A NEW PERSPECTIVE ON APPLYING MESOSCIENCE TO EXPLORE THE MODEL GENERALIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The black-box nature of machine learning (ML) models, particularly neural networks, poses a significant challenge to their broader application in engineering, especially in high-risk areas where decision-making transparency and interpretability are crucial. Understanding the generalizability of ML models remains a key topic in artificial intelligence (AI), yet a unified understanding of this issue has not been established. This study introduces the concept of compromise in competition (CIC) from mesoscience to elucidate ML model generalizability. In this work, a scale decomposition method is proposed from the perspective of training samples, and the CIC between memorizing and forgetting, refined as dominant mechanisms, is studied. Empirical studies on computer vision (CV) and natural language processing (NLP) datasets demonstrate that the CIC between memorizing and forgetting affects model generalizability significantly. Moreover, techniques like dropout and L2 regularization, traditionally used to combat overfitting, can be reinterpreted through the CIC between memorizing and forgetting. Collectively, this work proposes a new perspective to explain the generalizability of ML models, in order to provide inherent support for further applications of ML models in the field of engineering.

1 INTRODUCTION

ML models, particularly neural networks, have been widely implemented across various engineering fields, such as plasticity prediction Mozaffar et al. (2019), material discovery Hatakeyama-Sato et al. (2020), and fault diagnosis Qin & Zhao (2022), demonstrating robust generalizability. Nonetheless, these models are often criticized for their black-box nature, meaning their prediction processes lack transparency Hassija et al. (2024). In engineering domains where safety and prediction reliability are paramount, such as medicine Shehab et al. (2022), chemical engineering Wen et al. (2024), and autonomous driving Bachute & Subhedar (2021), the interpretability of models is critically important Zhu et al. (2022).

The interplay between model generalizability and interpretability has become a critical area of research in AI applications. In biomedical research and healthcare, particularly in cancer research, ML presents numerous opportunities, including cancer detection, diagnosis, subtype classification, treatment optimization, and the identification of novel therapeutic targets in drug discovery Elemento et al. (2021). While ML models can enhance the accuracy of cancer diagnoses, improve patient prognoses, and reduce medical costs, the challenge of explainable AI persists Elemento et al. (2021). Limited interpretability may lead to a lack of trust in these technologies among healthcare professionals Alshuhri et al. (2023).

Over the years, researchers have sought to identify key factors influencing model generalizability. The "bias-variance trade-off" Geman et al. (1992), historically viewed as a foundational principle for understanding generalizability Neal et al. (2018); Yang et al. (2020), suggests that test loss can be decomposed into bias and variance. However, bias and variance are merely outcomes on test datasets, not the underlying causes of generalizability. The "model complexity-data complexity" paradigm posits that optimal generalization is achieved when model complexity aligns with data complexity Myung (2000). Numerous studies have investigated how this relationship affects generalizability Hastie et al. (2022); Mei & Montanari (2022); Schaeffer et al. (2023), yet no unified quantitative standards for model complexity Hu et al. (2021) and data complexity Ho & Basu (2002);

Li et al. (2018a); Branchaud-Charron et al. (2019) have been established. A major challenge in unifying explanations for generalizability is the complexity of ML models and their training datasets. For instance, GPT-4, with 1.8 trillion parameters Raiaan et al. (2024), requires vast training data for robust generalization, making it difficult to describe the training process with precise mathematical or physical formulas. Thus, there is an urgent need for innovative approaches to enhance the interpretability of model prediction processes.

Recently, mesoscience (Ge et al. (2007); Li et al. (2018b)), which argues that the system complexity stems from the CIC between the two (or more) coexisting dominant mechanisms, has been proposed to cope with multilevel complexities. Instead of relying on traditional mathematical and physical formulas, mesoscience analyzes the CIC to realize the connections between system behaviors and underlying mechanisms. This approach involves performing scale decomposition, refining dominant mechanisms, and analyzing their CIC. Taking two dominant mechanisms in a system as an example, with the increasing dominance of mechanism B over mechanism A, three regimes can appear in turn: mechanism A dominates, mechanism A-mechanism B compromise, and mechanism B dominates, corresponding to different system behaviors, respectively Huang et al. (2018). The principle of mesoscience has been applied in multiple complex systems successfully, e.g., chemical engineering Li et al. (2016), life sciences Qian & Beltran (2022), geology Tordesillas et al. (2021). Guo et al. (2019) proposed a research paradigm of AI, which introduces the analytical principles of mesoscience into the design of deep learning models. This paradigm has led to the development of mesoscience-guided deep learning (MGDL), which has demonstrated remarkable improvements in terms of convergence stability and predictive accuracy Guo et al. (2024). Therefore, the application of mesoscience principles offers a promising methodological approach to explore the generalizability of ML models.

2 METHODOLOGY

This work employs the principles of mesoscience to elucidate the generalizability of neural networks, given their extensive applications. The research framework is depicted in Figure 1.

