The Data Efficiency of Deep Learning Is Degraded by Unnecessary Input Dimensions

Citation: D'Amario, Vanessa, Srivastava, Sanjana, Sasaki, Tomotake and Boix, Xavier. 2022. "The Data Efficiency of Deep Learning Is Degraded by Unnecessary Input Dimensions." Frontiers in Computational Neuroscience, 16.

As Published: 10.3389/fncom.2022.760085

Publisher: Frontiers Media SA

Persistent URL: https://hdl.handle.net/1721.1/139838

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fncom.2022.760085

The Data Efficiency of Deep Learning Is Degraded by Unnecessary Input Dimensions

Vanessa D'Amario[1,2*†], *Sanjana Srivastava*[2,3], *Tomotake Sasaki*[4] *and Xavier Boix*[1,2*]

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States, [2] Center for Brains, Minds and Machines, Cambridge, MA, United States, [3] Department of Computer Science, Stanford University, Stanford, CA, United States, [4] Artificial Intelligence Laboratory, Fujitsu Limited, Kawasaki, Japan

Biological learning systems are outstanding in their ability to learn from limited training data compared to the most successful learning machines, *i.e.,* Deep Neural Networks (DNNs). What are the key aspects that underlie this data efficiency gap is an unresolved question at the core of biological and artificial intelligence. We hypothesize that one important aspect is that biological systems rely on mechanisms such as foveations in order to reduce unnecessary input dimensions for the task at hand, *e.g.,* background in object recognition, while state-of-the-art DNNs do not. Datasets to train DNNs often contain such unnecessary input dimensions, and these lead to more trainable parameters. Yet, it is not clear whether this affects the DNNs' data efficiency because DNNs are robust to increasing the number of parameters in the hidden layers, and it is uncertain whether this holds true for the input layer. In this paper, we investigate the impact of unnecessary input dimensions on the DNNs data efficiency, namely, the amount of examples needed to achieve certain generalization performance. Our results show that unnecessary input dimensions that are task-unrelated substantially degrade data efficiency. This highlights the need for mechanisms that remove task-unrelated dimensions, such as foveation for image classification, in order to enable data efficiency gains.

Keywords: data efficiency, overparameterization, object recognition, object background, unnecessary input dimensions, deep learning

OPEN ACCESS

Edited by:
Omri Barak,
Technion Israel Institute of Technology, Israel

Reviewed by:
Wei Lin,
Fudan University, China
Mariofanna Milanova,
University of Arkansas at Little Rock, United States
Jeremy Bernstein,
California Institute of Technology, United States

***Correspondence:**
Vanessa D'Amario
vanessad@mit.edu
Xavier Boix
xboix@mit.edu

†Present address:
Vanessa D'Amario,
Fujitsu Research of America, Inc., Sunnyvale, CA, United States

Received: *17 August 2021*
Accepted: *03 January 2022*
Published: *31 January 2022*

Citation:
D'Amario V, Srivastava S, Sasaki T and Boix X (2022) The Data Efficiency of Deep Learning Is Degraded by Unnecessary Input Dimensions. Front. Comput. Neurosci. 16:760085. doi: 10.3389/fncom.2022.760085

1. INTRODUCTION

The success of Deep Neural Networks (DNNs) contrasts with the still distant goal of learning with few training examples as in biological systems, *i.e.,* in a data efficient manner (Hassabis et al., 2017). Understanding the principles that underlie such differential is a question at the core of both artificial and biological intelligence. In this paper, we introduce the hypothesis that an important aspect for data efficiency is that biological systems rely on mechanisms such as foveations in order to reduce unnecessary input dimensions, *e.g.,* background in object recognition, while state-of-the-art DNNs do not.

DNNs are usually trained on high dimensional datasets (*e.g.,* images and text), and many input dimensions of the DNN may be unnecessary to predict the ground-truth label as they are unrelated and/or redundant to the task at hand. Machine learning theory for linear and kernel methods

predicts that unnecessary input dimensions may degrade the DNN's data efficiency (Hastie et al., 2009), as the classifier may overfit to the unnecessary input dimensions if not enough training examples are provided to learn to discard them.

However, DNNs have challenged classic machine learning measures of complexity (*e.g.,* VC dimensions, Rademacher complexity) as they can achieve high test accuracy despite having a number of trainable parameters much larger than the number of training examples, *i.e.,* DNNs are overparameterized (Zhang et al., 2017; Nakkiran et al., 2020). Since unnecessary input dimensions lead to more overparameterization, it is unclear in what way DNNs suffer from unnecessary input dimensions and whether more data is needed to learn to discard them.

