Overparameterization: How much of Neural Networks are used?

Karan Ruparell

Abstract

In this paper we collate and discuss different theories on the true dimensionality of large neural networks. We also perform experiments to suggest that mislabelled data are treated as equally important as correctly labelled data in classifying similar images in the case of over-trained overparameterised models, negating the view of mislabelled data as being memorised exceptions . We find that there is a sharp *phase transition* around 5 - 5.5% of the data being mislabelled, at which the likelihood of Gaussian blurring to change classification is severely decreased for both mislabelled and correctly labelled data. We propose that at this point the model creates a firmer distinction between the two groups.

1 Introduction

The bias variance trade-off has been ubiquitous in statistical analysis, suggesting that there is an ideal number of parameters contained in a model that can learn the necessary features to solve a given task without providing enough redundancy to overfit to training data. Yet in deep learning researchers have obvserved that 'more is better', with larger parameters performing better far beyond the point of being overparameterized. One suggestion to explain the lack of a bias variance trade-off in machine learning is that only a small part of the complexity of a neural network is actually used to understand the data, and that further parameters only serve to increase the chances of finding good, but still low-dimensional, solutions.

One suggestion here is that extra parameters in neural networks contribute most strongly to having redundancy in the network, allowing more possibilities for the neural network to find generalisable, but still low parameter, solutions.

To evaluate this proposal, we will discuss three areas of research:

1. The Lottery Ticket Hypothesis: do subgraphs do all the work? [5]
2. Loss landscape: are there shallow paths between minima? If so, and are these increased with parameter size?
3. Model Preferences: Is there a bias towards more generalisable models?

New Work: We also ask what, if anything, neural networks learn from mislabelled data. More precisely, we ask if the classification of augmented training data is different depending on if the original data was mislabelled.

Conference on Neural Information Processing Systems (NeurIPS 2022), .

1.1 Definitions

Pruning: Pruning is a process by which neural networks can be made sparser. It involves taking a network and selectively removing, fixing to 0, different parameters based on a chosen measure of importance. We will discuss two different methods of pruning in our section on the lottery ticket hypothesis.

Kolmogorov Complexity: Kolmogorov complexity is a measure defined as the length of the smallest possible description of the object in question under some fixed universal language. It is an incomputable measure, but there are approximation for it and we will refer to a paper that uses it when evaluating the complexity of neural networks.

Gaussian Blurring: Gaussian blurring is a process that sets the value of each pixel in an image to match the weighted average of all the pixels in the kernel, a section of pixels surrounding it including itself. Each pixel has weighting

$$G(x,y) = \frac{1}{2\pi * \sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

Where x and y are the coordinate distances between the pixel being blurred and the influencing pixel. Small values of σ result in smaller contributions from distant elements and larger contributions to central elements, resulting in less blurring. We will apply Gaussian blurring in our experiments.

2 The Lottery Ticket Hypothesis (LTH): do subgraphs do all the work? [5]

The lottery ticket hypothesis was introduced in a paper by Frankle et al. in 2019, and provides some evidence to suggest that larger networks rely primarily on single subnetworks within it. They show that pruning uncovers natural subgraphs in the networks that perform just as well as if not better than the overall network when trained alone. This would suggest a bias towards models that are simpler in the sense that they can be expressed using fewer parameters, as the papers central claim is that the subnetworks they call lottery tickets perform just as well as the networks as a whole.

We posit that this is supported by Bai et al.'s recent work "Dual Lottery Ticket Hypothesis" [1], in which they attempt to transfer information from a network to one of its randomly chosen subnetworks. They do this via a process they call *Information Extrusion*, in which they train the whole network with an increasingly bigger regularisation term on the L_2 norm of the parameters outside of the sub-network. This slowly sets the other parameters to 0 as opposed to immediately as in the case of LTH, and the authors claim that this acts to motivate the neural network to transfer information towards the sub-network. Assuming this to be true, their method still only achieves a test accuracy of at most 0.6% more than LTH across all datasets they trained on. As both of these models also perform equally as well as the full model, this suggests that the vast majority of information in the network is already stored in the 'winning tickets'. This may also means that neural networks require much more computational power to learn models that require training all of it's parameters, and so tend towards more generalisable models that only take advantage of the winning tickets. In the case of mislabelling this would suggest that the whether or not the model treats the mislabelled data as exceptions would depend on the capacity of the winning ticket as opposed to the model itself, as it would not have the capacity to treat the mislabelled data as exceptions if the winning ticket were too small.

