Modeling Neural Variability in Deep Networks with Dropout

Xu Pan[1], **Ruben Coen-Cagli**[2,3,4], **and Odelia Schwartz**[1,4,*]

[1]Department of Computer Science, University of Miami, Coral Gables, FL, USA
[2]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, USA
[3]Dominick Purpura Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
[4]These authors jointly supervised this work: Ruben Coen-Cagli, Odelia Schwartz.
[*]odelia@cs.miami.edu

ABSTRACT

Convolutional neural networks (CNNs) have been used to model the biological visual system. Compared to other models, CNNs can better capture neural responses to natural stimuli. However, previous successes are limited to modeling mean responses; while another fundamental aspect of cortical activity, namely response variability, is ignored. How the CNN models capture neural variability properties remains unknown. Previous computational neuroscience studies showed that the response variability can have a functional role, and found that the correlation structure (especially noise correlation) influences the amount of information in the population code. However, CNN models are typically deterministic, so noise (and correlations) in CNN models have not been studied. In this study, we developed a CNN model of visual cortex that includes neural variability. The model includes Monte Carlo dropout, namely a random subset of units is silenced at each presentation of the input image, inducing variability in the model. We found that our model captured a wide-range of neural variability findings in electrophysiology experiments, including that response mean and variance scale together, noise correlations are small but positive on average, both evoked and spontaneous noise correlation are larger for neurons with similar tuning, and the noise covariance is low-dimensional. Further, we found that removing the correlation can boost trial-by-trial decoding performance in the CNN model.

Introduction

There has been an emerging trend using convolutional neural networks (CNNs) to model the biological visual system[1–9]. These studies have achieved great success in fitting neural responses to natural images and have revealed some representational similarities between visual areas of the brain and CNN models. However, previous studies only focused on the trial-averaged neural response while neglecting response variability across trials. Neural responses to an unchanged stimulus always fluctuate across trials and through time. This trial-by-trial variance can easily be larger than the explainable variance in regression models[3], suggesting that modeling neural variability should be at least as important as modeling mean response towards the goal of understanding neural codes.

Both empirical and theoretical studies indicate that neural variability can have an important role: the structure of the covariance—especially noise correlation, i.e. the trial-by-trial correlation coefficient between neuron pairs to a repeated same stimulus—influences the amount of information in the neural codes[10–12]. In neurophysiology studies, noise correlation can be modulated by various internal and external factors, such as attention[13–15], learning[15,16], adaptation[17], anesthesia[18], and surrounding context[19]. A general conclusion from these studies is that decrease of positive noise correlation is coupled with increase of behavior performance[13–17].

However, these studies used artificial stimuli instead of natural stimuli. This is partially because measuring information in recordings and models is straightforward for parametric inputs, such as orientation of sinusoidal grating. Thus there is a great interest in studying the role of noise correlation with natural stimuli, using models that can extract information from natural images.

CNNs have been very successful in processing natural images and performing complex tasks, such as image classification[20], object detection[21], and segmentation[22], with close to human-level performance on some tasks (though see also 23–25). So CNNs have great potential for modeling neural variability with natural inputs. However, the most common use of CNNs is deterministic (i.e. artificial neural responses do not fluctuate for an unchanged input). To introduce variability into artificial neural responses, we chose to use a class of neural networks named Bayesian neural networks[26]. Bayesian neural networks use Bayesian estimation to obtain the posterior distributions of network parameters (connection weights and biases), and then use them to infer a probability distribution over their outputs. Thus Bayesian neural networks can output both predictions and the uncertainty of the prediction[26]. If a sampling-based inference method is used, each independent sample can be seen as the response of the artificial neural population during an independent trial (Fig. 1B). Recent studies showed that the dropout layer, an existing neural networks tool which was originally designed for overcoming overfitting and filter coadaptation problems[27,28], has a variational inference interpretation that can make a non-Bayesian neural network Bayesian[29–32]. To obtain the inferred distribution of the network outputs, artificial neural activations are dropped (i.e. set to zero) in the testing phase for a random subset of units. This method is different from the traditional usage of dropout only during the training phase as a regularization technique, and is named Monte Carlo dropout[29].

One advantage of using a Bayesian neural network with dropout to model neural variability is that it is naturally consistent with the neural sampling hypothesis[33,34]. According to this hypothesis the brain approximates Bayesian inference, and neural responses represent samples from the inferred probability distribution; thus neural variability represents uncertainty. Empirical observations about neural variability and correlations can be captured by the uncertainty of inferences in Bayesian models[35–38]. Despite the good theoretical properties of Bayesian neural networks, it is unknown if they can capture empirical neural data.

In this study, we used Bayesian CNNs implemented with Monte Carlo dropout to test if CNNs can capture empirical observations on neural variance and covariance. We found that our model captured a wide-range of neural variability findings in electrophysiology experiments, including that response mean and variance scale together, noise correlations are small but positive on average, both evoked and spontaneous noise correlation are larger for neurons with similar tuning, and the noise covariance is low-dimensional. We also found that when the network is not fully trained, modulating dropout rate in the testing phase has a similar effect as attention on neural variability structure. Finally, we found that removing the correlation can boost trial-by-trial decoding performance in the CNN model.

Results

We built a 6-layer CNN and used the Cifar10 natural image dataset for training and testing (Fig. 1A). The channel numbers in the 6 convolutional layers are 64,64,128,128,256, and 256, which are also neuron numbers in the following analysis. The center neurons of each feature map after ReLU were used to model neural responses. The model has dropout layers with a dropout rate 0.5, but with a modification according to Monte Carlo dropout that drops activations in both the training and testing phase. Due to dropout in the testing phase, the feedforward computations are variable even when an identical stimulus is presented (Fig. 1B). The networks were trained for 100 epochs by the Adam optimizer with learning rate 0.0001 and L2-regularization coefficients 0.0001 (see online Methods for more details).

Our model achieved 82.88% accuracy on the Cifar10 testing set (Supplementary Fig. 1). Note that our model is not aimed to compete with benchmark models on classification accuracy. Rather, our goal is to construct a simple neural network model that is able to capture biological observations on neural variability while having a good classification accuracy.

The results are arranged as follows. We first show that the artificial neural variability has similar properties as observed in neurophysiology studies (Fig. 2, 3). Then we analyze the dimensionality of the artificial neural covariance and compare the effects of changing test-time dropout rate to visual attention (Fig. 4, 5). Finally, we study the role of noise correlations on the object classification task (Fig. 6).

CNN can capture basic properties of single-neuron variability

Although our main focus is covariability, as a sanity check, we first tested if the model can reproduce basic aspects of single-neuron variability. We found that, similar to cortical data[39], activation variance increases with the mean activation approximately monotonically (Fig. 2A).

A common way to characterize how much neural activity fluctuates compared to its mean is the Fano factor, i.e. variance divided by mean. The Fano factor is 1 for a Poisson process, which is the simplest statistical description of variability in the visual cortex, although cortical Fano factors often deviate from 1[18,39]. We found a similar order of magnitude for the Fano factors obtained in the CNN model (Fig. 2). The variabilities are inherent in our CNN model with dropout. We didn't heavily tune the model to match the amount of variability to the brain; instead we chose a default dropout rate of 0.5 which is a common value used in CNN models. To match empirical variability quantitatively, other Fano factor values can be obtained by changing the training dropout rate (Supplementary Fig. 2) and L2-regularization coefficient (Supplementary Fig. 3). After training, the Fano factor in the model is larger in deeper layers (Fig. 2B). This is consistent with some neurophysiology studies that showed higher areas in the sensory pathway have larger Fano factors[39,40]. This trend is also observed with different L2-regularization coefficients (Supplementary Fig. 3), but less obvious in a 12-layer CNN model (Supplementary Fig.4). During training, the Fano factor in the CNN model showed generally decreasing trends (Fig. 2C). Neurophysiology studies also showed that the Fano factor decreases after subjects learn the task[15,41,42]. This is expected in our model, because response variability presumably represents the uncertainty of inferences, and in Bayesian models training usually decreases the posterior variance of the inference. We find that the primary cause of the decreasing Fano factor in the CNN model is due to the L2-regularization prior (Supplementary Fig. 3), which penalizes large weights thus decreasing weights at the beginning of

