Neural Representations Reveal Distinct Modes of Class Fitting in Residual Convolutional Networks

Michał Jamroż, Marcin Kurdziel

AGH University of Science and Technology
Kraków, Poland

ML in PL 2023

Motivation

Our goal in this work is to characterize representations of classes in neural networks.

- Neural representations are inherently stochastic - a representation of some network input $\mathbf{x}$ can be seen as an outcome of sampling $\mathbf{x}$ from the data distribution.
- A reasonable notion of a *class representation* should capture the outcome of this sampling.

We therefore propose to leverage class-conditional distributions of inputs' representations as proxies to the neural representations of classes.

Motivation

Our goal in this work is to characterize representations of classes in neural networks.

- Concretely, we fit tractable class-conditional density models to sets of neural representations.
- We then use these models to characterize distributions of representations in classes.

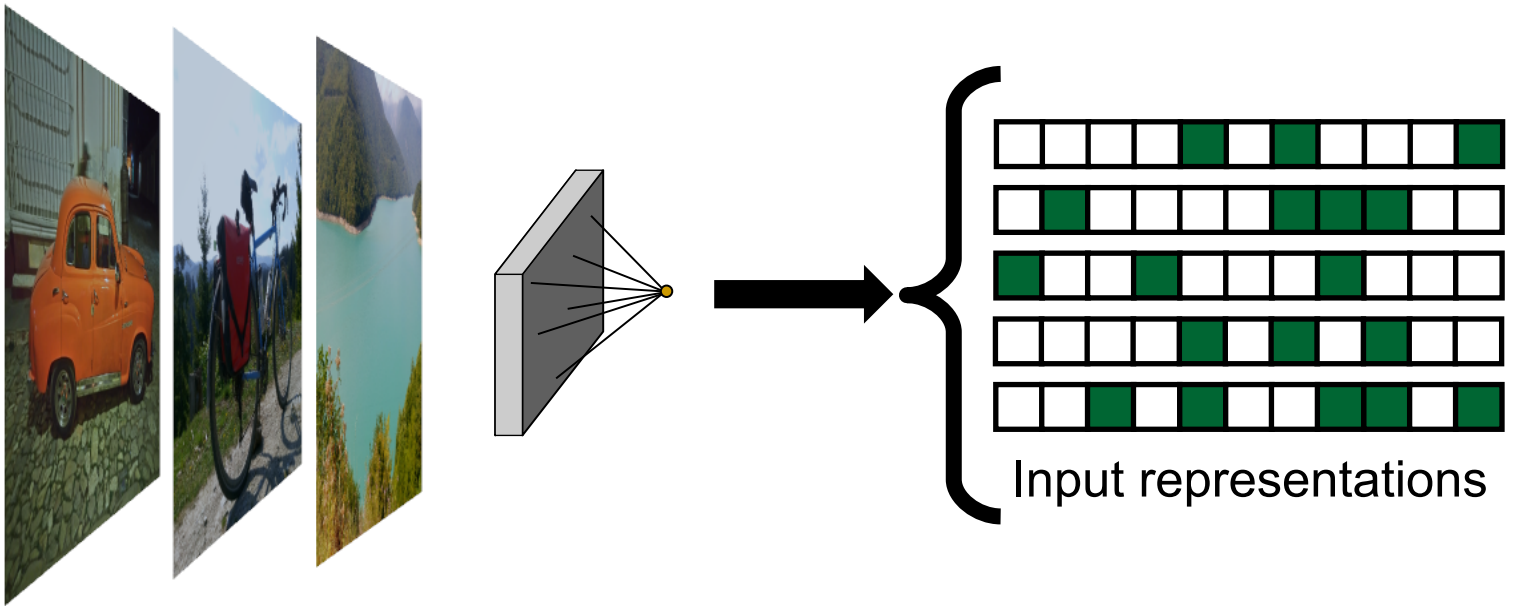
Surprisingly, our density models uncover distinct modes of class fitting in residual convolutional networks.

- This distinct modes of class fitting translate to marked differences in memorization of input examples and robustness to adversarial attacks.

Neural representations

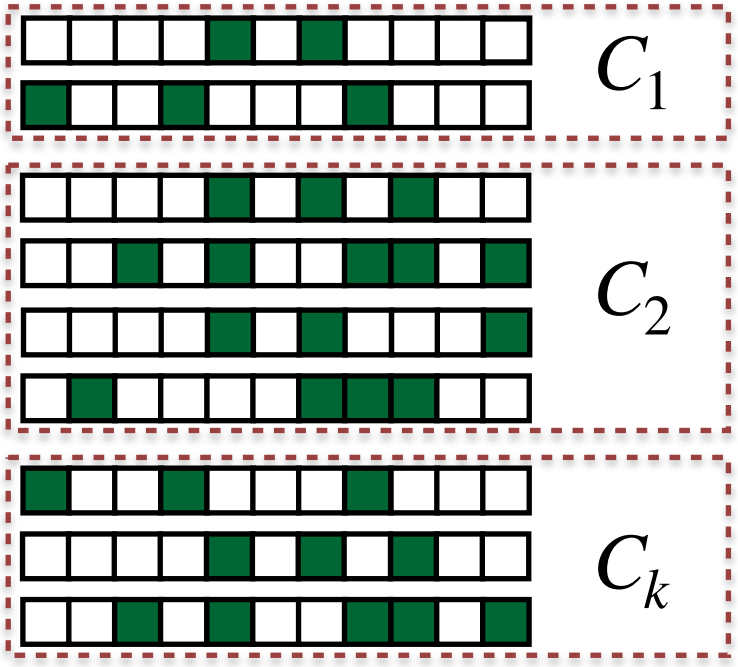
Inputs to classification heads are often constructed by pooling activations across spatial dimensions.

