Information Geometry? Exercises de Styles: Review

Ismail A Mageed[a], Kit Qichun Zhang[b]
[a,b]*Department of Computer Science, Faculty of Engineering and Informatics, University of Bradford, Bradford, BD7 1DP, UK*
[*]Corresponding author iammoham@bradford.ac.uk

***Abstract*— Information geometry, (IG), is an ever-growing area with a great scope of applications ranging from Probability & Statistics, Machine Learning (ML), Artificial Intelligence (AI), Signal Processing, Mathematical Programming, etc., There is always a great race to distil information from data to models. Since its inception, IG as a concept has been known under a variety of guises and been used in numerous contexts, establishing an almost rock-star status in both sciences and popular culture. The three most prominent "styles" which IG has been (re)told in and which have determined its popularity are Deep Learning, Statistical learning, and Machine Learning. Following the footsteps of the relentless hunt for the core of the concept that kindled this underlying development, connections with emergence of time combined with irreversibility, the elegant nature of probability and the generated information which add to its illusiveness as much as simulating its cross-contextual adoption and proliferation. In this review, we search and retrace through the five main perspectives from which IG has been regarded, emphasizing the motivations behind each application, their ramifications as well as the bridges that have been constructed to justify them. Consequently, this analysis of the foundations provides a beautiful panorama of several characteristic traits of the concept that underline its significance and exceptionality as an engine of conceptual progress**

Keywords- Information Geomerty, Geometric Deep Learning, Statistical Learning, Machine Learning.

I.INTRODUCTION

IG is the brainchild of the study on the invariant geometrical structure of a family of probability distributions. It is well agreed that invariance is a characterizing property of mathematical objects which remains unchanged upon an operation or symmetric transformation. Let's illustrate the following example to understand Invariance. In the study of circles, the ratio of perimeter and diameter (pi) remains unchanged with varying diameter values (scaling), which is a popular example of Invariance. In a similar fashion, comes the concept of Equivariance as a generalization to Invariance. That is applying transformation and applying function would produce the same result as computing function and then applying transformation. Now a legitimate question may arise. Why do we choose the info-geometric approach? IG efficiently enables us to study Invariance and Equivariance in a coordinate-free approach. It is a fact that more intuitive reasoning about the problems is significantly provided by IG. In addition to that, the obtained data portraits can be visualized and can be looked at as purely abstract objects.

What is IG? If the questioner's intention were to suddenly embarrass a combination of mathematicians, computer scientists or engineers a for a moment, that enquire would certainly achieve its aim—not because there is no answer, but because there are too many: a selection from the known sources, limited for example to *Deep Learning*, *Statistical Inference*, *Control Theory and Time Series Analysis*, *Machine Learning*, *Optimization and Neural Networks*, presents us already with several different, seemingly unrelated and not unambiguous definitions. Within science and beyond the mathematical realm, entropy is found in a profusion of other contexts: computation, mathematical physics, complexity theory, brain science, to name here but a few.

By analogy [1] to Information Theory (IT) (pioneered by Claude Shannon in his celebrated 1948 paper [2]) which considers primarily the communication of messages over noisy transmission channels, we may define Information Sciences (IS) as the fields that study "communication" between (noisy/imperfect) data and families of models (postulated as a priori knowledge). In short, information sciences seek methods to distil information from data to models. Therefore, information sciences encompass information theory but also include the fields of Probability and Statistics, Machine Learning (ML), Artificial Intelligence (AI), Mathematical Programming, just to name a few. Professor Shun-Ichi Amari, the founder of modern information geometry, defined information geometry in the preface of his latest textbook [3] as follows: "*Information geometry is a method of exploring the world of information by means of modern geometry.*"

Briefly, IG geometrically investigates information sciences. It is well agreed by the info-geometrists community that It is a mathematical endeavor to define and bound the term geometry itself as geometry is open-ended. Often, we start by studying the invariance of a problem (e.g., invariance of distance between probability distributions) and get as a result a novel geometric structure (e.g., a "statistical manifold"). Based on the scientific fact that a geometric structure is "pure" and thus may be applied to other application areas outside the scope of the original problem (e.g., use of the dualistic structure of statistical manifolds in mathematical programming [4]): the method of geometry [5] thus generates a pattern of abduction [6,7]. A narrower definition of IG can be formulated as the field that studies the geometry of decision making. This definition also includes model fitting (inference) which can be interpreted as a decision problem as illustrated in Figure 1; namely, deciding which model

parameter to choose from a family of parametric models. This framework was advocated by Abraham Wald [8,9,10] who considered all statistical problems as statistical decision problems.

