On Data-centric Myths

Antonia Marcu
Vision, Learning and Control
University of Southampton
am1g15@soton.ac.uk

Adam Prügel-Bennett
Vision, Learning and Control
University of Southampton
apb@ecs.soton.ac.uk

Abstract

The community lacks theory-informed guidelines for building good data sets. We analyse theoretical directions relating to what aspects of the data matter and conclude that the intuitions derived from the existing literature are incorrect and misleading. Using empirical counter-examples, we show that 1) data dimension should not necessarily be minimised and 2) when manipulating data, preserving the distribution is inessential. This calls for a more data-aware theoretical understanding. Although not explored in this work, we propose the study of the impact of data modification on learned representations as a promising research direction.

1 Motivation

In recent years, the crucial role of data has largely been shadowed by the field's focus on architectures and training procedures. As a result, there are no guiding principles for creating a good data set. *What are some intuitions that we get from the literature? Are they correct? What are promising future research directions?* In this paper we focus on the aspects of data quality as resulting from empirical methods for predicting generalisation, namely from Intrinsic Dimension (ID) based methods and Mixed Sample Data Augmentation (MSDA) based methods. We show that they provide misleading insights into how one should create and manipulate data to improve model performance.

Intrinsic Dimension: It is believed that data lies on a low-dimensional manifold. Manifold's dimension should reflect the minimum number of variables required to describe the true data. This is dependent on the task at hand. For reconstruction, we would expect a more complex representational space than for classification. In this paper we focus on the latter. While it is difficult to know the true ID, a number of estimates have been proposed [e.g. 7, 4, 5, 3]. Given a good model we can estimate the ID based on its representations by measuring how much its embedding space can be "compressed". While this is dependent on the quality of both the model and the estimator, in the paper we also provide a conceptual argument that abstracts away from these details. It has been claimed that the lower the dimension of the manifold, the easier it is to generalise. Based on this belief, train-time generalisation estimates have been proposed. Using the TWO-NN [5] algorithm, Ansuini et al. [1] estimate the global ID of last hidden layer manifolds using the train data. They then claim that the generalisation performance can be predicted based on this quantity. As such, better performance should correspond to a lower ID value. The intuition that results from this is that creating lower-dimensional data leads to better generalisation. Note that the *estimated* ID could be change by altering either the architecture or the data. Since in this paper we are interested in the role and attributes of the data, we analyse the data for a fixed model.

Mixed augmentation: In statistical learning, training with augmented data is seen as injecting prior knowledge about the neighbourhood of the data samples. The intuition behind augmentation caused researchers to interpret its effect through the similarity between original and augmented data distributions. This perspective is often challenged by methods which, despite generating samples that do not appear to fall under the distribution of natural images, lead to strong learners. This is particularly the case for MSDA, where two or more images are combined to obtain a new training

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Image 1 Image 2 MixUp FMix

Table 1: Examples of images obtained using MixUp and FMix augmentations with a 0.5 mixing coefficient.

Table 2: ID and generalisation performance on CIFAR-10 (top) and CIFAR-100 (bottom). RMixUp leads to worse generalisation but lower ID.

	ID	Accuracy
basic	$7.80_{\pm 17}$	$93.04_{\pm 0.17}$
MixUp	$\mathbf{9.14_{\pm 0.31}}$	$\mathbf{93.79_{\pm 0.18}}$
RMixUp	$7.80_{\pm 16}$	$92.40_{\pm 0.34}$
basic	$12.18_{\pm 1.30}$	$71.70_{\pm 0.37}$
MixUp	$\mathbf{14.11_{\pm 1.31}}$	$\mathbf{72.60_{\pm 0.63}}$
RMixUp	$10.71_{\pm 0.21}$	$69.00_{\pm 0.41}$

sample. Visual examples can be found in Figure 1. Gontijo-Lopes et al. [6] argue it is the *perceived* distribution shift that needs to be minimised, while maximising the sample vicinity. Formalising these concepts, they introduce augmentation "diversity" and "affinity". Diversity is defined as the training loss when learning with artificial samples, while affinity quantifies the difference between the accuracy on original test data and augmented test data for a reference model. The latter penalises augmentations that introduce artificial information to which the model is not invariant, implicitly assuming that training with that information is detrimental to generalisation. In other words, it implies that preserving the data distribution when distorting data is necessary.

