Understanding Generalization and Memorization in Deep Learning Models

Exploring Generalization in Deep Learning:
To delve deeper into the complexities of generalization in deep learning, a practical approach involves revisiting foundational works in the field. A significant step is to reproduce the results of influential studies that remain highly relevant. One such cornerstone is a study that necessitates rethinking how deep learning generalizes, serving as a pivotal starting point for exploring this intricate problem. Alongside this, there's an opportunity to reproduce visualizations from another well-regarded study, known for its insightful examination of memorization in deep networks.

Relevant Readings with Key Insights

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding Deep Learning Requires Rethinking Generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding Deep Learning (Still) Requires Rethinking Generalization. *Communications of the ACM, 64*(7), 107–115. doi:10.1145/3446776.

- **Challenge to Conventional Wisdom**: The paper challenges the traditional understanding of generalization in deep neural networks, demonstrating that large networks can have a small generalization error despite their capacity to fit even random labels or noise.
- **Central Finding - Network Capacity for Random Labels**: Deep neural networks can easily fit completely random labels, achieving zero training error. This occurs regardless of the model size, hyperparameters, or the optimizer used. This finding applies to various standard architectures like CIFAR10 and ImageNet.
- **Handling Noise**: Neural networks can fit data with zero training error even when true images are replaced with random noise. As noise levels increase, the generalization error steadily deteriorates, indicating the network's capacity to capture signal amidst noise.
- **Role of Explicit Regularization**: The study shows that explicit regularization methods like weight decay, dropout, and data augmentation, while potentially improving generalization performance, are neither necessary nor sufficient for controlling generalization error in neural networks.
- **Implicit Regularization and SGD**: The paper explores how stochastic gradient descent (SGD) acts as an implicit regularizer in neural networks, a perspective that contrasts with traditional views on explicit regularization. This exploration extends to the performance of linear models under SGD without explicit regularization.
- **Effective Capacity of Neural Networks**: The paper introduces a framework to understand the effective capacity of machine learning

models, highlighting that large neural networks are capable of memorizing training data, which raises questions about the nature of generalization in these models.

Arora, S., Cohen, N., Hu, W., & Luo, Y. (2019). Implicit Regularization in Deep Matrix Factorization. *In Advances in Neural Information Processing Systems (Vol. 32)*.

- **Depth Enhances Low-Rank Bias**: Adding depth to matrix factorizations enhances the implicit tendency towards low-rank solutions, often leading to more accurate recovery.
- **Questioning Norm-Based Regularization**: The study questions the adequacy of simple mathematical norms, like nuclear norm, in capturing the implicit regularization in deep matrix factorization, suggesting the need for alternative explanations.
- **Dynamic Nature of Implicit Regularization**: The paper characterizes the dynamic nature of implicit regularization in deep matrix factorization, showing that the evolution rates of singular values are proportional to their magnitude and influenced by the depth of the factorization.
- **New Perspective on Optimization Trajectories**: The research proposes a detailed analysis of optimization trajectories as essential for understanding generalization in deep learning, indicating that this approach may be crucial for analyzing implicit regularization in non-linear neural networks

Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., & Srebro, N. (2018). The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research, 19*(70), 1–57.

- **Summary:** This paper delves into the role of gradient descent's implicit bias in the generalization of deep learning models. It suggests that the trajectory taken by gradient descent in the non-convex landscape of neural network optimization plays a critical role in determining generalization, influenced by factors such as learning rate and momentum.

Feldman, V. (2019). Does Learning Require Memorization? A Short Tale About a Long Tail. *arXiv preprint arXiv:1906.05271*.

- **Memorization in Overparameterized Learning**: Demonstrates that state-of-the-art learning algorithms, especially in image recognition, tend to memorize training labels, including random labels, which is not accounted for by existing theoretical models.
- **Necessity of Memorization for Optimal Generalization**: Proposes that memorization of labels, including those of outliers and noisy data, is necessary to achieve close-to-optimal generalization error in natural data distributions, especially those following long-tailed distributions.
- **Contextualizing Implicit Regularization**: Discusses the limitations of implicit regularization theories in explaining the zero training error phenomenon in overparameterized models, despite high generalization error.
- **Interpolating Algorithms and Memorization**: Explores the generalization properties of interpolating algorithms, demonstrating that

they can generalize effectively while tolerating noise. However, these studies don't fully explain why state-of-the-art classifiers on many datasets interpolate training data, indicating that memorization occurs beyond mere interpolation.

Tishby, N., & Zaslavsky, N. (2017). Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*.

- **Summary:** This work uses information theory to analyze how deep neural networks learn and generalize. The authors suggest that during training, networks initially memorize data, then gradually 'forget' to generalize better. The paper provides a novel perspective on the training dynamics of deep networks through information-theoretic measures.

Bubeck, S., & Sellke, M. (2021). A Universal Law of Robustness via Isoperimetry. *arXiv preprint arXiv:2105.12806*.

- **Summary:** Bubeck and Sellke's paper presents a proof showing the necessity of overparameterization for a network's robustness, which they argue is essential for good generalization. They contribute to the understanding of inductive biases in neural networks and their role in generalization.

Yang, Z., Lukasik, M., Nagarajan, V., Li, Z., Rawat, A. S., Zaheer, M., Menon, A. K., & Kumar, S. (2023). ResMem: Learn what you can and memorize the rest. arXiv preprint arXiv:2302.01576v2. Retrieved from https://arxiv.org/abs/2302.01576.

- **Summary**: In the approach outlined in "ResMem: Learn what you can and memorize the rest", an initial base model is trained and its predictions are used to calculate residuals. These residuals are then utilized to train a K-NN regressor. For making final predictions, the strategy involves the combined use of the base model and the K-NN regressor, thereby leveraging the strengths of both predictive modeling and memorization.

Lukasik, M., Nagarajan, V., Rawat, A. S., Menon, A. K., & Kumar, S. (2023). What do larger image classifiers memorize? arXiv preprint arXiv:2310.05337. Retrieved from https://arxiv.org/abs/2310.05337.

- **Bi-modal Distribution of Memorization**: As model complexity grows, examples are increasingly divided into two groups: highly memorized and minimally memorized.
- **Four Memorization Trajectories**: Identified trajectories are (a) increasing memorization for ambiguous or mislabeled examples, (b) decreasing memorization, (c) cap-shaped memorization peaking at a certain model size, and (d) constant memorization for clear examples.
- **Impact of Distillation on Memorization**: Knowledge distillation inhibits memorization, especially in examples where memorization increases with model size, suggesting it improves generalization by reducing memorization of difficult examples.
- **Stability-Based Memorization Score**: This approach assesses whether a model genuinely understands a pattern or merely memorizes specific answers, based on its consistency in answering a question despite variations in other data.

- **Knowledge Distillation Process**: Involves transferring knowledge from a larger model to a smaller one, aiming to build more resource-efficient models.
- **Impact of Distillation on Memorization Tendency**: Distillation especially reduces memorization in examples that were heavily memorized by the standard model. This reduction is more noticeable when there's a significant size difference between the teacher and student models.
- **Effect on Challenging and Ambiguous Examples**: Distillation decreases memorization in these examples, indicating a shift towards learning general patterns rather than memorizing difficult cases.
- **Observations from Figures**: Figure 2 provides an intuitive understanding, while Figure 3 shows that larger models have more examples of high memorization, and an increase in bi-modality in distribution with more examples at low and high memorization scores.
- **Conclusion**: The study reveals that memorization varies with model size, challenging the previous notion of fixed memorization characteristics across varying model sizes, and highlighting the dynamic nature of what is memorized as model size changes.