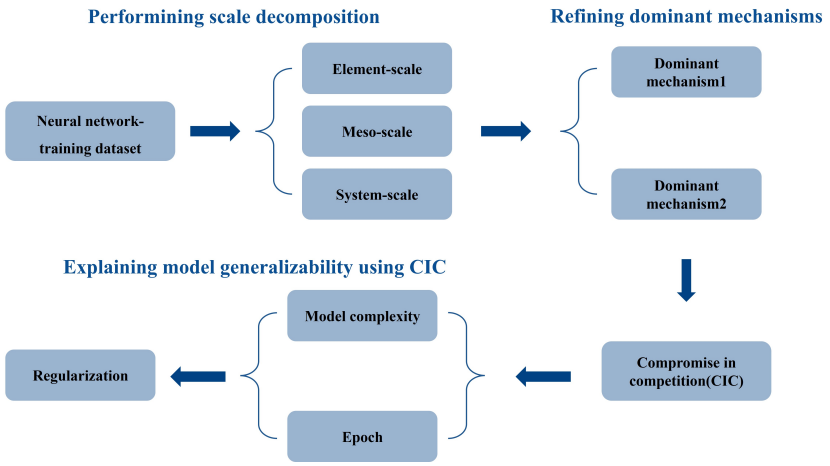

Figure 1: The research framework consists of four parts: performing scale decomposition; refining the dominant mechanisms; analyzing the CIC between them; explaining model generalizability using CIC.

For the study of complex systems, it is crucial to consider their multi-scale characteristics, particularly the quantification of meso-scale structures. A comprehensive understanding and control of system dynamics necessitate appropriate scale decomposition Ren et al. (2001), making it the initial step in mesoscience research Li & Huang (2014). This method should identify the characteristic scale that reflects the observed structure, based on the multi-scale properties of complex systems. For instance, in classical two-phase flow research, the element-scale, meso-scale, and system-scale correspond to the particle, cluster, and overall two-phase flow system, respectively, indicating that

characteristic scales must have clear physical meanings Li & Kwauk (2003). Additionally, meso-science research depends on refining dominant mechanisms. Complex systems may encompass multiple dominant mechanisms. Li & Huang (2014) proposed that it is important to group all dominant mechanisms into two integrated ones, each driving the system in opposing directions. While these mechanisms differ across systems, they adhere to the same principle of CIC. Notably, the CIC between dominant mechanisms varies across different scales Li & Huang (2014) and should be clarified. This study seeks to use CIC to uniformly explain changes in model generalizability induced by model complexity and the number of training epochs, both common factors influencing generalizability. Furthermore, the analysis of why various regularization methods mitigate overfitting will demonstrate CIC's effectiveness in explaining model generalizability.

2.1 Experiment setup

This work follows the well-established experiment setups of previous studies (Nakkiran et al. (2021); Han et al. (2020)).

- **Fully connected neural network (FCNN):** This architecture implementation is adopted from Nakkiran et al. (2021). with model complexity adjusted by modifying the width of the initial hidden layer (w) within the range [1, 10].
- **Four-layer convolutional neural network (Four-layer CNN):** This architecture implementation is adopted from Han et al. (2020). The models are formed by two convolutional layers and two fully connected layers. For all convolutional layers, the kernel size = 3, stride = 1, and padding = 0.
- **Five-layer convolutional neural network (Five-layer CNN):** This architecture implementation is adopted from Nakkiran et al. (2021). The models are formed by 4 convolutional stages of controlled base width [w, $2w$, $4w$, $8w$], for w in the range of [1, 10] and one fully connected layer.
- **Text convolutional neural network (TextCNN):** This architecture implementation is adopted from Han et al. (2020). The embedding dimension is 300, and the width of the convolutional layer (w) is 5.

In subsequent specific experiments, FCNNs and Four-layer CNNs are used to train MNIST LeCun et al. (1998), Five-layer CNNs are used to train CIFAR-10 Krizhevsky et al. (2009), and TextCNNs are used to train TREC Li & Roth (2002).

2.2 Scale decomposition

This study introduces a scale decomposition method for complex systems, comprising neural networks and training datasets. The individual training sample, which contains the necessary features for model training, is defined as the element-scale, while the entire dataset represents the system-scale. During training, the model updates its parameters in discrete batches, as illustrated in Figure 2, highlighting the significant impact of batch size and sample composition on model generalizability. Figure 2(a) demonstrates the trade-off between increased batch size and decreased model generalizability. This finding, derived from training a Five-layer CNN ($w = 3$) on the CIFAR-10 dataset with label noise ($p = 0.2$), is further supported by previous work Hoffer et al. (2017), which explored the effect of batch size on model generalizability more detailedly. Additionally, this study varies the composition of samples by adjusting random seeds, as shown in Figure 2(b), where such changes significantly affect generalizability. The batch's crucial role in training dynamics and generalization qualifies it as the meso-scale. As a case study, Figure 3 illustrates the scale decomposition of neural network-CIFAR-10 dataset system.

2.3 Dominant mechanisms

This study systematically examines various potential mechanisms, such as bias and variance, the number of parameters and the number of training samples, the number of clean data and the number of noisy data, etc. Ultimately, memorizing and forgetting are refined as the dominant mechanisms.

