Extended Abstract Track

On the Ricci Curvature of Attention Maps and Transformers Training and Robustness

Abstract

Transformer models have revolutionized machine learning, yet the underpinnings behind their success are only beginning to be understood. In this work, we analyze transformers through the geometry of attention maps, treating them as weighted graphs and focusing on Ricci curvature, a metric linked to spectral properties and system robustness. We prove that lower Ricci curvature, indicating lower system robustness, leads to faster convergence of gradient descent during training. We also show that a higher frequency of positive curvature values enhances robustness, revealing a trade-off between performance and robustness. Building on this, we propose a regularization method to adjust the curvature distribution and provide experimental results supporting our theoretical predictions while offering insights into ways to improve transformer training and robustness. The geometric perspective provided in our paper offers a versatile framework for both understanding and improving the behavior of transformers.

1. Introduction

Transformers (Vaswani et al., 2017) have become the cornerstone of modern machine learning, excelling in tasks across natural language processing (Devlin et al., 2018) and computer vision (Dosovitskiy et al., 2020). Recent studies have begun to explain the inner workings of transformers, analyzing their generalization properties (Li et al., 2023; Zhou et al., 2024), learning dynamics (Abbe et al., 2024; Dovonon et al., 2024), and robustness (Bhojanapalli et al., 2021). A rigorous understanding of how transformers process their input remains elusive, especially through mathematical tools offering practical solutions. To address this gap, we analyze transformers by treating attention mechanisms as graph operators and exploring for the first time the *geometry* of attention maps as weighted graphs. Focusing on graph Ricci curvature (Bauer et al., 2017; Ollivier, 2009) –closely tied to graph spectral properties and system robustness (Bauer et al., 2011; Pouryahya et al., 2017)– our findings reveal its role in shaping the training dynamics and robustness of transformers.

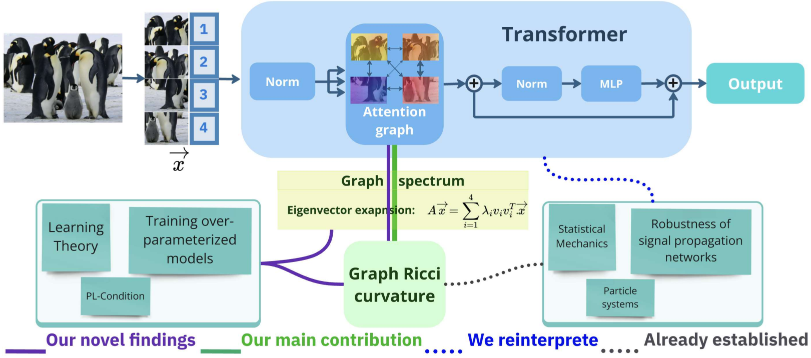

Figure 1: A visual summary of the main concepts and contributions, exemplified using a simple vision transformer with 4 tokens, though our findings are more general.

We analyze the relationship between attention graph geometry and transformer training and robustness, and prove lower Ricci curvature, indicative of reduced robustness, is linked to faster gradient descent convergence. Furthermore, by modeling attention as a system of interacting particles, we explore how the curvature distribution influences robustness. Visually summarized in Figure 1, this dual perspective provides a clear understanding of how the attention map geometry affects performance and robustness. We propose a regularization method informed by our findings to adjust the curvature distribution. Supporting our theory, our experiments show that increasing variance accelerates optimization, while reducing it improves robustness at the cost of generalizability, highlighting a trade-off.

Main contributions. This work establishes a theoretical connection between Ricci curvature of attention graphs and gradient descent convergence, showing that lower curvature accelerates training. We also reveal how the curvature distribution impacts robustness, particularly in response to input perturbations. Additionally, we introduce a method for curvature adjustment to balance performance and generalizability, supported by empirical results. By linking the geometry of attention to transformer behavior, this work offers new insights into improving training, performance, and robustness using geometric tools.

2. Background

Transformers and attention. The attention mechanism enables transformers to capture dependencies between tokens (Vaswani et al., 2017). Attention scores are computed from the query (Q) and key (K) representations, multiplied by the value matrix (V):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{1}$$

The attention matrix $A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$, can be viewed as the adjacency matrix of a graph with tokens as nodes and edge weights as attention scores.

Ricci curvature and system robustness. Ricci curvature captures the extent to which the local geometry induced by a Riemannian metric deviates from Euclidean space (Bauer et al., 2017), and has been linked to system robustness (Pouryahya et al., 2017), due to its connection with entropy. There exists a correlation between entropy and *system robustness*, which refers to the ability of a system to quickly return to its stationary state after a perturbation (Demetrius and Manke, 2005). The Fluctuation Theorem (Evans et al., 1993) establishes a positive correlation between robustness and entropy, implying a positive correlation with curvature (Pouryahya et al., 2017). Extended to graphs, we use Ollivier-Ricci curvature (ORC) (Ollivier, 2009), defined and explained in Appendix A.

3. Related work

Related work is discussed throughout the paper, with a full literature review in Appendix B. To our knowledge, the link we establish between attention map geometry, training dynamics, performance, and robustness in transformers is novel.

4. Graph geometry of learning representations in transformers

Attention mechanisms can be viewed as graph operators, with tokens as nodes and attention scores as edge weights. The geometric properties of this graph, particularly its spectrum, shape the representations learned by transformers. We show that lower attention Ricci curvature, reflecting lower system robustness, leads to faster convergence during training, while higher curvature enhances robustness to input perturbations. We adopt two complementary perspectives: a mathematical view linking ORC to gradient descent convergence, and a physics-based view treating transformers as interacting particle systems. The former explains how lower ORC accelerates convergence, while the latter offers insights into robustness. These perspectives together provide a comprehensive view of how attention geometry influences transformer behavior. For detailed proofs and discussions, see appendices C and D.

Geometry of attention and training transformers. Transformer models are typically over-parameterized (Fan et al., 2019; Liu et al., 2021; Wang and Tu, 2020), resulting in a non-convex optimization landscape (Bassily et al., 2018; Choromanska et al., 2015; Liu et al., 2022). Fast convergence in such settings relies on satisfying the Polyak-Łojasiewicz (PL) condition (Bassily et al., 2018; Polyak, 1963). We demonstrate that the probability of satisfying this condition is linked to the eigenvalues of the attention graph, which are linked to the ORC. Previous studies on neural tangent kernels show that the loss function in training over-parameterized models satisfies the PL condition if the minimum eigenvalue of the tangent kernel exceeds a threshold (Liu et al., 2022). In Lemma 1, we show that the probability p of exponential gradient descent convergence is positively correlated with the minimum eigenvalue of the attention matrix $\lambda^{A}{}_{1}$.

Lemma 1 $\frac{\partial p}{\partial \lambda^{A}{}_{1}} \geq 0$*, i.e., the minimum attention eigenvalue is positively correlated with* p*.*

Using this lemma, we prove Theorem 2, linking the learning behavior of transformers to the ORC, given its connection to the spectrum of the attention graph (Bauer et al., 2011). Proofs, assumptions, and more details are provided in Appendix C.1.

Theorem 2 *The probability* p *of exponential convergence of gradient descent on transformers is negatively correlated with the minimum ORC* k*.*

Graph geometry, statistical mechanics, and robustness of transformers. Extending our geometric analysis, we interpret the attention graph as a system of interacting particles, applying principles from statistical mechanics to study transformer robustness. This robustness, linked to system entropy and Ricci curvature (Demetrius and Manke, 2005; Pouryahya et al., 2017), helps maintain performance under noisy inputs. We observe a shift in the ORC distribution toward more positive values during a forward pass (Appendix Figure 5), which suggests a connection between attention dynamics and its geometric properties that influence transformer robustness. Ricci curvature thus serves as a geometric indicator of both training dynamics and robustness, offering insights for model improvement. Our empirical results in Section 6 further elucidate these diverse roles and their practical implications.

5. Training curvature-adjusted transformers

Section 4 discusses how curvature influences transformer behavior, suggesting that manipulating the curvature distribution of attention graphs can adjust their properties. While the Ricci flow (Hamilton, 1988) can flatten graphs, it is computationally expensive to apply during training (Jin et al., 2008; Ni et al., 2019). Instead, we propose an efficient method to control the ORC distribution through regularization allowing us to empirically validate our theory and improve transformers training and robustness. Experiments are conducted using *L-ViT* and BERT-Tiny models, described in Appendix E. We propose a regularization method, detailed in Appendix F, which adds a variance-dependent term to the loss function to adjust the ORC distribution of attention graphs. As we explain in Appendix A, increasing variance promotes smaller ORC values and decreasing variance eliminates them (appendix Figure 6).

Shown in figures 2 and 6, our experiments confirm that increasing variance leads to a lower ORC and accelerates loss minimization, validating Theorem 2. In addition to supporting our theory, this method allows us to enhance transformer training or robustness. Further results are included in Appendix F. Other results in Appendix G also show a trade-off between performance and robustness, where decreasing variance leads to better performance but decreases generalizability, while increasing variance has the opposite effect.

