Journal of ISOSS
2023 Vol. 9(1), 131-146

THE ROLE OF ARTIFICIAL INTELLIGENCE AND DIGITAL ERA IN MATHEMATICS

Maawish Bashir[1] and **Kehkashan Nizam**[2]

[1] Department of Mathematics, JS Public School and College Punjab,
Rawalpindi, Pakistan. Email: mahwish.jspublicschoolcollege@gmail.com

[2] Department of Business Administration, Iqra University, Karachi, Pakistan.
Email: kehkashan.60003@iqra.edu.pk

ABSTRACT

Artificial intelligence is currently enjoying astounding success in both research and everyday life in modern digital era. Nonetheless, it is still early in the process of creating a solid mathematical foundation. This comprehensive essay, which is based on a lecture given at the International Congress of Mathematicians 2022, will pay special attention to deep neural networks, the "workhorse" of artificial intelligence at the moment. We'll outline the main theoretical axes, show a few concrete findings, and talk about the main unresolved issues.

1. INTRODUCTION

Currently, artificial intelligence is producing success after breakthrough in a variety of fields, including the sciences, autonomous vehicles, speech recognition, and molecular dynamics. A record-breaking amount of research is also being done on artificial intelligence, particularly its theoretical underpinnings. According methodologies have the potential to fundamentally alter how we live in the future in a number of ways. But the concept of artificial intelligence is not new. In reality, McCulloch and Pitts began to develop algorithmic learning methods in 1943. They did this by creating artificial neurons that are connected and stacked in layers to create artificial neural networks, which replicate the operation of the human brain. They already had a plan in mind for how artificial intelligence might be used at the time. The community did not, however, fully appreciate neural networks' potential. As a result, this initial wave of artificial intelligence failed and disappeared (Bach et al., 2015).

Machine learning regained popularity about 1980, and numerous notable developments came out of that time period. Deep neural networks have exactly the same structure that McCulloch and Pitts proposed, which consists of several successive layers of artificial neurons. the two primary. Even while all of these changes appear to be hopeful, a word of caution is necessary. Together with the fact that neural networks are still seen as a "jack of all crafts" and that the practical constraints of approaches like deep neural networks have not even been remotely examined, it is alarming that there is no full theoretical groundwork. The phrase "Machine learning has become a type of alchemy" was made abundantly clear during the largest conference on artificial intelligence and machine learning, NIPS (now named NeurIPS), in 2017, when Ali Rahimi from Google received the Test of Time Prize.

In the case of deep neural networks, for example, this lack of mathematical underpinnings leads to a time-consuming search for an appropriate network architecture, a highly delicate trial-and-error-based (training) process, and the absence of error bounds for the trained neural network's performance. To address the multiple tough yet intriguing tasks in the field of artificial intelligence, practically all branches of mathematics are needed (Hornik, K., Stinchcombe, M., & White, H., 1989).

Artificial Intelligence Mathematical Foundations. This approach seeks to derive a profound understanding of mathematics. Based on this, it works to eliminate current challenges like a lack of robustness or theoretically bases the entire training process. Mathematical problems and artificial intelligence. This course emphasises the use of partial differential equations and inverse problems as settings for mathematical problem solving.

2. THE MATHEMATICAL SETTING OF ARTIFICIAL INTELLIGENCE

2.1. Definition of Deep Neural Networks

Artificial neurons form the fundamental components of this system. These neurons are modeled after the structure and functionality of neurons in the human brain (Cybenko, 1989). A neuron in the human brain consists of dendrites, which receive signals and transmit them to the soma. The signals are scaled and amplified as they pass through the dendrites. In the soma, the incoming signals are accumulated and a decision is made on whether or not to fire to other neurons, and with what strength (Donoho, 2001).

Definition 2.1. An artificial neuron with weights $w_1, \dots, w_n \in \mathbb{R}$, bias $b \in \mathbb{R}$, and activation function $\rho: \mathbb{R} \to \mathbb{R}$ is defined as the function $f: \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x_1, \dots, x_n) = \rho\left(\sum_{i=1}^{n} x_i w_i - b\right) = \rho(\langle x, w\rangle - b),$$

where $w = (w_1, \dots, w_n)$ and $x = (x_1, \dots, x_n)$.

Currently, there is a variety of activation functions available, with the most commonly known ones being:

(1) Heaviside function $\rho(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$

(2) Sigmoid function $\rho(x) = \frac{1}{1+e^{-x}}$.

(Andrade-Loarca et al., 2022) Rectifiable Linear Unit (ReLU) $\rho(x) = \max\{0, x\}$.

Artificial neurons are combined and organized into layers, similar to the structure of the human brain, resulting in an artificial feed-forward neural network. The structure of artificial neurons leads to a neural network consisting of affine linear maps and activation functions. The resulting function is traditionally defined as a deep neural network. However, different arrangements can lead to the same function, leading to the possibility of making a distinction between the architecture of a neural network and the corresponding realization function. For the purpose of this article, technical details of this nature will be avoided and the most standard definition will be presented.

