INTRACLASS CLUSTERING: AN IMPLICIT LEARNING ABILITY THAT REGULARIZES DNNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Several works have shown that the regularization mechanisms underlying deep neural networks' generalization performances are still poorly understood (Neyshabur et al., 2015; Zhang et al., 2017). In this paper, we hypothesize that deep neural networks are regularized through their ability to extract meaningful clusters among the samples of a class. Remarkably, this implies the emergence of a learning ability devoid of any supervision or explicit training mechanisms, hence constituting *an implicit form of regularization*. To support this hypothesis, we design four different measures of intraclass clustering, based on the neuron- and layer-level representations of the training data. We then show that these measures constitute accurate predictors of generalization performance across variations of a large set of hyperparameters (learning rate, batch size, optimizer, weight decay, dropout rate, data augmentation, network depth and width).

1 INTRODUCTION

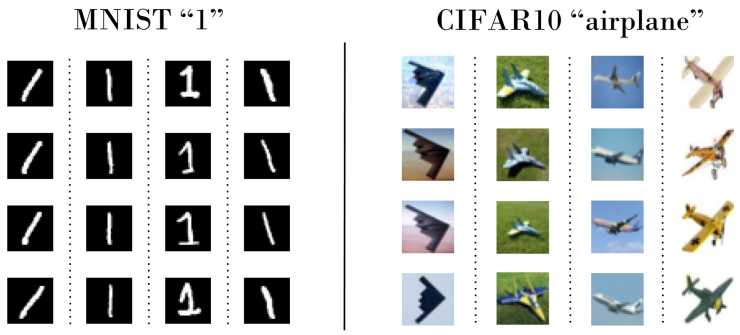

Figure 1: In standard image classification datasets, classes are typically composed of multiple clusters of similarly looking images. We call intraclass clustering the ability to differentiate such clusters despite their association to identical labels.

The generalization ability of deep neural networks remains largely unexplained. In particular, the traditional view that explicit forms of regularization (e.g. dropout, $\mathcal{L}_2$-regularization, data augmentation) could account for the generalization performance of state of the art neural networks has been experimentally invalidated (Neyshabur et al., 2015; Zhang et al., 2017). Today's conventional wisdom rather conjectures the presence of implicit forms of regularization, emerging from the interactions between neural network architectures, optimization and the inherent structure of the data itself (Arpit et al., 2017).

One structural component that seems to occur in most image classification datasets is the presence of multiple clusters amongst the samples of a class (or *intraclass clusters*, cfr. Figure 1). The extraction of such structure in the context of supervised learning is not self-evident, as today's standard training algorithms are designed to group samples from a class together, without any considerations for eventual intraclass clusters.

This paper hypothesizes that the identification of intraclass clusters emerges during standard supervised training of deep neural networks, notwithstanding the absence of supervision or explicit training mechanisms targeting this behaviour. Moreover, we suggest that this phenomenon improves the generalization ability of deep neural networks, hence constituting an implicit form of regularization.

We verify our hypotheses experimentally by evaluating the predictive power of four different measures of intraclass clustering with respect to a network's generalization performance. These measures are designed to capture intraclass clustering from four different perspectives, defined by the representation level (neuron vs. layer) and the amount of knowledge about the data's inherent structure (datasets with or without hierarchical labels). To evaluate these measures' predictive power, we train more than 500 models on CIFAR100 (whose 100 classes are hierarchically organized into 20 superclasses) and CIFAR10 datasets, varying standard hyperparameters in a principled way in order to generate a wide range of generalization performances. The measures are then evaluated qualitatively through visual inspection of their relationship with generalization and quantitatively through the granulated Kendall rank-correlation coefficient introduced by Jiang et al. (2020). Both evaluations reveal a tight connection between intraclass clustering measures and generalization ability, providing important evidence to support this work's hypotheses.

2 Related work

This work follows the line of inquiry initiated by Neyshabur et al. (2015); Zhang et al. (2017), and tries to unveil the phenomena underlying the generalization ability of deep neural networks. We follow an empirical approach based on the design of measures that correlate with generalization performance. This approach has been extensively used in recent years, leading to measures such as sharpness of minima (Keskar et al., 2017), sensitivity (Novak et al., 2018), reliance on single directions (Morcos et al., 2018) and stiffness (Fort et al., 2019). Jiang et al. (2020) recently introduced a principled way to evaluate how a measure correlates with generalization, which we reproduce in our experimental setup.