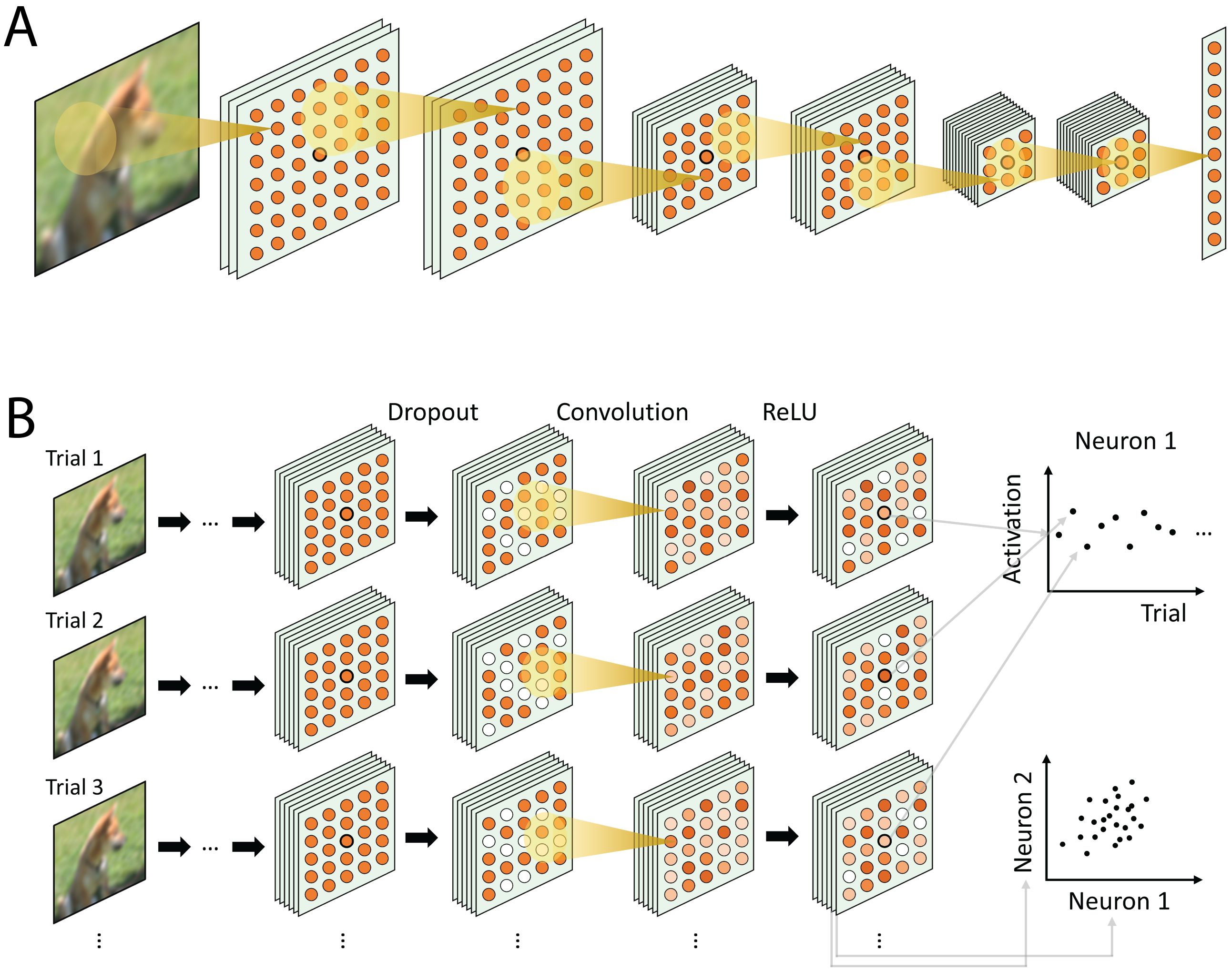

Figure 1. Neural network model and Monte Carlo dropout induced variability. (A) Schematic 6-layer neural network model. It contains widely used building blocks such as 2D convolution, ReLU, Batch Normalization and Max Pooling (see online Methods for more details). Center artificial neurons (thicker circle outline) in each feature map are used in the following analysis. (B) Illustration of sampling of the neural network. Dropout is used during both the training and testing phases. We treat each random dropout choice as a trial (a row in the diagram); thus each trial will elicit a unique activation pattern (last column, color indicates level of activation). An example of activations as a function of trials of one center artificial neuron is shown in the top right panel (pseudo data for illustrative purpose). Similarly, correlation between artificial neurons from different feature maps can be measured (bottom right panel).

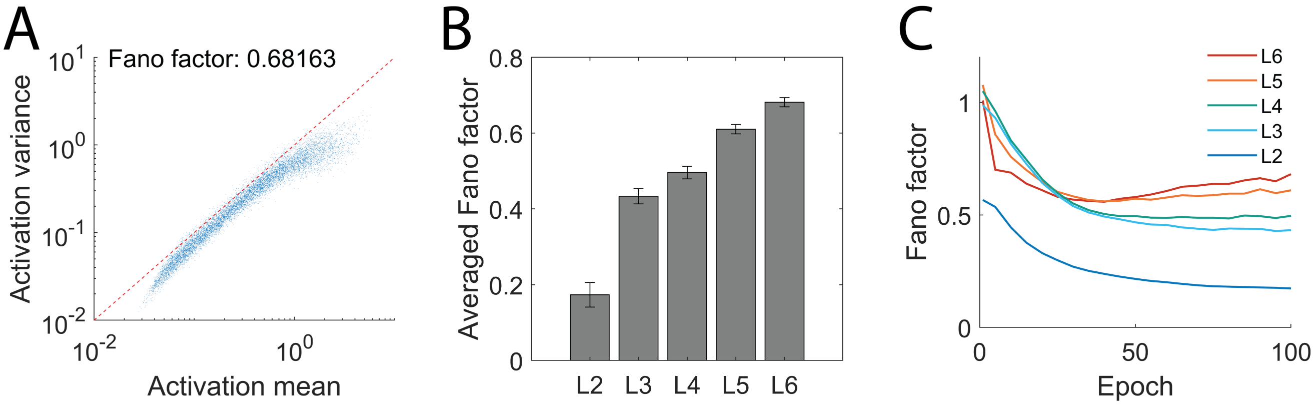

Figure 2. CNN captures basic properties of single-neuron variability. (A) Activation variance versus mean of layer 6 neurons. Each blue dot corresponds to one neuron and one stimulus. The geometric mean of the Fano factor is 0.68. The red dotted line corresponds to a Fano factor of 1. (B) Averaged Fano factors (geometric mean) of different layers. The errorbar indicates geometric standard error of the mean (SEM), i.e. the geometric Standard Deviation divided by the square root N (i.e. number of data points in A). (C) Fano factor in layer 2 (L2) to 6 (L6) changes during training.

the training (Supplementary Fig. 5). Larger L2-regularization coefficients induce a more rapid decrease of the Fano factor (Supplementary Fig. 3).

CNN can capture properties of covariability between neurons

In addition to variability in single neurons, variability shared between neurons, especially noise correlation, has been observed and widely studied in biological neural systems. We asked if the CNN model could capture properties of noise correlation in the visual cortex. In this study, we focused on variability between artificial neurons coding for different features at the same spatial location.

In our CNN model, noise correlations are small and positive on average at all layers (Fig. 3A), consistent with neurophysiology studies[44]. We further looked at how training affects noise correlation (Fig. 3B;C). For early layers (layers 2 and 3), noise correlation gradually increases (Fig. 3B). For late layers (layers 4, 5, and 6), noise correlation firstly increases, and then decreases after 15 epochs (Fig. 3B). The last layer showed the most drastic changes. In this later decreasing phase, testing accuracy increases while noise correlation decreases (Fig. 3C). This matches neurophysiology experiments that showed learning decreases noise correlation and increases behavioral performance[15,16]. The early rapid increase phase we observed in our model is unlikely to have biology implications, because an animal's brain is highly organized before training rather than random weighted connections as in the CNN model.

