Research Statement

Jang-Hyun Kim, PhD candidate, Computer Science Dept. at SNU

The dream goal of my research is to develop **sustainable AI systems that can learn and adapt efficiently over long periods**, much like humans do. Humans can learn from small amounts of data, forming long-term memories that evolve and improve over decades. In contrast, current neural network AI systems demand vast amounts of labeled data and immense computational resources to achieve human-level performance. Furthermore, most AI systems do not engage in inference based on long-term memory; instead, they rely on short-term context caching or periodic updates of model checkpoints. This data-hungry and memory-limited approach poses significant challenges to the sustainability of AI systems, especially regarding long-term interaction, computational costs, and labeling efforts.

To address these limitations, my research focuses on improving the efficiency of neural networks, both during training and inference, through a data-centric approach. Specifically, my work revolves around three key themes: (1) context and dataset compression, (2) synthetic training data generation, and (3) characterization of data using relational structures. These themes are integrated into **a data optimization framework that leverages feedback from the neural network itself**, thereby reducing the reliance on human supervision. Within this framework, I have introduced novel methods that enable memory-efficient inference and data-efficient training, pushing AI systems closer to the goal of sustainability. In the following sections, I will highlight the key contributions of each research theme and outline my ongoing projects, as well as future research directions.

Research Progress

1 Context/Dataset Compression

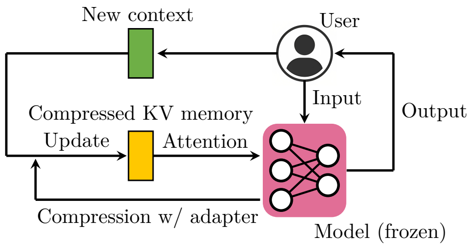
Humans process real-time information selectively, forming own compressed long-term memories. This capability allows for efficient learning and long-term adaptation to the environment. In contrast, current AI systems are inefficient in their data management. They require terabytes of storage for training data and tens of gigabytes for storing context features, as seen with Transformer models.

My primary research focuses on addressing these challenges by developing methods for compressing data and constructing long-term memory in neural networks [1, 2, 3]. I have proposed novel ideas within an integrated framework that leverages the capabilities of trained neural networks to compress information. These methods significantly enhance the efficiency of both inference and training, supporting the long-term usability of AI models in memory-constrained environments.

Compressed context memory for Transformers [1]. Transformer-based models, such as ChatGPT, excel at processing long-range dependencies in tokens, enabling coherent response generation based on extensive context. However, this capability comes with a significant challenge: the need to store key-value vectors for context tokens. While these vectors capture essential contextual information for response generation, their storage requirements can be prohibitive. For instance, in models like LLaMA-70B, key-value vectors for just 1,000 words demand tens of gigabytes of memory. As interactions progress, these storage demands grow rapidly, overwhelming the memory capacity of computing environments.

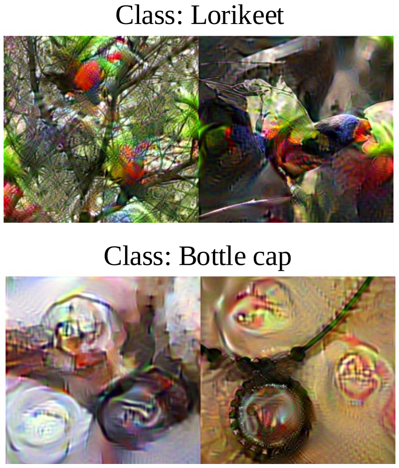
To address this challenge, I developed a novel online key-value compression method coupled with a compressed memory module for Transformer-based language models. My approach compresses newly introduced context into shorter key-value vectors, which are then updated in the compressed memory.

The language model accesses this compressed memory to generate contextually relevant responses. By leveraging the model's forward pass for compression, my approach seamlessly integrates the compression process into the model's inference pipeline. The results are promising: my method reduces the key-value cache size by 5×, while maintaining performance comparable to the original model without compression. This achievement opens up new possibilities for deploying Transformers in memory-constrained environments.



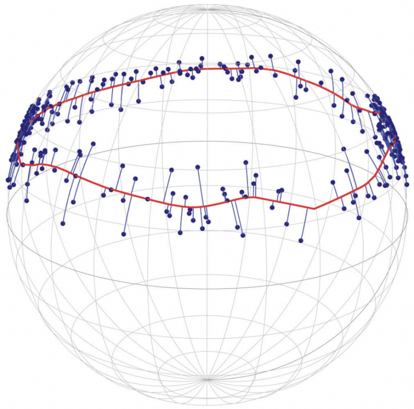
Compressed memory system for memory-efficient Transformer inference [1]. Colored boxes represent the attention key-value (KV) pairs.

