Multi-Scale Entropy for Transformers: Interpreting Training Dynamics and Guiding an Adaptive Training Pipeline

Jiahan Xie *

Department of Statistical Science, University College London, London, UK

* Corresponding Author Email: jiahan.xie.23@ucl.ac.uk

Abstract. The internal training dynamics of Transformer models remain poorly understood, limiting the development of both interpretable diagnostics and efficient training strategies. This paper proposes a unified multi-scale entropy framework that serves both as an analytical tool to interpret training dynamics and as a practical controller for adaptive optimization. It is demonstrated through a dual-track investigation. Analytically, it is used to dissect a BERT model's attention patterns on linguistic phenomena, revealing a clear layer-wise hierarchy of functional specialization. Second, the framework is operationalized to develop an entropy-guided adaptive training strategies and finally form a pipeline for Vision Transformers (ViT) which uses the stabilization of mid-layer attention entropy, a robust convergence signal identified through analysis, to trigger intelligent early stopping and staged optimization. Experiment shows the adaptive pipeline achieves performance comparable to a full training schedule while significantly reducing computational costs. Ultimately, this work establishes multi-scale entropy as a cohesive bridge between understanding how Transformers learn and controlling how they are trained, providing a versatile framework for enhancing both model interpretability and training efficiency.

Keywords: Transformer, deep Learning, information theory.

1. Introduction

The Transformer architecture, built on the multi-head attention mechanism, has been shown to improve the quality, efficiency, and scalability of deep learning (DL) models [1]. Despite its success not only in machine translation but also in other areas, such as natural language processing (NLP), computer vision (CV), and multimodal learning, the understanding of its internal training dynamics remains limited. This is largely due to the absence of unified, sensitive measures capable of characterizing convergence behavior across layers and over time. Besides, coarse, task-level measures like loss and gradient norm fail to capture finer-grained changes in information flow and model behavior. Therefore, this limit hinders the development of generalizable training diagnostics and adaptive strategies.

In contrast, a more detailed perspective can be obtained from model-internal measures based on information theory concepts. While gradient entropy describes how widely parameter updates are spread across the model, attention entropy captures how attention weights are distributed, reflecting changes in focus patterns across tokens and heads. Previous studies have demonstrated the value of entropy-based methods in understanding, diagnosing, and optimizing DL models. Entropy metrics are promising in providing more fine-grained and stable indicators of model behaviors, not only as descriptive diagnostic tools but also as active signals of strategies that guide training interventions.

Therefore, this study investigates the entropy patterns during training and evaluates the potential of entropy-based indicators for training guidance. As a key application, a set of entropy-guided adaptive training strategies guided by multi-scale entropy measures is introduced and demonstrated on the Vision Transformer (ViT) over CIFAR-10 dataset. One of the contributions of this work is a unified multi-scale entropy framework, based on gradient and attention entropy, is proposed to represent the Transformer training dynamics and enable earlier detection of phase transitions and overfitting. Under this framework, detailed analyses are provided for specific domains. For NLP tasks, sentence-level attention behavior is analyzed to clarify functions of different layers in BERT; for CV

tasks, entropy-guided adaptive training strategies are introduced, with model-agnostic strategies (e.g. early stopping, staged optimization) designed to improve training efficiency and prevent overfitting.

2. Literature Review

Information theory concepts provide insightful perspectives in understanding representation learning in deep neural networks. Shannon entropy, a fundamental measure of uncertainty in a probability distribution, serves as its foundation [2]. Representation entropies characterize information diversity at different model levels: gradient entropy quantifies parameter update uncertainty, attention entropy measures attention dispersion, and activation entropy reflects neuron output complexity. Besides, the information bottleneck (IB) principle posits that optimal models learn compressed input representations maximally informative about output targets [3]. By measuring entropy of various signals during training (e.g., neuron activations, attention weights), researchers can analyze how information flows and is compressed, enabling better prediction of learning processes and representation quality. The core innovation of Transformer architecture is its sole reliance on the self-attention mechanism, which produces probability distributions over input sequences [1], and therefore well-suited for information-theoretic analysis. As an insightful diagnostic metric, attention entropy reflects information distribution across tokens, heads, and layers, offering fine-grained insights into model behavior.

In 2019, Clark et al. found attention heads in BERT specialize in syntactic roles with distinct entropy patterns. [4] The work by Tian et al. in 2023 described a "Scan and Snap" process where attention evolves from uniform to focused distributions. [5] Yang et al. in 2024 used gradient flow analysis for word identification tasks. [6] Moreover, Chen et al. in 2024 provided theoretical proof for "generalized induction heads" in in-context learning. [7], while Saponati et al. in 2025 proved how training objectives naturally shape attention patterns. [8] In addition, analyses of training dynamics facilitate studies on generalization. The evolution from scanning to snapping and emergence of specialized modules [7] represent mechanisms for learning generalizable features, and such relationship can be quantified by information entropy. [5] Dong et al. in 2024, for instance, established that higher loss entropy correlates with larger generalization gaps. [9] while Mostafa et al. in 2024 demonstrated that lower activation entropy achieves better generalization. [10] Spadaro et al. in 2023 added that in transfer learning, appropriately pruned models with higher activation entropy can exhibit superior generalization. [11].

Traditional optimization strategies such as early stopping and learning rate schedules (e.g., step decay, cosine annealing) are widely applied. Several works have demonstrated in depth how optimization influences generalization. Hardt et al. proved SGD's stochasticity contributes to implicit regularization [12]. Also, Keskar et al. showed that large-batch training converges to sharp minima while smaller batches favor flatter, more generalized solutions [13]. While effective, these approaches rely on static thresholds and lack adaptability. In contrast, recent works explore internal signals, such as entropy measures, as adaptive indicators. Zhou et al. in 2025 found Absolute Gradient Entropy (AGE) tracks the "grokking" phenomenon, providing data-driven triggers for checkpointing. [14] Zhai et al. in 2023 diagnosed attention entropy collapse and proposed Sigmar Param to stabilize training. [15] Chen et al. in 2024 further attributing collapse to variance sensitivity in SoftMax. [16] Beyond stability, Zhou et al. in 2025 designed modules based on gradient entropy to accelerate generalization by actively lowering gradient complexity. [14] while Spadaro et al. in 2023 developed ENTROPI, an entropy-based pruning method that identifies optimal sub-networks for transfer learning [11].

3. Methodology

3.1. Diagnostic Metrics Framework

3.1.1. Entropy metrics

In general, gradient entropy and attention entropy are studied in this work. Gradient entropy captures the distribution of parameter updates during training. As model converging, gradients shift from broadly distributed to sparse and peaked, which suggests the shift from widespread adjustments to localized fine-tuning. Shannon entropy quantifies this distribution:

$$H_{grad} = -\sum_{i=1}^{N} p_i \log_2 p_i \quad (1)$$

Where p_i represents the normalized gradient magnitude. It can be normalized by its theoretical maximum, $\log_2 N$, to obtain a scale-free form:

$$\hat{H}_{grad} = \frac{H_{grad}}{\log_2 N} \quad (2)$$

For gradient entropy, higher values imply widespread parameter adjustments, while lower values reflect localized fine-tuning and refinement.

Attention entropy measures distribution of attention weights over tokens. For each query token, attention weights form a probability distribution, with entropy computed per head and per query:

$$H_{attn} = -\sum_{i=1}^{N} \alpha_i \log_2(\alpha_i + \varepsilon) \quad (3)$$

Where a_i represents the attention weight for the i^{th} token, and ε is a tiny constant added to prevent numerical instability. Higher attention entropy indicates diffuse attention and lower values suggest selective concentration. Unlike gradient entropy, attention entropy varies substantially across layers due to their distinct functions. Excessively low values may indicate attention collapse.

3.1.2. Auxiliary indicators

Apart from entropy metrics, several auxiliary indicators are also introduced to complement the diagnostic framework and offer additional information on model behavior, as shown in Table 1.

Table 1. Auxiliary Indicators

Indicator	Description
Top-k Mass	Measures how concentrated the output distribution is on the top-k predicted classes. A high value can indicate overconfidence.
Head Specialization	Calculates the average pairwise cosine similarity of attention distributions between different heads to indicate how functionally distinct they have become. Greater values suggest divergence between heads and potentially richer representational capacity.
Logits Margin	Captures the margin between the correct class logit and that of the highest-scoring incorrect class. A larger margin implies stronger classification confidence.

3.1.3. Robust statistics for stability

As entropy signals can be affected by noise, outliers, and fluctuations in gradient updates and mini-batch sampling, several robust-statistics techniques are employed to ensure stable interpretation. In epoch level, exponential moving averages (EMA) are applied to reduce short-term volatility in entropy curves. For gradient entropy, Shannon entropy is estimated on the resulting distribution by computing a log-binned histogram over $\log|g|$ on flattened gradients to deal with heavy tails and scale variation. For better summarization of general entropy trends, outputs across multiple random seeds are aggregated by the median with the 95% bootstrap confidence intervals (CI) reported to mitigate outlier impact. It needs to be noted that the full robustness pipeline was not applied uniformly to all experiments. Exploratory experiments used none of the methods above except simple moving average smoothing in visualization. For sentence-level analysis, attention entropy is computed per layer across epochs and mean-based aggregation across seeds with CI, instead of median-based

aggregation and CI, is introduced. Only in final experiments is the complete pipeline systematically implemented and tested.