A striking aspect across the CNN layers is that the noise correlation is highest for the last layer of the network and has the most drastic change during training (Fig. 3A;B). This also held for a 12-layer CNN model (Supplementary Fig. 4). We found this observation depends on the L2-regularization coefficients. With smaller L2-regularization coefficients, noise correlation in the last layer ended with a relatively larger value (Supplementary Fig. 3). Additionally, we looked for other potential factors that cause the last layer to have the highest noise correlation. We found that the computations after the layer can influence the amount of its noise correlation (by influencing the gradient backpropagation that changes weights in early layers). Two factors are if there is a following dropout layer, and how many feature maps are in the following layer (Supplementary Fig. 6).

Neural correlations can be measured under different conditions: conditioned on one stimulus (noise correlation), conditioned on all stimuli, i.e. measuring the correlation over all trials (signal correlation) which represents tuning similarity, and conditioned on no stimulus, i.e. measuring the correlation when the stimulus is a uniform gray input (spontaneous correlation). These different measures are shown to be positively related in experiments[44–47]. We found a similar relationship in our model. Both noise correlation (Fig. 3D) and spontaneous correlation are positively related to signal correlation (see also Supplementary Fig.7 for a more complete exploration). Thus neuron pairs with large signal correlation also have large noise correlation and spontaneous correlation (Fig. 3E; see also Supplementary Fig.7) as observed in primary visual cortex (V1) data[35,43] (Fig. 3F).

Low dimensionality of covariability in CNN

Neurophysiology studies showed that neural variability when conditioned on the stimulus is low-dimensional[48–50]. The direction of maximal variance in the shared covariance matrix (corresponding to the dominant eigenmode, with the largest eigenvalue) can account for about 70% of the total shared variance[48]. Therefore, we studied the dimensionality of the representation in the

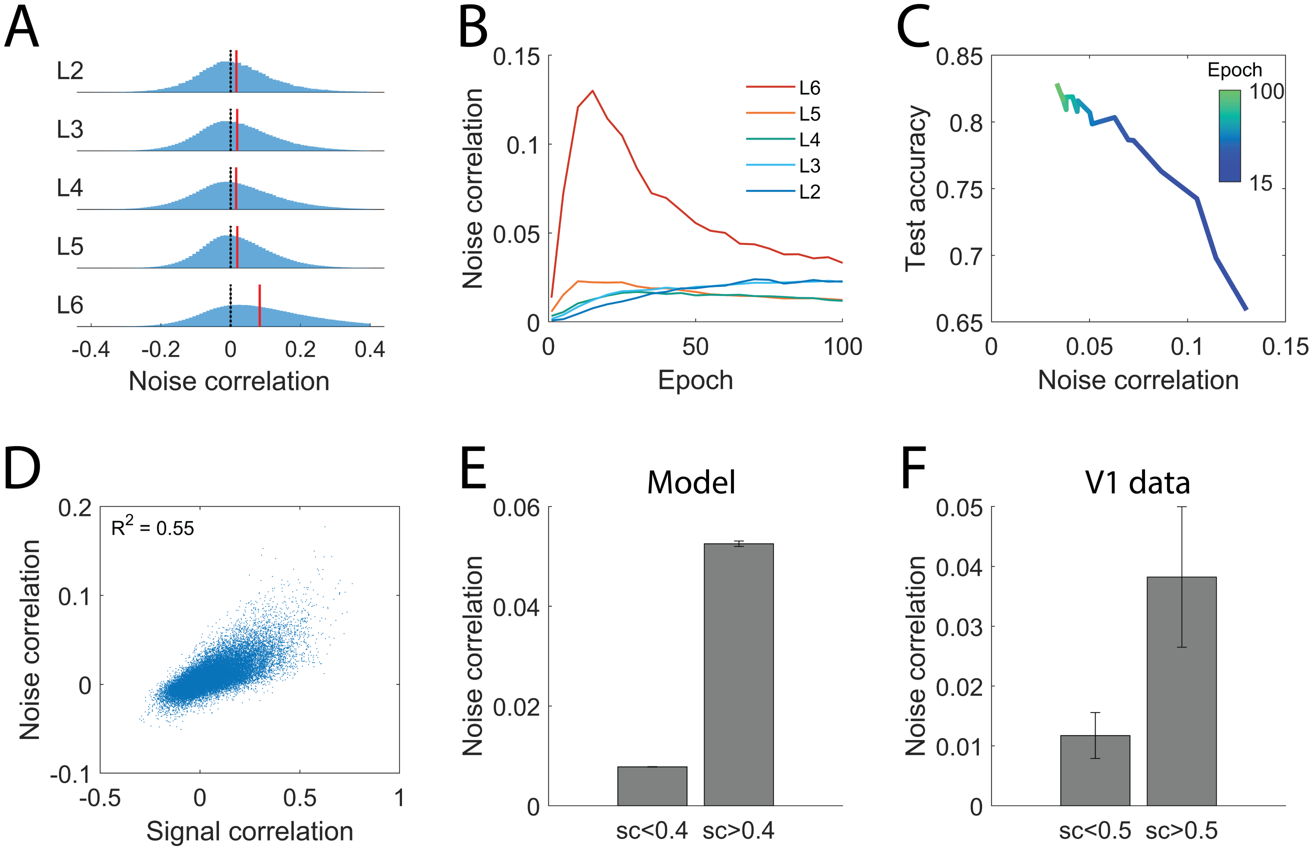

Figure 3. CNN captures properties of covariability between neurons (A) Histogram of the noise correlation in different layers at the end of the training. Black dotted line indicates 0 correlation and the red line indicates the mean. (B) Averaged noise correlation in different layers during training (C) Decrease of noise correlation in layer 6 during training is coupled with increase of model accuracy in later epochs. (D) Signal correlation is positively related to noise correlation and spontaneous correlation. Each data point is a neuron pair in layer 6 of the CNN. (E) Neuron pairs with larger (>0.4) signal correlation (sc) also have larger noise correlation and spontaneous correlation. Spontaneous correlation is measured with blank inputs to the model. (F) Similar observations are found in V1 (reproduced from 35,43). Error bars show SEM.

hidden layers of our model.

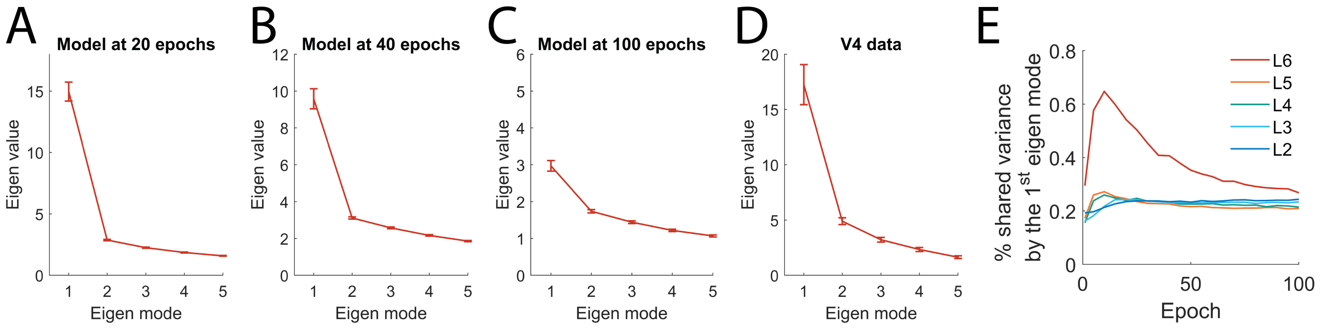

Figure 4. Low dimensionality of covariance in the CNN model versus neurophysiology studies. (A-C) The largest five eigenvalues (ordered from largest to smallest) of the shared component of the covariance matrix from layer 6 in the CNN model. Error bars show SEM. (D) Eigenvalues (corresponding to the variances) of the shared component of the covariance matrix from V4 data(reproduced from [28] [20]). (E) Percentage of the shared covariance explained by the dominant eigenmode in different layers (L2 - L6) at different epochs of training.