Curiously, the subnetworks produced here perform significantly better than networks of the same size and architecture perform under random initialisation. Frankle et al. conjecture that overparameterized networks perform better because they have more combinations of subnetworks, each with potential to be a winning ticket. This may explain why overparameterization helps test accuracy instead of diminishing it, as more parameters work directly to increase the chances of strong low dimension subnetworks to train.

A key concern in this hypothesis is that existing measures show that complexity increases with the number of hidden layers. They show that larger models do not stay simple under those measures of complexity. [9] Even so, the ability for smaller parts of a network to perform just as well as the network overall, suggests that there is redundancy in the model.

3 Loss landscape and model preferences

A key concern with overparameterization is that it may make us more likely to find, and be unable to move from, solutions that minimise the training loss but that do not generalise well. In response to this are two results suggested from different papers: (1) That there is a *low-loss landscape*, allowing for easy movement between different minima (2) The parameter-function map is biased towards simple functions which generalise better, so when available will tend towards those. Collectively, this would suggest that deep learning architectures have a preference towards better, more generalisable models, and becomes of the low-loss landscape are able to move towards them regardless of the initial paramterization.

3.1 A Low-Loss Landscape

In 2019 Fort et al [4] and Draxler et al. [3] showed the existence of low-loss sub-spaces connecting sets of solutions for a variety of architectures on the CIFAR datasets. These results suggest that there are routes between different minima, decreasing the odds that a training network may get stuck in a local but unideal minimum. In contrast, both papers also showed that this path is usually not the linear path between them. In fact, Fort et al. showed that a randomly chosen direction of travel from a point, regardless of if it is a local minimum, is likely to result in an increase in loss. They use this to describe the paths between minima as tunnels, and show that attempts to generate simpler models also serves to increase tunnel width. They use L2 regularisation, high learning rates, and dropout as ways to generate simpler models, constraining parameter values, precise location in the loss landscape, and the existence of different parameters respectively.

Crucially, this does not suggest that smaller networks would be expected to have wider tunnels, as there seems to be a substantial difference between generating simpler models from smaller networks and simpler models from larger ones. This can be seen in Li et al.'s work in 2018, [8] where they constrained the parameters of large neural network to satisfy random linear constraints. They found that while the networks could find very good approximations when constrained to very small sub-spaces, for example a 290 dimensional one for LeNet on MNIST for 90% of the baselines accuracy, these models significantly outperformed neural networks with this many initial parameters. This cannot be accounted for by the Lottery Ticket Hypothesis alone, as the randomised constraints reduced the effective dimension such that only the network as a whole has this dimension.

The conclusion is that there seem to be paths between different minima within deep learning architectures, and that these paths become more pronounced when more constraints are added to the parameter landscape that do not effect the overall architecture size. If it were also true that there were some attraction towards more generalisable minima, this would suggest that we need not be too concerned about getting stuck at a bad local minima.

3.2 The parameter-function map is biased towards simple functions which generalise better

In 2018, Perez et al. produced work in which they provide empirical evidence to suggest that various deep learning architectures are biased towards simple functions. They rely on the definitions of 'simple' and 'biased' provided by Dingle et al. [10, 2]. The paper by Dingle et al. shows that for many real world maps, including maps with function outputs, the probability of an output x decays exponentially with its Kolmogorov Complexity To show this empirically they used the approximate Kolmogorov complexity.

They showed this to hold in many real world cases, and Perez et al. furthered this to empirically suggest that it held when applied to CNNs and Fully Connected networks. If this applies to a more general case of networks, it suggests that for any two findable functions **x** and **y**, both optimisation methods advSGD and Adam are exponentially more likely to find the simpler function than the more complex function. A follow up paper by Mingard et al. also claims that SGD is closely related to the Bayesian posterior, in that the Bayesian posterior for the function generated seems to be the first order determinant of the functions generated under SGD. Further work needs to be done to have more confidence, as both papers only trained on Fully Connected (FC) and CNN models, however they do provide some evidence that simple models are preferred to complex ones.

