is important to note the maps are nonnegative but have different magnitudes. To facilitate comparison, both are scaled to $[0, 1]$ on a per image basis by dividing by the maximum pixel value for that map.

Visualizations of the maps are used to subjectively evaluate similarities. The saliency maps in the third column of Figure 1 are significantly more noisy than the reconstruction error maps in the last column. This is in part due to values in the saliency map being linearly related while the values in the reconstruction error map are quadratic. Therefore the square of the scaled map $s(x)$ is also presented for comparison between quadratic values.

To assess quantitative performance, we develop two metrics that measure *correspondence* between the top K elements of two maps, as detailed below. These metrics require as input the top K largest elements (pixels) of each map, which are stored as a set.

Metric 1: Top K agreement. We compute the *top K agreement* between two maps as the size of the intersection of their selected sets. Top K agreement has similar goals as the Spearman rank discussed in Sec. 2.3 but provides clear two advantages for this problem. First, the method only considers the top K elements instead of the entire image, mitigating the impact of points with little importance. Second, this method depends only on the ordering of values and does not compare values across maps. Therefore this method produces the same output under any transformation of its inputs that maintains the same sorted order of pixel values. Thus transformations like scaling and squaring are not necessary to make saliency maps and reconstruction error maps comparable.

Metric 2: Maximum distance to best match. Given the top K pixel sets of two maps, we compute the *maximum distance to best match* by identifying for each member of the reference set the closest matching pixel (by Euclidean distance) in the other set. As the reference set, we select the *saliency map*, as it is a more reliable indicator of which pixels should be identified as novel. We summarize all K matches via the maximum of such distances. This metric is motivated by a key limitation of top K agreement: it requires an exact pixel-level match and thus a shift of even one pixel reduces agreement substantially. Our distance metric provides a more gradual assessment developed for the intended use case of novelty localization, where a predicted map with small distances to the reference (ground truth) likely has milder consequences than one with large distances.

3.2 Experiment Setup

Experiments are conducted on the MNIST (LeCun & Cortes) dataset. The training set is split with 54,000 images for the training set and 6,000 images for a validation set. Digits 0, 1, 2, 3, and 4 are used as the normal classes while digits 5, 6, 7, 8, and 9 are used as novel classes. We use the convolutional autoencoder architecture for MNIST from Abati et al. (Abati et al., 2019), omitting their additional latent space model. The model is trained on the normal classes from the training set. The normal classes for the validation set were used to tune learning rate. Results were produced from the test set.

4 Results

We summarize our major findings in the paragraphs below.

Finding 1: Reconstruction error maps often miss key salient input pixels. Visualization of the maps produced for randomly selected normal and novel inputs in Fig. 1

reveals significant differences between saliency maps and reconstruction error maps. Saliency maps have more pixels with large values, revealing many pixels with high impact on the loss that are not found by the reconstruction error map. While using *squared* saliency maps reduces the number of pixels with middling values, the reconstruction error still only captures a subset of the intense pixels. This suggests that reconstruction quality is sensitive to a larger number of input pixels than the reconstruction error map indicates.

Finding 2: The correspondence between reconstruction error and saliency maps is weaker for novel images. The reconstruction error maps in Fig. 1 often resemble an outline of the digit, perhaps most clearly with the "0" in row 3. This outline can also often be seen in the saliency maps for normal data. In contrast, both the outlined digit and its correspondence in the saliency maps is less visible in all 3 novel examples in Fig. 1. This suggests that the strength of correspondence between reconstruction and saliency maps depends on whether data is normal or novel.

Finding 3: Top K agreement suggests reconstruction error does not closely correlate with saliency. Fig. 2 shows the frequency of top 10 agreement values across images in the test set. Total agreement is very rare; the most frequent level of agreement is 4 of 10 pixels for normal images and 2 of 10 pixels for novel images.

Finding 4: Novel images have noticeably lower top K agreement than normal images. Tab. 1 shows that only 26% of normal images have the majority of their top 5 pixels agree across the two maps. This drops to 19% for novel images. An even larger decrease is seen for top 10 agreement. Given the thousands of images in the test set, we expect this difference is significant.

Finding 5: Maximum distance metric reveals the best match for at least one salient pixel is poor. For top 5 agreement we find that across all images in the test set, the mean max distance to best match is 8.28 pixels for normal images and 9.25 pixels for novel images. When expanding to the top 10 agreement, these numbers become 8.14 pixels for normal images and 8.90 pixels for novel images. As MNIST images are 28 by 28 pixels, these distances indicate a clear separation between the top K sets of the two maps.