Definition 2.2. Let $d \in \mathbb{N}$ be the dimension of the input layer, L the number of layers, $N_0 := d, N_\ell, \ell = 1, \ldots, L$, the dimensions of the hidden and last layer, $\rho: \mathbb{R} \to \mathbb{R}$ a (non-linear) activation function, and, for $\ell = 1, \ldots, L$, let T_ℓ be the affine-linear functions

$$T_\ell: \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell},\ T_\ell x = W^{(\ell)} x + b^{(\ell)},$$

with $W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ being the weight matrices and $b^{(\ell)} \in \mathbb{R}^{N_\ell}$ the bias vectors of the ℓ th layer. Then $\Phi: \mathbb{R}^d \to \mathbb{R}^{N_L}$, given by

$$\Phi(x) = T_L \rho \left(T_{L-1} \rho (\ldots \rho(T_1(x))) \right),\ x \in \mathbb{R}^d$$

is called (deep) neural network of depth L.

Figure 1 provides a visualization of the multilayered structure of a deep neural network.

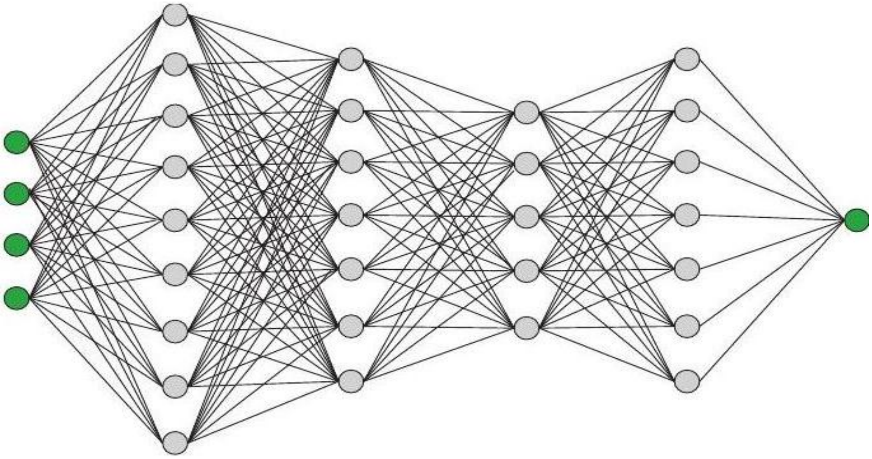


Figure 1: Deep Neural Network Φ: $\mathbb{R}^4 \to \mathbb{R}$ with Depth

2.2. Application of a Deep Neural Network

To identify the primary areas of mathematical research in deep neural networks, it is essential to comprehend the manner in which a deep neural network is employed within a given application context.

Step 1 (Train-test split of the dataset): We assume that we are given samples $\left(x^{(i)}, y^{(i)}\right)_{i=1}^{\tilde{m}}$ of inputs and outputs. The role of the deep neural network is to recognize the relationship between the data and the corresponding output. It should be noted that the network is not exposed to the test data set during the entire training process.

Step 2 (Choice of architecture): To prepare the learning algorithm, it is necessary to determine the architecture of the neural network. This involves selecting the number of layers L, in each layer the number of neurons $(N_\ell)_{\ell=1}^{L}$, and the function of activation ρ have to be selected.

Step 3 (Training): Next step is the process of actual training that including learning the affine-linear functions $(T_\ell)_{\ell=1}^{L} = \left(W^{(\ell)} \cdot + b^{(\ell)}\right)_{\ell=1}^{L}$. Risk minimization has accomplished.

$$\hat{\mathcal{R}}\left(\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}\right) := \frac{1}{m}\sum_{i=1}^{m}\left(\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Step 4 (Testing): Once the deep neural network has been trained, its performance, also known as generalization ability, is evaluated using the test data set $\left(x^{(i)}, y^{(i)}\right)_{i=m+1}^{\tilde{m}}$ by analyzing whether

$$\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}\left(x^{(i)}\right) \approx y^{(i)}, \text{ for all } i = m+1, \dots, \tilde{m}.$$

2.3. Relation to a Statistical Learning Problem

Based on the aforementioned procedure, the selection of architecture, optimization problem, and generalization ability can be identified as the primary research directions for the mathematical foundations of deep neural networks. Considering the entire learning process of a deep neural network as a statistical learning problem highlights these three research directions as the most relevant for analyzing the overall error.

Assume function $g: \mathbb{R}^d \to \mathbb{R}$ such that the training data $\left(x^{(i)}, y^{(i)}\right)_{i=1}^{m}$ is of the form $\left(x^{(i)}, g\left(x^{(i)}\right)\right)_{i=1}^{m}$ and $x^{(i)} \in [0,1]^d$ for all $i = 1, \dots, m$. One common approach to evaluate the success of the training is to consider the risk of a function, as viewed from a continuum perspective $f: \mathbb{R}^d \to \mathbb{R}$ given by

$$\mathcal{R}(f) := \int_{[0,1]^d} (f(x) - g(x))^2 dx$$

L^2-norm is used to measure the distance between f and g. The error between the trained deep neural network $\Phi^0 \left(:= \Phi_{(W^{(\ell)},b^{(\ell)})_\ell}\right) \in \mathcal{NN}_\theta$ and the optimal function g can then be estimated by

$$\mathcal{R}(\Phi^0) \leq \underbrace{\left[\hat{\mathcal{R}}(\Phi^0) - \inf_{\Phi\in\mathcal{NN}_\theta} \hat{\mathcal{R}}(\Phi)\right]}_{\text{Optimization error}} + \underbrace{2 \sup_{\Phi\in\mathcal{NN}_\theta} \left|\mathcal{R}(\Phi) - \hat{\mathcal{R}}(\Phi)\right|}_{\text{Generalization error}} + \underbrace{\inf_{\Phi\in\mathcal{NN}_\theta} \mathcal{R}(\Phi)}_{\text{Approximation error}}.$$

2.4. Main Research Threads

Two conceptually distinct research threads can be identified, the first one being dedicated to developing mathematical foundations for artificial intelligence, while the second aims to leverage artificial intelligence methodologies to tackle mathematical problems. It is fascinating to observe how both have already sparked a paradigm shift in some mathematical research areas, particularly in the field of numerical analysis (Bolcskei et al., 2019).