To foreshadow the results, we find that the DNNs' data efficiency depends on whether the unnecessary dimensions are *task-unrelated* or *task-related* (redundant with respect to other input dimensions). Namely, increasing the number of *task-unrelated* dimensions leads to a substantial drop of data efficiency, while increasing the number of *task-related* dimensions that are linear combinations of other *task-related* dimensions, helps to alleviate the negative impact of the *task-unrelated* dimensions. These results suggest that mechanisms to discard unnecessary input dimensions, such as foveations for object recognition, are necessary to enable data efficiency gains.

2. RELATED WORKS

We now relate our work with the effect of background on the generalization abilities of DNNs in object recognition, and also with the DNNs generalization abilities depending on the number of parameters of the network.

2.1. Object's Background and DNN Generalization

The data collection process is often biased (Torralba and Efros, 2011). One of the most prominent factors of such dataset bias is the background, such that some aspects of the background systematically co-occur with certain objects, *e.g.,* airplanes may tend to always appear in the sky. This co-occurrence is a confounding factor for the network, and the network may learn to associate the background with the object, *e.g.,* the sky may be regarded as part of the airplane. Previous works have shown that DNNs for image recognition fail to classify objects in novel and uncommon backgrounds (Choi et al., 2012; Volokitin et al., 2017; Beery et al., 2018; Tian et al., 2018). Remarkably, popular object recognition datasets are biased to such an extent that DNNs can predict the object category even when the objects are removed from the image (Zhu et al., 2017; Tian et al., 2018; Xiao et al., 2021). Barbu et al. (2019) introduced a new benchmark which addresses the biased co-occurrence of objects and background, among other types of bias. DNNs exhibit large performance drops in this benchmark compared to ImageNet (Deng et al., 2009). Recently, Borji has shown that a large portion of the performance drop comes from the bias in object's background, as classifying the object in isolation substantially alleviates the performance drop (Borji, 2021).

In contrast to previous works, we analyse the impact of the object's background to the DNN's generalization performance when the dataset is unbiased, *i.e.,* there is no significant correlation between the objects and backgrounds and the statistics of the object's background are the same between training and testing times. To the best of our knowledge, our work is the first to investigate the effects of object's background on DNNs when these are unbiased. We show that just the presence of background, even if it is unbiased, can degrade the data efficiency of the DNN.

2.2. Overparameterization and Data Dimensionality

A remarkable characteristic of DNNs is that the test error follows a double-descend when the DNN's width is increased by adding more hidden units. Thus, the test error decreases as the network's width is increased in both the underparameterized and overparameterized regimes, except in a critical region between these two where a substantial error increase can take place (Belkin et al., 2019; Advani et al., 2020; Nakkiran et al., 2020). The overparameterized regime has received a lot of attention because DNNs with many more parameters than training examples can achieve high test accuracy, and a theoretical understanding of this phenomenon is an active area of research. Robustness to overparametrization relates to unnecessary input dimensions because unnecessary input dimensions also increase the number of parameters of the network, albeit in the input layer rather than in the intermediate layers. As we show in the sequel, increasing the number of unnecessary input dimensions can have the opposite effect of increasing the number of hidden units in the test error.

A theoretical understanding of this phenomenon using mathematical tools is an open question. The PAC Bayes theory appears as a promising approach to describe the generalization capacity of DNNs [*e.g.,* (Dziugaite and Roy, 2017; De Palma et al., 2019; Bernstein and Yue, 2021)]. While these theoretical results provide insights about the trends of the behaviour of the DNN, an empirical, quantitative assessment of the effect of unnecessary dimensions to the DNN's data efficiency is missing. Our analysis derives from theoretical insights of the exact solution of a linear network in a regression task. In this way, we can relate and compare empirical results for DNNs with cases that are well understood theoretically.

Another strand of research relates the structure of the dataset with the generalization ability of the network. Several works in statistical learning theory for kernel machines relate the spectrum of the dataset with the generalization performance (Zhang, 2005). For neural networks, Ansuini et al. (2019), Recanatesi et al. (2019) define the intrinsic dimensionality based on the dimension of the data manifold. These works analyze how the network reduces the intrinsic dimension across layers. Yet, these metrics based on manifolds do not provide insights about how specific aspects of the dataset, *e.g.,* unnecessary dimensions, contribute to the intrinsic dimensionality.