3.2. Training Diagnosis Strategies

One strategy tested throughout the whole study is entropy-based early stopping, including two complementary variants. The more comprehensively explored variant is the joint early stopping strategy, which uses both loss and entropy signals as trigger signals to a more robust convergence signal than loss/entropy alone. Specifically, training is halted when validation loss fails to improve for a patience window and the mid-layer attention-entropy trend is flat (near-zero slope) in the later ViT experiments after initial explorations. In the final pipeline, the strategy combining mid-layer attention entropy with loss for early stopping is chosen, where "mid-layer" is defined as the central three layers of the encoder ($\frac{\mathrm{L}}{2}-1$, $\frac{\mathrm{L}}{2}$, and $\frac{\mathrm{L}}{2}+1$, where L is the number of layers) and the mean attention entropy across these layers is calculated. Meanwhile, an early stopping strategy triggered solely by entropy signals (preferably mid-layer attention entropy) is also explored in early experiments, which provides two signals to guide stop: S1 detects the first significant raw drop after a short warm-up and S2 marks stabilization via the earliest of either a smooth plateau or a sustained negative slope over a patience window. This variant is ultimately discarded because of its incompetence and retained here only for completeness. In early post-hoc analysis, conventional loss-based early stopping strategies, which stop training when the validation loss does not improve after a predefined number of epochs, are used as baseline.

Apart from early stopping, other entropy-based strategies are also introduced to detect meaningful shifts in entropy dynamics, thereby intervening mid-training. Possible methods include reducing the learning rate, checkpointing, tightening regularization/augmentation or freezing certain layers. For instance, the scheduler employed for the later ViT experiments triggers a stage transition when mid-layer attention entropy experience pattern, multiplicatively decaying the LR and freezing early encoder layers.

4. Results

4.1. Datasets and Models

Experiments are conducted on both text and vision datasets. For NLP tasks, a set of self-authored, phenomena-labelled sentences serves to analyze the attention behaviors of a pre-trained BERT on targeted linguistic phenomena (e.g. syntax, coreference, keywords). Meanwhile, for CV tasks, the primary dataset employed is CIFAR-10. Simpler MNIST dataset and more complicated CIFAR-100 dataset are also employed to examine signals in low-dimensional and more complex settings in exploratory stages respectively. As for model architecture, for the linguistic phenomena analysis, a pre-trained, frozen version of the$bert-base-uncased$model is used in a forward-only (inference) setup designed to analyze the model's inherent attention patterns, while a lightweight ViT serves as the primary architecture for CV experiments, which combines a patch embedding module, a multi-layer self-attention encoder, and a final linear classification head.

4.2. Experiment Setups

4.2.1. Linguistic-phenomena Analysis

A group of self-authored sentences targeting at specific linguistic phenomena is used for establishing how BERT distributes attention under targeted linguistic conditions. Each sentence is designed to isolate one factor (e.g., subject-object swaps). A pre-trained, frozen $\mathrm{bert}-\mathrm{base}-\mathrm{uncased}$ encoder is used without fine-tuning/training and no adapters, with standard WordPiece tokenisation set a fixed max length excludes attention on padding positions from analysis, so that reported attention patterns can be architectural and pre-training without biases of the checkpoint and tokenizer. Outputs include Layer-head attention-entropy heatmap, head specialization vs. layer depth

figure, three case studies analyses, and Jensen-Shannon divergence (JSD) figures between polysemous vs. disambiguated attention distributions across layers. Specific linguistic phenomena cases probed are summarized in Table 2.

Table 2. Linguistic-Phenomena Cases

Phenomenon	Analysis Objective	Test Case
Syntax (Voice)	Attention distribution shift on different voices	Active vs. Passive sentence pair ("The quick brown fox chases after the lazy dog." vs "The lazy dog is chased by the quick brown fox.")
Conference Resolution	Track attention from pronoun(s) to antecedent	Sentence with pronouns and a clear referent ("The dog wagged its tail because it was happy.")
Polysemy	Verify contextual disambiguation of an ambiguous word	Sentence using a word of two meanings ("The bank is on the river bank.")

4.2.2. Exploratory Entropy & Early Stopping Analyses

Table 3. ViT Settings

Model	Number of Layers	Hidden Dim
M_Light	4	64
M_Base	8	128
M_Deep	12	128
M_Wide	8	256

Table 4. Early Stopping Strategies Under Investigation

Strategy	Description (Triggering Condition)
Loss Early Stopping (ES)	Based on validation loss stagnation.
Gradient Entropy ES	Combines loss stagnation with global gradient entropy trend stabilization.
Global Attention ES	Combines loss stagnation with global average attention entropy trend stabilization.
Mid-Layer Attention ES	Combines loss stagnation with the average trend stabilization of mid-layer attention entropy.
Entropy-only ES	Detects the "first dip" and "plateau" of the attention entropy curve itself. Based on observations in entropy plots of the MNIST model

Experiments plan to map gradient/attention entropy trajectories and prototype early-stop ideas from CV experiments. There are four settings of lightweight ViT models, all shown in Table 3. Two Adam optimizers are covered, both with $\mathrm{LR} = 5 \times 10^{-4}$ one with $\mathrm{WD} = 1 \times 10^{-4}$ while the other without WD. During data processing, while MNIST undergones normalization only, CIFAR-10/100 experiments use extra data augmentation methods (e.g. $RandomCrop$, $RandomHorizontalFlip$) apart from normalization. For three datasets, training would take 30, 60 and 80 epochs respectively. The different early stopping strategies tested in this phase are defined in Table 4.

4.2.3. ViT Adaptive Training With Entropy-guided Strategies Pipeline

Afterwards, a pipeline for adaptive training is designed and its efficiency and performance are compared with what in the control experiments. The M_Base and M_Deep models are trained for completed 100 epochs as a baseline, comparing against the adaptive strategies defined in Section 3.2. Full validation and logging of internal metrics (attention/hidden entropies, head specialisation, logits margin, top-k mass) are performed every epoch.

4.3. Linguistic-Phenomena Analysis

The heatmap in Figure 1 reveals a hierarchical workflow of BERT model. At the lowest layers, attention heads attend to many words to gather foundational information, resulting in relatively high entropy. The closest head to the input $L1-H1$ exhibits particularly high entropy. Subsequently, some heads become highly specialized in the low-middle to middle layers with focused attention and lowering entropy. For the semantic processing stage of middle-to-high layers, the model requires different types of heads to work in synergy, such as key semantic information (low entropy), or integrating global contextual information (high entropy). Hence, entropy values for all heads vary significantly in such layers. Meanwhile, in the per-layer attention entropy curve shown in Figure 2, L1 has the highest entropy (approximately $2.5+$ bits), L6-8 have the lowest (~ 1.35 bits), and L9-12 see a rebound, which can be interpreted as diffuse attention in lower layers (broad scan), focused attention in middle layers (syntactic/semantic processing), and re-expanded attention in higher layers (global integration). This pattern suggests the IB theory, with intermediate representations compressing task-irrelevant information while preserving structural features.

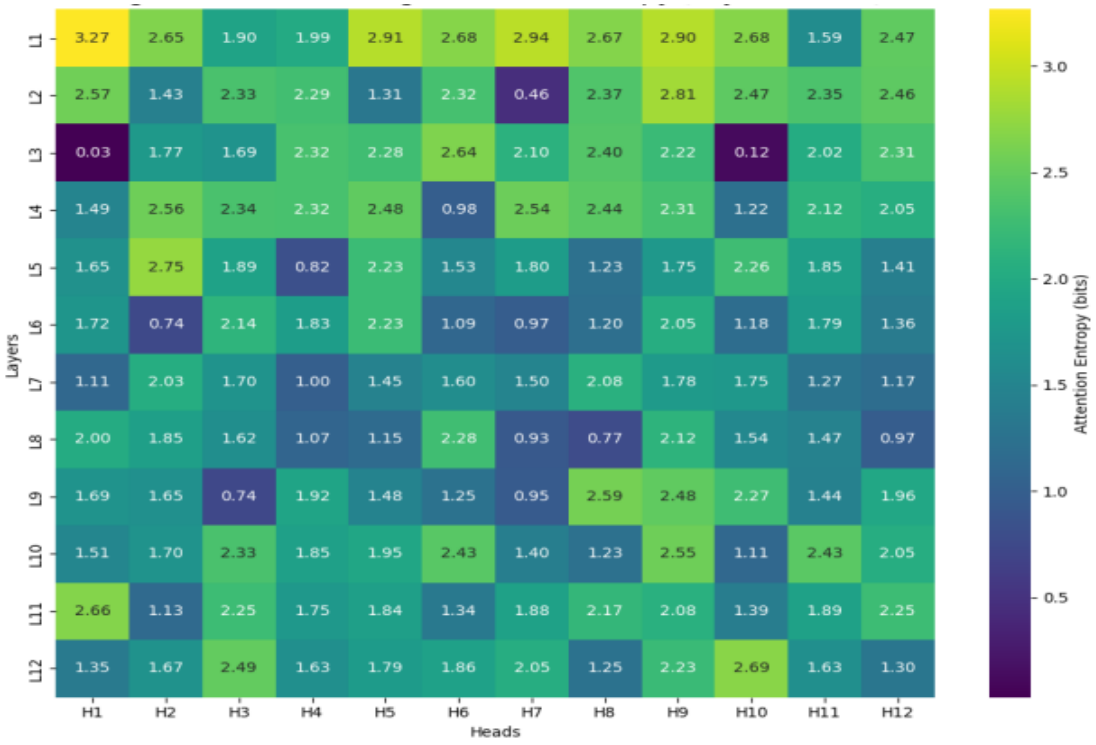


Figure 1. Heatmap of Global Average Attention Entropy (Layer $\times$ Head)