A measure related to our work is the selectivity of individual neurons for classes studied in Morcos et al. (2018)[1]. Since they do not observe a clear relationship between class selectivity and generalization, the authors suggest that individually selective neurons may not be necessary for strong network performance. One of the measures we study measures the neurons' selectivity for intraclass clusters instead of entire classes. This critical difference leads us to the opposite conclusion as the one drawn in Morcos et al. (2018): we observe empirically that models composed of individually selective neurons tend to generalize better.

A large body of work has studied the emergence of features in the internal representations of deep nets. Huh et al. (2016) shows that representations that discriminate classes naturally emerge when training on their corresponding superclasses. Several works have demonstrated that object- or feature-selective representations emerge in supervised image classification settings (Zeiler & Fergus, 2014; Simonyan et al., 2014; Yosinski et al., 2015; Zhou et al., 2015). Bau et al. (2017) has proposed a measure to quantify this phenomenon and studied its correlation with different training initializations, network depth and width, dropout and batch normalization. Our work can be interpreted as providing appropriate measures of this phenomenon that highlight its relationship with generalization.

3 Measuring intraclass clustering in internal representations

A challenge of our work resides in measuring a model's ability to differentiate intraclass clusters, without knowing which mechanisms underlie such ability. In such context, the design of multiple measures offers different perspectives that will help better characterize the phenomenon. Moreover, having multiple complementary measures reduces the risk that their potential correlation with generalization is induced by other phenomena independent of intraclass clustering. This section thus describes four complementary measures of intraclass clustering, differing in terms of representation

[1]The selectivity of a neuron for a set of samples here refers to the extent by which this set can be discriminated from the rest of the dataset based on the neuron's (pre-)activation values.

level (neuron vs. layer) and the amount of knowledge about the data's inherent structure (datasets with or without hierarchical labels).

3.1 TERMINOLOGY AND NOTATIONS

In the following sections we use D to denote the training dataset. Let I be the amount of classes in the dataset D, we denote the set of samples from class i by C_i with $i \in \mathcal{I} = \{1, 2, ..., I\}$. In the case of hierarchical labels, C_i denotes the samples from subclass i and $S_{s(i)}$ the samples from the superclass containing subclass i.

We denote by $\mathcal{N} = \{1, 2, ..., N\}$ and $\mathcal{L} = \{1, 2, ..., L\}$ the indexes of the N neurons and L layers of a network respectively. Neurons are considered across all the layers of a network, not a specific layer. The methodology by which indexes are assigned to neurons or layers does not matter in our context.

We further denote by $\text{mean}_{j \in \mathcal{J}}$, $\text{median}_{j \in \mathcal{J}}$ and $\max_{j \in \mathcal{J}}$ the mean, median and max operations over the index j respectively. Moreover, $\text{mean}^k_{j \in \mathcal{J}}$ corresponds to the mean of the top-k highest values, over the index j.

We call pre-activations the values preceding the application of the ReLU activation function (Nair & Hinton, 2010), and activations the values following it. In convolutional layers, the spatial dimensions of (pre-)activations are reduced through a global max pooling operation before applying our measures.

3.2 MEASURES BASED ON LABEL HIERARCHIES

The first class of measures takes advantage of datasets that include a hierarchy of labels. For example, CIFAR100 is organized into 20 superclasses (e.g. flowers) each comprising 5 subclasses (e.g. orchids, poppies, roses, sunflowers, tulips). We hypothesize that these hierarchical labels reflect an inherent structure of the data. In particular, we expect the subclasses to approximately correspond to different clusters amongst the samples of a superclass. Hence, measuring the extent by which a network differentiates subclasses when being trained on superclasses should reflect its ability to extract intraclass clusters during training.

3.2.1 NEURON-LEVEL SUBCLASS SELECTIVITY

The first measure quantifies how selective individual neurons are for a given subclass C_i with respect to the other samples of the associated superclass $S_{s(i)}$. Here, strong selectivity means that the subclass C_i can be reliably discriminated from the other samples of $S_{s(i)}$ based on the neuron's pre-activations. Let $\mu_{n,E}$ and $\sigma_{n,E}$ be the mean and standard deviation of a neuron n's pre-activation values taken over the samples of set E. The measure is defined as follows:

$$c_1 = \text{median}_{i \in \mathcal{I}} \max_{n \in \mathcal{N}} \frac{\mu_{n,C_i} - \mu_{n,S_{s(i)} \setminus C_i}}{\sigma_{n,C_i} + \sigma_{n,S_{s(i)} \setminus C_i}} \tag{1}$$

We aggregate over all the neurons of a network through a max operation, such that the neuron which is the most selective for subclass C_i is used as an indicator of the overall network's ability to differentiate this subclass C_i.

3.2.2 LAYER-LEVEL SILHOUETTE SCORE

The second measure quantifies to which extent the samples of a subclass are close together relative to the other samples from the associated superclass *in the space induced by a layer's activations*. In other words, we measure to what degree different subclasses can be associated to different clusters in the layer's representation. We quantify this by computing the pairwise cosine distances on the samples of a superclass and applying the Silhouette score (Kaufman & Rousseeuw, 2009) to assess the clustered structure of the subclasses. Let *silhouette*(a_l, E) be the mean silhouette score of the set of samples E based on layer l's activations, the measure is then defined as:

$$c_2 = \text{median}_{i \in \mathcal{I}} \max_{l \in \mathcal{L}} \textit{silhouette}(a_l, C_i) \tag{2}$$

3.3 MEASURES BASED ON VARIANCE

To establish the generality of our results, we also design two measures that can be applied in absence of hierarchical labels. We hypothesize that the discrimination of intraclass clusters should be reflected by a high variance in the representations associated to a class. If all the samples of a class are mapped to close-by points in the neuron- or layer-level representations, it is likely that the neuron/layer did not identify intraclass clusters.

3.3.1 VARIANCE IN THE NEURON-LEVEL REPRESENTATIONS OF THE DATA

The first variance measure is based on standard deviations of a neuron's pre-activations. If the standard deviation computed over the samples of a class is high compared to the standard deviation computed over the entire dataset, we infer that the neuron has learned features that differentiate samples belonging to this class. More formally, let $\sigma_{n,E}$ be the standard deviation of a neuron n's pre-activation values taken over the samples of set E. The measure is defined as:

$$c_3 = \text{mean}_{i \in \mathcal{I}} \, \text{mean}^k_{n \in \mathcal{N}} \, \frac{\sigma_{n,C_i}}{\sigma_{n,D}} \tag{3}$$

We use $k = 30$ in the experiments of this paper, but observed that any k in the range $[1, 100]$ leads to similar conclusions. A visual representation of the measure is provided in Figure 2.

3.3.2 VARIANCE IN THE LAYER-LEVEL REPRESENTATIONS OF THE DATA

The fourth measure transfers the neuron-level variance approach to layers by computing the standard deviations on the pairwise cosine distances between samples of a class or the entire dataset, calculated in the space induced by the layer's activations. Let $\Sigma_{l,E}$ be the standard deviation of the pairwise cosine distances' distribution, the distances being computed between the samples of set E in the space induced by layer l. The measure is defined as:

$$c_4 = \text{mean}_{i \in \mathcal{I}} \, \text{mean}^k_{l \in \mathcal{L}} \, \frac{\Sigma_{l,C_i}}{\Sigma_{l,D}} \tag{4}$$

We use $k = 5$ for residual networks and $k = 1$ for VGG networks in our experiments.

To compute this measure, we found it helpful to standardize the representations of different neurons. More precisely, we normalize each neuron's pre-activations to have zero mean and unit variance, then apply a bias and ReLU activation function (Nair & Hinton, 2010) such that 25% of the samples are activated.

4 EXPERIMENTAL SETUP

The purpose of our experimental endeavour is to assess the predictive power of the intraclass clustering measures with respect to generalization performance. To this end, we reproduce the methodology introduced by Jiang et al. (2020). First of all, this methodology puts emphasis on the scale of the experiments to improve the generality of the observations. Second, it tries to go beyond standard measures of correlation, and puts extra care to detect causal relationships between the measures and generalization performance. This is achieved through a systematic variation of multiple hyperparameters when building the set of models to be studied, combined with the application of principled correlation measures.

4.1 BUILDING A SET OF MODELS WITH VARYING HYPERPARAMETERS

Our experiments are conducted on three datasets and two network architectures. The datasets are CIFAR10, CIFAR100 and the coarse version of CIFAR100 with 20 superclasses (Krizhevsky & Hinton, 2009). The two network architectures are Wide ResNets (He et al., 2016; Zagoruyko & Komodakis, 2016) (applied on CIFAR100 datasets) and VGG variants (Simonyan & Zisserman, 2014) (applied on CIFAR10 dataset). Both architectures use batch normalization layers (Ioffe & Szegedy, 2015) since they greatly facilitate the training procedure.

In order to build a set of models with a wide range of generalization performances, we vary hyperparameters that are known to be critical. Since varying multiple hyperparameters improves the

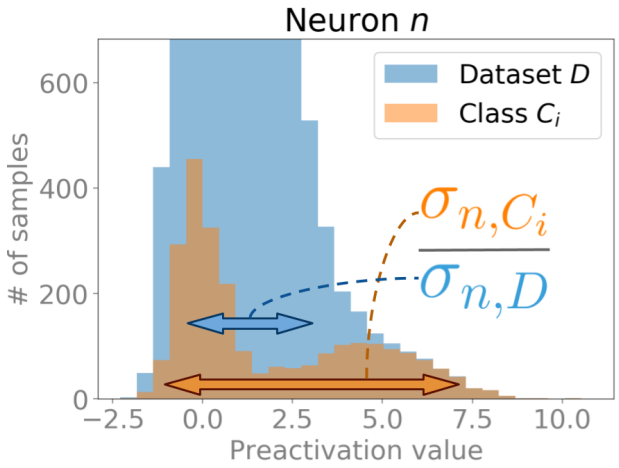

Figure 2: Our simplest measure (denoted c_3) quantifies intraclass clustering through the ratio of standard deviations σ_{n,C_i} and $\sigma_{n,D}$. Intuitively, a high ratio means that the neuron relies on features that *differentiate* samples from C_i although they belong to the same class. This illustration is based on a neuron from our experimental study, and the associated ratio is 2.47. Despite its simplicity, our results in Section 5 suggest a remarkably strong connection between c_3 and generalization performance.

identification of causal relationships, *we vary 8 different hyperparameters*: learning rate, batch size, optimizer (SGD or Adam (Kingma & Ba, 2015)), weight decay, dropout rate (Srivastava et al., 2014), data augmentation, network depth and width. In total, we elaborated a set of 192 hyperparameter configurations for each dataset-architecture pair[2].

4.2 Evaluating the measures' predictive power

Jiang et al. (2020) provides multiple criteria to evaluate measures of generalization. We opted for the granulated Kendall coefficient for its simplicity and intuitiveness. This coefficient compares the *ranking* of models provided by a measure to the ranking of models by their generalization performance. The Kendall coefficient is computed across variations of each hyperparameter independently. The average over all hyperparameters is then computed for the final evaluation score. The goal of this approach is to better capture causal relationships by penalizing measures that correlate with generalization only when specific hyperparameters are tuned.

We compare intraclass clustering-based measures to sharpness-based measures. The latter constituted the most promising measure family from the large scale study presented in Jiang et al. (2020). Many different sharpness measures exist. We decided to opt for magnitude-aware versions that measure sharpness through random and worst-case perturbations of the weights (denoted by $\frac{1}{\sigma'}$ and $\frac{1}{\alpha'}$ respectively in Jiang et al. (2020)). We also include the application of these measures with perturbations applied on kernels only (i.e. not on biases and batch normalization weights) with batch normalization layers in batch statistics mode (i.e. not in inference mode). We observed that these alternate versions often provided better estimations of generalization performance. We denote these measures by $\frac{1}{\sigma''}$ and $\frac{1}{\alpha''}$.

[2]The final set of VGG variants trained on CIFAR10 result from only 144 hyperparameter configurations as the network could not reach $\sim 100\%$ training accuracy with high dropout rates.

5 RESULTS

5.1 VISUALIZING THE MEASURES' RELATIONSHIPS WITH GENERALIZATION

We start our assessment of the measures' predictive power by visualizing in Figure 3 their relationship with generalization performance. *We observe a clear trend across datasets, network architectures and measures, suggesting a tight connection between intraclass clustering and generalization.*

It is worth noting that the neuron-level measures predict generalization performance as well as the layer-level measures in our experiments. The behaviour of some carefully selected neurons thus quite accurately predicts properties of the entire neural network they belong to. Moreover, while the importance of individually selective neurons was questioned in previous work (Morcos et al., 2018), our results rather suggest that neurons selective in terms of intraclass clusters are a key constituent of models with good generalization performance.

Moreover, it is also worth noting that the variance-based measures predict generalization performance as well as the measures based on label hierarchies, despite their complete lack of knowledge about the intraclass clusters' constitutions. Our work thus provides simple and generally applicable measures of intraclass clustering, which we hope will stimulate further research towards a more comprehensive understanding of this phenomenon.

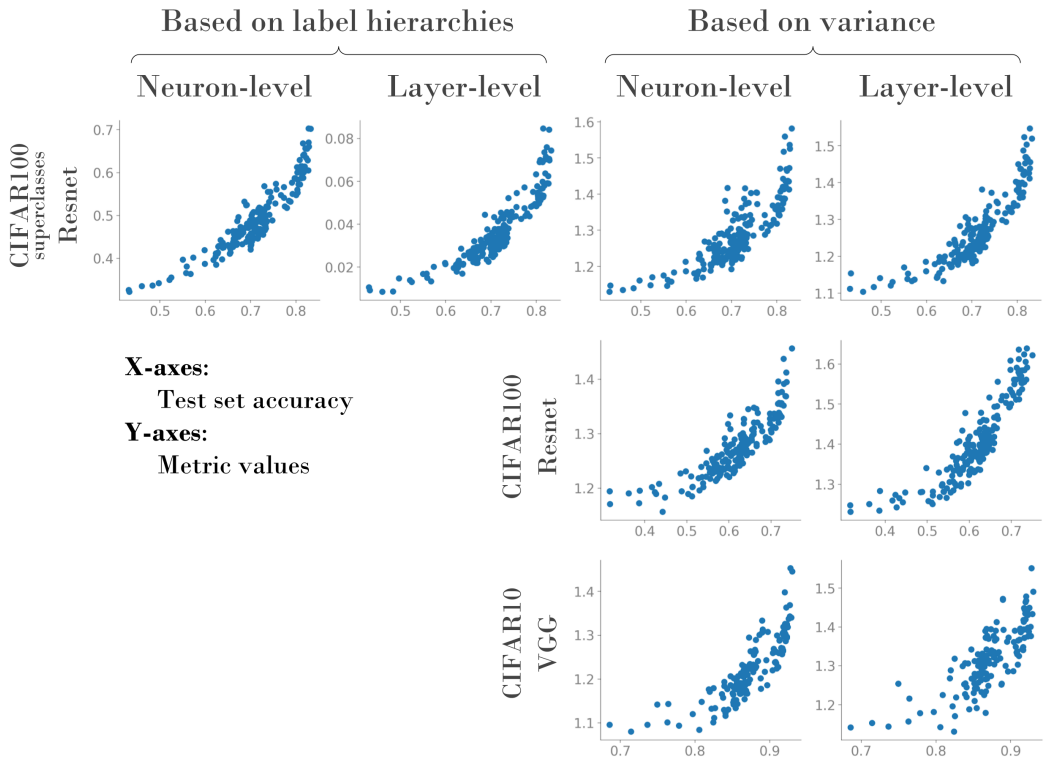

Figure 3: Visualization of the relationship between the four intraclass clustering measures and generalization performance, across datasets and network architectures. The four columns correspond to c_1, c_2, c_3 and c_4 measures respectively. All measures display a tight connection with generalization performance, suggesting a crucial role for intraclass clustering in the implicit regularization of deep neural networks.

5.2 EVALUATING MEASURES THROUGH THE GRANULATED KENDALL COEFFICIENT

To further support the conclusions of our visualizations, we evaluate the measures' predictive power through the granulated Kendall rank-correlation coefficient (cfr. Section 4.2). This evaluation criterion introduced in Jiang et al. (2020) constitutes a more principled assessment of the measures'

Table 1: Kendall coefficients for resnets trained on CIFAR100 superclasses.

		learning rate	batch size	weight decay	optim.	dropout rate	data augm.	width	depth	total score
Intraclass clustering	c_1	0.86	0.4	0.54	0.6	0.85	**1.0**	0.62	0.62	0.69
	c_2	0.83	0.56	**0.71**	0.44	**0.89**	**1.0**	0.19	0.38	0.62
	c_3	0.88	0.6	0.46	0.62	0.81	**1.0**	**0.81**	0.66	**0.73**
	c_4	**0.89**	0.69	0.62	0.65	0.86	**1.0**	0.44	0.69	**0.73**
Sharpness	$1 / \sigma'$	0.81	0.51	0.31	0.69	0.28	-0.58	0.67	0.61	0.41
	$1 / \sigma''$	0.86	0.58	0.17	0.4	-0.05	0.42	0.69	**0.72**	0.47
	$1 / \alpha'$	0.88	**0.94**	0.29	0.26	0.6	0.08	-0.03	-0.09	0.37
	$1 / \alpha''$	0.85	0.8	0.48	**0.71**	0.16	-0.08	0.08	0.34	0.42

Table 2: Kendall coefficients for resnets trained on CIFAR100.

		learning rate	batch size	weight decay	optim.	dropout rate	data augm.	width	depth	total score
Intraclass clustering	c_3	0.94	0.65	**0.62**	0.58	**1.0**	**1.0**	**1.0**	**0.78**	**0.82**
	c_4	0.93	0.62	**0.62**	0.21	**1.0**	**1.0**	0.91	**0.78**	0.76
Sharpness	$1 / \sigma'$	0.88	0.68	0.17	**0.8**	0.4	-0.62	0.94	0.61	0.48
	$1 / \sigma''$	0.92	0.61	0.12	0.35	-0.06	0.31	0.94	0.53	0.47
	$1 / \alpha'$	**0.96**	**0.96**	0.17	0.25	0.54	0.15	-0.16	-0.23	0.33
	$1 / \alpha''$	**0.96**	0.91	0.42	0.64	0.12	-0.25	0.17	0.14	0.39

Table 3: Kendall coefficients for VGG networks trained on CIFAR10.

		learning rate	batch size	weight decay	optim.	dropout rate	data augm.	width	depth	total score
Intraclass clustering	c_3	0.92	0.83	**0.67**	0.51	**0.92**	**1.0**	**1.0**	0.88	**0.84**
	c_4	0.86	0.75	0.33	0.29	**0.92**	**1.0**	0.54	0.92	0.7
Sharpness	$1 / \sigma'$	0.86	0.62	-0.25	0.6	-0.04	-0.27	**1.0**	0.85	0.42
	$1 / \sigma''$	0.9	0.67	0.11	**0.69**	**0.92**	0.19	**1.0**	**0.94**	0.68
	$1 / \alpha'$	**0.94**	**0.89**	0.61	0.53	0.67	0.77	0.15	0.06	0.58
	$1 / \alpha''$	0.93	0.67	0.36	0.54	0.69	-0.02	0.15	-0.15	0.4

relationship with generalization, and is designed to better capture causality. Moreover, it facilitates comparison with previous work by providing a quantitative evaluation. Tables 1, 2 and 3 present the granulated Kendall rank-correlation coefficients applied on intraclass clustering and sharpness-based measures, for the three dataset-architecture pairs respectively.

Our results reveal strong correlations between intraclass clustering measures and generalization performance *across all hyperparameters*. In terms of overall score, intraclass clustering measures surpass the sharpness-based measures variants by a large margin across all dataset-architecture pairs. Our results thus suggest that intraclass clustering plays a major role in the regularization of deep neural networks, and constitutes a promising research direction for future work.

5.3 Evolution of the measures over the course of training

To complete our experimental study, we provide a small step towards the understanding of the intraclass clustering *dynamics* by visualizing the evolution of the measures over the course of training in Figure 4.

A first interesting observation comes from the comparison of models with high and low generalization performances. It appears that *the differences in terms of intraclass clustering measures arise essentially during the early phase of training*. This observation makes an interesting link with previous work that highlighted the criticality of the early phase of training with respect to regularization (Golatkar et al., 2019), hessian eigenspectra (Gur-Ari et al., 2018), training data perturbations (Achille et al., 2019) and weight rewinding (Frankle et al., 2020a;b).