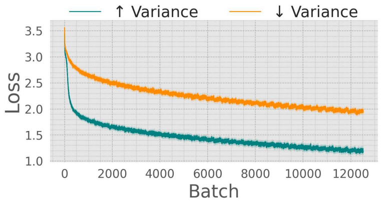

Figure 2: Training loss minimization on $L\text{-}ViT$ for CIFAR10, with regularization methods increasing (↑) and decreasing (↓) variance. Increasing variance promotes faster loss reduction and low-ORC edges.

6. Robustness and curvature distribution in transformers

Building on our particle system interpretation of attention and the established link between ORC and system robustness (Bauer et al., 2017; Ollivier, 2009), we empirically explore the relationship between curvature and transformer robustness. Sections 2 and 4 suggest that positive ORC values enhance robustness, which we empirically confirm by investigating pre-trained models' robustness to input perturbation (Appendix H) and the impact of ORC on model forgetfulness (Appendix I.1). Our results show that more robust models consistently exhibit higher frequencies of larger ORC values. For example, ViT shows $\sim 20\%$ less accuracy drop on perturbed MNIST inputs compared to DeiT, with similar improvements of $\sim 14\%$ and $\sim 62\%$ on Fashion-MNIST and CIFAR10. Similarly, for language models, ELECTRA-Small is $\sim 17\%$ to $\sim 49\%$ more robust than BERT-Tiny across datasets. The more robust models in each pair consistently display more positive-leaning ORC distributions (Figure 3), confirming the anticipated correlation. Details of these experiments are provided in Appendix H, with additional experiments in Appendix I supporting these findings.

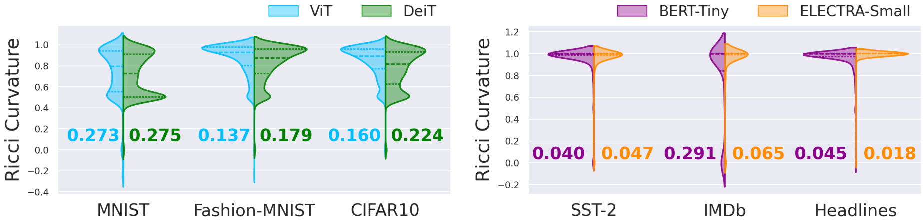

Figure 3: ORC distributions in the last attention layers of ViT, DeiT, BERT-Tiny, and ELECTRA-Small for batches of image and text datasets. ViT (blue) and ELECTRA-Small (orange) show more positive curvature than DeiT (green) and BERT-Tiny (purple), quantified by the Wasserstein distances from a Dirac mass at 1 (smaller is more positive leaning).

7. Conclusion

We analyzed transformers through graph geometry, showing that the Ricci curvature of attention maps directly influences both training dynamics and robustness. Lower curvature accelerates convergence, while more positive curvature enhances robustness, highlighting a trade-off. We proposed a curvature-adjusted regularization method to validate our theory and improve training or generalizability. By linking attention geometry to transformer behavior, this work offers new tools for improving their training, performance, and robustness.

References

Emmanuel Abbe, Samy Bengio, Enric Boix-Adsera, Etai Littwin, and Joshua Susskind. Transformers learn through gradual rank increase. Advances in Neural Information Processing Systems, 36, 2024.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. arXiv preprint arXiv:1811.02564, 2018.

Frank Bauer, Jürgen Jost, and Shiping Liu. Ollivier-Ricci curvature and the spectrum of the normalized graph laplace operator. arXiv preprint arXiv:1105.3803, 2011.

Frank Bauer, Bobo Hua, Jürgen Jost, Shiping Liu, and Guofang Wang. The geometric meaning of curvature: Local and nonlocal aspects of Ricci curvature. Modern Approaches to Discrete Curvature, pages 1–62, 2017.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10231–10241, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33: 1877–1901, 2020.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in Neural Information Processing Systems, 31, 2018.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In International Conference on Artificial Intelligence and Statistics, pages 192–204. PMLR, 2015.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In International Conference on Learning Representations, 2020.

Lloyd Demetrius and Thomas Manke. Robustness and network evolution—an entropic principle. Physica A: Statistical Mechanics and its Applications, 346(3-4):682–696, 2005.

Li Deng. The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6):141–142, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, and Michael M Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In International Conference on Machine Learning, pages 7865–7885. PMLR, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: transformers for image recognition at scale. In International Conference on Learning Representations, 2020.

Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. arXiv preprint arXiv:2401.04301, 2024.

Denis J Evans, Ezechiel Godert David Cohen, and Gary P Morriss. Probability of second law violations in shearing steady states. Physical Review Letters, 71(15):2401, 1993.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In International Conference on Learning Representations, 2019.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. arXiv preprint arXiv:2312.10794, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

Richard S Hamilton. The Ricci flow on surfaces, mathematics and general relativity. Contemporary Mathematics, 71:237–261, 1988.

Pierre-Emmanuel Jabin and Sebastien Motsch. Clustering and asymptotic behavior in opinion formation. Journal of Differential Equations, 257(11):4165–4187, 2014.

Prateek Jain, Vivek Kulkarni, Abhradeep Thakurta, and Oliver Williams. To drop or not to drop: Robustness, consistency and differential privacy properties of dropout. arXiv preprint arXiv:1503.02031, 2015.

Miao Jin, Junho Kim, Feng Luo, and Xianfeng Gu. Discrete surface Ricci flow. IEEE Transactions on Visualization and Computer Graphics, 14(5):1030–1043, 2008.

Jürgen Jost and Shiping Liu. Ollivier's Ricci curvature, local clustering and curvature-dimension inequalities on graphs. Discrete & Computational Geometry, 51(2):300–322, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. Communications in Difference Equations, 2000:227–236, 2000.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In International Conference on Learning Representations, 2023.

Yong Lin, Linyuan Lu, and Shing-Tung Yau. Ricci curvature of graphs. Tohoku Mathematical Journal, 63(4):605–627, 2011.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. Applied and Computational Harmonic Analysis, 59:85–116, 2022.

Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Do we actually need dense over-parameterization? in-time over-parameterization in sparse training. In International Conference on Machine Learning, pages 6989–7000. PMLR, 2021.

Yang Liu, Chuan Zhou, Shirui Pan, Jia Wu, Zhao Li, Hongyang Chen, and Peng Zhang. Curvdrop: A Ricci curvature based approach to prevent graph neural networks from over-smoothing and over-squashing. In Web Conference, pages 221–230. Association for Computing Machinery, 2023.

John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. Annals of Mathematics, pages 903–991, 2009.

Jian Luo, Jianzong Wang, Ning Cheng, and Jing Xiao. Dropout regularization for self-supervised learning of transformer encoder speech representation. arXiv preprint arXiv:2107.04227, 2021.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, 2011.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In Proceedings of the 18th ACM International Conference on Multimedia, pages 1485–1488, 2010.

Rishabh Misra and Prahal Arora. Sarcasm detection using news headlines dataset. AI Open, 4:13–18, 2023. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2023.01.001. URL https://www.sciencedirect.com/science/article/pii/S2666651023000013.

Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. SIAM Review, 56(4):577–621, 2014.

Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with Ricci flow. Scientific Reports, 9(1):1–12, 2019.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. Journal of Functional Analysis, 256(3):810–864, 2009.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 2019.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 3(4):643–653, 1963.

Maryam Pouryahya, James Mathews, and Allen Tannenbaum. Comparing three notions of discrete ricci curvature on biological networks. arXiv preprint arXiv:1712.02943, 2017.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. arXiv preprint arXiv:2002.10716, 2020.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. arXiv preprint arXiv:2103.15670, 2021.

Jayson Sia, Edmond Jonckheere, and Paul Bogdan. Ollivier-Ricci curvature-based method to community detection in complex networks. Scientific Reports, 9(1):9800, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, 2013.

Joshua Southern, Jeremy Wayland, Michael M. Bronstein, and Bastian Rieck. Curvature filtrations for graph generative model evaluation. In Advances on Neural Information Processing Systems, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In International Conference on Learning Representations, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In International Conference on Machine Learning, pages 35151–35174. PMLR, 2023.

Phil Wang. vit-pytorch. https://github.com/lucidrains/vit-pytorch, 2020.

Wenxuan Wang and Zhaopeng Tu. Rethinking the value of transformer components. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6019–6029, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45. Association for Computational Linguistics, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Huan Xu and Shie Mannor. The robustness-performance tradeoff in markov decision processes. Advances in Neural Information Processing Systems, 19, 2006.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. Advances in Neural Information Processing Systems, 33:8588–8601, 2020.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33:17283–17297, 2020.

Lin Zehui, Pengfei Liu, Luyao Huang, Junkun Chen, Xipeng Qiu, and Xuanjing Huang. Dropattention: A regularization method for fully-connected self-attention networks. arXiv preprint arXiv:1907.11065, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, pages 7472–7482. PMLR, 2019.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In International Conference on Learning Representations, 2024.