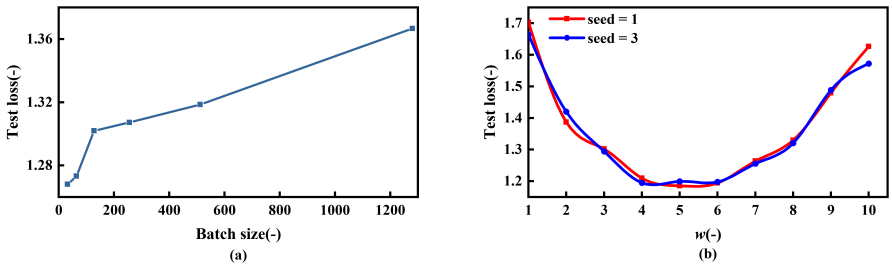

Figure 2: Training batch affects model generalizability: (a) Increasing batch size leads to a rise in test loss, suggesting reduced generalizability; (b) Variations in random seed, affecting sample composition in batches, are shown to affect the generalizability of ML models with constant architecture but varying network width (w).

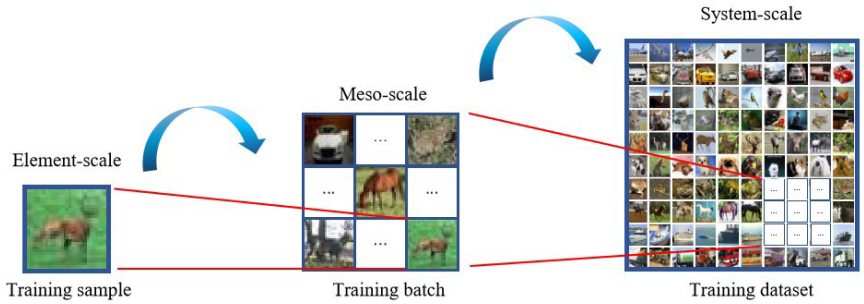

Figure 3: Taking the CIFAR-10 dataset as an example, scale decomposition is performed from the perspective of training samples. The individual training sample is the element-scale, the training batch is the meso-scale, and the entire CIFAR-10 training dataset is the system-scale.

Artificial neural networks (ANNs) discern features and formulate decisions through hierarchical abstraction, mimicking the integrative process of neuronal activity in the brain Uyanik et al. (2022). Neuroscience has long inspired advancements in AI, including CNNs LeCun et al. (1989) and reinforcement learning (RL) Mnih et al. (2015), as noted by Zador et al. (2023). Cognitive neuroscience reveals that humans continuously gather external data through sensory inputs, accumulating vast information. Like ML models, the brain processes information and makes decisions while facing risks of underfitting and overfitting. Since the brain constructs experiences and anticipates future scenarios through memorizing of learned data, insufficient memorizing may lead to underfitting. Conversely, excessive memorizing of details can result in overfitting, where reliance on past experiences hinders adaptation to new environments. Overfitting in the brain is linked to disorders such as post-traumatic stress disorder, depression, schizophrenia, and obsessive-compulsive disorder Sha et al. (2024). Gravitz (2019) demonstrates that forgetting, as an adaptive learning form, aids humans in adapting their experiences, thus preventing experiential overfitting. The importance of memorizing and forgetting extends beyond neuroscience, garnering interest from scholars in education, ecology, and linguistics Sha et al. (2024). Similar to the human brain, ML models undergo processes of memorizing and forgetting during training, significantly impacting their generalizability. For instance, scaling laws suggest that larger model sizes enhance memorization capabilities, thereby improving generalizability Kaplan et al. (2020). However, not all memorizing positively contributes to model generalizability. For example, excessive memorizing in large language models (LLMs) can lead to hallucinations Huang et al. (2023), undermining content reliability. Traditionally, forgetting is viewed as detrimental to model performance McCloskey & Cohen (1989), but recent insights from neuroscience suggest that beneficial forgetting is an adaptive function that enhances model generalizability Peng et al. (2021). Therefore, this study refines memorizing and forgetting as the dominant mechanisms, given their crucial role in influencing model generalizability.

Over time, researchers have adapted various definitions tailored to specific problems Carlini et al. (2023); Pondenkandath et al. (2018); Stern & Weinshall (2023). This study builds upon and extends the definitions proposed by Toneva et al. (2018): during training, if a model fails to accurately predict a training sample at time t, but succeeds at the subsequent time step $t+1$, it indicates that the model has memorized the sample. Conversely, if the model accurately predicts a training sample at time t but fails at $t+1$, it indicates that the model has forgotten the sample. Training samples can be categorized into three distinct sets: S_1, which includes samples that are neither memorized nor forgotten; S_2, which includes samples that are only memorized and remain unforgettable subsequently (the unforgettable examples, as defined by Toneva et al. (2018)); S_3, which includes samples that are both memorized and forgotten at least once. These sets are mutually exclusive, and their union constitutes the entire training dataset S.

A list synchronized with the training epochs tracks the model's predictions for individual samples throughout training, where zeros denote inaccurate predictions and ones indicate accurate predictions, as illustrated in Figure 4. Training samples in S_1 are not the focus of this study, as they are neither memorized nor forgotten by the model. According to Toneva et al. (2018), samples in S_2 carry limited information and thus have a negligible effect on model generalizability. Experiments on datasets without label noise, such as MNIST and CIFAR-10, confirm that removing S_2 does not significantly impact model generalizability. In contrast, training samples in S_3 have a substantial effect on model generalizability Toneva et al. (2018).

