A Generalized Approach to Determine Confident Samples for Deep Neural Networks on Unseen Data

Min Zhang[1(✉)], Kevin H. Leung[1,2], Zili Ma[1], Jin Wen[1], and Gopal Avinash[1]

[1] GE Healthcare, San Ramon, CA 94583, USA
chinazm@gmail.com
[2] The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Abstract. Deep neural network (DNN) models are widely applied in biomedical image studies since DNN models take advantage of massive data to provide improved performance over traditional machine learning models. However, like any other data-driven models, DNN models still face generalization limitations. For example, a model trained on clinical data from one hospital may not perform as well on data from another hospital. In this work, a novel approach is proposed to determine confident samples from unseen data on which a DNN model will have improved performance. Confident samples are defined as inliers identified by an outlier detector, which is based on projection of training data onto a standard feature space (e.g. ImageNet feature space). The hypothesis of the proposed method is that in a standard feature space, a DNN model will perform better on the inlier data samples and more poorly on the outliers. While projecting the unseen data to a standard feature space, if data points are detected as inliers, then the model will likely have consistent performance on those inliers as those patterns have already been "seen" from the training dataset. To validate our hypothesis, experiments were conducted using publicly available digit image datasets and chest X-ray images from three unseen datasets collected across U.S. and Canada hospitals. The experimental results showed consistently improved performance across various DNN models on all confident samples from unseen datasets.

Keywords: Deep neural network · Feature extraction · Outlier detection · Confident samples

1 Introduction

Deep neural network (DNN) methods have been successfully applied in healthcare domain in recent years. With significant advances in improving the performance of DNN models, it is also crucial to find or detect whether the model has produced a

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32689-0_7) contains supplementary material, which is available to authorized users.

H. Greenspan et al. (Eds.): CLIP 2019/UNSURE 2019, LNCS 11840, pp. 65–74, 2019.
https://doi.org/10.1007/978-3-030-32689-0_7

confident prediction, especially for unseen data. The lack of confidence in model predictions may result in poor clinical decisions, which is especially relevant in the medical diagnosis field [1].

In the literature, researchers have estimated the model confidence by quantifying the uncertainty of the model using Monte Carlo dropout [2]. This approach can be applied to image segmentation as in [3] where overall segmentation accuracy improved when pixels with high uncertainty were dropped. Bayesian dropout was also used to estimate prediction uncertainty in diagnosing diabetic retinopathy using fundus images [4]. Diagnostic performance was improved when estimates of uncertainty were used to filter out samples [4].

Other recent work on classification using a filtering strategy for DNNs has been done to filter out instances where the base model prediction is not confident [5]. This method used the outputs of model softmax layer to determine a threshold to optimize sensitivity while preserving precision at a given confidence level [5]. Hendrycks et al. [6] reported a method to measure the confidence score of unseen examples by finding the difference between output probabilities from the final softmax activation layer and true probabilities. In the study, they performed outlier detection between popular computer vision datasets, such as MNIST. However, the out-of-domain examples are visually very different from the in-domain examples in their study. In real world applications, image data from different data sources are often visually indistinguishable. As shown in Fig. 1, the hospital of origin cannot be determined by visual assessment of examples of chest X-ray images from different datasets used in the present study.