Appendix

Appendix A. Graph Ricci curvature and weight distributions

Extending Ricci curvature to graphs, the Ollivier-Ricci curvature (ORC) for any given edge in the graph depends on the geodesic distances of nodes in the neighborhood of the edge. Figure 4 visualizes positive, negative, and near-zero curvature edges in a weighted complete graph, relevant to full-attention, which is studied in this work. Specifically, ORC is defined as

$$\kappa_{OR}(v,u) \coloneqq 1 - \frac{W_1(\mu_v,\mu_u)}{d_G(v,u)}, \tag{2}$$

for an edge $(v,u) \in E$ on a graph $G=(V,E)$, where μ_v and μ_u are probability measure on the nodes anchoring (v,u), $d_G(.)$ represents a distance metric on V, and W_1 denotes the 1-Wasserstein distance (Lin et al., 2011; Jost and Liu, 2014). To better explain this point let us expand the Wasserstein distance in Equation 2 to write the ORC in the form

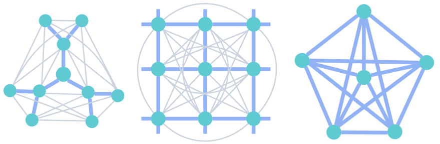

Figure 4: Three complete weighted graphs with highlighted larger weight edges (in blue). Different edge weight distributions result in negative (left), near zero (middle), and positive (right) Ricci curvature. Curvature in a weighted graph is negative/zero/positive when removing low-weight edges (in gray) yields a tree/grid/complete graph, respectively.

$$\kappa_{OR}(v,u) \coloneqq 1 - \frac{1}{d_G(v,u)} \inf_{\gamma\in\Gamma(\mu_v,\mu_u)} \int_{\mathcal{N}_v\times\mathcal{N}_u} d_G(v',u')\, d\gamma(v',u'), \tag{3}$$

where $d_G(\cdot,\cdot)$ is the geodesic distance over G, $\mathcal{N}_v$ denotes the neighborhood of node v, μ_v and μ_u are probability measures over $\mathcal{N}_v$ and $\mathcal{N}_u$ and $\Gamma(\mu_v,\mu_u)$ is the set of all couplings over them.

Curvature and clustering. Note that for any pair of nodes w and w', $d_G(w,w')$ is the weighted shortest path between w and w'. Hence, given two edges (v,u) and (v',u'), similarity of weight distributions over the neighborhoods of (v,u) and (v',u') leads to the similarity of geodesics in their neighborhoods, and considering the definition in Equation 3, this results in the proximity of the values of $\kappa_{OR}(v,u)$ and $\kappa_{OR}(v',u')$. Geshkovski et al. (2023) find that in a forward pass, attention maps exhibit clustering signaled by the cosine similarity of pairs of nodes, which corresponds to the similarity of weight distributions in their neighborhood. Given the discussion above, this in turn leads to a change in the ORC pattern, as a group of edges with similar ORC values form. This indeed matches the observation in Figure 5, showing the emergence of more concentrated modes in the distribution of ORC following attention dynamics through a forward pass.

Curvature and weight variance. Recall that the ORC definition provided in Equation 3 expands the Wasserstein distance from Equation 2, which is the cost of optimal transport of a mass distribution from $\mathcal{N}_v$ to $\mathcal{N}_u$. Recognizing this, we can observe that, fixing any (unweighted) connectivity structure, the higher the imbalance in the neighborhoods, the larger their distributional distances, which leads to smaller curvature values considering the definition of ORC. Note that the connectivity in full attention is trivially fixed, since the graph is always fully connected. It follows that the curvature in this case depends on the (im)balance in weight distributions. Therefore, it is the *variance* of weight distribution that determines the curvature values in attention maps. This intuition informs the regularization proposed in Section 5 and Appendix F for training curvature-adjusted transformer models.

Appendix B. Related work

In this work, we investigate the connections between the geometry of attention maps in transformers, their training, performance, and robustness properties. While related work in transformers, geometric deep learning, and robustness is referenced throughout the paper where relevant, we now provide additional discussion of the most related works in each area here.

Transformers and attention mechanism. Transformers (Vaswani et al., 2017) have revolutionized deep learning, achieving state-of-the-art performance across various domains (Brown et al., 2020; Devlin et al., 2018; Dosovitskiy et al., 2020). Recent works have begun to shed light on their inner workings, providing insights into their expressive power (Zaheer et al., 2020), generalization properties (Li et al., 2023), in-context learning (Von Oswald et al., 2023), learning dynamics (Abbe et al., 2024; Dovonon et al., 2024), and robustness characteristics (Bhojanapalli et al., 2021; Shao et al., 2021). However, many questions regarding the fundamental principles governing the behavior of transformers remain unanswered. Our work contributes to this growing body of research by investigating the connection between the geometry of attention maps and the properties of transformers.

Graph Ricci curvature. Ricci curvature captures the deviation of the local geometry from Euclidean space (Bauer et al., 2017). When extended to discrete structures like graphs, it has proven to be a powerful tool for various deep learning tasks (Di Giovanni et al., 2023; Liu et al., 2023; Southern et al., 2023; Topping et al., 2021). We employ the Ollivier-Ricci curvature (ORC) (Ollivier, 2009), an optimal transport formulation of Ricci curvature on graphs, to study the geometry of attention maps in transformers. To the best of our knowledge, this is the first work to explore this connection.

Curvature and robustness. Ricci curvature has been linked to the robustness of systems (Pouryahya et al., 2017), owing to its relationship with entropy. This connection is highlighted by a fundamental result from optimal transport theory (Lott and Villani, 2009), which provides bounds on entropy based on a lower bound of the Ricci curvature. The Fluctuation Theorem (Evans et al., 1993) further establishes a positive correlation between the robustness and entropy of a system, implying that robustness and curvature are positively correlated (Pouryahya et al., 2017). Leveraging this connection, we explore the implications of the ORC distribution for the robustness of vision and language transformers. In summary, our key contribution is uncovering insightful patterns in the curvature distribution of attention maps that evolve during training and through the layers of transformers, establishing a connection between the geometry of attention and the properties of these models. To the best of our knowledge, this is the first work to explore this connection, opening up new avenues for developing more robust and interpretable transformers.

Appendix C. Ricci curvature and training of transformers

In this appendix, we provide the full technical details and formal proofs regarding the connection between curvature and convergence of gradient descent in transformers, discussed in Section 4. The derivations and theoretical steps leading to the key results are elaborated here, complementing the informal discussion in Section 4. For completeness, we restate some of the results, which the technical discussion in this appendix builds up to. These results explain the relationship between ORC and gradient descent convergence in transformers.

C.1. Ricci curvature and convergence of gradient descent

In Section 4 we informally explained our main theoretical findings on the connection between the Ricci curvature of attention and gradient descent-based training of transformer models. Here we elaborate on the steps leading to this result, which rely on a connection we establish between the eigenvalue spectrum of transformers and optimization of over-parameterized models. Following standard practices in the theoretical analysis of transformers (see, e.g., Dovonon et al. (2024); Von Oswald et al. (2023); Abbe et al. (2024)), we derive our theoretical results in an idealized setting with a set of simplifying assumptions, stated here and in Appendix C.2. Our empirical investigation further demonstrates that the key implications of our theoretical findings extend beyond these idealized conditions, proving relevant and applicable to common transformer models in practice.

Eigenvalues of transformers. Consider a single-head transformer of depth L, with each block containing an attention layer, a skip-connection, and a projection. For this analysis, we assume the feed-forward layer in the attention block is a projection by $W_{p,l}$. Hence, the output of each layer is

$$X_{l+1} = X_l + A_l X_l W_{V,l} W_{p,l}. \tag{4}$$

Following Dovonon et al. (2024), we further assume that all attention blocks are identical and remove the layer indexing. Stacking the column to obtain a vectorized input, $\mathbf{x}$, we can write

$$F(\mathbf{x}) = \mathbf{x}_L = (I + H \otimes A)^L \mathbf{x}, \tag{5}$$

$$\lambda^F{}_{ij} = \left(1 + \lambda^A{}_i \lambda^H{}_j\right)^L, \tag{6}$$

where $F(\cdot)$ and $\lambda^F{}_{ij}$ denote the transformer and its eigenvalues, $H \coloneqq {W_p}^T {W_V}^T$, and $\lambda^A{}_i$ and $\lambda^H{}_j$ are the eigenvalues of A and H. Note that relaxing the above-mentioned assumptions is rather straightforward, though distracting and not amenable to the here provided analysis.

Tangent kernel and convergence of gradient descent. Transformers are often over-parameterized (Fan et al., 2019; Liu et al., 2021; Wang and Tu, 2020) and hence essentially non-convex (Bassily et al., 2018; Choromanska et al., 2015). The Polyak-Łojasiewicz (PL) condition (Bassily et al., 2018; Polyak, 1963) implies exponential convergence of gradient descent in non-convex regimes (Liu et al., 2022). Given a model $f_\theta(\cdot)$ parameterized by θ, we say the loss function $\mathcal{L}(\theta)$ is μ-PL if $\|\nabla \mathcal{L}(\theta)\|^2 \geq \mu L(\theta)$. Let $K_\theta \coloneqq (\nabla f_\theta)(\nabla f_\theta)^T$ and $\lambda_1(K_\theta)$ be the tangent kernel of f and its minimum eigenvalue. For any region on the loss manifold $\mathcal{S}$, the following holds.