Dissimilarities (also loosely called distances among others) play a crucial role not only for measuring the goodness-of-fit of data to model (say, likelihood in statistics, classifier loss functions in ML, objective functions in mathematical programming or operations research, etc.) but also for measuring the discrepancy (or deviance) between models.

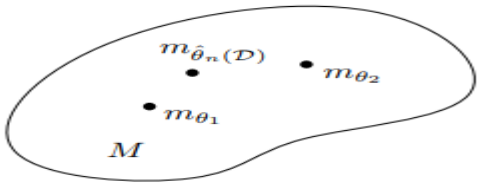

Figure 1. We can interpret the parameter inference $\hat{\theta}$ of a model from data D as a decision-making problem: building an algorithm which guarantees that the choice of parameter of a parametric family of models $M = \{m_\theta\}_{\theta\in\Theta}$ suits the "best" the data. IG provides a differential-geometric structure on manifold M which is so powerful for designing and studying statistical decision rules.

A breakthrough of Info-geometric Queueing Theory (IGQT) is devised by Mageed and Kouvatsos [11,12]. The strength of this novel approach is clearly demonstrated by the fact that it derives for the first time ever the exact stability and instability phases of the underlying $M/G/1$ queueing system. The beauty of our novel approach that revolutionizes Queueing Theory, is looking at a queue as a manifold, in which case, the parameter of curvature as well as being the connection parameter of the underlying stable $M/G/1$ queue manifold.

II. IG APPLICATIONS TO GEOMETRIC DEEP LEARNING

The deep Learning [13,14] technologies, for example, the convolutional neural networks [15], have achieved unprecedented, good results in some machine learning applications such as object detection [16-18], image classification [19], speech recognition [20], and machine translation [21]. Different from traditional neural networks, the deep neural networks, especially convolutional neural networks, make use of the basic statistical characteristics of data including local stationarity and multi-scale component structure to capture deeper local information and features. Although deep learning technology is very successful in processing traditional signals such as image, sound, video or text, the current research on deep learning still mainly focuses on the data mentioned above which are defined in the Euclidean domain, namely grid-like data. With the emergence of larger data scale and more powerful GPU computing ability, people begin to be more and more interested in processing data in non-Euclidean domain, such as graphs and manifolds. This type of data is ubiquitous in real life. It is of great significance to study deep learning techniques in non-Euclidean domains. This is called geometric deep learning.

The geometric deep learning (GDL) primarily studys graph and manifold data, where the graph is made of nodes and edges of the network structure data. For instance, in social network, each node represents a person's information and the edge represent the relationship between people. These edges are either directed or undirected based on the relationship of the connecting vertices. The Manifold data are usually used to describe geometric shapes, such as surface of objects returned by radar scanning. These geometric data are irregularly arranged and randomly distributed, which makes it difficult for people to find out the underlying pattern. Specifically, it is difficult to find the neighbour nodes of a certain point in the data, or the number of a node's neighbour is different in [22]. As a direct consequence, this makes it difficult to define convolution operations like those on images. On the other hand, data like images in the Euclidean domain can be regarded as a special graph data, with vertices arranged in a regular way. Another issue is that non-Euclidean data usually has extraordinarily large scale. For example, molecular graph can have hundreds of millions of nodes. For this case, it is unlikely to use the traditional deep learning technology to carry out analysis and prediction tasks. Therefore, deep learning is so important in the field of geometric data.

