Beware of the beginnings: intermediate and higher-level representations in deep neural networks are strongly affected by weight initialization

Johannes Mehrer (johannes.mehrer@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge
15 Chaucer Road, CB2 7EF Cambridge, UK

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Zuckerman Mind Brain Behavior Institute, Columbia University
3227 Broadway, L3-064 New York City, NY, USA

Tim C. Kietzmann (tim.kietzmann@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge
15 Chaucer Road, CB2 7EF Cambridge, UK

Abstract:

Deep neural networks (DNNs) excel at complex visual recognition tasks and have successfully been used as models of visual processing in the primate brain. Because network training is computationally expensive, many computational neuroscientists rely on pre-trained networks. Yet, it is unclear in how far the obtained results will generalize, as different weight initializations might shape the learned features (despite reaching similar testing performance). Here we estimate the effects of such initialization while keeping the network architecture and training sequence identical. To investigate the learned representations, we use representational similarity analysis (RSA), a technique borrowed from neuroscience. RSA characterizes a network's internal representations by estimating all pairwise distances across a large set of input conditions – an approach that is invariant to rotations of the underlying high-dimensional activation space. Our results indicate that differently initialized DNNs trained on the same task converged on indistinguishable performance levels, but substantially differed in their intermediate and higher-level representations. This poses a potential problem for comparing representations across networks and neural data. As a path forward, we show that biologically motivated constraints, such as Gaussian noise and rate-limited tanh activation functions can substantially improve the reliability of learned representations.

Keywords: RSA, consistency of representations

Introduction

To date DNNs are the best model class for predicting activity in multiple regions of the primate visual cortex. Network internal representations result from training on millions of images, and are shaped by network architecture, input statistics, learning algorithm, and objective function (Kietzmann, Mcclure, & Kriegeskorte, 2017). Apart from these main driving forces, the initial assignment of random weights may affect network internal features, despite oftentimes having a negligible effect on test performance. Such random effects could potentially raise problems for comparisons of internal representations between different DNNs, or, as in computational neuroscience, between representations in artificial and neural networks. The overall question therefore is, how consistent network internal features are across different weight initializations, and whether specific training parameters exist that alleviate potential problems. The latter include, among others, the type of activation function (here: ReLU vs. tanh, but for biologically more plausible examples, see Bhumbra, 2018), as well as the level and type of dropout (Gaussian or Bernoulli) during training and test. Moreover, it may matter where the noise is applied: to the activations ("drop-out"), to the weights ("drop-connect"), or to both (e.g. "Spike-and-Slab Dropout"; (McClure & Kriegeskorte, 2016; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). To start a systematic investigation, we here first estimate the overall magnitude of the effect. We then test for factors that may influence the consistency of learned representations. The size of the effect is calculated by training multiple identical networks with different weight initialization. The effect size is then compared to the effects of different input statistics in terms of image-set and category-selection. Finally, we investigate in how far limits imposed on the activation levels and activation noise can constrain training outcomes to lead to more consistent representations.

Methods

Representational similarity analysis

To compare different network instances, we here use Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008), a widely used neuroscience method to compare representations within and across measurement modalities. RSA is based around the concept of representational dissimilarity matrices (RDMs), which, across a large set of input stimuli, store all pairwise distances between stimulus-driven pattern activations. The resulting matrix characterizes the representational space spanned by the network solution, as it depicts the geometric relations of all different input stimuli w.r.t. each other. By relying on distances, RSA is invariant to rotations in the input space. It therefore directly offers itself to comparisons across deep neural network instances. As a distance measure, we use correlation distances of the layer activations.

Experimental design, and DNN architecture and training

Our experiments are based on the idea of representational consistency. Given two network instances, we first compute network RDMs for each layer using a large set of 1000 images, and subsequently test how similar the corresponding RDMs are. Given the properties of RDMs, if two networks span the same space, but one is a rotated (or scaled) version of the other, then the RDMs will be highly correlated, i.e. they will exhibit large representational consistency. In the first experiment, we investigate how the initial set of weights (random seed 1 vs. 2) can affect internal representations and how the effects compare to using different input statistics via different training sets for the same category structure or different output categories. We then test in how far the network consistency is affected when Gaussian noise is introduced to the activations during training and when using a range-limited activation function (tanh) rather than ReLU (experiment 2).

The architecture used throughout the paper is reminiscent of VGG-S (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014), but the fully-connected layers were replaced by convolutional layers to reduce the amount of trainable parameters by ~90%. In addition, we adapted the amount of maps (96, 128, 256, 512, 512, 1024, 1024) and the kernel sizes (7, 5, 3, 3, 3, 3, 3). Training was performed on CIFAR 10, which consists of 10 categories with 5.000 training, and 1000 test images each.

Experiment I We first investigated how RDM consistency is influenced by the initial set of weights, holding all other training aspects constant. To be able to judge the size of the effect, we relate it to the differences in consistency that result from training on different image sets while starting from identical weights. First, we tested for the effects of training on the same categories, albeit with different input images (using 10 categories, each with 2,500 of the 5000 training images). We refer to these subsets of CIFAR 10 as “CIFAR 10, set 1 and 2”. Following a 2x2 design (random seed 1 and 2 vs. “CIFAR 10, set 1 and 2”), four DNNs were trained for 250 epochs each, using a ReLU activation function. One step further, we asked how training on different image- and category sets influences RDM consistency. For this, we split the training set of CIFAR 10 into two sets of five categories each. Each category contained the full 5,000 training images, such that the overall amount of training images was identical to the experiments with different image-sets, but a different category structure was used. We refer to these subsets of CIFAR 10 as “CIFAR 5, set 1 and 2”. In a 2x2design (random seed 1 and 2 vs. “CIFAR 5, set 1 and 2”), we again trained four DNNs using ReLU as the activation function.

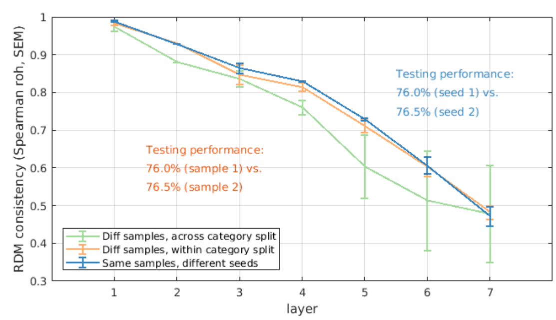
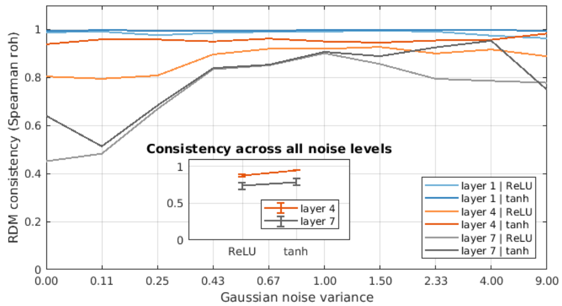
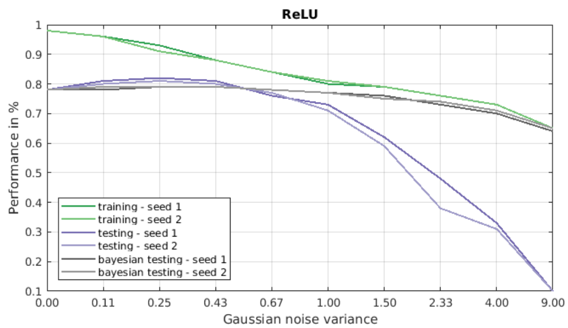


Figure 1. RDM consistency across different seeds decreases with layer depth (blue). The effect is comparable to training on a completely different set of images with the same (orange) or a different category structure (green).

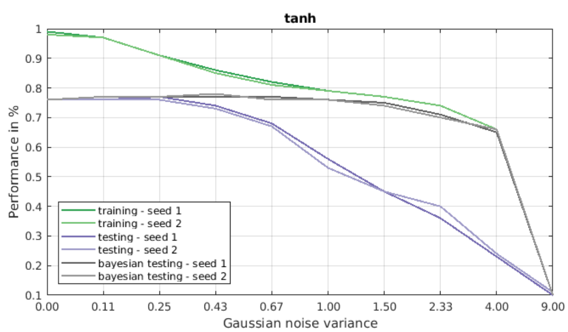
Experiment II Following a first characterization of the problem in Experiment 1, we explored in how far adding noise to the activations and a biologically more plausible type of activation function (rate-limited tanh instead of ReLU) might affect RDM consistency. CIFAR 10 was not split, but the entire set was used for training. In a 2x2x10 design (random seed 1 and 2 vs. ReLU and tanh vs. 10 levels of noise) 40 DNNs were trained for 250 epochs. Experiments with activation noise are based on multiplicative noise, which follows a Gaussian distribution centered on 1, with a variance σ^2 scaling between 0 and 9.



a)



b)



c)

Figure 2. Effects of activation-noise and –function on consistency and task performance **a)** The decrease in RDM consistency is negligible for low-level representations (layer 1, dark and light blue) and can be reduced to a minimum for intermediate and higher-level representations (layer 4, red and orange; layer 7, black and grey) **b)** Only high levels of noise (variance $\sigma^2 > 0.67$) affect testing performance for DNNs trained with ReLU. When using Dropout at test time, performance remains robust even at very high noise levels **c)** A similar pattern of results can be observed for a tanh activation function.