- We use a similar construction for neural representations of network inputs.



Neural representations

To construct class-conditional density models, we estimate an independent probability density model for neural representations in each class.



Density model for neural representations

Our density model for neural representations is a mixture of gaussian components with a Dirichlet process prior (DP-GMM).

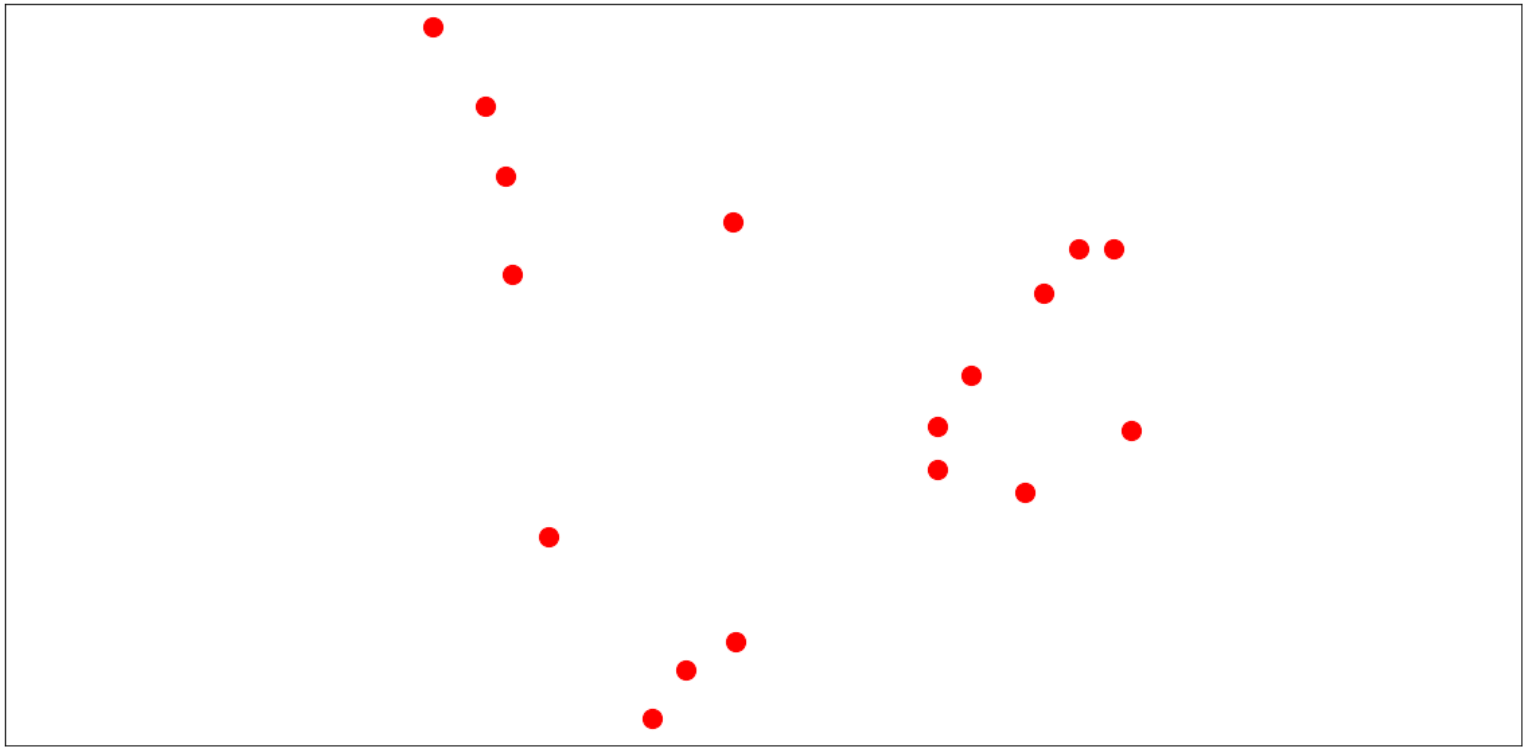
- This is a nonparametric Bayesian model, in which the number of components adapts to the explained data.

$$\begin{aligned} \alpha &\sim Gamma(1,1), \\ G \mid \alpha &\sim DP(NIW(\boldsymbol{\theta}_0), \alpha), \\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim G, \\ \boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned}$$

- In our previous work [1] we used this model characterize distributions of neural representations in networks that memorize inputs and networks that can exploit patterns in data.