Consequently, this study focuses on the memorizing and forgetting of training samples in S_3 by neural networks. To quantify the model's overall memorizing and forgetting of these samples, the degree of memorizing ($\mathcal{M}$) and the degree of forgetting ($\mathcal{F}$) are introduced. Specifically, $\mathcal{M}$ denotes the fraction of samples in S_3 that have been memorized by the end of training, while $\mathcal{F}$ denotes the fraction of those that have been forgotten in S_3.

$$\mathcal{M} = \frac{N_{acc=1}}{N_{acc=1} + N_{acc=0}} \tag{1}$$

$$\mathcal{F} = \frac{N_{acc=0}}{N_{acc=1} + N_{acc=0}} \tag{2}$$

where $N_{acc=1}$ denotes the number of training samples in S_3 predicted accurately by the end of training. $N_{acc=0}$ denotes the number of training samples in S_3 cannot predicted accurately by the end of training.

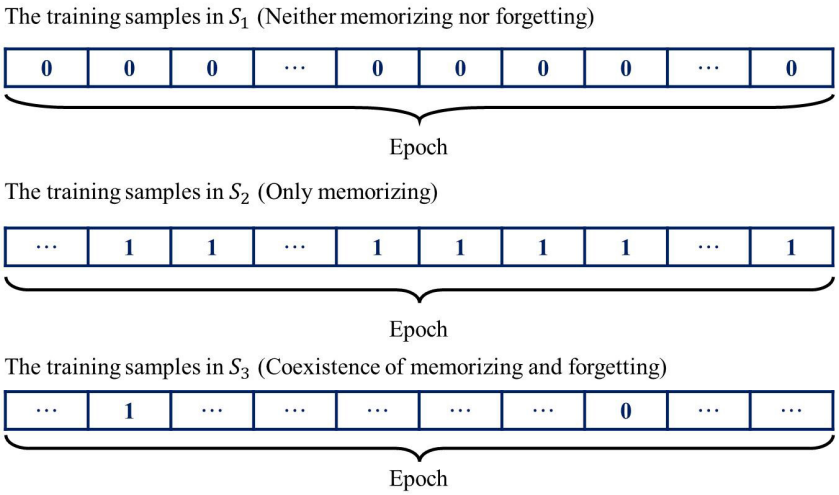

Figure 4: The model's memorizing and forgetting of samples in S_1, S_2, S_3 during training. S_1 includes the samples that are neither memorized nor forgotten; S_2 includes the samples that are only memorized and remain unforgettable; S_3 includes the samples that are both memorized and forgotten at least once, where zeros denote inaccurate predictions and ones indicate accurate predictions.

2.4 The CIC between memorizing and forgetting on different scales

The CIC between dominant mechanisms exhibits variations on different scales Li & Huang (2014), and thus should be clarified. The analysis process, exemplified by training a Five-layer CNN ($w = 3$) on the CIFAR-10 dataset with label noise ($p = 0.2$), is as follows: when memorizing is dominant, $\mathcal{M}$ tends towards its maximum, whereas when forgetting is dominant, $\mathcal{F}$ tends towards its maximum. Figure 5 illustrates that on the element-scale (points A and B), the extremum tendencies of $\mathcal{M}$ and $\mathcal{F}$ can be realized only instantaneously and alternatively. Spatially, at a specific moment, such as the completion of the 106th epoch, the model's forgetting of point A is dominant, indicating the extremum tendency of $\mathcal{F}$ is realized, while the model's memorizing of point B is dominant, indicating the extremum tendency of $\mathcal{M}$ is realized. Evidently, stability conditions do not exist on the element-scale. In the meso-scale region M (batch index = 1), the extremum tendencies of $\mathcal{M}$ and $\mathcal{F}$ still cannot be realized simultaneously. However, in this region, a spatio-temporal compromise occurs between these dominant mechanisms. Over time, $\frac{\mathcal{M}}{\mathcal{F}}$ gradually converges to a constant C_1, with fluctuations of a certain amplitude due to the competition between memorizing and forgetting. This analysis suggests the CIC between memorizing and forgetting on the meso-scale. In the system-scale region G, the CIC is even more pronounced, with $\frac{\mathcal{M}}{\mathcal{F}}$ converging to another constant C_2 over time and the fluctuations diminishing.

3 The CIC between memorizing and forgetting explains model generalizability

3.1 Changes in model complexity

In the context of training FCNNs (for w in the range of [1, 10]) on the MNIST dataset with label noise ($p = 0.4$), Figure 6(a) illustrates that training loss decreases with model complexity, while the test loss decreases initially and increases subsequently, indicative of overfitting at excessive model complexity. Figure 6(b) indicates the dynamics between $\mathcal{M}$ and $\mathcal{F}$ in the neural network-training dataset system. When the model complexity is too low ($w = 1$), forgetting is dominant relative to memorizing absolutely. The fact that the model's capacity for memorizing sufficient training data is limited, and the data memorized are prone to being quickly forgotten results in poor performance on both training and test datasets, leading to underfitting. The dominance of memorizing over forgetting continuously increases with the model complexity, and the system transitions from being forgetting-dominated to being memorizing-dominated. When $w = 3$, $\frac{\mathcal{M}}{\mathcal{F}} = 1.08$, the model exhibits the best

