Critical Learning Periods in Deep Neural Networks

UCLA-TR-170017

Alessandro Achille[a,1,2], **Matteo Rovere**[b,1], **and Stefano Soatto**[a]

[a]Department of Computer Science, University of California, Los Angeles, 405 Hilgard Ave, Los Angeles, 90095, CA, USA; [b]Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

This manuscript was compiled on May 16, 2018

Critical periods are phases in the early development of humans and animals during which experience can irreversibly affect the architecture of neuronal networks. In this work, we study the effects of visual stimulus deficits on the training of artificial neural networks (ANNs). Introducing well-characterized visual deficits, such as cataract-like blurring, in the early training phase of a standard deep neural network causes a permanent performance loss that closely mimics critical period behavior in humans and animal models. Deficits that do not affect low-level image statistics, such as vertical flipping of the images, have no lasting effect on the ANNs' performance and can be rapidly overcome with further training. In addition, the deeper the ANN is, the more pronounced the critical period. To better understand this phenomenon, we use Fisher Information as a measure of the strength of the network's connections during the training. Our information-theoretic analysis suggests that the first few epochs are critical for the creation of strong connections across different layers, optimal for processing the input data distribution. Once such strong connections are created, they do not appear to change during additional training. These findings suggest that the initial rapid learning phase of ANN training, under-scrutinized compared to its asymptotic behavior, plays a key role in defining the final performance of networks. Our results also show how critical periods are not restricted to biological systems, but can emerge naturally in learning systems, whether biological or artificial, due to fundamental constrains arising from learning dynamics and information processing.

Critical Period | Deep Learning | Information Theory | Artificial Neuroscience | Information Plasticity

The term "critical period" refers to a phase in brain development during which the effects of experience lead to deep and irreversible remodeling of neural circuits (1). Similarly, the expression "sensitive period" is used to describe a time of heightened, yet reversible, plasticity in response to experience (2). The concept was introduced by Hubel and Wiesel in the 1960s, as part of their seminal work on the architecture and development of the visual system (3). In their classical experiments using monocularly deprived kittens, they studied the effects of the deficit on ocular dominance in the primary visual cortex (V1). They found that kittens monocularly blinded by lid suture in the first 3 months after birth remained blind in the deprived eye, while no effects were observed in adult cats subjected to analogous or more severe deficits (4, 5). Critical periods were later reported for a variety of animal species, including humans (6, 7), and are not limited to visual or, for that matter, to sensory systems. Either "critical" or "sensitive" periods have been observed, among others, in auditory space processing (8), filial imprinting (9), song learning in songbirds (10), language proficiency (11), and behavioral development (12). The evolutionary advantage of all these phenomena has been attributed to the delicate balance between adaptivity and robustness that has to be struck in order to maximize fitness in a mutable environment, using early experience as a guide (2, 13).

The biological mechanisms underlying the regulation of critical periods are manifold and depend on the species being studied, the brain regions involved and the nature of the experience (14). In this work we show that, by introducing visual deficits during the initial training epochs (*i.e.*, iterations over the training data) of common *artificial* deep neural networks, we observe trends that closely mimic those of humans and animal models. This may come as a surprise, since contemporary artificial neural networks (ANNs) are only loosely inspired by biological systems (15).

Most studies to date have focused either on the behavior of networks at convergence (Representation Learning) or on the asymptotic properties of the numerical scheme used to get there (Optimization). The role of the initial transient, especially its effect in biasing the network towards "good" regions of the complex and high-dimensional optimization problem, is rarely addressed. To study this initial learning phase of ANNs, we replicate experiments performed in animal models and find that the responses to early deficits are remarkably similar, despite the large underlying differences between the two systems. In particular, we show that the quality of the solution depends only minimally on the final, relatively well-understood, phase of the training process or on its very first epochs; instead, it depends critically on the period prior to initial convergence.

In animals, sensory deficits introduced during critical periods induce changes in the architecture of the corresponding areas (3, 7, 16). To determine whether a similar phenomenon exists in ANNs, we compute the Fisher Information of the weights of the network as a proxy to measure its "effective connectivity", that is, the density of connections that are effectively used by the network in order to solve the task. Like

Significance Statement

Similar to humans and animals, we find that artificial neural networks (ANNs) exhibit critical periods during which a temporary stimulus deficit can impair the development of a skill. The extent of the impairment depends on the inception and length of the deficit window, as in animal models, and on the size of the neural network. Our work shows how layer-wise changes in the effective connectivity, measured using the Fisher Information of the connections' weights, underpin critical periods in ANNs, and could be one of the mechanisms underlying biological critical periods as well. We refer to this phenomenon in deep networks as "Information Plasticity".

[1] Alessandro Achille and Matteo Rovere contributed equally to this work and are listed in alphabetical order.

[2] To whom correspondence should be addressed. E-mail: achille@cs.ucla.edu

others before us (17), we observe two distinct phases during the training, first a "learning phase" in which the Fisher Information of the weights increases as the network learns from the data, followed by a "consolidation" or "compression" phase in which the Fisher Information decreases and stabilizes. Sensitivity to critical-period-inducing deficits is maximal exactly when the Fisher Information peaks.

A layer-wise analysis of the network's effective connectivity shows that, in the tasks and deficits we consider, the hierarchy of low-level and high-level features in the training data is a key aspect behind the observed phenomena. In particular, our experiments suggest that the existence of critical periods in deep neural networks depends on the inability of the network to change its effective connectivity pattern in order to process different information (in response to deficit removal). We call this phenomenon, which is not mediated by any external factors, a loss of the "Information Plasticity" of the network.

Results

Deep Artificial Neural Networks exhibit critical periods. Studies of the critical period for monocular deprivation in humans rely on cohorts of patients affected by stimulus-deprivation amblyopia (reduced visual acuity in one eye), either in their infancy or childhood, caused by spontaneous or traumatic unilateral cataracts (18, 19). After surgical correction of the cataracts, the ability of the patients to regain normal acuity in the affected eye depends both on the duration of the deficit and on its age of onset. Earlier and longer deficits, as in animal studies of monocular deprivation (5, 20), lead to increasingly severe effects.

In order to replicate this experimental setup in ANNs, we train a standard convolutional network (CNN) to classify objects in small 32×32 RGB images from the CIFAR-10 dataset (21) in 10 classes. Recognition performance in the image classification task is akin to optotype symbol identification, employed in optometric acuity testing. To simulate the effect of cataracts, for the first t_0 epochs the images in the dataset are downsampled to 8×8 and then upsampled back to 32×32 using bilinear interpolation, in practice blurring the image and destroying small-scale details.[*] After that, the training continues for 300 more epochs, giving the network enough time to converge and ensuring it is exposed to the same number of uncorrupted images as in the control ($t_0 = 0$) experiment.