2.4.1. Mathematical Foundations for Artificial Intelligence

Following up on the discussion in Subsection 2.3, we can identify three research directions which are related to the three types of errors which one needs to control in order to estimate the overall error of the entire training process.

- Expressivity. This direction aims to derive a general understanding whether and to which extent aspects of a neural network architecture affect the best case performance of deep neural networks. More precisely, the goal is to analyze the approximation error $\inf_{\Phi \in \mathcal{N}\mathcal{N}_\theta} \mathcal{R}(\Phi)$ from 2.4, which estimates the approximation accuracy when approximating g by the hypothesis class $\mathcal{N}\mathcal{N}_\theta$ of deep neural networks of a particular architecture. Typical methods for approaching this problem are from applied harmonic analysis and approximation theory.
- Learning/Optimization. The main goal of this direction is the analysis of the training algorithm such as stochastic gradient descent, in particular, asking why it usually converges to suitable local minima even though the problem itself is highly non-convex.
- Generalization. This direction aims to derive an understanding of the out-of-sample error, namely,
- Explainability. This direction considers deep neural networks, which are already trained, but no knowledge about the training is available; a situation one encounters numerous times in practice.

2.4.2. Artificial Intelligence for Mathematical Problems

Similar to the previous subsection on mathematical foundations of artificial intelligence, we can distinguish two research threads related to the use of methods of artificial intelligence for mathematical problem settings. The first one focuses on developing mathematical tools and techniques that leverage the capabilities of deep learning and other AI algorithms to tackle mathematical problems in various fields, from scientific computing and numerical analysis to algebraic geometry and optimization. This research direction aims to understand the strengths and limitations of AI-based methods in mathematical contexts, to design new algorithms and models that exploit the intrinsic structures and properties of mathematical objects, and to provide rigorous analysis and validation of their performance.

The second research thread concerns the application of AI methods to specific mathematical problems, ranging from solving differential equations and inverse problems to discovering new conjectures and proving theorems. This research direction involves identifying suitable representations and architectures for the input data and the target output, designing appropriate loss functions and training algorithms, and exploring the interplay between AI and traditional mathematical methods. The ultimate goal is to develop new computational tools and insights that can accelerate and enhance the discovery and verification of mathematical knowledge, as well as to open up new avenues for interdisciplinary research and applications.

- Inverse Problems. The research in this direction aims to enhance classical model-based approaches for solving inverse problems by utilizing methods of artificial intelligence. To avoid disregarding domain knowledge such as the physics of the problem, current approaches aim to optimally combine model- and data-driven methods to take the best of both worlds.
- Partial Differential Equations. In the area of partial differential equations, the objective is to enhance classical solvers of partial differential equations by utilizing concepts from artificial intelligence.

3. MATHEMATICAL FOUNDATIONS FOR ARTIFICIAL INTELLIGENCE

This section will provide an overview of the key research directions focused on developing a mathematical foundation for artificial intelligence. We will present the problem settings, highlight some notable achievements, and explore current open questions.

3.1 Expressivity

Expressivity is an active area of research that has produced many mathematical results. The central question is: given a function class/space $\mathcal{C}$ and deep neural networks class $\mathcal{NN}_\theta$, approximation accuracy if elements approximating of $\mathcal{C}$ by networks $\Phi \in \mathcal{NN}_\theta$ relate to the complexity of such Φ ? To answer this question, a complexity measure for deep neural networks needs to be introduced. The most common complexity measure is the memory requirements of the network. However, other complexity measures may exist. It is important to note that the $\|\cdot\|_0$-"norm" counts the non-zero components numbers.

3.2 Optimization

This area of research aims to analyze optimization algorithms that are used to solve the learning problem described in 2.1) or 2.2). A common approach is to use gradient descent, as the gradient of the loss function (or optimized functional) with respect to the weight matrices and biases can be computed exactly. The process of computing these gradients is known as backpropagation, which is essentially an efficient application of the chain rule. However, given that the number of training samples is typically in the millions, it is computationally infeasible to compute the gradient on each sample. Therefore, in each iteration, only one or a small batch of randomly selected gradients are computed, leading to the stochastic gradient descent algorithm [25].

The convergence of stochastic gradient descent can be guaranteed in convex settings. However, in the context of neural networks, the optimization problem is non-convex, which makes it difficult to analyze even when using a non-random version of gradient descent. The problem becomes more challenging when including randomness, as shown in Figure 2 where the two algorithms converge to different local minima.