In this paper we show that data sets which have a higher intrinsic dimension could lead to better performance than their low-dimension counterparts. Thus, *minimising data dimension is not a relevant goal when creating and refining data sets*. Further, we construct empirical counter-examples which disprove common beliefs in the literature and highlight the importance of understanding the changes MSDAs introduce. We show that, in contrast to what is widely assumed, *not preserving the data distribution can lead to learning better representations*. A direct consequence is that when dealing with limited data, the focus of the practitioners should be on understanding the changes that mixed augmentations cause, rather than choosing the augmentation that produces the smallest distribution shift. Correctly understanding the impact of the increasingly popular mixed-sample augmentation is essential for trusting its usage in sensitive applications where the data can be out of distribution. But most importantly, we believe this could set a new direction in capturing the relationship between data and learned representations, which could ultimately play a small role in understanding generalisation and creating better data sets.

We focus on two MSDAs, MixUp [18] and FMix [8]. MixUp interpolates between two images to obtain a new sample. FMix masks out a region of an image with the corresponding region of another image, sampling the mask from Fourier space. We refer to models by the augmentations they were trained with and use "basic" for models trained without MSDA. We do 5 runs of each experiment.

2 Should we aim to obtain a data set with minimum ID?

Ansuini et al. [1] observe a correlation between data dimension and generalisation capacity. If such a correlation exists, it could be used to characterise and improve data sets. The method they use to compute ID requires training a model, making it impractical for the purpose of data creation and refinement. However, if intrinsic dimension can indeed be used as a generalisation capacity predictor, the effort of the community could be steered towards building more efficient estimators. But is lower ID the driving factor of stronger learners or is this correlation coincidental? In this section we show that higher ID representations can lead to better generalisation performance, thus disproving the above correlation. We train the same model architecture on data with different ID and compute the estimate of representation dimension. Following Ansuini et al. [1], we use the TWO-NN estimator introduced by Facco et al. [5], where the ratio of the distances to the closest two neighbours of each point are used to approximate manifold dimension. We then argue that even if the representation ID estimator was not entirely reflective of true data dimension, minimising data ID does not imply better generalisation and is not a relevant objective when creating data sets.

Table 3: Augmentation comparison on CIFAR-10. We consider two variants when calculating diversity. One is the cross-entropy loss using the label of the majority class (Diversity), as for mixing in [11]. The alternative, MixDiversity, takes a linear combination of the two cross-entropy losses.

	Affinity	Diversity	MixDiversity
MixUp	$\mathbf{-12.58_{\pm 0.14}}$	$\mathbf{0.41_{\pm 0.01}}$	$\mathbf{0.84_{\pm 0.00}}$
FMix	$-25.55_{\pm 0.26}$	$0.34_{\pm 0.01}$	$0.65_{\pm 0.00}$

To obtain data with different properties in a controlled manner, we make use of augmentation. In addition to the two MSDAs introduced above, we also create a data set with a variation of MixUp equivalent to an objective reformulation [10, 8] which we label RMixUp. This consists of creating a new sample by interpolating two images but unlike MixUp, the new sample is assigned the label of only one of the source images. Ignoring one of the targets when mixing inputs is expected to create a data set where the instances can be represented in a more compressed manifold, decreasing separability at the same time. Table 2 shows the results we obtain for the VGG16 [16] network on the CIFAR-10/100 [12] data sets. MixUp has the highest test accuracy, while having a significantly higher ID compared to the RMixUp model. *This directly contradicts the idea that a minimum ID data set is necessarily better*.

One question that is immediately raised is if our conclusion would still hold given a more accurate method of capturing manifold dimension. We argue that even with further estimator refinements, this hypothesis lacks a strong basis and it is unlikely to hold in practice. To see this more clearly, we can think of a binary image classification, where a data collection artefact is present for one of the classes such as a specific small group of identifying pixels. In this case, a learner that classifies entirely based on this spurious rule would achieve very high compression with no real generalisation abilities. Thus, we should not seek to minimise the intrinsic dimension when processing data.

