**Scientific Background**

Deep learning has witnessed unprecedented advancements in recent years, propelling artificial intelligence (AI) to new heights and enabling a wide range of applications, from computer vision and natural language processing to autonomous vehicles and healthcare diagnostics. Understanding deep learning calls the following questions to be addressed: (i) optimization—the effectiveness of gradient-based algorithms in solving neural-network training programs that are non-convex and thus seemingly difficult—and (ii) generalization—the phenomenon of deep learning models that do not overfit the data, even when they have many more parameters than examples to learn from. Existing analyses of generalization typically adopt the language of classical learning theory, abstracting away many details from the considered setting. Despite the remarkable strides in the field of deep learning, a striking gap exists in our understanding of how to consistently achieve robust and effective deep model generalization. This gap in the current literature underscores the critical need for further research, development, and innovation in this area. In this grant, I will study model generalization and its relationship to the key blocks of normalization, optimization, and regularization. This endeavor promises not only to enhance our understanding of deep learning generalization but also to empower the deployment of more resilient, adaptable, and capable deep learning systems across diverse domains.

## **Deep Learning Generalization**

Deep model generalization refers to the ability of trained neural networks to make accurate predictions on data not encountered during training. The aim is to learn generic feature representations agnostic to domains (or datasets) and enable trained models to perform well in completely new domains. Achieving this goal demands the development of models that capture valuable information shared across domains and recognize semantically related but visually distinct examples. In essence, it is the litmus test for a deep learning model's practical utility, as real-world applications necessitate not just training data memorization but understanding of the underlying patterns to facilitate confident decision-making in new situations. Existing challenges include the need for strategies for data-efficient generalization in scenarios with limited data, enhanced model robustness, and techniques for efficient domain adaptation. Moreover, as deep models grow in complexity, the imperative for model generalizability and resource efficiency grows, as the concern is that higher model complexity increases overfitting to the training data.

Generalization techniques in deep learning can be broadly categorized into various groups based on their approach. Some algorithms revolve around the creation of innovative loss functions that facilitate the learning of domain-agnostic representations [1–4]. These loss functions are designed to minimize the influence of domain-specific characteristics during training. Another set of techniques focuses on deep neural network architectures that are inherently robust to domain shifts, thereby promoting generalization [5–7]. The research proposed in this grant falls into the latter category, specifically emphasizing the enhancement of network architectures to improve generalization capabilities. In particular, our research delves into and introduces novel adaptive models for three key deep-learning components: data normalization, deep optimization and network regularization. These components play pivotal roles in the training process and the overall performance of deep neural networks. **These adaptive models enhance both performance and generalizability, striking a delicate balance between model simplicity and complexity. In other words, our adaptive models, as proposed, empower us to achieve the same task accuracy for simple models as that obtained by complex models while mitigating the risk of overfitting, which is common in complex models but not in simple ones.**

## Data Normalization

Data normalization and deep model generalization are closely related in the context of deep neural-network training. Data normalization scales and centers the input data to the network or a specific hidden layer, whereas deep model generalization refers to a model's ability to perform well on unseen data. Normalization affects model generalization for the following reasons:

1. **Training Stabilization**: Deep learning models, particularly neural networks with many layers, are sensitive to the scale and distribution of input features. When these features have vastly different scales, some network weights may update much more slowly than others during training, leading to convergence issues and longer training times. Data normalization scales all input features to a similar range, which helps stabilize the training process.
2. **Faster Convergence**: Normalizing the data often results in faster convergence during training. This is because the optimization algorithm can more effectively navigate the loss landscape when the features are on a similar scale. Faster convergence leads to better generalization since the model does not spend as much time fitting the training data noise.
3. **Overfitting Mitigation:** Data normalization also helps prevent overfitting. Overfitting occurs when a model becomes too complex and starts fitting the noise in the training data rather than capturing the underlying patterns. When input features are normalized, overfitting becomes less likely because the model focuses on the relevant patterns rather than noise.
4. **Improved Gradient Flow:** Normalizing the data can improve gradient flow during backpropagation; that is, the gradients used to update the model's weights are less likely to explode or vanish. Improved gradient flow can help the model generalize better because it can learn more effectively from the training data.
5. **Transferability**: When a deep model is trained on normalized data, it is more likely to transfer well to new and unseen datasets. Since the model learns patterns invariant to the scale and distribution of the input features, it is better equipped to handle a broader range of data, improving its generalization capabilities.

Several existing normalization techniques address these challenges. For instance, Batch Normalization (BN) [8] conducts global normalization across the batch dimension. However, its sensitivity to batch size and the impracticality of batch-wise normalization during inference necessitate the use of pre-computed mean and variance values derived from the training set via running averages [8]. These pre-computed statistics may become unreliable when the distribution of the target data shifts, creating disparities between the training and testing phases. To mitigate such issues, alternative normalization methods [9-11] have surfaced. Layer Normalization [9] operates over all channels along the layers dimension, Instance Normalization [10] conducts calculations akin to BN on individual samples, and Weight Normalization [11] introduces filter weight normalization. Despite their merits, these techniques are not as accurate as BN in various visual recognition tasks. A more recent technique, Group Normalization (GN), segments channels into groups and normalizes features within them [12]. GN initializes channel groups in sequential order in the first epoch and maintains these groups throughout training. However, this method relies on the premise that a deep network will progressively learn optimal channel weights, which is not universally valid. The effectiveness of this method is variable.

