Assessment of Manifold Unfolding in Trained Deep Neural Network Classifiers

Štefan Pócoš, Iveta Bečková, Tomáš Kuzma, Igor Farkaš

Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava

Abstract

Research on explainable artificial intelligence has progressed remarkably in the last years. In the subfield of deep learning, considerable effort has been invested to understanding deep classifiers that have proven successful in training on various benchmark datasets. Within the methods focusing on geometry-based understanding of the trained models, an interesting, *manifold disentanglement hypothesis* has been proposed. This hypothesis, supported by quantitative evidence, suggests that the class distributions become gradually reorganized over the hidden layers towards lower inherent dimensionality and hence easier separability. In this work, we extend our results, concerning four datasets of low and medium complexity, and using three different assessment methods that provide robust consistent support for manifold untangling. In particular, our quantitative analysis supports the hypothesis that the data manifold becomes flattened, and the class distributions become better separable towards higher layers.

Keywords: neural networks, manifold unfolding, embedding complexity

Introduction

Deep neural networks, albeit data greedy, have proven successful in learning various complex, mostly supervised end-to-end tasks [13]. On the other hand, these black-box models are primary candidates for the need to apply to them techniques, allowing the users to understand the functioning of the trained model [14]. Poorly understood functionality of models can have serious consequences, as for example the sensitivity to adversarial examples (explained and investigated in [18, 5]), bad generalisation and overall low credibility of used models.

Research towards interpretable and explainable AI (XAI) has progressed considerably in the last years and has become widely acknowledged as a crucial feature for the practical deployment of AI models (for an extensive overview, see e. g. [1, 19, 4]). Within XAI focusing on deep learning, various analytical or visualization methods have been proposed trying to shed light on models' behavior and the internal causal processes [11, 10, 15].

One of the research goals in this direction is to understand the neuron activations in trained deep classifiers, typically from a layer perspective. The function of a deep neural network mathematically corresponds to a cascade of smooth nonlinear transformations, each composed of a linear mapping (mediated by the complete matrix of individual neuron weights) and a subsequent nonlinearity (activation functions). This lends itself to an interpretation allowing us to examine the process of classifying an input from a manifold perspective. *The manifold untangling hypothesis* was proposed in [2], also supported by quantitative evidence, suggesting that the process of mappping the inputs across layers can be interpreted as the flattening of manifold-shaped data in high-dimensional spaces (hidden layers). The data is often assumed to lie on a *low-dimensional manifold* that becomes eventually partitioned in the output space where the final layer has one neuron for each class, and the classification is determined by which neuron has the largest activation [12]. This process of manifold disentangling may be tightly related to the aforementioned problem of adversarial examples, as they were studied in the context of their distances to (input data) manifolds (for details, see [3, 17]).

In [7] we proposed a method to assess the extent of the class separation effect by testing several measures based on the *embedding complexity* of the internal representations, of which the t-distributed stochastic neighbour embedding (t-SNE) turned out to be the most suitable method. Using t-SNE we then demonstrated the validity of the disentanglement hypothesis by measuring embedding complexity, classification accuracy and their relation on a sample of image classification datasets. The process was observed to be very robust, occurring for various activation functions and regardless of the number of hidden layers.

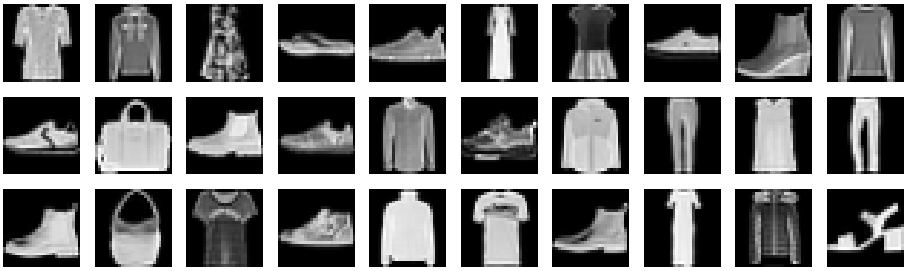
In this work we extend the validity of our work in two ways, by adding two more datasets into computational analysis and by adding two methods that help shed light on the manifold untangling hypothesis.