Finding 6: Use squared saliency maps when reconstruction maps use squared error. We find that both visually and quantitatively, squared saliency maps have better correspondence to reconstruction error maps than original (non-squared) saliency maps. When switching from saliency to squared saliency, the average mean-squared-error to reconstruction maps drops from 0.0303 to 0.0071 for normal images and from 0.0305 to 0.0072 for novel ones. When the reconstructions are based on squared error, it makes sense that a corresponding square is applied to the saliency map.

Top 5 Agreement	Normal Data	Novel Data
0-2	0.74	0.81
3-5	0.26	0.19
Top 10 Agreement	Normal Data	Novel Data
0-5	0.81	0.92
6-10	0.19	0.08

Table 1. The frequency of top 5 agreement and top 10 agreement values across normal and novel images in the MNIST test set. Frequencies are aggregated to show accuracy at identifying the majority of important input pixels. Both normal and novel images fail to exceed 26% accuracy, with novel images significantly worse at top 10 agreement with only 8% accuracy.
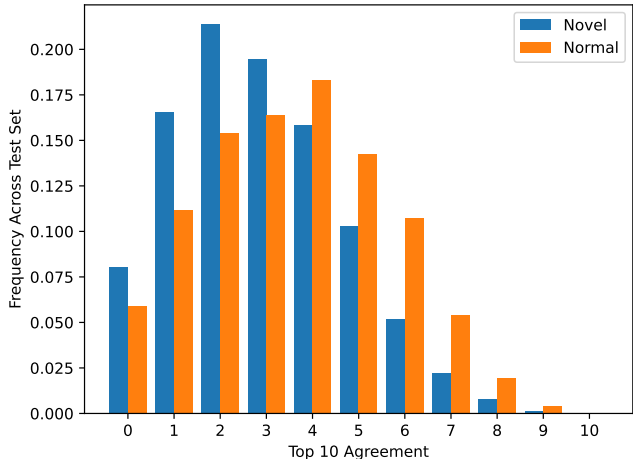

Figure 2. The frequency of possible top 10 agreement values (0 = no agreement, 10 = a perfect match), observed across all images in the MNIST test set. Overall low agreement suggests that the top values of reconstruction error maps do not closely correlate with those of saliency maps, especially for novel images.

5   Discussion & Conclusion

Our results find that reconstruction error maps do not closely correlate with the importance of pixels in the input images. Visually the maps are significantly different, although scaling the saliency map by squaring slightly improves this. The significant reduction of the mean squared error suggests squared saliency maps should be used over saliency maps when using techniques for image similarity sensitive to the scaling of pixel values.

The top K agreement clearly shows that reconstruction error maps are not identifying important pixels in the input, especially in the case of novel data. This suggests that while reconstruction error maps provide visually appealing novelty localization maps, they are not always representative of which parts of the input are novel and contribute most to reconstruction error. Further work and more experiments are needed to evaluate whether saliency maps themselves might provide a suitable replacement for novelty detection and localization.

References

Abati, D., Porrello, A., Calderara, S., and Cucchiara, R. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2020.

Arun, N. T., Gaw, N., Singh, P., Chang, K., Hoebel, K. V., Patel, J., Gidwani, M., and Kalpathy-Cramer, J. Assessing the validity of saliency maps for abnormality localization in medical imaging. *arXiv preprint arXiv:2006.00063*, 2020.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Bourlard, H. and Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.*, 59(4–5):291–294, 1988. URL https://doi.org/10.1007/BF00332918.

Geng, C., Huang, S.-j., and Chen, S. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Japkowicz, N., Myers, C., Gluck, M., et al. A novelty detection approach to classification. In *IJCAI*, volume 1, pp. 518–523. Citeseer, 1995.

LeCun, Y. and Cortes, C. Mnist handwritten digit database.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Luo, W., Liu, W., and Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 341–349, 2017.

Oza, P. and Patel, V. M. C2ae: Class conditioned autoencoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Perera, P., Nallapati, R., and Xiang, B. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

Richter, C. and Roy, N. Safe visual navigation via deep learning and novelty detection. *Robotics: Science and Systems XIII*, 2017.

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., and Klette, R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2018.02.006. URL https://www.sciencedirect.com/science/article/pii/S1077314218300249.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2015.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008. URL https://doi.org/10.1145/1390156.1390294.

Zimmerer, D., Kohl, S. A. A., Petersen, J., Isensee, F., and Maier-Hein, K. H. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.