Theorem 3 (Liu et al. (2022)) *$\mathcal{L}(\theta)$ is μ-PL if $\lambda_1(K_\theta) \geq \mu$ for all $\theta \in \mathcal{S}$.*

It follows that the learning behavior of transformers depends on the minimum eigenvalue of their tangent kernel. This suggests a connection between ORC and the training of transformers, given the link between ORC and the eigenvalue spectrum. We specify this connection next.

Ricci curvature and convergence of gradient descent. Given the tangent kernel of the transformer, $\varphi \coloneqq (\nabla F)(\nabla F)^T$, from equations 5 and 6, we can compute the eigenvalues of φ. Thus, using the findings from (Dovonon et al., 2024) on ${\lambda^A}_i$ and ${\lambda^H}_j$ and in light of the implications of Theorem 3, under a set of standard assumptions (Appendix C.2), we prove the following lemma.

Lemma 4 *Let p denote the probability of exponential convergence of gradient descent on the transformer. Then, $\frac{\partial p}{\partial {\lambda^A}_1} \geq 0$, i.e., the minimum attention eigenvalue is positively correlated with p.*

Meanwhile, ORC bounds the spectrum of the normalized Laplacian (Bauer et al., 2011) (see Appendix C.2). Since the attention scores are outputs of the softmax activation, as we describe in Appendix C.2, there is a simple connection between the eigenvalues of A and its Laplacian, which in turn implies that k controls the spectrum of A. Theorem 2 then follows from Lemma 4, showing a negative correlation between the minimum ORC value of the attention matrix and p. This theorem suggests the ORC distribution bears implications for training transformers via gradient descent-based algorithms. Next, we propose a novel training heuristic to manipulate the ORC distribution in transformers, and we empirically validate the theoretical implications of Theorem 2.

C.2. Assumptions and proofs of results

We establish a theoretical link between the ORC of the attention map and the convergence of gradient descent on transformers. As we discuss in Section C.1, under a set of assumptions, this connection follows from the learning theory of non-convex optimization of over-parameterized neural networks, the eigenvalue spectrum of transformers, and the ORC bound on the graph spectrum. Our Theorem 2, which follows from Lemma 4, states that the minimum ORC of A is positively correlated with the probability of exponential convergence of gradient descent on F. In this appendix, we specify the assumptions leading to these lemma and theorem and provide a proof.

Assumptions The majority of our assumptions follow the setup and assumptions in Dovonon et al. (2024). We however introduce two other assumptions that we discuss further in this section. All assumptions are listed below.

A0. The transformer uses a self-attention mechanism, and the keys and queries share a projection matrix, i.e. $K = Q$.

A1. A is positive and invertible, and all its eigenvalues are real.

A2. Eigenvalues of H are positive and bounded, i.e., $0 < {\lambda^H}_1 < \ldots < {\lambda^H}_d < \infty$

A3. $\lambda^{H}{}_{d} < |\lambda^{A}{}_{1}|^{-1}$ where $\lambda^{A}{}_{1}$ is the smallest eigenvalue of A.

A4. Let $M \coloneqq I + H \otimes A$, and let θ be the vector of the weights in F, consisting of the entries of the weight matrices W_Q, W_K, W_V, and W_p. Then $\nabla_\theta M$ and M are simultanously diagonalizable by an orthogonal matrix P, i.e., there exists an orthogonal matrix P and diagonal matrices D and $\tilde{D}$, such that $M = P^{-1}DP$ and $\nabla_\theta M = P^{-1}\tilde{D}P$. Furthermore, we assume $\tilde{D}$ is approximately constant with respect to D.

A5. The minimum eigenvalue of A is negative.

Note that assumptions **A1** and **A2** follow the setup by Dovonon et al. (2024), and imply

$$-1 < \lambda^{A}{}_{1} \leq \ldots \leq \lambda^{A}{}_{n} = 1. \tag{7}$$

Proofs. Considering that $F(\mathbf{x}) = (I + H \otimes A)^L \mathbf{x}$, we can write $F(\mathbf{x})$ as a linear operator $F(\mathbf{x}) = F\mathbf{x}$ where $F = (I + H \otimes A)^L$. Hence, the eigenvalues of F are $\left(1 + \lambda^{A}{}_{i}\lambda^{H}{}_{j}\right)^L$. Let $\varphi \coloneqq (\nabla_\theta F)(\nabla_\theta F)^T$ be the tangent kernel of F. We can write $\varphi(\mathbf{x}) = \nabla_\theta M^T BB^T \nabla_\theta M$ where $M \coloneqq I + H \otimes A$ and $B \coloneqq M^{L-1}$. By assumption **A4**, we can write $\varphi(\mathbf{x}) = P^{-1}\tilde{D}D^{2L-2}\tilde{D}P$. Since diagonal matrices are commutative, it follows that the eigenvalues of φ are of the form

$$\lambda_m^\varphi = \eta_m^2 \left(1 + \lambda_i^A \lambda_j^H\right)^{2L-2}, \tag{8}$$

where η_m is a diagonal entry of $\tilde{D}$. On the other hand, it follows from **A2**, **A4**, and **A5** that the minimum eigenvalue of φ is $\lambda_1^\varphi = \eta_m(1 + \lambda^{A}{}_{1}\lambda^{H}{}_{d})^{2L-2}$ for some m, since by **A2** and **A4** we know $0 < 1 + \lambda^{A}{}_{1}\lambda^{H}{}_{j} < 1$ and achieves its minimum values for $j = d$. Given μ, let p be the probability that F is $\mu-$PL, i.e., the PL theorem holds as described in Appendix C.1. From Theorem 3, we have

$$p = \mathbb{P}\left[\lambda_1^\varphi \geq \mu\right]. \tag{9}$$

Note that we can equivalently write $p = \mathbb{E}\left[\mathbf{1}_{\lambda_\infty^\varphi > \mu}\right]$, where $\mathbf{1}$. denotes the indicator function. Thus, we can write,

$$\begin{aligned}
\frac{\partial p}{\partial \lambda^{A}{}_{1}} &= \frac{\partial p}{\partial \lambda_1^\varphi}\frac{\partial \lambda_1^\varphi}{\partial \lambda^{A}{}_{1}} \\
&= \mathbf{1}_{\lambda_\infty^\varphi > \mu}\left(2L - 3\right)\eta_1^2\left(1 + \lambda^{A}{}_{1}\lambda^{H}{}_{d}\right)^{2L-3} \\
&\geq 0,
\end{aligned}$$

where the last line follows since all terms on the right-hand side of the previous line are non-negative for $L > 1$. This completes the proof of Lemma 4. It follows that,

$$\Delta p \times \Delta\lambda^{A}{}_{1} \geq 0, \tag{10}$$

which establishes a positive correlation between p and $\lambda^{A}{}_{1}$. Meanwhile Bauer et al. (2011) show that $\lambda^{L}{}_{1}$ is bounded below by the minimum ORC, where $\lambda^{L}{}_{1}$ denotes the minimum eigenvalue of the normalized Laplacian of A. Notice that A is row stochastic (Dovonon et al., 2024), which implies that its normalized Laplacian is in fact $I - A$. Therefore, $\lambda^{L}{}_{1} = 1 - \lambda^{A}{}_{1}$, which means that the minimum ORC also bounds $1 - \lambda^{A}{}_{1}$ from below. This in turn, establishes a negative correlation between minimum ORC and $\lambda^{A}{}_{1}$. Considering Equation 10, it follows immediately that minimum ORC is negatively correlated with p, completing the proof of our Theorem 2. □

Discussion on the assumptions. The self-attention mechanism is in fact a common setup for transformers, hence the only constraint assumption **A0** imposes beyond common practice is sharing the projection matrix between keys and queries. This yields a symmetric attention, while restricting the keys and queries projections to the column space of a shared weight matrix. Assumptions **A3** and **A5** are purely technical assumptions for our proof. Assumption **A4** on the other hand, though plausible, has more consequences. First, note that the order in which the weight matrices are combined to parameterize F with θ can be completely flexible. Therefore, it suffices if assumption **A4** holds for any arrangement of the weights in θ. Despite this flexibility, being simultaneous diagonalizable restricts the degrees of freedom in M and $\nabla_\theta M$. This assumption implies that the two matrices share an eigenstructure, though they could have different eigenvalues. From a geometric perspective, this assumption limits the degrees of freedom that the changes in M have at each gradient step to the directions of its eigenvectors. While this is a restrictive assumption, we should keep in mind that transformers are often over-parameterized with a high degree of freedom, and hence the assumption is still plausible. Furthermore, we assume **A4** for technical reasons in order to simplify the proof of the theorem, and otherwise, the assumption is not a conceptual necessity. In other words, while assuming **A4** allows us to state our proof, it is only a tool to simplify a step of the proof, not the only tool to show the main conclusion, and the implications may still hold without it.