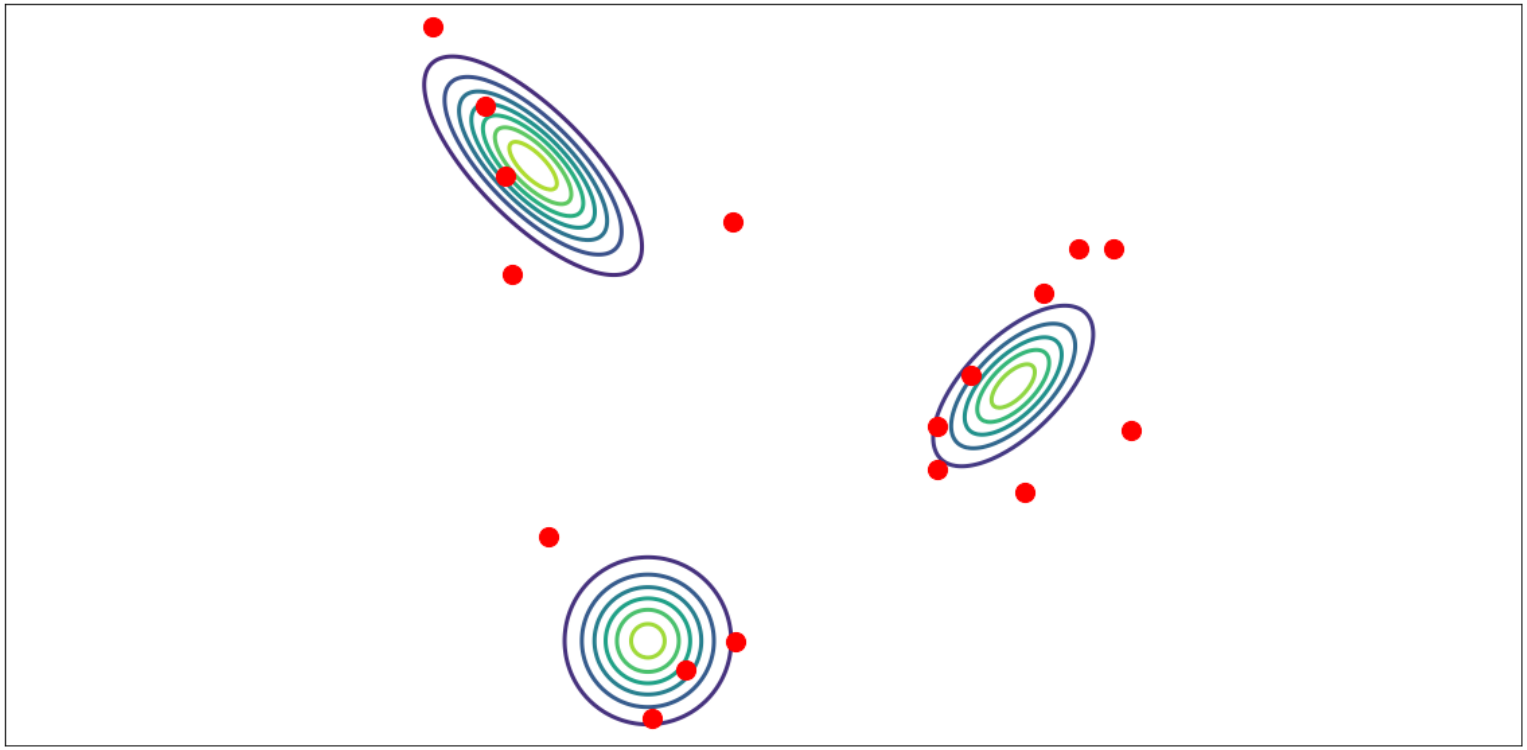
Density model for neural representations

Our density model for neural representations is a mixture of gaussian components with a Dirichlet process prior (DP-GMM).



Density model for neural representations

Our density model for neural representations is a mixture of gaussian components with a Dirichlet process prior (DP-GMM).



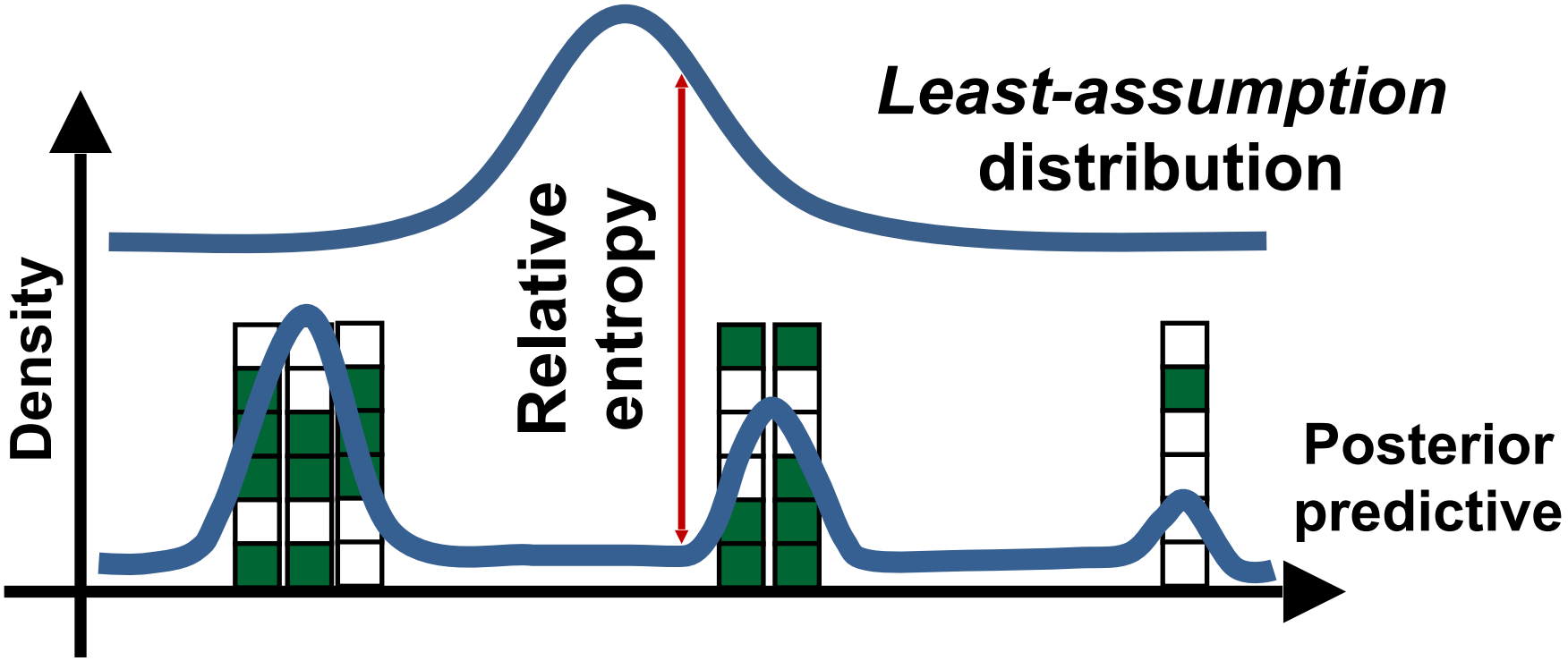
Complexity measure for neural representations

Importantly, under DP-GMM we can estimate the KL divergence (or relative entropy) between posterior predictive distribution and another distribution with a tractable density.

