Research Statement

Jang-Hyun Kim

The core focus of my research is the pivotal role of data in shaping deep learning models. Data captures the complexities of the world in various forms and serves as the foundation for how deep learning models perceive and predict reality. These models have achieved remarkable success, driven by scaling laws that highlight the transformative power of vast amounts of data. My research explores a critical question at the frontier of deep learning: How can models continue to learn effectively in data-limited regimes? Conversely, how can models efficiently handle an infinite stream of newly introduced data? These questions have become increasingly urgent, as much of the internet's data has already been consumed for training. Furthermore, as we move toward a future where AI agents engage in dynamic interactions with humans, the ability to process evolving information streams becomes essential.

My research addresses these challenges by enabling deep learning systems to process data more efficiently and adaptively. From a model perspective, my work has advanced methods for continual inference within limited memory and has achieved effective learning with limited human-labeled data [1, 2, 5, 6]. From a data perspective, I have developed techniques to identify problematic data and uncover causality within vast amounts of observations [7, 8]. My research themes can be broadly categorized into three areas: (1) data/context compression, (2) synthetic data generation, and (3) data characterization. At the core of these themes is a feedback-driven approach, where trained models iteratively improve data processing in a self-reinforcing manner while minimizing human intervention. Below, I delve into each of these themes in detail and outline future directions for my research.

Research Progress

1 Data/Context Compression

Humans efficiently process streamed information, adapting to dynamic environments and acquiring new knowledge. Inspired by these cognitive abilities, my research focuses on developing machine learning systems that leverage compression to achieve more efficient information processing. My work encompasses both dataset compression, aimed at optimizing succinct datasets for neural network training [2], and representational compression, designed to enable sequence processing with linear complexity [1]. These compression processes contribute to constructing long-term memory in neural networks, allowing them to process streamed data effectively. Furthermore, compression facilitates learning by uncovering the underlying structure of data. Specifically, I have proposed statistical and data-driven compression methods that enhance modeling capabilities [1, 3].

Linear-complexity sequence processing [1]. Transformer-based models, such as ChatGPT, excel at processing long-range dependencies in sequences, enabling contextually coherent response generation. However, this capability relies on storing all key-value features for context tokens, which results in quadratic attention complexity and imposes significant memory demands. For instance, in LLaMA-70B, storing key-value features for 1,000 words requires tens of gigabytes of memory, a requirement that grows rapidly as contexts lengthen in interactive scenarios.

To address this challenge, I proposed a novel method that dynamically compresses key-value features into shorter representations, which are subsequently updated in a limited-size memory. The language model accesses this compressed memory to generate contextually coherent responses with reduced mem-

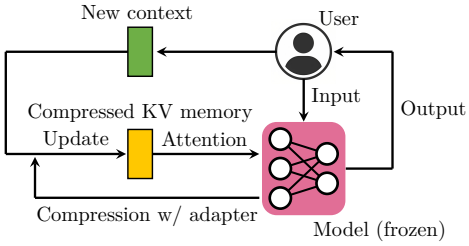
ory requirements and linear attention complexity. Notably, my approach seamlessly integrates the compression process into the model's inference pipeline by leveraging the model's forward pass for compression. The results are promising: my method reduces the key-value cache size by 5× while maintaining performance comparable to the original model without compression. This achievement significantly enhances the feasibility of deploying Transformer models in memory-constrained environments.

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, my approach demonstrated effectiveness in continual learning scenarios by forming condensed example memories, further highlighting its practical utility.

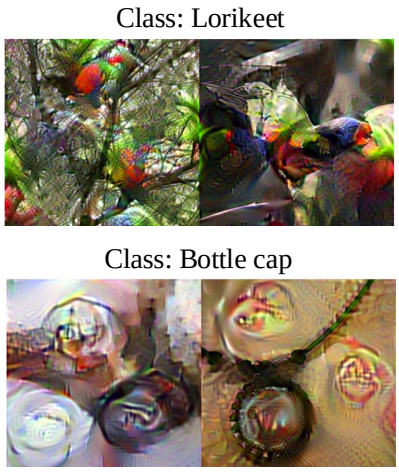
Spherical data dimension reduction [3, 4]. Skeleton structures, such as human keypoints or molecules, are often represented as Cartesian products of spherical surfaces. Capturing the temporal evolution of these structures is crucial in fields like human motion analysis and molecular dynamics. However, 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.

2 Synthetic Training Data Generation

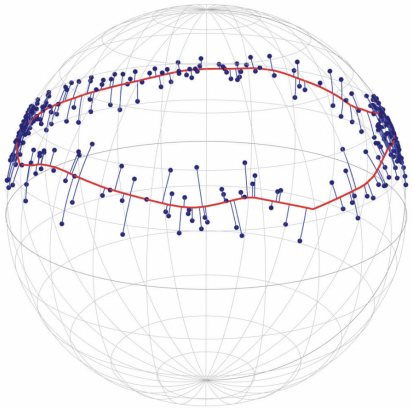
High-quality labeled data is crucial for training high-performing AI models, but it is costly and scarce in certain domains. One promising research direction I have pursued is the use of AI models to synthesize training data. I established principles for defining effective training data and developed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



Compressed context memory [1] enables continual inference of Transformers within a limited memory space.