Appendix D. Graph geometry and statistical mechanics of transformers

In Section 4 we briefly discussed the value of looking at transformers from the angle of networks/graphs and its consequences for training and robustness of transformers. The mathematical implications of this perspective for training transformers are provided in Appendix C.1. Continuing in this direction from the perspective of network dynamics, in this Appendix we elaborate on the physics that follow as an input signal is processed by a series of attention blocks (graph operators), which enables us to analyze the robustness properties of transformers from a networked system point of view. As we mentioned in Section 4, we interpret attention as a network of interacting particles and study the statistical mechanics of transformers in light of established results on the interplay between graph geometry and the dynamics and robustness of networked systems.

Attention as a particle system and transformer dynamics. A forward pass through attention blocks in transformers can be interpreted as dynamics of a system of interacting particles where each attention score represents the strength of the interaction between a pair of particles (tokens). Leveraging the mathematical framework of Geshkovski et al. (2023), in the continuous approximation of a forward pass through deep residual network layers as evolution through time (Chen et al., 2018), the dynamics of the representations for token i, z_i can be modeled as

$$\dot{z}_i(t) = \mathbf{P}_{z_i(t)} \left[\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} \exp\left\{ \beta \left[W_Q(t) z_i(t) \right]^T W_K(t) z_j(t) \right\} W_V(t) z_j(t) \right], \qquad (11)$$

where, $Z_{\beta,i}(t)$ is the partition function, β the system's *temperature*, and for d-dimensional z and u, $\mathbf{P}_z u = u - z^T u z$ is the projection of u onto the tangent space of a $(d-1)$-dimensional sphere in R^d at z (Geshkovski et al., 2023). The similarity between particle systems and attention mechanisms is evident from comparing equations 1 and 11. Thus, the output of the simplified transformer studied here can be approximated as the solution of Equation 11 as $t \to +\infty$. This framework sets the stage for the subsequent empirical investigation into how ORC distribution informs the behavior of transformers and their robustness properties in light of the discussion in Section 2.

Forward pass and attention clustering. Geshkovski et al. (2023) demonstrate that when the query, key, and value projection matrices are the identity, the dynamics modeled by this equation are similar to the Krause model (Krause et al., 2000), which shows particles clustering when $t \to +\infty$ (Jabin and Motsch, 2014). Clustering behavior is also observed in other similar models (Geshkovski et al., 2023; Motsch and Tadmor, 2014). Despite the differences between transformer models used in practice and the idealized model commonly assumed for analysis, the interpretation above leads to questions on attention clustering in transformers. Geshkovski et al. (2023) show empirical evidence that in commonly used pre-trained transformers the attention map becomes more clustered through a forward pass. Given this connection between ORC and clustering in networks (Jost and Liu, 2014; Sia et al., 2019), which we expand upon in Appendix A, the forward pass dynamics of attention also corresponds to an evolving pattern of ORC distribution.

Attention dynamics and Ricci curvature. Computing the ORC distribution in pre-trained and fine-tuned vision and language transformers, our experiments confirm our theoretically anticipated connection between attention ORC distributions and forward pass dynamics in transformers. We empirical observations reveal that a forward pass in transformers leads to a shift in the ORC distribution towards more positive values, or significantly reduces the frequency of smaller ORC values. This is visualized for two vision and two language transformers with six standard datasets in Figure 5, and quantified by the Wasserstein distance of the distribution from a Dirac mass at 1 – maximum ORC value. A smaller distance indicates a more positive-leaning distribution in the final attention layer compared to the first layer, as observed in most cases. In the only exceptions where this Wasserstein distance value is greater – DeiT model on CIFAR10 and BERT-Tiny on IMDb – we still observe that the forward pass eliminates the smaller peak in the ORC distribution, shifting the smaller values upwards. As we explain in Section 2, ORC is a known indicator of system robustness. Meanwhile, the forward pass in transformers is known to play an important role in their remarkable performance, for instance, in their in-context learning (Von Oswald et al., 2023). Hence, these evolving patterns of ORC which follow transformers' dynamics reveal evidence on the geometric signature of both the robustness and performance of transformers.

Appendix E. Implementation details

E.1. *L-ViT* configuration

The *L-ViT* model utilizes a transformer architecture for image classification borrowed from ViT-PyTorch library (Wang, 2020). The input image is divided into 7x7 patches for the MNIST and Fashion-MNIST datasets, and 8x8 patches for the CIFAR10 dataset. These are projected into a 64-dimensional space within a single channel for the MNIST and Fashion-MNIST datasets, and 3 channels for the CIFAR10 dataset. An 8-layer deep transformer

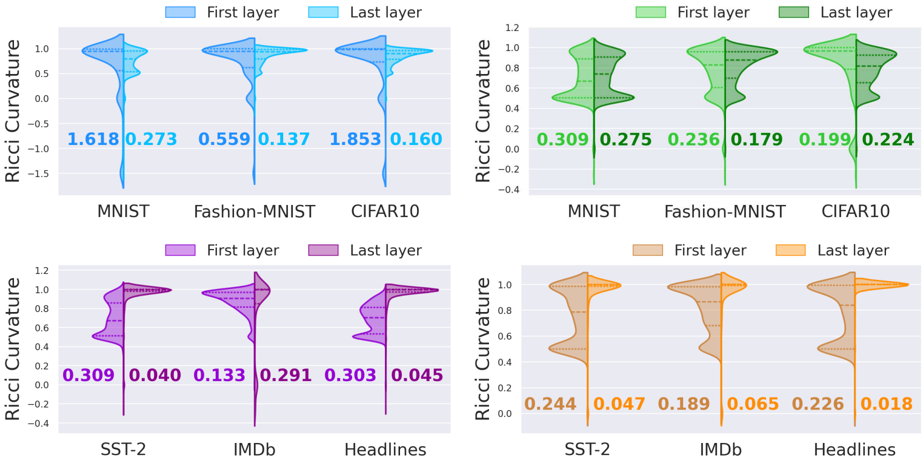

Figure 5: The ORC distributions of the first and last attention layers of ViT (blue) and DeiT (green) for a batch of MNIST, Fashion-MNIST, and CIFAR10 images, and of BERT-Tiny (purple) and ELECTRA-Small (orange) for a batch of SST-2, IMDb, and Headlines data. The curvature distribution shifts towards more positive values through a forward pass; quantified by the Wasserstein distances of the ORC distributions from a Dirac mass at the ORC value of 1, annotated next to each distribution (smaller means more positive leaning).

encoder with a single attention head per layer processes these patch embeddings. The MLP block within each layer has a hidden dimension of 128. Dropout with a rate of 0.1 is applied both to activation functions and embeddings during training to enhance model robustness and prevent overfitting.

E.2. Training and evaluation details

We use a learning rate of $5 \cdot 10^{-4}$ for training $L\text{-}ViT$ with a batch-size of 64 images, and a learning rate of $5 \cdot 10^{-6}$ for training BERT-Tiny with a batch-size of 16 sentences. The pre-trained model weights were downloaded from Huggingface Transformers (Wolf et al., 2020). For fine-tuning pre-trained vision models, we use a learning rate of $2.5 \cdot 10^{-4}$ with a batch-size of 64 images, and for pre-trained language models, these values are $2 \cdot 10^{-5}$ and 16.

The perturbations for experiments in Section 6 include flipping, blurring, and rotation of images with probabilities 0.2 and 0.4, and word swap in texts with probability 0.4. For the experiments in Section 5 and Appendix F, we trained $L\text{-}ViT$ for 20 epochs on MNIST and Fashion-MNIST images and 40 epochs on CIFAR10 images, and same experiments on BERT-Tiny (in Appendix F) were performed with 20 epochs for all three datasets. For the $L\text{-}ViT$ experiments in Appendix I.3, models were trained with an early stopping condition at accuracies of 97%, 87%, and 50%, over MNIST, Fashion-MNIST, and CIFAR10 validation sets, respectively. BERT-Tiny models mentioned in Appendices I.2 and I.3 were trained with an early stopping condition at accuracies of 88%, 84%, and 79%, on the IMDb, Headlines, and SST-2 validation sets, respectively. In Appendix I.3 experiments, we perturb the test set images by applying elastic transformation with $\alpha = 75$ and random perspective transformation with a distortion scale of 0.5, with a probability of $p = 0.5$. Text test set sentences are perturbed via characters or complete word substitution. These substitutions were chosen randomly, potentially resulting in nonsensical combinations. This process aimed to introduce noise and potentially disrupt the inherent coherence and structure of the sentences. Both character-level and word-level perturbations were applied independently with a fixed perturbation rate of $r = 0.2$ for each.

Consistent subset sizes were employed for all experiments involving ORC computation over subsets. Specifically, a subset of 1,024 images was utilized for image datasets tested with the L-ViT model. A subset of 16 sentences was employed for text datasets tested with the BERT-Tiny model.

E.3. Other implementation details and system specification

All ORC computations in this study were conducted using the GraphRicciCurvature (Ni et al., 2019) Python library.