- This can be used as a complexity measure for the posterior predictive density.

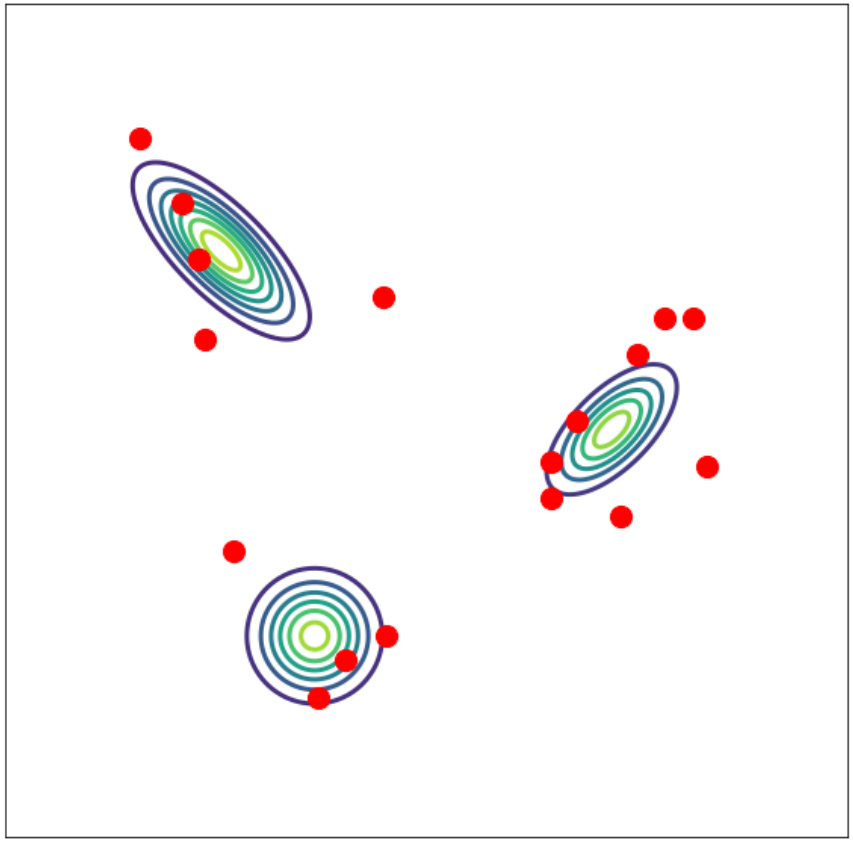
Complexity measure for neural representations

Importantly, under DP-GMM we can estimate the KL divergence (or relative entropy) between posterior predictive distribution and another distribution with a tractable density.



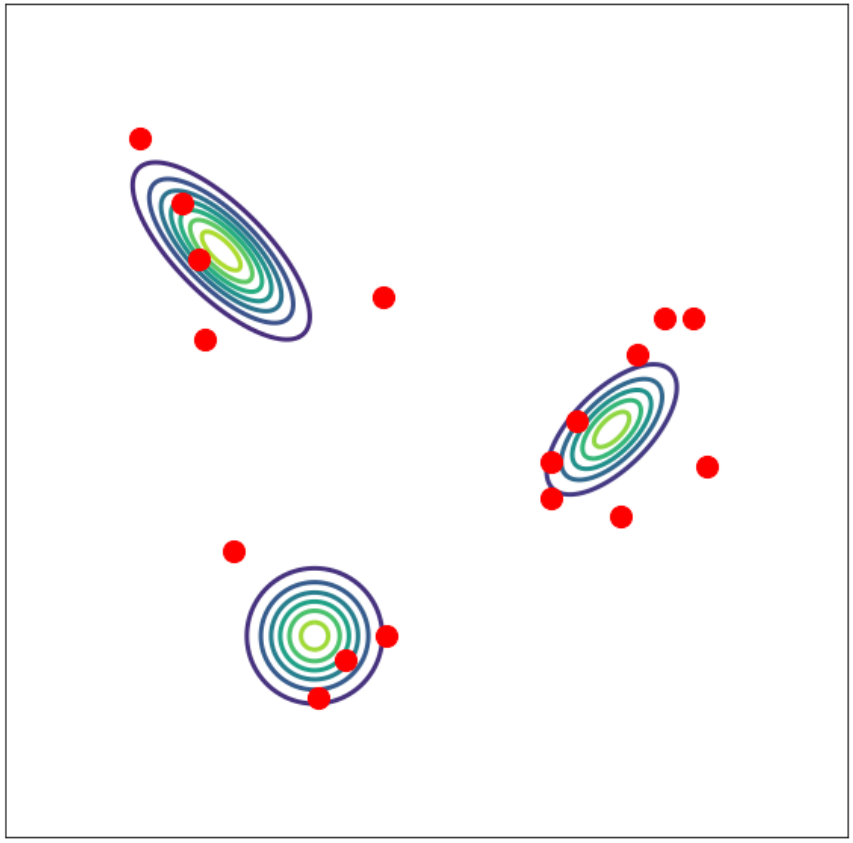
Complexity measure for neural representations