As early as in 2005, M. Gori et al. first proposed a graph neural network (GNN) to process graph data [23] such as directed graphs, undirected graphs, labelled graphs, and recurrent graphs. The work of [24] published by Scarselli et al.in 2009 brought back the graph neural network model to the public's horizon, defined a function that can map graph and any node to a dimensional Euclidean space, and proposed an algorithm to estimate the neural network model parameter with supervised learning. In the work of [25] proposed spectral convolutional neural networks on graphs. Work of [26] extended the spectral network by combining a graph estimation process. diffusion convolutional neural network (DCNN) was next proposed in [27] to learning diffusion-based representation from graph data for node classification. The work of [22], like image based convolutional network operating on the input locally connected region, proposed a general method to extract the locally connected region from the graph. In 2016, M. Defferrard et al. proposed ChebNet [28], and then a simplified version GCN (graph convolutional network) was proposed [29]. One year later, CayleyNet was proposed by Levie et al. [30]. All the above research results were based on the idea of convolutional network. Besides graph convolution model, there are other similar studies conducted in parallel, such as graph attention networks, graph generative networks, and graph auto-encoders.

At the same time, research of deep learning theory on manifold data are also carried out. There have been two traditional research methods on manifolds, one is to fill 3D shapes with many voxel grids (cube blocks), and each voxel can be processed by 3D CNN operation, called 3D volumetric CNN. The other is to take photos of 3D objects from multiple angles to increase the data source of the same object, which is called multi-view CNN. A framework of Geodesic CNN was proposed by [31], which is the promotion of convolutional

neural network (CNN) paradigm on non-Euclidean manifold. Later, in the work [32], the authors proposed Anisotropic CNN framework in the study of intrinsic dense correspondences between deformable shapes on the experiment of the results over [31]. This method generalized convolutional neural networks to the non-Euclidean domain by replacing the traditional convolution operations by projections on a set of oriented anisotropic diffusion kernels. Related work is [33], which proposed SyncSpecCNN network, where the kernel is parameterized in the spectral domain spanned by the Laplacian feature basis. D. Localized Spectral CNN (LSCNN) has been proposed in [34], in which the model structure is based on local frequency analysis with a windowed Fourie transform to manifold data. This method can be used for deformable shapes. A new framework called FMNet was introduced to learn the dense correspondence between deformable 3D shapes [35].

Many works have been conducted to find better approach to generalize convolution-like operations of convolution neural networks to the non-Euclidean domain. For example, the work [36] proposed a unified CNN framework MoNet and declared that the previous various CNN models can be unified within the framework.

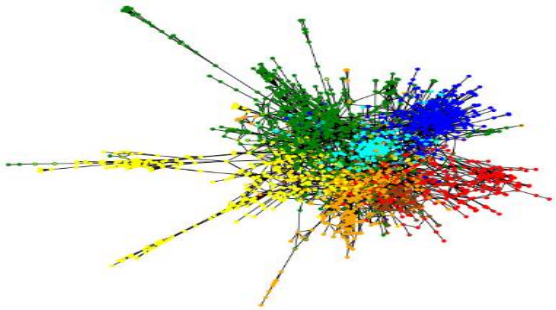

Figure 2. Example graph network of Cora dataset. Marker fill colour represents the predicted class, marker outline colour represents the ground truth class (c.f., [36]).

In addition, many researchers have tried to apply the above methods to a wide range of practical problems, from biochemistry [37] and skeleton-based human motion recognition task [38] to the recommender systems [39].

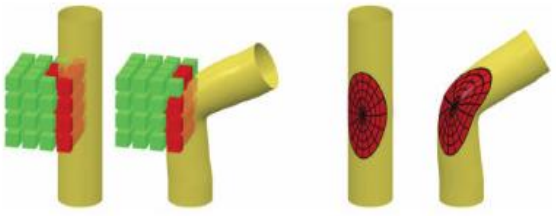

Figure 3. Illustration of the difference between extrinsic (left) and intrinsic (right) deep learning methods on geometric data. Intrinsic methods work on the manifold rather than its Euclidean realization and are isometry-invariant by construction.

The work of [41,42] put forward a most up-to-date survey on deep learning for graphs, which partially updated the work in [40]. Still, it did not cover those studies on graph generative and graph attention networks and methods on manifolds. Most recently, based on the above two surveys, Wu et al. put forward the current network structures on graphs, including spatio-temporal networks [43].