All loss optimizations were performed via Adam (Kingma and Ba, 2014), and the training and testing implementations use the PyTorch library (Paszke et al., 2019), with the default data split from the Torchvision package (Marcel and Rodriguez, 2010). While visualizing distributions, outliers and zeros were removed as needed for clarity of the distributions in the figures. The system specification for the computations is provided in Table 1.

Table 1: System specifications for the computations.

CPU	Intel(R) Xeon(R) CPU @ 2.40GHz
GPU	Nvidia A100 SXM4 40GB
OS	AlmaLinux 9.3
Architecture	x86_64

Due to computational costs associated with the resource demands of graph ORC computations, evaluations of several experiments were conducted on randomly selected subsets of the MNIST, Fashion-MNIST, and CIFAR10 image datasets and the IMDb, Headlines, and SST-2 text datasets. Additionally, the pre-trained models were fine-tuned through 1 epoch on each dataset, which already yields over 90% accuracy.

Appendix F. Training curvature-adjusted transformers

In this appendix, we provide the detailed formulation and empirical results of the regularization method we propose, mentioned in Section 5, which allows us to adjust the curvature distribution of attention maps in transformers. The technical details, loss functions, and all experimental results used to validate our theory are described here. For completeness, key results and discussions in the main text are restated.

Section 4 discusses the novel connections we establish between the curvature of attention maps and various aspects of transformer behavior, such as learning dynamics and robustness. This raises an intriguing possibility: manipulating the curvature distribution of the attention graph could allow us to adjust their behavior and improve their capabilities. A well-established tool to diffuse the curvature over a manifold is the Ricci flow (Hamilton, 1988), and its discrete variation could be iteratively applied to flatten the graph and reduce the frequency of edges with highly negative curvature (representing unstable networks) (Jin et al., 2008; Ni et al., 2019). However, each iteration of the discrete Ricci flow requires computing the curvature distribution, and hence it is computationally taxing to incorporate in the training. To overcome this challenge via a computationally efficient proxy, we introduce a simple, efficient, and effective method to manipulate the ORC distribution of attention maps through regularization by a function of the variance of attention scores. Aiming to either promote or reduce the frequency of low-curvature edges, this regularization strategy not only provides a means to empirically validate our theory, but also offers a practical mechanism to enhance the training and robustness of transformers by adjusting curvature. The experiments included in this section use the *L-ViT* and BERT-Tiny models, described in Appendix E.

Encouraging or discouraging low-curvature edges via regularization Recognizing the correspondence between the weight distributions and ORC in a complete graph (of full attention), we propose a regularization method to control the attention variance to adjust the ORC distribution. Negative ORC corresponds to weight imbalance, while a weight distribution with a small variance corresponds to a positively-curved graph. Hence, we expect variance increase to promote smaller ORC values, while variance decrease will lead to their elimination. In Appendix A, we elaborate on how the formulation of the ORC implies its connection with weight variance. To put this in practice, we regularize the loss function with a variance-dependent term $h(\sigma^2)$, such that $\frac{\partial h}{\partial \sigma^2} < 0$ for promoting low-curvature edges and $\frac{\partial h}{\partial \sigma^2} > 0$ for reducing their frequency. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \gamma h(\sigma_A^2), \tag{12}$$

where $\mathcal{L}_{\text{task}}$ denotes the usual loss associated with the task (e.g., classification loss in our experiments), γ the regularization coefficient, and σ_A^2 the variance of attention scores, averaged over all layers. In our implementation, we use $h(x) = \exp(-ax + b)$ and $h(x) = \log(ax + b)$ to increase and decrease variance (respectively), with $a > 0$ and b serving as tunable scale and shift parameters, helping to place the regularization value in an impactful range. Furthermore, the choice of exponential and logarithmic functions ensures decreasing marginals, limiting adjustments to the loss and avoiding extreme outcomes where attention either becomes uniformly distributed (losing discriminative power) or overly imbalanced (leading to instability or absurd attention).

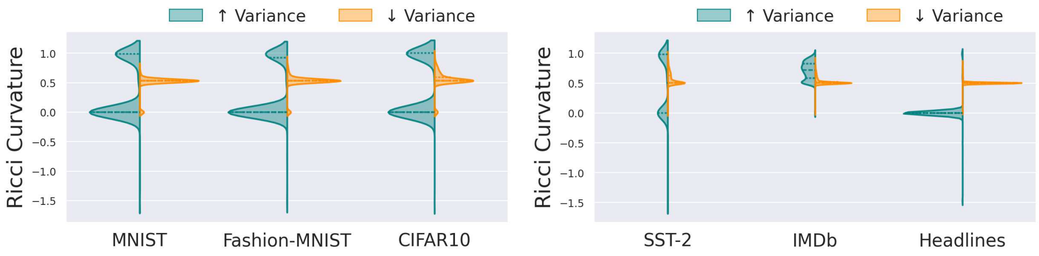

Figure 6: Final layer ORC distributions in *L-ViT* and BERT-Tiny trained via regularization methods increasing (↑) and decreasing (↓) variance, exhibiting a connection with the direction of variance regularization.

Empirical results on curvature-adjusted training. Our experiments on standard image and text datasets using *L-ViT* and BERT-Tiny confirm that our proposed regularization influences the ORC distribution as anticipated, illustrated in Figure 6. Equipped with this method, we now empirically validate Theorem 2 by comparing transformers with smaller and larger minimum values of attention ORC, corresponding to increasing and decreasing variance. Our experimental results in Figure 7 indicate that increasing variance not only accelerates loss minimization but also leads to earlier convergence. Note that we observe a larger loss variance across trials in language models compared to vision models, and due to computational constraints, reducing this variance is beyond the scope of this project, especially considering strong empirical validations already observed using *L-ViT*. Having said that, the average over 30 trials on BERT-Tiny is consistent with what we observe for *L-ViT*, though the confidence intervals overlap. Considering clearly significant trends observed with *L-ViT* and similar trends for the trial averages for the BERT-Tiny model, these results overall support the implications of our Theorem 2, which states that a smaller minimum ORC value corresponds to a higher probability of exponential convergence of gradient descent, suggesting faster convergence. This finding also suggests that our proposed regularization can enhance the training of transformer models. In Appendix G we further empirically demonstrate the impact of this curvature adjustment on transformers, effectively influencing the learning dynamics, performance, and robustness characteristics of the model.

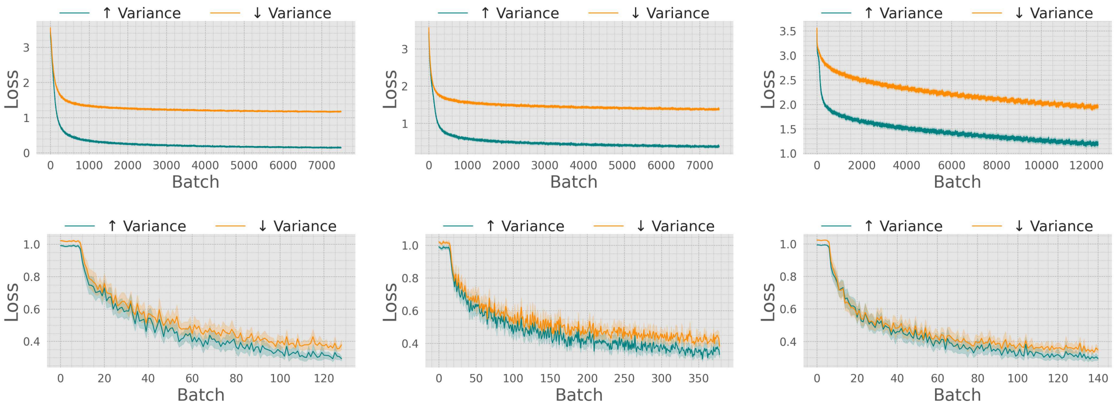

Figure 7: Training loss minimization on *L-ViT* (top) and BERT-Tiny (bottom) with regularization methods increasing (↑) and decreasing (↓) variance for MNIST (left), Fashion-MNIST (middle), and CIFAR10 (right). Higher variance can promote faster loss reduction and achieve convergence in fewer iterations.

Appendix G. Robustness and performance trade-off

In this appendix, we provide empirical results highlighting a trade-off between performance and robustness observed in transformers trained with our proposed variance-based regularization. We examine how increasing or decreasing the frequency of low-curvature edges affects training, accuracy, and generalizability, presenting further empirical results using our proposed regularization method discussed in Section 5 and Appendix F. Our findings highlight a potential contrast between the advantages of negative ORC for optimization and positive ORC for the robustness of transformers. The trade-off between robustness and performance is an established phenomenon studied in various contexts (Raghunathan et al., 2020; Xu and Mannor, 2006; Zhang et al., 2019). In light of the geometric underpinning of the attention mechanism unveiled in this work, here we further explore this trade-off in the context of transformers.

Table 2: Training and test accuracies and their gap using the regularization methods increasing and decreasing variance for six image and text datasets. The average and 95% confidence intervals over 30 trials are shown.