3. UNNECESSARY INPUT DIMENSIONS AND DATA EFFICIENCY

We aim at analyzing the effect of unnecessary input dimensions on the data efficiency of DNNs. Let $\mathbf{x}$ be a vector representing a data sample, and let $\mathbf{y}$ be the ground-truth label of $\mathbf{x}$. We define $f(\mathbf{x}) = \mathbf{y}$ as the target function of the learning problem. Also, we use $[\mathbf{x}; \mathbf{u}]$ to denote the data sample $\mathbf{x}$ with unnecessary input dimensions appended to it. The unnecessary dimensions do not affect the target function of the learning problem, *i.e.,* $g([\mathbf{x}; \mathbf{u}]) = f(\mathbf{x}) = \mathbf{y}$, where g is the target function of the learning problem with unnecessary input dimensions. Each sample can have a different set of dimensions that are unnecessary, *e.g.,* one sample could be $[\mathbf{x}_1; \mathbf{u}_1]$ and another be $[\mathbf{u}_2; \mathbf{x}_2]$. Note that this variability is present in object recognition because the dimensions representing the object's background are unnecessary and vary across data samples, as the object can be in different image locations.

We define two types of unnecessary input dimensions: *task-unrelated* and *task-related*. Unnecessary input dimensions are *task-unrelated* when they are independent of $\mathbf{x}$, *i.e.,* they can not be predicted from $\mathbf{x}$, as in unbiased object's background. Otherwise, the unnecessary dimensions are *task-related*, which are equivalent to redundant dimensions. An example that leads to more *task-related* unnecessary dimensions is upscaling the image.

To study the effect of unnecessary input dimensions, we measure the test accuracy of DNNs trained with different amounts of unnecessary input dimensions and training examples. Given a DNN architecture and a dataset with a fixed amount of unnecessary dimensions, we define the *data efficiency* of the DNN as the Area Under the Test Curve (AUTC) for the DNN trained with different number of training examples. The curve is monotonically increasing, as more training examples lead to higher test accuracy, and the AUTC measures the area under it. We normalize the AUTC to be between 0 and 1, where 1 is the maximum achievable, and it corresponds to 100% test accuracy for all number of training examples. In the experiments where the number of training examples spans several orders of magnitude, we calculate the AUTC by converting the number of training examples in logarithmic scale, such that all orders of magnitude are equally taken into account.

4. DATASETS AND NETWORKS

We now introduce the datasets and networks we use in the experiments (refer to **Appendices 1**, **2** for additional details).

4.1. Linearly Separable Dataset

We use a linearly separable dataset for binary classification, as it facilitates relating results of classic machine learning and DNNs. We generate a binary classification dataset of 30 input dimensions, which follow a Gaussian distribution with $(\mu = 0, \sigma = 1)$. The ground-truth label is the output of a linear classifier, such that the dataset is linearly separable with a hyperplane randomly chosen. Unnecessary input dimensions are appended to the data samples. *Task-unrelated* dimensions follow a Gaussian distribution with $(\mu = 0, \sigma = 0.1)$. The *task-related* dimensions are linear combinations of the dimensions of the original dataset samples.

We evaluate the following linear and Multi-Layer Perceptron (MLP) networks: linear network trained with square loss (pseudo-inverse solution), MLP with linear activation functions trained with either square loss or cross entropy loss, and MLP with ReLU trained with cross entropy loss.

4.2. Non-linearly Separable Dataset With Different Noise Distributions

To further evaluate the generality of results on data distributions that are not linearly separable, we use a mixture of Gaussians to generate non-linearly separable datasets for binary classification. Each class consists of three multivariate Gaussians of dimensions $p = 30$. We generate a sample by randomly selecting with the same probability one of the three distribution. To give a more comprehensive evaluation on the effect of different types of noise, we generate unnecessary dimensions using Gaussian distributions with different variance, and we also evaluate two other noise distributions, namely, Gaussian noise with $\Sigma_{ii} = 1$, $\forall i$, with $\Sigma_{ij} = 0.5$, $\forall i \neq j$, and salt and pepper noise, where each vector component can assume value (0, or u), based on a Bernoullian distribution on $\{-1, 1\}$.

We consider the MLP with ReLU and soft-max with cross-entropy loss because among the different variants it is the only well suited to fit non-linearly separable data.

4.3. Object Recognition Datasets

We evaluate object recognition datasets based on extensions of the MNIST dataset (LeCun et al., 1998) and the Stanford Dogs dataset (Khosla et al., 2011).

Synthetic and Natural MNIST. We generate two datasets based on MNIST: the synthetic MNIST and the natural MNIST, which have synthetic and natural background, respectively. In both datasets, the MNIST digit is always at the center of the image and normalized between 0 and 1.

In the synthetic MNIST dataset, the *task-unrelated* dimensions are sampled from a Gaussian distribution with $(\mu = 0, \sigma = 0.2)$ and the *task-related* dimensions are the result of upscaling the MNIST digit. We also combine *task-related* and *unrelated* dimensions by fixing the size of the image and changing the ratio of *task-related* and *unrelated* dimensions by upscaling the MNIST digit.