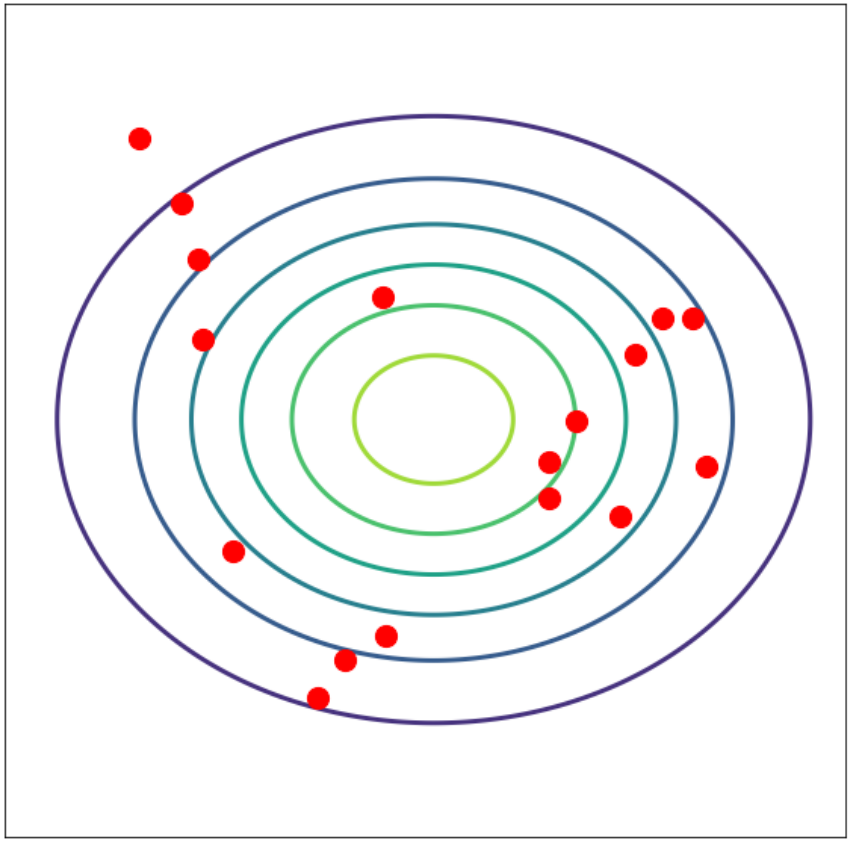
As a *least-assumption* distribution we adopt a maximum entropy distribution that explains just the location and scale of the data.



Complexity measure for neural representations

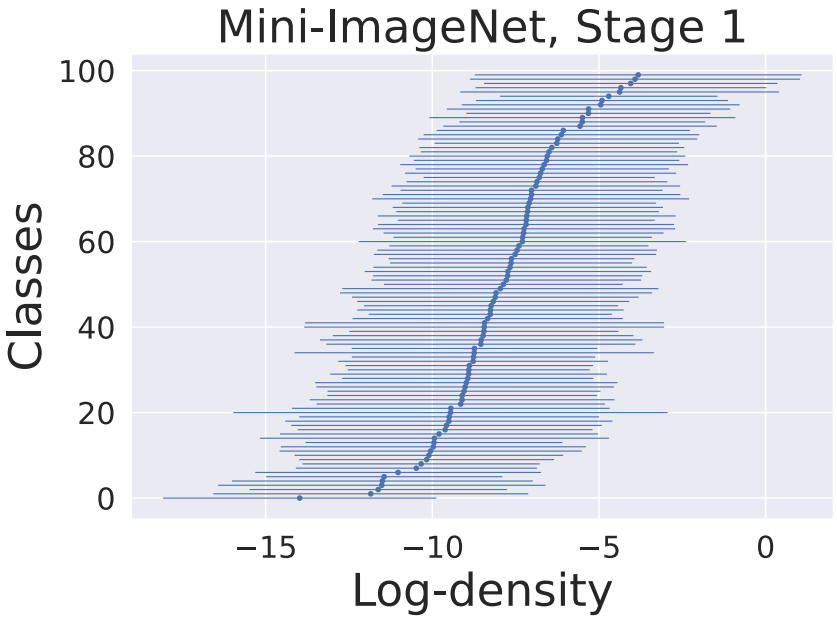
As a *least-assumption* distribution we adopt a maximum entropy distribution that explains just the location and scale of the data.

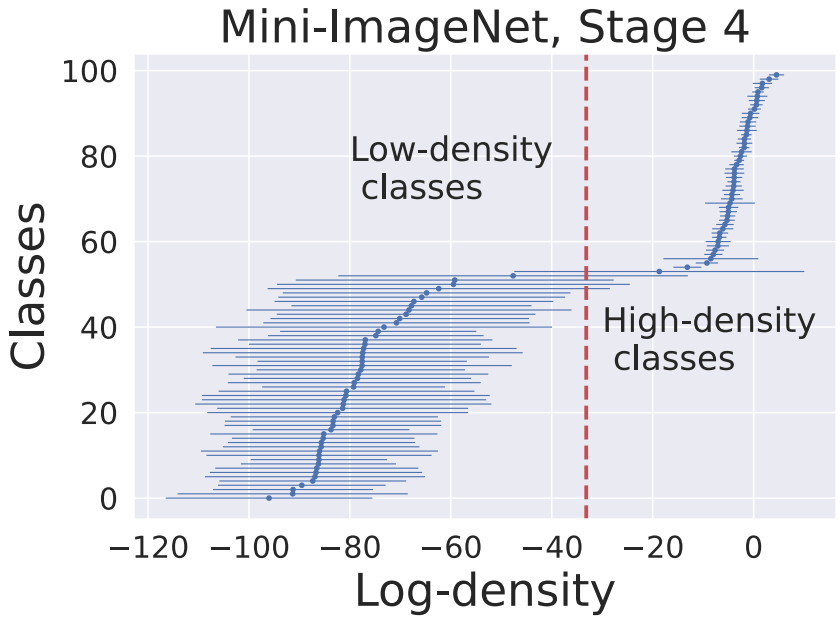




How residual ConvNets fit classes?