So what should one seek when manipulating data? A line of work that tries to address this is augmentation analysis. There is no unifying framework for understanding augmentation. Despite the lack of consensus in the field, there is one undisputed belief that *the smaller the distribution gap between original and manipulated data, the better the model generalises*. In the following section we focus on this belief which impacts not only augmentation, but data manipulation as a whole.

3 Is the magnitude of the distribution shift important when manipulating data?

Traditionally it was believed that a good augmentation should have minimal distribution shift. Most recently, it has been argued that it is the degree of the *perceived* shift that determines augmentation quality [6]. We show that *the magnitude of the distribution shift does not determine augmentation quality*. We start with the perceptual gap of training with MSDA, as proposed in Gontijo-Lopes et al. [6]. Reiterating, this is given by the difference between the performance of the baseline model when presented with original test data and augmented test data and is termed "affinity". Subsequently, we address the gap in the wider sense, as is often sought in prior art. We first argue that high affinity and high diversity are not necessarily desirable. Indeed, on CIFAR-10, we find FMix, a better performing augmentation, to have both lower affinity and lower diversity than MixUp (Table 3). For diversity, we compute the cross-entropy loss where the label is taken to be that of the majority class. Similar results are obtained with the MixUp loss, where a weighted average of the true labels is taken.

While intuitively for a high level of affinity, high diversity could correspond to better methods, the converse does not hold. We argue this is because affinity is rather an analysis of the learnt representations of the reference model and cannot give an insight into the quality of the augmentation or its effect on learning. As such, an augmentation will have a lower affinity if it introduces artefacts that could otherwise lead to learning better representations when used in the training process. We believe this issue extends to other approaches that aim to motivate the success of MSDA through reduced distribution shift. Henceforth, we focus on bringing further supporting evidence that the importance lies in the invariance introduced by the shift and its interaction with the given problem rather than its magnitude.

Table 4: Accuracy on CIFAR-10 (left) and CIFAR-100 (right) upon mixing with samples from a different data set. The baseline is the accuracy when training with a single data set. CIFAR-110 is used to refer to mixing with CIFAR-100 when training on the CIFAR-10 problem and vice-versa.

	MixUp	FMix	MixUp	FMix
baseline	$94.18_{\pm 0.34}$	$94.36_{\pm 0.28}$	$74.68_{\pm 0.37}$	$75.75_{\pm 0.31}$
CIFAR-110	$94.70_{\pm 0.27}$	$94.80_{\pm 0.32}$	$72.36_{\pm 1.04}$	$74.80_{\pm 0.55}$
Fashion	$92.28_{\pm 0.28}$	$95.03_{\pm 0.10}$	$66.40_{\pm 1.86}$	$74.46_{\pm 0.57}$

3.1 If it is not the magnitude that matters, is it the direction?

We use empirical evidence to argue against previous assumptions behind the success of MSDA and propose the study of introduced bias as a more informative research direction. We use the term "bias" to refer to a drift in the learnt representations introduced by the change in the training procedure. A fundamental difference to classical training is that the samples are no longer independent when augmenting. Mixed-sampling takes this even further. An immediate question is, does the added correlation lead to more meaningful representations? It is claimed that the strength of MixUp lies in causing the model to behave linearly between two images [18] or in pushing the examples towards their mean [2]. Both of these claims rely on the combined images to be generated from the same distribution. Performing inter-dataset augmentation we show that this is not necessary for a successful augmentation. The same experiment further shows that by distorting the data distribution by the same magnitude we can obtain two different results depending on the direction of the introduced bias.

We once again use the reformulated objective setting, where two images are mixed without mixing targets as well. This allows us to apply MSDA between data sets. Thus, for training a model on a data set, we use an additional one whose targets will be ignored. As an example, a model that is learning to predict CIFAR-10 images will be trained on a combination of CIFAR-10 and CIFAR-100 images, with the target of the former. This scenario breaks the added correlation between training examples.