In the natural MNIST dataset, the background is taken from the Places dataset (Zhou et al., 2014), as in Volokitin et al. (2017). The size of the image is constant across experiments (256×256 pixels), and the size of the MNIST digits determines the amount of *task-related* and *unrelated* dimensions.

We use the MLP with ReLU and cross entropy loss, and also Convolutional Neural Networks (CNNs). The architecture of the CNN consists of three convolutional layers each with max-pooling, followed by two fully connected layers. Since the receptive field size of the CNN neurons may have an impact on the data efficiency, we evaluate different receptive field sizes. We use a factor r to scale the receptive field size, such that the convolution filter size is $(r \cdot 3) \times (r \cdot 3)$ and the pooling region size is $(r \cdot 2) \times (r \cdot 2)$. We experiment by

either fixing r to a constant value or adapting r to the scale of the MNIST digit, such that the receptive fields of the neurons capture the same object region independently of the scale of the digit.

Stanford Dogs. Recall our analysis focuses on unnecessary input dimensions that are unbiased. We use the Stanford Dogs dataset (Khosla et al., 2011) as it is reasonable to assume that the bias between breeds of dogs and background is negligible. This dataset contains natural images (227 × 227 pixels) of dogs at different image positions. The amount of *task-unrelated* dimensions is determined by the dog size, which is different for each image. To evaluate the effect of unnecessary input dimensions, we introduce the following five versions of the dataset. Case 1 corresponds to the original image. In case 2, we multiply by zero the pixels of the background, which reduces the variability of the *task-unrelated* dimensions. In case 3, the dog is centered in the image. In case 4, we fix the ratio of *task-related/unrelated* dimensions by centering the dog and scaling it to half of the image size. In case 5, we remove the background by cropping and scaling the dog.