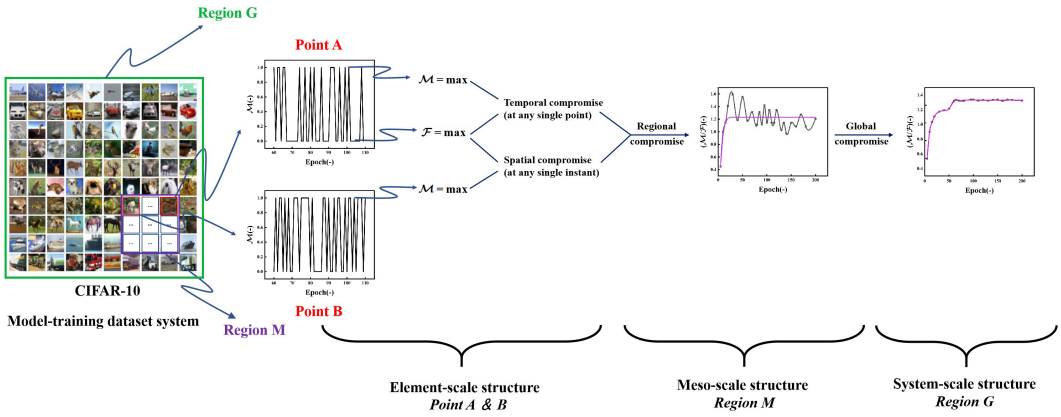

Figure 5: The CIC between memorizing and forgetting on different scales (batch index = 1). On the element scale, there is no stability conditions, while on the meso-scale and system-scale, there is CIC between memorizing and forgetting. The black line represents the change in $\mathcal{M}$ (element-scale) or $\frac{\mathcal{M}}{\mathcal{F}}$ (meso-scale and system-scale) with the number of training epochs, and the purple line represents the trend line (This analytical approach is inspired by Li et al. (2004)).

generalizability on the test dataset. On one hand, the model is capable of memorizing sufficient and accurate information. On the other hand, although the model may memorize some details (such as label noise), it forgets them eventually, thus preventing further harm to model generalizability. However, when the model complexity is too high (w = 10), memorizing is dominant relative to forgetting absolutely. The model has the capability to memorize a vast amount of details from the training dataset. Additionally, since the extremum tendency of $\mathcal{F}$ is inhibited, these memorized details are difficult for the model to forget, resulting in excellent performance on the training dataset but poor generalizability on the test dataset, leading to overfitting.

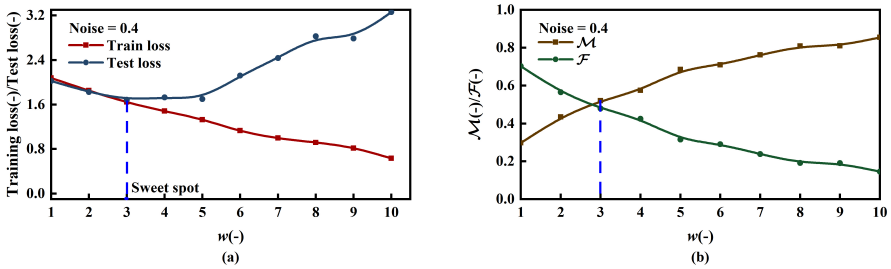

Figure 6: The change in model complexity affects the dominance of memorizing over forgetting: (a) the model tends to exhibit overfitting with the model complexity; (b) The dominance of memorizing over forgetting continuously increases with the model complexity.

The above analysis reveals that the neural network-training dataset system transitions through three distinct regimes with model complexity: When forgetting is dominant absolutely, the model performs poorly on both training and test datasets, exhibiting underfitting; When neither memorizing nor forgetting can dominate absolutely, the model shows a U-shaped test loss curve with model complexity; When memorizing is dominant absolutely, the model excels on training dataset but performs poorly on test dataset, exhibiting overfitting, as shown in Figure 7. Moreover, label noise in training datasets significantly impacts model generalizability. This study examines the effects of varying label noise on $\mathcal{M}$ and $\mathcal{F}$, as depicted in Figure 8. The findings indicate that increased label noise hinders the model's memorizing of training data, but enhances the forgetting of memorized data.

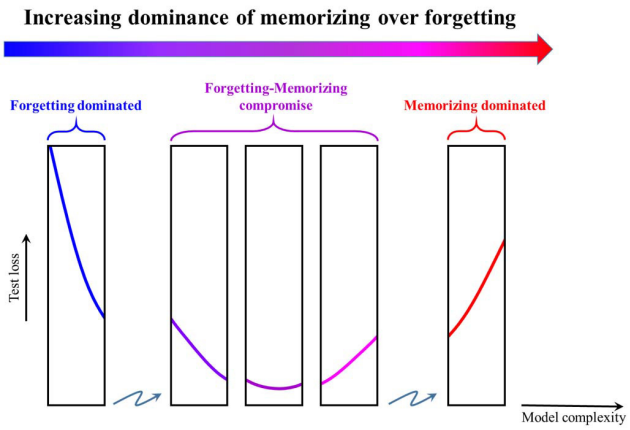

Figure 7: Three regimes occur successively with increasing dominance of memorizing over forgetting. They are the forgetting-dominated regime, the forgetting-memorizing compromising regime, and the memorizing-dominated regime, respectively.