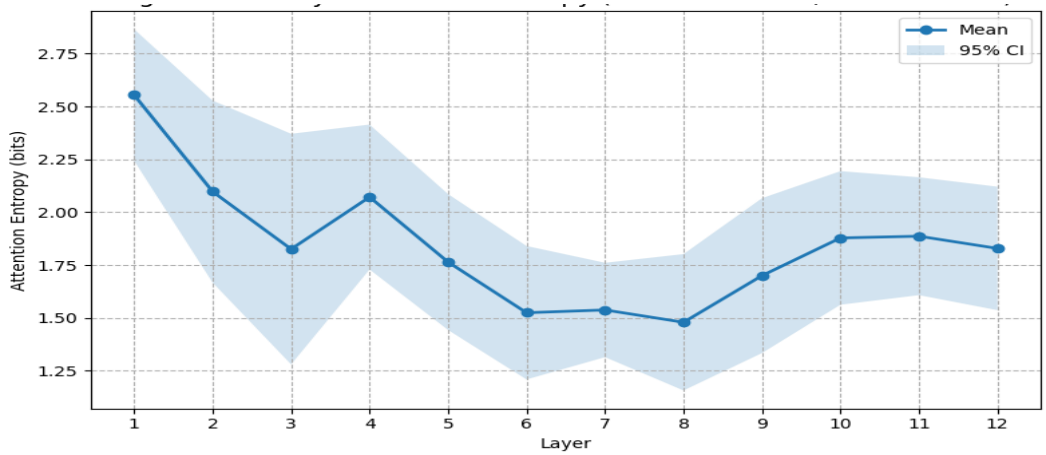


Figure 2. Per-layer Attention Entropy (Mean $\pm$ 95% CI, across Heads)

Experiments across case studies demonstrate BERT's layer-specific functions. Overall, lower layers prioritize local syntax and function words, middle layers construct main structure and long-distance dependencies, and high layers perform information summarization and discourse-level integration. Throughout this process, attention allocation reflects hierarchical syntactic preference and sensitivity to semantic roles (with agent bias), and more complex structures are accompanied by higher entropy.

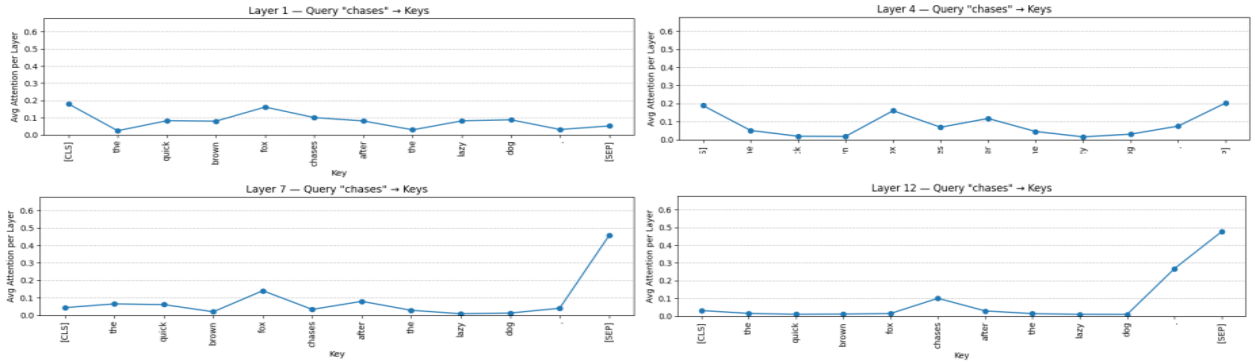


Figure 3. Case 1 Attention Entropy Line Charts in Selected Layers: Active Voice (“chases” to keys)

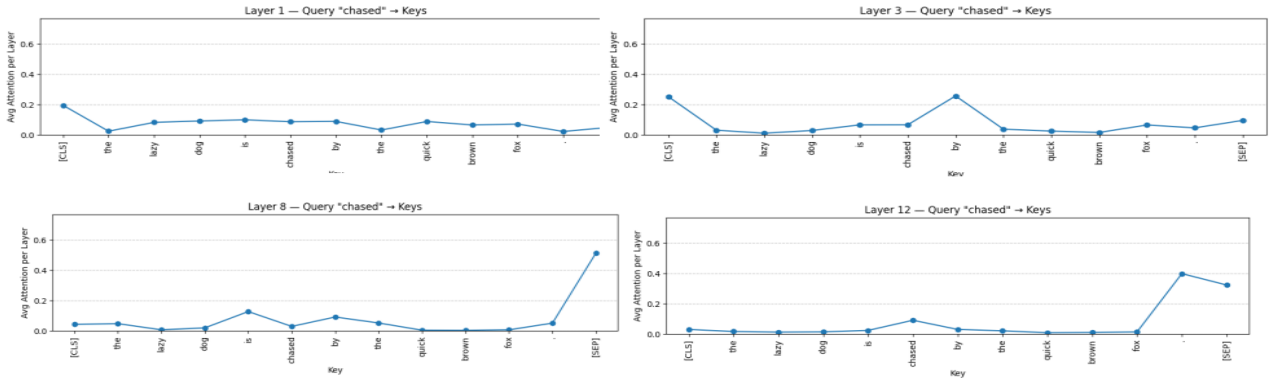


Figure 4. Case 1 Attention Entropy Line Charts in Selected Layers: Passive Voice (“chased” to keys)

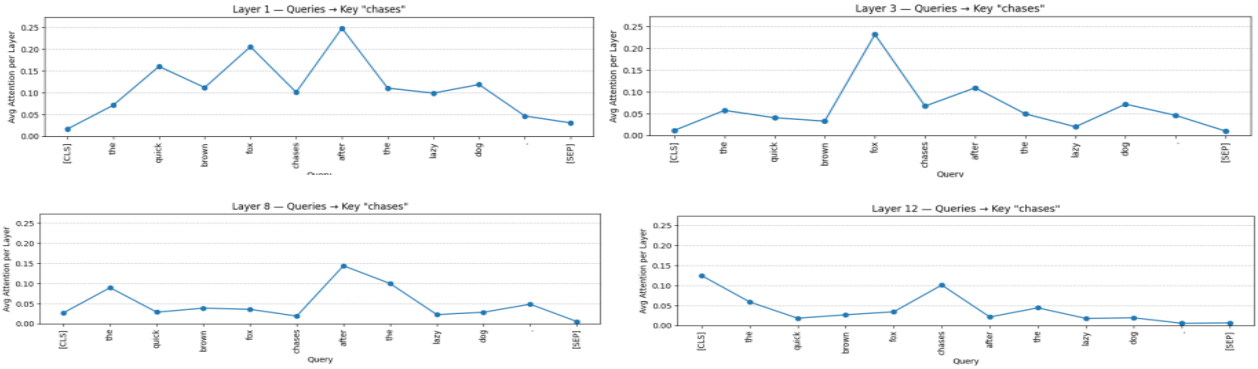


Figure 5. Case 1 Attention Entropy Line Charts in Selected Layers: Passive Voice (queries to “chases”)

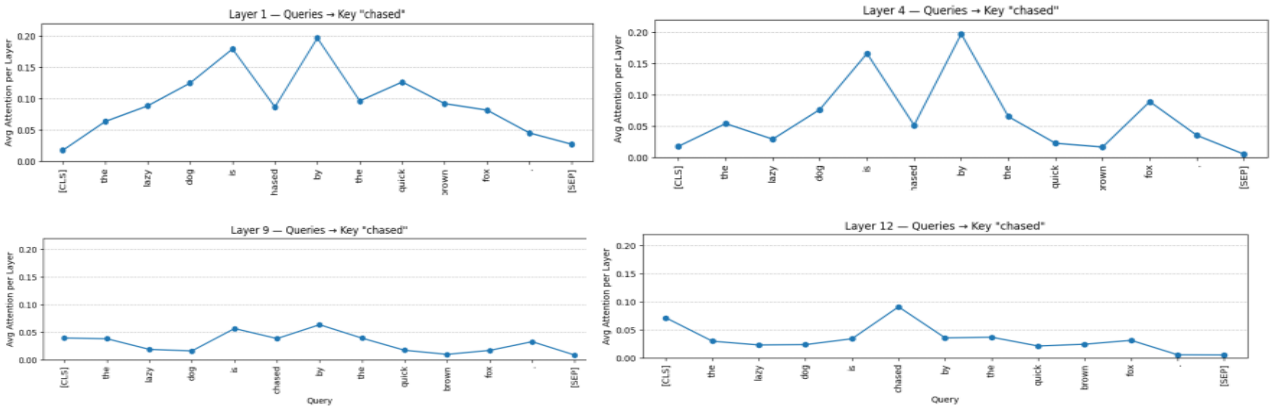


Figure 6. Case 1 Attention Entropy Line Charts in Selected Layers: Passive Voice (queries to "chased")

In Case 1 (Fig. 3-6), attention in lower layers distributes significantly to auxiliary verbs, prepositions, and phrasal collocations, tracking the grammatical skeleton. For both active and passive sentences, attention on the agent noun significantly exceeds that on the patient noun in lower layers, with the main verb also receiving focused attention. However, this mutual focus disappears in deeper layers, indicating subject-verb relationship analysis completes by middle layers. This agent-preference bias persists regardless of surface syntactic position, demonstrating BERT prioritizes semantic role identification over word order.

