Cumulant Attention in Vision Transformers (Student Abstract)

Yuto Morimoto[1*], Zhipeng Wang[1*], Koji Yasuda[1*,2*]

[1]Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan

[2]Institute of Materials and Systems for Sustainability, Nagoya University

morimoto.yuto.b7@s.mail.nagoya-u.ac.jp, wangzp0611@gmail.com, yasudak@imass.nagoya-u.ac.jp

Abstract

Transformer models have achieved remarkable success across diverse deep learning fields, including natural language processing (NLP) and computer vision (CV). One drawback of these models is that the computational cost of the softmax attention, the core component of the transformer, exhibits quadratic complexity in both time and memory. As data scales up various attempts have been reported to overcome this bottleneck. The objective of this study is to propose a novel attention mechanism, "Cumulant Attention", that systematically balances efficiency and accuracy. This proposal introduces a statistical-mechanics perspective and a reliable approximation based on cumulant expansion into the attention layer. The low-order variant reduces computational complexity to linear order, similar to the linear attention, while keeping nonlinearity of the softmax attention. We evaluate several variants on CV tasks, including image classification with ViT on ImageNet-100 and video classification with ViViT on UCF-101. Experimental results demonstrate that the cumulant attention outperforms the linear attention and achieves accuracy comparable to the softmax attention. These findings validate the effectiveness of our approach and highlight future directions, including scaling to larger models, extending to other modalities, and optimizing implementations for GPU hardware.

Code — https://github.com/MorimotoYuto/Cumulant-Attention-in-ViT.git

Introduction

Transformer architecture has advanced the state of the art across various domains including NLP, CV, speech, graph learning, and multimodal learning. Its core mechanism, scaled dot-product attention, provides stable relevance modeling through softmax normalization but suffers from quadratic computational and memory complexity with respect to sequence length. This bottleneck becomes critical when dealing with large-scale data (Vaswani et al. 2017).

To address this, we design a novel attention mechanism inspired by statistical mechanics, aiming to reduce computational cost while preserving expressiveness. We show that attention can be interpreted as a "force" in statistical physics, enabling analysis and improvement through physical principles. The proposed method achieves superior accuracy on CV tasks.

Related Work

A representative approach for reducing computational cost is **linear attention** (Katharopoulos et al. 2020; Qin et al. 2022), which uses specially designed kernels to capture context efficiently. However, studies report shortcomings such as limited spatial selectivity in image recognition and weak emphasis on important local regions, reducing local feature extraction (Han et al. 2023; Fan et al. 2025; Fan, Huang, and He 2025).

Methodology

We show that attention can be understood as the free-energy gradient of a physical system, and we derive a systematic approximation by using the cumulant expansion technique. Let's consider a physical system: a random variable X uniformly drawn from set of tokens $\{x^l \mid l = 1, \dots, L\}$ as the magnetic moment of a particle, and a vector $p = d^{-1/2} W_K^T W_Q y$ as the (negative of) applied magnetic field. The statistical average $Z(p) = \langle \exp(p \cdot X) \rangle$ is called the partition function (under unit temperature). The derivative of (negative) Helmholtz free energy, $F(p) = \log Z(p)$ with respect to the external field p reproduces the attention as

$$\mathrm{attn}(q) = \frac{\sum_{j=1}^{L} \exp(q \cdot k^j)\, v^j}{\sum_{j=1}^{L} \exp(q \cdot k^j)} = L W_V \frac{\partial F(p)}{\partial p}$$

$$= \frac{1}{Z(p)} \sum_{l=1}^{L} W_V x^l \exp\left[\left(d^{-\frac{1}{2}} W_Q y\right) \cdot (W_K x^l)\right]$$

Cumulants are the numbers representing (higher-order) correlations among random variables, such as the covariance matrix. They appear as the coefficients of the power series.

$$F(p) = F(0) + \sum_i p_i \langle X_i \rangle + \sum_{ij} \frac{p_i p_j}{2} \big(\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \big) + \sum_{ijk} \frac{p_i p_j p_k}{3!} \langle X_i X_j X_k \rangle_c + \cdots$$

$$\begin{aligned} &\langle X_i X_j X_k \rangle_c \\ &= \langle X_i X_j X_k \rangle - \langle X_i X_j \rangle \langle X_k \rangle - \langle X_j X_k \rangle \langle X_i \rangle \\ &\quad - \langle X_k X_i \rangle \langle X_j \rangle + 2 \langle X_i \rangle \langle X_j \rangle \langle X_k \rangle \end{aligned}$$