Similar to the methods used in previous studies, we partitioned the covariance matrix into independent and shared variability by factor analysis[48], then focused on the shared part of total covariance. To analyze the dimensionality, we eigen decomposed the shared component of the covariance matrix. We used the relative amount of the largest eigenvalue to other eigenvalues as an indication of dimensionality (the larger the percentage, the lower the dimensionality). We found that the largest eigenvalue dominates others, which indicates the variability is low dimensional in the CNN model (Fig. 4A;B;C). In other words, the first eigen mode can account for a large proportion of the total covariance. Similar eigenvalue spectrum is found in neurophysiology studies[14,49] (Fig. 4D).

We then studied how dimensionality changes with learning, to test if it reflects the trends seen in noise correlations (Fig 3). Intuitively, when correlations are large on average we can expect a low dimensionality. This is confirmed by Fig. 4D, where we observe the shared variance explained by the dominant eigenmode has a similar trend to the noise correlation as a function of the number of epochs (Fig. 3B). Layer 6 shows the most dominant eigenmode, accounting for 65% of the total shared variance which peaks at 10 epochs (Fig. 4E). Other layers have a moderate change of dimensionality during training, where the dominant eigenmode accounts for about 20% of the total shared variance, which is less pronounced than layer 6 but still a large number considering the total numbers of neurons in the analysis (64 in L2, 128 in L3 and L4, 256 in L5 and L6). We also found that the dominant eigenmode accounts for a large portion of the positive covariance, by reconstructing the covariance matrix from the corresponding eigenvector (Supplementary Fig. 8).

Neurophysiology studies showed that when an animal is attending to the stimulus, noise correlation decreases, especially through a reduction of the amplitude of the dominant eigenmode[15,48,49]. We considered the effect of testing time dropout rate on the noise correlation and the dimensionality of the covariance matrix as an analogy to attention. Conceptually, decreasing testing dropout rate will increase the number of artificial neurons used for the feedforward computation; since each of them is a source of variability, the dimensionality of variability of output artificial neurons could increase. We found that changing the dropout rate from 0.5 to 0.2 decreases noise correlation of L6 at early epochs (before 50 epoch) but not at later epochs (Fig. 5A). Decreasing the testing dropout rate also decreases the proportion of the largest eigenmode, especially at the early epochs of training (Fig. 5B). In neurophysiology studies, when the stimuli were attended to, only the dominant eigenmode showed a decrease in the magnitude (Fig. 5F). We found a similar pattern in layer 6 of the CNN model, especially at early epochs (Fig. 5 C;D;E). The pattern was less pronounced in the early layers of the CNN model (Supplementary Fig. 9).

The influence of the testing dropout rate decreasing the dimensionality of the shared covariance (relating to the attention effect) is more aligned with neurophysiology studies when the CNN is not fully trained. One possibility is that reducing testing dropout rate and extensive training have a similar role (analogous to learning and attention effects in the visual cortex[15]), i.e. they both increase performance and decrease noise correlations . But when the two factors are combined, the effect is not additive since the system already reached the limits by either of these two factors.

The alignment of reducing testing dropout rate and attention is not limited to noise correlation and dimensionality of the covariance matrix. We also found that reducing testing dropout rate in the CNN model can reduce the Fano factor (Supplementary Fig. 10), similar to neurophysiology studies reporting that attention decreases the Fano factor[14,15]. Behaviorally, attention increases subjects' task performance[14,15]. We found that during early epochs, reducing testing dropout rate can increase testing accuracy (Supplementary Fig. 11). Combined with the result shown in Figure 3C, we therefore found two factors that can affect

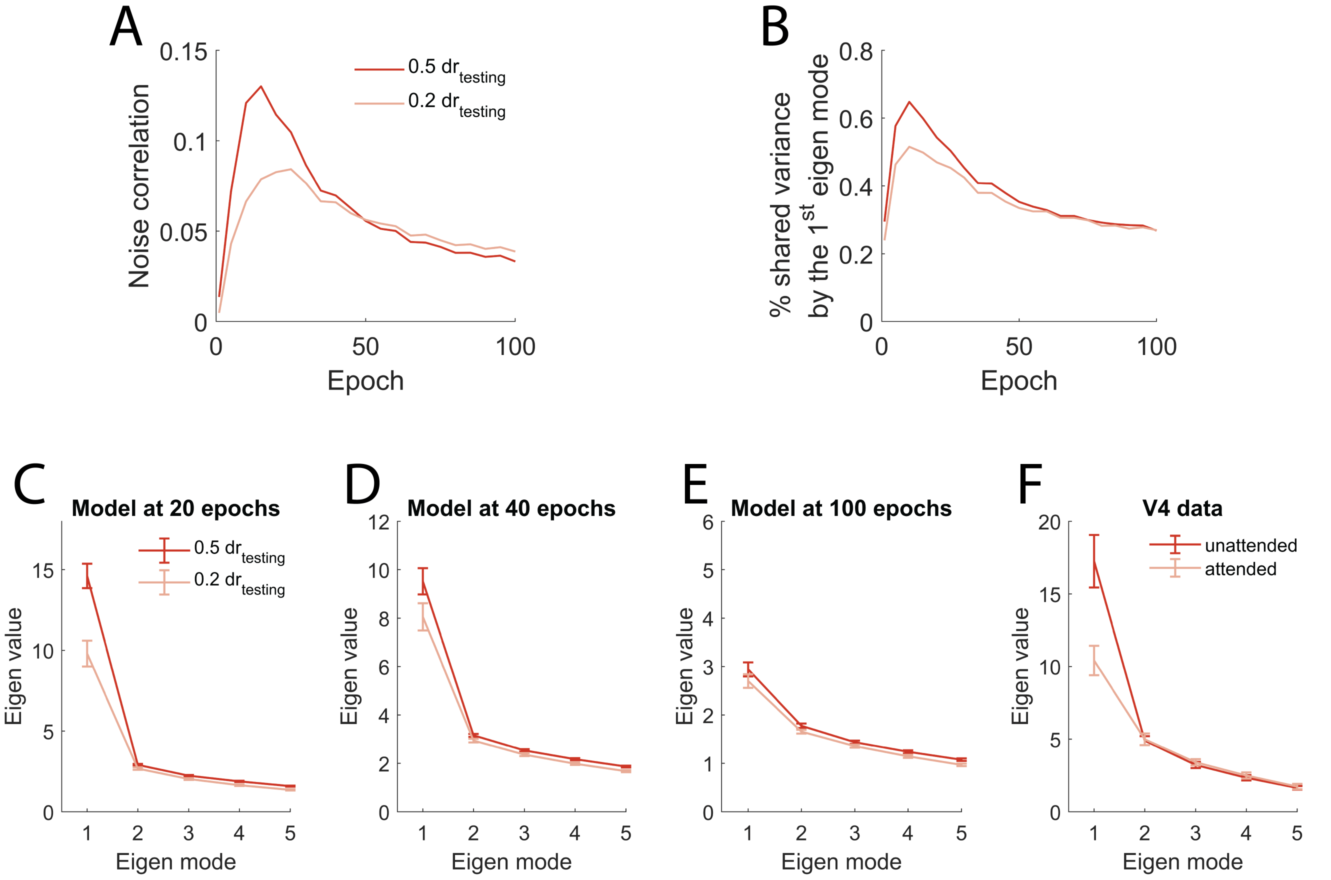

Figure 5. Modulating testing dropout rate reveals a similar change on the neural covariability as attention effect in animal studies. (A) Reducing testing dropout rate from 0.5 to 0.2 changes the percentage of the shared covariance explained by the first eigenmode. (B) Reducing testing dropout rate from 0.5 to 0.2 changes noise correlation. (C) The eigenvalues of the shared covariance matrix of the CNN layer 6 responses at 20 epochs. Error bars show SEM. (D, E) Same as (C), but measured at 40 and 100 epochs. (F) The eigenvalues of the shared covariance matrix of V4 are modulated by attention (reproduced from [14, 49]).

the classification accuracy and noise correlation: training and reducing testing dropout rate increase classification accuracy, while decreasing noise correlation.