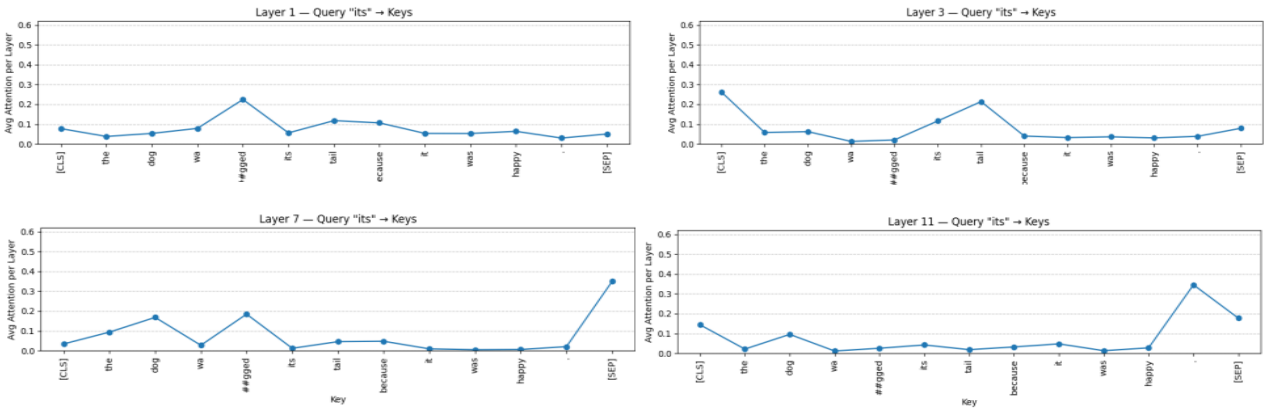


Figure 7. Case 2 Attention Entropy Line Charts in Selected Layers: "its" to keys

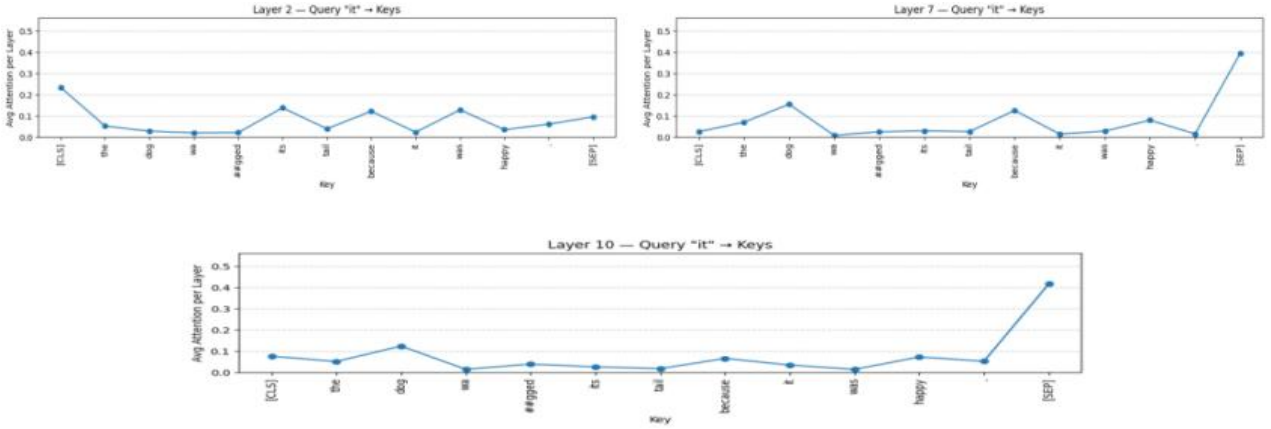


Figure 8. Case 2 Attention Entropy Line Charts in Selected Layers: "it" to keys

For Case 2 (Figs. 7-8), pronouns "its" and "it" experience a shift from local to global attention distribution. In low layers, they complete immediate syntactic tasks ("its" to its governed noun, "it" to clause-internal elements). Middle layers introduce logical connectives. Finally, in high layers, both

tokens' attention converges on the same antecedent. This local-to-global trajectory implies coreference resolution is a multi-stage reasoning process requiring high-layer synthesis.

Case 3, illustrated in Figure 9, demonstrates BERT's semantic disambiguation mechanism. When bank_1 (financial institution) serves as key, signals come from syntactic role, confirming “subject-copula-complement structure”. When bank_2 (riverbank) is the key, the meaning-determining word "river" drives attention, giving bank_2 the highest attention and forming the phrase "on the river bank," where semantic cues dominate. Attention from "river" to bank_2 is around 2.5 times that to bank_1, indicating the suppression of irrelevant co-occurrences. Therefore, to deal with strong semantic cues, the model prioritizes semantic pathways, otherwise falls back to syntactic pathways. This demonstrates a semantic-first, syntactic-fallback disambiguation strategy.

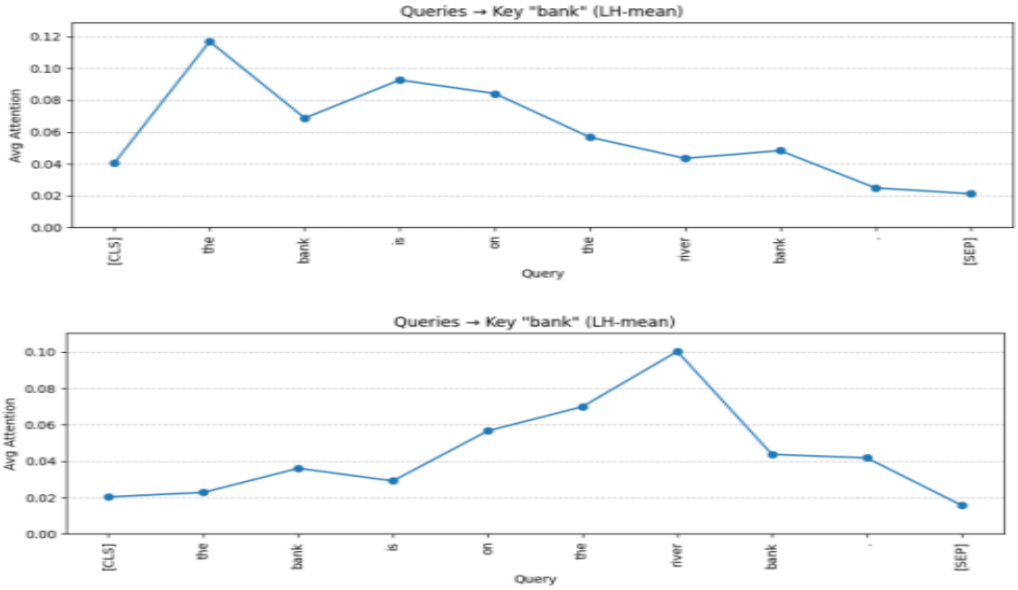


Figure 9. Case 3 Attention Entropy Line Charts in Selected Layers: queries to bank

Furthermore, polysemy analysis using within-phenomenon and across-phenomena JSD (Figs. 10-12) demonstrates that semantic differentiation requires deeper layers.

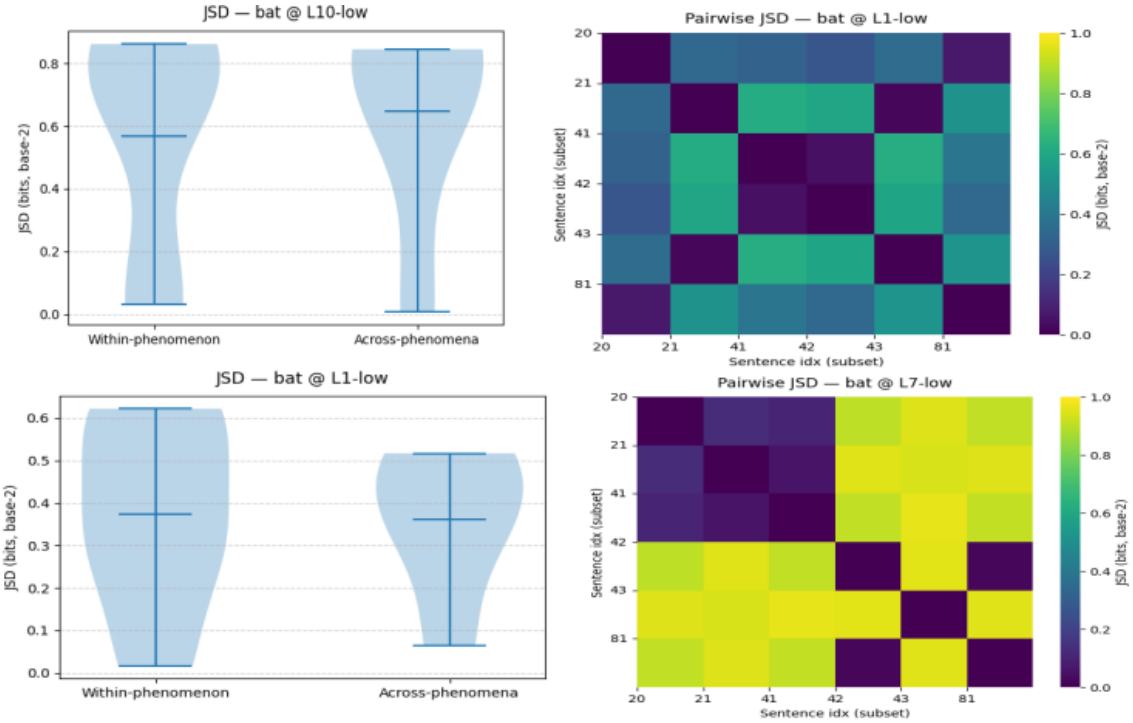


Figure 10. Selected Pairwise, Within Phenomenon & Across Phenomena JSD: Polysemy - bat

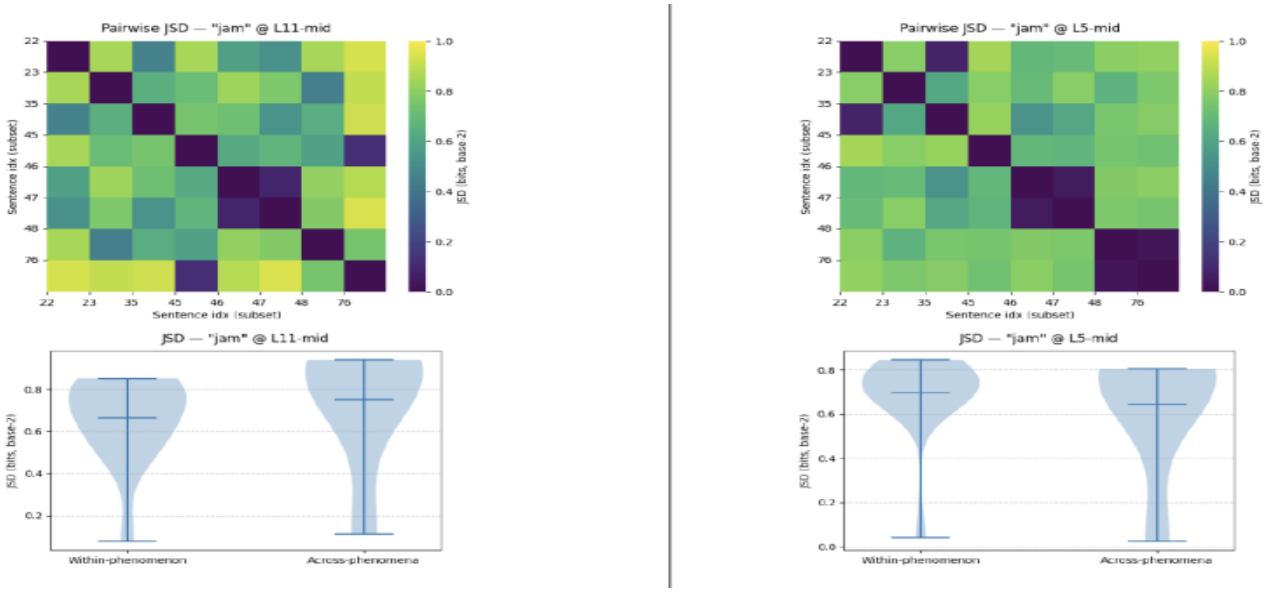


Figure 11. Selected Pairwise, Within Phenomenon & Across Phenomena JSD: Polysemy - jam

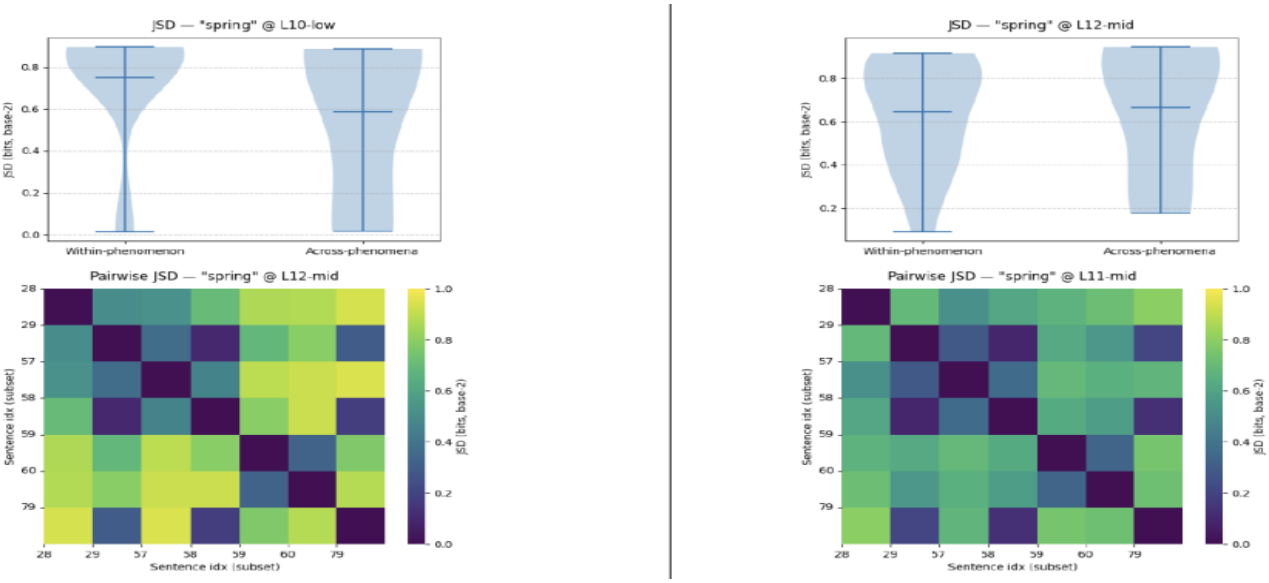


Figure 12. Selected Pairwise, Within Phenomenon & Across Phenomena JSD: Polysemy - spring

Most polysemous words show no differentiation in lower layers; within and across JSD values are close, indicating the model operates at word form level. By middle layers (L5-7), easily distinguishable words (e.g., "bat", "coach") begin differentiating, with across JSD increasing while within JSD decreases. Differentiation for almost all words only becomes apparent in deep layers (L10-12), where across JSD significantly exceeds within JSD, demonstrating sense disambiguation while maintaining within-sense consistency. Notably, Figure 12 shows that the model has persistent disambiguation difficulty for "spring", with high within JSD even in the deepest layers. These JSD experiments confirm that while phenomena resolved early require only lower-middle layers, complex operations like polysemy require high-layer attention for generalization and consistency.

4.4. Exploratory Entropy & Early Stopping Analyses

Due to the ViT model's strong dependence on data when training from scratch, its performance decreases significantly as dataset complexity increases. In Figure 13, while all results achieve approximately 98.6% accuracy on MNIST, accuracy decreases to 81-83% on CIFAR-10 and 56-58% on CIFAR-100, indicating over-parameterization for MNIST and insufficient capacity for CIFAR-100. Hence, CIFAR-10 is selected to test final pipeline.

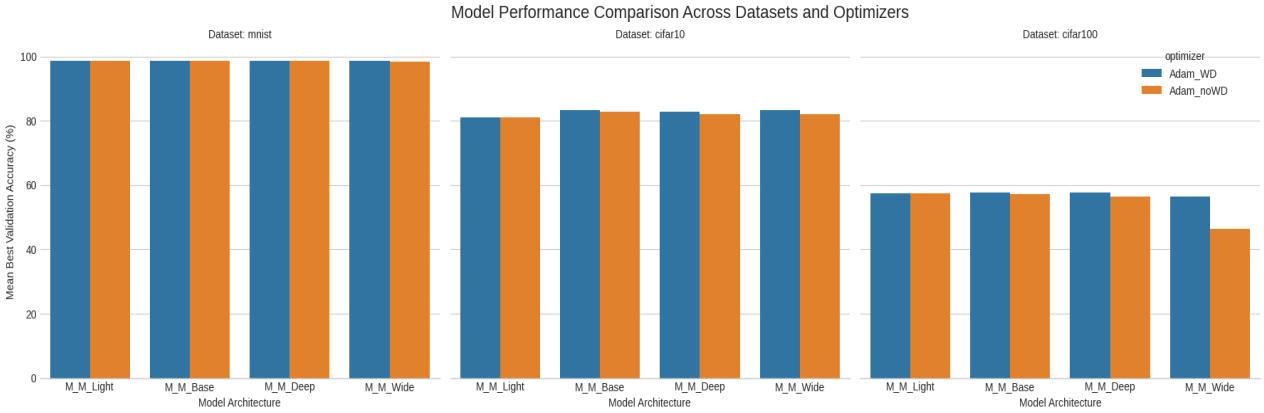


Figure 13. Model Performance Comparison Across Datasets and Optimizers

It is also proved that entropy trajectories can serve as diagnostic signatures for capacity mismatch. Excessive capacity (Figure 14) produces rapid convergence within 5-10 epochs with validation accuracy exceeding 98% before plateauing. Overfitting emerges around epoch 9 as validation loss ascends. Average global attention entropy exhibits a rise-and-fall pattern peaking after validation loss inflection, and gradient entropy peaks later, consistent with transition from broad updates to fine-tuning. Conversely, insufficient capacity (Figure 15) yields suppressed accuracies and non-plateauing attention entropy with unstable fluctuations, signaling failure to identify stable learning directions.