In Figure 1, we graph the final performance of the network (described in Materials and Methods) as a function of the epoch at which the deficit is corrected (t_0). We clearly observe the existence of a critical period for this deficit in the ANN: if the blur is not removed within the first 60 epochs, the final performance is severely decreased when compared to the baseline (from a test error of $\sim 6.4\%$, in the absence of a deficit, to more than 18% when the blur is present over 140 epochs, a $\sim 300\%$ increase). Once rescaled to account for the arbitrary time units (training epochs vs. days), the profile of the curve is also strikingly similar to the one obtained in kittens monocularly deprived from near birth and whose visual acuity upon eye-opening was tested and plotted against the length of the deficit window (23). Just like in humans and animal models (where critical periods are characteristic of early development), the critical period in the DNN also arises during the initial rapid learning phase. At this stage, the network is quickly learning a solution before the test error plateaus and the longer asymptotic convergence phase begins.

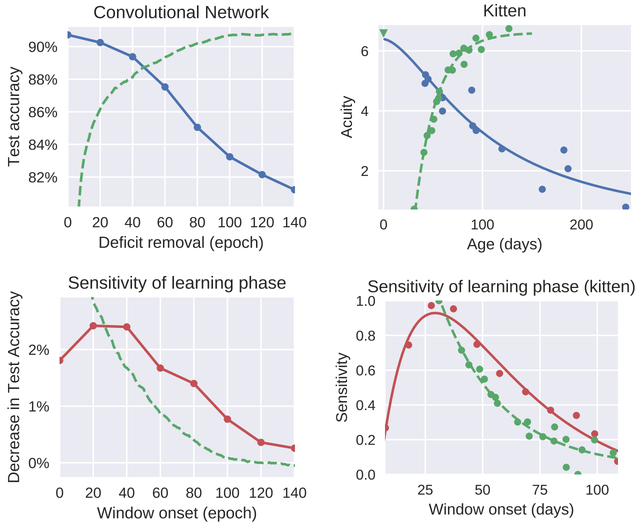

Fig. 1. (Top Left) Final test accuracy (solid line) achieved by a CNN trained in the presence of a deficit (image blur) plotted against the epoch at which the deficit is removed. The performance of the network is impaired if the blur is not removed early enough in the training. The critical period is centered in the early rapid learning phase of the network (the dashed line shows the rapid increase of test accuracy as a function of the training epoch, when training the CNN in the absence of deficit). **(Top Right)** Development of visual acuity in the cat as a function of age (dashed line), compared with the critical period for monocular deprivation (solid line; the visual acuity values plotted are those of kittens monocularly deprived since birth and tested at the time of eye-opening, while the green triangle marks the acuity of non-deprived kittens). Adapted from (22) and (23). Despite the profound differences between the two systems, the trends observed in animal models are remarkably similar to the ones obtained from a CNN. **(Bottom Left)** Here the deficit is introduced in a short window of 40 epochs, varying the epoch of onset, and the final test accuracy of the network is plotted as a function of the onset. The decrease in the CNN's performance can be used to probe the sensitivity of each epoch to the deficit: the most affected epochs are in the middle of the early rapid learning phase, before the test error (dashed line) begins to plateau. After the test error plateaus, the network is largely unaffected by the deficit. **(Bottom Right)** Degree of functional disconnection (normalized numbers of V1 monocular cells disconnected from the contralateral eye, 0 in a non-deprived kitten) plotted against the kittens' age at the onset of the monocular deprivation deficit window of 10-12 days (solid line). Adapted from (20). Overlaid on the critical period profile is the the plot of the percentage missing visual acuity at each age, normalized to the value attained by a mature cat (dashed line, data from (22)). Again, a similar experiment performed on animal models shows results analogous to the ANN's performance.

To quantify more accurately the sensitivity of the ANN to image blurring throughout its early learning phase, we introduced the deficit in a short constant window (40 epochs), starting at different epochs, and then measured the decrease in the ANN's final performance compared to the baseline. In Figure 1, we plot the final testing error of the network against the epoch of onset of the deficit. We observe that the network's sensitivity to blurring peaks in the central part of the early rapid learning phase (around 30 epochs), while later deficits produce little or no effect. A similar experiment was also performed on kittens by Olson and Freeman, using a window of 10-12 days during which the animals were monocularly deprived and using it to "scan" the first 4 months after birth to obtain a sensitivity profile (20). The results obtained on the DNN are, again, in accord with the trend observed in animal

[*] We employed this method, instead of a simpler Gaussian blur, since it has a very similar effect and makes the quantification of information loss clearer.

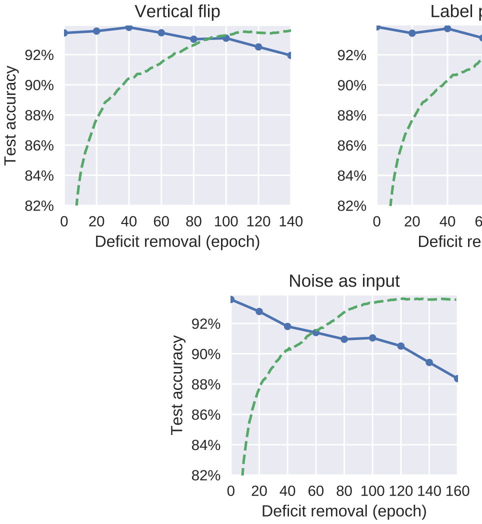

Fig. 2. High-level perturbations do not induce a critical period. **(Left)** Final performance of a CNN trained on vertically flipped images for the first t_0 epochs, plotted against t_0 and **(Right)** the same experiment performed when using permutation of the labels as the deficit. The solid blue curves are the test error at the end of the training as a function of the epoch of deficit removal, whereas the green dashed curves are the test error during training of a network in the absence of deficit. When the deficit only affects high-level features (as in the vertical flip) or the last layer of the CNN (label permutation), the network does not present a critical period (test accuracy remains largely flat). High-level features have also been shown to remain plastic well into adulthood in humans and animal models (24, 25). These findings suggest that the critical period in DNNs is closely related to the depth of the network and the hierarchy of image features. **(Bottom)** Sensory deprivation causes a less severe deficit response during the critical period. We simulate sensory deprivation by replacing the input images with Gaussian noise; remarkably, a critical period still exists, but it is longer and less pronounced than the one obtained in the case of image blur (Figure 1). A similar phenomenon has been reported in humans and animal models, where early sensory deprivation prolongs the plasticity of neuronal networks (26, 27).

models (Figure 1).†

High-level deficits are not associated with a critical period. Critical periods of neuronal networks dealing with high-level aspects of sensory processing end later than their lower-level counterparts, with some nuclei remaining plastic even in adulthood (7). Although the inversion of the visual field has not been systematically tested on young animals (or humans), it is well-reported that adult humans can quickly adapt to it (24, 30), suggesting the absence of a critical period. In addition to the behavioral recovery from the deficit, the anatomical plasticity in the adult has also been confirmed in animal models (25).