We begin characterization of class representations by comparing log-densities estimated for inputs from each class.

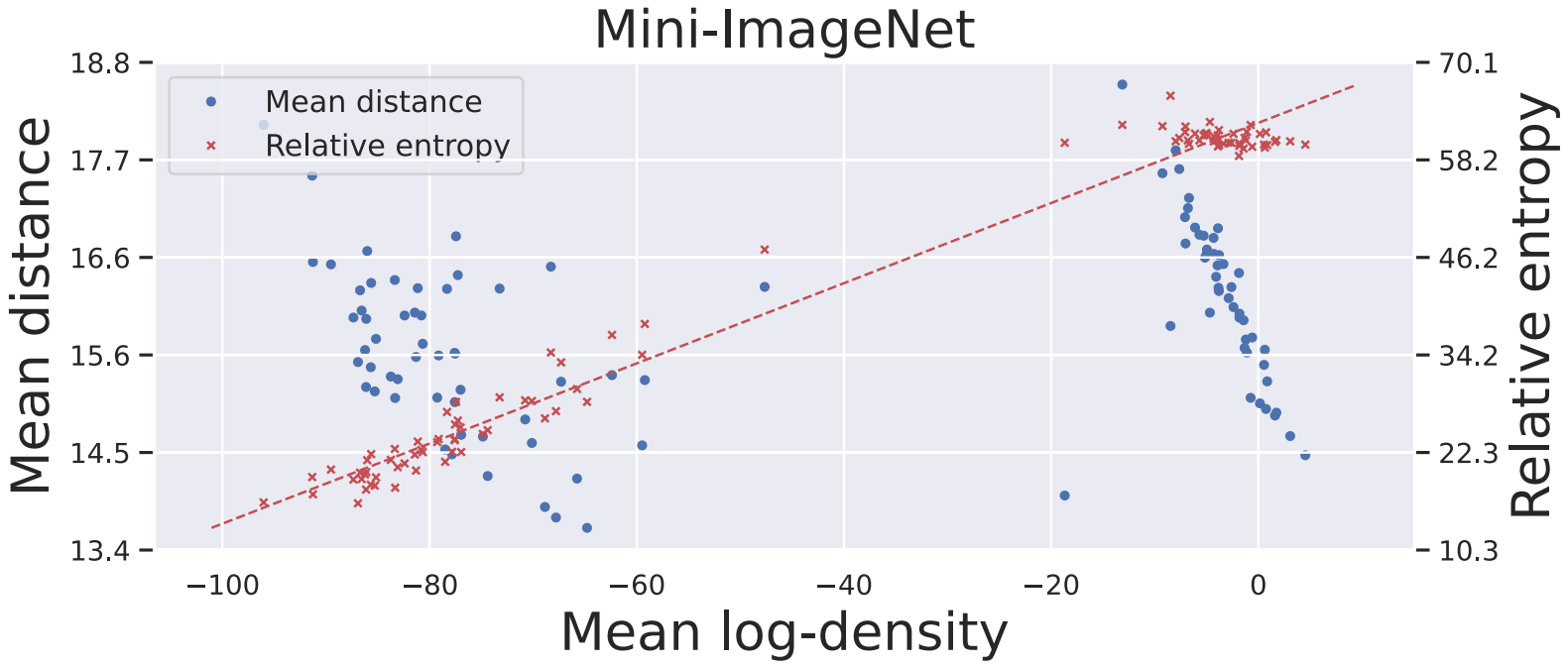




How residual ConvNets fit classes?

One obvious explanation for this structure could be that input representations in the high-density classes are simply more similar to each other.

- Our results shows that this simple explanation is incorrect.



How residual ConvNets fit classes?

One obvious explanation for this structure could be that input representations in the high-density classes are simply more similar to each other.

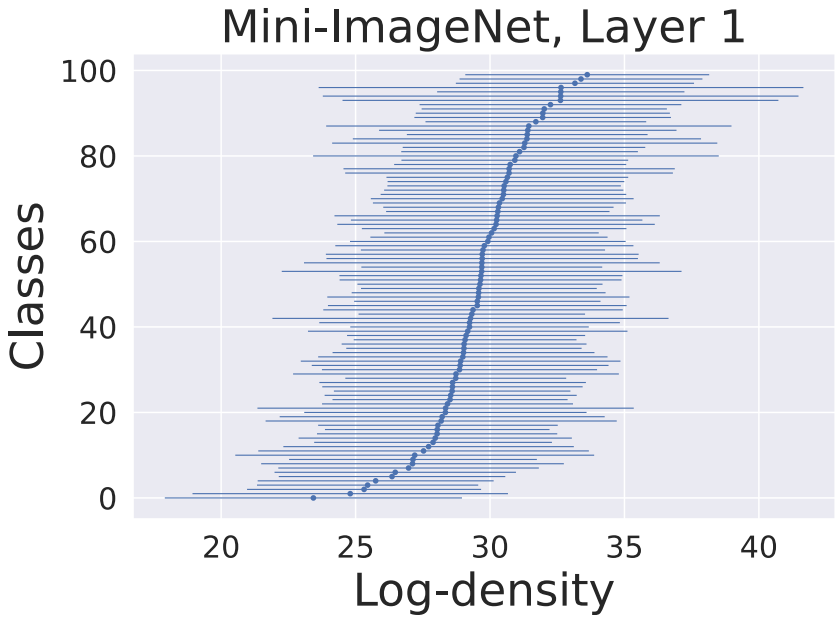
- Our results shows that this simple explanation is incorrect.
- High-density cases are not more spatially compact at the neural representation level than the low-density classes.
- However, posterior predictive distributions for these classes are vastly more complex than posterior predictive distributions for the low-density classes.
- In other words, these classes have vastly more non-gaussian distributions of neural representations.

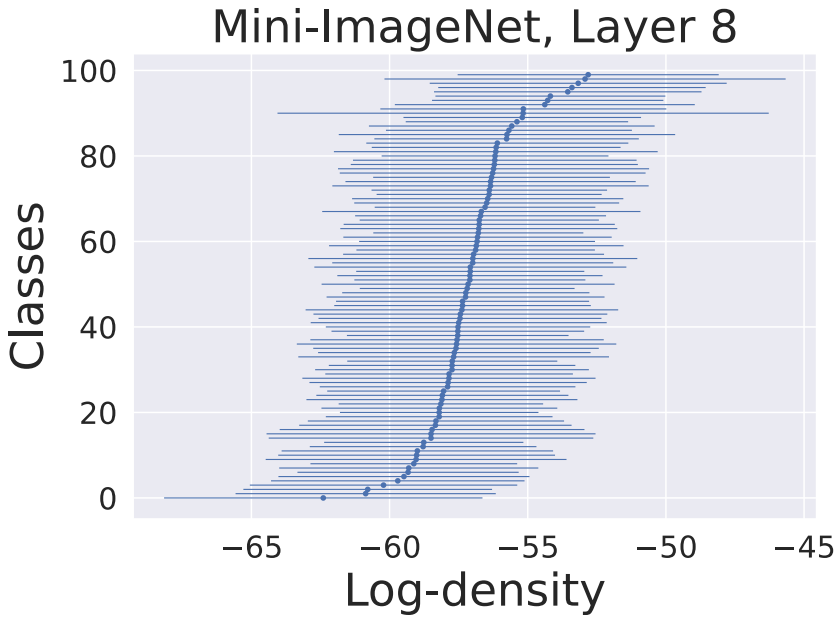
Together, these findings suggests that at the neural representation level high-density classes are formed by compact but spatially separated components.

How residual ConvNets fit classes?

Can this structure be simply a product of the datasets commonly used to train neural nets?

- We do not observe distinct modes of class fitting in plain ConvNets. The architecture therefore plays a role in the process that we observe!





Class representations correlate with memorization

Feldman et al. [2] showed that in a long-tailed data distribution minimization of the generalization errors requires memorizing some of the input examples.

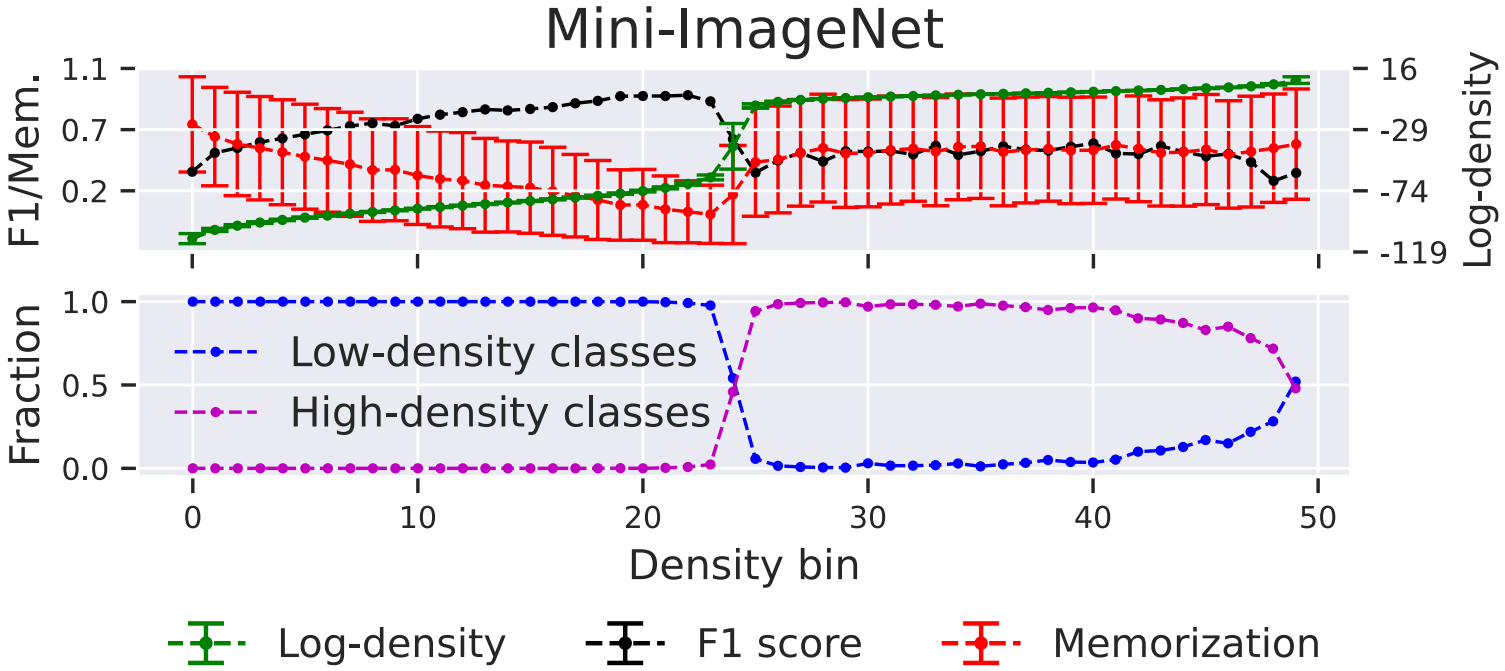
Next, Feldman and Zhang [3] proposed a tractable proxy to the measure of input memorization.