Condensing training dataset [2]. Current neural networks require vast amounts of training data, raising the key question: *What is the actual dataset size required for effective training?* If similar performance can be achieved with a smaller dataset, it would dramatically reduce the computational costs and storage required for training. To address this question, I conducted research on optimizing small synthetic datasets that match the training performance of the original dataset. Using a novel differentiable data-parameterization technique, I have successfully compressed large-scale datasets such as ImageNet with high efficiency. Notably, my approach achieves 90% of the full dataset's training performance with only 1% of its storage size, significantly outperforming traditional data subset selection methods. Additionally, I demonstrated the effectiveness of my approach in continual learning scenarios by forming condensed example memories, further highlighting its practical utility.



Condensed ImageNet samples [2]. My approach constructs compressed form of training data information, enabling efficient training.

Spherical data dimension reduction [3, 4]. Skeleton structures, such as human keypoints or molecules, are often represented as Cartesian products of spherical surfaces. Capturing the temporal evolution of these structures is crucial in fields like human motion analysis and molecular dynamics. However, conventional representation learning methods struggle with the inherent non-Euclidean geometry of spheres, making this task challenging. To address this issue, I developed an algorithm that optimizes a curve on spherical surfaces by minimizing the sum of geodesic distances to the data points. This approach effectively captures the underlying structure of spherical data and models the temporal evolution by projecting data onto a curve. Building on this research, I aim to further explore dataset compression methods for temporal and structured data, broadening the scope and potential impact of my work.



Spherical principal curve [3] encodes the temporal evolution of spherical data through dimension reduction.

2 Synthetic Training Data Generation

High-quality labeled data is crucial for training high-performing AI models, but it is costly and scarce in certain domains. One promising research direction I have pursued is the use of AI models to synthesize training data. I established principles defining effective training data and developed model-driven data synthesis frameworks that create a self-reinforcing loop, enhancing overall training performance. This approach holds significant potential to drive sustained improvements in AI systems while reducing human labeling efforts.

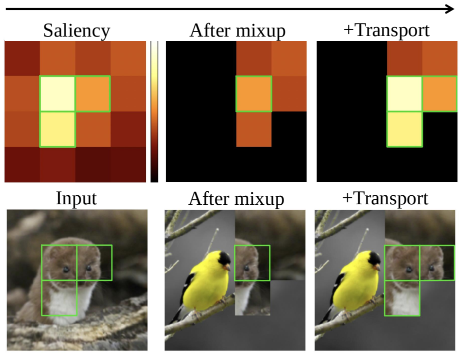
Saliency-guided data augmentation [5]. Conventional data augmentation methods often rely on predefined functions that randomly alter data without accounting for its individual characteristics, such as object locations in images. This results in the generation of data that are inconsistent with the assigned labels, providing false supervisory signals. To overcome this challenge, I developed a novel approach that leverages the data saliency maps obtained by the model under training. Using this saliency information, I designed an augmentation technique that mixes a data pair while preserving the most informative regions. This feedback-driven method not only improves the model's generalization performance but also guides the model more robust to input noise. The impact of my work extends beyond image processing, with successful applications in various domains including natural language, graph, and point cloud data.

Batch-level data augmentation [6]. In my previous work, I explored the challenge of generating informative data from given data pairs. Building on this, I posed a new question: *What combination of input data is optimal for augmentation?* This question not only opens a fresh research direction but also leads to a novel batch-level approach to data augmentation. Optimizing input combinations, however, presents an NP-hard problem due to the combinatorial complexity involved. Furthermore, if the objective is solely to maximize the saliency of each output, the result could be a lack of variety, producing identical outputs. To address this issue, I formalized the problem to optimize output diversity and developed an efficient combinatorial algorithm capable of synthesizing batch-level data within a few milliseconds. The synthesized data not only enhance neural network training performance but also yield notable improvements in uncertainty calibration for image/speech classifiers.

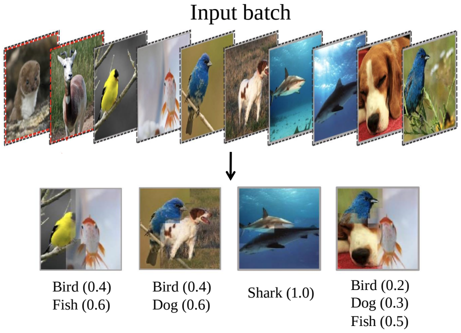
3 Characterization of Data using Relational Structure

I have frequently engaged with machine learning practitioners to understand their challenges. One surprising insight was the significant amount of time devoted to data preprocessing. The vast amount of data collected from diverse sources presents complex issues, including misalignment, outliers, and incorrect labels.