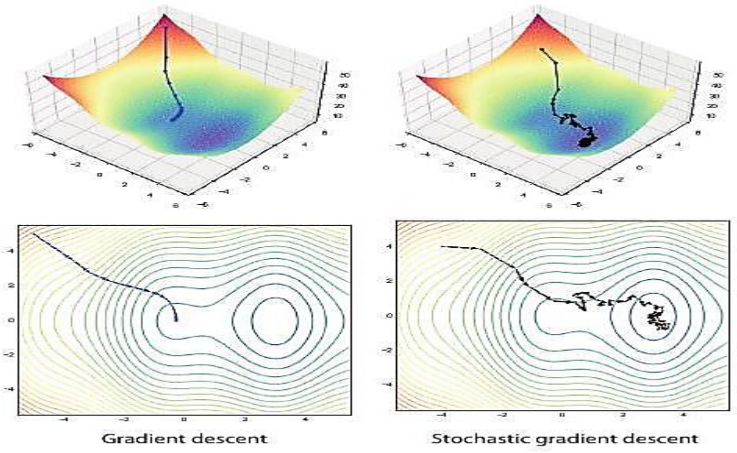


Figure 2: Gradient Descent Versus Stochastic Gradient Descent [6]

This area is by far less explored than expressivity. Recently focus on shallow neural networks, and for a survey as in [6].

3.3 Generalization

This research direction is concerned with understanding the out-of-sample error of deep neural networks and is often referred to as the "holy grail" of understanding deep neural networks. It targets the out-of-sample error

$$\sup_{\Phi \in \mathcal{N}\mathcal{N}_\theta} |\mathcal{R}(\Phi) - \hat{\mathcal{R}}(\Phi)|$$

as described in Subsection 2.4.1.

Unlike expressivity and optimization, generalization is the least explored area. The goal is to determine how well a trained model can generalize to new, unseen data. It is a mystery why highly overparameterized deep neural networks, with a high complexity, do not overfit, which refers to the problem of fitting the training data too tightly and consequently failing to correctly classify new data. Overfitting can be seen in Figure 3.

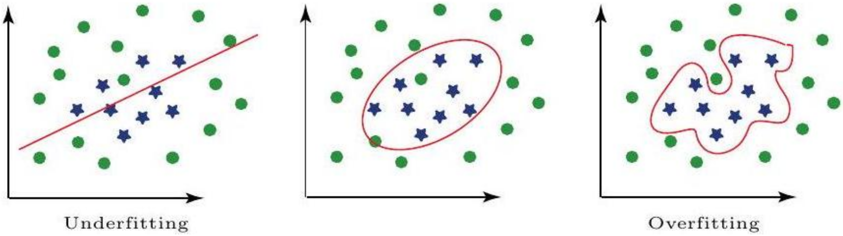


Figure 3: Phenomenon of Overfitting for the Task of Classification with Two Classes

Now, generalization error in (3.1) in a depth. For a large number m of training samples the law of large numbers show high probability $\hat{\mathcal{R}}(\Phi) \approx \mathcal{R}(\Phi)$ for each neural network $\Phi \in \mathcal{N}\mathcal{N}_\theta$. Bounding the complexity of the hypothesis class $\mathcal{N}\mathcal{N}_\theta$ by the VC-dimension, the generalization error can be bounded with probability $1 - \delta$ by

$$\sqrt{\frac{\text{VCdim}(\mathcal{N}\mathcal{N}_\theta) + \log(1/\delta)}{m}}$$

Research on generalization in deep neural networks is still in its early stages, and it is considered the most challenging research direction. The focus is on out-of-sample error, as explained in Subsection 2.4.1. One of the interesting phenomena of deep neural networks is that highly overparameterized networks do not overfit even though they have high complexity. For classes of highly over-parametrized neural networks, i.e., where $\text{VCdim}(\mathcal{N}\mathcal{N}_\theta)$ is very large. Overfitting refers to the problem of fitting the training data too tightly, which can lead to incorrect classification of new data. However, it is still not well-understood why this happens. When dealing with classes of highly over-parameterized neural networks with a large VC dimension, a considerable amount of training data is needed to control generalization error. It is surprising that numerical experiments have shown the phenomenon of a double descent curve (Belkin et al., 2019). After passing the interpolation point, the test error decreases, followed by an increase consistent with statistical learning theory (see Figure 4).

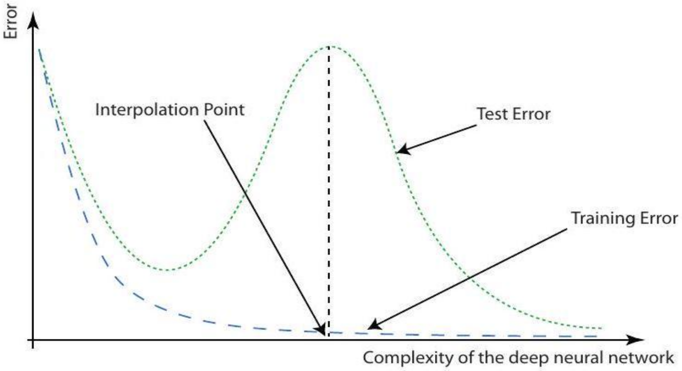


Figure 4: Double Descent Curve

3.4 Explainability

The area of explainability aims to understand the decision-making process of deep neural networks and provide explanations for their decisions. This involves providing relevance scores for features of the input data, and most approaches focus on image classification tasks. There are various types of approaches, including gradient-based methods, propagation of activations in neurons, surrogate models, and game-theoretic approaches.

One specific approach is based on information theory and is known as the rate-distortion explanation (RDE) framework. It allows for an extension to different modalities, such as audio data, and can analyze the relevance of higher-level features. This framework was introduced in 2019 and has since been extended to non-canonical input representations. A survey paper on this topic can be found in [15].