4 Experiments and Results

4.1 Methods

One experiment to provide insight on the effect of overparameterisation on neural networks is how they respond to mislabelled data. We do this in order to test the models preferences towards simpler and more generalisasble models stated in the last section. We expect that a function that had a simple model for classification and a look up table of exceptions would have a lower Kolmogorov complexity than one that tried to incorporate the random noise from the mislabellings into a model. We asked whether the mislabelled data are treated as isolated 'mistakes', where they are fully memorised along with their class, or if they form part of the overall classification model for new data. To test this we mislabelled some MNIST data before the training process, and over-trained a ResNet 18 model. As ResNet18 has over a million parameters, far more than our observations, we expect that the winning ticket which is usually 10-20 percent the size of the overall network will be more than able to store the mislabelled data as exceptions, and so we will achieve good training accuracy on these points. We mislabelled an equal amount of data from each class to control for the proportion of images mislabelled in each class. We then observed how likely the model was to classify altered training data in comparison to unaltered training data when they have undergone Gaussian blurring of different standard deviations. By comparing the difference in label change rate with respect to the standard deviation, and looking at the change rates for mislabelled and correctly labelled data separately, we hope to see if the two are treated significantly differently.

We chose a kernel size equivalent to the whole image so that the limit with respect to σ is a single colour image for all images, containing no information on the original image.

We expect that if the mislabelled data is treated as exceptions, the model would be more likely to classify points close to correctly labelled data with the same label as those close to mislabelled data. This is because the model would be less likely to generalise information learned from mislabelled images if it treated them as exceptions.

This relies on a 'nearest neighbours' approach, that points close to each other are likely to be labelled similarly. It is already known that deviation from an image in some directions lead to strong misclassification, as in Goodfellow et al.'s work on the Fast Gradient Sign Method [6]. This method however requires the direction of movement to be close to the gradient, in such a way that is unlikely to occur in such a high dimensional space. We notice that Gaussian blurring decreases accuracy at a continuous rate, relieving this concern.

All models were trained to at least 99.9% training accuracy over the course of 40 epochs, where accuracy is defined as identifying the labels we assigned to the images as opposed to the their true classes. MNIST has been shown to already contain mislabelled data, and we studied models with additional mislabelling rates ranging from 0.8% to 10%.

4.2 Results

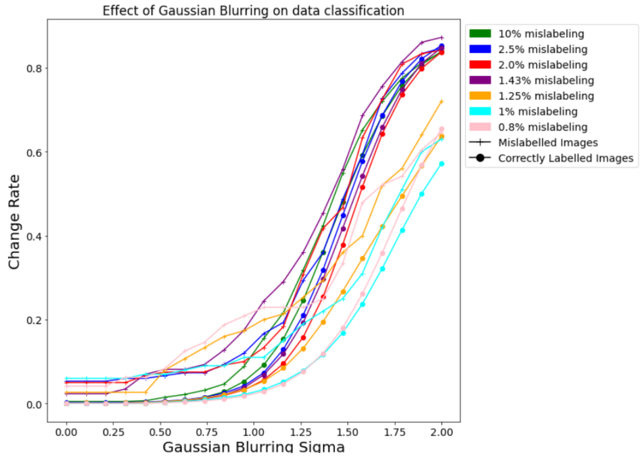


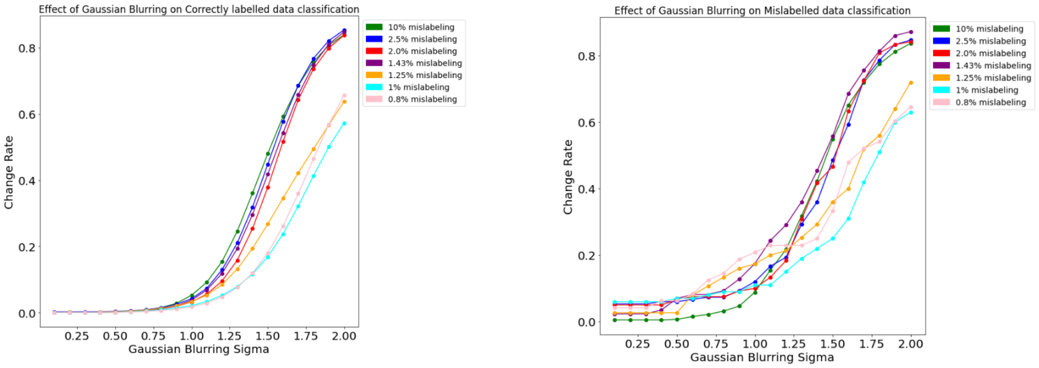
Figure 1: Gaussian Blurring on all data, split by correctly labelled and mislabelled data