Datasets

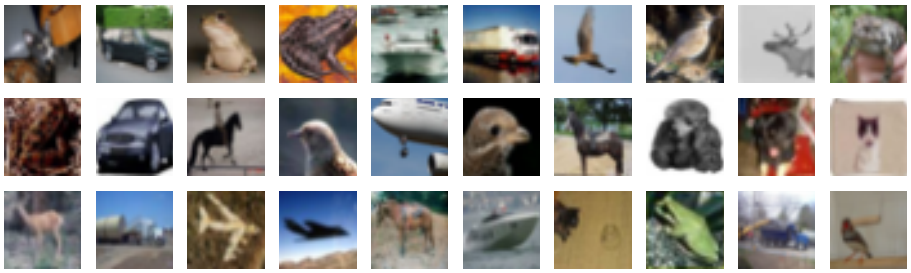
To study the process of untangling class manifolds, datasets of medium complexity are required. The complexity should be high enough so that the problem cannot be solved in the input space by a simple classifier, but a complex transformation, such as by an artificial neural network, is necessary. However, at the same time, the untransformed or partially transformed inputs cannot be inscrutable to available embedding methods as to remain interpretable. These restrictions led us to select four suitable datasets, all inadvertently being visual tasks.

In [7] we tested MNIST and SVHN datasets. MNIST [8] is a basic well-known dataset for optical character recognition (handwritten grayscale digits). The SVHN (StreetView House Numbers dataset) is a more challenging task for digit recognition, which adds color, distracting surroundings, blurring and oblique perspectives. In this work, we add two more datasets of moderate complexity: F(ashion)-MNIST and CIFAR-10. F-MNIST [20] is a dataset of a higher complexity compared to MNIST. It consists of 60 000 training and 10 000 testing

images in ten classes. Data are 28×28 grayscale images of clothing or accessories. Some examples:



The CIFAR-10 [6] dataset provides a variety of 32×32 color images with medium difficulty of classification. As in previous cases, the dataset consists of exclusive 10 classes, but here they depict animals or vehicles. Training and testing subsets contain 50 000 and 10 000 images, respectively. A few random examples:



Models

Following [7], we employ simple (deep) feed-forward networks that are minimally powerful enough to satisfactorily classify the selected datasets. By default, we use fully-connected layers of 100 neurons and use the same activation function at each hidden layer. In [7] we tested four activation functions: logistic sigmoid, hyperbolic tangent, softsign, and rectified linear units (ReLU). Of these, ReLU worked best for t-SNE, so we only focus on it henceforth.

For evaluation of CIFAR-10 we use a more complex architecture to get satisfactory classification accuracy. The architecture consists of 1-to-4 VGG [16] type blocks (two convolutional layers followed by pooling) and a fully-connected hidden layer of 100 neurons before the output layer. In order to achieve higher accuracy we also use dropout.

The final classification layer has a neuron for each class with a *softmax* activation. All of these networks can be satisfactorily trained within 100 epochs using a simple stochastic gradient descent with momentum.

Method

To assess the progress of the manifold disentanglement process we measure the *embedding complexity*, i.e. how difficult is to embed the activation vectors for a balanced sample of training inputs to a lower dimensional space. For the purpose of visualization we use 2D (or 3D) output space. In [7] we examined several popular embedding methods (in order of increasing sophistication): PCA, LLE, MDS, Isomap and t-SNE [9], of which the last one turned out to be the most useful. Here, we hence only use t-distributed Stochastic Neighbour Embedding, a popular non-linear embedding method, which is based on preserving the *stochastic neighbourhood* of elements. This stands in contrast to more conventional methods which usually use a fixed neighbourhood, either using an adjustable parameter of the algorithm (e.g. k nearest neighbours in LLE or Isomap), or optimized to satisfy an internal condition, or factor all data points into consideration (e.g. MDS or PCA). The disparity between the distributions of probabilities in the input and output spaces in t-SNE is quantified by *KL-divergence* and the desired embedding is then produced by minimizing this divergence with respect to the placement of points in the output space.