	Training accuracy (↑)		Test accuracy (↑)		Training accuracy - Test accuracy (↓)	
	Decrease variance	Increase variance	Decrease variance	Increase variance	Decrease variance	Increase variance
CIFAR10	**0.704 ± 0.001**	0.613 ± 0.009	**0.609 ± 0.002**	0.557 ± 0.005	0.095 ± 0.002	**0.056 ± 0.005**
Fashion-MNIST	**0.903 ± 0.000**	0.893 ± 0.001	**0.890 ± 0.001**	0.883 ± 0.001	0.013 ± 0.001	**0.010 ± 0.001**
MNIST	**0.981 ± 0.000**	0.980 ± 0.000	**0.984 ± 0.000**	0.982 ± 0.000	**-0.003 ± 0.001**	**-0.002 ± 0.000**
SST-2	**0.986 ± 0.001**	0.985 ± 0.001	0.826 ± 0.003	**0.831 ± 0.001**	0.160 ± 0.003	**0.154 ± 0.002**
IMDb	**0.994 ± 0.000**	0.993 ± 0.000	0.836 ± 0.003	**0.838 ± 0.002**	0.158 ± 0.003	**0.155 ± 0.002**
Headlines	0.965 ± 0.001	**0.967 ± 0.001**	**0.915 ± 0.002**	0.910 ± 0.002	**0.050 ± 0.002**	0.057 ± 0.001

Using the training techniques described in Section 5 and Appendix F, we observe divergent impacts on performance and robustness from the opposing directions of variance regularization. Increasing the frequency of low-curvature edges by increasing variance speeds up optimization. Meanwhile, when this regularization introduces a second mode of highly positive curvature values, which is observed on all image datasets and two of the three text datasets in our experiments (Figure 6), it also improves generalizability, as evidenced by smaller train-test performance gaps. Conversely, decreasing variance enhances accuracy in most cases, but reduces generalizability. These effects are evident from Table 2, which reports the train and test accuracy values, and the difference between the two. Note that our proposed regularization method globally impacts the attention scores and has limitations in targeted curvature adjustment, as, for instance, the simultaneous emergence of high and low curvature modes in some of our datasets suggests. While Table 2 reports the results of the regularization methods, together with the impact of each regularization on ORC distribution (figures 6), they allow us to understand the impact of ORC distribution on accuracy and generalization. In particular, the results highlight the trade-off between high performance and robustness across different distributions, corresponding directly to the curvature distribution of the attention graph.

Appendix H. Ricci curvature distribution and robustness to input perturbation

In this appendix, we provide detailed discussion and empirical results for the connection between the ORC of attention graphs and the robustness of transformers to input perturbation, discussed in Section 6. For completeness, we repeat key results from Section 6 to provide full context and ensure clarity in the interpretation of our findings. We conduct a series of experiments exploring the relationship between ORC and robustness in transformers. The discussions in sections 2 and 4 suggest that more positive ORC values, thereby more system robustness, should enhance the robustness of transformers. We empirically investigate this link through two focused studies: first, by examining the curvature and robustness of pre-trained transformers to input perturbation; and second, by assessing the impact of the ORC distribution on model forgetfulness. In line with the theoretical insight, our experiments confirm that more robust transformers feature a higher frequency of larger ORC values in their attention graphs. Additional experiments in Appendix I further support these findings.

Models and data. The experiments in this section and Appendix I are conducted on MNIST (Deng, 2012), Fashion-MNIST (Xiao et al., 2017), and CIFAR10 (Krizhevsky et al., 2009) images for the vision models, and on the IMDb Movie Reviews (Maas et al., 2011) (IMDb), Stanford Sentiment Treebank v2 (Socher et al., 2013) (SST-2) and the News Headlines Sarcasm Detection (Misra and Arora, 2023) (Headlines) datasets for the language models. We fine-tune pre-trained vision and language transformers on these datasets, comparing ViT (Dosovitskiy et al., 2020) against DeiT (Touvron et al., 2021) for vision, and BERT-Tiny (Turc et al., 2019) against ELECTRA-Small (Clark et al., 2020) for language models. All implementation details are reported in Appendix E.

Curvature and robustness of pre-trained transformers to input perturbation. We now compare the performance of fine-tuned models on perturbed and unperturbed test sets (further details in Appendix E). We quantify robustness by metric ρ.

$$\rho \coloneqq \text{accuracy}_{\text{unperturbed}} - \text{accuracy}_{\text{perturbed}}. \tag{13}$$

We find that more positive-leaning ORC distributions correspond to smaller ρ values, indicating more robust models. For instance, ViT yields a ρ value of 0.067 on MNIST images, which is ∼20% than the ρ value of 0.084 obtained from DeiT. Similarly, the ρ values for Fashion-MNIST and CIFAR10 datasets are ∼14% and ∼62% smaller in ViT than DeiT. Meanwhile, the ORC distribution in ViT is more positive-learning, Figure 3, as quantified by the Wasserstein distance of the distribution from a δ distribution at the ORC upper bound of 1. This pattern holds in language transformers as well, where ELECTRA-Small exhibits a more positive-leaning ORC distribution than BERT-Tiny, and is ∼ 17%, ∼ 49%, and ∼ 23% more robust to input perturbation on SST-2, IMDb, and Headlines datasets. Experiments on *L-ViT* and BERT-Tiny from random initialization (Appendix I.3) show a similar trend, where training a more robust model corresponds to eliminating negative ORC values leading to a distribution concentrated around a positive mode.

Appendix I. Additional experiments

I.1. Forgetfulness and curvature distribution

As a part of our empirical analysis of the robustness properties of transformers, in Section 6 we investigated the connection between ORC distribution and the robustness of transformers to input perturbation. Exploring robustness from a different perspective, here we look into *forgetfulness* of pre-trained transformers and their connection with the ORC distribution. Forgetfulness, where models lose previously learned information when exposed to new tasks or data, offers a complementary perspective on robustness, alongside robustness to input perturbation. These experiments further link ORC in attention graphs to transformer robustness, showing that models with more positive-leaning ORC distributions are less prone to forgetfulness.

We sequentially fine-tune each pre-trained model on a primary dataset and then a secondary one, tracking the performance decay on the primary dataset to measure how quickly each model "forgets" due to subsequent training. A fast decay in performance on the primary dataset implies a more *forgetful* model. The results reveal a significant drop in classification accuracy –from above 80% to below 50%– highlighting a pronounced forgetting effect. Specifically, our results indicate that DeiT and BERT-Tiny tend to forget learned patterns faster than ViT and ELECTRA-Small, signaling higher robustness of ViT and ELECTRA-Small to training on new distributions. Meanwhile, as shown in Figure 3, the ORC distributions of attention maps in ViT and ELECTRA-Small lean further towards more positive ORC values. Together, these observations support our arguments regarding the positive correlation between the ORC of attention maps and the robustness of transformers, as the more positive-leaning distribution belongs to the less forgetful model. These findings also align with our observations on perturbation robustness: ViT and ELECTRA-Small, with more positive-leaning ORC distributions and higher system robustness, show greater resistance to forgetting. In this appendix, we include the results of the forgetfulness experiment on all pairs of primary and secondary datasets. Figure 8 demonstrates these results for vision transformers, and Figure 9 for language transformers, both exhibiting the same general trend. It is worth noting that the forgetting phenomenon is not observed in some of our experiments on language models, which is likely due to semantic similarities between the tasks and datasets in hand.

I.2. Attention pruning and performance

Attention pruning and performance over image datasets. To better understand the role of ORC in transformer performance versus robustness, we investigate the impact of selective attention pruning based on ORC values. We independently prune attention edges with low and high ORC values across all layers of the trained $L\text{-}ViT$ during testing (details in Appendix E.2). This allows us to evaluate the dependence of model performance on attention edges characterized by different ORC values (see Figure 10). Pruning low-ORC edges results in a faster performance decay compared to pruning high-ORC edges, indicating that negative-ORC edges are more critical for model performance (Figure 10). Similar experiments with BERT-Tiny are presented in the following section. Additionally, our experiments in Appendix I.3 show that training a more robust transformer corresponds to concentrating ORC distribution around a positive value, eliminating negative edges. These observations further strengthen the evidence supporting the previously identified trade-off between model robustness and performance.