Let now $\Phi: \mathbb{R}^d \to \mathbb{R}^n$ be a trained neural network, and $x \in \mathbb{R}^d$. The goal of RDE is to provide an explanation for the decision $\Phi(x)$ in terms of a sparse mask $s \in \{0,1\}^d$ which highlights the crucial input features of x. This mask is determined by the following optimization problem:

$$\min_{s \in \{0,1\}^d} \mathop{\mathbb{E}}_{v \sim \mathcal{V}} d(\Phi(x), \Phi(x \odot s + (1-s) \odot v)) \quad \text{subject to} \quad \| s \|_0 \leq \ell,$$

where $\odot$ denotes the Hadamard product, d is a measure of distortion such as the ℓ_2-distance, $\mathcal{V}$ is a distribution over input perturbations $v \in \mathbb{R}^d$, and $\ell \in \{1, \ldots, d\}$ is a given sparsity level for the explanation mask s. The key idea is that a solution s^* is a mask marking few components of the input x which are sufficient to approximately retain the decision $\Phi(x)$. This viewpoint reveals the relation to rate-distortion theory, which normally focusses on lossy compression of data.

Since it is computationally infeasible to compute such a minimizer (see [30], a relaxed optimization problem providing continuous masks $s \in [0,1]^d$ is used in practice:

$$\min_{s\in[0,1]^d} \mathop{\mathbb{E}}_{v\sim\mathcal{V}} d(\Phi(x), \Phi(x \odot s + (1-s) \odot v)) + \lambda \parallel s \parallel_1,$$

where $\lambda > 0$ determines the sparsity level of the mask. The minimizer now assigns each component x_i of the input - in case of images each pixel - a relevance score $s_i \in [0,1]$. This is typically referred to as Pixel RDE.

Extensions of the RDE framework have allowed for the incorporation of different distributions V that are better adapted to the data distributions. Additionally, relevance scores can be assigned to higher-level features, such as those arising from wavelet decomposition, which led to the development of the CartoonX approach. A comparison between Pixel RDE and CartoonX is shown in Figure 5, which also demonstrates the ability of the higher-level explanations of CartoonX to provide insights into what the neural network saw when misclassifying an image.

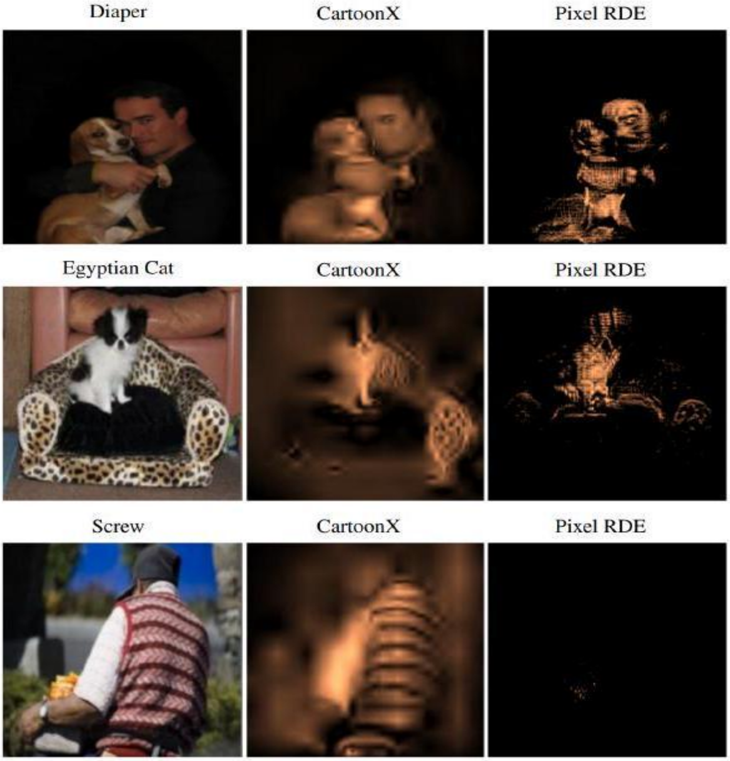


Figure 5: Pixel RDE versus CartoonX for Analyzing Misclassifications of a Deep Neural Network

4. ARTIFICIAL INTELLIGENCE FOR MATHEMATICAL PROBLEMS

We now turn to the research direction of artificial intelligence for mathematical problems, with the two most prominent problems being inverse problems and partial differential equations. As before, we will introduce the problem settings, showcase some exemplary results, and also discuss open problems.

4.1 Inverse Problems

The field of artificial intelligence, specifically deep neural networks, has had a significant impact on inverse problems. One current trend is to optimally combine classical solvers with deep learning to benefit from both worlds. To understand such results, let's review some basics about solvers of inverse problems. An inverse problem is considered ill-posed.

$$Kf = g$$

where $K: X \to Y$ is an operator and X and Y are, for instance, Hilbert spaces. Drawing from the area of imaging science, examples include denoising, deblurring, or inpainting (recovery of missing parts of an image). Most classical solvers are of the form (which includes Tikhonov regularization)

$$f^{\alpha} := \underset{f}{\operatorname{argmin}}[\underbrace{\| Kf - g \|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\mathcal{P}(f)}_{\text{Penalty/Regularization term}}],$$

where $\mathcal{P}: X \to \mathbb{R}$ and $f^{\alpha} \in X, \alpha > 0$ is an approximate solution of the inverse problem 4.1). One very popular and widely applicable special case is sparse regularization, where $\mathcal{P}$ is chosen by

$$\mathcal{P}(f) := \|(\langle f, \varphi_i \rangle)_{i \in I}\|_1$$

and $(\varphi_i)_{i \in I}$ is a suitably selected orthonormal basis or a frame for X.