Results

We train five independent networks (or runs) for one to seven layers on F-MNIST and one to four VGG blocks on CIFAR-10. For F-MNIST dataset the increasing number of layers did not improve the accuracy (~ 88 %) but in case of CIFAR-10 the accuracy increased from 71 % (using one block) to 82 % (using four blocks). We then sample activations of each hidden neuron for 100 randomly selected input samples for each of the 10 input classes, yielding 1000 activation vectors for each layer. In each independent run, we then embed those activation into two dimensions using t-SNE and measure the resulting KL-divergence as the hardness score. For a quantitative overview, we plot the result of all runs into a single bar chart, with the averaged value shown as the bold line and individual runs as translucent overlapping rectangles (this is an alternate version of a boxplot, which puts the greatest emphasis on the mean value).

The phenomenon of decreasing KL-divergence is also visible in a more complex (convolutional) network, trained on CIFAR-10. KL-divergence scores for progressive layers of a chosen network are depicted in Fig. 1. As in the case of simpler networks, after using t-SNE we observe clustering of data which belongs to the same class. Moving to higher layers, individual clusters become more discernible and separable. The t-SNE visualization of layer activations in a 4-block VGG network in Fig. 2 illustrates the unfolding process at several stages. Fig. 3 confirms the intuition that the model benefits from the higher number of layers.

The idea of measuring the KL-divergence after manifold disentanglement using t-SNE to two dimensions can be extended to three or more dimensions. Since t-SNE is computationally demanding, we only additionally test embedding to

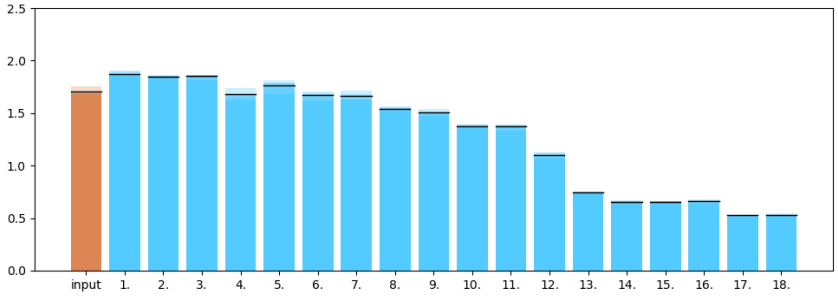


Fig. 1: KL-divergence on a 4-block VGG-type network, trained on CIFAR-10.

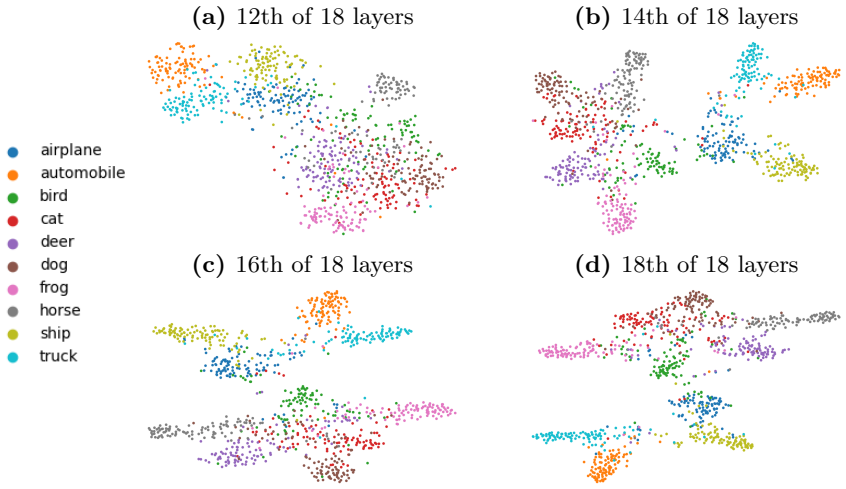


Fig. 2: t-SNE visualisation of activations on a 4-block VGG-type network trained on CIFAR-10. The higher (deeper) the layer, the clearer clusters are obtained.

three dimensions. Results show us that embeddings to three dimensions are somewhat closer to original (high-dimensional) manifolds as they have a lower KL-divergence score. Fig. 4 provides a comparison of KL-divergence scores for a randomly chosen network trained on F-MNIST dataset, whereas in Fig. 5 we can find an example of 2D and 3D embeddings on the same network.