**In this research grant, our aim is to delve deeper into the normalization–generalization relationship, studying the impact of hybrid normalization on the generalization of deep networks. To do that, we will introduce an innovative adaptive strategy that overcomes the constraints of existing normalization methods.**

## Deep Optimization

Deep learning architectures rely heavily on adaptive and non-adaptive optimization algorithms, which improve convergence towards the global minimum through techniques such as momentum, individual learning rates, and Exponential Moving Averages (EMAs). Example optimizers are Stochastic Gradient Descent (SGD), RMSprop, Adam, AdaGrad, AdamW, AdaHessian, and others that combine SGD and Adam [13–14]. These optimizers mostly depend on EMAs, which improve generalization for the following reasons:

1. **Training Stabilization**: EMA-based optimizers introduce stability into the training process by maintaining a smoothed version of the model's weights over time, reducing rapid fluctuations during training. This stabilization can prevent the model from overfitting to noise (and help identify real data patterns, as it should), improving generalization.
2. **Noise Reduction**: During deep neural network training, the optimization process can encounter noisy updates, especially when batch sizes are small or the loss landscape is complex. Deep EMA-based optimizers, by averaging parameter values over time, reduce the impact of such noise. This leads to more consistent weight updates and helps the model generalize better by focusing on the underlying patterns in the data.
3. **Regularization**: EMA-based optimizers effectively act as a form of regularization. By maintaining an exponential moving average of the model's weights, they encourage the model to explore more stable regions of parameter space, which often correspond to better generalization. This regularization effect can help prevent overfitting.
4. **Memory of Past Information**: EMA-based optimizers remember past parameter values. This benefits generalization, especially on large, noisy datasets, because the model can employ useful information from earlier stages of training even as it adapts to current data.
5. **Improved Exploration-Exploitation Balance**: EMA-based optimizers help strike a better balance between exploration (trying out new weight configurations) and exploitation (refining the current best-known configuration). This can improve generalization by ensuring that the model explores a wider range of solutions before settling on a final one.

These benefits make EMA-based optimizers a valuable tool for improving the ability of deep neural networks to generalize well to unseen data. However, EMA-based optimizers also have several limitations. Existing adaptive optimization methods often struggle to quickly adapt to changing data trends during optimization. **To overcome this, our financial-inspired approach integrates a robust technical indicator into adaptive methods. Our method is used as an active indicator, affecting the optimized weights of the network directly, in contrast to the passive role that a technical indicator performs in finance.**

## Network Regularization

The dependence between regularization and deep model generalization is a fundamental concept in deep learning. Regularization techniques are employed to improve a model's generalization performance, which refers to its ability to make accurate predictions or classifications without depending on specific neurons. This improvement is thanks to the following:

1. **Overfitting Prevention**: Regularization techniques, such as L1 and L2 regularization, dropout, and early stopping, are designed to prevent overfitting. Regularization introduces constraints or penalties on the model's parameters, discouraging it from becoming overly complex. Hence, regularization helps the model generalize better because the model focuses on the relevant patterns rather than the noise.
2. **Smoother Decision Boundaries**: Regularization methods encourage the model to learn smoother decision boundaries. Instead of fitting the training data points exactly, the model learns to make predictions based on broader patterns and trends in the data. This results in more stable and generalizable predictions on new, unseen data.
3. **Weight Pruning**: Techniques like L1 regularization encourage sparsity in model parameters by driving some weights to zero. This process effectively prunes unnecessary connections, reducing the model’s capacity to memorize training data and improving generalization.
4. **Dropout as a Form of Noise Injection**: Dropout randomly deactivates a fraction of the neurons during training. This introduces noise and variability into the training process, making a model more robust and less prone to overfitting. Dropout improves generalization by preventing a model from relying too heavily on any single neuron or feature.

By striking a balance between fitting the training data and capturing underlying patterns, regularization helps deep models perform well on new, unseen data, improving robustness and reliability in practice. To improve dropout performance, researchers have mainly explored adaptive approaches based on prior network knowledge. For instance, [15] uses SGD to learn a strength parameter to guide the dropout regularization of each node, and [16] normalizes the dropout probability at each layer and training batch to maintain a consistently effective drop rate.