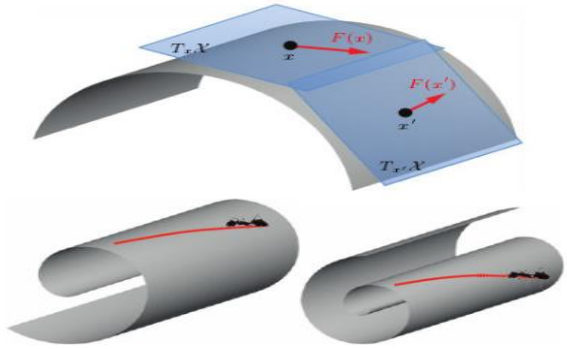

Figure 4. Top: tangent space and tangent vectors on a two-dimensional manifold (surface). Bottom: Examples of isometric deformations (c.f., [40]).

III. IG APPLICATIONS TO STATISTICAL LEARNING

A geometry [44] associated with U -divergence including ideas of U -models, U -loss functions of two versions has been presented. This geometric consideration leads to a special application to statistical pattern recognition.U -Boost algorithm associates with iteratively the U -divergence projection onto U -model evolving by one dimension according to one iteration. U -Boost algorithm of the version without the probability constraint, typically AdaBoost is shown to perform the novel property of statistical property beyond the notion of Fisher consistency. We discuss the property invariant over the coset of the Bayes rule with respect to the equivalence relation a natural requirement of pattern recognition. In the research area of statistical learning theory, the method of support vector machine has been developed parallel to the boosting method, [45,46]. Basically, the two paradigms have different objectives, in which an approach is recently proposed to connect the two methods based on the idea of soft margin [47]. As a future project we mention an embedding of U -loss function to a kernel space, which is closely related with a problem of characterization of U -divergence class. It needs a study of infinite-dimensional analysis on U -model as done by the use of the theory of Orlicz space in [48]. Recently there appear a vast of data sets of higher dimension along rapidly growing research activities in the genome sciences. For example, the micro-array technology enables to the simultaneous observations to gene expressions for a large group of genes. This information from the gene expression data should be related with difficult diseases, sensitivity for medication, and so that the problem is directly formulated as that of pattern recognition in which the feature vector is vector of gene expression, and, for example, class-label denotes the occurrence of considerable drug sensitivity. However, it is known that there is an unbalance relation between the number p and the sample size n, which leads to spurious discover for the relation of the particular gene expression and disease. The problem is addressed as $n \ll p$, which motivates a variety of approaches. It is interesting that we find a solution of the problem in the class of U -Boost algorithms.

IV. IG APPLICATIONS TO MACHINE LEARNING

The optimization algorithm [49] typically used in machine learning is stochastic gradient descent. Amari introduced the so-called natural gradient, mentioned in [50], which is meant to better capture the direction of steepest descent in a parameter space of probability distributions, compared to the standard gradient. In [51], it was shown that even when gradient descent is used, one can compute the expected change in output with respect to a change in the parameters by leveraging the Fisher information matrix. It has been shown that [49] momentum-based gradient descent algorithms can be extended to a Riemannian setting. Finally, we mention the drawbacks of these techniques, and why they are not widespread.

An artificial neural network (ANN) is a function with many parameters, mapping input vectors to output vectors. They have extensively applied to the areas of regression, computer vision, and speech processing [52, 53]. In principles, the primary building block of a neural network is the neuron, which is loosely related to a biological neuron. It consists of several components**:** Input, Weight, Bias, and Activation Function.

A large part of machine learning research [49] is studying how to adjust the parameters of an ANN to make it perform well at a certain task. These processes of learning the parameters are generally separated into two major categories: supervised learning and unsupervised learning. In supervised learning, we use a sample of desired input-output pairs, called a training set. An example of this is an ANN that learns to recognize stop signs in images, by using a dataset of images that have stop signs and a dataset of images that do not. In this case, we define a way to measure the performance of the ANN based on how close its outputs are to the expected outputs in the training set. Unsupervised learning covers techniques that do not use a training set. For example, data compression algorithms could use an ANN, by learning to transform an input x into a lower-dimensional representation such that little to no information about x is lost.
The back-propagation algorithm [54, 55] was conceived as a way to compute the gradient decent more efficiently, using only one forward pass of the input and one backward pass that determines how much contribution each parameter had in the error.