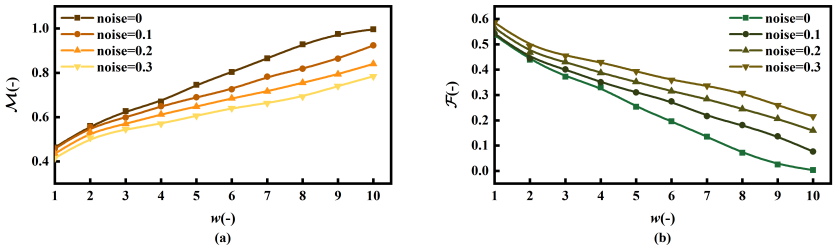

Figure 8: Increasing levels of label noise across different neural networks with different widths(w)-CIFAR-10 dataset system consistently result in decrease in $\mathcal{M}$ and increase in $\mathcal{F}$, indicating that label noise can decrease the relative dominance of memorizing over forgetting during the training process.

3.2 CHANGES IN EPOCH

In the field of NLP, the training of a TextCNN with width (w) = 5 on the TREC dataset with label noise ($p = 0.2$) illustrates how the changing dominance of memorizing over forgetting affects model generalizability. Figure 9(a) shows that training loss decreases with epochs, while test loss initially decreases and then increases, indicating overfitting as the number of epochs becomes excessive. Figure 9(b) reveals that $\mathcal{M}$ and $\mathcal{F}$ in the neural network-training dataset system evolve continuously with epochs. With insufficient epochs, limited parameter updates lead to the absolute dominance of forgetting, resulting in inadequate memorizing and rapid forgetting, thus poor performance on both training and test datasets, indicative of underfitting. The system transitions from being forgetting-dominated to being memorizing-dominated with the number of training epochs. When the ninth epoch is completed, the model achieves optimal generalizability on the test dataset by prioritizing the fitting of regular data Arpit et al. (2017), enabling it to memorize intrinsic patterns while forgetting non-generalizable details. However, with excessive epochs, memorizing becomes absolutely dominant, causing the model to retain too many non-generalizable details. The suppression of the extremum tendency of $\mathcal{F}$ makes it difficult to forget these details, resulting in excellent training performance but poor test performance, indicative of overfitting.

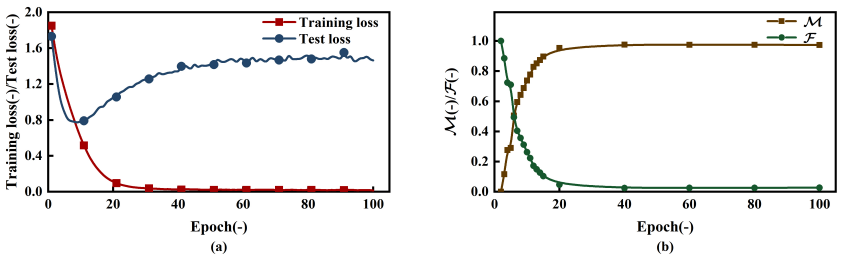

Figure 9: The changes in epoch affect the dominance of memorizing over forgetting: (a) The model tends to exhibit overfitting with the number of epochs; (b) The dominance of memorizing over forgetting continuously increases with the number of epochs.

3.3 REGULARIZATIONS

This study explores the effectiveness of regularization techniques, such as dropout and L2 regularization, in mitigating overfitting. In neural networks, neurons form co-adaptation relationships through interconnections and signal transmissions, capturing intrinsic patterns in training data. The dropout technique temporarily removes random neurons during training, reducing sensitivity to training data. Figure 10(a) shows that without dropout (dropout rate = 0), the model tends to overfit. An optimal dropout rate, like 0.7, enhances generalizability, whereas an excessive dropout rate, such as 0.95, can cause underfitting. Figure 10(b) indicates that without dropout, memorizing is absolutely dominant over forgetting, leading to overfitting. Moderate dropout rates, which disrupt some co-adaptations, compel the model to focus on common features of the training dataset while forgetting specific details, leading to optimal generalizability. Conversely, the excessive dropout rate increases the dominance of forgetting over memorizing which makes the model struggle to memorize but forget easily the effective information in the training dataset, leading to underfitting.

Introducing an additional penalty term, such as L2 regularization, to the loss function during training is a common method to constrain model complexity. L2 regularization reduces model complexity by adding a penalty term for the L2 norm of weight parameters to the loss function. The regularization parameter λ is used to control the strength of the regularization term. A larger λ increases the degree of regularization, forcing the model to adopt a simpler form and thereby reduce its complexity. For instance, in training a Four-layer CNN on MNIST with label noise ((p=0.2)), Figure 11 illustrates that similar to dropout, L2 regularization modulates the model generalizability by controlling the relative dominance between memorizing and forgetting.

