Detecting the Doubt Effect
Using overparameterized Deep Neural Networks and Observers' Pupillary Responses

Australian National University, School of Computing, Australian National University, Canberra ACT 2601, Australia

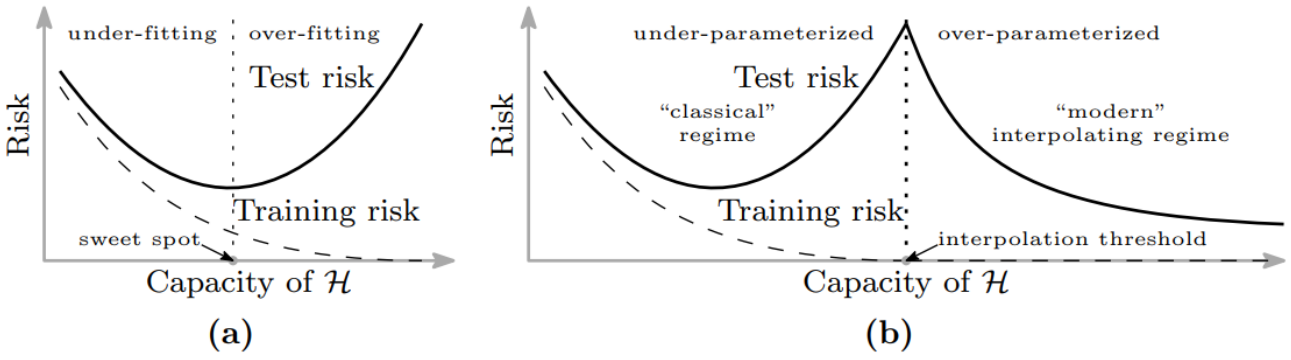
Abstract. I investigated the application of deep neural networks trained on pupillary responses to identify manipulated beliefs or doubt, which outperformed human veracity judgments. Humans struggle to recognize dishonesty consciously, with an average accuracy of 54%. The pivotal observation emerges when we consider the power of overparameterization taking a deep learning approach. This study suggests that deep neural networks that are overparameterized may unlock significantly enhanced predictive capabilities. Future research should explore this phenomenon further, potentially yielding results that outperform previous research in this field.
Keywords: neural networks, deep neural networks, pupillary responses, overparameterization, data analysis, veracity judgments.

1 Introduction

The paper written by Zhu et al. titled "Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers' Pupillary Responses", it was discovered that neural networks trained on a subject's pupillary response to stimulus videos achieved a higher accuracy in differentiating the doubt/manipulated belief compared to the observers' own veracity judgments. In a study by Bond and Depaulo [2][3] when being asked to provide direct veracity judgments of the question such as 'Is that person lying or telling the truth?', people are barely better than chance at consciously recognizing a dishonesty, with an average accuracy of 54%. This accuracy improves slightly when people's judgments are assessed indirectly even though they may not be aware that they are being lied to. Zhu achieved better performance than 54% utilizing a simple neural network meaning that there are still lots of improvements that can be made.

Belkin, M et al. wrote that Advancements in machine learning are rapidly reshaping both the field of science and broader society. However, our foundational comprehension of this technology has not kept pace with its progress. Specifically, a core concept in this field, the bias-variance trade-off, seems to clash with how modern machine learning methods operate. This trade-off suggests that a model should find a balance between underfitting and overfitting. The model should be complex enough to capture the underlying data patterns but simple enough to avoid fitting noise. Yet, in current machine learning practice, often, we train highly complex models like neural networks to precisely match (or interpolate) the data. Traditionally, such models would be considered overfit, yet they frequently achieve good accuracy on test data. In this paper, I apply the ideas of overparameterization to reach the second half of the unified performance curve. This curve, known as the "double descent" curve, incorporates the traditional U-shaped bias-variance trade-off and demonstrates that increasing model complexity beyond the point of interpolation can lead to improved performance.

Fig. 1. Curves for training risk (dashed line) and test risk (solid line). (a) U-shaped risk curve from bias-variance trade-off. (b) Double descent risk curve incorporates (a) as well as high-capacity function classes separated by the interpolation threshold.



When the capacity of a function class falls below the interpolation threshold, the learned predictors follow the classic U-shaped curve illustrated in Figure 1(a). In this context, function class capacity refers to the number of parameters needed to define a function within that class. The bottom point of the U-shaped curve represents the ideal balance between fitting the training data accurately and avoiding overfitting. To the left of this point, predictors underfit the data, while to the right, they overfit it. If we boost the function class capacity significantly by increasing the number of features or the size of the neural network architecture, the learned predictors can achieve nearly perfect fits to the training data, a phenomenon known as interpolation [1]. Interestingly, predictors obtained at the interpolation threshold often carry a high risk. However, by increasing the function class capacity beyond this threshold, the risk decreases, typically dropping below the risk attained at the sweet spot in the "classical" scenario [1].

2 Method

2.1 Data Inspection

For Zhu's paper, columns 2 through to 80 were not used as they contain data related to blood volume pulse (BVP), Galvanic Skin Response (GSR), and Skin Temperature (ST) which is irrelevant given the paper we are trying to build on is solely focused on predictions using pupillary dilation. This leaves column 1 (participant id and the id of the video) and columns 81 through to 123 where the last column contains the label of whether the presenter in this video has doubted their belief in the video. In Zhu's paper, they developed an ensemble NN using the observer's responses as the prediction rather than the ground truth labels. This, however, was not able to be reproduced as only the ground truth labels were present in the data meaning the following tasks will need to be adjusted to account for the missing data.

2.2 Data preparation

From Zhu's paper [1], the following methods have been employed to pre-process the data published:

- The extracted pupil size data were normalised across all videos viewed by each observer to reduce the effect of individual bias due to the naturally varying pupil sizes among participants, since the significant signal is the variation in the pupil size for an individual, not the magnitude of the pupil size.
- Linear interpolation was applied to missing pupil size data caused by occasional eye blinks. This procedure was employed on the pupil data of left and right eyes separately.
- The interpolated pupil size values were averaged to obtain a single pupillary response signal for each participant, and their minimum and mean of processed pupillary data during each video watching session were extracted as features.

As the data has already been pre-processed to a sufficient point, I did not deem it necessary to do further pre-processing.

2.3 Model Description

Two sets of models were trained on the observer's pupillary responses to predict the presenter's subjective beliefs jointly. All neural networks used a sigmoid activation function, 39 input neurons, two output neurons, using the Adam optimizer, with the cross-entropy loss function. The first set of model's train did not implement a dropout layers, where the second set of models trained implemented a dropout layer after each activation function (excluding final output layer. The reason for implementing two sets of models was to ensure that the effects of overfitting are minimized, while maintaining accuracy and the results can then be compared. For the first set of models (no dropout), the number of hidden units vary: 1, 5, 10, 30, 39, 50, 100, 1000, and 10000 with the number of hidden layers varying from 1 -5 (Only had 1 layer with 10000 units). For the second set of models (with dropout), the number of hidden units vary: 1, 3, 5, 10, 20, 30, 39, 50, 1000, 10000 with the number of hidden layers varying from 1 – 9 (Only had 5 layers with 1000, and 2 layers with 10000). The reason for these numbers is because it approaches overparameterization in different ways. For example, the network with 10 hidden units will take 4 layers to achieve overparameterization (10 hidden units multiplied by 4 layers is 40 units which is greater than 39 parameters thus achieving overparameterization).

As all hyperparameters have been set in this instance, validation is not necessary. As the data in question is human data, it is not always true that one segment of data will reflect a human's response thus using random splits for data will not be correct in the context of classification models. Considering this, by employing a method of leaving all data for

one human out is called leave-one-participant-out, which was used in this study. Pupillary data from one observer was used as the testing set, and those from the remaining participants formed the training set, and repeated for all, averaging to calculate the result reported.

2.4 Hyperparameters

Two hyperparameters that were tuned in the previous paper are the number of epochs and the magnitude of the weight decay. For the purpose of reasonable computation time, these have been fixed to 6000 epochs, and a weight decay magnitude of 1e-5. The reason for the number epochs is because in Belkin, M et al 2019 paper it was used, additionally, due to the high number of parameters and low amount of data, a high number of epochs are required to properly train the model. The reason for the weight decay value is because this value had good results from the previous paper.

3 Results and Discussion

As the accuracies of the network developed by Zhu cannot be directly compared in most cases, (due to not having access to the same data) some other way of comparison must be considered. However, in Zhu's paper, the relationship between Pupillary Size, Doubt Effect and Subjective Beliefs was presented. Additionally, as the verbal response (subjective belief) was reported to be a total of 50%, there should be a minimal effect on the model if this feature was removed. The total overall accuracy of the best performing model can be observed (see Table 1), achieving an accuracy of 83.1% which is far greater than the model described in Zhu's paper that achieved an accuracy of 58.3%. The main reason behind the difference in accuracy would come down to the number of epochs the model was trained on. Even though the number of epochs were not explicitly detailed in Zhu's paper, the magnitude for the number of epochs is unconventional given the context, so it is unlikely that the number used is comparable.

Table 1. Results for total accuracy of the best performant model

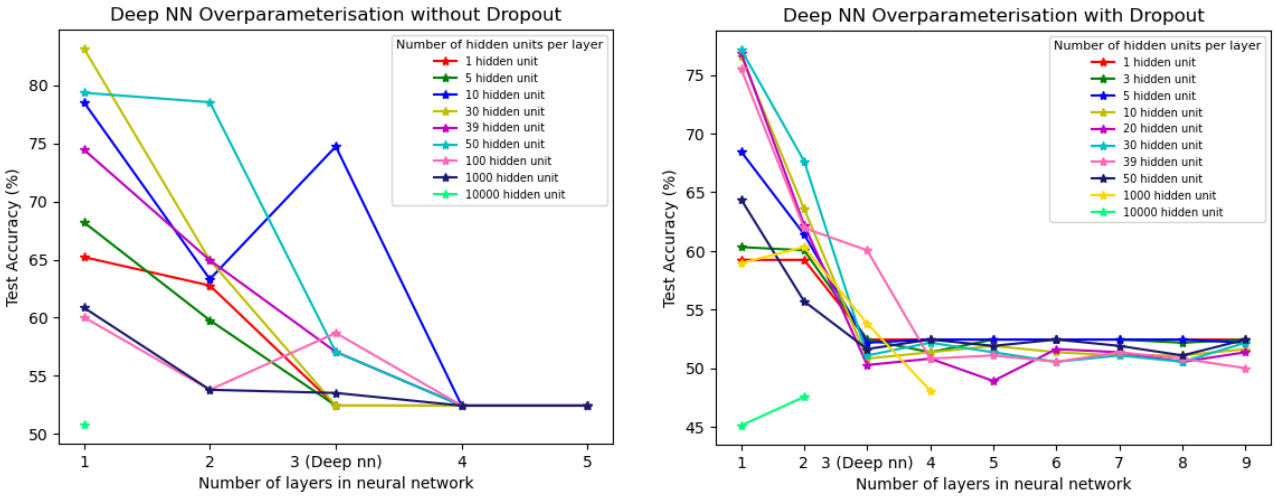
Model	Total accuracy
Zhu's model	58.3 %
NN with 30 hidden units, 1 layer, no dropout layer	83.1 %

The total accuracy of the model created is much higher than the accuracy of the model Zhu created in his paper. From an overpreparation perspective, it is the case that the best performing model is not overparameterized but is rather in the 'sweet spot' (see figure 1).

The results of increasing number of hidden layers and number of units per layer of the two sets of models trained are presented in figure 2.