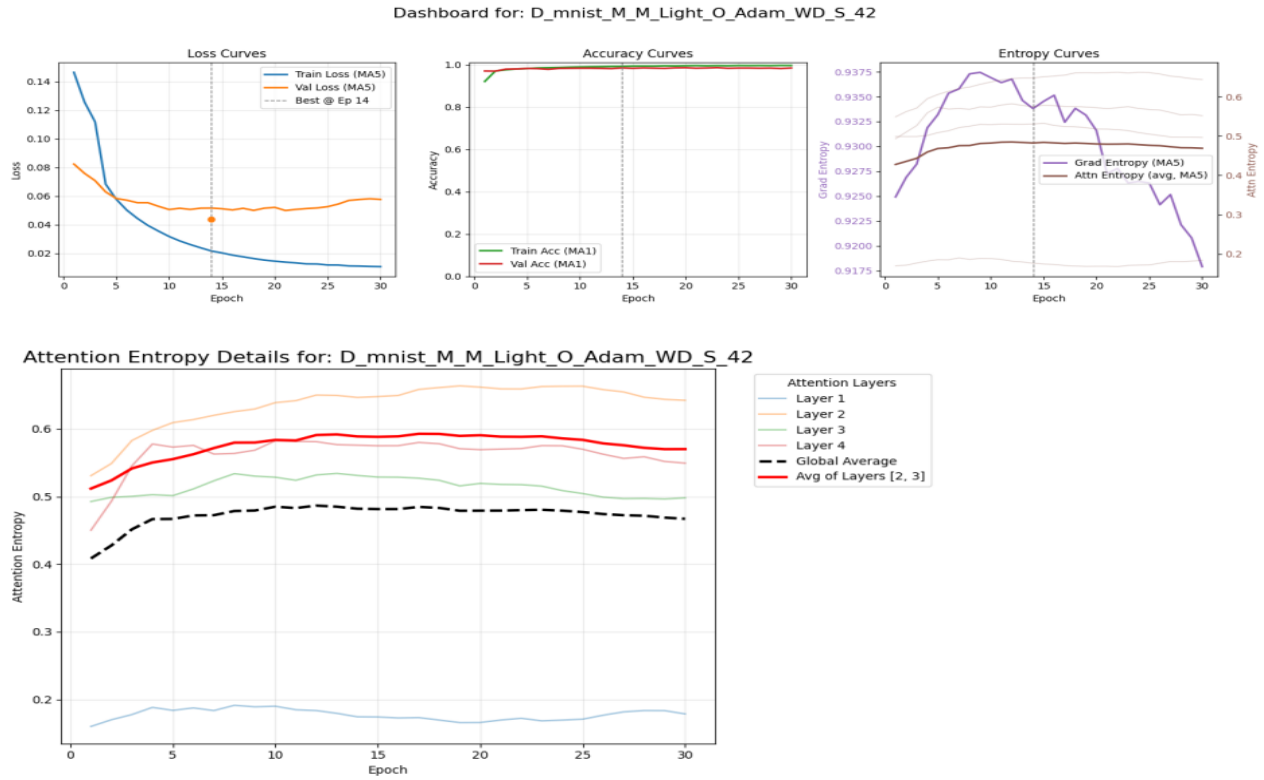


Figure 14. Example: Model Volume Much Larger Than Needed (MNIST, no WD, M_Light)

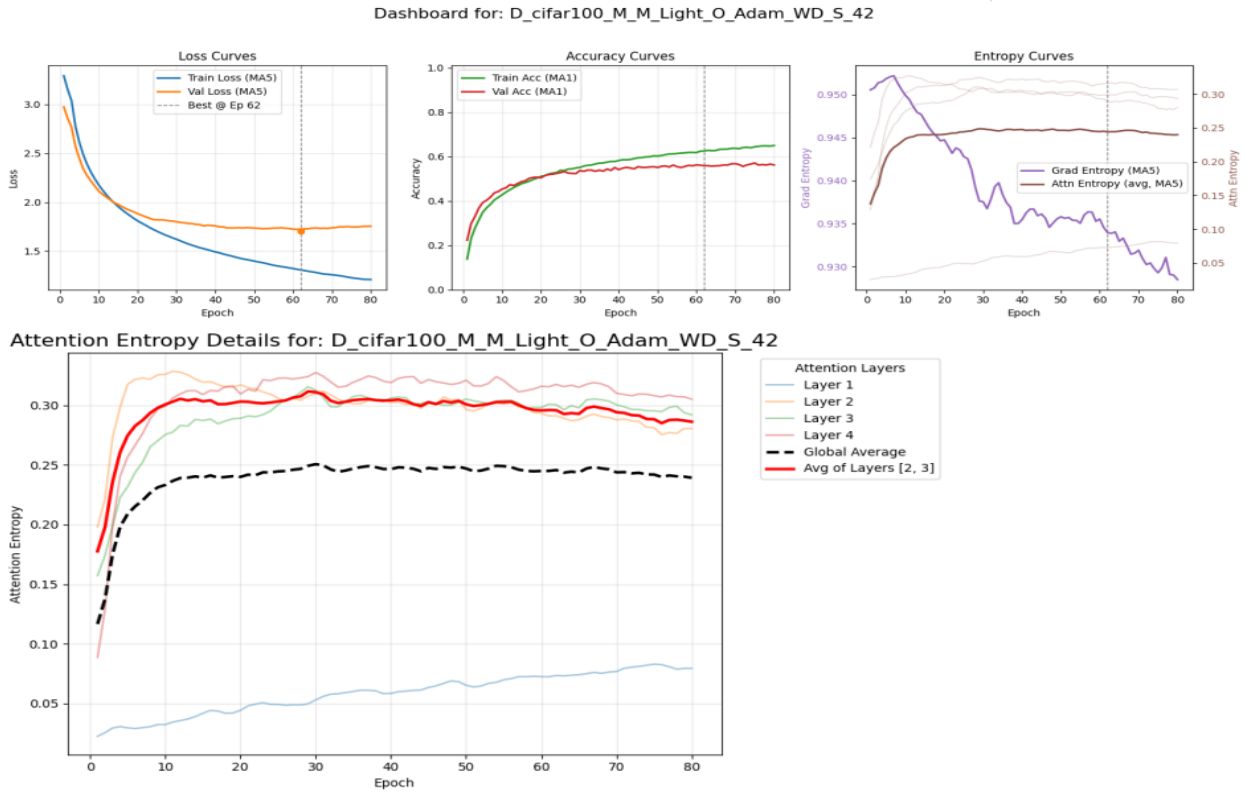


Figure 15. Example: Model Volume Much Less Than Needed (CIFAR-100, WD, M_Light)