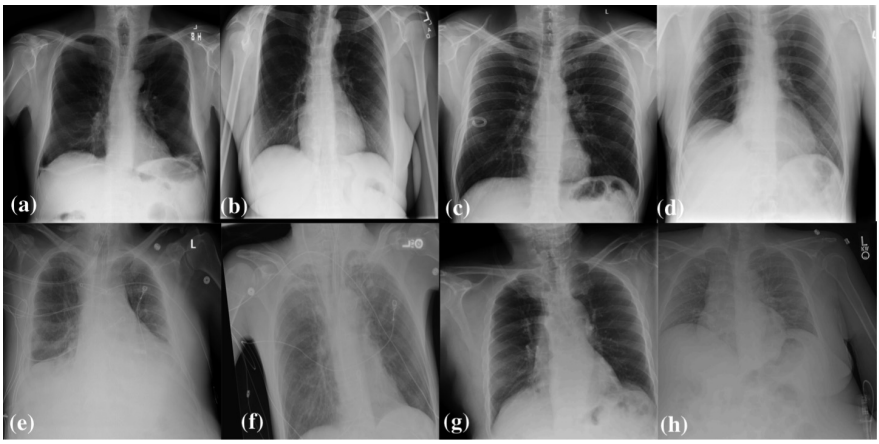

Fig. 1. Good and bad lung field data collected from different data sources. (a) NIH good lung field image, (b) NIH bad lung field image, (c) Source 1 good lung field image, (d) Source 1 bad lung field image, (e) Source 2 good lung field image, (f) Source 2 bad lung field image, (g) Source 3 good lung field image and (h) Source 3 bad lung field image.

In this paper, we propose a generalized novel approach to find the confident predictions when applying DNN models. Borrowing the ideas from Inception Score (IS) [7] and Fréchet Inception Distance (FID) [8] to measure similarity across image datasets, the proposed method projects correctly predicted training data samples onto a standard feature space (ImageNet [9] feature space based on the VGG16 [10] network). An outlier detection model is then trained and built based on a standard feature space to identify the inlier samples on the training dataset. Such inlier samples are considered as confident samples of the model as defined by the projected features from the training dataset. Unseen data is projected onto a standard feature space and passed through the outlier detector. The unseen data is not touched during training and only given to the model for model prediction. The model is considered to be confident on predictions on inlier samples as those samples are expected to be similar to the training dataset. Otherwise, the model is not expected to be confident, and we cannot rely on the model's prediction. To evaluate the proposed approach, several datasets are used in this paper: the Modified National Institute of Standards and Technology (MNIST) handwritten digits [11] as training data with the United States Postal Service (USPS) [12] and Street View House Number (SVHN) [13] digit datasets as test data, and the National Institutes of Health (NIH) Chest X-ray dataset [14] as training data with private test datasets across three U.S. and Canada hospitals. Three state-of-art DNN networks including VGG16 [10], ResNet50 [15] and DenseNet121 [16] were employed to evaluate the generalizability of this approach.

2 Methods

For an unseen dataset, a DNN model is expected to maintain consistent performance on data that are similar or close to training data. In order to compare training data to unseen data, it is beneficial to project those datasets onto a standard feature space as shown in the literature [7]. In this study, we utilize the ImageNet feature space as our standard feature space. A model is expected to have consistent performance on data that are close to the training data in the ImageNet feature space. The workflow of proposed method is shown in Fig. 2. When a new unseen sample is given, first, it is projected onto the ImageNet feature space to generate a feature vector. Second, the outlier detector predicts whether the unseen data sample is an inlier (close to training dataset) or an outlier (far from training dataset). It is noted that a data-driven model learns common patterns from most samples within a group or class and that outlier samples are more often associated with wrong predictions. We expect a DNN model to maintain consistent performance on inliers when compared to the performance on test samples from the same distribution as the training data.

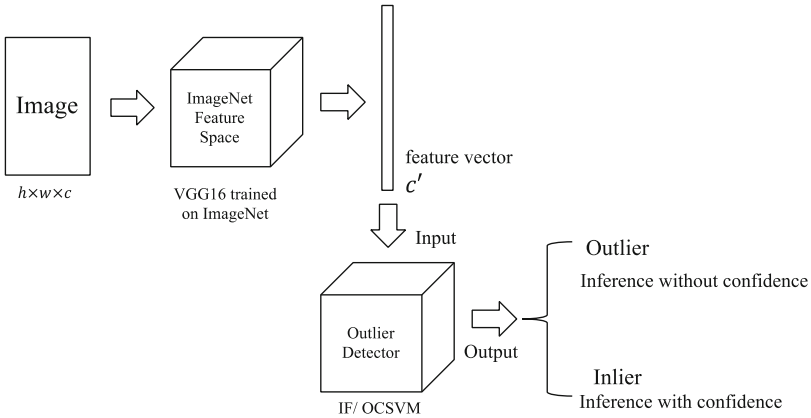

Fig. 2. Illustration of the proposed method.

