Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221272

Gaps in Generalization Theory of Neural Networks

Eshan PANDEY[a,1] and Santosh KUMAR[a]
[a] *ABES Engineering College, Ghaziabad, Uttar Pradesh-201009, India*

Abstract. In the last decade, neural networks have become exceptionally powerful. But there is a huge difference in practical achievement and theoretical understanding. One such underdeveloped theoretical concept is generalization of the deep nets. Many studies have been conducted to develop a better understanding of generalization in deep neural networks. In this paper a survey of studies has been done for exploring the generalization concept. The contradiction and overlapping with the statistical learning theory, conventional wisdom and unconventional results has also been highlighted. A sound understanding of generalization will enable researchers to model more cost effective and powerful networks. Demystifying generalization will also result in more informed architecture design decisions

Keywords. Deep Neural Networks, Generalization, Loss Landscape, Regularization, Overparameterized Networks.

1. Introduction

Deep neural networks try to approximate a function *F** that can appropriately represent the data samples. The modern neural network algorithms can easily match human level precision in tasks such as detection and classification. The objective of any ML algorithm is to make a prediction on unseen data. While training data is available for the model to learn and approximate the function *F**, test data is not seen by the algorithm/model. For this reason, there is often a gap between the training error *(TE)* and testing error *(TE')*. An absolute difference between training error *(TE)* and testing error *(TE')* is called generalization error *(GE)*. $GE = | TE - TE' |$

There are many motivations to study generalization. Following are some mentioned reasons. The fundamental reasons for generalization of deep neural networks are not known [12]. Given a network and network parameters can one predict the maximum and minimum generalization gap? Statistical learning theory, finite sample expressivity, universal approximation theorem [12, 13] provides some loose bounds but fails to capture practicality in the context of deep learning. Developing theoretical data dependent generalization bounds, independent from all other network parameters is also a challenge. VC Dimensions, PAC Learning, Rademacher Complexity provide an initial generalization framework for the traditional ML algorithms.

[1] Corresponding Author.

2. Factors affecting the Generalization

There are multiple components that may influence the generalization capability of deep neural networks. Data sets, distribution of samples in data sets, the nature of loss function, optimizers, model architecture, regularization, stability, complexity of network, capacity control, robustness, activation, etc. The degree to which each factor contributes to generalization is highly debatable, and less understood. In [12], authors challenge the fundamental understanding about the learnability of overparameterized networks. They show that modern architectures with SGD can fit partially corrupted labels, fully corrupted labels, partially corrupted images and fully random pixels. Most theoretical concepts derived from statistical learning theory do not seem to justify the mysterious behavior of deep neural networks. [17] suggests that the disconnect between the mysterious behavior of deep nets and statistical learning theory is just a case of misinterpretation. However, most theoretical bounds are too loose to be applied in real life scenarios. [2,9,14] present some advances real life applications.

2.1. Learnability, Data and Architectures

Most modern neural networks are overparameterized, where the total count of training data sample points is not dependent on the number of parameters. VC dimension learning theory cannot be used to study such settings because the VC dimensions grow with the number of parameters. In [12], authors further show that the overparameterized networks can memorize completely random noise, while being able to generalize to true data. They suggest that randomizing data is a data transformation problem and does not affect the learning algorithm. In such settings the notion of stability derived from PAC bayes learning has limited context. Thus, a theorem independent from such constraints/assumptions about network design, parameters and data distributions is more desirable.

Universal approximation theorem defines upper bounds in the approximation capability of a 2-layered network. Any continuous and bounded function can be modeled using a 2-layered network having nonlinear activation. In [13], authors argued with experiments that size of the network may affect the capacity of neural networks, but size is not the primary form of capacity control. Further suggesting that there might be a different form of capacity control altogether. In [15], authors explore an interesting question that “are all layers created equal?” They conclude each layer can be categorized into two categories, i.e., robust and critical. Robust layer hardly changes throughout the entire process of training. They also suggest that robust layers can be altogether eliminated implying there is minimal to no contribution in learning and generalization. They conclude with experimental evidence that deep nets automatically adjust their capacity; when big networks are trained on easy tasks, only a few layers play a critical role. Their work is a strong indication that mere parameter counting is not sufficient to develop generalization theory.