Noise correlations in CNNs reduce classification performance

Noise correlations can impact the amount of information encoded in population codes[10–12, 51–54]. Theoretical studies showed that if noise correlation has a component that is aligned with the coding direction, the noise correlation is information-limiting[11], i.e. the amount of information saturates with increasing population size. Neurophysiology studies have observed this information-limiting effect in large neuron populations[55–57]. To study if noise correlation in CNN models is information-limiting, we stacked a SVM classifier after layer 6 (to replace the dense-layer classifier). The SVM classifiers were trained and tested under two conditions: leaving layer 6 unchanged (correlated), and permuting by assigning artificial neural activations random trial numbers to break the correlations (shuffle). As more neurons are used in the classifier, the SVM classifiers trained and tested on correlated activation show faster saturation of accuracy and discriminability index (i.e. d prime) than those trained and tested on the permuted activation (Fig. 6). This faster saturation on correlated activation is also seen during the training (Supplementary Fig.12). This indicates noise correlation in the CNN is harmful. This may be related to the information-limiting effects, because noise correlations in our model originate from the feedforward propagation of dropout “noise”[54], although we cannot rule out other factors, given that discriminability for the shuffled data also appears to saturate.

To further show that noise correlation can be harmful in the CNN, instead of training new SVM classifiers on top of the CNN model, we kept the CNN model unchanged (even without fine tuning) and either kept the layer 6 activation unchanged or shuffled. The single trial prediction accuracy (rather than the Monte Carlo sampling which averages across trials) increased

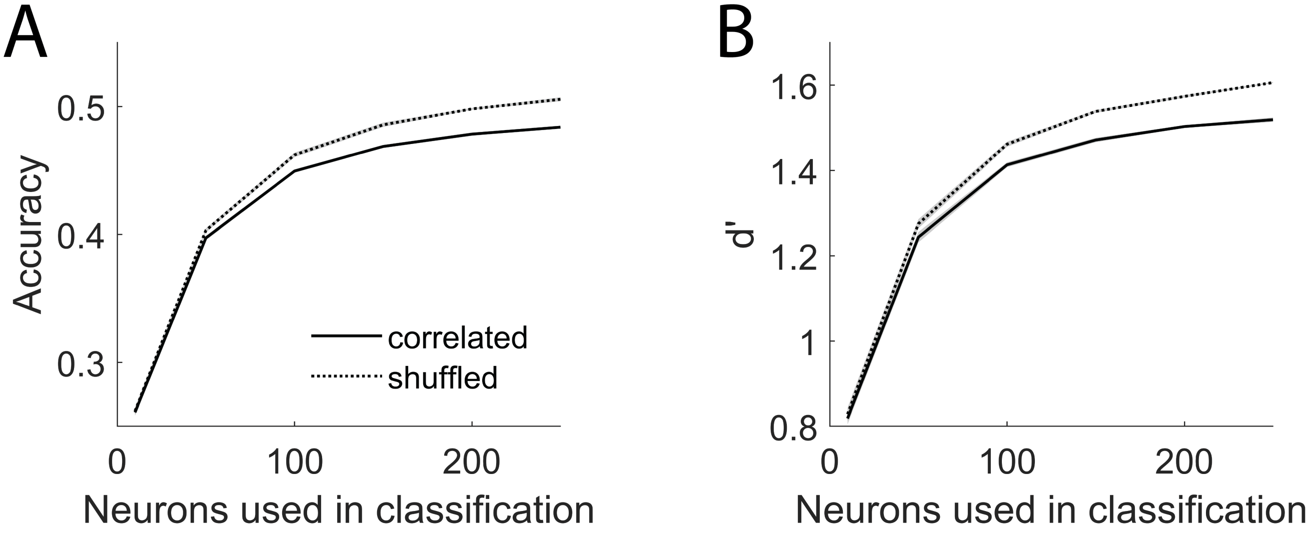

Figure 6. Decoder performance plateaued higher when correlation is eliminated (A) Accuracy of SVM classifiers trained on different numbers of correlated L6 center neurons(solid line) or permuted to break correlations (shuffled condition; dotted line). When more neurons are recruited, the shuffled condition plateaued at a higher accuracy.(B) Same as (A) but measuring d prime.

from 85.88% ± 1.56% to 88.38% ± 1.51% ($p < 10^{-8}$, by 6-fold cross validation and t-test) when layer 6 activation was permuted. This result further indicates that noise correlation in the CNN is harmful. It is surprising that we can boost the performance of a fully trained neural network by only permuting activations without any retraining or fine tuning. This result is counterintuitive, since the decoder is trained on correlated data but performs better on shuffled data. We think this observation results from the balance of two opposing effects. First, if the decoder is optimal for correlated data but suboptimal for shuffled data, performance on shuffled data should be worse[10]. Second, if noise correlations partly align with the coding direction, shuffling should increase performance. Our result indicates that either the dense layer decoder, even if trained on correlated data, is not optimal; or the noise correlations are strongly aligned with the coding direction, such that the optimal decoders for correlated and shuffled data are similar, and so the beneficial effect of shuffling dominates.

Discussion

Without heavily tuning hyperparameters of the model, our Bayesian CNN model with the Monte Carlo dropout implementation captured a variety of properties of neural variability found in the brain. These properties include that the variance and the mean of responses scale together (Fig. 2), both evoked and spontaneous noise correlation are correlated with signal correlation (Fig. 3), and population activity is low dimensional (Fig. 4). We further found that reducing the testing dropout rate has a similar effect on noise correlation and covariance dimensionality as attention in neurophysiology studies (Fig. 5). Furthermore, we found that noise correlation in CNNs harms the representation, and removing noise correlations by trial-shuffling yields better classification accuracy (Fig. 6).

Using CNN as a brain model

CNNs have shown great success in solving Computer Vision tasks[20–22]. They have also been recognized as powerful computational models for visual processing[1–7,9], and indeed are inspired by the hierarchical structure and linear-nonlinear computations in the brain. Compared to other models, representations in CNN models are more similar to cortical representations, and can better fit single neuron responses to natural images (but see also a discussion on model failures in 23–25). Furthermore, CNN based generative algorithms can synthesize images that drive a specific neural population response to a target value[58,59]. However, despite this trend of modeling neural responses with the CNN, the main focus has been on the trial-averaged responses of single neurons, and no study has explored the potential to model the variability of neural responses and how it is shared between neurons. Without taking the trial-by-trial variance into account, regression models can only explain a small portion of the total variance[3]. Modeling neural variability can help explain total neural variance.

Although existing CNN models of biological visual processing are typically deterministic and thus lack variability, not all types of CNN designs are deterministic. In Bayesian neural networks, the posterior distribution of network parameters is estimated, so that during inference, artificial neural responses are represented as distributions. Then, trial-by-trial variability can be defined by using Monte Carlo sampling based inference methods. In addition to introducing trial-by-trial variability in

CNNs, this approach also gives them the ability to compute uncertainty. Therefore, Bayesian CNNs with Monte Carlo sampling could be used to extend and test predictions of the neural sampling theory of uncertainty representation in the brain[33–36]. In this study, we used an existing deep neural networks technique, namely dropout, and its relatively new Bayesian interpretation[29–32]. Our model, though simple, can already capture a wide-range of properties of variability in the brain.