We observe a similar behavior in ANNs: in Figure 2 we perform the same experiment as before, this time using vertical flipping of the training images as the deficit, instead of blurring. We observe that the ANN is largely unaffected by vertical flipping and that the network quickly recovers its baseline performance after deficit correction. An analogous result is also obtained when the high-level deficit is a permutation of the task labels (Figure 2). These findings suggest that the existence of a critical period depends on the structure of the input data and the nature of the deficit applied, and, as such, cannot be interpreted solely as an artifact of the optimization process or the network's architecture. We have further explored these aspects in later experiments.

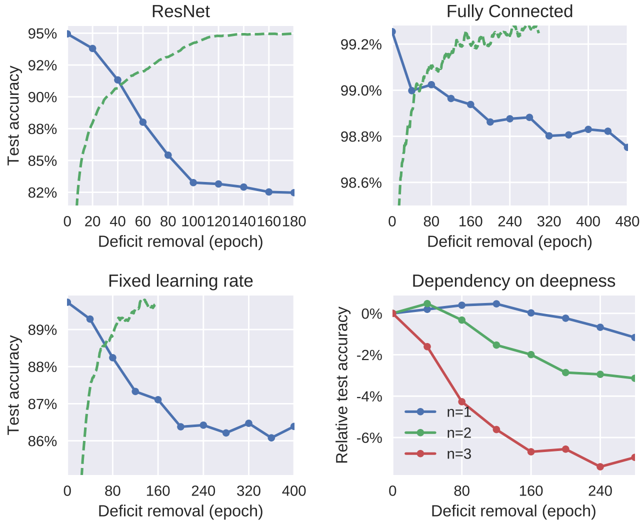

Fig. 3. Critical periods for an image blur deficit **(Top Left)** in a ResNet architecture trained on CIFAR-10 with learning rate annealing and **(Top Right)** in a deep fully-connected network trained on MNIST with a fixed learning rate. Different architectures, using different optimization methods and trained on vastly different datasets, still exhibit similar critical period behavior. **(Bottom Left)** Same experiment as in Figure 1, but in this case the network is trained with a fixed learning rate instead of using an annealing scheme. Although the time scale of the critical period is longer, the qualitative trends are the same. This finding supports the notion that having a critical period is an intrinsic property of deep networks and cannot be explained solely in terms of the loss landscape and energy functions, without taking into account the network's architecture. **(Bottom Right)** Dependence of the critical period profile on the network's depth: adding more convolutional blocks to the network increases the effect of the deficit during its critical period.

Sensory deprivation causes a less severe critical period. It has been observed in animal models, and recently indirectly confirmed in humans (26), that a period of early sensory deprivation (dark-rearing) can lengthen the critical period and thus cause less severe effects (at the same age) than those documented in light-reared animals (27). Similarly, both in humans and animal models, the effects of binocular cataracts have been characterized as being, somewhat counter-intuitively, less severe than those reported in patients and animals suffering from monocular deficits (31, 32). These findings, though, can be explained considering that, since visual experience is necessary to shape neuronal circuits, lack of experience can be, under certain circumstances, less damaging to the development of visual skills than defective experience (2).

Simulating complete visual deprivation in a neural network is not as simple as feeding a constant stimulus: a network presented with a constant blank input will rapidly become trivial and thus unable to train on new data. This is to be expected, since a blank input is a perfectly predictable stimulus and thus the network can quickly learn the (trivial) solution to the task. We instead want to model an uninformative stimulus, akin to noise. Moreover, even when the eyes are sutured or maintained in the darkness, there will be background excitation of photoreceptors that is best modeled as noise. To account for

† Olson and Freeman here employ a neuroanatomical readout to measure the impact of monocular deprivation on visual development. While excellent correlations between anatomy, physiology and behavior are generally observed (7, 16, 28, 29), direct comparisons with the ANN's performance in the image classification task are less warranted in this case. We later return to this aspect and characterize the "connectivity" of ANNs during training in *Information, synaptic strength, critical periods.*

this, we simulate sensory deprivation by replacing the input images with a dataset composed of (uninformative) random Gaussian noise. This way the network is trained on solving the highly non-trivial task of memorizing the association between the finitely-many noise patterns and their corresponding labels.

Figure 2 shows that this extreme form of deficit has a much less severe effect, during its critical period, than the one obtained by only blurring images. Using Gaussian noise during the early training phase of the ANN does not provide any information on the natural images. Yet its effects are milder than those caused by a deficit (*e.g.*, image blur), which instead conveys *some* information, but leads the network to (incorrectly) learn that no fine structure is present in the images. It is also interesting to note how a network trained to memorize noise patterns, while training on a completely different array of features than those needed to classify natural images, can still learn to identify images correctly with relative ease (Figure 2).

Dependence on the optimization algorithm and architecture. Deep networks are commonly trained using stochastic gradient descent (SGD) with an annealing scheme decreasing the learning rate over the duration of the training, in order to obtain convergence despite the noise introduced by the stochastic gradients. One possible explanation for the existence of critical periods in DNNs would thus be that, once the network enters a sub-optimal local minimum because of the deficit introduced, it cannot then (when the deficit is removed) escape due to the decreased learning rate. In Figure 2 we have already observed that in the case of several deficits the network manages to escape such a hypothetical local extremum despite the decreased learning rate. In Figure 3, we confirm that a critical period exists even when the learning rate, which can be thought of as regulating the temperature of the system undergoing gradient descent (34), is kept constant.

Figure 3 also shows that deep fully-connected networks, trained on the MNIST digit classification dataset, present a critical period for the image blur deficit; therefore the convolutional structure is not necessary, nor is the use of natural images. Similarly, a ResNet-18 trained on CIFAR-10 also has a critical period, sharper than the one found in a standard convolutional network (Figure 1). This is especially interesting, since ResNets allow for easier backpropagation of gradients to the lower layers, thus suggesting that the critical period is not caused by vanishing gradients.