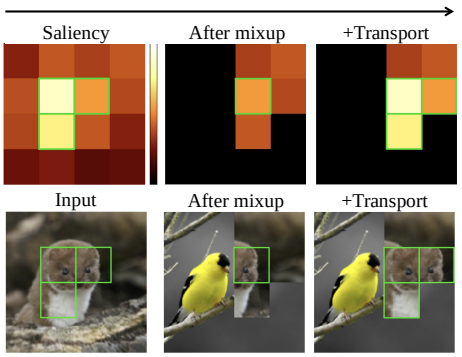


Condensed ImageNet samples [2]. My approach optimizes the compressed form of datasets, enabling efficient model training.



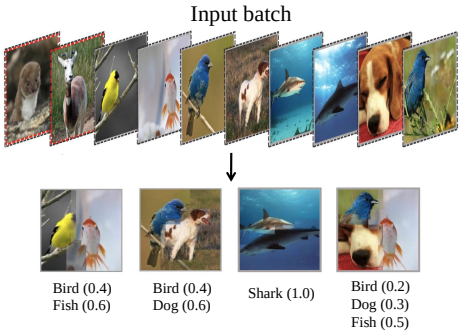
Spherical principal curve [3] encodes the 1D temporal evolution of spherical data.

Saliency-guided data augmentation [5]. Previous 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 is not informative and has inconsistencies with its 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 preserving informative parts. This feedback-driven method improves the model's generalization performance on test data from new domains. The impact of my work extends beyond image processing, with successful applications in various modalities including natural language, graph, and point cloud data.



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

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 jointly optimize output diversity and saliency, and developed an efficient combinatorial algorithm for synthesizing batch data. The resulting diverse and salient data not only enhance the model's generalization performance but also yield notable improvements in the uncertainty calibration of model predictions.

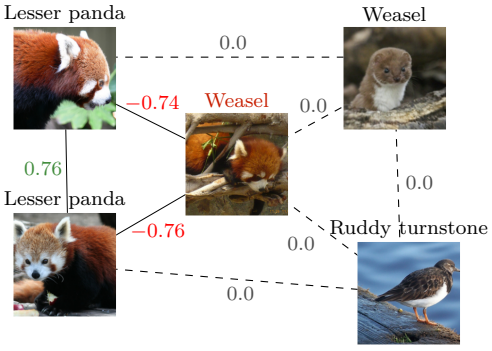


Co-Mixup [6] generates a batch of synthetic data, jointly optimizing the diversity and saliency of outputs.

3 Data Characterization

In my interactions with ML practitioners, I frequently observed that they devote a substantial portion of their efforts to data preprocessing. This stage is often complicated by the large volumes of data accompanied by complex issues such as label errors, outliers, and misalignment. Addressing these challenges is critical to ensuring robust model performance but requires substantial labor and cost.

To streamline this preprocessing stage and facilitate model development, I explored principled approach for data characterization using pretrained models. Moving beyond conventional single-feature metrics like prediction error, I proposed a method that analyzes relationships among data using 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 relational structure encodes rich information of data, enabling effective data characterization. Additionally, I designed a visualization tool that transforms these relationships into an intuitive 2D plot, enabling users to interactively explore and preprocess datasets.



Neural relation graph [7] encodes relational structures within data. 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 processing in AI systems to enable (1) efficient memory management during inference and (2) effective learning from limited human-labeled data. I am eager to deepen and broaden my research to achieve tangible impacts in real-world applications and sciences through the following research topics:

- I am interested in enabling neural networks to *efficiently process and store infinite streams of information*. To achieve this, I aim to examine the trade-off between the memory size allocated for context features and model performance. My focus is on various performance aspects, such as safety and bias, extending beyond factuality to explore the optimal memory structure. Furthermore, by analyzing recurrent memory-attention hybrid models, I seek to determine whether the memory architecture established during training plays a significant role or if the optimal memory structure and size are driven by the data itself.
- 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. The goal is to develop AI systems capable of long-term video interactions, thereby enhancing the AI's utility in promising areas such as robotic agents.
- I am also interested in empowering AI to uncover novel insights from data in scientific domains. In my recent work on *Targeted Cause Discovery* [8], I explored a data-driven learning approach for discovering causality among human genes, which led to the successful identification of novel causal relationships involving specific cancer-related genes. Moving forward, I am passionate about bridging AI with scientific fields to drive transformative, multi-disciplinary advancements.

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.