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 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 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 (“dropout”), to the weights (“dropconnect”), or 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.

Keywords: RSA, consistency of representations

Introduction
Methods

Representational similarity analysis

Experimental design, and DNN architecture and training
Results and Discussion

Experiment I. 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). 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. A similar pattern of results can be observed for a tanh activation function.

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 $\sigma^2$ of 1.0 appears to yield maximal RDM consistency across intermediate (figure 2a, layer 4, orange) and higher level (figure 2a, layer 7, grey) representations. In contrast, lower-level RDM representations show overall high levels of consistency and are not strongly affected by varying the noise (figure 2a, 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 2b and c, “training”, dark and light green).
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 $\sigma^2$ 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).