2.1 Standard Feature Space

To evaluate the similarity between two image datasets, Salimans et al. [7] first proposed the Inception Score (IS), which is computed by projecting both image datasets onto the ImageNet label space using an ImageNet pre-trained Inception model [17]. Images are projected onto this standard space because ImageNet provides 1000 categories of classification labels and it has been shown that it is more reliable to calculate entropies over classification labels in a standard space. Later, Heusel et al. [8] proposed the Fréchet Inception Distance (FID) to improve the consistency of the IS. They used the features from coding layers instead of classification labels to obtain vision-relevant features. The Fréchet distance is calculated from this standard feature space to improve the consistency of similarity measurements across two image datasets.

Borrowing the same idea, to better measure the similarity between the training data and the unseen data, we project our images onto the ImageNet feature space. Due to the page limit, only the results of the standard feature space built on the VGG16 network pretrained on ImageNet are shown in this paper. Results using the ResNet50, DenseNet121 and InceptionV3 pre-trained ImageNet standard feature spaces are given in the Supplementary Material.

2.2 Outlier Detector

Both IS and FID explicitly measure the distribution similarity between two image sets. However, a large number of images, on the order of thousands of samples, is required to accurately estimate the similarity measurement. In this paper, we aim to identify the similarity of each individual case from an unseen data source against the training dataset. As such, IS and FID cannot be applied here. Therefore, we employ an outlier detection method to classify the data into inliers and outliers to inexplicitly measure the distance between individual cases from an unseen data source and the training dataset.

State of art outlier detector methods in the literature include Isolation Forest (IF) [18] and one-class support vector machine (OCSVM) [19]. IF algorithm is a

powerful unsupervised ensemble method for outlier detection based on decision trees. The averaged random tree path length determines whether a data point is an inlier or an outlier. OCSVM identifies the smallest hypersphere consisting of all the data. Data points that fall inside the hypersphere are considered to be inliers, whereas data points outside the hypersphere are considered to be outliers.

In this study, we employed and compared both IF and OCSVM to detect inliers and outliers from unseen data against the training dataset. To construct the outlier detector, only correctly predicted training samples using DNN models are projected onto a standard feature space (ImageNet feature space). The outlier detector is built on these extracted features and is given an outlier ratio, which is defined as the proportion of outliers present in the training data. The outlier ratio helps the outlier detector define the decision boundary to determine outlier data points.

Samples classified as inliers by the outlier detector are considered as confident samples for the DNN models. For the unseen data, model performance is evaluated on the confident samples to evaluate performance improvement on such samples. Detailed experiments will be discussed in next section.

3 Experiments

To validate the proposed method, two sets of experiments were conducted on the MNIST classification [11] and NIH chest X-ray lung field classification tasks [20]. To address these classification tasks, state of the art DNN models including VGG16, ResNet50 and DenseNet121 networks were retrained on the training datasets. The ResNet50, DenseNet121 and VGG16 models were initialized with pre-trained ImageNet weights. In each case, the training process followed the standard training process in Keras [21] with both horizontal and vertical random flipping augmentation during training. The Adam optimizer [22] with a default learning rate of 0.001 was used during training. After the DNN models were trained and tested, the correctly predicted images from the training dataset were projected into the standard space to extract features as stated in Sect. 2. IF and OCSVM outlier detector models were constructed with those extracted features and used to remove outliers. Outlier ratios of 0.1, 0.2, 0.3, 0.4 and 0.5 were selected. Other settings of IF and OCSVM were based on default settings from the Python Scikit-Learn package.

3.1 MNIST Classification

MNIST classification is a classical machine learning task that classifies handwritten digit images into 10 classes (numbers 0 to 9). First, we built MNIST classification models based on VGG16, ResNet50 and DenseNet121. Second, we evaluated those models separately on the MNIST test, USPS and SVHN datasets. As shown in Fig. 3, the USPS dataset [12] is more visually similar to the MNIST dataset while the SVHN dataset [13] is very different. The results reported in Table 1 indicate that the models trained on MNSIT dataset performed reasonably well on USPS data but performed poorly on SVHN dataset.

Fig. 3. Sample images from (a) MNIST, (b) USPS, and (c) SVHN datasets.