We now turn to deep learning approaches to solve inverse problems, which might be categorized into three classes:

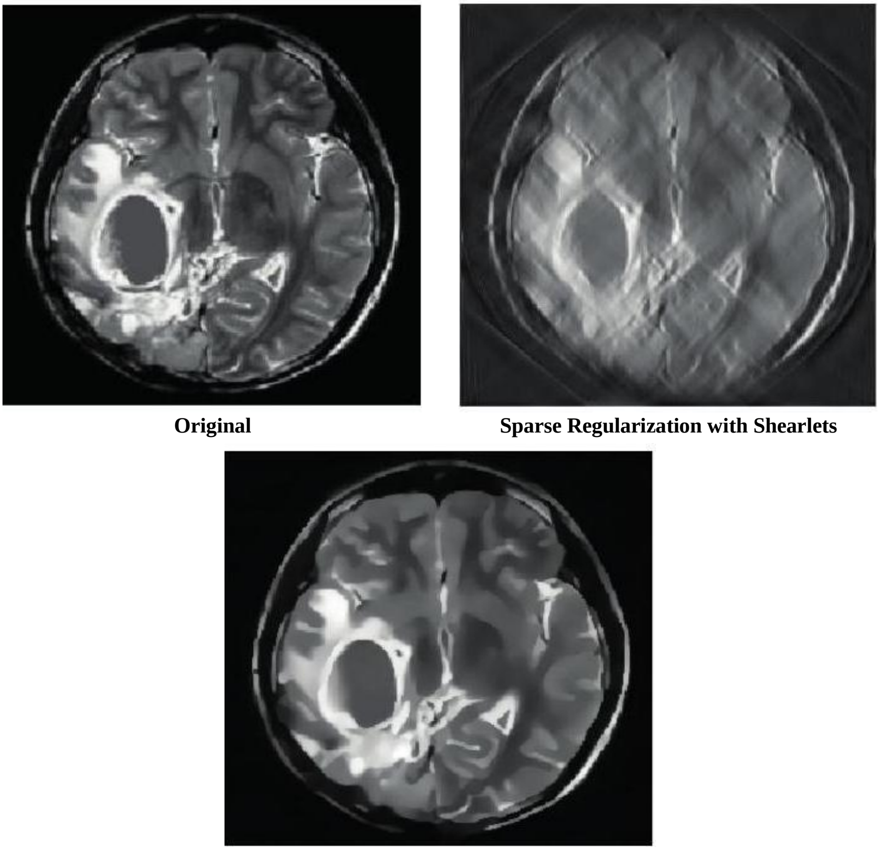
- Supervised approaches. An ad-hoc approach in this regime is given in (Jin, K. H., McCann, M. T., Froustey, E., & Unser, M., 2017), which first applies a classical solver followed by a neural network to remove reconstruction artifacts. More sophisticated approaches typically replace parts of the classical solver by a custom-build neural network [26] or a network specifically trained for this task (Adler & Öktem, 2017).
- Semi-supervised approaches. These approaches encode the regularization as a neural network with an example being adversarial regularizers 20.
- Unsupervised approaches. A representative of this type of approaches is the technique of deep image prior [29. This method interestingly shows that the structure of a generator network is sufficient to capture necessary statistics of the data prior to any type of learning.

Aiming to illustrate the superiority of approaches from artificial intelligence for inverse problems, we will now focus on the inverse problem of computed tomography (CT) from medical imaging. The forward operator K in this setting is the Radon transform, defined by

$$\mathcal{R}f(\phi, s) = \int_{L(\phi,s)} f(x) dS(x)$$

where $L(\phi, s) = \{x \in \mathbb{R}^2 : x_1 \cos(\phi) + x_2 \sin(\phi) = s\}, \phi \in [-\pi/2, \pi/2),$ and $s \in \mathbb{R}$. Often, only parts of the so-called sinogram $\mathcal{R}f$ can be acquired due to physical constraints as in, for instance, electron tomography. The resulting, more difficult problem is termed limited-angle CT. One should notice that this problem is even harder than the problem of low-dose CT, where not an entire block of measurements is missing, but the angular component is "only" undersampled.

The most prominent features in images f are edge structures. This is also due to the fact that the human visual system reacts most strongly to those. These structures in turn can be accurately modeled by microlocal analysis, in particular, by the notion of wavefront sets $WF(f) \subseteq \mathbb{R}^2 \times \mathbb{S}$, which-coarsely speaking - consist of singularities together with their direction. Basing in this sense the application of a deep neural network on microlocal considerations, in particular, also using a deep learning-based wavefront set detector (Andrade-Loarca et al., 2019) in the regularization term, the reconstruction performance significantly outperforms classical solvers such as sparse regularization with shearlets (see Figure 6, we also refer to (Andrade-Loarca et al., 2022) for details). Notice that this approach is of a hybrid type and takes the best out of both worlds in the sense of combining model- and artificial intelligence based approaches.