Although our results on variability are qualitative, a promising future direction is to fit Bayesian CNN models to both the mean and variance of the neural responses to stimuli. This goal may require a more capable Bayesian model than the Monte Carlo dropout; in other words, a more flexible variational distribution. There is an argument that Monte Carlo dropout cannot estimate uncertainty effectively[60]. This is due to the fact that the variational distribution of weights, if the dropout rate is fixed, is a one-parameter discrete distribution (Bernoulli distribution) for which the variance and mean are coupled during updating the only one parameter.

Modeling variability and uncertainty representation in the brain

An advantage of our CNN model for neural variability is that it combines a number of properties: it is hierarchical, uses natural stimuli, is trained on a supervised task, and has a Bayesian interpretation.

Mechanistic models showed that recurrent dynamics can reveal cortical-like neural variability[61,62] and covariability[63–66]. Some of these works also showed that network dynamics can be interpreted as approximating Markov chain Monte Carlo sampling to generate samples of the posterior distribution in Bayesian models. However, these studies did not address natural images and did not relate the structure of variability to performance on a complex visual task, as we did here. To our knowledge, our work is the first to study the effects of noise correlations on object classification performance in a sampling based Bayesian neural network. Our finding that noise correlations impair classification performance can be understood by the observation that input noise that propagates through the feedforward connections limits population information[54]. Since our CNN model can be considered as having multiplicative Bernoulli noise in the inputs of each layer, the noise correlation observed in our model is likely to be information-limiting, consistent with our results (Fig. 6). Note that our model does not imply a mechanistic implementation of dropout in the nervous system. The corresponding variability for instance might result from recurrent dynamics in neural circuits. However, a recent study suggested that the brain may indeed use a similar representation to dropout in deep learning[67].

A different approach to explaining neural responses to natural inputs involves developing generative models of image statistics, and comparing neural responses to the inferences. In particular, image statistics models based on the Gaussian scale mixture have reproduced many aspects of neural variability and noise correlations[35,36,63]. One advantage of our CNN model is that it is trained on natural images for classification, and unlike previous models of neural variability, is therefore goal-oriented and performs a natural image task. This opens up a new type of model and neural variability study direction that explores the role of neural correlation in complex tasks. Another advantage of supervised training of the model on a visual task is that all the model parameters are optimized end-to-end, and therefore could be more easily adjusted to other training conditions.

Similar to GSM-based models of variability[35,36,63], our model could also relate neural variability to the representation of uncertainty via sampling, consistent with the neural sampling hypothesis. The Monte Carlo dropout layers give the CNN model a Bayesian interpretation[29,30]. The posterior distribution is estimated by sampling. Therefore, to test if the Fano factor in our model is related to uncertainty as predicted by the neural sampling hypothesis, we ran two simulations in the testing phase: (i) modulating the input contrast; (ii) adding Gaussian noise to the input (Supplementary Fig. 13). Low contrast and high input noise presumably mean high model uncertainty. We found that higher input noise and contrast leads to higher Fano factor (Supplementary Fig. 13). The result of adding noise is consistent with our hypothesis that the Fano factor indicates model uncertainty. However, modulating contrast does not match the uncertainty expectation. This might be because the network lacks computations such as local divisive normalization and recurrent connections[36,63,68–70]. Normalization has been shown to capture a number of effects relating to neural variability and contrast, and arises as part of the inference in the GSM model[36,69].

Attention effects on neural variability

Attention is a sophisticated psychological phenomena that involves interactions in multiple brain areas, especially top-down modulation[71,72]. Our model is not aimed to reveal a biological mechanism of attention and it is unlikely to do so since it is purely feedforward. However, the model could capture some empirical consequences of visual spatial attention. Studies showed that task-relevant cortical areas have more activity, and that neurons increase their firing rate[73,74] and probably excitability[75,76], when attention is involved. Our dropout model has a subtle effect on neural excitability: lower dropout rates allow more neurons to be active; but for each neuron, the activation is scaled down. The combined effect is that reducing the dropout rate slightly increases individual artificial neural activation (if they are not dropped out and have non-zero activation) (Supplementary Fig. 10). If there is a biological correspondence, it could be that during the attended state synaptic connections are more reliable (more neurons are active), but the synaptic weights are tuned down (less gain). As shown in the results, dropout in our model captured some attention effects on neural variability. Studies showed that attention reduces the Fano factor and neural correlation[14,15]. We observed both these effects in our model (Fig 5, Supplementary Fig. 10).

Conclusion

While CNN models have been influential in modeling cortical neural responses, they have only focused on capturing mean neural responses. Our study proposed that the CNN model with variability is a better model of visual cortical neurons, and could account for a range of cortical variability data. This provides a much-needed framework to understand the relation between neural variability and performance in complex natural tasks.

Methods

Neural network model

Exact Bayesian models usually have intractable posterior problems; so does the Bayesian Neural Network which could have millions of parameters. Two main routes to solve this problem are Markov chain Monte Carlo sampling[77] and variational inference[78]. Variational inference uses a parametric variational distribution to approximate the posterior distribution. The objective of learning is then to minimize the divergence between the variational distribution and the posterior distribution. Recent studies showed that Dropout can be interpreted as a variational inference method that makes any neural network model bayesian[29–31]. Model neuron activations are dropped in both training and testing phase. In the Bayesian interpretation, this procedure is Monte Carlo sampling from a Bernoulli variational distribution, so that it is named Monte Carlo dropout[29]. Furthermore, applying L2 regularization loss for kernels during training is equivalent to applying a Gaussian prior over weights under the Bayesian interpretation[29].

The neural network model is implemented using the TensorFlow framework. The main results are based on a 6-layer network trained and tested on the Cifar10 image classification dataset (32 x 32 pixels, 10 classes). The channel numbers of 6 convolutional layers are 64,64,128,128,256,256. The kernel weights are initialized using the Xavier uniform method. The kernel size of all convolutional layers is 3 by 3 with “same” padding. The network includes 2 by 2 max pooling layers after the second, fourth and last convolutional layer. Each convolutional layer is followed by ReLU activation, batch normalization and dropout. If there is a max pooling layer, the dropout layer is placed after the max pooling; this design can better retain the model average[79,80]. Following previous dropout studies, we do not add a dropout layer after the last convolution layer. The networks are trained for 100 epochs by the Adam optimizer with learning rate 0.0001 and L2-regularization coefficients 0.0001. We also consider how different L2-regularization coefficients affect neural variability in the CNN model (Supplementary Fig. 5). Input images are rescaled so that pixel values are from -1 to 1. A simple data augmentation is applied to the training set (15 degree random rotation, 10% random shift, and random horizontal flip.).

The dropout layer is modified so that it drops activations at both the training and testing phase according to Monte Carlo dropout[29]. Only with this modification, the activations in the CNN can have variability to unchanged inputs. As dropout drops activations at both training and testing phase, we can use different dropout rates at each of the two phases. We chose a common dropout rate, 0.5, in most of the experiments.

To exclude the correlations due to overlapping receptive fields with identical feature tuning, we only use the center neuron of each feature map in the analysis. This also applies to the SVM experiments, in which only the center neurons in layer 6 are used in the SVM classifiers.

Neural variability analysis

Intermediate layer activations are used for the neural variability analysis. We specifically take the neural activations after the ReLU operation to account for non-negative neural activity. Because we do not use responses immediately after the dropout layer, the variability is not a direct result of multiplicative Bernoulli noise. A better way to think about it is that the variability of each layer’s activation is due to the multiplicative Bernoulli noise to its input. A biological view of this can be that the unstable synapses failed to activate a neuron.

Note that the first dropout layer is after the first convolutional layer, so the activations of layer 1 neurons are not variable and not included in the analysis.

We focus on two types of variability: single neuron variability (Fano factor) and covariability between neuron pairs. To compute variability measures, 100 random selected images from the test set are used. Each image is passed to the model 1000 times. Variability measures are computed stimulus-wise, and then averaged (geometric mean when computing mean Fano factor). When analyzing dimensionality, to keep the artificial neuron population size same across layers, we randomly include 100 neurons (except for the second layer which only has 64 neurons) in the analysis. Because some stimuli do not induce a non-zero activation of some neurons, we exclude neurons that have no activity in more than 90% of the total trials. This thresholding process is after 100 neurons are randomly selected.