Another issue is that gradient descent tends toward a local minimum, where ideally, we want the global minimum. This is highly dependent on the how the network was initialized. We can modify the gradient descent algorithm to improve this: by adding a momentum term to the update [54] or using an adaptive learning rate [56]. These methods also tend to have faster convergence compared to unmodified gradient descent. In practice, convergence speed is also improved by estimating the cost of the entire training set by only using a fraction of the set per update. The sample of the training data used for an update is called a mini-batch. Many studies have shown that performing an update after every example (i.e., using a mini-batch size of 1) is effective [57]. When this technique is used, gradient descent is referred to as stochastic gradient descent.

Looking at the issue of overfitting, we are measuring the performance of the ANN based on how well it can classify data in the training set. Based on this criterion, the real goal is to create an ANN that can classify data well, even if it has never been seen before. Since ANNs have many parameters, they are prone to overfitting: that is, they have low error on training data, but high error on data outside of the training set. A study showed that ANNs with many neurons trained using back-propagation yielded similar results to ANNs with less neurons [58]. A common way of reducing overfitting is to use a validation set and early stopping [58]. At the beginning of training, a portion of the training data is set aside as a validation set, which is never used to update the parameters. Periodically during training, the ANNs performance is measured on the validation set. When the error on validation set starts to increase, it indicates that overfitting is occurring and that the training should stop. Overfitting is also reduced by artificially augmenting the training data with random distortions, and by a technique called Dropout, where inputs to individual neurons are randomly zeroed out during the training process [59].

VI. CONCLUSION AND PROSPECT

This survey reports the developments of GDL, statistical inference, machine learning and neural networks. There are still many challenges to extend and uncover new result to widen the horizon of the applicability of entropy in these disciplines**.** Several developments can be achieved to extend the applicability of entropy to unexplored disciplines. The recent development of IG applications reported in the survey will enable the reader to get knowledge in a bird's eye-view. By now, the reason for the title should be clear: as with Queneau's story in Exercises de style [70], entropy was retold and reinterpreted in manifold ways. These were all different stories, and yet the same story in that they shared a common core. Considering its conceptual developments, IG suggests us that perhaps the essential quality of a fruitful and thought-provoking concept is, at least in the eyes of some questioners, to call for ever more fundamental definitions, to stimulate IG reinterpretations, to allow for some ambiguity and plasticity to remain rather than dispel them entirely, and yet to give us one core to refer to and use as a base for our retellings.

The methodology of natural gradient was used by Frédéric and Nielsen in [60], for the study of blind separation of mixed signals, but the theory of Riemannian metrics in statistical settings already existed well before this (see [61] and [62]). Computing the inverse of the Fisher information matrix can be costly, but there are many ways of approximating it [64]. In a statistical framework, this algorithm was shown to be theoretically more efficient for estimating parameters than stochastic gradient descent [64].

When ANNs have many layers, they are difficult to train. By changing the parameters on neurons in the early layers, the possible inputs that are seen in later layers are also changed.

This effect is sometimes referred to as internal covariate shift. This effect is especially a problem when the neurons have activation functions that map R to a bounded interval (such a function is called a saturating non-linearity), as in the case of the sigmoid function [65, 66]. A method known as batch normalization was conceived by [65] to reduce the effect of internal covariate shift. This stabilizes the learning process, allowing for larger learning rates to be used. To stabilize the distributions of inputs to a layer, the real-valued inputs to the activation functions can be individually normalized over the training set. In a typical ANN, for a neuron in any given layer, we take the dot product of the input vector *x* from the previous layer with the weight vector of the neuron w and add a bias scalar b, before using it as input to an activation function f (see Figure 4).

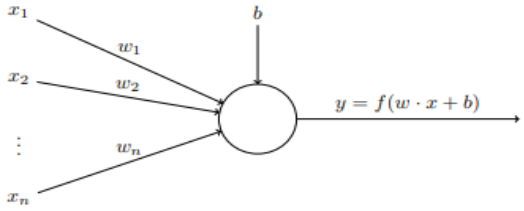

Figure 4. A single neuron

Information geometry can be applied to understand how much a gradient update will change the output distribution of the ANN. In the case of batch normalization, it becomes even more obvious that distances in the parameter space do not correspond with the magnitude of change in the output, since re-scaling the weights does not change the output at all. For an ANN with multiple neurons and layers, the impact of changing the parameters can be analyzed by using a block-diagonal approximation of the Fisher information matrix, where each block corresponds to the parameters of a single neuron. A detailed overview of this analysis can be found in [67].