Assessing class distributions

Another way to assess manifold disentanglement is to measure the complexity (in terms of data distribution) of the manifolds themselves. This can be done by performing the singular value decomposition (SVD) and evaluating, how many eigenvalues (and their corresponding principal components) explain 95 % of variance in the given data, in this case per-class activations on hidden layers. We chose to not perform this testing on models trained on CIFAR-10, as the number

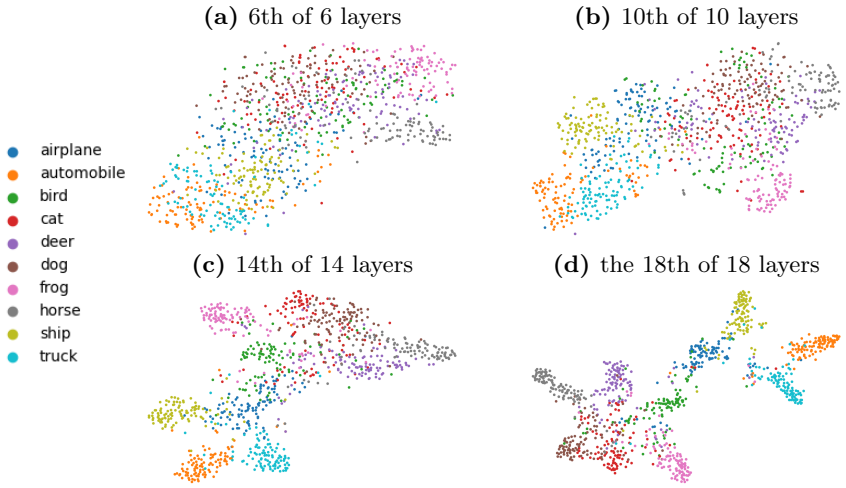


Fig. 3: t-SNE visualisation of classes on 1-to-4 block VGG-type network performed on the last hidden layer. It is evident that more blocks are beneficial for better unfolding of the classes, which leads to higher classification rate.

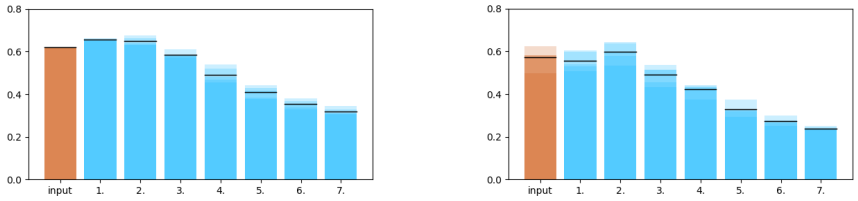


Fig. 4: KL-divergence scores for t-SNE embedding into two (left) and three (right) dimensions on a 7-layer network trained on F-MNIST.

of neurons on hidden layers varies rapidly, due to the use of convolution and pooling. Thus the results might be harder to interpret, for example to determine, to what extent were the changes of complexity driven by the change of number of hidden neurons.

The other three datasets provided the following observations. Models trained on MNIST and SVHN datasets show the same trend, regardless the class or the number of hidden layers. In case of MNIST, complexity of manifolds decreases through the network. In case of SVHN it increases rapidly through first 2 layers, then slowly decreases (as shown in Fig. 6). This difference between MNIST and SVHN might be caused by the different input size (784 vs. 3072), but also different task complexity with SVHN being much more demanding, thus resulting in a lower classification accuracy. Models trained on F-MNIST are slightly different and it is possible to observe two distinct behaviors dependent on the class,

(a) 2D embedding **(b)** 3D embedding

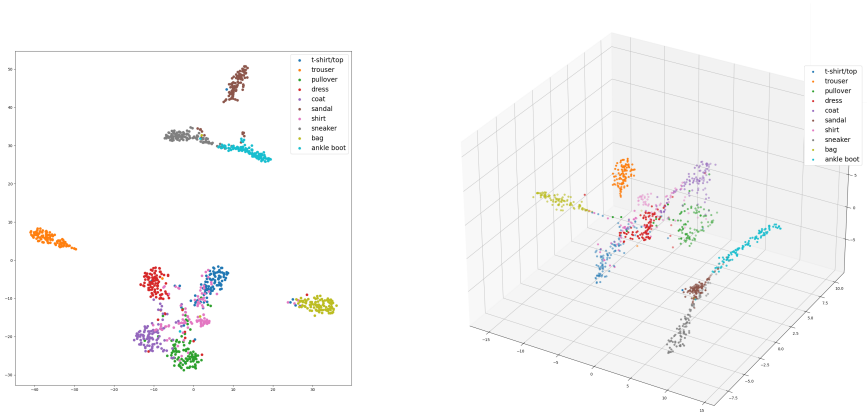


Fig. 5: Examples of t-SNE embeddings of activations (on the 7th layer) into various dimensions on a feed-forward network with 7 hidden layers trained on F-MNIST.