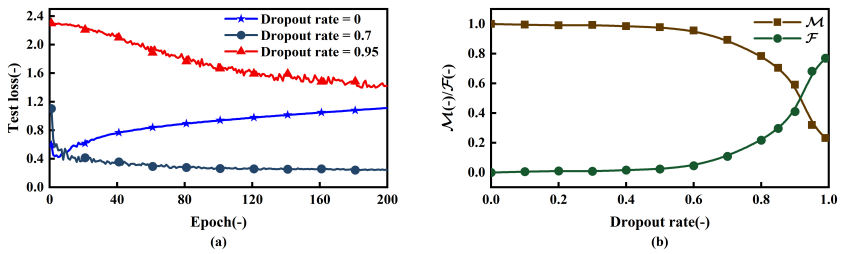

Figure 10: Dropout changes the relative dominance between memorizing and forgetting: (a) The model gradually transitions from underfitting to overfitting with dropout rate increasing from 0 to 0.95; (b) The dominance of memorizing over forgetting continuously increases with dropout rate.

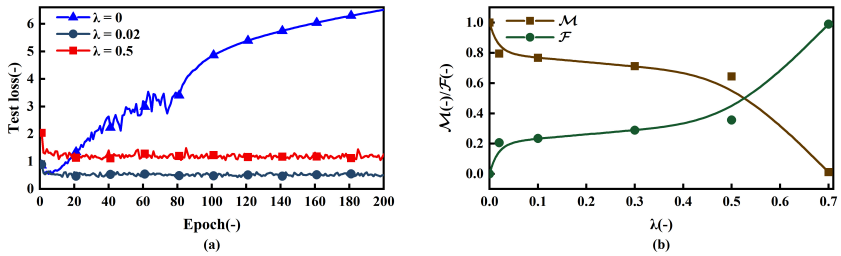

Figure 11: L2 regularization changes the relative dominance between memorizing and forgetting: (a) The model gradually transitions from underfitting to overfitting with λ increases from 0 to 0.5; (b) The dominance of memorizing over forgetting continuously increases with λ.

4 Conclusion

This work explains the generalizability of ML models based on the principle of mesoscience, focusing on the model's memorizing and forgetting of training samples during training, and analyzes the CIC between memorizing and forgetting. Additionally, this work proposes $\mathcal{M}$ and $\mathcal{F}$ to quantify the relative dominance between memorizing and forgetting. The following conclusions are as follows:

(1) The individual training sample is considered as the element-scale, where memorizing and forgetting only compete during training; a batch of training samples is considered as the meso-scale, where memorizing and forgetting exhibit spatio-temporal compromise; and the entire training dataset is considered as the system-scale, where the spatio-temporal compromise is more evident.

(2) The increase of model complexity and the number of training epochs both can promote the extremum tendency of $\mathcal{M}$ and inhibit the extremum tendency of $\mathcal{F}$, which make the ML model-training dataset system transition from being forgetting-dominated to being memorizing-dominated gradually.

(3) Regularization methods such as dropout, L2 regularization, although proposed from different research perspectives, control the relative dominance between memorizing and forgetting to improve model generalizability essentially.

References

Mohammed Alshuhri, Sada Ghalib Al-Musawi, Ameen Abdulhasan Al-Alwany, Herlina Uinarni, Irodakhon Rasulova, Paul Rodrigues, Adnan Taan Alkhafaji, Asim Muhammed Alshanberi, Ahmed Hussien Alawadi, and Ali Hashim Abbas. Artificial intelligence in cancer diagnosis: Opportunities and challenges. *Pathology-Research and Practice*, pp. 154996, 2023.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Mrinal R Bachute and Javed M Subhedar. Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, 6:100164, 2021.

Frederic Branchaud-Charron, Andrew Achkar, and Pierre-Marc Jodoin. Spectral metric for dataset complexity assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3215–3224, 2019.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Olivier Elemento, Christina Leslie, Johan Lundin, and Georgia Tourassi. Artificial intelligence in cancer research, diagnosis and therapy. *Nature Reviews Cancer*, 21(12):747–752, 2021.

Wei Ge, Feiguo Chen, Jian Gao, Shiqiu Gao, Jin Huang, Xiaoxing Liu, Ying Ren, Qicheng Sun, Limin Wang, Wei Wang, et al. Analytical multi-scale method for multi-phase complex systems in process engineering—bridging reductionism and holism. *Chemical Engineering Science*, 62(13): 3346–3377, 2007.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Lauren Gravitz. The importance of forgetting. *Nature*, 571(July):S12–S14, 2019.

Li Guo, Jun Wu, and Jinghai Li. Complexity at mesoscales: a common challenge in developing artificial intelligence. *Engineering*, 5(5):924–929, 2019.

Li Guo, Fanyong Meng, Pengfei Qin, Zhaojie Xia, Qi Chang, Jianhua Chen, and Jinghai Li. A case study applying mesoscience to deep learning. *Engineering*, 2024.

Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Kan Hatakeyama-Sato, Toshiki Tezuka, Momoka Umeki, and Kenichi Oyaizu. Ai-assisted exploration of superionic glass-type li+ conductors with aromatic structures. *Journal of the American Chemical Society*, 142(7):3301–3305, 2020.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):289–300, 2002.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63:2585–2619, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023. doi: 10.48550/ARXIV.2311.05232. URL https://doi.org/10.48550/arXiv.2311.05232.