Most existing dropout methods remove individual activations within each unit independently, either with a fixed or adaptive probability. Clearly, visual structures in input images activate the corresponding regions in convolution feature maps. This suggests that in adaptive dropout, feature maps with similar activation patterns should be stochastically dropped to reduce co-adaptations between neurons. Nevertheless, these feature maps also encode information about intra-class variation in latent semantic features. This intriguing insight leads us to explore adaptive dropout approaches. The authors in [17] proposed group-wise dropout, a method that adapts to latent semantic variations while simulating dynamic sparseness in the network, ultimately improving object recognition performance. In [18], spatial dropout was proposed, in which a random subset of feature-map activations are dropped independently to reduce spatial correlations. Poernomo and Kang introduced cross-map dropout, which simultaneously drops or retains elements at the same coordinates on different feature maps [19]. Zhang et al. developed region dropout by considering salient regions with fixed size and relative positions for training [20]. However, these salient regions are not always available in general object-recognition problems. Wang and Manning demonstrated a Gaussian approximation to adaptive dropout [21], whereas Kingma et al. proposed variational dropout, which connects global uncertainty with dropout rates to optimize a generalized Gaussian dropout [22]. In [23], dropout training was introduced as a form of adaptive regularization, whereas [24] updated the elementwise probability for mask matrix generation based on activation output. Reference [25] extends the overlaid model to learn adaptive dropout rates for different neurons or groups of neurons. **Nonetheless, the primary hurdle faced by these methods stems from their departure from randomness. The work on adaptive knowledge-based dropout strives to thwart co-adaptations between neurons through various strategies. However, if these efforts fall short, a significant chance of overfitting exists because of the absence of randomness.**

# Research Objectives and Expected Significance

**This research aims to comprehensively study the generalization abilities of deep learning architectures. Its primary objective is to thoroughly understand the underlying factors driving generalization in deep neural networks, enabling high accuracy in simple models (comparable to that of complex models) but without the risk of overfitting common in complex models.** To achieve this goal, we will delve into the practical implications of three core elements: data normalization, optimization, and regularization. Through systematic analysis, we aim to unearth insights that enhance our comprehension of deep network generalization.

The expected significance of this work is multifaceted and extends across various domains, with the potential to bring about transformative advancements in the field of machine learning and AI. An expanded perspective on the key points is as follows:

1. **Improved Model Performance**

* Enhanced Accuracy: By pioneering adaptive strategies for data normalization, optimization, and regularization, this research aims to elevate the performance of deep learning models. Through more generalizable methods, models can increase accuracy, reducing errors and misclassifications.
* Enhanced Reliability: Improved model performance is not just about accuracy but also reliability. Models will become more consistent in their predictions, leading to greater trust and usability in real-world applications.

1. **Enhanced Generalization**

* Better Adaptation to Unseen Data: The insights derived from this research have the potential to revolutionize the way models generalize to data they have not encountered during training. Models will be more adaptable, making them highly proficient in handling diverse datasets and real-world scenarios.
* Cross-Domain Applicability: Generalization improvements will render models versatile, capable of seamless adaptation to various domains, ensuring their applicability in a wide array of industries and problem spaces.

1. **Expansion of Knowledge about Generalization**

* Novel Concepts: This study introduces groundbreaking concepts that expand the understanding of deep model generalization. It delves into the intricacies of how models learn and generalize, shedding light on previously unexplored facets.
* Implications and Significance: By unraveling the implications and significance of our models, this research will help demystify the 'black box' nature of deep learning, enabling us to make more informed decisions.
* Underlying Factors: Understanding the underlying factors that contribute to generalization provides valuable insights into model behavior. This knowledge can lead to the development of more robust, interpretable, and controllable models.

1. **Real-world Applications**

* The potential for real-world applications is vast, spanning various fields. In computer vision, models could recognize objects and patterns with unmatched accuracy. In natural language processing, they could become better at understanding context and generating human-like text. In audio analysis, models could distinguish sounds and speech with remarkable precision, benefiting industries such as security and entertainment. Many further applications are yet to be fully explored, making this research a pioneering force in the advancement of AI.

In summary, the expected significance of this work is profound, with the potential to push the boundaries of machine learning, impacting everything from model performance and generalization to the expansion of our knowledge in this field and its widespread real-world applications. This research has the power to reshape how we approach and utilize AI, unlocking new possibilities for innovation and problem-solving.

The specific research goals are as follows:

* ***AIM 1: Investigating Adaptive Strategy for Hybrid Data Normalization***

Our first objective revolves around exploring the potential of **hybrid adaptive data normalization techniques integrating knowledge-based and data-driven approaches**.

Diverging from the common fixed-grouping GN approach, our method harnesses the concept of channel similarity to **dynamically re-order the data**, thereby elevating both the adaptability and generalizability of the normalization process. Moreover, our framework empowers flexibility in group sizes (in contrast to the original GN), enabling better tailoring of normalization to specific data characteristics. **Through this research, we aim to study the impact, strengths, and potential limitations of this hybrid model on data normalization procedures.** **Additionally, we will explore the influence of our adaptive normalization on the generalization capabilities of the learning model.**

* ***AIM 2: Exploring High-Order EMA Optimizers***

Building on the success of our recent work introducing the Triple EMA (TEMA) optimizer [26], we strive to push the boundaries further by **exploring higher-order EMA-based optimization techniques to enhance the performance and stability of deep optimization**. This endeavor entails a thorough examination of higher-order EMAs and their potential effect on the overall performance of deep learning models, with a particular emphasis on their effects on model generalizability and accuracy. Our study encompasses an evaluation of the stability of these higher-order frameworks across a diverse spectrum of datasets, architectural configurations, and training scenarios. **Key research questions include whether higher-order EMAs improve generalization because of their enhanced ability to accurately identify and track gradient trends with minimal lagging, whether there exists an optimal order beyond which generalization diminishes, and whether adaptation of the optimizer order to network characteristics (such as depth) affect generalizability.**