The 1- and 2-cumulants are the mean and the covariance, respectively. The cross-cumulant among statistically independent variables vanishes. We have the cumulant (2nd) attention by keeping only 1- and 2-cumulants. The major correction to it comes from the diagonal part of 3-cumulant $\langle K_i K_j V \rangle_c$,

$$\mathrm{attn}_{2\mathrm{nd}} = L\bar{v} + \sum_{l=1}^{L} \big[q \cdot \big(k^l - \bar{k} \big) \big] \big(v^l - \bar{v} \big)$$

$$\mathrm{attn}_{3\mathrm{rd}} = \mathrm{attn}_{2\mathrm{nd}} + \frac{1}{2} \sum_{l=1}^{L} \sum_i q_i^2 \big(k_i^l - \bar{k}_i \big)^2 \big(v^l - \bar{v} \big)$$

where $\bar{k}$ and $\bar{v}$ are average key and value vectors, respectively. Due to the linked-cluster theorem our formula is free of denominator. Cumulants describe correlations among embedding dimensions, not tokens. Keeping up to second-order cumulants is equivalent to assuming random variables obey the multivariate normal, and the higher-order expansion provides systematic approximation to it.

Experiment

We evaluated Cumulant Attention in CV tasks. We employed the ViT (Dosovitskiy et al. 2021) architecture using the ImageNet-100 dataset for image classification. On the test data, the softmax attention achieved 69.1% accuracy, while the linear attention dropped to 43.0%. In contrast, the cumulant attention reached **63.6%** with the second-order and **64.0%** with the third-order expansion, improving over linear attention and approaching softmax. Both the second- and third-order variants consistently outperformed conventional linear attention and approached the performance of the softmax attention. We attribute this improvement to the selective capability of the cumulant attention, which is absent in linear attention. Indeed, visualization of the attention scores revealed more localized patterns (Figure 1).

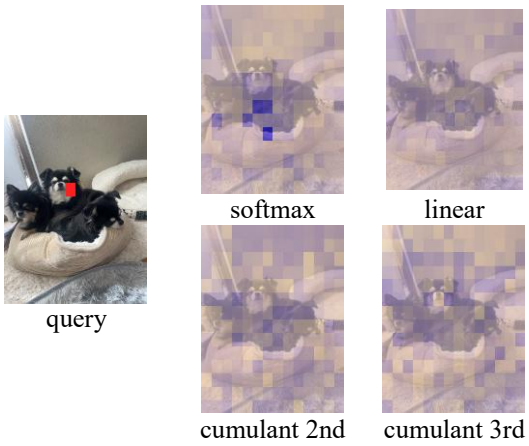

Figure 1: Visualization of attention scores. The cumulant attention scores showed more localized patterns than the linear one.

For video classification, we adopted the ViViT (Arnab et al. 2021) architecture with the UCF-101 dataset. Consistent with the image classification results, differences among attention types were observed, and the effectiveness of Cumulant Attention was confirmed. However, the model exhibited a tendency to overfit, due to smallness of the dataset.

Conclusion and Future Work

In this work, we proposed an efficient attention mechanism based on cumulants from statistical mechanics and validated it on CV tasks. Our method outperforms the linear attention and achieves accuracy close to that of softmax attention. Future directions include scaling to larger models and datasets, extending to graph and multimodal applications, and optimizing the design for efficient GPU acceleration.

References

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the 37th International Conference on Machine Learning. PMLR.

Qin, Z.; Sun, W.; Deng, H.; Li, D.; Wei, Y.; Lv, B.; Yan, J.; Kong, L.; and Zhong, Y. 2022. CosFormer: Rethinking Softmax in Attention. In Proceedings of the 10th International Conference on Learning Representations. OpenReview.net.

Fan, Q.; Huang, H.; and He, R. 2025. Breaking the Low-Rank Dilemma of Linear Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE/CVF.

Fan, Q.; Huang, H.; Ai, Y.; and He, R. 2025. Rectifying Magnitude Neglect in Linear Attention. arXiv preprint. arXiv:2507.00698.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations. OpenReview.net.

Han, D.; Pan, X.; Han, Y.; Song, S.; and Huang, G. 2023. FLatten Transformer: Vision Transformer using Focused Linear Attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF.