The Oracle of DLphi

Dominik Alfke
University of Chemnitz

Weston Baines
Texas A&M

Jan Blechschmidt
University of Chemnitz

Mauricio J. del Razo Sarmina
FU Berlin

Amnon Drory
Tel Aviv University

Dennis Elbrächter
University of Vienna

Nando Farchmin
Physikalisch-Technische Bundesanstalt

Matteo Gambara
ETH Zürich

Silke Glas
University of Ulm

Philipp Grohs
University of Vienna

Peter Hinz
ETH Zürich

Danijel Kivaranovic
University of Vienna

Christian Kümmerle
TU Munich

Gitta Kutyniok
TU Berlin

Sebastian Lunz
University of Cambridge

Jan Macdonald
TU Berlin

Ryan Malthaner
Texas A&M

Gregory Naisat
University of Chicago

Ariel Neufeld
ETH Zürich

Philipp Christian Petersen
University of Oxford

Rafael Reisenhofer
University of Vienna

Jun-Da Sheng
UC Davis

Laura Thesing
University of Cambridge

Philipp Trunschke
TU Berlin

Johannes von Lindheim
TU Berlin

David Weber
UC Davis

Melanie Weber
Princeton University

Abstract

We present a novel technique based on deep learning and set theory which yields exceptional classification and prediction results. Having access to sufficiently large amount of labelled training data, our methodology is capable of predicting the labels of the test data almost always even if the training data is entirely unrelated to the test data. In other words, we prove in a specific setting that as long as one has access to enough data points, the quality of the data is irrelevant.

1 Introduction

This paper takes aim at achieving nothing less than the impossible. To be more precise, we seek to predict labels of unknown data from entirely uncorrelated labelled training data. This will be accomplished by an application of an algorithm based on deep learning, as well as, by invoking one of the most fundamental concepts of set theory.

Estimating the behaviour of a system in unknown situations is one of the central problems of humanity. Indeed, we are constantly trying to produce predictions for future events to be able to prepare ourselves. For example, the benefits of accurate weather forecast include allowing us to make the right choice of clothing for the day, deciding if we should go by bike or take the bus to work, or if we should seek shelter from a natural catastrophe [4]. Election predictions give companies the possibility of funding the most promising candidates and gain in influence. Analysing the history of class conflicts allows making reliable predictions of our future economic systems [27]. Finally, precise knowledge of the behaviour of the stock market enables banks to make profits and strategically destabilise the economy [10].

*All authors are aware of this work.

One of the most promising and successful techniques for such predictions are data-driven methods, most importantly *deep learning* (DL). Nonetheless, it has been conventional wisdom that not even deep learning techniques can predict labels of unseen data points if they have nothing to do with the training data. In this work, we challenge this paradigm and show that there is no need to use training data that is particularly correlated with the data points that should be predicted. Indeed, we will see below, that as long as one has a sufficient amount of training data, one can accurately predict *everything*, even events that are entirely independent of every data point used for training! In other words, not the quality of the training data is important but its sheer amount.

1.1 Related work

We will recall a couple of relevant articles that highlight the current efficiency of deep neural networks and then cite a number of our articles for no apparent reason. Neural networks were initially introduced in the 1940s by McCulloch and Pitts [28] in an attempt to mathematically model the human brain. Later, this framework was identified as a flexible and powerful computational architecture which then led to the field of deep learning [12, 23, 38]. Deep learning, roughly speaking, deals with the data-driven manipulation of neural networks. These methods turned out to be highly efficient to the extent that deep learning based methods are state-of-the-art technology in all image classification tasks [20, 41, 22]. They have revolutionised the field of speech recognition [16, 7, 47] and have achieved a level of skill in playing games that humans or the best alternative algorithms cannot match anymore [40, 43, 48].

Many mathematicians have since been trying to understand why these algorithms are superior to the classical approaches. It was observed that deep networks naturally produce invariances, which might help to explain their excellent behaviour as a high-dimensional classifier [5, 46]. The learning procedure is very unlikely to end up in bad local minima as has been observed in [44, 21]. Alternative approaches focus on approximation theory; we mention [6, 19, 2, 29, 30, 34, 39, 49] which is undoubtedly the second most biased selection of references we are about to make in this article.

A couple of additional papers that the reader should be aware of are [1, 13, 34, 9, 8, 33, 25, 35, 37, 32, 14, 11, 17, 3, 26, 45, 31]. If not for their relevance to deep learning, these articles at least give the reader an impression of the general interests of the authors.

1.2 Notation

For a set Ω, we denote by $\mathcal{P}(\Omega)$ the power set of Ω. In the following we will restrict out attention to the space $[0,1]^d$ for $d \in \mathbb{N}$ and unless specified differently we equip it with the Lebesgue measure, which in this case is a probability measure. For $1 \leq p < \infty$, we denote by $L^p([0,1]^d)$ the classical Lebesgue spaces. Finally, we denote by $[0,1]^{([0,1]^d)}$ the set of all indexed sets $(y_i)_{i\in[0,1]^d}$ where $y_i \in [0,1]$ or equivalently the set of all maps from $[0,1]^d$ to $[0,1]$.

2 Neural Networks

We start by introducing neural networks as functions that can be written as an alternating application of affine linear maps and an activation function.

Definition 2.1. *For $d, L \in \mathbb{N}$, a* neural network with activation function $\varrho : \mathbb{R} \to \mathbb{R}$ *is a function $\Phi : [0,1]^d \to \mathbb{R}$ such that there exist $N_0, \dots, N_L \in \mathbb{N}$, where $N_0 \coloneqq d$ and $N_L \coloneqq 1$ as well as affine linear maps $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ such that*

$$\Phi(x) = T_L(\varrho(T_{L-1}(\dots \varrho(T_1(x))))), \text{ for all } x \in [0,1]^d,$$

where ϱ is applied coordinate-wise. The set of all neural networks which have a representation as above is denoted by $\mathcal{NN}_{d,\varrho,L}$. Additionally, we denote

$$\mathcal{NN}_{d,\varrho} \coloneqq \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{d,\varrho,L}.$$

We recall one of the main results in neural network theory: the universal approximation theorem. The original statements are due to Hornik [18] and Cybenko [6], but we instead state a more general version describing approximation with respect to L^p norms.

Theorem 2.2 ([24]). *Let $d \in \mathbb{N}$, $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous and assume that ϱ is not a polynomial. Then, for all $L \geq 2$ and all $1 \leq p < \infty$, $\mathcal{NN}_{d,\varrho,L}$ is dense in $L^p([0,1]^d)$. In particular, $\mathcal{NN}_{d,\varrho}$ is dense in $L^p([0,1]^d)$, for $1 \leq p < \infty$.*

3 Main Results

As already announced in the introduction, we aim to construct a prediction algorithm that can produce a correct classification of unseen data based on uncorrelated training data. This formulation is sufficiently vague to be considerably impressive; however, it lacks the necessary rigour to be considered a mathematical statement. To make the statement more precise, we introduce the following definitions.

Definition 3.1. *Let μ be the Lebesgue measure on $[0,1]^d$ (which corresponds to the uniform probability distribution on $[0,1]^d$) and let ν be any discrete probability distribution on $\mathbb{N}$. We call a set X obtained as the union of $k \in \mathbb{N}$ samples $X_1, \ldots, X_k$ drawn i.i.d. according to μ a* random set of size at most k. *We call a set X generated by the following sampling procedure a* random set of finite size (or simply a finite random set)*: First draw $k \in \mathbb{N}$ according to ν, then generate a random set of size at most k as before.*

Note that for the uniform distribution μ we will almost surely generate a set of size exactly k (as opposed to at most k) by drawing k i.i.d. samples for every $k \in \mathbb{N}$. It is not hard to see that for any finite $S \subset [0,1]^d$ and any random finite set $X \subset [0,1]^d$ the probability of S and X having non-empty intersection is zero. This is independent of ν as long as the uniform measure μ is used for the sampling.

Definition 3.2. *Let $d \in \mathbb{N}$. A map $L : \mathcal{P}([0,1]^d) \times [0,1]^{([0,1]^d)} \to [0,1]^{([0,1]^d)}$ is said to possess the* precise prediction property *if for every set of labelled data $(y_i)_{i \in [0,1]^d} \in [0,1]^{([0,1]^d)}$ it holds that, for almost all finite random sets $X \subset [0,1]^d$,*

$$L\left(X, \left(y_i^X\right)_{i \in [0,1]^d}\right) = (y_i)_{i \in [0,1]^d},$$

where $y_i^X \coloneqq y_i$ if $i \notin X$ and $y_i^X \coloneqq 0$ if $i \in X$.

An honest description of Definition 3.2 is that given a ground truth $(y_i)_{i \in [0,1]^d}$ of labelled data, possibly without any inherent structure, there exists a learning algorithm that, when presented with a large subset of the labelled data (namely with $(y_i)_{i \in [0,1]^d \setminus X}$ for a generic finite $X \subset [0,1]^d$), can predict the missing data almost surely.

It turns out that there is an algorithm based on deep neural networks, that possesses the precise prediction property. We shall present this result in the following theorem. What is more is that the proof below is semi-constructive. All that is required to perform the construction of a map with the precise prediction property is an unwavering belief in the axioms of set theory, a potentially significant amount of book-keeping, and the solution of a moderate regression problem. Naturally, conventional wisdom suggests that the explicit construction of the proof can be replaced by training a sufficiently large and deep neural network via stochastic gradient descent.

Theorem 3.3. *Let $\varrho : \mathbb{R} \to \mathbb{R} : \varrho(x) \coloneqq \max\{0, x\}$. Then, for all $\epsilon > 0$, there exists a map $\mathcal{L} : \mathcal{P}([0,1]^d) \times [0,1]^{([0,1]^d)} \to \mathcal{NN}_{d,\varrho}$ such that: For every set of labelled data $(y_i)_{i \in [0,1]^d} \in [0,1]^{([0,1]^d)}$ it holds that, for almost all finite random sets $X \subset [0,1]^d$,*

$$\mathcal{L}\left(X, \left(y_i^X\right)_{i \in [0,1]^d}\right)(j) = y_j, \text{ for all } j \in X. \tag{3.1}$$

If, additionally, the map

$$f : [0,1]^d \to [0,1] : j \mapsto y_j$$

is integrable, then we have that

$$\int_{[0,1]^d} \left| \mathcal{L}\left(X, \left(y_i^X\right)_{i \in [0,1]^d}\right)(j) - f(j) \right| \, dj < \epsilon. \tag{3.2}$$

In particular, the map

$$L\colon \mathcal{P}\left([0,1]^d\right) \times [0,1]^{([0,1]^d)} \to [0,1]^{([0,1]^d)}$$
$$L(X,(y_i)_{i\in[0,1]^d})(j) \coloneqq \begin{cases} \mathcal{L}\left(X,\left(y_i^X\right)_{i\in[0,1]^d}\right)(j) & \text{if } j \in X \\ y_j & \text{else} \end{cases}$$

has the precise prediction property.

Proof. We define an equivalence relation $\sim$ on $[0,1]^{([0,1]^d)}$ by defining $(y_i)_{i\in[0,1]^d} \sim (z_i)_{i\in[0,1]^d}$ if $y_i = z_i$ for all but finitely many $i \in [0,1]^d$. We denote the equivalence class containing $(y_i)_{i\in[0,1]^d}$ by $[[(y_i)_{i\in[0,1]^d}]]$. By the axiom of choice, there exists a function λ that maps every equivalence class $\Xi \subset [0,1]^{([0,1]^d)}$ to an associated representative $(y_i)_{i\in[0,1]^d} \in \Xi$, i.e.,

$$\lambda\left([[(y_i)_{i\in[0,1]^d}]]\right) \sim (y_i)_{i\in[0,1]^d}$$

for all $(y_i)_{i\in[0,1]^d} \in [0,1]^{([0,1]^d)}$.

Now let $\epsilon > 0$. Let $X \subset [0,1]^d$ (not necessarily finite) and let $(y_i)_{i\in[0,1]^d} \in [0,1]^{([0,1]^d)}$. Let $(z_i)_{i\in[0,1]^d} \coloneqq \lambda([[(y_i)_{i\in[0,1]^d}]])$ be the representative of the equivalence class $[[(y_i)_{i\in[0,1]^d}]]$. We proceed by defining several neural networks and consider the case of infinite and finite X separately.

First, in the case of infinite X we simply set $\Phi_{\epsilon,X} = 0$.

Second, in the case of finite X let

$$g\colon [0,1]^d \to [0,1]\colon i \mapsto z_i.$$

If g is measurable and integrable then by Theorem 2.2 there exists a neural network Φ_ϵ such that

$$\int_{[0,1]^d} |g(j) - \Phi_\epsilon(j)|\, dj < \frac{\epsilon}{2}.$$

Otherwise, we set $\Phi_\epsilon = 0$. Now further, for every $k \in X$ set $r_k \coloneqq g(k) - \Phi_\epsilon(k)$ and define

$$\Phi_{k,n} : [0,1]^d \to \mathbb{R} :$$
$$j \mapsto r_k \varrho\left(\sum_{\ell=1}^{d} \varrho\left(n\left(j_\ell - k_\ell + \frac{1}{n}\right)\right) - 2\varrho\left(n\big(j_\ell - k_\ell\big)\right) + \varrho\left(n\left(j_\ell - k_\ell - \frac{1}{n}\right)\right) - (d-1)\right), \quad (3.3)$$

where j_ℓ and k_ℓ denote the ℓ-th coordinate of j and k respectively. It is not hard to see that $\Phi_{k,n}(k) = r_k$ and that there exists n^* such that for all $n \geq n^*$

$$\int_{[0,1]^d} |\Phi_{k,n}(j)|\, dj < \frac{\epsilon}{2|X|} \text{ and } \operatorname{supp}\, \Phi_{k,n} \cap \operatorname{supp}\, \Phi_{k',n} = \varnothing \text{ for } k \neq k'.$$

Now define

$$\Phi_{\epsilon,X} \coloneqq \Phi_\epsilon + \sum_{k\in X} \Phi_{k,n^*}.$$

Finally, we define

$$\mathcal{L}(X,(y_i)_{i\in[0,1]^d}) \coloneqq \Phi_{\epsilon,X}.$$

Clearly, $\mathcal{L}(X,(y_i)_{i\in[0,1]^d})$ is a neural network. Moreover, by construction, if X is finite and $k \in X$,

$$\mathcal{L}\left(X,\left(y_i^X\right)_{i\in[0,1]^d}\right)(k) = \Phi_\epsilon(k) + \Phi_{k,n^*}(k) = \lambda\left(\left(y_i^X\right)_{i\in[0,1]^d}\right)_k$$

and $\lambda\left(\left(\left(y_i^X\right)_{i\in[0,1]^d}\right) \sim \left(y_i^X\right)_{i\in[0,1]^d} \sim (y_i)_{i\in[0,1]^d}\right.$. Therefore, the set on which $\lambda\left(y_i^X\right)_{i\in[0,1]^d}\big)_k$ and $(y_i)_{i\in[0,1]^d}$ differ is finite and thus the probability of it intersecting X is zero if X is a finite random set. This concludes the first part of the theorem.

For the second part observe that if f is integrable, then since $g = f$ except on a finite set of measure zero, also g is integrable and thus by the triangle inequality

$$\int_{[0,1]^d} \left| \mathcal{L}\left(X, \left(y_i^X\right)_{i\in[0,1]^d}\right)(j) - f(j) \right| \, dj \leq \int_{[0,1]^d} |\Phi_\epsilon(j) - f(j)| \, dj + \sum_{k\in X} \int_{[0,1]^d} |\Phi_{k,n^*}(j)| \, dj < \epsilon.$$

This completes the proof. □

4 Discussion

Theorem 3.3 implies that deep neural networks are, in principle, able to predict data, that is completely uncorrelated with the training data. This observation might appear unintuitive, stunning, and maybe unbelievable at first, but the same can be said for most other deep learning results. We shall address a number of limitations and observations about Theorem 3.3 below.

In a sense, the algorithm requires an uncountable number of data points. Such a situation is often excluded in applications despite the current trend of big data. It is, however, expected that the amount of data that we need to handle will increase with time. Following this reasoning, scholars have speculated that we might soon enter the era of *very big data*.

Many machine learning algorithms suffer from overfitting, i.e., their predictions are overly adapted to the training data, and they do not generalise to unseen data. Equation (3.1) shows that the neural network produced by $\mathcal{L}$ predicts the labels of the points in X correctly with probability 1. However, on the training set, the accuracy is a bit worse as demonstrated by Equation (3.2). In other words, $\mathcal{L}$ is not overfitting, but deliberately underfitting while the prediction remains accurate.

It has often been conjectured that deep learning techniques can remove the curse of dimensionality that plagues many high-dimensional approximation tasks. Careful inspection of the statement of Theorem 3.3 indeed reveals that the quality of prediction appears to be independent of the ambient dimension d. Another piece of conventional wisdom is that deep networks are more efficient than their shallow counterparts, [42, 36]. The construction of (3.3) requires at least two hidden layers. Thus, the network is technically not a shallow network. It is fair say that this observation gives yet another illustration of the power of depth.

Finally, we mention that the construction is inspired by a prediction strategy for ordered data, presented in [15]. The discussion in that paper indicates that the same result is most likely not valid in classical Zermelo-Fraenkel set theory. It might not be wise to prominently suggest that the current deep learning revolution and the overwhelming empirical success prove that the axiom of choice is an empirically reasonable assumption, but we leave the reader to her own conclusion/s.

Acknowledgements

All authors except for Philipp Grohs are indebted to Philipp Grohs for giving an especially lovely lecture at the Oberwolfach Seminar "Deep Learning" in October of 2018. This lecture led the participants to profound insights and culminated in this article and hopefully many more. The authors are grateful for the hospitality of the Mathematisches Forschungsinstitut Oberwolfach which hosted this fantastic event.

Disclaimer

We concede that this manuscript has mostly humouristic value. However, while the main theorem is based on a dubious application of the axiom of choice, it is a correct mathematical statement. Therefore, this manuscript at least highlights the dangers of applying mathematical theory to real-world applications blindly.

References

[1] D. Alfke, D. Potts, M. Stoll, and T. Volkmer. NFFT meets Krylov methods: Fast matrix-vector products for the graph Laplacian of fully connected networks. *arXiv preprint arXiv:1808.04580*, 2018.

[2] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

[3] J. Blechschmidt and R. Herzog. Improving policies for hamilton-jacobi-bellman equations by postprocessing. 2018.

[4] M. E. Brown. *Famine early warning systems and remote sensing data*. Springer Science & Business Media, 2008.

[5] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[6] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal*, 2(4):303–314, 1989.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Audio, Speech, Language Process.*, 20(1):30–42, 2012.

[8] M. Del Razo and R. LeVeque. Numerical methods for interface coupling of compressible and almost incompressible fluids. *SIAM J. Sci. Comput., submitted*, 2016.

[9] A. Drory, C. Haubold, S. Avidan, and F. A. Hamprecht. Semi-global matching: a principled derivation in terms of message passing. In *German Conference on Pattern Recognition*, pages 43–53. Springer, 2014.

[10] I. M. Fund. Global financial stability report: A bumpy road ahead. 2018. Washington, DC, April.

[11] S. Glas, A. Mayerhofer, and K. Urban. Two ways to treat time in reduced basis methods. In *Model Reduction of Parametrized Systems*, pages 1–16. Springer, 2017.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[13] P. Grohs and G. Kutyniok. Parabolic molecules. *Foundations of Computational Mathematics*, 14(2):299–337, 2014.

[14] A. Hansen and L. Thesing. On the stable sampling rate for binary measurements and wavelet reconstruction. *Applied and Computational Harmonic Analysis*, 2018.

[15] C. S. Hardin and A. D. Taylor. A peculiar connection between the axiom of choice and predicting the future. *The American Mathematical Monthly*, 115(2):91–96, 2008.

[16] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012.

[17] P. Hinz and S. van de Geer. A framework for the construction of upper bounds on the number of affine linear regions of ReLU feed-forward neural networks. *arXiv preprint arXiv:1806.01918*, 2018.

[18] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[19] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, 1989.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[21] K. Kawaguchi. Deep learning without poor local minima. In *Adv. Neural Inf. Process. Syst.*, pages 586–594, 2016.

[22] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[23] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[24] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[25] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. *arXiv preprint arXiv:1805.11572*, 2018.

[26] J. Macdonald and L. Ruthotto. Improved susceptibility artifact correction of echo-planar mri using the alternating direction method of multipliers. *Journal of Mathematical Imaging and Vision*, 60(2):268–282, 2018.

[27] K. Marx. Das Kapital: Kritik der politischen ökonomie. *Germany: Verlag von Otto Meisner*, 1885:1894, 1867.

[28] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.

[29] H. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80, Feb 1993.

[30] H. N. Mhaskar and C. Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Adv. Appl. Math.*, 13:350–373, 1992.

[31] B. Monshi, M. Vujic, D. Kivaranovic, A. Sesti, W. Oberaigner, I. Vujic, S. Ortiz-Urda, C. Posch, H. Feichtinger, M. Hackl, et al. The burden of malignant melanoma–lessons to be learned from Austria. *European Journal of Cancer*, 56:45–53, 2016.

[32] A. Neufeld, M. Nutz, et al. Superreplication under volatility uncertainty for measurable claims. *Electronic journal of probability*, 18, 2013.

[33] D. Perekrestenko, P. Grohs, D. Elbrächter, and H. Bölcskei. The universal approximation power of finite-width deep ReLU networks. *arXiv preprint arXiv:1806.01528*, 2018.

[34] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

[35] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018.

[36] I. Safran and O. Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2979–2987, 2017.

[37] N. Saito and D. S. Weber. Underwater object classification using scattering transform of sonar signals. In *Wavelets and Sparsity XVII*, volume 10394, page 103940K. International Society for Optics and Photonics, 2017.

[38] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[39] U. Shaham, A. Cloninger, and R. Coifman. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.* to appear.

[40] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] M. Telgarsky. Benefits of depth in neural networks. arXiv:1602.04485.

[43] N. Usunier, G. Synnaeve, Z. Lin, and S. Chintala. Episodic exploration for deep deterministic policies: An application to StarCraft micromanagement tasks. *arXiv preprint arXiv:1609.02993*, 2016.

[44] R. Vidal, J. Bruna, R. Giryes, and S. Soatto. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017.

[45] M. Weber, E. Saucan, and J. Jost. Coarse geometry of evolving networks. *Journal of Complex Networks*, 6(5):706–732, 2017.

[46] T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Trans. Inf. Theory*, 64(3):1845–1866, 2018.

[47] C. Wu, P. Karanasou, M. J. Gales, and K. C. Sim. Stimulated deep neural network for speech recognition. Technical report, University of Cambridge, 2016.

[48] G. N. Yannakakis and J. Togelius. *Artificial Intelligence and Games*. Springer, 2017.

[49] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103 – 114, 2017.