* ***AIM 3: Studying Multi-level Dropout for Enhanced Regularization***

Conventional dropout regularization techniques and their adaptations are routinely employed to handle feature co-adaptation within individual hidden units. Inspired by concepts like group dropout, we propose a novel approach for multi-level dropout regularization. This technique aggregates neurons into distinct groups, guided by spatial and semantic correlations among neighboring neurons, which is at knowledge-based level. We then introduce a randomness-based level by selecting a random neuron from each group. In this way, we enjoy both worlds—we improve dropout by incorporating knowledge into the procedure while mitigating overfitting by randomly selecting neurons from each group. **Key research questions include whether our multi-level regularization outperforms traditional complete random dropout and whether it enhances model generalizability.**

By addressing these issues, we aim to contribute significantly to the advancement of deep learning, ultimately empowering the development of more reliable, generalizable, and adaptable deep learning models with broad applicability across various domains. To en our proposed models, most tested deep architectures have a number of parameters that is much higher than the number of training examples.

# Detailed Research Plan

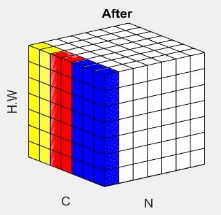
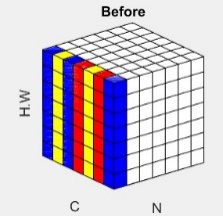
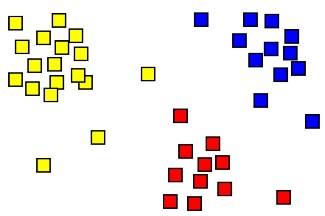
* ***AIM 1: Investigating Adaptive Strategy for Hybrid Data Normalization***

Our research introduces a novel hybrid normalization strategy that integrates domain knowledge into data-driven learning. Hybrid models are extensively employed across various domains. **However, our innovation lies in introducing a hybrid model for data normalization. Furthermore, our model employs a distinctive approach featuring a non-trivial blend of online and offline phases.** Therefore, our approach marks a significant leap in deep-learning normalization techniques. Our method is based on conventional GN, but instead of relying on fixed channel-grouping, we **leverage channel similarity to** **dynamically re-organize the channel data**. This dynamic organization both augments the efficiency of the normalization process and enhances its adaptability to diverse datasets, increasing generalizability.

**Preliminary work**

Our proposed Similarity-based GN (SGN) employs **online and offline phases**. The whole channel re-clustering occurs offline, while the core of the learning procedure is performed online.

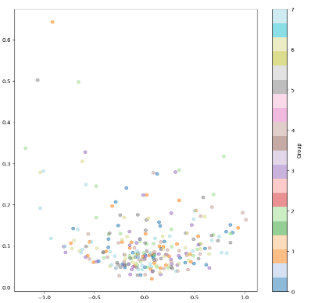
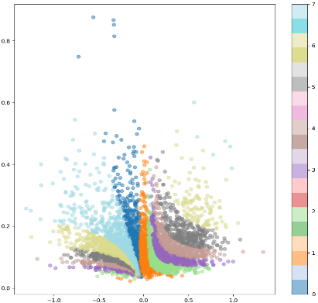
* **Offline K-means clustering and outlier exclusion:** We calculate the mean and standard deviation for each deep channel. Then, the isolation forest algorithm is used to exclude channels that are outliers (using feature means and variances), followed by K-means clustering on the inlier channels. Figure *1* illustrates channel clustering in SGN, and Figure 2 presents its effects.
* **Online integration of the re-ordered channels into learning:** After establishing the ***offline*** *channel clusters*, all channels are grouped into individual clusters and re-integrated into the deep network. Training then continues until the next re-clustering procedure.



**Figure 1:** Proposed similarity-based group normalization approach

Two grouping strategies are explored: *SGN-V1* employs a single mini-batch re-clustering in which the channels of different images in the whole minibatch are clustered. This offers flexibility by clustering channels from different images, but requires identical batch sizes in training and inference. *SGN-V2* uses single-image re-clustering, in which K-means is performed on each image, enabling testing with various batch sizes and making it suitable for single-image applications, albeit at the potential cost of optimal channel group statistical alignment. Our experiments suggest SGN-V2 slightly underperforms SGN-V1 in downstream tasks.

We re-cluster every 10th epoch because of evolving weights, which affect image features differently as training progresses. This adaptation prevents overfitting, which can occur if we maintain the static groupings based on initial training statistics.

****תמונה שמכילה צילום מסך, צבעוני

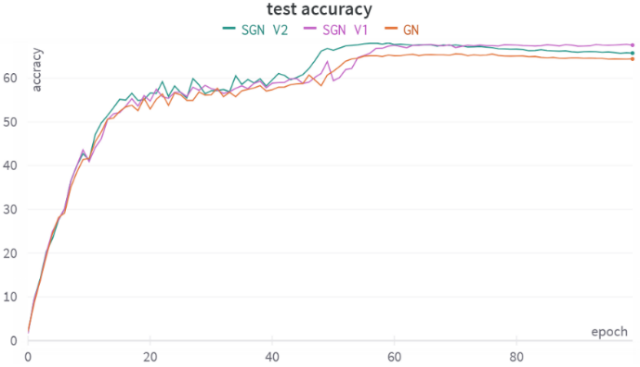
התיאור נוצר באופן אוטומטי

1. (b) (c)

Figure 2: Each point indicates the mean (x-axis) and variance (y-axis) of individual channels. (a) Original GN: channel groups are highly mixed. (b) SGN-V1. (c) SGN-V2.

**Preliminary results**

To evaluate our technique, we first tested it on the CIFAR-100 benchmark. We employed the common ResNet50 architecture, which incorporates convolutional layers and pooling layers to mitigate overfitting. The training used 100 epochs for all models and a learning rate of 0.001. Mini-batches of 8–16 images were used. Figure 3 presents the experimental results for three normalization methods: GN, SGN-V1, and SGN-V2. SGN-V1 achieves a substantial increase in accuracy (3.15\%) with respect to GN. Furthermore, SGN-V2 obtains a further improvement of 1.32\% over SGN-V1. These findings underscore the advantages of SGN over conventional GN.



Maybe to replace that with a graph without V2

Figure 3: Comparison of test accuracy with respect to number of training epochs obtained using SGN-V1, SGN-V2, and GN on the CIFAR-100 dataset using the ResNet50 architecture. This figure provides a clear indication of the performance dynamics of each method during training.

**Ongoing and Future Work**

In our ongoing SGN analysis research, we plan to further introduce two key innovations:

1. Variable Grouping: Unlike traditional methods with fixed group sizes, we leverage K-means to create diverse group sizes for data channels. This approach enables a more comprehensive exploration of channel grouping's impact and significance, offering valuable insights into our data dynamics by accommodating varying group sizes.
2. Dynamic Standard Deviation Recalculation: Prior to each re-clustering iteration, we calculate the standard deviation of the mean values within each group of channels. Groups with standard deviations below a predefined threshold are automatically excluded from subsequent channel re-clustering, which focuses our efforts on groups with standard deviations higher than the threshold. This strategic approach optimizes computational resources, enhancing re-clustering efficiency and reducing computational overhead.

These enhancements will bolster the generalizability and efficiency of our SGN analysis, enabling more meaningful patterns and insights to be extracted from our data. To assess the strengths and limitations of our approach, we will conduct comprehensive evaluations by integrating it into various deep neural networks. We will rigorously analyze its performance across a range of publicly available benchmarks spanning both computer vision and natural language processing domains. Evaluation metrics will be tailored to specific downstream tasks, which may include but are not limited to object detection, image classification, sentiment analysis, and language translation.

Following the methodology evaluations, our research endeavor raises intriguing questions:

1. Domain Knowledge and Normalization: One of the core elements we are exploring is the infusion of domain knowledge into the normalization process. We are driven to understand how this infusion impacts the efficiency and adaptability of the process. Does the contextual wisdom brought by domain knowledge lead to more effective and context-aware normalization, potentially improving model generalization?
2. Adaptive Normalization and Data Imbalance: Our venture into adaptive normalization, marked by dynamic data re-organization and adjustable group sizes, opens the door to addressing common challenges tied to data imbalance and heterogeneous datasets. Can this approach address these challenges, thus leveling the field for machine learning algorithms?
3. Trade-offs in Normalization Strategies: As we tread into this uncharted territory, we must also consider the trade-offs involved. Are there scenarios where traditional normalization techniques, such as group normalization, still outperform our hybrid approach? In understanding these nuances, we aim to refine our strategy for the most optimal results.
4. Measuring Improved Generalization: The ultimate goal of our research is to enhance the generalization capabilities of deep learning models. The question then arises: Does the adaptability introduced by our novel normalization strategy indeed lead to better generalization, and if so, to what extent? We are keen to quantify and measure this improvement across a wide spectrum of tasks and domains, revealing the path towards more robust and versatile AI solutions.

**Pitfalls and Alternative Approaches**

We minimize the risks for pitfalls by presenting very promising preliminary results. Obviously, testing our proposed method on a larger and more diverse cohort may present new challenges. In the event that the similarity-based method will not deliver satisfactory results, we will explore a random-based re-grouping approach, which we have already begun to explore. In contrast to our similarity-based approach presented above, the channels will be re-grouped randomly. The motivation for this is our assumption that generating a "mess" in the channels order every *nth* epoch will enforce the learning architecture to be more robust and strengthen its abilities.

* ***AIM2 – Exploring High-Order EMA Optimizers***

We hold a strong conviction that the implementation of a higher-order *EMA* optimization, capable of swiftly adapting to gradient variations and trends while minimizing lag, will substantially enhance model generalization. This, in turn, will diminish our reliance on diverse learning architectures and downstream tasks.