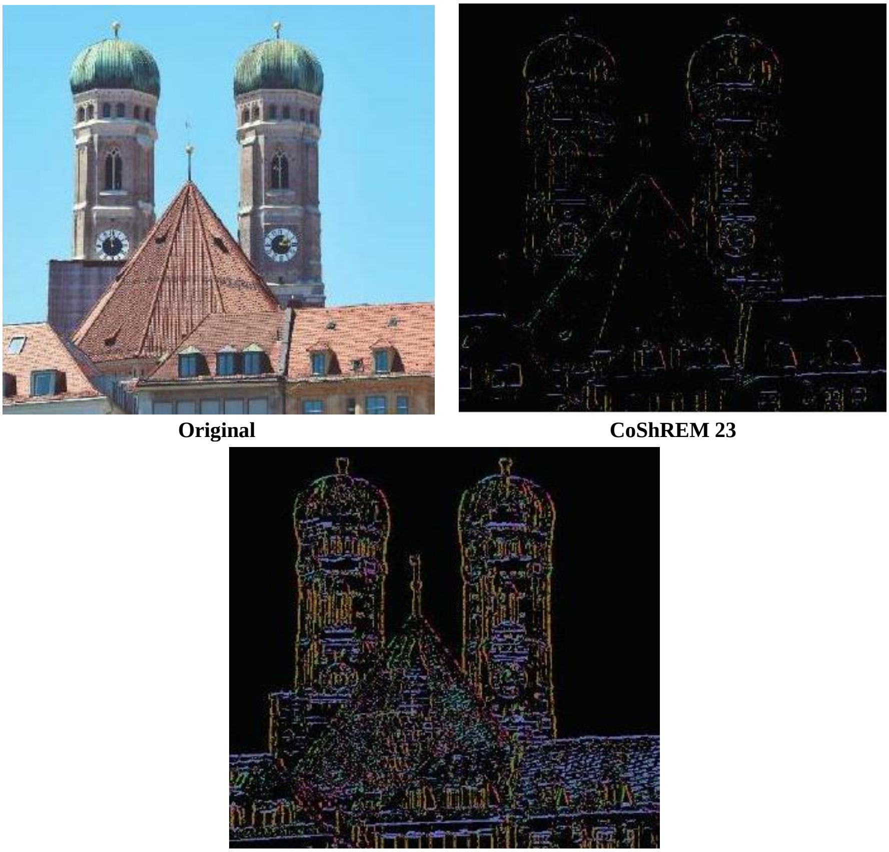
Original

Sparse Regularization with Shearlets

Deep Microlocal Reconstruction 3

Figure 6: CT Reconstruction from Radon Measurements with a Missing Angle of 40°

Finally, the deep learning-based wavefront set extraction itself is yet another evidence of the improvements on the state-of-the-art now possible by artificial intelligence. Figure 7 shows a classical result from [23, whereas Andrade-Loarca et al. (2019) uses the shearlet transform as a coarse edge detector, which is subsequently combined with a deep neural network.



Original **CoShREM 23**

DeNSE (Andrade-Loarca et al., 2019)

Figure 7: Wave front Set Detection by a Model-Based and a Hybrid Approach.

4.2 Partial Differential Equations

The second main range of mathematical problem settings, where methods from artificial intelligence are very successfully applied to, are partial differential equations. Although the benefit of such approaches was not initially clear, both theoretical and numerical results show their superiority in high-dimensional regimes.

The most common approach aims to approximate the solution of a partial differential equation by a deep neural network, which is trained according to this task by incorporating the partial differential equation into the loss function. More precisely, given a partial differential equation $\mathcal{L}(u) = f$, we train a neural network Φ such that

$$\mathcal{L}(\Phi) \approx f$$

Since 2017, research in this general direction has significantly accelerated. Some of the highlights are the Deep Ritz Method (Yu, 2018) and Physics Informed Neural Networks [22], or a very general approach for high-dimensional parabolic partial differential equations (Han, J., Jentzen, A., & E, W, 2018).

One should note that most theoretical results in this regime are of an expressivity type and also study the phenomenon whether and to which extent deep neural networks are able to beat the curse of dimensionality. In the sequel, we briefly discuss one such result as an example. In addition, notice that there already exist contributions - though very few-which analyze learning and generalization aspects. Let $\mathcal{L}(u_y, y) = f_y$ denote a parametric partial differential equation with y being a parameter from a highdimensional parameter space $\mathcal{Y} \subseteq \mathbb{R}^p$ and u_y the associated solution in a Hilbert space $\mathcal{H}$. After a high-fidelity discretization, let $b_y(u_y^h, v) = f_y(v)$ be the associated variational form with u_y^h now belonging to the associated high-dimensional space U^h, where we set $D := \dim(U^h)$. We moreover denote the coefficient vector of u_y^h with respect to a suitable basis of U^h by u_y^h. Of key importance in this area is the parametric map given by

$$\mathbb{R}^p \supseteq \mathcal{Y} \ni y \mapsto \mathrm{u}_y^\mathrm{h} \in \mathbb{R}^D \text{ such that } b_y(u_y^h, v) = f_y(v) \text{ for all } v$$

which in multi-query situations such as complex design problems needs to be solved several times. If p is very large, the curse of dimensionality could lead to an exponential computational cost.

We now aim to analyze whether the parametric map can be solved by a deep neural network, which would provide a very efficient and flexible method, hopefully also circumventing the curse of dimensionality in an automatic manner. From an expressivity viewpoint, one might ask whether, for each $\epsilon > 0$, there does exist a neural network Φ such that

$$\|\Phi(y) - \mathrm{u}_y^\mathrm{h}\| \leq \epsilon \text{ for all } y \in \mathcal{Y}.$$

The ability of this approach to tackle the curse of dimensionality can then be studied by analyzing how the complexity of Φ depends on p and D. A result of this type was proven in [18, the essence of which we now recall.

Theorem 4.1. There exists a neural network Φ which approximates the parametric map, i.e., which satisfies 4.2), and its dependence on $C(\Phi)$ on p and D can be (polynomially) controlled.