A large drawback of gradient descent is that when optimizing for non-convex loss functions, we can get stuck in a poor local minimum. Additionally, there are versions of gradient descent that do not use a normalized gradient in the update rule. We expect improvements to slow near local minima, as the algorithm tends toward a solution, but this also results in slow optimization around other critical points, including saddle points and local maxima, since the gradient is close to 0. The associated problems with momentum -based optimization are partly caused by the fact that the direction of the gradient update can change instantaneously. To deal with these challenges, it is useful to incorporate some form of momentum into the gradient update rule, to avoid the optimization from slowing down when the recent changes were large. Intuitively, it is similar to how a ball gains speed rolling down a hill and is able to climb back up over small bumps

Cho and Lee [68] showed that neurons with n weight parameters and batch normalization can be identified with the set of orthonormal vectors in $\mathbb{R}^n$ (the Stiefel manifold $v(1,n)$), or alternatively, they can (almost) be identified with the set of 1-dimensional subspaces of $\mathbb{R}^n$ (the Grassmannian manifold $g(1,n)$). There exists promising work, showing that the Fisher information matrix can be approximated using a low-rank, block-diagonal matrix [69], but for Fisher information and the natural gradient to be feasible in machine learning, further work must be done towards approximating these efficiently.

REFERENCES

[1] F. Nielsen, "An elementary introduction to information geometry", 2020, Entropy 22.10: 1100.

[2] C.E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J,27, 1948,pp. 623–656.

[3] S. Amari, "Information geometry and its applications", Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016.

[4] S.Kakihara, et al, " Information geometry and interior-point algorithms in semidefinite programs and symmetric Cone Programs",J.Optim.TheoryAppl.,157,2013,pp.749–780, doi:10.1007/s10957-012-0180-9.

[5] S. Amari and H. Nagaoka, "Methods of information geometry", American Mathematical Society: Providence, RL, USA,2007.

[6] C.S.Peirce, "Chance, Love, and Logic: Philosophical Essays", University of Nebraska Press: Lincoln, NE, USA. 1998.

[7] G.Schurz, "Patterns of abduction", Synthese, 164,2008, pp. 201–234.

[8] A. Wald, "Statistical decision functions", Ann. Math. Stat., 1949, pp. 165–205.

[9] A.Wald, " Statistical decision functions",Wiley: Chichester, U.K., 1950.

[10] A.G.Dabak, " A Geometry for Detection Theory",1993,Ph.D. Thesis, Rice University, Houston, USA.

[11] I. A. Mageed, and D.D.Kouvatsos, "Information Geometric Structure of Stable M/G/1 Queue Manifold and its Matrix Exponential", Proceedings of the 35th UK Performance Engineering Workshop, School of Computing, University of Leeds, Edited by Karim Djemame,2019, p.123-135. Available online at: https://sites.google.com/view/ukpew2019/home

[12] I.A.Mageed, and D.D. Kouvatsos, "The impact of information geometry on the analysis of the stable M/G/1 queue manifold", Major extension of paper [11], In Proceedings of the 10th International Conference on Operations Research and Enterprise Systems - Volume 1: ICORES, ISBN 978-989-758-485-5,2021, pp.153-160. DOI: 10.5220/0010206801530160.

[13] C. Wenming, et al. "A comprehensive survey on geometric deep learning", IEEE Access 8: 2020, pp.35929-35949.

[14] Y. LeCun, et al, ''Deep learning'', Nature, vol. 521, no. 7553,2015, pp. 436–444.

[15] Y. LeCun and Y. Bengio, ''Convolutional networks for images, speech, and time series'', in The Handbook of Brain Theory and Neural Networks. Cambridge, MA, USA: MIT Press,1998, pp. 255–258.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation'', in Proc.IEEE Conf. Comput. Vis. Pattern Recognit, 2014,pp. 580–587.

[17] R. Girshick, ''Fast R-CNN'', in Proc. IEEE Int. Conf. Comput. Vis. (ICCV),2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks'', IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, 2017,pp. 1137–1149.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks'', Commun. ACM, vol. 60, no. 6, 2017, pp. 84–90.

[20] G. Hinton, et al, ''Deep neural networks for acoustic modeling in speech recognition'', IEEE Signal Process. Mag., vol. 29, no. 6, 2012, pp. 82–97.