shown in Fig. 7. Classes 'trouser', 'sandal', 'sneaker', 'bag' and 'ankle-boot' show the same trend as in case of MNIST and complexity monotonically decreases. Classes 't-shirt/top', 'pullover', 'dress', 'coat' and 'shirt' are similar to SVHN and complexity initially rises and decreases afterwards. This seemingly correlates with per-class classification accuracy (Fig. 8) and even t-SNE embedding of the model activations, as classes 'trouser', 'sandal', 'sneaker', 'bag' and 'ankle-boot' are classified with a higher accuracy and also form clusters easier than the remaining classes. These observations show possible correlation between model accuracy and the development of activation complexity, offering interesting ideas for future research.

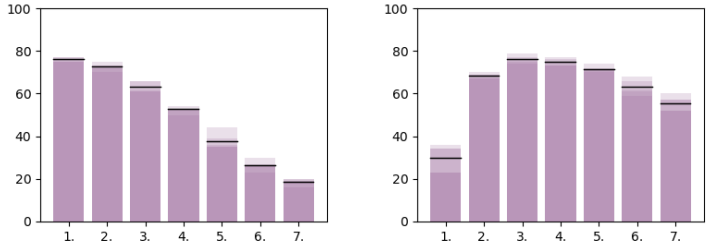


Fig. 6: Number of eigenvalues explaining 95 % of variance through the hidden layers for class 5 from MNIST (left) and class 7 from SVHN (right).

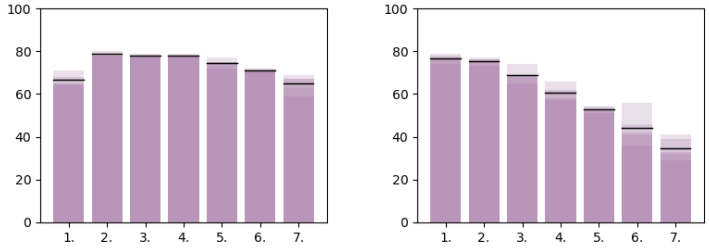


Fig. 7: Number of eigenvalues explaining 95 % of variance through the hidden layers for classes 'pullover' (left) and 'sandal' (right) from F-MNIST.

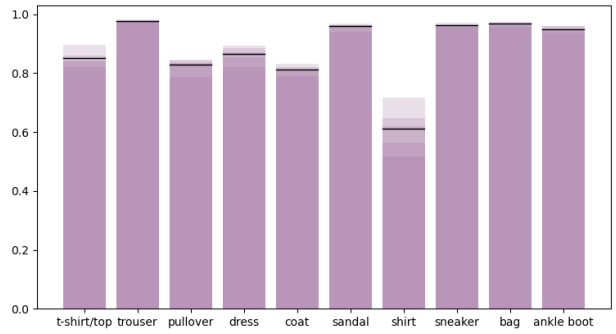


Fig. 8: Per-class accuracy of a model with 7 hidden layers trained on F-MNIST.

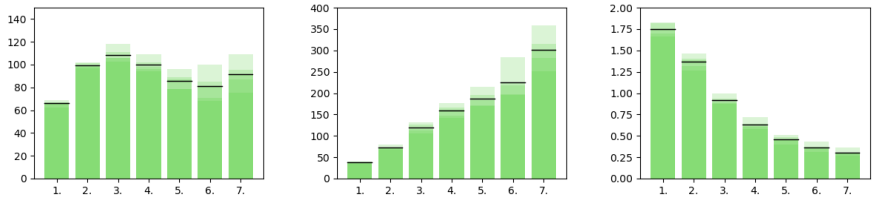


Fig. 9: Intra-class (left), inter-class (middle) and the ratio (right) of mean distances computed from the internal activation vectors on a model with 7 hidden layers trained on F-MNIST.