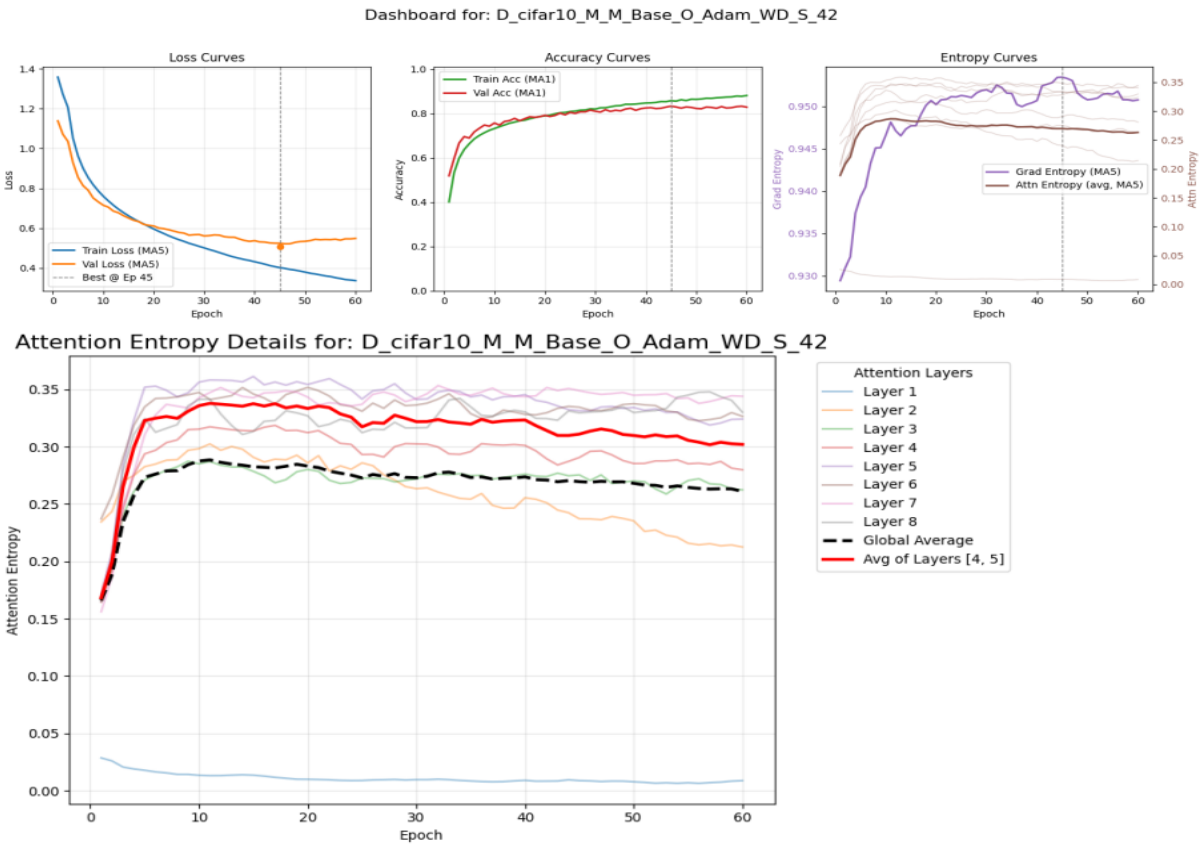


Figure 16. Visualization of a Model Relatively Perform Better (CIFAR-10, WD, M_Base)

Meanwhile, for balanced model-dataset pairings, such as what is shown in Figure 15, entropy dynamics shift distinctly, with the rise-and-fall pattern replaced by a noisy near-plateau trajectory.

Meanwhile, gradient entropy plateaus substantially earlier than convergence as shown in Figure 16, indicating sustained an active, high-intensity learning throughout the training. Absolute attention entropy values decline compared to MNIST, which suggests the shift from diffuse to sparse, peaked attention distribution concentrating on critical patches. This confirms full utilization of training duration for the model to effectively learn from the given data.

For each seed and model-WD-dataset combination, a post-hoc analysis of the given five early-stopping strategies were conducted, with selected results shown in Table 5.

Table 5. Post-hoc Analysis: Triggered Epoch (Selected)

Combination	Loss ES	Gradient Entropy ES	Global Attention ES	Mid-Layer Attention ES	Entropy-only ES	Best Accuracy
CIFAR_10 + M_Base + Adam_WD + seed = 16	47	47	47	58	23(S1), 43(S2)	54
MNIST + M_Base + Adam_noWD + seed = 16	30	30	Not Triggered	Not Triggered	23(S1), 24(S2)	20
CIFAR_10 + M_Base + Adam_noWD + seed = 123	47	47	52	53	23(S1), 38(S2)	57
CIFAR_10 + M_Deep + Adam_noWD + seed = 2025	26	26	36	30	22(S1), 49(S2)	31

The results reveal validation loss-based stopping consistently terminates prior to peak accuracy, as cross-entropy loss already reaches inflection when accuracy continues refining decision boundaries. Mid-layer attention entropy emerges as most effective, with stabilization points closely aligned with peak accuracy epochs. This implies that mid-layer's construct intermediate representations, converge last, so they are therefore the most reliable signal. The entropy-only strategy is proved to be hyperparameter-dependent with premature triggering on complex tasks, resulting in the adoption of hybrid loss-entropy strategies for adaptive pipeline. Several limitations still exist, including occasional strategy failures, fixed hyperparameter constraints, and inability to measure real-time computational savings due to its post-hoc nature. Subsequent experiments therefore narrow model-dataset combinations, extend training duration, and implement modified entropy calculation methods with comprehensive robustness protocols to deal with curve noise and reliability concerns.

4.5. ViT Adaptive Training with Entropy-guided Strategies Pipeline

Figure 17-19 visualizes the aggregated outputs from the training process of M_Base + Adam_WD combination. The adaptive pipeline achieves convergence at approximately half the epochs required by the control group, with validation accuracy and loss plateauing at termination. This demonstrates that entropy-based early stopping pipeline substantially reduces computational cost while maintaining performance.

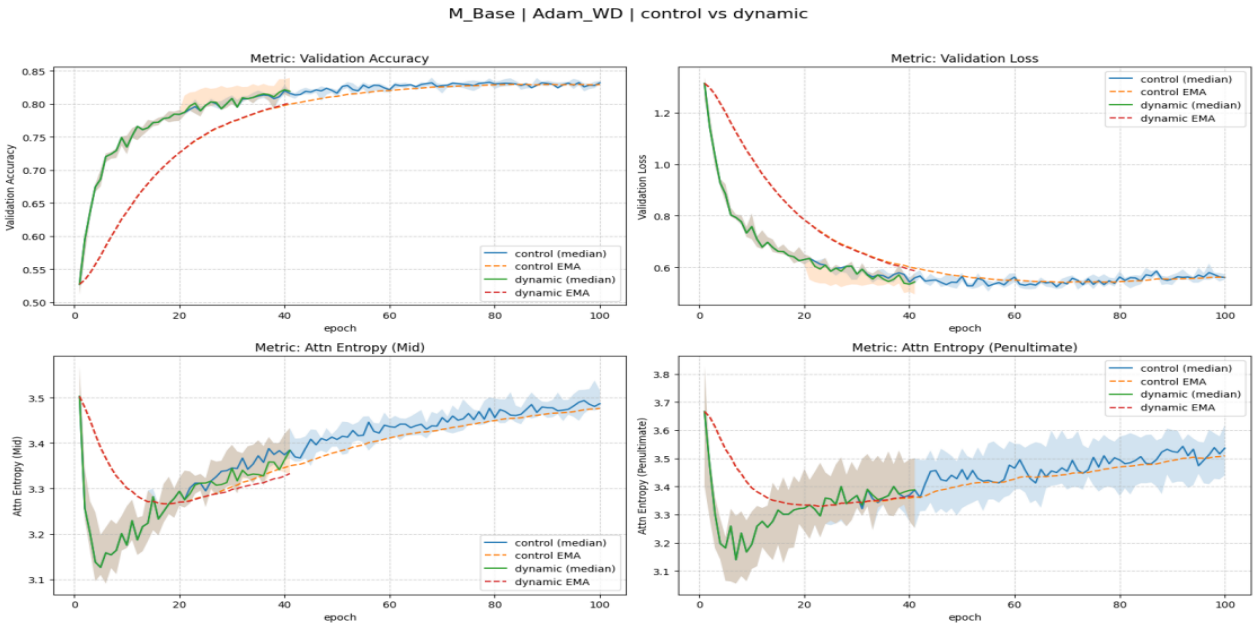


Figure 17. Curve of Evaluation Metrics for Control and Adaptive Training (Part 1)

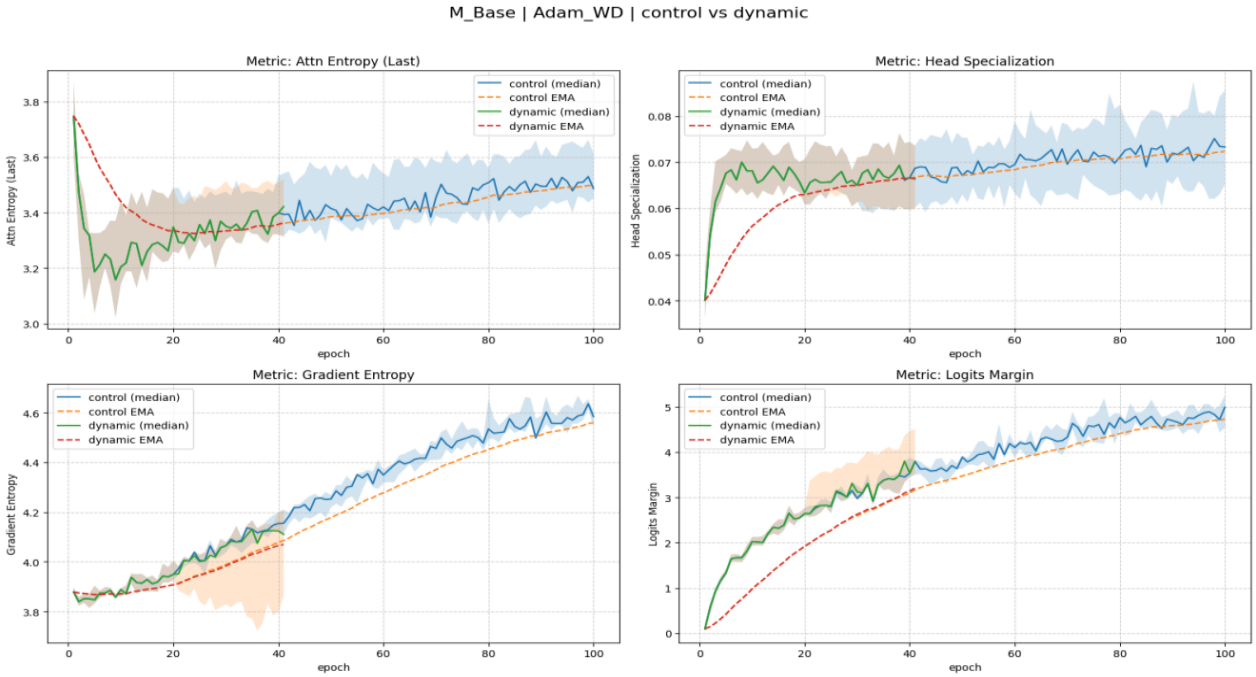


Figure 18. Curve of Evaluation Metrics for Control and Adaptive Training (Part 2)