4 Pupillary Responses

Fig. 2. (A) Deep NN Overparameterization without dropout varying number of hidden layers and size of hidden layers (data in Appendix A). (B) Deep NN Overparameterization with dropout varying number of hidden layers and size of hidden layers. (data in Appendix B)

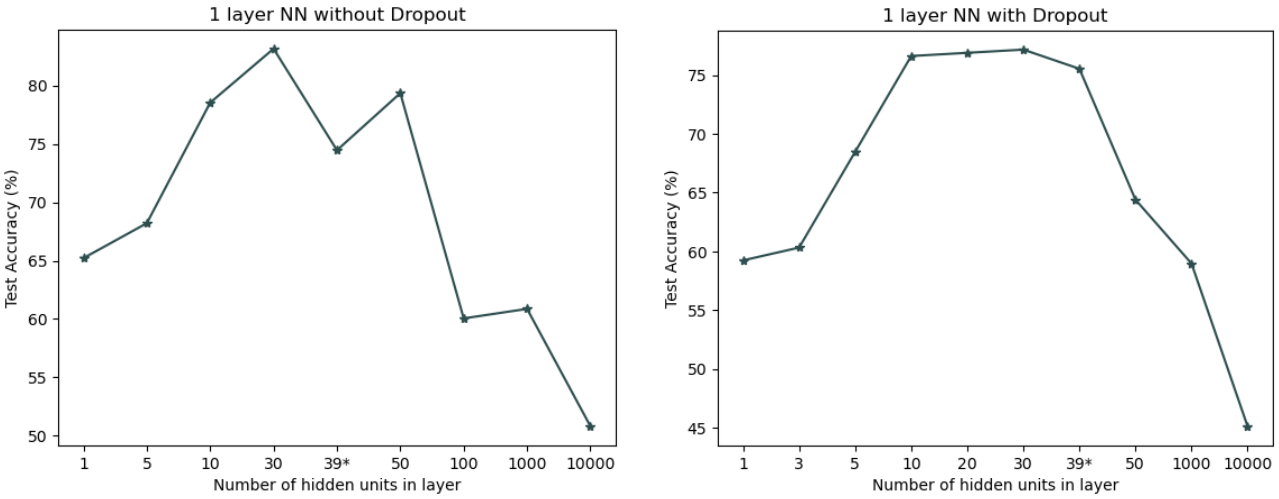


A general trend in (A) and (B) is that increasing the number of layers in the network decreases the overall accuracy of the mode. Rekha and Tyagi (2020, January) propose in their paper 'Challenges of Applying Deep Learning in Real-World Applications' that deep neural networks (3 or more layers) need a larger amount of data to perform well compared to their machine learning counterpart (less than 3 layers). In this research, it was hypothesized that a double decent curve (Fig 1) would be formed regardless of the hidden units being in a single layer or multiple layers. Instead, the double decent curve was not observed meaning that overparameterization in this instance was unsuccessful (Fig 3). An additional reason for this is because human data is complex. As deep NN's tend to extract lower-level features at the input layer and have higher lever features at the output layer [5], it might not be possible to do that in this case.

Another general trend of both (A) and (B) is that the performance of each model remains approximately the same (approximately 52%) after there is 3 layers when the model becomes a deep neural network, this is much more evident in (B) with the only exception being the model with 39 units per layer. This result is particularly interesting because it shows that in this case, the number of hidden units only effect the performance of a model with less than 3 layers.

The second set of models trained with a dropout layer had the effect of decreasing the model's performance. This is an interesting result as the high number of epochs was likely to cause overfitting as weight decay in the Adam optimiser was the only thing combatting it in the models without dropout. Additionally, the models with dropout had a higher variance in the plateau accuracy (accuracy after 3 layers). A reason for this is because there is a trade of between generalisation and overfitting reduction [1]. To discuss the results of overparameterization more accurately without deep NN's, the figures in fig 3 have been produced.

Fig. 3. (A) One-layer NN without dropout with varying number of hidden units on the x-axis and test accuracy on the y-axis (data in Appendix A). (B) One-layer NN with dropout with varying number of hidden units on the x-axis and test accuracy on the y-axis. Additionally, note that 39 has been marked with an asterisk to emphasize when the model is deemed over parameterized. (data in Appendix B)



If the data were like [1], figure 3 (in particular (A)) should be very similar to an inverted double decent curve (figure 1) where the accuracy takes a negative parabolic shape into a log shape. Instead, it was observed that the general trend of a negative parabolic shape without increase after a 'sweet spot'. This result is likely due to the complexity of the data, or lack of number of samples to train on.

