Interpreting Intentionally Flawed Models with Linear Probes*

Mara Graziani, Henning Müller, Vincent Andrearczyk
University of Applied Sciences of Western Switzerland (HES-SO Valais), Switzerland
University of Geneva (UNIGE), Geneva, Switzerland
mara.graziani@hevs.ch, henning.mueller@hevs.ch, vincent.andrearczyk@hevs.ch

Abstract

The representational differences between generalizing networks and intentionally flawed models can be insightful on the dynamics of network training. Do memorizing networks, e.g. networks that learn random label correspondences, focus on specific patterns in the data to memorize the labels? Are the features learned by a generalizing network affected by randomization of the model parameters? In high-risk applications such as medical, legal or financial domains, highlighting the representational differences that help generalization may be even more important than the model performance itself. In this paper, we probe the activations of intermediate layers with linear classification and regression. Results show that the bias towards simple solutions of generalizing networks is maintained even when statistical irregularities are intentionally introduced.

1. Introduction

In this paper, we investigate the representational differences between Deep Neural Networks (DNNs) that learn to generalize and those that do not. Understanding the generalization properties of DNNs can ensure that their deployment in high-risk daily practices will lead to reliable decisions [7, 8, 11, 20]. The link between learning and generalization is still unclear, with over parametrized networks being able to achieve the best generalization performances and fit pure noise at the same time [5, 29, 30]. To shed some light about the learning behaviors of generalizing and non-generalizing models, we analyze the optimization bias towards simple solutions even when statistical irregularities are intentionally introduced (e.g. randomization of the training labels). Are there patterns in the data that are learned by both generalizing and memorizing networks? This paper proposes to consider the activation of an intermediate layer l as a geometric space and to look at linear combinations of the neuronal directions, which we call *linear probes* [2], as clues for the interpretation. Given a model M trained on the main task (e.g. DNN trained on image classification), an interpreter model M_i (e.g. the linear probe) is trained on an interpretability task in the activation space of layer l (hence M_i^l). In addition to the generalizing networks trained on correct data, two types of intentionally flawed models are used for the main task model M: i) networks with random initialization of the trainable parameters as in [1], which we call *random networks*; ii) networks trained on image datasets with different fractions of randomized labels (i.e. closed-set noise) as in [22, 28, 29]. For a given fraction N of corrupted labels, we refer to such network as *N-memorizing network*. The interpreter model M_i^l computes linear probes in the activation space of a layer l. The task of M_i^l consists of learning either linear classifier probes [2], Concept Activation Vectors (CAV) [16] or Regression Concept Vectors (RCVs) [12, 13]. Each technique gives different insights about the learned representations. Linear classifier probes measure the linear separability of the classes at intermediate layers of the DNN. CAVs interpret the DNN internal state in terms of human-friendly concepts. RCVs extend the original definition of CAVs from linear classification of binary concepts (e.g. presence or absence of a concept) to the linear regression of continuous-valued concept measures (e.g. the area of an object). In this work, we focus on linear classifier probes and RCVs.

1.1. Main contributions

Our main contributions and findings are the following:

- We propose an analysis of intentionally flawed models, i.e. random and N-memorizing networks by linearly probing the internal activation space with linear classifier probes [2] and RCVs [12, 13].
- We show in Sec. 3 that network training increases the linear separability of the classes in the activation space. Moreover, simple concepts become linearly regressable after training.
- Experiments in Sec 4 suggest that simple concepts are learned at early layers to solve the memorization

[*] The first workshop on Statistical Deep Learning for Computer Vision, in *Seoul, Korea*, 2019. Copyright by Author(s).

task. These concepts are then passed on to deep layers, where the random mapping is learned.

- We show in Sec. 5 that DNNs learn the non-corrupted data distribution earlier than the strong statistical irregularities artificially introduced by label corruption.

Differently from previous works on memorizing [5, 21, 22, 30] and randomly initialized networks [1, 24], the internal activations are interpreted with linear probes. In particular, our experiments focus on the representational differences between generalizing and faulty models in terms of simple concepts such as first (color) and second order (texture) statistics computed on the image pixels[1].