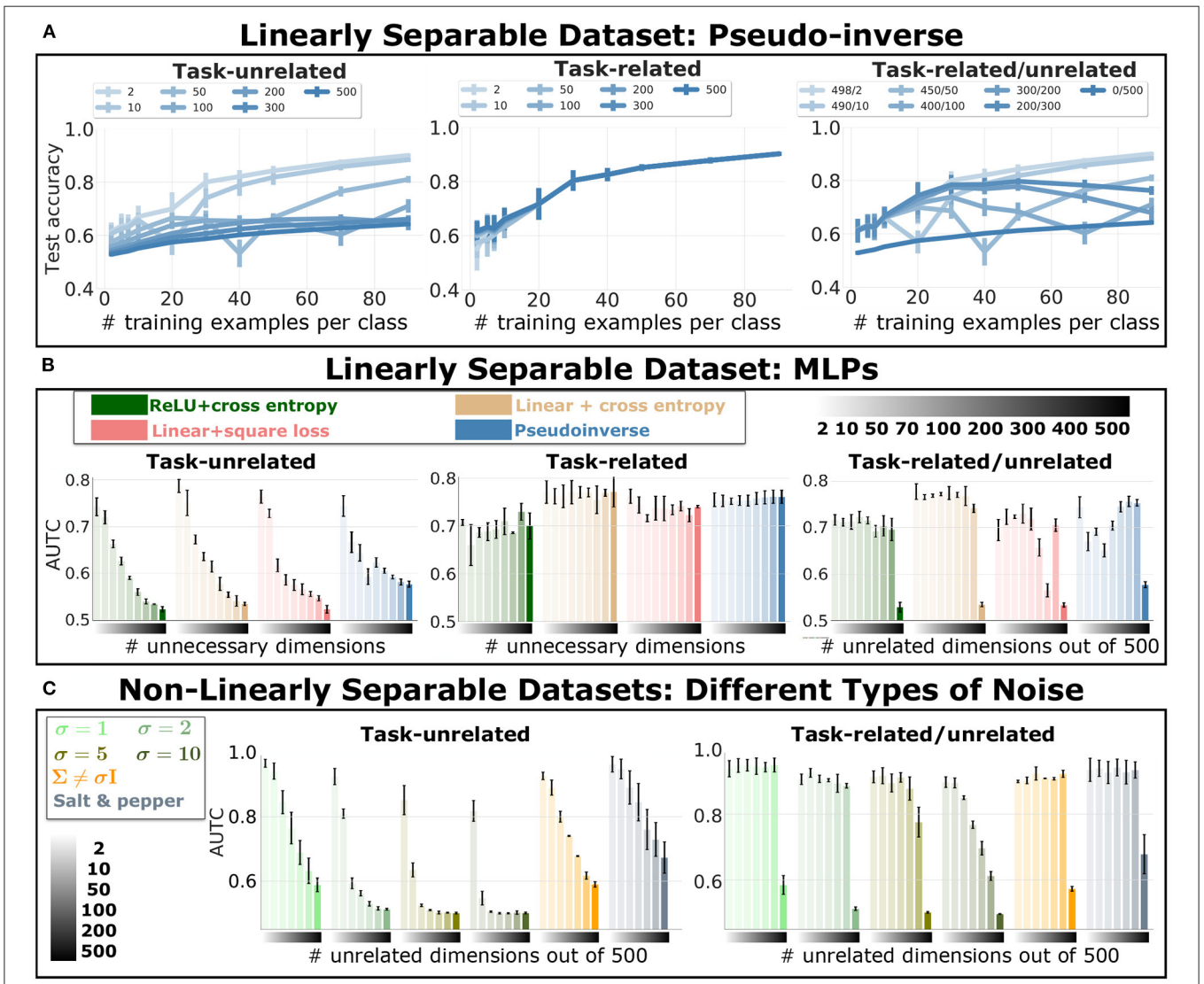

FIGURE 1 | *Data Efficiency of Linear and Fully Connected Networks Trained on Dataset For Binary Classification.* Data efficiency for different amount of unnecessary input dimensions. Error bars indicate standard deviation across experiment repetitions. **(A)** Test accuracy of the pseudo-inverse solution as function of different amount of training examples. Each curve corresponds to a different amount of additional *task-unrelated* (left plot), *task-related* (middle plot) and *task-related/unrelated* dimensions, as reported in the legend. **(B)** We report the Area Under the Test Curve (AUTC) of the accuracy for different amount of training examples. We indicate with the gradient bar on the *x*-axis the amount of additional *task-unrelated* dimensions (left and right plot), and additional *task-related* dimensions, in the middle plot. The legend indicate the different networks trained and tested on the dataset. **(C)** AUTC values for an MLP with ReLU activation and cross entropy loss, for different types of *task-unrelated* dimensions: Gaussian independent components with increasing variance, Gaussian with non diagonal covariance, and salt and pepper noise. The gradient on the *x*-axis indicates the number of *task-unrelated* dimensions.

We use a ResNet-18 (He et al., 2016), following the standard pre-processing of the image used in ImageNet.

5. RESULTS

In this section, we report results, first on the linearly separable datasets and then, on the object recognition datasets.

5.1. Linearly Separable Dataset

Figure 1A shows the test accuracy of the pseudo-inverse solution for different number of training examples and unnecessary input dimensions. **Figure 1B** reports the data efficiency of all networks tested for different number of unnecessary input dimensions. Recall that the data efficiency is measured with the AUTC and summarizes the test accuracy as a function of the amount of training examples, *e.g.,* for the pseudo-inverse solution, the curves in **Figure 1A** are summarized by the AUTC in **Figure 1B**. We observe that increasing the amount of *task-unrelated* dimensions harms the data efficiency, *i.e.,* the AUTC drops. Also, the *task-related* dimensions alone do not harm data efficiency, and they alleviate the effect of the *task-unrelated* dimensions.

These results clarify the difference between robustness to overparameterization in intermediate layers and unnecessary input dimensions. Note that the effect on the test accuracy of increasing the number of hidden units is the opposite of increasing the number of *task-unrelated* input dimensions, *i.e.,* DNNs are not robust to all kinds of overparameterization.

Analytical results for linear regression using the square loss predicts an analogous effect of *task-unrelated* dimensions on the solution. For sake of clarity, we retrace these results in the **Appendix 3.1**, where we outline the effect of additional *task-unrelated* dimensions that are Gaussian-distributed on the pseudo-inverse solution. There, we show that *task-unrelated* dimensions lead to the pseudo-inverse, Tihkonov-regularized solution calculated in the dataset without *task-unrelated* dimensions. Since in this case the regularization can not be tuned or switched off as it is fixed by the number of *task-unrelated* dimensions, it is likely to harm the test accuracy, as we have observed.

The regularizer is beneficial in some specific cases. Following from regularization theory (Hastie et al., 2009), **Appendix 3.2** highlights a noisy regression problem in which certain amounts of *task-unrelated* dimensions help to improve generalization. In a classification problem, Tiknhonov regularization may also be beneficial in some cases. This can be seen in **Figure 1A**, where we observe that for a given number of training examples, increasing the number of *task-unrelated* dimensions improves the test accuracy in some cases. This specific trend relates to the aforementioned double descend of DNNs (Belkin et al., 2019; Advani et al., 2020). As shown in Nakkiran et al. (2020), the location of the critical region is affected by the number of training examples and the complexity of the model. Here, the complexity of the model is affected by the number of *task-unrelated* dimensions due to its regularization effect.