Overall, it is confirmed models trained on pupillary responses for content were able to predict the veracity of the content better than humans can consciously predict.

4 Conclusion and Future Work

In our study, I looked at methods to improve Zhu's methods on using physiological signals to identify what we call the doubt effect. This is when someone's personal belief in certain information is manipulated or changed. Neural networks were trained using two pupil-related features, and the networks were surprisingly similar given that there was missing data. By increasing the number of epochs to a very large number, a large increase in the accuracy was observed from the previous paper (achieving a result of 52%). It was found that delving into the realm of overparameterization did not give any increases to accuracy. Even though a worse result was observed when the model was overparameterized, there is still future work that including getting a large data set and seeing if deep neural networks achieve as good/better results. Additionally, for future research it would be beneficial to get access to all the data used in the original experiment so that a more comprehensive analysis of the effects of overparameterization can be observed. As there is a relatively small amount of data it would be assumed that there needs to be a relatively small number of parameters to achieve good results. However, as human data is complex, I believe the model needs to have much more capacity in remembering the training set. By training a model with perhaps 10004 parameters, the model should have the capacity to achieve much better results, perhaps even better than Zhu's model without the additional data.

6 **Pupillary Responses**

5 Appendix

5.1 A Model results without dropout

#layers #parameters	1	5	10	30	39	50	100	1000	10000
1	65.2	68.2	78.5	83.1	74.4	79.3	60.0	60.9	50.8
2	62.8	59.8	63.3	64.9	64.9	78.5	53.8	53.8	N/A
3	52.4	52.4	74.7	52.4	57.1	57.1	58.7	53.5	N/A
4	52.4	52.4	52.4	52.4	52.4	52.4	52.4	52.4	N/A
5	52.4	52.4	52.4	52.4	52.4	52.4	52.4	52.4	N/A

5.2 B Model results with dropout

#layers #parameters	1	3	5	10	20	30	39	50	1000	10000
1	59.2	60.3	68.5	76.6	76.9	77.2	75.5	64.4	59.0	45.1
2	59.2	60.0	61.4	63.6	62.2	67.8	61.9	55.7	60.3	47.5
3	52.4	52.4	52.2	50.8	50.3	51.1	60.0	51.6	53.8	N/A
4	52.4	51.3	52.4	51.3	50.8	52.2	50.8	52.4	48.1	N/A
5	52.4	52.4	52.4	51.9	48.9	51.3	51.1	51.9	N/A	N/A
6	52.4	52.4	52.4	51.3	51.6	50.5	50.5	52.4	N/A	N/A
7	52.4	52.4	52.4	51.1	51.3	51.1	51.3	51.9	N/A	N/A
8	52.4	52.2	52.4	51.1	50.5	50.5	50.8	51.1	N/A	N/A
9	52.4	52.4	52.2	51.6	51.3	52.2	50.0	52.4	N/A	N/A

References

1. Belkin, M., Hsu, D., Ma, S., Mandal, S. (2019, September) Reconciling modern machine learning practice and the bias-variance trade-off
2. Bond Jr., C.F., DePaulo, B.M.: Accuracy of deception judgments. Pers. Soc. Psychol. Rev. 10, 214–234 (2006)
3. DePaulo, B.M., Bond Jr., C.F.: Beyond accuracy: bigger, broader ways to think about deceit. J. Appl. Res. Mem. Cogn. 1, 120–121 (2012)
4. Zhu, X., Qin, Z., Gedeon, T., Jones, R., Hossain, M. Z., & Caldwell, S. (2018, December). Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers' Pupillary Responses. In International Conference on Neural Information Processing (pp. 610-621). Springer, Cham.
5. Rekha, G., Tyagi, A., (2020, January). Challenges of Applying Deep Learning in Real-World Applications