Table 4 contains the results of this experiment, showing that an accuracy similar to or better than that of regular MSDA can be obtained by performing inter-dataset MSDA. This invalidates the argument that the power of MixUp resides in causing the model to act linearly between samples. Another observation is that for FMix and MixUp, introducing elements from CIFAR-100 when training models on the CIFAR-10 problem does not harm the learning process. The reciprocal, however, does not hold. Hence, the "distribution shift" is more intimately linked to the problem at hand and aiming to characterise an augmentation based on the distance from the original distribution is a limiting approach. This experiment shows that shifting two distributions by the same amount can have different effects on the model performance. Thus, *the specifics of the bias introduced could be more important than its magnitude*. While some level of data similarity has to be preserved when performing MSDA, it is far from being the objective of such data-distorting approaches which should be rather seen as forms of regularisation (see Appendix A for experiments and discussion on the increase in data complexity added by MSDA).

In this paper we demonstrate that the shift in learnt representations can lead to better models and simply quantifying the distribution shift can be misleading. An open question remains: How can we better capture the bias that is introduced and measure its quality? We believe understanding how a relatively small change in the data distribution impacts learnt representations could lead the way to characterising the relationship between data and model generalisation.

4 Conclusions

Starting from generalisation studies, we empirically disprove the hypothesis that lower data dimension is necessarily associated with better performance on unseen data. We then show that the purpose of data manipulation is not to leave the distribution unchanged, but to modify it in a principled and constructive manner. The focus of the community must be on analysing the introduced bias rather than its elimination. Correctly interpreting this bias is important not only for making the models trustable but also for injecting more informed prior knowledge in future applications. Beyond their practical benefits, we believe MSDAs have the potential to help characterise the interplay between data and learnt representations.

References

[1] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *arXiv preprint arXiv:1905.12784*, 2019.

[2] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

[3] Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. Distributional results for model-based intrinsic dimension estimators. *arXiv preprint arXiv:2104.13832*, 2021.

[4] Leo L Duan and David B Dunson. Bayesian distance clustering. *arXiv preprint arXiv:1810.08537*, 2018.

[5] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

[6] Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

[7] Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific reports*, 6(1):1–12, 2016.

[8] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Understanding and enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.

[9] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9995–10006. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/71e9c6620d381d60196ebe694840aaaa-Paper.pdf`.

[10] Ferenc Huszár. mixup: Data-dependent data augmentation, 2017. URL `http://www.inference.vc/mixup-data-dependent-data-augmentation/`.

[11] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

[12] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[13] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[14] Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L Edelman, Fred Zhang, and Boaz Barak. SGD on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604*, 2019.

[15] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf`.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[17] Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rye4g3AqFm`.

[18] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

A MSDA increases data complexity

We believe that MSDA training could help bypass some of the simplicity bias. The simplicity bias refers to the tendency of deep models to find simple representations and has been used to justify the success of deep models [14, 17]. Recent research shows that this propensity causes models to ignore complex features that explain the data well in favour of elementary features, even when they lead to worse performance [15, 9].

Although it could seem natural that MSDAs increase the complexity of the problem, we design an experiment to support this claim. Similarly to Shah et al. [15], we combine CIFAR-10 and MNIST [13] samples. Since they have the same number of classes, we can easily associate each class of one data set with a corresponding one from the other. Thus, we stack a padded image from the kth class of MNIST on top of a sample from the kth class of CIFAR-10, such that a $3 \times 64 \times 32$ image is obtained. We then randomly combine the test images and separately compute the accuracy with respect to the targets of each data set.

The predictions with respect to the CIFAR-10 labels are no better than random ($10.04_{\pm 0.11}$), while the accuracy with respect to the MNIST images remains high ($99.57_{\pm 0.72}$). Thus, models trained on this combination are mostly relying on MNIST images to make predictions. Similar behaviours have previously been associated with simplicity bias. Subsequently, when training, we perform FMix only on MNIST images and observe that this is enough to reverse the results. Evaluating against the CIFAR-10 label gives an accuracy of $86.60_{\pm 0.34}$, while testing against the MNIST label only gives $11.61_{\pm 0.30}$. We find that this also holds true for the other MSDAs. Thus, performing these distortions on the simpler data set increases its complexity to the point where it surpasses that of CIFAR-10.