Noise correlation is the stimulus-wise averaged Pearson correlation coefficient between neuron pairs. Signal correlation is the Pearson correlation coefficient between neuron pairs, regardless of stimulus. Spontaneous correlation is the Pearson correlation coefficient between neuron pairs, when the input is an all-zero image.

Dimensionality analysis

We use the factor analysis method (Matlab's implementation) to estimate the dimensionality of neural covariance[48,49,81]. Factor analysis partitions variance into a shared component and an independent component, and thus best serves our purpose of characterizing the neural covariance structure. Since the active neuron number varies across stimuli, we choose a constant loading factor number, 8. If the algorithm does not converge on a stimulus (rare occurrence), we exclude this stimulus in the analysis. The shared covariance matrix is constructed from the outer product of the loading factors. The shared covariance matrix is eigen decomposed, where the proportion of eigenvalues can be interpreted as the percentage of shared covariance explained by that eigenmode.

Information saturation analysis

We use two methods to test if noise correlation leads to faster performance saturation. First, we stack a SVM classifier on the top of layer 6 of the CNN model. Center neurons in layer 6 are used as the inputs of the SVM model. Three trials are run, with different random subsets of neurons selected. The number of neurons are chosen to be 10, 50, 100, 150, 200, and 250, to reveal how performance changes according to increasing population size. We either permute layer 6 activations to break trial-by-trial correlation or keep layer 6 activations unchanged. To train the SVM classifier, we randomly select 10000 out of 50000 images from the training set. Each image is repeated 300 times, which is slightly larger than the max neuron number so that it is sufficient to break any correlation structure. The performance of the SVM classifier is tested on 10000 images of the testing set with the same repetition process. If a SVM model is trained on permuted activation, then it is tested on permuted activation. If a SVM model is trained on correlated activation, then it is tested on correlated activation. Discriminability index (d prime) is calculated in one-versus-all fashion for each class, and then averaged to get a scalar value for each model: $d' = \frac{1}{10}\sum_{i=1}^{10} \frac{|\mu_i^i - \mu_{\Omega-i}^i|}{\sqrt{(\sigma_i^{i\,2} - \sigma_{\Omega-i}^{i}{}^{2})/2}}$, where μ_i^j is the mean of the predicted probability of class i with true label j; $\sigma_i^{j\,2}$ is the variance of the predicted probability of class i with true label j; $\Omega - i$ is a set of all samples except those with true label i. Note that only center neurons are used in the SVM model. This causes the SVM models to have a lower accuracy than the full CNN model which has a dense layer classifier.

Second, we test if only permuting the layer 6 activation and keeping everything else unchanged can boost performance. Unlike the SVM method, the model (dense layer classifier) is trained on correlated activation, but tested on permuted activation. In this simulation, all the neurons of layer 6 are permuted and used in the analysis. Each image from the testing set is repeated 1000 times. The accuracy we measure is trial-wise accuracy, rather than based on the averaged output of 1000 repetitions. We did a 6-fold cross validation by dividing the train and test set in 6 ways and trained 6 CNN models. The final result shown in the main text is the averaged accuracy with double-side paired t-test. This is different from the original Monte-Carlo dropout method which predicts based on the trial average, but we think trial-wise accuracy has better biology implication where the system needs to predict based on the current activation.

References

1. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4, DOI: 10.3389/neuro.06.004.2008 (2008).

2. Cadieu, C. F. *et al.* Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput. Biol.* **10**, 1003963, DOI: 10.1371/journal.pcbi.1003963 (2014). 1406.3284.

3. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897, DOI: 10.1371/journal.pcbi.1006897 (2019).

4. Kindel, W. F., Christensen, E. D. & Zylberberg, J. Using deep learning to probe the neural code for images in primary visual cortex. *J. Vis.* **19**, 1–12, DOI: 10.1167/19.4.29 (2019). 1706.06208.

5. Kubilius, J. *et al.* Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. Tech. Rep. (2019).

6. Deza, A., Liao, Q., Banburski, A. & Poggio, T. Hierarchically Compositional Tasks and Deep Convolutional Networks. (2020). 2006.13915.

7. Wallis, T. S. *et al.* A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *J. Vis.* **17**, 1–29, DOI: 10.1167/17.12.5 (2017).

8. Mcintosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S. & Baccus, S. A. Deep Learning Models of the Retinal Response to Natural Scenes. Tech. Rep. (2016).

9. Abbasi-Asl, R. *et al.* The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv* 465534, DOI: 10.1101/465534 (2018).

10. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation, DOI: 10.1038/nrn1888 (2006).

11. Moreno-Bote, R. *et al.* Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417, DOI: 10.1038/nn.3807 (2014).

12. Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. Correlations and Neuronal Population Information. *Annu. Rev. Neurosci.* **39**, 237–256, DOI: 10.1146/annurev-neuro-070815-013851 (2016).

13. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron* **63**, 879–888, DOI: 10.1016/j.neuron.2009.09.013 (2009).

14. Cohen, M. R. & Maunsell, J. H. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594–1600, DOI: 10.1038/nn.2439 (2009).

15. Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J. & Cohen, M. R. Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463–465, DOI: 10.1126/science.aao0284 (2018).

16. Gu, Y. *et al.* Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750–761, DOI: 10.1016/j.neuron.2011.06.015 (2011).

17. Gutnisky, D. A. & Dragoi, V. Adaptive coding of visual information in neural populations. *Nature* **452**, 220–224, DOI: 10.1038/nature06563 (2008).

18. Ecker, A. S. *et al.* State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248, DOI: 10.1016/j.neuron.2014.02.006 (2014).

19. Snyder, A. C., Morais, M. J., Kohn, A. & Smith, M. A. Correlations in v1 are reduced by stimulation outside the receptive field. *J. Neurosci.* **34**, 11222–11227 (2014).

20. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Tech. Rep. (2012).

21. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587, DOI: 10.1109/CVPR.2014.81 (IEEE Computer Society, 2014). 1311.2524.

22. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *IEEE Transactions on Pattern Analysis Mach. Intell.* **42**, 386–397, DOI: 10.1109/TPAMI.2018.2844175 (2020). 1703.06870.

23. Turner, M. H., Sanchez Giraldo, L. G., Schwartz, O. & Rieke, F. Stimulus- and goal-oriented frameworks for understanding natural vision, DOI: 10.1038/s41593-018-0284-0 (2019).

24. Geirhos, R. *et al.* Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *CoRR* **abs/1811.12231** (2018). 1811.12231.

25. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015). 1412.6572.

26. Tishby, N., Levin, E. & Solla, S. A. Consistent inference of probabilities in layered networks: Predictions and generalization. In *IJCNN Int Jt Conf Neural Network*, 403–409, DOI: 10.1109/ijcnn.1989.118274 (Publ by IEEE, 1989).

27. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Tech. Rep. 56 (2014).

28. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. (2012). 1207.0580.

29. Gal, Y. & Zoubin, G. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (2016).

30. Gal, Y. & Ghahramani, Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. (2015). 1506.02158.

31. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick. Tech. Rep. (2015).

32. Maeda, S. A bayesian encourages dropout. *CoRR* **abs/1412.7003** (2014). 1412.7003.

33. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations, DOI: 10.1016/j.tics.2010.01.003 (2010).

34. Hoyer, P. O. & Hyvärinen, A. Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior. Tech. Rep. (2002).

35. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron* **92**, 530–543, DOI: 10.1016/j.neuron.2016.09.038 (2016).

36. Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nat. Commun.* **12**, 1–11 (2021).

37. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* **90**, 649–660, DOI: 10.1016/j.neuron.2016.03.020 (2016).

38. Lange, R. D., Shivkumar, S., Chattoraj, A. & Haefner, R. M. Bayesian Encoding and Decoding as Distinct Perspectives on Neural Coding. *bioRxiv* 2020.10.14.339770, DOI: 10.1101/2020.10.14.339770 (2021).