[21] I. Sutskever, O. Vinyals, and Q. V. Le, ''Sequence to sequence learning with neural networks'',in Proc. Adv. Neural Inf. Process.Syst.,2014, pp. 3104–3112.

[22] M. Niepert, M. Ahmed, and K. Kutzkov, ''Learning convolutional neural networks for graphs'', in Proc. Int. Conf. Mach.Learn.,2016.

[23] M. Gori, G. Monfardini, and F. Scarselli, ‘‘A new model for learning in graph domains’’ , in Proc. IEEE Int. Joint Conf. Neural Netw., vol. 2, 2005, pp. 729–734.
[24] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, ‘‘The graph neural network model’’, IEEE Trans. Neural Netw., vol. 20, no. 1,2009, pp. 61–80.
[25] J. Bruna, et al, ‘‘Spectral networks and locally connected networks on graphs’’ 2013,arXiv:1312.6203. [Online]. Available: http://arxiv.org/abs/1312.6203
[26] M. Henaff, J. Bruna, and Y. LeCun, ‘‘Deep convolutional networks on graph-structured data’’ 2015, arXiv:1506.05163. [Online]. Available: http://arxiv.org/abs/1506.05163
[27] J. Atwood and D. Towsley, ‘‘Diffusion-convolutional neural networks’’, in Proc. Adv. Neural Inf. Process. Syst., 2016,pp. 1993–2001.
[28] M. Defferrard, X. Bresson, and P. Vandergheynst, ‘‘Convolutional neural networks on graphs with fast localized spectral filtering’’, in Proc. Adv.Neural Inf. Process. Syst., 2016,pp. 3844–3852.
[29] T. N. Kipf and M. Welling, ‘‘Semi-supervised classification with graph convolutional networks’’ 2016, arXiv:1609.02907. [Online]. Available:http://arxiv.org/abs/1609.02907
[30] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, ‘‘CayleyNets: Graph convolutional neural networks with complex rational spectral filters’’, IEEE Trans. Signal Process., vol. 67, no. 1, 2019,pp. 97–109.
[31] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst, ‘‘Geodesic convolutional neural networks on Riemannian manifolds’’ , in Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW), 2015, pp. 37–45.
[32] D. Boscaini,et al, ‘‘Learning shape correspondence with anisotropic convolutional neural networks’’, in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 3189–3197.
[33] L. Yi, H. Su, X. Guo, and L. Guibas, ‘‘SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation’’, in Proc. IEEE Conf. Comput.Vis. Pattern Recognit. (CVPR),2017, pp. 2282–2290.
[34] D. Boscaini, et al, ‘‘Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks’’, Comput. Graph. Forum, vol. 34, no. 5,2015, pp. 13–23.
[35] O. Litany, et al, ‘‘Deep functional maps: Structured prediction for dense shape correspondence’’, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 5659–5667.
[36] F. Monti, et al, ‘‘Geometric deep learning on graphs and manifolds using mixture model CNNs’’, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),2017, pp. 5115–5124.
[37] D. K. Duvenaud, et al, ‘‘Convolutional networks on graphs for learning molecular fingerprints’’, in Proc. Adv. Neural Inf. Process.Syst.,2015, pp. 2224–2232.
[38] Z. Huang, et al, ‘‘Deep learning on lie groups for skeleton-based action recognition’’ , in Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), 2017, pp. 6099–6108.
[39] F. Monti, et al, ‘‘Geometric matrix completion with recurrent multi-graph neural networks’’, in Proc. Adv. Neural Inf. Process. Syst.,2017, pp. 3697–3707.
[40] M. M. Bronstein, et al, ‘‘Geometric deep learning: Going beyond Euclidean data’’, IEEE Signal Process. Mag., vol. 34, no. 4,2017, pp. 18–42.
[41] Z. Zhang, P. Cui, and W. Zhu, ‘‘Deep learning on graphs: A survey’’ 2018, arXiv:1812.04202. [Online]. Available: http://arxiv.org/abs/1812.04202
[42] J. Zhou, et al., ‘‘Graph neural networks: A review of methods and applications’’ 2018,arXiv:1812.08434. [Online]. Available: http://arxiv.org/abs/1812.08434
[43] Z. Wu, et al. , ‘‘A comprehensive survey on graph neural networks,’’ 2019, arXiv:1901.00596.[Online]. Available: http://arxiv.org/abs/1901.00596