**Preliminary work**

*EMA* is widely-used in deep learning optimization. It applies exponentially decreasing weighting factors to past data, smoothing out short-term fluctuations; thus, it is used to denoise data during optimization. Let be a data sequence up to time ; *EMA* recursively combines the current and previous data values by

Here, is a tuning parameter that determines its responsiveness to data changes, where lower values provide faster response but less effective denoising.

*Double EMA (DEMA)* and *Triple EMA (TEMA)* are extensions of *EMA* used in finance for trend assessment [27]. They incorporate lag-correcting terms to enhance noise reduction and trend identification while maintaining smoothness.

In our recent paper [26], we introduced a novel optimizer based on *TEMA*. It is designed to accurately estimate the first and second moments of the gradient to estimate the true gradient at each time step with less bias than adaptive EMA-based methods like Adam. This is achieved by *TEMA* 's ability to reduce lag in the gradient-moment estimation while still performing effective denoising (Figure 4). The main contributions of this paper are as follows:

* **High-order EMAs**: To our best knowledge, we were the first to exploit the strong potential of high-order EMAs to overcome the limitations of first-order EMA such as lag and insufficient adaptation to data trends. In [26], only third-order EMAs were explored. Our current aim is to investigate a series of high-order EMAs
* **Active Optimization Guidance**: Our *TEMA* optimizer is a finance-inspired technical indicator. However, in finance, such indicators are used as passive indicators, merely analyzing market changes *without* directly affecting them. By contrast, our *TEMA* optimizer *actively* guides the optimization process and affects network weights, extending its usefulness in high-dimensional optimization problems.

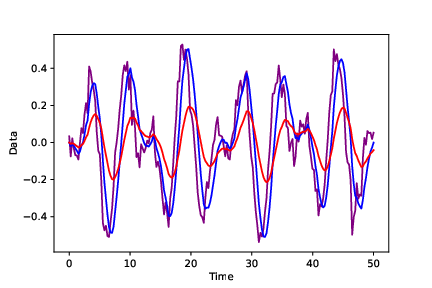
******

Figure 4: Simulated demonstration of gradient trend estimation and lagging. Ground truth (purple), TEMA-based estimation (blue), and EMA-based estimation (red).

***Preliminary results***

We have compelling evidence that the *TEMA*-based optimizer outperforms conventional EMA-based optimizers in terms of stability, robustness, generalizability, and accuracy. This finding represents a significant validation of the concepts underpinning the current grant. Our rigorous evaluation of the TEMA-based optimizer encompasses a diverse set of domains and tasks. We conducted tests on six datasets (CIFAR-10, CIFAR-100, PASCAL-VOC, MS-COCO, Cityscapes and ImageNet) using 15 architectural configurations and various computer vision tasks such as object detection, image classification, and semantic understanding. We compared our optimizer with SGD with momentum, Adam, AdamW, AdaHessian, AdaBound, and AdaGrad, and the results were averaged over two different weight initializations. The results are clear—our optimizer consistently outperforms the competitors, establishing it as a state-of-the-art optimizer for neural network models across a wide range of domains and applications.

On CIFAR-10 and CIFAR-100, our TEMA-based optimizer consistently outperformed Adam and SGD in most architectures. It also achieved a higher accuracy than the remaining methods. Across the CIFAR-10/100 experiments, our optimizer increased average classification accuracy by up to 16.33% and yielded smoother test-accuracy curves than standard EMA (Figure 5), thanks to its lower bias. On ImageNet, our optimizer obtained average improvements of 1.9% over Adam, SGD, and AdamW. On Pascal-VOC, our optimizer excelled in YOLOv5-s and YOLOv5-m, improving mAP@0.5 by 6.56% and 7.76%, respectively. On MS-COCO, our optimizer significantly outperformed SGD, Adam, and AdamW, with average improvements of 16.9% for mAP@0.5, 16.1% for Precision, 15.7% for Recall, and 15.16% for F1-score. On the Cityscapes benchmark, our optimizer enhanced the mean IoU results by 5.3%, 1.4%, and 1.9% compared with SGD, Adam, and AdamW, respectively.

**The comprehensive evaluation underscores our optimizer's generalizability and showcases its ability to provide less noisy gradient predictions. Importantly, the accuracy achieved by our simple network is on par with, if not superior to, the accuracy attained by other optimizers using more complex models from the same family (e.g., ResNet or DenseNet).**

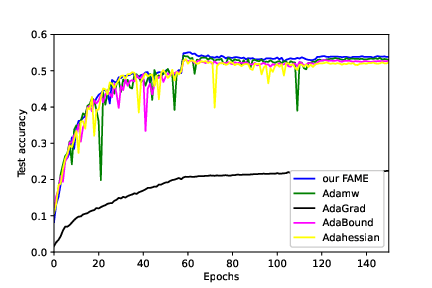
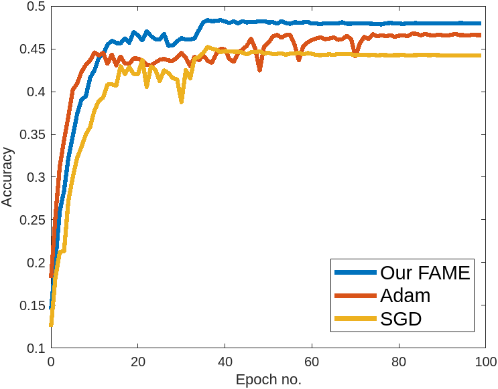
1. **** (b)

Figure 5: Performance stability of (a) our optimizer, Adam, and SGD on CIFAR-100 using a Reversible ViT and (b) our optimizer vs. other optimizers on CIFAR-100 using an EfficientNet-b3 (trained from scratch).