Assessing class separation

Finally, we support manifold disentanglement hypothesis by appreciating the degree of class separability. We do this by monitoring the inter-class and intra-class distances between activation vectors on hidden layers of trained networks. Let us denote the hidden neuron activations (of j-th sample) belonging to a class k as x_j^k, where $k = 1, 2, ..., K$. Let N be the number of all samples, and N^k be the number of samples belonging to class k. Let c^k be the centroid of k-th

class and C the arithmetic mean of all samples. The mean inter-class distance is defined as a mean squared (Euclidean) distance of class centroids from C and the mean intra-class distance as an average of squared distances of all class samples from the corresponding class centroid, i.e.

$$Dist_{\text{Inter}} = \frac{1}{K}\sum_{k=1}^{K}\|c^k - C\|^2 \qquad Dist_{\text{Intra}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N^k}\sum_{j=1}^{N^k}\|x_j^k - c^k\|^2 \quad (1)$$

For this task, we chose a model with 7 hidden layers, trained on F-MNIST.* This dataset uses classes with an equal number of samples, so there is no need to perform weighted average while evaluating average distances. At the same time, all hidden layers have the same dimensionality, which allows consistent comparisons across layers.

From results depicted in Fig. 9 it can be seen that the mean intra-class distance changes non-monotonically, at smaller magnitudes, but the inter-class distance grows steadily, dominating the trend. Thus, their decreasing ratio indicates that the classes become gradually unfolded, hence allowing their better separation on higher layers, which supports the manifold disentanglement hypothesis.

Conclusion

In this paper, we extend our results focused on visual and computational analysis of trained deep neural network classifiers. The qualitative and quantitative results related to four chosen datasets of a simple and medium complexity point towards the data manifold unfolding process that explains geometric changes in data distribution. This drives the formation of high-dimensional activation vectors that step-by-step, given the available number of hidden layers, appear to contribute to better separability of classes. This process is typically associated with a gradual decrease of the manifold inherent dimensionality, which is quantified by the quality of (decreasing) complexity embedding into 2D (or 3D), as measured by KL divergence of t-SNE visualization method. This robust process has consistently been confirmed by three different methods we used in this paper. At the same time, these provide cues for further research, for instance, what is the relationship between manifold unfolding and the network accuracy, or where are the adversarial examples situated relatively to the manifolds of each layer.

Acknowledgment. This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215, and in part by Slovak national projects VEGA 1/0796/18 and KEGA 042UK-4/2019.

* For the same reason as in SVD, we did not analyse here CIFAR-10 dataset.

Bibliography

[1] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)

[2] Brahma, P.P., Wu, D., She, Y.: Why deep learning works: A manifold disentanglement perspective. IEEE Transactions on Neural Networks and Learning Systems **10**(27), 1997–2008 (2016)

[3] Gilmer, J., Metz, L., Faghri, F., Schoenholz, S.S., Raghu, M., Wattenberg, M., Goodfellow, I.: Adversarial spheres (2018), arXiv:1801.02774 [cs.CV]

[4] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: International Conference on Neural Information Processing. pp. 378–385 (2020)

[5] Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)

[6] Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. Rep. TR-2009, University of Toronto (2009)

[7] Kuzma, T., Farkaš, I.: Embedding complexity of learned representations in neural networks. In: Proceedings of 28th International Conference on Artificial Neural Networks (ICANN). vol. 2, pp. 518–528 (2019)

[8] LeCun, Y., Cortes, C.: MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/ (2010)

[9] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)

[10] Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)

[11] Montúfar, G.F., Pascanu, R., Cho, K., Bengio, Y.: On the number of linear regions of deep neural networks. In: Advances in Neural Information Processing Systems. pp. 2924–2932 (2014)

[12] Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., Shea-Brown, E.: Dimensionality compression and expansion in deep neural networks (2019), arXiv:1906.00443 [cs.LG]

[13] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85–117 (2015)

[14] Schubbach, A.: Judging machines: philosophical aspects of deep learning. Synthese (2019). https://doi.org/10.1007/s11229-019-02167-z

[15] Schulz, A., Hinder, F., Hammer, B.: DeepView: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence. pp. 2305–2311 (2020)

[16] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large scale image recognition. In: International Conference on Learning Representations (2015)

[17] Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization (2019), arXiv:1812.00740 [cs.CV]

[18] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)

[19] Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review (2020), aXiv:2006.00093 [cs.AI]

[20] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms (2017), arXiv:abs/1708.07747 [cs.LG]