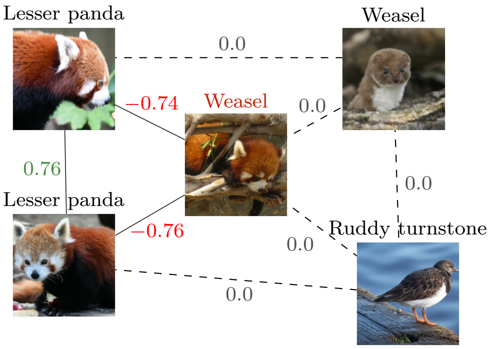
To tackle these critical challenges in real-world applications, I developed a framework to systematically identify complex data issues. Moving beyond conventional single-feature metrics like prediction error, I proposed a method that analyzes the relationships between data points. Utilizing pretrained models, I introduced *Neural Relation Graph* [7], a fully-connected graph of data points, where each edge weight quantifies the degree of complementarity or conflict between data points. This graph provides rich information of individual data points, which I leveraged to develop an algorithm that effectively identifies label errors as well as outliers. Additionally, I designed a visualization tool that enables users to intuitively explore these relationships in a 2D plot, thereby enhancing interpretability and usability of my method.



Puzzle Mix [5] generates synthetic training data by mixing salient regions from data pairs. I extract the saliency map from the model under training, creating a self-reinforcing loop in the training process.



Co-Mixup [6] generates a batch of synthetic data, jointly optimizing the diversity and saliency of outputs.



Neural relation graph [7] encodes relational structures within data using neural networks. By characterizing each data point within the graph, my algorithm effectively identifies outliers and mislabeled data.

Ongoing and Future Direction

In summary, my research has focused on optimizing data usage in AI systems to develop sustainable systems capable of efficient memory management during inference and effective learning from limited human-labeled data. I am eager to deepening and broadening my research to achieve tangible impacts in real-world applications with following research topics:

- *What is the most effective memory structure in neural networks?* I am interested in enabling neural networks to store new information with minimal capacity requirements. This approach is akin to identifying the minimum description length of information within the neural network's memory space (*i.e.*, model weights or data features). The objective is to determine the most concise memory structure, thereby allowing the efficient storage of vast amounts of information in neural network memory and enabling lifelong inference.
- Another major ambition of mine is to create *multi-modal memory systems*, drawing on my research experience across various domains, including image [2, 5, 6, 7], speech [2, 10], natural language [1, 7], and multi-joint dynamics [3, 9]. A particular focus is video data which presents substantial challenges due to its extensive size and information sparsity, requiring innovative compression techniques. My goal is to expand my research to develop AI systems capable of long-term video interactions, thereby enhancing the AI's utility in real-world applications.
- I am also interested in enabling AI to *identify causal relationships among data*. This capability deepens the AI's understanding about the environment, which is crucial for developing interactive AI models. In my recent work on *Targeted Cause Discovery* [8], I explored a data-driven learning approach for causal discovery. Moving forward, I aim to expand this research to refine neural networks' capability to accurately identify causal relationships within data, enhancing their memory and reasoning abilities.

References

[1] **Jang-Hyun Kim**, Junyoung Yeom, Sangdoo Yun, Hyun Oh Song. "Compressed Context Memory for Online Language Model Interaction". In *ICLR*. 2024.

[2] **Jang-Hyun Kim**, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, Hyun Oh Song. "Dataset Condensation via Efficient Synthetic-Data Parameterization". In *ICML*. 2022.

[3] Jongmin Lee*, **Jang-Hyun Kim***, Hee-Seok Oh (*: equal contribution). "Spherical Principal Curves". In *TPAMI*. 2021.

[4] Jongmin Lee, **Jang-Hyun Kim**, Hee-Seok Oh. "spherepc: An R Package for Dimension Reduction on a Sphere". In *R Journal*. 2022.

[5] **Jang-Hyun Kim**, Wonho Choo, Hosan Jeong, Hyun Oh Song. "Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup". In *ICML*. 2020.

[6] **Jang-Hyun Kim**, Wonho Choo, Hosan Jeong, Hyun Oh Song. "Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity". In *ICLR* (**oral presentation**). 2021.

[7] **Jang-Hyun Kim**, Sangdoo Yun, Hyun Oh Song. "Neural Relation Graph: A Unified Framework for Identifying Label Noise and Outlier Data". In *NeurIPS*. 2023.

[8] **Jang-Hyun Kim**, Claudia Skok Gibbs, Sangdoo Yun, Hyun Oh Song, Kyunghyun Cho. "Targeted Cause Discovery with Data-Driven Learning". *arXiv* preprint (under review). 2024.

[9] Gaon An*, Seungyong Moon*, **Jang-Hyun Kim**, Hyun Oh Song (*: equal contribution). "Uncertainty-Based Offline Reinforcement Learning with Diversified Q-ensemble". In *NeurIPS*. 2021.

[10] **Jang-Hyun Kim***, Jaejun Yoo*, Sanghyuk Chun, Adrian Kim, Jung-Woo Ha (*: equal contribution). "Multi-Domain Processing via Hybrid Denoising Networks for Speech Enhancement". *arXiv* preprint. 2018.