Information, synaptic strength, critical periods

Deficits that result in critical periods reflect in changes of the brain architecture in the associated areas (7). This is inevitably different in artificial networks, as their connectivity is formally fixed at all times during training. However, not all the connections have the same importance for the final output of the network: This fact is captured, for example, by the Fisher Information Matrix (FIM) (35), which may be considered the artificial network equivalent of population-level measures of synaptic strength and connectivity in neuronal networks (36) (but see also (37) for alternative metrics). In this section, we briefly recall some properties of the FIM, and apply it to the study of the connections' strength of an artificial neural network trained in the presence and absence of deficits. Using this approach we aim to determine whether the (global) response of the ANN to deficits correlates with (local) changes in connectivity, similarly to how behavioral and physiological features correlate in experience-deprived animal models (3, 16, 28, 29).

The Fisher Information Matrix. Generally, an artificial neural network encodes a distribution $p_w(y|x)$ — parametrized by the weights w — of the task variable y (*e.g.*, an image label in CIFAR-10), given an input image x. Intuitively, the importance of a specific connection can be estimated by perturbing the corresponding weight and looking at the magnitude of the change in the final distribution. Let us consider then a perturbation $w' = w + \delta w$ of the weights, which results in a perturbed distribution $p_{w'}(y|x)$. The discrepancy between the original distribution and the perturbed one can be measured by their Kullback-Leibler divergence, which, to second-order approximation, is given by:

$$\mathbb{E}_x \operatorname{KL}(\, p_{w'}(y|x) \,\|\, p_w(y|x)\,) = \delta w \cdot F \delta w + o(\delta w^2),$$

where the expectation over x is computed using the empirical data distribution $\hat{Q}(x)$ given by the dataset, and $F := \mathbb{E}_{x\sim\hat{Q}(x)}\mathbb{E}_{y\sim p_w(y|x)}[\nabla_w \log p_w(y|x)\nabla_w \log p_w(y|x)^T]$ is the Fisher Information Matrix (FIM). The FIM can thus be considered a local metric measuring how much the perturbation of a single weight (or a combination of weights) affects the output of the the network (38). In particular, weights with low Fisher information can be removed (pruned) without affecting the network's performance, which suggests that the Fisher Information can be used, in an artificial network, as a measure of the effective connectivity or, more generally, of the strength of a connection (36). For these reasons, especially when drawing comparison with neuronal networks, we will sometimes use "connection strength" and Fisher Information interchangeably. As its name suggests, the FIM can also be considered as a measure of the quantity of information that the model contains about the training data (35). Indeed, we expect the strength of the connections to increase as we learn more information from experience. Finally, the FIM is also a semi-definite approximation of the Hessian of the loss function (39) and hence of the curvature of the loss landscape at a particular point w during training, which provides a connection between the FIM and the optimization procedure (38).

We will employ the trace of the Fisher Information Matrix to measure the global or layer-wise connection strength, since it can be calculated efficiently despite the large size of our networks. We also considered computing the log-determinant of the matrix using the Kronecker-Factorized approximation (40, 41) in order to capture the behavior of off-diagonal terms, but we observed the same qualitative trend as the trace. We use ResNets to compute the FIM because, being a local measure, it is very sensitive to the irregularities of the loss landscape and ResNets have a relatively smooth landscape (42). For other network architectures we use a more stable estimator of the FIM based on the injection of noise in the weights (43) (also see Materials and Methods).

Information changes during training. In Figure 4 we estimate the global changes to the connections' strength in a ResNet by plotting the trace of the FIM during the network training. We observe that during an initial phase the network acquires information about the data, which results in a large increase in the strength of the connections. However, once the performance

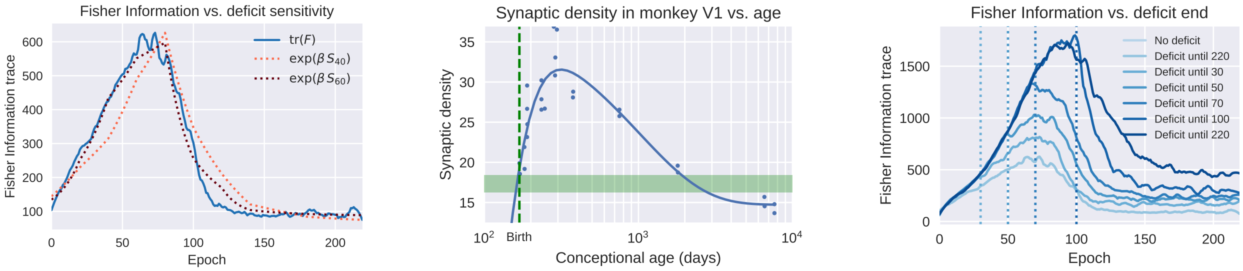

Fig. 4. **(Left)** Plot of the trace of the Fisher Information Matrix (FIM) of the network weights as a function of the training epoch (solid blue line). We observe two distinct phases during the training: First, the network employs increasingly more connections in order to solve the task (and consequently information grows). Once a good solution is found, the network starts to consolidate the connections (and information decreases), so that fewer resources (information, or connections if the weak ones were pruned) are needed to solve the task. This follows the same trend as the network's sensitivity to critical periods, as computed in Figure 1 and fitted to the Fisher Information using a simple exponential (see Materials and Methods), suggesting that the two quantities are linked. We show the result using both a window size of 40 and 60 to compute the sensitivity curve. **(Center)** The FIM can be interpreted as a measure of the density of the effective network connections. Synaptic density during development (here shown for the visual cortex of macaques; the green dashed line marks birth and the green stripe is the average synaptic density in a healthy adult, adapted from (33)) follows a similar trend where, after a sharp increase and a peak during early development, progressive elimination of synapses (pruning) occurs and continues throughout life. ANNs, though, appear to have both a different timescale and a more abrupt decrease in the effective connectivity after peaking. **(Right)** Effects of deficit (image blur) on the FIM. In the presence of a critical-period-inducing deficit, the ANN employs a much larger set of connections to solve the task. This can be explained by the fact that the network, unable to find a general rule to classify the corrupted images, is forced to memorize them in order to solve the task. This reflects in increased use of the network's resources, correctly captured by the FIM, it being a measure of the ANN's information content.

in the task begins to plateau, the network starts reducing the overall strength of its connections while keeping its performance unchanged. This can be seen as a "consolidation" or "compression" phase, during which redundant connections are eliminated and non-relevant memorized information is discarded. It must be noted how a related observation on the training procedure consisting of two phases has been suggested by (17), who, however, focus on the compression of the *activations* of the network, rather than on the information in the *weights* and their connectivity. There is a fundamental difference between these two analyses: The relation between the information in the weights and the activations is non-trivial and articulated in (43). In particular, compression of the weights *can* imply compression of the activations, under suitable non-generic assumptions.

Interestingly, this change in the connection strength is closely related to the sensitivity to critical-period-inducing deficits as image blur, computed using the sliding window method as in Figure 1. In Figure 4 we see that the sensitivity is maximal precisely when the Fisher Information peaks. Moreover, we observe that the exponential of the sensitivity closely fits the Fisher Information, which is remarkable since the FIM is a local quantity computed at a single point during the training of a network without deficit, while the sensitivity during a critical period is computed using the test data at the end of the training of a network in the presence of a deficit. It is also interesting to see the effect of deficit introduction on the Fisher Information. From Figure 4 (Right), we see that the FIM trace grows much more in the presence of the deficit (image blur) and remains substantially higher even after it is removed. When the data are so corrupted as to be classified incorrectly, the network is forced to memorize the labels, therefore increasing the quantity of information needed to perform the same task.

Notice that a similar phenomenon is observed in the development of neuronal networks (Figure 4, Center), where the density of synaptic connections rapidly peaks during early development and slowly decreases afterwards as a results of synaptic pruning (44, 45). However, while deficits have been shown to affect the synaptic density in animal models (45), an increase in the overall number of connections is not observed. This difference can be partly attributed to the fact that in our experiments ANNs are trained using a finite dataset, making pure memorization a viable strategy to solve the task and therefore encouraging the storage of unnecessary information.

Layer-wise effects of deficits. We can gain additional insights on what is happening in an ANN affected by a deficit by looking at the layer-wise changes of the Fisher Information. In order to do so, we used the All-CNN architecture, which has a clearer subdivision in layers than a ResNet. When the network is trained without deficits (Figure 5, Left), the strongest (most important) connections are in the intermediate layers. This reflects the fact that most of the information in CIFAR-10 images is present at an "intermediate" scale, with low level information (*e.g.*, edges) alone being insufficient for identification, and global/contextual information (*e.g.*, background, position of objects) providing only limited benefit to the task performance. Therefore, we expect the network to focus its resources on the intermediate layers, since their receptive fields are optimal for this task.

On the other hand, if the network is initially trained on blurred data (Figure 5, Top Center and Top Right), the strength of the connections is dominated by the top layer (Layer 6). This is to be expected, since the low-level and mid-level structures of the images are destroyed and all of the remaining information is contained in the global features of the image, which can be processed by Layer 6, but not by the low-level layers with their smaller receptive fields. When the deficit is removed early in the training (Figure 5, Top Center), the network manages to "re-organize", reducing the information contained in the last layer, and, at the same time, increasing the Fisher Information in the intermediate layers, in order to process the new features available. We refer to these changes in the Fisher Information as "Information Plasticity". However, if the deficit is removed after the beginning of the "consolidation" phase, the quantity of information contained in the mid-level layers does not change significantly because of the network's loss of Information Plasticity.

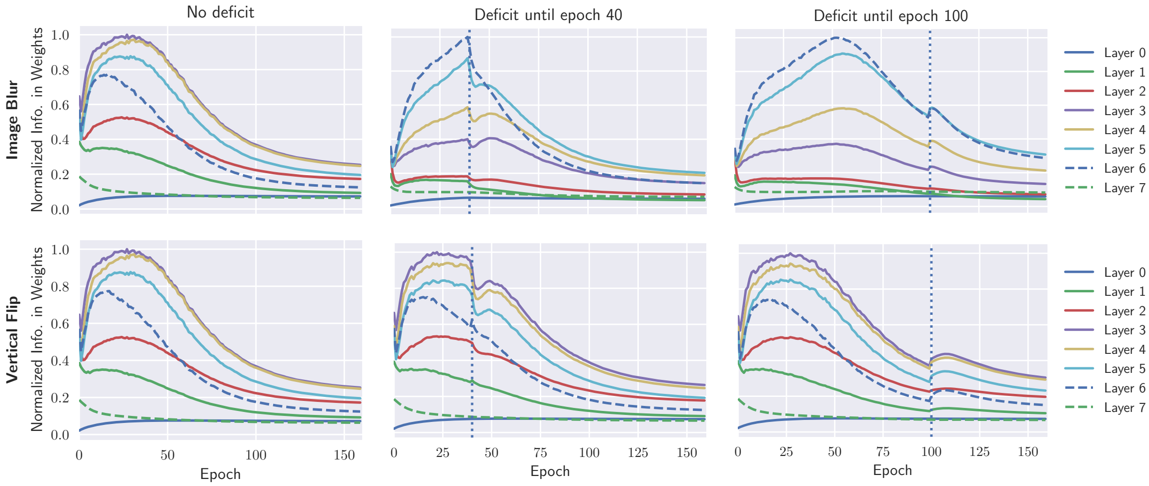

Fig. 5. Normalized quantity of information contained in the weights of each layer as a function of the training epoch. **(Top Left)** In the absence of deficit, the network initially acquires a large amount of information, stored in the connections (weights), and uses it to quickly learn the new task (the training error also decreases rapidly, with possible overfitting). As training progresses, the network explores more efficient solutions to the task, requiring less resources (the "compression phase" to optimize the information regularizer described in (43)). It must be noted that most of the resources of the network (*i.e.*, information in the weights) are allocated to the middle layers (layers 3-4-5). **(Top Right)** In the presence of an image blur deficit until epoch 100, the task is significantly more difficult, and the network allocates more resources in order to solve it (see also Figure 4). Most importantly, more resources are allocated to the higher layers (layer 6) rather than to the middle layers (as it was the case in the absence of deficit). This can be explained by the fact that the deficit destroys the low- and mid-level features that could be processed by those layers and leaves only the global features of the image, which can only be processed by the higher layers. When the deficit is removed, after a small transient, the layers maintain the same content of information instead of allocating more resources to the middle layers. **(Top Center)** On the other hand, when the deficit is removed at an earlier epoch, the layers are, at least partially, reconfigured (layer 6 rapidly loses information, relative to other layers), without long-term consequences. The most sensitive period for learning a task appears to coincide with the start of a "consolidation phase", after the initial fast-changing transient. **(Bottom Left, Center, Right)** The layer-wise normalized quantity of information is shown for an ANN trained in the presence of a vertical flip deficit, which does not induce a critical period (from **Left** to **Right**: no deficit, vertical flip until epoch 40, vertical flip until epoch 100). As expected, the relative quantity of information in the layers is not affected.