**Ongoing and Future Work**

The work of this grant proposal involves a thorough investigation of higher-order EMAs (above third order) and their potential impact on the overall performance of deep learning models, with a specific emphasis on their influence on model generalization and accuracy. Our study encompasses a comprehensive assessment of the stability of these higher-order frameworks across diverse datasets, architectural configurations, and training scenarios. We are particularly interested in exploring several intriguing research questions:

* Do higher-order EMAs exhibit improved generalization abilities owing to their enhanced capacity to accurately identify and track gradient trends?
* Is there a specific order beyond which the generalization performance starts to diminish, providing insights into the optimal order for practical use?
* Should the selection of the order be tailored to the characteristics of the network, such as depth, to maintain consistent generalizability across various network architectures?

**Through this research, we aim to not only expand EMA-based optimization but also provide insights into the selection the appropriate order of the EMA, ultimately advancing our understanding of optimization techniques and their impact on model generalization.**

While the simple in (1) denoises the data effectively, it also introduces lag in the estimation, which may result in inadequate gradient updates, and consequently suboptimal performance. denotes the application of EMA on data sequence . To reduce lag, the and add lag correcting terms to (see [26] for details).

In this research, we will introduce and explore the generalization of *EMA* to general order *EMA*, called . Here, denotes recursive applications of *EMA* on . We define

(2)

(3)

Let , and for *k=1,2,3*…

(4)

can be seen as a -order lag correction operator. The calculation below shows that

(5)

Higher-order *KEMAs* sacrifice smoothness in exchange for more aggressive lag reduction. Thus, in practice, the order should be selected to balance reducing the lag and denoising the gradients. The appropriate order highly depends on the problem at hand. Evidently, a larger *k* entails larger coefficients in the expansion for , which may increase sensitivity to hyperparameter tuning. This affects the choice of order. Because the *EMA* operator is linear, we may w.l.g. expand and as the sums

By definition

(9)

Defining for convenience and plugging in the last sum into (3) gives

Matching the coefficients of in (11) to that in (7) gives the recursive equation,

Recalling that by definition and , the solution to (12) is as in (5), since

To show the expansion of in (6), we have

Matching the coefficients of to that in (10), we obtain the recursive equation,

Since and by definition , we obtain the solution in (6) since

Using the developed equations, we will study the performance of high-order optimizers while carefully considering their stability. We expect high-order EMAs to be a bit unstable because the scale of increases as the order increases*.* We will study the generalization of high-order *EMAs* and other *EMA-*based methods in terms of architecture and downstream tasks.

**Pitfalls and Alternative Approach**

We have introduced promising preliminary results. Scaling up to a larger and more diverse dataset may introduce complexities; however, we stand well-prepared to tackle them directly. Should our proposed method fall short of expectations, we have contingency plans. These include the possibility of introducing a new family of advanced high-order equations (we outlined a specific family in this proposal) tailored to better match the specific data under analysis. Moreover, we note that our proposal aims to study the effects of higher-order optimizers, while our findings already demonstrate that *TEMA* outperforms the original EMA. If additional higher-orders optimizers do not further enhance TEMA's performance, this would be a legitimate conclusion, addressing a key research question regarding the optimal order that can be applied.

* ***AIM3 - Studying Multi-level Dropout for Enhanced Regularization***

Expanding upon the notion of adaptive dropout, our research introduces an innovative adaptive model that leverages both knowledge and randomness to accommodate varying data characteristics. **This research aims to shed light on the nuanced interplay between regularization techniques and model generalization, exploring the inherent strengths of knowledge in contrast to full randomness or their combination.**

We employ a hybrid framework that combines Convolutional Neural Networks (CNNs) with the Neural Additive Model (NAM) [28] to better distill the contribution of each individual neuron. Our current architecture utilizes ResNet-34 as the CNN structure. We seamlessly integrate the NAM model in the middle layer of ResNet, which serves a dual purpose:

1. Individual Neuron Analysis: NAM helps dissect individual neuron contributions, offering deeper insights into model behavior.

2. Parameter Efficiency: NAMs also reduce the number of network parameters, making computations more efficient for quicker training and inference.

After applying the CNN-NAM approach, we will employ a multi-level analysis: 1) a **similarity-based** clustering according to the significance of individual neurons for the given task and their prominence within a particular feature map, and 2) a **random** selection of a predetermined percentage of neurons from each cluster to undergo dropout in each training epoch. These chosen neurons (from each group) will be shut down.