39. Goris, R. L., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865, DOI: 10.1038/nn.3711 (2014).

40. Poland, E., Donner, T. H., Müller, K. M., Leopold, D. A. & Wilke, M. Thalamus exhibits less sensory variability quenching than cortex. *Sci. Reports* **9**, 1–12, DOI: 10.1038/s41598-019-43934-9 (2019).

41. Qi, X. L. & Constantinidis, C. Variability of prefrontal neuronal discharges before and after training in a working memory task. *PLoS ONE* **7**, e41053, DOI: 10.1371/journal.pone.0041053 (2012).

42. Qi, X. L. & Constantinidis, C. Neural changes after training to perform cognitive tasks, DOI: 10.1016/j.bbr.2012.12.017 (2013).

43. Ecker, A. S. *et al.* Decorrelated neuronal firing in cortical microcircuits. *Science* **327**, 584–587, DOI: 10.1126/science.1179867 (2010).

44. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations, DOI: 10.1038/nn.2842 (2011).

45. Schulz, D. P., Sahani, M. & Carandini, M. Five key factors determining pairwise correlations in visual cortex. *J. Neurophysiol.* **114**, 1022–1033, DOI: 10.1152/jn.00094.2015 (2015).

46. Kohn, A. & Smith, M. A. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J. Neurosci.* **25**, 3661–3673 (2005).

47. Smith, M. A. & Kohn, A. Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.* **28**, 12591–12603 (2008).

48. Williamson, R. C. *et al.* Scaling Properties of Dimensionality Reduction for Neural Populations and Network Models. *PLoS Comput. Biol.* **12**, 1005141, DOI: 10.1371/journal.pcbi.1005141 (2016).

49. Huang, C. *et al.* Circuit Models of Low-Dimensional Shared Variability in Cortical Networks. *Neuron* **101**, 337–348.e4, DOI: 10.1016/j.neuron.2018.11.034 (2019).

50. Mendels, O. P. & Shamir, M. Relating the structure of noise correlations in macaque primary visual cortex to decoder performance. *Front. Comput. Neurosci.* **12**, 12, DOI: 10.3389/fncom.2018.00012 (2018).

51. Sompolinsky, H., Yoon, H., Kang, K. & Shamir, M. Population coding in neuronal systems with correlated noise. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **64**, 11, DOI: 10.1103/PhysRevE.64.051904 (2001).

52. Shamir, M. & Sompolinsky, H. Implications of neuronal diversity on population coding. *Neural Comput.* **18**, 1951–1986, DOI: 10.1162/neco.2006.18.8.1951 (2006).

53. Ecker, A. S., Berens, P., Tolias, A. S. & Bethge, M. The effect of noise correlations in populations of diversely tuned neurons. *J. Neurosci.* **31**, 14272–14283, DOI: 10.1523/JNEUROSCI.2539-11.2011 (2011).

54. Kanitscheider, I., Coen-Cagli, R. & Pouget, A. Origin of information-limiting noise correlations. *Proc. Natl. Acad. Sci. United States Am.* **112**, E6973–E6982, DOI: 10.1073/pnas.1508738112 (2015).

55. Rumyantsev, O. I. *et al.* Fundamental bounds on the fidelity of sensory cortical coding. *Nature* **580**, 100–105, DOI: 10.1038/s41586-020-2130-2 (2020).

56. Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Information-limiting correlations in large neural populations. *J. Neurosci.* **40**, 1668–1678, DOI: 10.1523/JNEUROSCI.2072-19.2019 (2020).

57. Kafashan, M. *et al.* Scaling of sensory information in large neural populations shows signatures of information-limiting correlations. *Nat. communications* **12**, 1–16 (2021).

58. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, DOI: 10.1126/science.aav9436 (2019).

59. Ponce, C. R. *et al.* Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* **177**, 999–1009.e10, DOI: 10.1016/j.cell.2019.04.005 (2019).

60. Osband, I. Risk versus uncertainty in deep learning : Bayes , bootstrap and the dangers of dropout (NIPS Bayesian Deep Learning Workshop, 2016).

61. Van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724–1726, DOI: 10.1126/science.274.5293.1724 (1996).

62. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* **7**, 1002211, DOI: 10.1371/journal.pcbi.1002211 (2011).

63. Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nat. Neurosci.* **23**, 1138–1149, DOI: 10.1038/s41593-020-0671-1 (2020).

64. Kanashiro, T., Ocker, G. K., Cohen, M. R. & Doiron, B. Attentional modulation of neuronal variability in circuit models of cortex. *eLife* **6**, DOI: 10.7554/eLife.23978 (2017).

65. Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M. & Miller, K. D. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron* **98**, 846–860.e5, DOI: 10.1016/j.neuron.2018.04.017 (2018).

66. Landau, I. D. & Sompolinsky, H. Coherent chaos in a recurrent neural network with structured connectivity. *PLoS Comput. Biol.* **14**, e1006309, DOI: 10.1371/journal.pcbi.1006309 (2018).

67. Yoshida, T. & Ohki, K. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nat. communications* **11**, 1–19 (2020).

68. Giraldo, L. G. S. & Schwartz, O. Integrating Flexible Normalization into Mid-Level Representations of Deep Convolutional Neural Networks. *Neural computation* **31**, 2138–2176 (2018). 1806.01823.

69. Coen-Cagli, R. & Solomon, S. S. Relating Divisive Normalization to Neuronal Response Variability. *J. Neurosci.* **39**, 7344–7356, DOI: 10.1523/JNEUROSCI.0126-19.2019 (2019).

70. Pan, X., Kartal, E., Gonzalo, L., Giraldo, S. & Schwartz, O. Brain-inspired weighted normalization for CNN image classification. *ICLR Brian2AI workshop* (2021).

71. Harris, K. D. & Thiele, A. Cortical state and attention, DOI: 10.1038/nrn3084 (2011).

72. Noudoost, B., Chang, M. H., Steinmetz, N. A. & Moore, T. Top-down control of visual attention, DOI: 10.1016/j.conb.2010.02.003 (2010).

73. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron* **55**, 131–141, DOI: 10.1016/j.neuron.2007.06.018 (2007).

74. Luck, S. J., Chelazzi, L., Hillyard, S. A. & Desimone, R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* **77**, 24–42, DOI: 10.1152/jn.1997.77.1.24 (1997).

75. Bestmann, S., Ruff, C. C., Blakemore, C., Driver, J. & Thilo, K. V. Spatial Attention Changes Excitability of Human Visual Cortex to Direct Stimulation. *Curr. Biol.* **17**, 134–139, DOI: 10.1016/j.cub.2006.11.063 (2007).

76. Reynolds, J. H., Pasternak, T. & Desimone, R. Attention increases sensitivity of V4 neurons. *Neuron* **26**, 703–714, DOI: 10.1016/S0896-6273(00)81206-4 (2000).

77. Neal, R. M. Bayesian Learning via Stochastic Dynamics. Tech. Rep. (1992).

78. Hinton, G., Hinton, G. & Van Camp, D. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. *IN PROC. OF THE 6TH ANN. ACM CONF. ON COMPUTATIONAL LEARNING THEORY* 5—-13 (1993).

79. Wu, H. & Gu, X. Towards dropout training for convolutional neural networks. *Neural Networks* **71**, 1–10, DOI: 10.1016/j.neunet.2015.07.007 (2015).

80. Wang, S. I. & Manning, C. D. Fast dropout training. In *ICML* (2013).

81. Santhanam, G. *et al.* Factor-analysis methods for higher-performance neural prostheses. *J. Neurophysiol.* **102**, 1315–1330, DOI: 10.1152/jn.00097.2009 (2009).

Acknowledgements (not compulsory)

R.CC. was supported by NIH grant EY030578. O.S. was supported by NSF grant 1715475.

Author contributions statement

X.P. and O.S. conceived the initial project. All authors contributed to the experiments design. X.P. conducted all the experiments and analysis. All authors reviewed this manuscript.

Competing interests

The authors declare no competing interests.