Discussion

ANNs and neuronal networks are vastly different systems, nonetheless both exhibit critical periods during their early learning phases and these share a number of common features. In particular, we have shown that while cataract-like deficits, such as severe image blurring, present a critical period, high-level impairments, like image flipping, have little or no effect on the DNNs' performance. These findings are in agreement with the characterizations of critical period responses to deficits in humans and animal models. In addition, we use the Fisher Information to study the development of the DNNs' effective connections during training and recognize an early phase where the Fisher Information increases sharply, followed by a decrease. The maximum sensitivity to critical-period-inducing deficits corresponds to the maximum of the Fisher Information.

It must be noted that critical periods depend on the task explored, the deficit applied, and the readout used: Due to the fundamental differences between neuronal networks and ANNs, it is challenging to perfectly match experiments to animal models. Since visual pathways are extensively studied in the neurobiology of critical periods and DNNs are particularly suited for image classification, we chose to focus our analysis on this task.

"Behavioral" readouts upon deficit removal, both in ANNs and neuronal networks, can potentially be confounded by deficit-coping changes at different levels of the visual pathways (2, 7). Since deficits in deprived animals are mirrored by abnormalities in the circuitry of the visual pathways, we chose to also characterize their effects on network connections. While the connectivity of an ANNs is fixed by design, its "effective connectivity", *i.e.*, the connections that are actually employed by the network to solve the task, is not: We quantify it from the Fisher Information Matrix of the network's weights. Sensitivity to critical periods and the trace of the Fisher Information peak at the same epoch, in accord with the evidence that skill development and critical periods in neuronal networks are modulated by changes (generally experience-dependent) in synaptic plasticity (2, 14). Our layer-wise analysis of the Fisher Information (Figure 5) also shows that, while in a non-deprived network connections are stronger in the intermediate layers, upon introduction of an image blurring deficit, the ANN reinforces higher layers to the detriment of intermediate layers, leaving low-level layers virtually untouched. If the deficit is removed after the ANN's critical period ends, the network is not able to reverse these effect. Although the two systems are radically different, it is worth noting how similarities to such response to the introduction of visual deficits can be found in the visual pathways of animal models (and, in part, in humans). Lower levels (*e.g.*, retina, lateral geniculate nucleus) are little (or not at all) affected by deficits (4, 16, 46). Higher-level visual areas *e.g.*, V2 and post-V2) also show little remodeling upon deprivation (47–49), possibly because their plasticity is maintained into adulthood (7) and thus most of the neuroanatomical and physiological changes due to deprivation cluster, with varying severity, in the different layers of V1 (which we can compare to an intermediate-processing level) (3, 16).

In ANNs, the Fisher Information can be interpreted as an approximation of the local curvature of the loss landscape (39). According to this view, Figure 4 (Left) suggests that SGD encounters two different phases during the network training: At first, the network moves towards high curvature regions of the loss landscape, while in the second phase the curvature decreases and the network eventually converges to a flat minimum (as also observed in (50)). We could interpret these as the network crossing a narrow bottleneck during its training before entering a flat region of the loss surface. When combining this assumption with our deficit sensitivity analysis, we can hypothesize that the critical period occurs precisely upon crossing of this bottleneck. A different two-phase interpretation of SGD has also been proposed in (17): In the authors' analysis, the network first learns to encode the task in the network's activations, while in a second phase it minimizes the quantity of information, while keeping enough of it to maintain good task performance. The relationship between the biphasic behaviors we observe in the weights, and those observed by (17) in the activations, is non-trivial, as described in (43) and even defining information quantities for the activations presents a number of technical challenges.

There is evidence that convergence to flat minima (minima with low curvature) in a DNN correlates with a good generalization performance (42, 50–52). Indeed, by recalling the link between the Fisher Information and curvature, we can deduce from Figure 4 (Right) that networks impaired by the deficit converge to sharper minima than unaffected networks. However, we have found that the performance of the network is determined by an early sensitive phase, during which the network crosses high-curvature region of the loss landscape. This suggests that the final sharpness at convergence may be an epiphenomenon rather than the cause of good generalization.

Applying a deficit at the beginning of the training may be compared to the common practice of pre-training and fine-tuning an ANN, whereby the network is initially trained on a similar task, for which data are plentiful, and then fine-tuned to solve a specific task. This is generally found to improve the performance of the network. In (53) a similar practice is analyzed, but in the different setting of layer-wise unsupervised pretraining, seldom used in current practice, and suggests that the pre-training may act as a regularizer by moving the weights of the network towards an area of the loss landscape closer to the attractors for good solutions. The authors observe that the initial examples have a stronger effect in steering the network towards particular solutions. Here, we have shown that pre-training on blurred data can have the opposite effect; *i.e.*, it can severely decrease the final performance of the network. Following the loss landscape interpretation, this would suggest that, surprisingly, a configuration of parameters able to classify blurred images is farther apart from a configuration able to correctly classify real images. On the other hand, a configuration able to memorize noise patterns can access a better solution to the original task. However, it must be remembered that the interpretation of the deficit's effect as moving the network close to a bad attractor is difficult to reconcile with the smooth transition observed in the critical periods, since the network would either converge to this attractor, and thus have low accuracy, or not. A richer theoretical framework may therefore be needed to study critical periods purely in terms of the geometry of the loss landscape.

A peculiar interpretation of critical periods in humans and animal models is proposed in (2). Knudsen claims that the initial connectivity profile of neuronal networks is unstable and easily changed by early experiences (highly plastic). As more "samples" are observed, the connections change and reach a more stable configuration, which is not as easily modified by additional experience. Learning can, however, still happen within the newly created connectivity pattern. This is largely compatible with our findings: Sensitivity to critical-period-inducing deficits peaks when the remodeling of connections is at its maximum (Figure 4, Left) and different connectivity profiles are observed in networks trained with and without a deficit (Figure 5). This also agrees with the fact that high-level deficits such as image-flipping and label permutation, which do not require restructuring of the network's connections in order to be corrected, do not exhibit a critical period.