**Our first hybrid knowledge-random dropout approach ensures that dropout is not fully random but is also based on knowledge of the data.** We will explore the following aspects:

* Knowledge-Enhanced Generalization: To what extent does the knowledge-driven grouping and dropout strategy enrich the model's performance by preserving crucial feature representations? Does this approach lead to more consistent and superior generalization than conventional random dropout, and if so, in what scenarios or network architectures is this effect most pronounced?
* Stability and Overfitting Mitigation: Does the randomness carefully introduced within the adaptive group dropout framework effectively mitigate overfitting to the training set? How does this approach balance model stability and the risk of increased overfitting? Can it enhance the model's resistance to noise and perturbations in real-world data?
* Optimal Grouping Strategies for higher Generalization: What are the key factors influencing the success of different neuron grouping strategies? Are there specific spatial or semantic relationships among neurons that are particularly conducive to robust generalization? How do these strategies adapt to varying dataset characteristics and network complexities?
* Transferability across domains and tasks: To what degree can the benefits of adaptive group dropout be transferred across different domains, datasets, and tasks? Does this technique offer a transferable improvement in the generalization ability of deep learning models, and are there limitations to its applicability in specific contexts?

**By addressing these research questions, we aim to both advance the understanding of adaptive group dropout and provide actionable insights into its role as a regularization tool for enhancing model generalization.** We aspire to unravel the intricate interplay between knowledge-driven dropout strategies, model stability, overfitting mitigation, and the broader context of deep learning, contributing to the evolution of techniques that empower deep neural networks to generalize effectively across diverse datasets and real-world scenarios.

**Preliminary results**

Preliminary findings reveal a noteworthy 1.5% enhancement in classification accuracy on CIFAR-100 using ResNet34. These results also indicate a substantial 92% reduction in both the number of parameters and computational load within the learning architecture. To put this into perspective, a standalone ResNet-34 configuration typically requires 21.8M parameters, but integrating it with NAM results in a leaner 1.8M parameter architecture. These preliminary findings pertain to the application of the combined CNN-NAM network and were achieved without employing the multi-level approach *yet*, which involves grouping and random choice. Based on the robust performance of our similarity-based normalization techniques and findings in [17], which unequivocally demonstrate the positive impact of similarity-based knowledge, we hold the strong conviction that our proposed suggestions have the potential to further enhance performance. These aspects will be explored in this grant.

**Pitfalls and alternative approaches**

**Bibliography**

1. Dou, Q., Coelho de Castro, D., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems*, *32*.‏
2. Li, Y., Gong, M., Tian, X., Liu, T., & Tao, D. (2018, April). Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).‏
3. Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 5715-5725).‏
4. Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning* (pp. 10-18). PMLR.‏
5. D’Innocente, A., & Caputo, B. (2019). Domain generalization with domain-specific aggregation modules. In *Pattern Recognition: 40th German Conference, GCPR 2018, Proceedings 40* (pp. 187-198). Springer International Publishing.‏
6. Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 5542-5550).‏
7. Matsuura, T., & Harada, T. (2020). Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11749-11756).‏
8. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.‏
9. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.‏
10. Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.‏
11. Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, *29*.‏
12. Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).‏
13. Keskar, N. S., & Socher, R. (2017). Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*.‏
14. Landro, N., Gallo, I., & La Grassa, R. (2020). Mixing Adam and SGD: a combined optimization method. *arXiv preprint arXiv:2011.08042*.‏
15. Keshari, R., Singh, R., & Vatsa, M. (2019). Guided dropout. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 4065-4072).‏
16. Wang, S., Zhou, T., & Bilmes, J. (2019, May). Jumpout: improved dropout for deep neural networks with ReLUs. In *International conference on machine learning* (pp. 6668-6676). PMLR.‏
17. Ke, Z., Wen, Z., Xie, W., Wang, Y., & Shen, L. (2020). Group-wise dynamic dropout based on latent semantic variations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11229-11236).‏
18. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 648-656).‏
19. Poernomo, A., & Kang, D. K. (2018). Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural networks*, *104*, 60-67.‏
20. Choe, J., Han, D., Yun, S., Ha, J. W., Oh, S. J., & Shim, H. (2021). Region-based dropout with attention prior for weakly supervised object localization. *Pattern Recognition*, *116*, 107949.‏
21. Wang, S., & Manning, C. (2013). Fast dropout training. In *international conference on machine learning* (pp. 118-126). PMLR.
22. Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, *28*.‏
23. Wager, S., Wang, S., & Liang, P. S. (2013). Dropout training as adaptive regularization. *Advances in neural information processing systems*, *26*.‏
24. Ba, J., & Frey, B. (2013). Adaptive dropout for training deep neural networks. *Advances in neural information processing systems*, *26*.‏
25. Zhuo, J., Zhu, J., & Zhang, B. (2015, July). Adaptive Dropout Rates for Learning with Corrupted Features. In *IJCAI* (pp. 4126-4133).‏
26. Peleg, R., Weiss, R., & Hoogi, A. (2023). Leveraging the Triple Exponential Moving Average for Fast-Adaptive Moment Estimation. *arXiv preprint arXiv:2306.01423*.‏
27. Mulloy, P. G. (1994). Smoothing data with faster moving averages. Stocks & Commodities, 12(1):11–19.
28. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, *34*, 4699-4711.‏