2. Related work

Intensive research focused on the comparison between generalizing and non-generalizing models [1, 4, 5, 18, 21, 22, 24, 28–30]. Part of these [1, 24, 27] suggested the existence of an ”architecture prior”, that is the impact that the architecture with randomly initialized parameters has on the learned representations, hence on the search space of the optimization. The analysis of models trained on noise [5, 18, 21, 22, 29, 30] showed that a sufficiently large DNN can fit data distributions with strong statistical irregularities, such as random labels [29]. Research in label noise modeling achieved some robustness to noise, particularly with mixup data augmentation [31] and loss correction [4]. Qualitative differences between learning noise and natural images, however, showed that DNNs are biased towards learning simple patterns before memorizing the out-of-distribution samples.

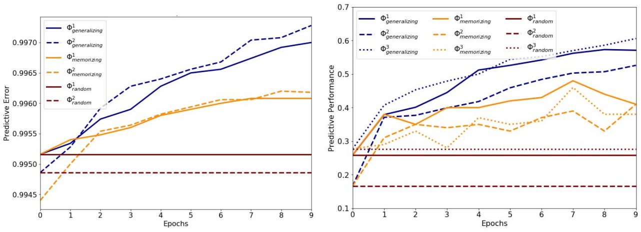
On a parallel side, an increasing number of studies has been addressing the challenging task of understanding what makes the representations learned by DNNs so successful, of which extensive surveys can be found in [8, 20]. Post-hoc interpretability methods are particularly suited to the analysis of flawed models since they allow to interpret the representations without the need for retraining or modifying the optimization task. Linear classifier probes [2] and CAVs [16] showed that the internal activations of a layer can be interpreted in terms of linear classifiers (of the class labels in the former and of the binary presence or absence of a human-friendly concept in the latter). RCVs extended the interpretability task of CAVs to learning continuous valued concept measures by linear regression [12, 13]. This method was insightful in the interpretation of DNNs for tasks in the field of computer vision and in the medical domain [13, 14]

This paper attempts to link the research on randomly initialized and memorizing networks to the interpretation of the learned representations with linear models.

3. Linear probes improve over training

In this paper, the model for the main task M is either a Multi-Layer Perceptron[2] (MLP), a shallow Convolutional Neural Network[3] (shallow-CNN) or an InceptionV3 network [25], trained on different image classification datasets. The MLP is trained for 1,000 epochs with Stochasitc Gradient Descent (SGD) and learning rate 0.01 as in [5]. The shallow-CNN also follows the setup in [5] and is trained for 100 epochs with SGD and learning rate 0.01. InceptionV3 is trained for 1,000 epochs with the Adam optimizer and standard parameters (learning rate 0.01, β_1 0.9 and β_2 0.999). Note that all the N-memorizing networks converge to a full overfit of the training data. The model choices are based upon relevant research in understanding deep learning [1, 2, 5, 29].

The MLP is trained on the dataset of handwritten digits MNIST [19], while the shallow-CNN is trained from scratch on a small subset of ImageNet [10]. The latter, referred to as ImageNet10, contains fewer well separated classes to better enhance the differences between generalizing and flawed networks. Five texture-like classes with high texture appearance (namely bookshop, butcher, chainlink fence, cliff dwelling and confectionery) and five object-like classes (namely acoustic guitar, ambulance, chihuahua, golden retriever, ladybug) are retained following the distinction between texture-like and object-like classes proposed in [3]. As a first analysis, we use linear classifier probes as the interpreter model M_i to evaluate the linear separability of the classes during training. Fig. 1 shows the predictive performance of the linear classifier probes on the activations ϕ^l of layer l in generalizing and flawed models. Evidently, training increases the linear separability of the classes in the learned internal representations.



(a) MNIST

(b) ImageNet10

Figure 1: Predictive performance of linear probes against training epochs on a held-out validation set for (a) MNIST and (b) ImageNet10; for the three types of networks: randomized, 0.4-memorizing and generalizing. Best on screen.