5.2. Non-linearly Separable Datasets With Different Distributions of *Task-Unrelated* Dimensions

In **Figure 1C**, we show the data efficiency of MLP with ReLU trained with cross entropy loss on non-linearly separable datasets. On the left of the quadrant, we report different distributions of the *task-unrelated* dimensions: Gaussian noise with different σ (corresponding to multiplicative factor applied on the identical covariance matrix), Gaussian noise with non-diagonal covariance matrix and salt and pepper noise. The amount of *task-unrelated* dimensions is reported through the colored bar (indicated with a gradient). We observe that, similarly to previous results in the linearly separable dataset (**Figure 1B**), *task-unrelated* dimensions harm data efficiency. Also, as expected, data efficiency deteriorates as the variance of Gaussian noise increases. The combination of *task-related/unrelated* dimensions alleviates the detrimental effect of *task-unrelated* dimensions. These empirical results on MLPs show a similar trend to the one predicted for linear networks, with exception of a less pronounced effect of the double descent behavior.

5.3. Object Recognition Datasets

Figure 2A shows the log-AUTC for the MLP and the CNN for different amount of unnecessary dimensions (an increasing amount as we move left to right), for the synthetic MNIST dataset. In **Appendix 4.1**, we report the test accuracy for different number of unnecessary dimensions, which further strengthens the results of **Figure 1**. Conclusions are consistent with the previous results in the linearly separable dataset. Also, we observe that CNNs are overall much more data efficient than MLPs, which is expected because of their more adequate inductive bias given by the weight sharing of the convolutions.

Figure 2B shows results in natural MNIST dataset for different ratios of *task-related/unrelated* dimensions. The plots compare CNNs with different receptive field sizes, represented by the factor r (see Section 4.3). Since the CNN achieves high accuracy with few examples, the mean and standard deviation of the log-AUTC (left plot) hardly show any variation when computed on more than 20 training examples per class. Yet, the gap of the testing accuracy is considerable for 20 training examples per class (right plot). These results confirm that *task-unrelated* dimensions degrade data efficiency independently of the receptive field sizes (see **Appendix 4.2** for additional results further supporting these conclusions).

Figure 2C shows results on the Stanford Dogs dataset, namely the AUTC score across the five cases of unnecessary dimensions that we evaluate. This dataset serves to assess a more realistic scenario, where the objects can appear at different positions and scales. We observe that the *task-unrelated* dimensions, which come from the background, harm the data efficiency (cases 1 to 4 versus case 5). Putting to zero the unnecessary dimensions improves the data efficiency of models trained on the original dataset (cases 2 to 4 vs. case 1). This is because the *task-unrelated* dimensions become redundant as they all take the same value in all images. We also observe that removing the variability of the position and scale of the object hardly affects the data efficiency