A second observation is that strong increases of the intraclass clustering measures systematically co-occur with strong increases of the training accuracy. This suggests that supervised training could act as a necessary driver for intraclass clustering ability, despite not explicitly targeting such behaviour.

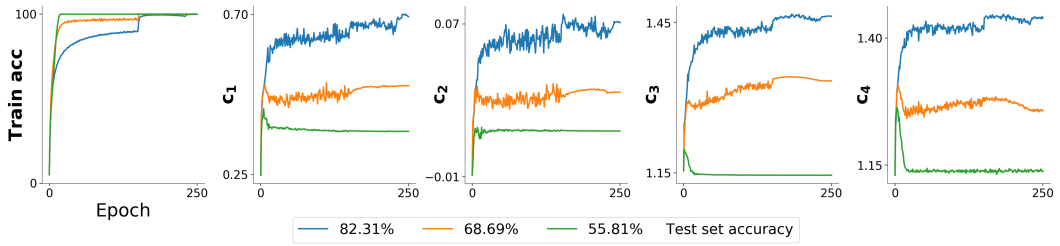

Figure 4: Evolution of the intraclass clustering measures over the course of training. We observe that the differences between models with high and low generalization performance arise essentially in the early phase of training. This enables an interesting link with previous work on this seemingly critical phase of neural network training (Golatkar et al., 2019; Gur-Ari et al., 2018; Achille et al., 2019; Frankle et al., 2020a;b).

6 Conclusion

Our work starts from the observation that classes of samples from standard image classification datasets are usually structured into multiple clusters. Since learning structure present in the data is commonly thought to improve generalization performance, it is natural to ask if state of the art models are capable of identifying such intraclass clusters.

While no supervision or training mechanisms are explicitly programmed into deep neural networks for targeting this behaviour, it is possible that this learning ability implicitly emerges during training. Such an hypothesis is especially compelling as recent work on generalization conjectured the emergence of implicit forms of regularization during deep neural network training.

In order to support this hypothesis, our work provides four different attempts at quantifying intraclass clustering in trained deep neural networks. We then conduct a large scale experimental study to identify potential relationships between the measures and generalization performance. The results reveal strong correlations across datasets and network architectures, suggesting a crucial role for intraclass clustering in the regularization of deep neural networks.

Important open questions remain to further support this work's hypothesis. In particular, the specific mechanisms that induce intraclass clustering in deep neural networks are not yet identified. Moreover, finding an explanation as to why the studied hyperparameters (learning rate, dropout, batch size,..) affect intraclass clustering would help confirm the causal relationship between intraclass clustering and generalization.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical Learning Periods in Deep Neural Networks. In *ICLR*, pp. 1–14, 2019.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *ICML*, 2017.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *CVPR*, 2017.

Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. Stiffness: A New Perspective on Generalization in Neural Networks. *arXiv 1901.09491*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In *ICML*, 2020a.

Jonathan Frankle, David J Schwab, and Ari S Morcos. The Early Phase of Neural Network Training. In *ICLR*, 2020b.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time Matters in Regularizing Deep Networks : Weight Decay and Data Augmentation Affect Early Learning. In *NeurIPS*, 2019.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient Descent Happens in a Tiny Subspace. *arXiv:1812.04754*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pp. 770–778, 2016.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML*, pp. 448—-456, 2015. ISSN 0717-6163. doi: 10.1007/s13398-014-0173-7.2. URL `http://arxiv.org/abs/1502.03167`.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. In *ICLR*, 2020.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *ICLR*, 2017.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Ari S Morcos, David G T Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *ICLR*, 2018.

Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, pp. 807—-814, 2010.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR*, 2015.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Sensitivity and generalization in neural networks: an empirical study. In *ICLR*, 2018.

K. Simonyan, A. Vedaldi, and A Zisserman. Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps. *ICLR*, 2014.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 (1):1929–1958, 2014.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. In *Deep Learning Workshop at ICML*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *BMVC*, 2016.

Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *ECCV*, pp. 818–833, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires re-thinking generalization. In *ICLR*, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in Deep Scene CNNs. In *ICLR*, 2015.