4.3 Overall Differences Between Mislabelled and Correctly Labelled Data

As shown at the start of the curves in figure 11, for all percentages of data mislabelled, correctly labelled training data was significantly more likely to have unchanged classifications for low σ than for mislabelled data. For mislabelling rates greater than or equal to 1.43% mislabelled data was 10 times more likely to be incorrectly classified than correctly labelled data, and 100 times more likely for lower mislabelling rates. In all cases howevver, both mislabelled and correctly labelled data are correctly classified with an error of less than 10%, and so the model has learned to classify images from both groups. This provides evidence in support of the theory that mislabelled data is memorised, with a clear difference in the learnability of mislabelled and correctly labelled data. However, accounting for the difference in correct classification rate at the start, there seems to be only a minor difference between the change rates of mislabelled and correctly labelled data. This suggests that both groups are treated equally by the model when classifying nearby data. Figures 2(a) and 2(b)2 show that mislabelled data had curves that we far less smooth than for correctly labelled data. Since even in the case of 1% mislabelling the curves had 600 observations, this is unlikely to be solely due to the smaller observation size. Instead, the neural network may have learned a different and less smooth decision rule for classifying the mislabelled data, further suggesting latent knowledge of mistakes in the data.

4.4 The effect of the percentage of mislabelled data on classification change

Figure 11 shows a clear separation when the percentage of mislabelled data changed from 1.43%, where every 70th image was misclassified, to 1.25%. When the misclassification rate is higher an increase in σ quickly increases the likelihood for the edited image to have a different classification, and all 4 percentages checked above or at 1.43% show nearly identical curves. Below this point there is more variation, but a clearly slower change in classification after a sigma of 1.3, visualised in figure 33. This level of blurring is noticeable but still not significant to an observer. This suggests a phase transition between 1.43% and



(a) Change rate for Correctly Labelled Data

(b) Change rate for Mislabelled Data

Figure 2: Change on classification under Gaussian Blurring

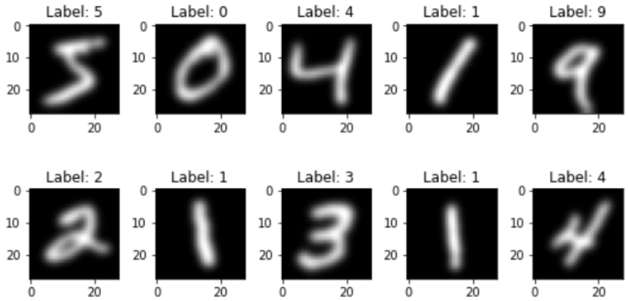


Figure 3: Gaussian Blurring with $\sigma = 1.3$

1.25%. This dramatic increase in performance in classification in the labelling, of both correctly labelled and mislabelled data, is surprising and could suggest a turning point at which mislabelled and correctly labelled data are wholly distinguished. While outside the scope of this project, this could be tested by comparison with a completely unadulterated test set, to see if it follows the same curves shown below the turning point. One suggest for this sharp transition is that the percentage of randomly mislabelled data is a *non-robust feature* in the sensed discussed in Ilyas et al.'s 2019 paper [7], however more research would need to be done here.

It should be noted that as given roughly 4% of MNIST labels are themselves wrong, so the percentage of mislabelling we give above are roughly 4% below the net mislabelling. Thus the point of change is likely to be between 5.25 and 5.4% for MNIST.

5 Conclusion

We have discussed existing literature suggesting that, even in an overparameterized regime, neural networks seem to tend to find 'low complexity' solutions. We showed the existence of a phase transition on the effect of Gaussian blurring on image classification depending on the percentage of data that is mislabelled, suggesting that at a certain point the mode learns to treat the mislabelled and correctly labelled groups more distinctly.

Surprisingly, we also found evidence to suggest that mislabeled and correctly labelled training images are treated equally in the classification of images similar to them. More research needs to be done to test this on other models, data, and forms of image transformation. In particular, it may be beneficial to test this on networks that are designed to be robust to adversarial examples, and how this treatment changes test performance.

6 References

References

[1] Yue Bai, Huan Wang, Zhiqiang Tao, Kunpeng Li, and Yun Fu. Dual lottery ticket hypothesis. *arXiv preprint arXiv:2203.04248*, 2022.

[2] Kamaludin Dingle, Chico Q Camargo, and Ard A Louis. Input–output maps are strongly biased towards simple outputs. *Nature communications*, 9(1):1–7, 2018.

[3] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

[4] Stanislav Fort and Stanislaw Jastrzebski. Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2019.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[8] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[9] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[10] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.