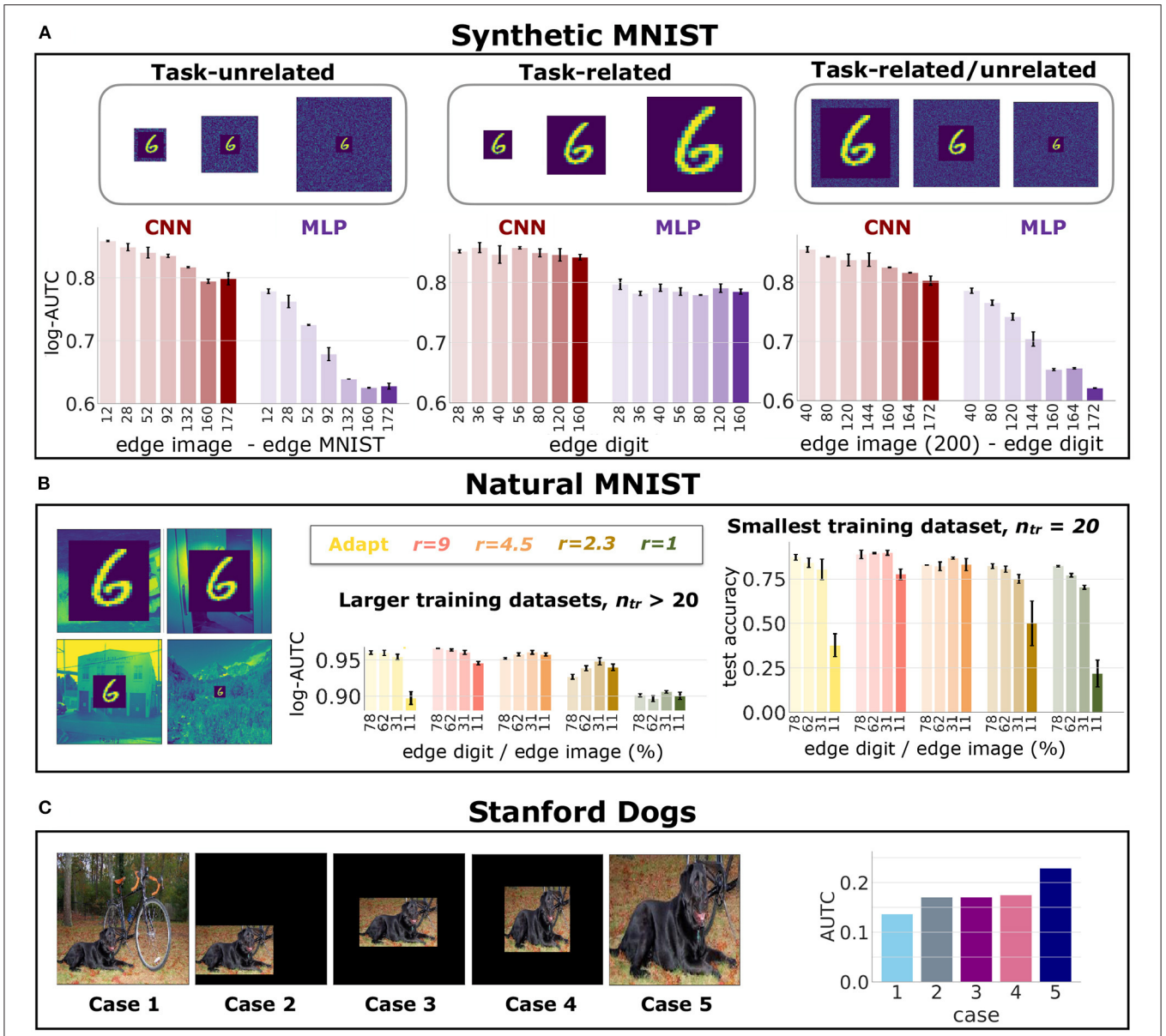

FIGURE 2 | *Data Efficiency in Object Recognition Datasets*. Results of DNNs' data efficiency for different amount of unnecessary input dimensions, tasks and networks. Error bars indicate standard deviation across experiment repetitions. **(A)** log-AUTC for CNNs and MLPs trained on synthetic MNIST, for different number of unnecessary dimensions. **(B)** Left plot: log-AUTC for networks trained on Natural MNIST for larger amount of training examples; Right: test accuracy on the smallest training set. **(C)** AUTC on the Stanford Dogs dataset for the five cases shown on the left of the panel.

(case 2 to 4). Thus, learning to discard the background requires more training examples than learning to handle the variability in scale and position of the object.

6. CONCLUSIONS

We have analyzed the effect of unnecessary input dimensions (*e.g.,* object's background). We found that *task-unrelated* dimensions harm the data efficiency, while increasing the number of *task-related* dimensions that are linear combinations of other *task-related* dimensions help to alleviate the negative effect of *task-unrelated* dimensions. These results demonstrate that the robustness of DNNs to overparameterization is limited, as increasing the number of *task-unrelated* input dimensions is a form of overparameterization that degrades the accuracy. Also, our results add to the growing body of works in object recognition that shows that bias in the object's background can undermine the reliability of DNNs. Here we have shown that the problem runs far deeper, as the object's background negatively affects the network even when there is no bias.

Taken together, these results suggest that data efficiency gains could be enabled by mechanisms that remove *task-unrelated* dimensions, such as foveation for image classification (Luo et al., 2016; Akbas and Eckstein, 2017), or also by adapting to DNNs

regularization techniques that encourage predictions from a sparse subset of input dimensions (*e.g.,* ℓ_1 regularization for linear regression Hastie et al., 2019). Also, our results can be extended to other domains, such as natural language processing and clinical tasks, as the effect of unnecessary dimensions may have been investigated, *e.g.,* (Laksana et al., 2020), but their effects in the data efficiency remain largely unexplored.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Stanford Dogs dataset: https://www.tensorflow.org/datasets/catalog/stanford_dogs; MNIST dataset: https://www.tensorflow.org/datasets/catalog/mnist; PLACES dataset: http://places.csail.mit.edu/; Synthetic datasets can be generated using the following code: https://github.com/vanessadamario/data_efficiency/blob/main/synthetic_framework/main.py; The code supporting the conclusions of this article is publicly accessible in the following github repository: https://github.com/vanessadamario/data_efficiency.git.