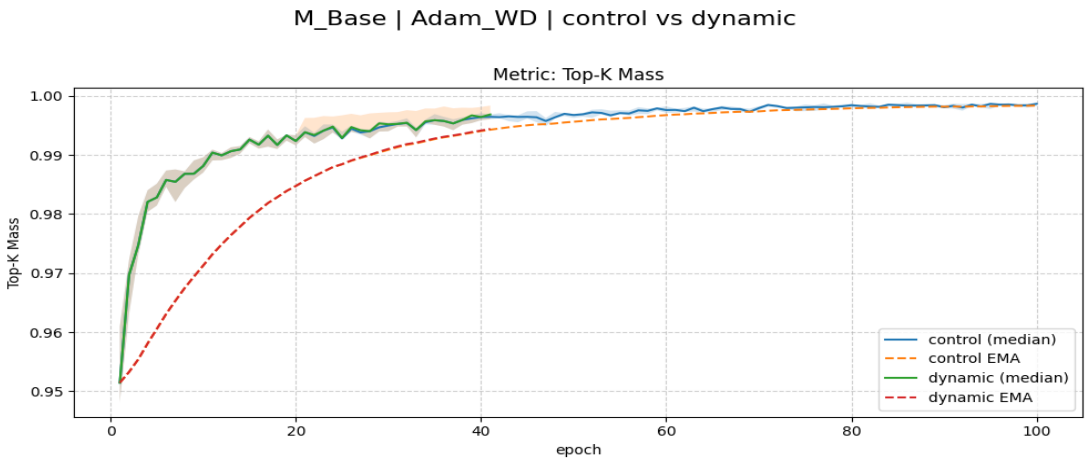


Figure 19. Curve of Evaluation Metrics for Control and Adaptive Training (Part 3)

Regarding entropy metrics, smoothed multi-scale attention entropy curves exhibit consistent patterns across mid-layer mean, penultimate layer, and final layer: initial decline followed by gradual ascent towards plateau. Mid-layer entropy displays the greatest fluctuation, confirming its suitability for guiding adaptive training. In contrast, gradient entropy increases continuously throughout control training, rendering it unsuitable as a stopping trigger despite its utility for characterizing internal updates. This implies the pipeline halts whilst minor gradient refinements persist, though overall convergence is assured, and the potential of combining both gradient and attentional entropy metrics with loss. In addition, the newly introduced metrics, top-k mass and head specialization, also nearly reach their plateau at the early-stopping epoch, indicating that the model has already developed sharp, concentrated attention distributions and highly specialized heads. However, similar to gradient entropy, the logit margin in the control group continues to rise even after the early-stopping trigger point. This can be attributed to the margin between correct and incorrect classes continuing to be refined for a period after the model has generally converged, which explains the hysteresis phenomenon, where validation accuracy increases slightly after attention entropy stabilization, and suggest future work could integrate gradient entropy and logit margin into hybrid stopping criteria.

4.6. Discussion

Multi-scale entropy analysis across BERT and ViT reveals that mid-layer attention entropy stabilization provides more reliable convergence signals than loss inflection. This motivates the Mid-Layer Stabilization Hypothesis (MSH): mid-layers, assembling local elements into reusable intermediate representations, stabilize last as learning transitions from global exploration to local fine-tuning. This phenomenon suggests information compression begins earlier in intermediate layers, consistent with IB theory. This reconciles divergent observations across domains. In NLP, upper layers consolidate semantic capabilities (evidenced by declining within-phenomenon JSD), while mid-layers reach stationarity last, rendering them optimal for operational control. In CV, penultimate and final layers show similar entropy trajectories but suffer temporal lag, suggesting mid-layers superior providing entropy signal for early stopping. For other entropy metrics, global average attention entropy has potential to provide capacity diagnostics: over-capacity manifests as early peaks with rapid plateaus, while under-capacity produces persistent noise without stabilization. This signal, though requiring broader validation in future, operates at a different analytical scale than layer-specific entropy and suggests attention mechanisms undergo phase transitions from diffuse to sparse feature selection. Gradient entropy proves unsuitable as primary stopping criterion due to optimization sensitivity and premature plateaus. However, it captures late-stage boundary, which lags behind loss convergence, and therefore promising as a complementary signal for current hybrid stopping rules. Practical deployment requires addressing constraints. Entropy plateaus may fail under underfitting or appear spuriously under distribution shift. Robust implementation necessitates

duration thresholds, stability tests, and bootstrap CIs, so that portability across architectures can be ensured and entropy measures from diagnostic can be used for principled training control system.

5. Conclusion

In conclusion, this work demonstrates that multi-scale entropy is a powerful tool for observing and interpreting the training dynamics of Transformer models. By providing a fine-grained view into the model's internal state, these entropy-based metrics can serve as effective signals for practical applications, including real-time training diagnostics, entropy-aware early stopping, and staged optimization without changing model architecture or task setup. Across different domains, mid-layer attention entropy generally tends to enter a low-variance steady regime near convergence, while deeper layers remain helpful for explaining capability consolidation (e.g. greater within-phenomenon consistency in the linguistic probes). In ViT, using mid-layer attention entropy as the primary control signal supports an adaptive pipeline that combines early stopping with staged optimization, without changing the architecture or the task. Auxiliary indicators usually reach a plateau around the same epoch and help contextualize decisions, while gradient entropy remains informative about update activity, although it is less dependable as a sole trigger. For future work, a key priority is to extend our analysis to larger and more complex scenarios, including large-scale datasets like ImageNet and the GLUE benchmark, as well as state-of-the-art models such as the GPT and BERT families. Furthermore, while this study suggests correlation, future work should aim to establish a more rigorous quantitative relationship between specific entropy measures that is not yet comprehensively explored (e.g. average global attention entropy) and their potential usage.

References

[1] Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All You Need. Advances in Neural Information Processing Systems, 2017, 30: 5998 - 6008.

[2] Cover T. M., Thomas J. A. Elements of Information Theory (2nd ed.). John Wiley & Sons, 2006.

[3] Tishby N., Pereira F. C., Bialek W. The information bottleneck method. arXiv preprint, 2000.

[4] Clark K., Khandelwal U., Levy O., Manning C. D. What does BERT look at? An analysis of BERT's attention. arXiv preprint, 2019.

[5] Tian Y., Chen Y., Zhang B., Li Y., Ramea K., Daskalakis C. Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer. Advances in Neural Information Processing Systems, 2023, 36: 72338 - 72353.

[6] Yang H., Li Y. Training Dynamics of Transformers to Recognize Word Co-occurrence via Gradient Flow Analysis. Advances in Neural Information Processing Systems, 2024.

[7] Chen S., Sheen H., Wang T., Yang Z. Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers. Advances in Neural Information Processing Systems, 2024.

[8] Saponati M., Filippi A., Ansuini A., Micheli A., Bacciu D., Livi L. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in Transformer training. International Conference on Machine Learning, 2025.

[9] Dong Y., Shu H., Li P. Rethinking Information-Theoretic Generalization: Loss Entropy Induced PAC Bounds. International Conference on Learning Representations, 2024.

[10] Mostafa S., Islam M., Karia R. Leveraging Neuron Activation Patterns to Explain and Improve Deep Learning Classifiers. International Conference on Learning Representations, 2024.

[11] Spadaro G., Bacco M., Bertini F., Passaro G., Livi L., Alippi C. Shannon Strikes Again! Entropy-Based Pruning in Deep Neural Networks for Transfer Learning Under Extreme Memory and Computation Budgets. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2023: 3600 - 3609.

[12] Hardt M., Recht B., Singer Y. Train faster, generalize better: Stability of stochastic gradient descent. International Conference on Machine Learning, 2016: 1225 - 1234.

[13] Keskar N. S., Nocedal J., Mudigere D., Smolyansky M., Tang P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. International Conference on Learning Representations, 2017.

[14] Zhou X., Ji Y., Liu Z., Du S., Wei K. NeuralGrok: Accelerate Grokking by Neural Gradient Transformation. arXiv preprint, 2024.

[15] Zhai S., Likhomanenko T., Littwin E., Busbridge D., Ramapuram J., Zhang Y., Gu J., Susskind J. Stabilizing transformer training by preventing attention entropy collapse. arXiv preprint, 2023.

[16] Chen Z., Zhang J., Wu K., Han X., Liu Z. Variance Sensitivity Induces Attention Entropy Collapse and Instability in Transformers. arXiv preprint, 2024.