As a further analysis, InceptionV3 is trained to classify

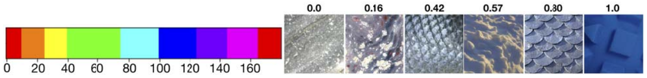
[1] Code for reproducibility at `github.com/maragraziani/intentionally_flawed_models`

[2] fc(4,096)-fc(4,096)-fc(#classes) with relu activations

[3] conv(200,5,5)-mp(3,3)-conv(200,5,5)-mp(3,3)-gap-fc(384)-fc(192)-fc(#classes) with bn and relu activations

the Describable Texture Dataset (DTD) [9]. DTD is a collection of 5,640 textural images organized in 47 categories inspired by human-centric attributes of perceptual properties of textures. The training images of original sizes ranging between 300 × 300 and 640 × 640 pixels are randomly cropped during training to the standard input size of InceptionV3 (299 × 299).

We extract concept measures of first and second order statistics from the image pixels. The *colorfulness* metric, based on opponent color spaces, is computed as in [15]. Besides, individual measures of the percentage of a specific color in the image are computed by applying the color quantization of the HSV (Hue, Saturation, Value) space shown in Fig. 2a. The HSV colorspace is closer to the human representation of hue ranges than RGB. For each of the eight bin quantizations, we define a distinct concept measure. For example, the *blue-ness* of the image is computed as $\frac{\#bluepixels}{\#pixels}$. Images of the DTD dataset sorted for increasing values of blue-ness can be inspected in Fig. 2b. The same technique is applied to the saturation ranges to obtain measures of white-ness and black-ness.



(a) HSV quantization

(b) Increasing blue-ness

Figure 2: Measuring the presence of individual colors in the image. (a) Quantization of the HSV color space (b) Examples of DTD images sorted from low to high blue-ness.

In this experiment, the interpreter model M_i is the RCV linear probe computed for a concept of interest. We analyze 11 concepts of color (i.e. the eight hue ranges, white-ness, black-ness and colorfulness) and six concepts of texture (i.e. energy, ASM, dissimilarity, homogeneity, contrast and correlation). Fig. 3 shows the determination coefficient R^2 of the RCV probes against the training epochs. For best presentation, we select two concepts of color and two concepts of texture and discuss similarity and differences of the trends with the remaining concepts.The increasing values of R^2 illustrate the learning of the concepts during training. We observe two main trends in the results. For some concepts, namely orange (Fig 3a left), dissimilarity (Fig. 3b left), contrast, correlation, homogeneity, red and colorfulness, the R^2 of the probes in the 1-memorizing network is markedly below the R^2 in the generalizing network. The R^2 of the RCVs for the concepts green (Fig 3a right), energy (Fig. 3b right), ASM, cyano, magenta, purple, yellow, black and blue, however, do not show significant differences between the two networks. A singular case is observed with white-ness, which reaches high R^2 after only 50 epochs for both the generalizing and memorizing network. The R^2 remains almost constant over training, suggesting that whiteness is quickly learned at the beginning of training and then remains easy to regress in both networks.

We further evaluate the RCVs by computing the Mean Squared Prediction Error (MSPE) on 376 data points that were not used in the estimation of the regression coefficients. The MSPE of random networks, 1-memorizing networks and generalizing networks are compared in Table 1. Network training drastically reduces the MSPE of both 1-memorizing and generalizing networks, as shown by the comparison with the random network.

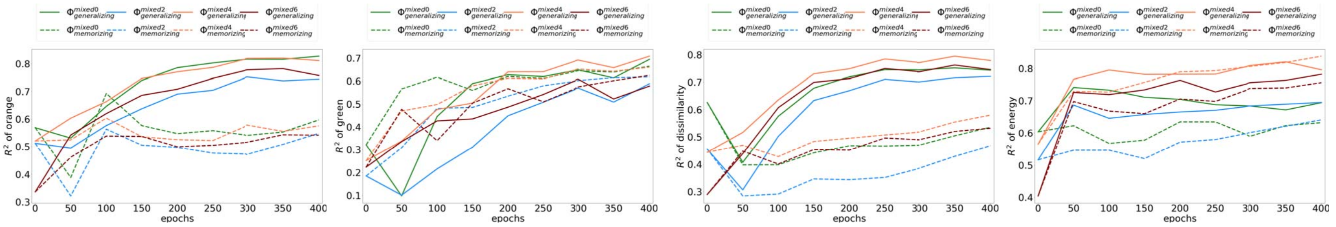
Table 1: MSPE for concepts of texture and color in mixed0 of InceptionV3 trained on DTD. Lower MSPE reveals higher predictive performance of the RCV.