Finally, we want to emphasize how our goal here is not so much to investigate the human (or animal) brain through ANNs, as to understand fundamental information processing phenomena, both in their biological or artificial implementations. It is also not our goal to suggest that, since they both exhibit critical periods, ANNs are necessarily a valid model of neurobiological information processing, although recent work has emphasized this aspect (15, 54, 55). We engage in an "Artificial Neuroscience" exercise in part to address a technological need to develop "explainable" artificial intelligence systems whose behavior can be understood and predicted. While traditionally well-understood mathematical models were used by neuroscientists to study biological phenomena, information processing in modern artificial networks is just as poorly understood as in biology, so we chose to exploit well-known biological phenomena as probes to study information processing in artificial networks. We have shown how our approach can help unveil how information is represented, acquired and transformed in complex networks of relatively simple components, whether biological or artificial.

Materials and Methods

In all of the experiments, unless otherwise stated, we use the following architecture, adapted from (56):

$$\texttt{conv 96 - conv 96 - conv 192 s2 - conv 192 - conv 192 -}$$
$$\texttt{conv 192 s2 - conv 192 - conv1 192 - conv1 10 -}$$
$$\texttt{avg. pooling - softmax}$$

where each `conv` block consists of a 3×3 convolution, batch normalization and ReLU activations. `conv1` denotes a 1×1 convolution. The network is trained with SGD, with a batch size of 128, learning rate starting from 0.05 and decaying smoothly by a factor of .97 at each epoch. We also use weight decay with coefficient 0.001. In the experiments with a fixed learning rate, we fix the learning rate to 0.001, which we find to allow convergence without excessive overfitting. For the ResNet experiments, we use the ResNet-18 architecture from (57) with initial learning rate 0.1, learning rate decay .97 per epoch, and weight decay 0.0005. When experimenting with varying network depths, we use the following architecture:

$$\texttt{conv 96 - [conv } 96 \cdot 2^{i-1} \texttt{ - conv } 96 \cdot 2^{i} \texttt{ s2]}_{i=1}^{n} \texttt{ - conv } 96 \cdot 2^{n}$$
$$\texttt{- conv1 } 96 \cdot 2^{n} \texttt{ - conv1 10}$$

In order to avoid interferences between the annealing scheme and the architecture, in these experiments we fix the learning rate to 0.001. The Fully Connected network used for the MNIST experiments has hidden layers of size $[2500, 2000, 1500, 1000, 500]$. All hidden layers use batch normalization followed by ReLU activations. We fix the learning rate to 0.005. Weight decay is not used. We use data augmentation with random translations up to 4 pixels and random horizontal flipping. For MNIST, we pad the images with zeros to bring them to size 32×32.

To compute the trace of the Fisher Information Matrix, we use the following expression derived directly from the definition:

$$\begin{aligned} \operatorname{tr}(F) &= \mathbb{E}_{x \sim \hat{Q}(x)} \mathbb{E}_{y \sim p_w(y|x)} [\operatorname{tr}(\nabla_w \log p_w(y|x) \nabla_w \log p_w(y|x)^T)] \\ &= \mathbb{E}_{x \sim \hat{Q}(x)} \mathbb{E}_{y \sim p_w(y|x)} [\|\nabla_w \log p_w(y|x)\|^2], \end{aligned}$$

where the input image x is sampled from the dataset, while the label y is sampled from the output posterior. Expectations are approximated by Monte-Carlo sampling. Notice, however, that this expression depends only on the local gradients of the loss with respect to the weights at a point $w = w_0$, so it can be noisy when the loss landscape is highly irregular. This is not a problem for ResNets (42), but for other architectures we use instead a different technique, proposed in (43). More in detail, let $L(w)$ be the standard cross-entropy loss. Given the current weights w_0 of the network, we find the diagonal matrix Σ that minimizes:

$$L' = \mathbb{E}_{w \sim N(w_0, \Sigma)}[L(w)] - \beta \log |\Sigma|,$$

where β is a parameter that controls the smoothness of the approximation. Notice that L' can be minimized efficiently using the method in (58). To see how this relates to the Fisher Information Matrix, assume that $L(w)$ can be approximated locally in w_0 as $L(w) = L_0 + a \cdot w + w \cdot Hw$. We can then rewrite L' as

$$L' = L_0 + \mathrm{tr}(\Sigma H) - \beta \log |\Sigma|.$$

Taking the derivative with respect to Σ, and setting it to zero, we obtain $\Sigma_{ii} = \beta / H_{ii}$. We can then use Σ to estimate the trace of the Hessian, and hence of the Fisher information.

Fitting of sensitivity curves and synaptic density profiles from the literature was performed using GraphPad Prism 7 (GraphPad Software, La Jolla, CA), using:

$$f(t) = \mathrm{e}^{-(t-d)/\tau_1} - k\mathrm{e}^{-(t-d)/\tau_2}$$

as the fitting equation. t is the age at the time of sampling and τ_1, τ_2, k and d are unconstrained parameters (59).

The exponential fit of the sensitivity to the Fisher Information trace uses the expression $F(t) = a \exp(cS_k(t)) + b$, where a, b and c are unconstrained parameters, $F(t)$ is the Fisher Information trace at epoch t of the training of a network without deficits and S_k is the sensitivity computed using a window of size k. That is, $S_k(t)$ is the increase in the final test error over a baseline when the network is trained in the presence of a deficit between epochs t and $t + k$.

ACKNOWLEDGMENTS. Supported by ONR N00014-17-1-2072, ARO W911NF-17-1-0304, AFOSR FA9550-15-1-0229 and FA8650-11-1-7156.

1. Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ (2013) *Principles of Neural Science*. (McGraw-Hill, New York, NY), 5th edition.
2. Knudsen EI (2004) Sensitive periods in the development of the brain and behavior. *Journal of cognitive neuroscience* 16(8):1412–25.
3. Wiesel TN, Hubel DH (1963) Single-Cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of neurophysiology* 26:1003–17.
4. Wiesel TN, Hubel DH (1963) Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *Journal of Neurophysiology* 26(6):978–993.
5. Wiesel TN (1982) Postnatal development of the visual cortex and the influence of environment. *Nature* 299(5884):583–91.
6. Berardi N, Pizzorusso T, Maffei L (2000) Critical periods during sensory development. *Current opinion in neurobiology* 10(1):138–145.
7. Daw NW (2014) *Visual Development*. (Springer, New York, NY), 3rd edition.
8. Knudsen EI (2002) Instructed learning in the auditory localization pathway of the barn owl. *Nature* 417(6886):322–8.
9. Hess E (1973) *Imprinting: early experience and the developmental psychobiology of attachment*, Behavioral Science Series. (Van Nostrand Reinhold Co., New York, NY).
10. Konishi M (1985) Birdsong: from behavior to neuron. *Annual review of neuroscience* 8(1):125–70.
11. Newport EL, Bavelier D, Neville HJ (2001) Critical thinking about critical periods: Perspectives on a critical period for language acquisition. in *Language, brain and cognitive development: Essays in honor of Jacques Mehler*, ed. Doupoux E. (MIT Press, Cambridge, MA), pp. 481–502.
12. Fox SE, Levitt P, Nelson CA (2010) How the timing and quality of early experiences influence the development of brain architecture. *Child development* 81(1):28–40.
13. Bateson P, Gluckman P (2011) *Plasticity, Robustness, Development and Evolution*. (Cambridge University Press, Cambridge, UK).
14. Hensch TK (2004) Critical period regulation. *Annual review of neuroscience* 27(1):549–79.
15. Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired artificial intelligence. *Neuron* 95(2):245–258.
16. Hendrickson AE, et al. (1987) Effects of early unilateral blur on the macaque's visual system. II. Anatomical observations. *The Journal of neuroscience* 7(5):1327–39.
17. Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.
18. Vaegan, Taylor D (1979) Critical period for deprivation amblyopia in children. *Transactions of the ophthalmological societies of the United Kingdom* 99(3):432–9.
19. von Noorden GK (1981) New clinical aspects of stimulus deprivation amblyopia. *American journal of ophthalmology* 92(3):416–21.
20. Olson CR, Freeman RD (1980) Profile of the sensitive period for monocular deprivation in kittens. *Experimental Brain Research* 39(1):17–21.
21. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images, (University of Toronto), Technical report.
22. Giffin F, Mitchell DE (1978) The rate of recovery of vision after early monocular deprivation in kittens. *The Journal of physiology* 274(1978):511–37.
23. Mitchell DE (1988) The extent of visual recovery from early monocular or binocular visual deprivation in kittens. *The Journal of physiology* 395:639–60.
24. Kohler I (1964) *The formation and transformation of the perceptual world*, Psychological Issues Monographs. (International Universities Press, New York, NY).
25. Sugita Y (1996) Global plasticity in adult visual cortex following reversal of visual input. *Nature* 380(6574):523–6.
26. Kalia A, et al. (2014) Development of pattern vision following early and extended blindness. *Proceedings of the National Academy of Sciences* 111(5):2035–2039.
27. Mower GD (1991) The effect of dark rearing on the time course of the critical period in cat visual cortex. *Brain research. Developmental Brain Research* 58(2):151–8.
28. Kiorpes L, et al. (1987) Effects of early unilateral blur on the macaque's visual system. I. Behavioral observations. *The Journal of neuroscience* 7(5):1318–26.
29. Movshon JA, et al. (1987) Effects of early unilateral blur on the macaque's visual system. III. Physiological observations. *The Journal of neuroscience* 7(5):1340–51.
30. Stratton GM (1896) Some preliminary experiments on vision without inversion of the retinal image. *Psychological Review* 3(6):611–617.
31. Maurer D, Lewis TL (1993) Visual outcomes after infantile cataract. in *Early visual development, normal and abnormal.*, ed. Simons K. (Oxford University Press, New York, NY), p. 454–484.
32. Wiesel TN, Hubel DH (1965) Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *Journal of neurophysiology* 28(6):1029–40.
33. Rakic P, Bourgeois JP, Eckenhoff MF, Zecevic N, Goldman-Rakic PS (1986) Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science* 232(4747):232–5.
34. Welling M, Teh YW (2011) Bayesian learning via stochastic gradient langevin dynamics in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 681–688.
35. Fisher RA (1925) Theory of statistical estimation in *Mathematical Proceedings of the Cambridge Philosophical Society*. (Cambridge University Press), Vol. 22, pp. 700–725.
36. Kirkpatrick J, et al. (2017) Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114(13):3521–3526.
37. Zenke F, Poole B, Ganguli S (2017) Continual learning through synaptic intelligence in *International Conference on Machine Learning*. pp. 3987–3995.
38. Amari Si, Nagaoka H (2007) *Methods of information geometry*. (American Mathematical Soc.) Vol. 191.
39. Martens J (2014) New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*.
40. Martens J, Grosse R (2015) Optimizing neural networks with kronecker-factored approximate curvature in *International conference on machine learning*. pp. 2408–2417.
41. Grosse R, Martens J (2016) A kronecker-factored approximate fisher matrix for convolution layers in *International Conference on Machine Learning*. pp. 573–582.
42. Li H, Xu Z, Taylor G, Goldstein T (2017) Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*.
43. Achille A, Soatto S (2017) Emergence of Invariance and Disentangling in Deep Representations. *ArXiv e-prints*.
44. Huttenlocher PR, de Courten C, Garey LJ, Van der Loos H (1982) Synaptogenesis in human visual cortex – evidence for synapse elimination during normal development. *Neuroscience letters* 33(3):247–52.
45. Huttenlocher PR (2002) *Neural Plasticity: The Effects of Environment on the Development of the Cerebral Cortex*. (Harvard University Press, Cambridge, MA).
46. Sherman SM, Stone J (1973) Physiological normality of the retinal in visually deprived cats. (1):224–30.
47. Sincich LC, Jocson CM, Horton JC (2012) Neuronal projections from V1 to V2 in amblyopia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32(8):2648–56.
48. Gandhi T, Kalia A, Ganesh S, Sinha P (2015) Immediate susceptibility to visual illusions after sight onset. *Current Biology* 25(9):R358–R359.
49. Gandhi TK, Singh AK, Swami P, Ganesh S, Sinha P (2017) Emergence of categorical face perception after extended early-onset blindness. *Proceedings of the National Academy of Sciences of the United States of America* 114(23):6139–6143.
50. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP (2017) On large-batch training for deep learning: Generalization gap and sharp minima.
51. Hochreiter S, Schmidhuber J (1997) Flat minima. *Neural Computation* 9(1):1–42.
52. Chaudhari P, Choromanska A, Soatto S, LeCun Y (2016) Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*.
53. Erhan D, et al. (2010) Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 11(Feb):625–660.
54. Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* 19(3):356.
55. Kriegeskorte N (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science* 1:417–446.
56. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
57. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778.
58. Kingma DP, Salimans T, Welling M (2015) Variational dropout and the local reparameterization trick in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15. pp. 2575–2583.

59. Banks MS, Aslin RN, Letson RD (1975) Sensitive period for the development of human binocular vision. *Science* 190(4215):675–7.