Benchmark Experimental Results. We benchmark the performance of the DNN models on unseen data. Benchmark accuracy is defined as the classification accuracy of the retrained DNN model on the entire unseen dataset before outlier detection. From Table 1, we observe that the model performance dropped substantially on the unseen USPS and SVHN datasets when compared to the MNIST test dataset. For the ResNet50 model, the accuracy dropped by 16.23% and 77.94% from the MNIST test dataset to the USPS and SVHN datasets, respectively. For DenseNet121, the accuracy dropped by 8.18% and 80.79% from MNIST test dataset to the USPS and SVHN datasets, respectively. For VGG16, the accuracy dropped by 6.54% and 65.97% from MNIST test dataset to the USPS and SVHN datasets, respectively. This reduction in model performance indicates a failure to generalize to unseen data.

Table 1. Performance of the proposed method on the MNIST test, USPS and SVHN datasets

MNIST test data		ResNet50 (99.53%)					DenseNet121 (99.67%)					VGG16 (99.66%)				
Outlier ratio		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
IF	In	99.54	99.57	99.61	99.59	99.69	99.68	99.72	99.68	99.66	99.73	99.67	99.68	99.66	99.69	99.72
	Out	99.47	99.36	99.33	99.43	99.36	99.53	99.48	99.65	99.69	99.61	99.57	99.59	99.66	99.62	99.59
OCSVM	In	99.53	99.57	99.55	99.62	99.61	99.69	99.68	99.66	99.69	99.69	99.66	99.66	99.66	99.68	99.66
	Out	99.49	99.38	99.48	99.39	99.45	99.49	99.64	99.69	99.64	99.65	99.7	99.65	99.66	99.63	99.66
USPS benchmark		ResNet50 (83.30%)					DenseNet121 (91.49%)					VGG16 (93.12%)				
Outlier ratio		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
IF	In	83.53	83.82	84.34	85.09	85.87	91.61	91.65	91.77	91.93	92.34	93.07	93.06	93.12	93.28	93.63
	Out	81.88	81.6	81.14	80.87	80.91	90.64	90.92	90.85	90.84	90.62	93.4	93.29	93.11	92.91	92.66
OCSVM	In	83.61	83.63	84.41	85.25	86.59	91.37	91.09	90.9	90.93	91.54	92.98	92.81	92.71	92.9	93.54
	Out	81.87	82.55	81.63	81.4	81.26	92.04	92.44	92.36	92.03	91.46	93.8	93.86	93.74	93.33	92.86
SVHN benchmark		ResNet50 (21.59%)					DenseNet121 (18.88%)					VGG16 (33.69%)				
Outlier ratio		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
IF	In	50	100	–	–	–	33.33	100	–	–	–	60	–	–	–	–
	Out	21.56	21.58	21.59	21.59	21.59	18.87	18.87	18.88	18.88	18.88	33.68	33.69	33.69	33.69	33.69
OCSVM	In	83.33	100	100	–	–	75	66.67	100	–	–	100	100	100	–	–
	Out	21.55	21.57	21.58	21.59	21.59	18.86	18.87	18.87	18.88	18.88	33.65	33.68	33.68	33.69	33.69

Experimental Results on Inliers. When applying the proposed method, results in Table 1 show that model performance generally improved on confident samples (inliers) when using both IF and OCSVM outlier detectors across different outlier ratios. For the IF outlier detector when compared to benchmark, ResNet50 accuracy improved from 83.30% up to 85.87% on USPS data and 21.59% up to 100% on SVHN data. Similarly, for the OCSVM detector, ResNet50 accuracy improved from 83.30% up to 86.59% on USPS data and 21.59% up to 100% on SVHN data. Similar trends were observed for DenseNet121 and VGG16 (Table 1). The above results empirically support our hypothesis that a generic model will perform consistently well on inlier data. The number of inliers and outliers detected for each dataset are given in the Supplementary Material (Tables S4–S6).

3.2 Chest X-ray Lung Field Classification

The chest frontal X-ray lung field classification was chosen to evaluate the performance of the proposed method. Determination of the acceptability of a patient positioning is one of the key components for successful Quality Control (QC) in a radiology department workflow. All clipped chest X-ray lung field images were considered to belong to the bad lung field class. Figure 1 shows examples of good and bad lung field chest X-ray images from different data sources.

As in the MNIST experiment, the ResNet50, DenseNet121 and VGG16 were selected and trained on the NIH chest X-ray dataset. Three additional datasets from different hospitals across the U.S. and Canada were collected as unseen datasets to evaluate the proposed method. Due to the data contracts, the names of those three hospitals were anonymized to Source 1, Source 2 and Source 3 in this paper.

The NIH dataset had over 112,000 of images from more than 30,000 patients, including many with advanced lung diseases. We selected and manually annotated 7,856 images with good lung field where the whole lung field was clearly visible without any clipping. Manual annotation was performed with the help of Radiology Technologists. In the same manner, 4,356 images with bad lung field were also selected. All the models were trained only on the NIH X-ray dataset with a random split of 75%/15%/15% for training/validation/testing. The three other datasets from Sources 1, 2 and 3 were selected and manually annotated with the same criteria and used as unseen datasets.

Benchmark Experimental Results. From Fig. 4, we observe that the model benchmark accuracy dropped substantially on the unseen datasets from Source 1, 2 and 3 when compared to the NIH test dataset benchmark accuracy. For the ResNet50 model, the accuracy performance dropped by 8.19%, 20.09% and 7.86% from the NIH test dataset on Sources 1, 2 and 3, respectively. For DenseNet121, the accuracy performance dropped by 8.31%, 19.64% and 6.03% from the NIH test dataset on Sources 1, 2 and 3, respectively. For VGG16, the accuracy performance dropped by 7.80%, 14.13% and 7.02% from the NIH test dataset on Sources 1, 2 and 3, respectively.