model	cyano	energy	orange	correlation
random	2×10^5	2×10^6	7×10^5	8×10^5
1-mem.	**0.032**	**0.017**	0.29	0.10
gener.	0.070	0.030	**0.21**	**0.08**

4. Early layers focus on simple concepts

We define the complexity of a concept according to its position in the hierarchical structure of visual categories [6]. Simple concepts include low-level visual attributes of color and texture, while attributes of material, object parts, full objects and scenes have increasing complexity as they represent a progression to more abstract concepts. The results in Table 1 show that texture and color are learned by early layers of 1-memorizing networks. If we consider the fact that the main task of 1-memorizing networks is to learn the random mapping between the data points and the corrupted labels, we can conclude that learning texture and color is useful to the task. In other words, these simple concepts are learned to simplify the clustering of the learned representations to match the random labeling scheme. However, the conjecture proposed by Tishby in [26] claims that DNN training consists of an initial fitting phase and a subsequent compression phase. One could further analyze with linear classifier probes at different network depths whether the separability of the classes in the activation space happens before or after these concepts are learned. In the former scenario, the concepts are likely used to memorize the samples. In the latter, compression happens after memorization as suggested in [26]. Future work will address this point.

In the next experiment, we train N-memorizing MLPs with label corruption ratios ranging from zero to one on the MNIST dataset. We increase the MLP depth to six hidden layers and we probe simple concepts of shape for the MNIST task such as area, eccentricity and perimeter as in [12]. All concepts are best regressed in the representation space of the first layer, ϕ^1 (see Fig. 4 showing the RCV



(a) Color: green and orange

(b) Texture: dissimilarity and energy

Figure 3: R^2 of the regression of concepts of color (1st order statistics of pixel values) and texture (2nd order statistics) at intermediate layers of InceptionV3 (1-memorizing and generalizing) for DTD images. Best seen on screen.

probe of area. The regression of other concepts presents comparable behaviors). Increasing the fraction of label corruption does not affect the learning of the concepts in the first layer. By probing the layers at different depths, we inspect the representational differences introduced by the increasingly enforced memorization. As depth increases, the R^2 is more and more impacted by label corruption. We find that these results underline the importance of depth in memorizing network, already discussed in [21,23]. In particular, depth seems to play a fundamental role in the rearranging of the internal clustering of the data points to match the statistical irregularities introduced by the random labeling. Linear classifier probes could be used to further confirm this hypothesis.

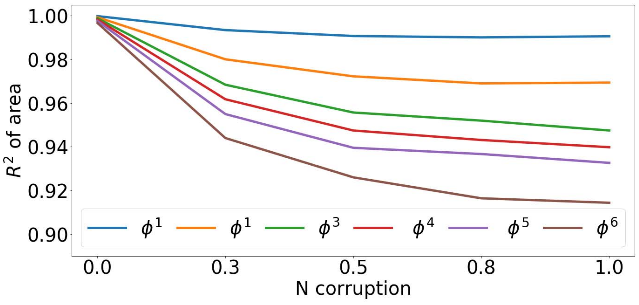


Figure 4: R^2 of the RCVs of area against label corruption at each of the 6 hidden layers of the MLP trained on MNIST.

5. True labels are learned before random labels

We monitor the convergence of a 0.5-memorizing InceptionV3, separating the performance on the true and the corrupted labels. We use a single training set up rather than different settings as in [5,29]. We find this approach more representative of a real-case scenario where unintended memorization may happen on a fraction of the original dataset. In Fig. 5, we show that the network learns more easily the true data distribution than the corrupted one. As we expected, the underlying distribution of natural images is easier to fit during training than randomly labeled images. Our results align with the work in [4], which models the training loss as a bimodal distribution[4]. Similar results were obtained on the CIFAR10 dataset [17].

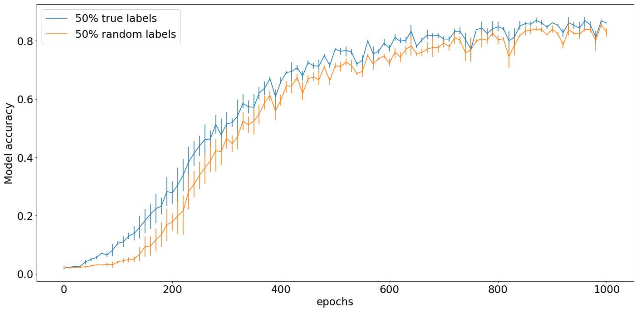


Figure 5: Accuracy on the true and random labels of InceptionV3 (0.5-memorizing) on DTD.

6. Conclusion

In this paper, we analyzed the differences in the representations learned by flawed and generalizing models. The analysis of the activations of intermediate layers with linear probes (classifiers, CAVs or RCVs) adds a new viewpoint to previous works [5,21,29,30] by interpreting model flaws with human-friendly concepts. Simple concepts are learned already at early layers, even in fully memorizing networks. Monitoring the learning curves on portions of data, rather than on the entire dataset, highlighted the slower convergence of memorization, particularly at early epochs. We believe that these observations can help to notice the memorization of incorrectly-labeled samples or outliers. This is particularly relevant in medical applications, where imprecise labels may affect the learning of the true underlying distribution of the data.

[4] [4] was not yet published at the time of the experiments.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9524–9535, 2018.

[2] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *ICLR Workshop*, 2016.

[3] V. Andrearczyk and P. F. Whelan. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69, 2016.

[4] E. Arazo, D. Ortego, P. Albert, N. OConnor, and K. Mcguinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321, 2019.

[5] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242, 2017.

[6] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

[8] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, et al. Interpretability of deep learning models: a survey of results. In *IEEE Smart World Congress 2017 Workshop: DAIS*, 2017.

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large–scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3):50–57, 2017.

[12] M. Graziani, V. Andrearczyk, S. Marchand-Maillet, and H. Müller. Concept attribution with regression concept vectors. *(submitted) IEEE transactions on Multimedia*, 2020.

[13] M. Graziani, V. Andrearczyk, and H. Muller. Regression concept vectors for bidirectional explanations in histopathology. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops*, 2018.

[14] M. Graziani, J. Brown, V. Andrearczyck, V. Yildiz, J. P. Campbell, D. Erdogmus, S. Ioannidis, M. F. Chiang, J. Kalpathy-Kramer, and H. Muller. Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. *Medical Imaging 2019: Computer-Aided Diagnosis*, 2019.

[15] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. International Society for Optics and Photonics, 2003.

[16] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

[17] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[18] D. Krueger, N. Ballas, S. Jastrzebski, D. Arpit, M. S. Kanwal, T. Maharaj, E. Bengio, A. Fischer, and A. Courville. Deep nets don't learn via memorization. *ICLR Workshop*, 2017.

[19] Y. LeCun and C. Cortes. The MNIST database of handwritten digits. 1998.

[20] Z. C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, Sept. 2018.

[21] A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pages 5727–5736, 2018.

[22] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick. On the importance of single directions for generalization. *ICLR*, 2018.

[23] A. Radhakrishnan, M. Belkin, and C. Uhler. Memorization in overparameterized autoencoders. *ICML 2019 Workshop on Deep Phenomena*, 2019.

[24] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. In *International Conference on Machine Learning*, pages 1089–1096. Omnipress, 2011.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[26] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pages 1–5, 2015.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[28] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia. Iterative learning with open-set noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8688–8696, 2018.

[29] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

[30] C. Zhang, S. Bengio, M. Hardt, and Y. Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *ICML 2019 Workshop on Deep Phenomena*, 2019.

[31] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.