2.2. Geometric Nature of Loss function and Optimizers

Training a neural loss function is NP hard. The shape of the loss function is often non-convex and high dimensional. Local minima are very likely to have error values near to global minima [3]. For large sized networks most, local minima are equivalent and perform equally good, it is hard to find bad local minima for small size networks and this

difficulty increases with network size and looking for global minima is not recommended and may lead to overfitting [4]. In contrast, [5] introduces Entropy-SGD that favors wide valleys situated far deep in the loss-landscape having much lesser empirical loss value compared to the local minima found by SGD, while comparing favorably with most other algorithms in terms of generalization

Studies suggest that a small training batch with SGD optimizer generates flat minimizers which generalize well but a large training batch with SGD generate sharp minima that fails to generalize [6]. In [7] authors experimentally verify that large batches with SGD converge to sharp minima that lead to a bad generalization. However, it is possible to achieve better generalization with large batch sizes. [8] suggests several techniques to improve the generalization gap while training with large batches. However, contrary to the former studies,[9] argues that the notion of flatness and sharpness cannot directly explain generalization. [9] agrees that the algorithms that generalize tend to be flatter at minima. However, flat minima generalizes better than sharp minima cannot be universally true. And so flat minima are not the fundamental reason for generalization.

2.3. Regularization leads to generalization

Any modification with an intent to bring down the generalization error, but training error, is regularization. One or more regularization techniques can be combined to achieve better results. Data Augmentation is a regularization technique that requires training on more data. Creating some fake data may help. Rotating, scaling and translating has also been observed to be useful. Dropout is a widely used and cost-efficient alternative to bagging. However, dropout does not yield significant performance gains when less training data is available. Early Stopping is arguably the most simple and effective regularization technique.

While popular regularization techniques might help in generalization, they are not the fundamental reason for generalization. [12] suggests that greater generalization can be achieved by modifying the network architecture. By increasing the number of hidden units, generalization capability of the network is improved even when the training error does not decrease [13]. This might indicate how implicit regularization can be helpful and might be better than explicit regularization techniques.

3. Why more study on generalization is needed?

Question 1: How can overparameterized networks generalize? One major dispute is the learnability of over parameterised networks as shown by [12] Over parameterised networks, as suggested by conventional wisdom and statistical learning theory should fail to generalize. Observations of Zhang et al., [12] challenges the conventional wisdom about deep learning and principles of statistical learning theory. [17] addresses this problem and suggests that the observations of [12] do not contradict the principles of statistical learning theory; the conflict arises due to misinterpretation. Assume $p :=$ "small complexity" and $q :=$ "small generalization"; statistical learning theory suggests $p \Rightarrow q$. However, this does not mean $q \Rightarrow p$. Thus, it is possible to have low generalization error despite large complexity of hypothesis, instability and non-robustness of learning algorithm or existence of sharp minima. [17] further suggests that small capacity, low complexity, flat minimum, stability and robustness and is not fundamental to

generalization and this phenomenon is not specific to deep nets, overparameterized linear models can generalize too.

Question 2: How, a network that can fit random noise, generalize to true labels? [12] suggests that randomization of data is only data transformation and does not affect learning algorithms. This further suggests the need to develop theoretical bounds independent from data distribution and parameter counting. In [21], authors numerically evaluated a generalization error bound from the PAC-Bayes framework showing that it can forecast the difference in generalization capability of the networks trained on true labels vs the networks trained on random labels.

Question 3: What architecture generalizes better? Zhang [12] highlighted the question why some networks can generalize better than others. Modification in network architecture can achieve better generalization than explicit regularizers. Residual Networks [22] achieves better generalization than simple feed forward networks. Deeper network achieves better generalization when compared to shallow networks.