Wenlai Huang, Jinghai Li, and Peter P Edwards. Mesoscience: exploring the common principle at mesoscales. *National science review*, 5(3):321–326, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *Handbook of Systemic Autoimmune Diseases*. Toronto, ON, Canada, 2009.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018a. URL https://openreview.net/forum?id=ryup8-WCW.

Jinghai Li and Huang. *Towards mesoscience: the principle of compromise in competition*. Springer, 2014.

Jinghai Li and Mooson Kwauk. Exploring complex systems in chemical engineering—the multi-scale methodology. *Chemical Engineering Science*, 58(3-6):521–535, 2003.

Jinghai Li, Jiayuan Zhang, Wei Ge, and Xinhua Liu. Multi-scale methodology for complex systems. *Chemical engineering science*, 59(8-9):1687–1700, 2004.

Jinghai Li, Wei Ge, Wei Wang, Ning Yang, and Wenlai Huang. Focusing on mesoscales: from the energy-minimization multiscale model to mesoscience. *Current Opinion in Chemical Engineering*, 13:10–23, 2016.

Jinghai Li, Wenlai Huang, Jianhua Chen, Wei Ge, and Chaofeng Hou. Mesoscience based on the emms principle of compromise in competition. *Chemical Engineering Journal*, 333:327–335, 2018b.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Mojtaba Mozaffar, Ramin Bostanabad, W Chen, K Ehmann, Jian Cao, and MA Bessa. Deep learning predicts path-dependent plasticity. *Proceedings of the National Academy of Sciences*, 116(52): 26414–26420, 2019.

In Jae Myung. The importance of complexity in model selection. *Journal of mathematical psychology*, 44(1):190–204, 2000.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *CoRR*, abs/1810.08591, 2018. URL http://arxiv.org/abs/1810.08591.

Jian Peng, Xian Sun, Min Deng, Chao Tao, Bo Tang, Wenbo Li, Guohua Wu, Qing Zhu, Yu Liu, Tao Lin, and Haifeng Li. Learning by active forgetting for neural networks. *CoRR*, abs/2111.10831, 2021. URL https://arxiv.org/abs/2111.10831.

Vinaychandran Pondenkandath, Michele Alberti, Sammer Puran, Rolf Ingold, and Marcus Liwicki. Leveraging random label memorization for unsupervised pre-training. *CoRR*, abs/1811.01640, 2018. URL http://arxiv.org/abs/1811.01640.

Haili Qian and Adriana Sujey Beltran. Mesoscience in cell biology and cancer research. *Cancer Innovation*, 1(4):271–284, 2022.

Ruoshi Qin and Jinsong Zhao. Adaptive multiscale convolutional neural network model for chemical process fault diagnosis. *Chinese Journal of Chemical Engineering*, 50:398–411, 2022.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

Jinqiang Ren, Qiming Mao, Jinghai Li, and Weigang Lin. Wavelet analysis of dynamic behavior in fluidized beds. *Chemical Engineering Science*, 56(3):981–988, 2001.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W. Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. *CoRR*, abs/2303.14151, 2023. doi: 10.48550/ARXIV.2303.14151. URL https://doi.org/10.48550/arXiv.2303.14151.

Alyssa Shuang Sha, Bernardo Pereira Nunes, and Armin Haller. "forgetting" in machine learning and beyond: A survey. *CoRR*, abs/2405.20620, 2024. doi: 10.48550/ARXIV.2405.20620. URL https://doi.org/10.48550/arXiv.2405.20620.

Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H Gandomi. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

Uri Stern and Daphna Weinshall. Relearning forgotten knowledge: on forgetting, overfit and training-free ensembles of dnns. *CoRR*, abs/2310.11094, 2023. doi: 10.48550/ARXIV.2310.11094. URL https://doi.org/10.48550/arXiv.2310.11094.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Antoinette Tordesillas, Sanath Kahagalage, Lachlan Campbell, Pat Bellett, Emanuele Intrieri, and Robin Batterham. Spatiotemporal slope stability analytics for failure estimation (sssafe): linking radar data to the fundamental dynamics of granular failure. *Scientific Reports*, 11(1):9729, 2021.

Cihan Uyanik, M Ahmed Khan, Iris C Brunner, John P Hansen, and Sadasivan Puthusserypady. Machine learning for motor imagery wrist dorsiflexion prediction in brain-computer interface assisted stroke rehabilitation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 715–719. IEEE, 2022.

Kaijie Wen, Li Guo, Zhaojie Xia, Sibo Cheng, and Jianhua Chen. A hybrid simulation method integrating cfd and deep learning for gas–liquid bubbly flow. *Chemical Engineering Journal*, 495:153515, 2024.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

Li-Tao Zhu, Xi-Zhong Chen, Bo Ouyang, Wei-Cheng Yan, He Lei, Zhe Chen, and Zheng-Hong Luo. Review of machine learning for hydrodynamics, transport, and reactions in multiphase flows and reactors. *Industrial & Engineering Chemistry Research*, 61(28):9901–9949, 2022.