Results and Discussion

Experiment I. When varying only the initial set of weights, the RDM consistency between different DNN instances decreases substantially with layer depth (figure 1, blue) despite almost identical test performance (76 vs. 76.5% for seed 1 and 2, respectively). Surprisingly, this effect is qualitatively similar to when two DNNs with the same set of initial weights are trained on independent image sets originating from the same category structure (figure 1, orange). Even when training an entirely different category structure (figure 1, green) leads to only somewhat lower consistency.

These results demonstrate the strong effect of different random seeds on a DNN's intermediate and especially on higher-level representations, despite all other parameters being held constant. This finding has potential implications for comparing representations across DNNs or, as in computational neuroscience, to neural data, as observed differences could be solely due to weight initialization.

Experiment II – RDM consistency. We next explored biologically motivated constraints for their ability to yield more robust internal representations. We considered multiplicative Gaussian noise in the unit activations, and a rate-limited activation function (tanh instead of ReLU). Like before, we computed the RDM consistency across two random weight initializations. Figures show consistency estimates for three exemplary layers (early, middle and high-level layers) together with network performances on the test data.

Our results suggest that varying the noise level influences the consistency of both intermediate and higher-level representations. For ReLU-DNNs and CIFAR 10, noise with a variance σ^2 of 1.0 appears to yield maximal RDM consistency across intermediate (figure 2 a, layer 4, orange) and higher level (figure 2 a, layer 7, grey) representations; for tanh-DNNs noise with a variance σ^2 of 4.0 appears to yield maximal RDM consistency across intermediate (figure 2 a, layer 4, red) and higher level (figure 2 a, layer 7, black) representations. In contrast, lower-level RDM representations show overall high levels of consistency and are not strongly affected by varying the noise (figure 2 a, dark and light blue). Across noise levels, tanh results in more consistent representations, compared to ReLU (figure 2A inset).

Experiment II – task performance. Varying the level of activation-noise and -function not only affects RDM consistency, but also task performance. Due to increasingly strong regularization, training performance decreases with increasing noise level independent of whether ReLU or tanh was used (figure 2 b and c, "training", dark and light green). Test performance in ReLU-DNNs is relatively unaffected up to noise variance $\boldsymbol{\sigma}^2$ of 0.67. At higher levels, testing performance depends on whether dropout is applied at test time (figure 2 b, “Testing” (no dropout) vs. “Bayesian testing” (dropout), dark and light purple vs. black and grey). The same overall observations can be made for tanh-DNNs, where increasing the noise variance $\boldsymbol{\sigma}^2$ above 0.43 leads to decreased testing performance if dropout is not applied at test time (figure 2 c, “Testing” (no dropout) vs. “Bayesian testing” (dropout), dark and light purple vs. black and grey). Yet, dropout leads to robust test performance even at high noise levels.

In sum, these results suggest that the consistency of higher-level representations in DNNs across random weight initialization can be maximized by Gaussian activation noise and by using tanh as activation function. While Bayesian testing remains comparably stable, test performance without dropout may be considerably reduced when networks are optimized for consistency.

Conclusions

We use RSA, an analysis framework borrowed from neuroscience, to investigate the consistency of learned representations in DNNs. We find that random weight initialization most affected intermediate and higher-level representations. Surprisingly, the effect is qualitatively similar to training on different sets of images with the random seeds held constant. The addition of Gaussian activation noise during training, and a rate-limited activation function (tanh) resulted in increased, at times almost perfect consistency of intermediate and higher-level representations.

While these analyses and results are important for machine learning and computational neuroscience, they are derived from a relatively small dataset (CIFAR 10). The number of training instances (50.000 across 10 categories) is small compared to the amount of parameters of the DNN used here (~18 mio.). Thus, it remains to be established how representational consistency is affected when using larger datasets, such as Imagenet or ecoset (Mehrer, Kietzmann, & Kriegeskorte, 2017; Russakovsky et al., 2015). As an addition or even alternative to using dropout, it will be interesting to test for the effects of implicit regularization, as introduced via training with heavy data augmentation (Hernández-García & König, 2018). Finally, experiments with Bayesian testing, which was largely unaffected across noise levels, will provide important insights into network consistency under uncertainty.

Acknowledgments

We would like to thank our funders for supporting our work: Vice Chancellor’s Award from the Cambridge Trust (J.M.), DFG research fellowship (T.C.K).

References

Bhumbra, G. S. (2018). Deep learning improved by biological activation functions, 1–11.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets, 1–11.

Kietzmann, T. C., Mcclure, P., & Kriegeskorte, N. (2017). Deep Neural Networks in Computational Neuroscience.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4.

McClure, P., & Kriegeskorte, N. (2016). Representing inferential uncertainty in deep neural networks through sampling, (Mcmc), 1–14.

Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. (2017). Deep neural networks trained on ecologically relevant categories better explain human IT. *Cognitive Computational Neuroscience Meeting*, *1*, 1–2.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.