They demonstrated that contemporary neural nets memorize training data to a non-trivial degree.

These results suggests that the structure we observe in class representations may correlate with memorization of input examples.

Class representations correlate with memorization

And indeed, we observe that the transition from the low-density to the high-density regime correlate with a marked increase in the degree of input memorization.



Class representations correlate with adversarial robustness

Compact and spatially separated components in the high-density classes should—intuitively—be less robust to an adversarial attack.

- In particular, a relatively small input perturbation may move the representation of the attacked example outside of its component.

One way to verify this hypothesis could be to compare the high- and the low-density classes against a selection of adversarial attack.

However, given the large number of attacks proposed so far, we instead opt for an attack-agnostic comparison.

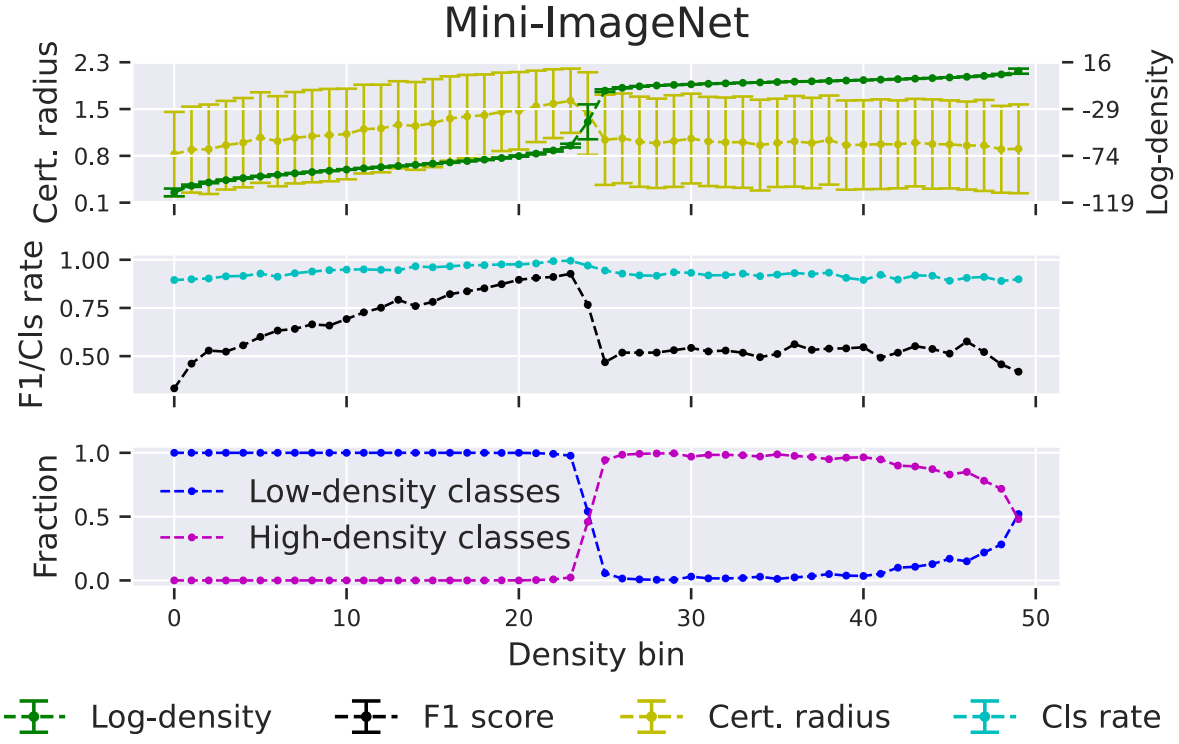
Class representations correlate with adversarial robustness

To this end, we compare the low- and the high-density classes w.r.t the performance of a classifier certifiably robust to l_2 perturbations.

- We use the randomized smoothing-based certification procedure proposed by Cohen et al. [4].
- We measure the robustness to an adversarial attack with F_1 score of the smoothed classifier, the estimated certified radius and the fraction of inputs for which certified classifier did not abstain from prediction.

Class representations correlate with adversarial robustness

We observe that the transition from the low- to the high-density regime correlate with a marked decrease in adversarial robustness.



Conclusions

We characterized distributions of representations in classes learned by residual convolutional networks.

- Surprisingly, we found that ResNets do not fit classes in an uniform way.
- Previous observations showed that as training progresses the intra-class variance of neural representations becomes small relative to inter-class variance — a so called *neural collapse*.
- Our results demonstrate that despite this increasing class separation during the final stages of training, classes in residual networks still retain non-trivial internal structure.

Conclusions

We observe two distinct modes of class fitting that differ in mean and variance of the log-densities estimated for class members, namely, high- and low-density classes.

- This observation is not explained by the input data — a similar structure is missing in plain ConvNets, indicating that the network architecture plays a role in the observed process.
- We demonstrate that the high-density classes correlate with increased memorization of input examples.
- We also demonstrate that the high-density classes are less robust to an adversarial attack.

Conclusions

See extended versions of our paper for additional results, including preliminary results for MLP-Mixer and Vision Transformer architectures.

https://arxiv.org/abs/2212.00771

References

1. Jamroż, M., Kurdziel, M., & Opala, M. (2020). A Bayesian nonparametrics view into deep representations. *Advances in Neural Information Processing Systems*, *33*, 1440-1450
2. Feldman, V. (2020). Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing,* 954-959
3. Feldman, V., & Zhang, C. (2020). What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, *33*, 2881-2891
4. Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning,* pp. 1310-1320