Question 4: loss landscape? Loss landscape is very crucial in the study of generalization, section 2.2 discusses this in detail. The notion of flatness and sharpness is difficult to ignore while studying generalization. However, flatness and sharpness, as is, cannot be universally used to make an inference about generalization [9].

Question 5: Does batch size affect generalization? Yes. Small batch with SGD produces flat minimizes while large batch size with SGD produces sharp minima that fails to generalize [6,7]. However, it is possible to achieve good generalization with a large batch size [8,10]. Generally, algorithms that generalize tend to be flatter at minima, however, this is not universally true [9]. So, batch size is not fundamental to generalization. However, batch normalization can be used as a regularization technique as and when needed.

Question 6: What can affect generalization? Small capacity, low complexity, flat minimum, stability and robustness is not fundamental to generalization, although either can be sufficient. [17].

Table 1. Critical points in the literature. Studies performed, novel approaches, observations and critical analysis of the literature.

Objective	Observation	Dispute & settlements (Critical Analysis)
[3] Introduce a novel second order optimization technique	Global minimum will most likely have error values close to the local minima.	Local minima are good.
[4] Compare various minima	It's hard to find a bad local minimum. For large sized networks most, local minima are equivalent and perform equally good.	looking for global minima may lead to overfitting.
[5] Introduce Entropy-SGD	Entropy-SGD can find wide and deeper valleys with lower empirical loss (compared to local minima achieved by SGD)	Contradicts the observation of [4], i.e., the existence of multiple local minima with similar loss.
Relate batch size with generalization gap [6]	A small batch size with SGD optimizer produces flat minimizers which generalize well.	Large batch size with SGD produces sharp minima which fails to generalize well
Experimental Validated of [6] by [7]	Large batches trained with SGD converge to sharp minima, leading to high generalization error.	Support the claims of [6]

[8] Introduced Ghost batch normalization	Generalization gap arises from small iterations of updates and not from the batch size.	Suggests training methods to achieve good generalization with large batch size. Counter the claims of [6, 7]
Generalization with large batch size [10]	Large batch cause optimization difficulty but when addressed, trained networks can achieve good generalization	Counter the claims of [6, 7]
[11] Propose Big Batch SGD	By choosing larger batches with less noise, it is possible to maintain descent directions on each iteration and uphold fast convergence.	Counter the claims of [6, 7]
Generalization and geometry [9]	Algorithms that generalize tend to be flatter at minima. Concept of flatness and sharpness isn't enough to describe the generalization capability of deep nets.	Flat minima generalities better than sharp minima. However, this cannot be universally true.
Loss landscape of networks [16]	Skip connection promotes flat minimizers. It prevents the transition into chaotic behavior.	Resnet like structures outperform simple feed forward networks.
Capacity Control for neural networks [13]	Size of the network is not the main reason for its expressive power.	Highlights the limitations of VC Dimension.
[12] Explores what leads to generalization	Deep nets can fit completely random noise (huge capacity). Despite huge model capacity, deep nets can generalize. Small generalization improvement due to regularizations	Reasons for generalizations are not known. Better results can be achieved by modifying the network design. Regularization is not the fundamental reason for generalization.
[15] Study the contribution of individual layer in overall network	Critical layers contribute to generalizations.	Not all layers are equal, some layers can be all together removed. Deep nets automatically adjust their capacity.
[17] Study the disputes and questions raised by [12]	Small capacity, low complexity, flat minimum, stability and robustness is not necessary for generalizations. But either one is sufficient.	Observations of [12] are in accordance with statistical learning and it's a matter of interpretations
[18] Study generalization under various circumstances	For overparameterized networks, complexity measures based on total number of parameters cannot explain generalization. Sharpness is not sufficient to explain generalization.	Partly address the dispute raised by [12] Sharpness combined with PAC bayes analysis and weight norms can be used to obtain complexity measures.
[1] Study over parameterized networks	Existing complexity measures are proportional to the number of hidden units	Complexity measures does not sufficiently explain generalization in over parameterized networks
[20] Study over parameterized networks	Over parameterization helps in optimization. Weight-decay helps in generalization.	Over-parameterization allows to find global optima, when combined with weight decay, the solution also generalizes well.