Analyzing the learning procedure and the generalization ability of the neural network in this setting is currently out of reach. Aiming to still determine whether a trained neural networks does not suffer from the curse of dimensionality as well, in (Geist et al., 2021) extensive numerical experiments were performed, which indicates that indeed the curse of dimensionality is also beaten in practice.

5. CONCLUSION: SEVEN MATHEMATICAL KEY PROBLEMS

Let us conclude with seven mathematical key problems of artificial intelligence as they were stated in (Berner et al., (2021). Artificial intelligence and the digital era have had a significant impact on the field of mathematics. Here are some of the ways in which they have influenced mathematics that including with the help of AI and digital tools, mathematicians can automate complex calculations and computations that were previously impossible or very time-consuming. This has allowed them to explore complex mathematical problems and find new solutions. AI has helped mathematicians to solve complex mathematical problems by analyzing large amounts of data and identifying patterns that humans may not have been able to see. This has led to new insights and discoveries in the field. The digital era has enabled mathematicians to explore new areas of research, such as computational mathematics and data science. This has led to the development of new mathematical models and techniques that are now widely used in many fields. The digital era has made it easier for mathematicians to collaborate and share ideas with each other. With the help of online platforms, mathematicians can now share their work with a global audience, get feedback from other experts, and collaborate on research projects in real-time. AI and digital tools have also revolutionized the way mathematics is taught. Today, students can access online resources, interactive simulations, and digital textbooks that make learning mathematics more engaging and accessible. Overall, the role of artificial intelligence and the digital era in mathematics has been transformative, providing new insights, tools, and techniques that have advanced the field in significant ways. Those constitute the main obstacles in Mathematical Foundations for Artificial Intelligence with its subfields expressivity, optimization, generalization, and explainability as well as in Artificial Intelligence for Mathematical Problems which focusses on the application to inverse problems and partial differential equations.

1) What is the role of depth?
2) Which aspects of a neural network architecture affect the performance of deep learning?
3) Why does stochastic gradient descent converge to good local minima despite the non-convexity of the problem?
4) Why do large neural networks not overfit?
5) Why do neural networks perform well in very high-dimensional environments?
6) Which features of data are learned by deep architectures?
7) Are neural networks capable of replacing highly specialized numerical algorithms in natural sciences?

REFERENCES

1. Adler, J. and Öktem, O. (2017). Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12), 124007. DOI 10.1088/1361-6420/aa9581
2. Andrade-Loarca, H., Kutyniok, G., Oktem, O. and Petersen, P.C. (2019). Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM Journal on Imaging Sciences*, 12(4), 1936-1966.
3. Andrade-Loarca, H., Kutyniok, G., Öktem, O. and Petersen, P. (2022). Deep microlocal reconstruction for limited-angle tomography. *Applied and Computational Harmonic Analysis*, 59, 155-197.

4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R. and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.
5. Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849-15854.
6. Berner, J., Grohs, P., Kutyniok, G. and Petersen, P. (2021). *The modern mathematics of deep learning*. Cambridge University Press
7. Bolcskei, H., Grohs, P., Kutyniok, G. and Petersen, P. (2019). Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1), 8-45.
8. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303-314.
9. Donoho, D.L. (2001). Sparse components of images and optimal atomic decompositions. *Constructive Approximation*, 17, 353-382.
10. Yu, B. (2018). The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1), 1-12.
11. Geist, M., Petersen, P., Raslan, M., Schneider, R. and Kutyniok, G. (2021). Numerical solution of the parametric diffusion equation by deep neural networks. *Journal of Scientific Computing*, 88(1), 22.
12. Han, J., Arnulf, J. and Weinan, E. (2018). Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34), 8505-8510.
13. Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359-366.
14. Jin, K.H., McCann, M.T., Froustey, E. and Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9), 4509-4522.
15. Kolek, S., Nguyen, D.A., Levie, R., Bruna, J. and Kutyniok, G. (2022, April). A rate-distortion framework for explaining black-box model decisions. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 91-115). Cham: Springer International Publishing.
16. Kutyniok, G. and Labate, D. (2012). *Multiscale analysis for multivariate data*. Springer Science & Business Media.
17. Kutyniok, G. and Lim, W.Q. (2011). Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163(11), 1564-1589.
18. Kutyniok, G., Petersen, P., Raslan, M. and Schneider, R. (2022). A theoretical analysis of deep neural networks and parametric PDEs. *Constructive Approximation*, 55(1), 73-125.
19. Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
20. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S. and Sohl-Dickstein, J. (2017). On the expressive power of deep neural networks. In *international conference on machine learning* (pp. 2847-2854). PMLR.

21. Raissi, M., Perdikaris, P. and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378, 686-707.
22. Reisenhofer, R., Kiefer, J. and King, E.J. (2016). Shearlet-based detection of flame fronts. *Experiments in Fluids*, 57, 1-14.
23. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
24. Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400-407.
25. Romano, Y., Elad, M. and Milanfar, P. (2017). The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4), 1804-1844.
26. Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536.
27. Smilkov, D., Thorat, N., Kim, B., Viégas, F. and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
28. Ulyanov, D., Vedaldi, A. and Victor, S. (2018). Lempitsky. Deep Image Prior. In *Computer Vision and Pattern Recognition (CVPR)* (Vol. 1).
29. Wäldchen, S., Macdonald, J., Hauch, S. and Kutyniok, G. (2021). The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70, 351-387.
30. Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks,* 94, 103-114.