[44] S.Eguchi, “Information geometry and statistical pattern recognition”, Sugaku Expositions 19.2; 2006,pp. 197-216.
[45] G. James, et al., “Statistical learning, An introduction to statistical learning”, Springer, New York, NY, 2013, pp.15-57.
[46] M.Pastore, et al., “Statistical learning theory of structured data”,Physical Review E 102.3: 032119.2020.
[47] S.Chen, et al., “A strong machine learning classifier and decision stumps based hybrid AdaBoost classification algorithm for cognitive radios”, Sensors 19.23: 5077.2019.
[48] G.Pistone and C. Sempi, “An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one”, The annals of statistics: 1995, pp.1543-1561.
[49] J. d’Eon, “Applications of Information Geometry to Machine Learning”, 2019, MSC Thesis, University of Waterloo.
[50] S. Amari, “Natural Gradient Works Efficiently in Learnin”, Neural Comput. 10.2, 1998,pp. 251–276. issn: 0899-7667. doi: 10.1162/089976698300017746. url: http://dx.doi.org/10. 1162/089976698300017746
[51] J. Lei Ba et al., “Layer Normalization”, 2016. eprint: arXiv: 1607.06450.
[52] D.S. Bulgarevich, et al. “Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures”, Scientific reports 8.1:; 2018,pp.1-8.
[53] L.G. Maryellen , “Deep learning”, High-Dimensional Fuzzy Clustering.2021.
[54] J.L.Suárez, et al., “A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges”, Neurocomputing 425; 2021,pp.300-322.
[55] M. Alexander, et al. “A theoretical framework for target propagation” ,Advances in Neural Information Processing Systems 33; 2020, pp. 20024-20036.
[56] X. Yanchun, et al. “Competitive search algorithm: a new method for stochastic optimization”, Applied Intelligence;2022, pp. 1-24.
[57] S. Xiaoyu, et al. “Large-scale and scalable latent factor analysis via distributed alternative stochastic gradient descent for recommender systems", IEEE Transactions on Big Data . 2022.
[58] B.Lengerich, et al. “On dropout, overfitting, and interaction effects in deep neural networks”,2020.
[59] Z.Yicheng, et al. “Breeds classification with deep convolutional neural network.”, Proceedings of the 2020 12th International Conference on Machine Learning and Computing.2020.
[60] B. Frédéric and F. Nielsen, “Differential geometrical theory of statistics”, MDPI AG, Basel, Switzerland.2017.
[61] S.Sergey and J.Mikeš, “A brief review of publications on the differential geometry of statistical manifolds”,COJ Tech. Sci. Res 2; 2020.
[62] C.R.Rao, “Information and the Accuracy Attainable in the Estimation of Statistical Parameters”, A Tribute to the Legend of Professor CR Rao. Springer, Singapore; 2021, pp.1-13.
[63] L. Huang, et al. “Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks”, Thirty-Second AAAI Conference on Artificial Intelligence.2018.
[64] P.Zhao, et al. “Towards query-efficient black-box adversary with zeroth-order natural gradient descent”, Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04; 2020.
[65] Y.Greg, et al. “A mean field theory of batch normalization”, 2019, arXiv preprint arXiv:1902.08129.
[66] I.Pavel, et al. “Dangers of Bayesian model averaging under covariate shift”, Advances in Neural Information Processing Systems 34; 2021.
[67] X. Ruibin, et al. “On layer normalization in the transformer architecture”,International Conference on Machine Learning. PMLR; 2020.
[68] M. Cho and J.Lee, “Riemannian approach to batch normalization”, 2017. eprint: arXiv: 1709.09603.
[69] T.Zedong, et al. “AsymptoticNG: A regularized natural gradient optimization algorithm with look-ahead strategy”, 2020, arXiv preprint arXiv:2012.13077.
[70] B. Sırma, and P. Yıldız., “A Spatial Translation On he Text Of Raymond Queneau’s “Exercises in Style”, Iconarp Internatioqal Journal of Architecture and Planning 8.1; 2020, pp. 211-240.