4. Conclusion

Some modern deep neural network architectures may have up to a billion parameters and are massively resource and cost hungry. Under such scenarios understanding and studying subjects like generalization becomes essential. The paper has attempted to

summarize the generalization theory of deep neural networks, it also highlights the gaps in the theory and present the critical analysis of the same. It is observed that while many studies explore the reasons for generalization, the elementary reason for generalization in deep neural networks remain undiscovered. The contradiction in the literature is also discussed. Few unanswered questions have been identified that can act as potential future research topics. Answering generalization will enable researchers to model the deep net architectures with precision and make more informed decisions.

References

[1] Neyshabur B, Li Z, Bhojanapalli S, LeCun Y, Srebro N. Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks. InInternational Conference on Learning Representations (ICLR) 2019 Jan.

[2] Nainvarapu, R., Tummala, R. B., & Singh, M. K. (2022). A Slant Transform and Diagonal Laplacian Based Fusion Algorithm for Visual Sensor Network Applications. In High Performance Computing and Networking (pp. 181-191). Springer, Singapore.

[3] Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. Advances in neural information processing systems. 2014;27.

[4] Choromanska A, Henaff M, Mathieu M, Arous GB, LeCun Y. The loss surfaces of multilayer networks. InArtificial intelligence and statistics 2015 Feb 21 (pp. 192-204). PMLR.

[5] Chaudhari P, Choromanska A, Soatto S, LeCun Y, Baldassi C, Borgs C, Chayes J, Sagun L, Zecchina R. Entropy-sgd: Biasing gradient descent into wide valleys. Journal of Statistical Mechanics: Theory and Experiment. 2019 Dec 20;2019(12):124018.

[6] Hochreiter S, Schmidhuber J. Flat minima. Neural computation. 1997 Jan 1;9(1):1-42.

[7] Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PT. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836. 2016 Sep 15.

[8] Hoffer E, Hubara I, Soudry D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. Advances in neural information processing systems. 2017;30.

[9] Dinh L, Pascanu R, Bengio S, Bengio Y. Sharp minima can generalize for deep nets. InInternational Conference on Machine Learning 2017 Jul 17 (pp. 1019-1028). PMLR.

[10] Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677. 2017.

[11] De S, Yadav A, Jacobs D, Goldstein T. Automated inference with adaptive batches. InArtificial Intelligence and Statistics 2017 Apr 10 (pp. 1504-1513). PMLR.

[12] Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM. 2021 Feb 22;64(3):107-15.

[13] Neyshabur B, Tomioka R, Srebro N. In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv preprint arXiv:1412.6614. 2014 Dec 20.

[14] Padma, U., Jagadish, S., & Singh, M. K. (2021). Recognition of plant's leaf infection by image processing approach. Materials Today: Proceedings.

[15] Zhang C, Bengio S, Singer Y. Are all layers created equal? arXiv preprint arXiv:1902.01996 (2019).

[16] Li H, Xu Z, Taylor G, Studer C, Goldstein T. Visualizing the loss landscape of neural nets. Advances in neural information processing systems. 2018;31.

[17] Kawaguchi K, Kaelbling LP, Bengio Y. Generalization in deep learning. arXiv preprint arXiv:1710.05468. 2017.

[18] Neyshabur B, Bhojanapalli S, McAllester D, Srebro N. Exploring generalization in deep learning. Advances in neural information processing systems. 2017;30.

[19] Satya, P. M., Jagadish, S., Satyanarayana, V., & Singh, M. K. (2021, October). Stripe Noise Removal from Remote Sensing Images. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 233-236). IEEE.

[20] Du S, Lee J. On the power of over-parametrization in neural networks with quadratic activation. In International conference on machine learning 2018 Jul 3 (pp. 1329-1338). PMLR.

[21] Dziugaite GK, Roy DM. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008. 2017.

[22] Frei S, Cao Y, Gu Q. Algorithm-dependent generalization bounds for overparameterized deep residual networks. Advances in neural information processing systems. 2019;32.