AUTHOR CONTRIBUTIONS

VD'A implemented the experiments and carried out the analysis, with contributions of SS and XB. VD'A and XB conceived the experiments with contributions of SS and TS. VD'A and XB wrote the manuscript with contributions of TS. XB and TS supervised the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been supported by the Center for Brains, Minds, and Machines (funded by NSF STC award CCF-1231216), XB by the R01EY020517 grant from the National Eye Institute (NIH) and XB and VD'A by Fujitsu Laboratories Ltd. (Contract No. 40008819) and the MIT-Sensetime Alliance on Artificial Intelligence.

ACKNOWLEDGMENTS

We would like to thank Pawan Sinha and Tomaso Poggio for useful discussions and insightful advice provided during this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2022.760085/full#supplementary-material

REFERENCES

Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Netw.* 132:428–446. doi: 10.1016/j.neunet.2020.08.022

Akbas, E. and Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLOS Comput. Biolo.* 13:e1005743. doi: 10.1371/journal.pcbi.1005743

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). "Intrinsic dimension of data representations in deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)* (Vancouver, BC), 6109–6119.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., et al. (2019). "ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems (NeurIPS)* (Vancouver, BC), 9448–9458.

Beery, S., Van Horn, G., and Perona, P. (2018). "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 456–473.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15849–15854. doi: 10.1073/pnas.1903070116

Bernstein, J., and Yue, Y. (2021). On the implicit biases of architecture & gradient descent. *arXiv preprint* arXiv:2110.04274.

Borji, A. (2021). "Contemplating real-world object classification," in *Proceedings of the International Conference on Learning Representations (ICLR)*.

Choi, M. J., Torralba, A., and Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recogn. Lett.* 33, 853–862. doi: 10.1016/j.patrec.2011.12.004

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM.

De Palma, G., Kiani, B. T., and Lloyd, S. (2019). "Random deep neural networks are biased towards simple functions," in *Advances in Neural Information Processing Systems (NeurIPS)* (Vancouver, BC).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Miami, FL), 248–255.

Dziugaite, G. K., and Roy, D. M. (2017). "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)* (Sydney, NSW).

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical Learning With Sparsity: the Lasso and Generalizations*. London: CRC.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. (2011). "Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available online at: http://vision.stanford.edu/aditya86/ImageNetDogs/ (accessed November 15, 2021).

Laksana, E., Aczon, M., Ho, L., Carlin, C., Ledbetter, D., and Wetzel, R. (2020). The impact of extraneous features on the performance of recurrent neural network models in clinical tasks. *J. Biomed. Inf.* 102:103351. doi: 10.1016/j.jbi.2019.103351

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791 Available online at: http://yann.lecun.com/exdb/mnist/ (accessed: November 15, 2021).

Luo, Y., Boix, X., Roig, G., Poggio, T., and Zhao, Q. (2016). *Foveation-Based Mechanisms Alleviate Adversarial examples. Technical Report CBMM Memo No. 44*. Center for Brains, Minds and Machines. Available Online at: https://cbmm.mit.edu/sites/default/files/publications/cbmm_memo_044.pdf

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020). "Deep double descent: where bigger models and more data hurt," in *Proceedings of the International Conference on Learning Representations (ICLR)*. (Addis Ababa).

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., and Shea-Brown, E. (2019). Dimensionality compression and expansion in deep neural networks. *arXiv preprint* arXiv:1906.00443.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). "A generalized representer theorem," in *Proceedings of the International Conference on Computational Learning Theory (COLT)* (Amsterdam), 416–426.

Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., and Wang, X. (2018). "Eliminating background-bias for robust person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT), 5794–5803.

Torralba, A., and Efros, A. A. (2011). "Unbiased look at dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, CO), 1521–1528.

Volokitin, A., Roig, G., and Poggio, T. A. (2017). "Do deep neural networks suffer from crowding?" in *Advances in Neural Information Processing Systems (NIPS)* (Long Beach, CA), 5628–5638.

Xiao, K. Y., Engstrom, L., Ilyas, A., and Madry, A. (2021). "Noise or signal: the role of image backgrounds in object recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). "Understanding deep learning requires rethinking generalization," in *Proceedings of the International Conference on Learning Representations (ICLR)* (Toulon).

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* 17, 2077–2098. doi: 10.1162/089976605432 3008

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). "Learning deep features for scene recognition using Places database," in *Advances in Neural Information Processing Systems (NIPS)* (Montreal, QC), 487–495. Available online at: http://places.csail.mit.edu/ (accessed November 15, 2021).

Zhu, Z., Xie, L., and Yuille, A. L. (2017). "Object recognition with and without objects," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Melbourne, QC), 3609–3615.

Conflict of Interest: This study received funding from Fujitsu Laboratories Ltd. The funder through TS had the following involvement with the study: conception of the experiment, writing of this article, and supervision of the study.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.