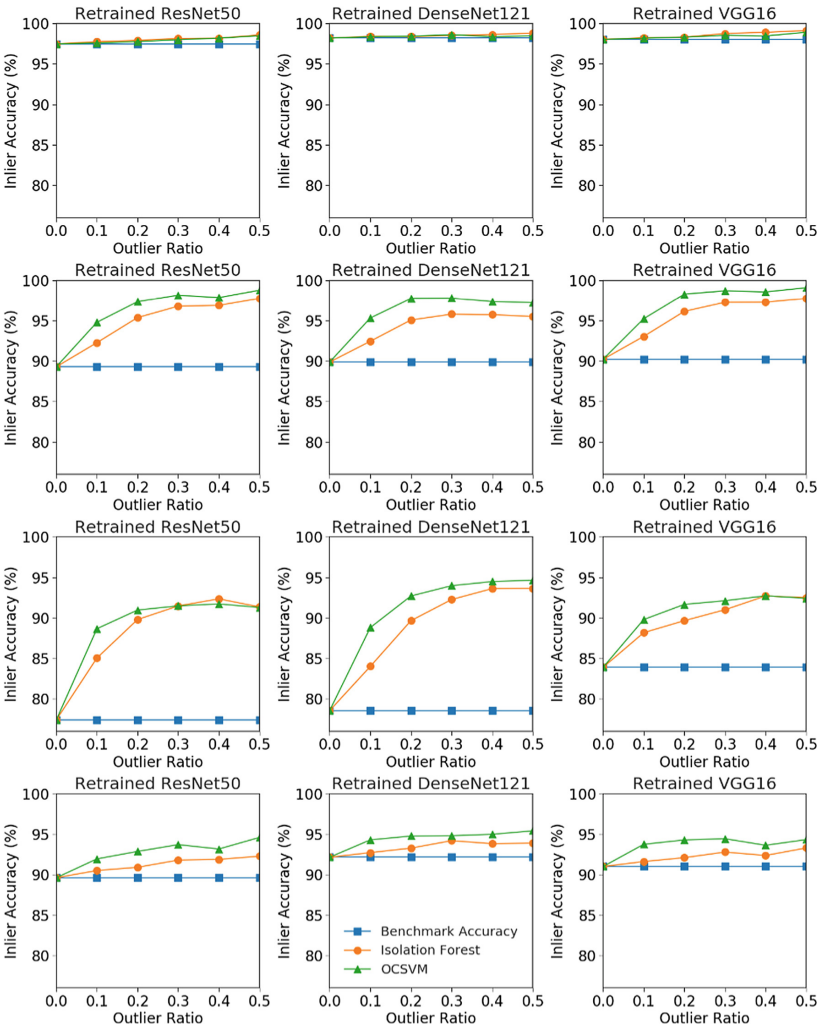

Fig. 4. Inlier classification accuracy (%) as a function of outlier ratio across X-ray image datasets originating from (a) NIH test, (b) Source 1, (c) Source 2 and (d) Source 3. Benchmark accuracy is defined as the classification accuracy of the retrained DNN model on the entire unseen dataset before outlier detection.

Experimental Results on Inliers. While applying the proposed method, results in Fig. 4 show that the performance of all models improved for both IF and OCSVM across different outlier ratios. When using IF as the outlier detector, the ResNet50 accuracy improved from 89.30%, 77.40% and 89.63% up to 97.75%, 92.35% and 92.29% on Sources 1, 2 and 3, respectively. DenseNet121 accuracy improved from 89.89%, 78.56% and 92.17% up to 95.75%, 93.64% and 94.21% on Sources 1, 2 and 3, respectively. VGG16 accuracy improved from 90.23%, 83.90% and 91.01% up to

97.75%, 92.71% and 93.31% on Sources 1, 2 and 3 respectively. By using OCSVM, the best accuracy that ResNet50 can achieve is up to 98.77%, 91.73% and 94.61% on Sources 1, 2 and 3, respectively. The best accuracy that DenseNet121 can achieve is up to 97.78%, 94.66% and 95.42% on Sources 1, 2 and 3, respectively. The best accuracy that VGG16 can achieve is 99.08%, 92.73% and 94.45% on data Sources 1, 2 and 3, respectively. Similar to the experimental results on the MNIST digit images, the experimental results on medical chest X-ray images also show that a generic model performed better on its confident samples (inliers) across different outlier detectors. The number of inliers and outliers detected for each dataset are given in the Supplementary Material (Tables S11–S14).

4 Discussion and Conclusion

As observed in both Table 1 and Fig. 4, the model has improved performance on the inliers of unseen data despite the fact that the model would typically not be confident on most of the samples in the unseen datasets. This can help clinicians prioritize and allocate more resources to cases where the model is not confident. However, the quality of proposed method heavily depends on the robustness of feature extractors and the outlier detectors that are used. In this paper, we consider the feature space constructed from the VGG16 network trained on ImageNet. Additionally, based on our preliminary experiments, we have noticed that when the outlier detector was not well constructed, the inliers identified were not implying correct predictions. One potential way to resolve these issues, which is one of our future work is to build an end-to-end framework incorporating both Bayesian approach (resolves uncertainty) and outlier detection (resolves domain-shift).

In this study, we proposed a novel method to determine the confident samples. An outlier detection model was constructed based on the features extracted from the standard feature space. Inliers from the detection model were considered as the confident samples from the unseen data. Experimental results showed that the performance of the ResNet50, DenseNet121 and VGG16 models improved on all unseen datasets after filtering out outliers.

References

1. Chen, T., Navrátil, J., Iyengar, V., Shanmugam, K.: Confidence scoring using whitebox meta-models with linear classifier probes. arXiv preprint arXiv:1805.05396 (2018)
2. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
3. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 655–663. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_74

4. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. **7**, 17816 (2017)
5. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, pp. 4878–4887 (2017)
6. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
7. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
9. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
12. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. **16**, 550–554 (1994)
13. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
14. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
18. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
19. Moya, M.M., Hush, D.R.: Network constraints and multi-objective optimization for one-class classification. Neural Networks. **9**, 463–474 (1996)
20. Competition, B.P.: AAPM spring clinical meeting-abstracts SATURDAY, APRIL 7. J. Appl. Clin. Med. Phys. **19**, 370–407 (2018)
21. Chollet, F.: Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek. MITP-Verlags GmbH & Co. KG (2018)
22. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)