Attention pruning and performance over text datasets. Similar to the experiments using $L\text{-}ViT$, we conduct attention pruning experiments using BERT-Tiny over subsets of the IMDb Maas et al. (2011), SST-2 Socher et al. (2013), and Headlines datasets. In these experiments, we independently prune attention edges with low (negative) and high (positive) ORC values over all layers of BERT-Tiny models during testing. Our findings deviate from the observations on image datasets. As illustrated in Figure 11, pruning edges with high ORC values leads to a more rapid degradation in performance compared to pruning edges with low ORC values. Furthermore, the accuracy exhibits abrupt transitions in response to minor variations in the pruning rate. The results should be interpreted with caution due to

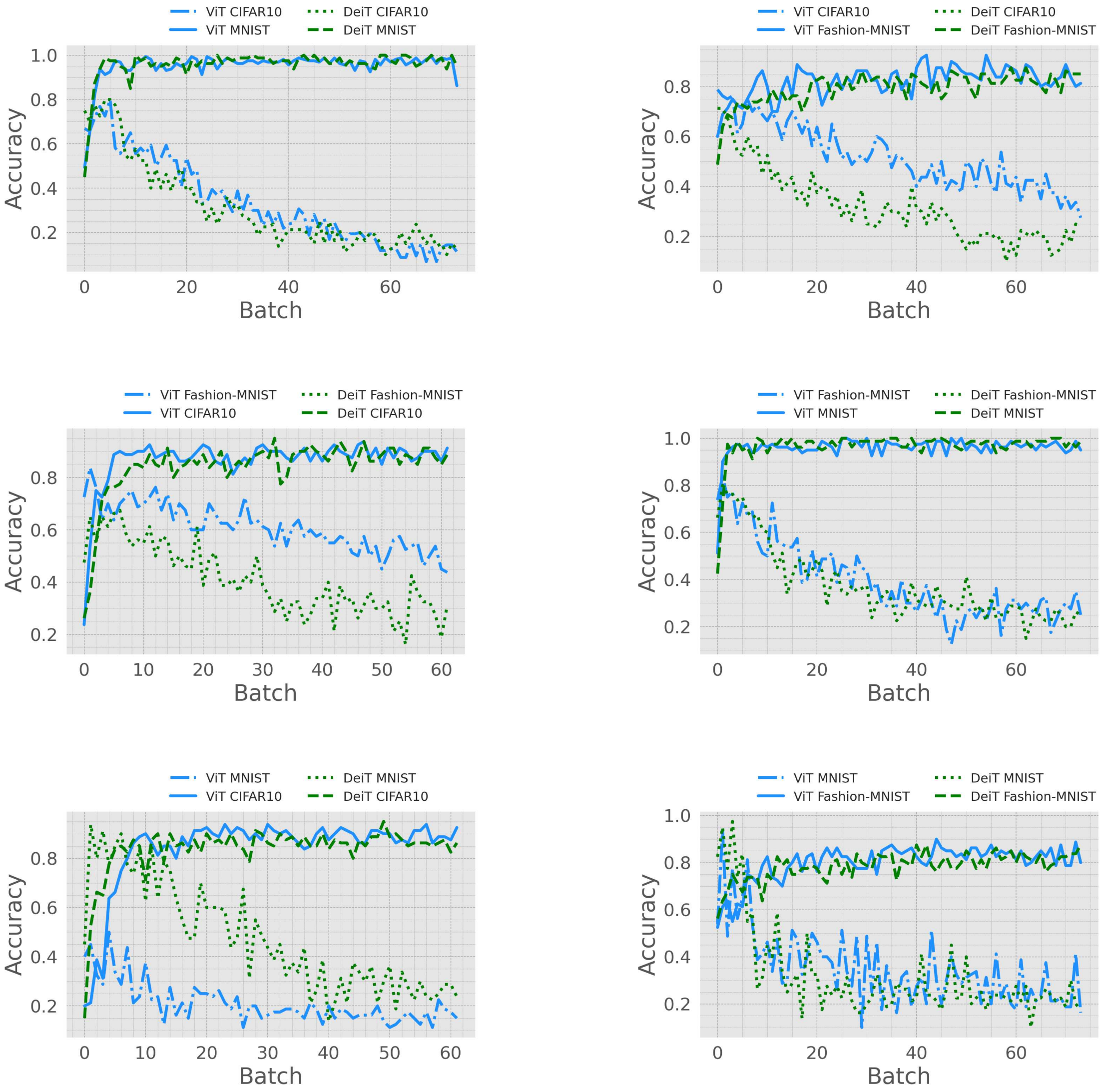

Figure 8: Test set accuracy of ViT (blue) and DeiT (green) models while fine-tuning on a secondary dataset after fine-tuning on a primary dataset. The primary and secondary datasets are: (a) CIFAR10 and MNIST; (b) CIFAR10 and Fashion-MNIST; (c) Fashion-MNIST and CIFAR10; (d) Fashion-MNIST and MNIST; (e) MNIST and CIFAR10; (f) MNIST and Fashion-MNIST. DeiT demonstrates a faster degradation in performance compared to ViT following subsequent fine-tuning.

the limited sample size. This is reflected in the observed variability and jitter in the results and the absence of a clear, monotonic decreasing trend.

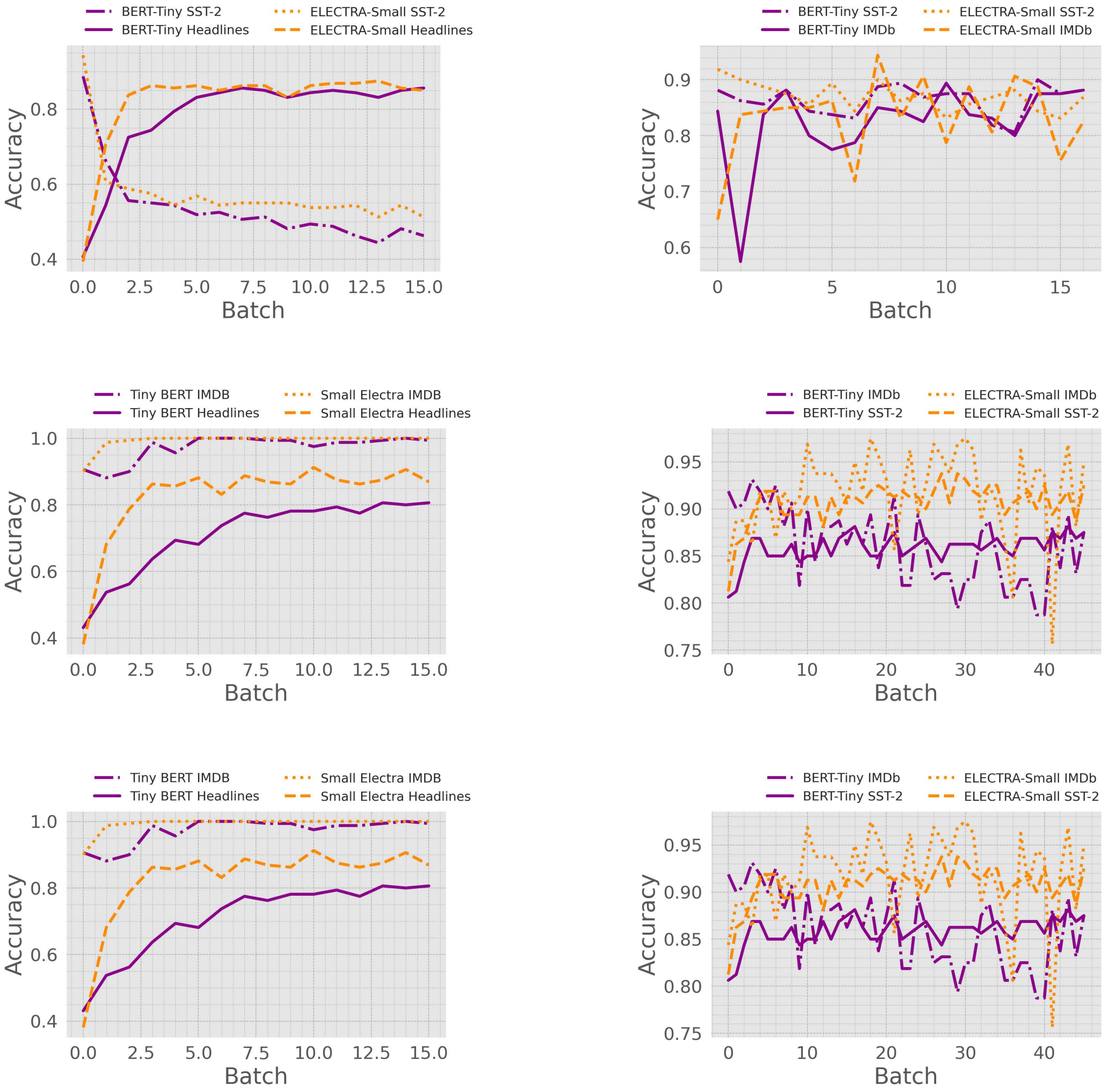

Figure 9: Test set accuracy of BERT-Tiny (purple) and ELECTRA-Small (orange) models while fine-tuning on a secondary dataset after fine-tuning on a primary dataset. The primary and secondary datasets are: (a) Headlines and SST-2; (b) Headlines and IMDb; (c) SST-2 and Headlines; (d) SST-2 and IMDb; (e) IMDb and Headlines; (f) IMDb and SST2. In cases where forgetting happens, BERT-Tiny demonstrates a faster degradation in performance compared to ELECTRA-Small following subsequent fine-tuning.

I.3. Training robust transformers

Given the established success of connection-dropping methods in training robust neural networks (Goodfellow et al., 2014; Jain et al., 2015; Srivastava et al., 2014; Yang et al., 2020), we also investigate the impact of random dropout of attention connections on the ORC distribution in *L-ViT* and BERT-Tiny. During models training, a Bernoulli mask was

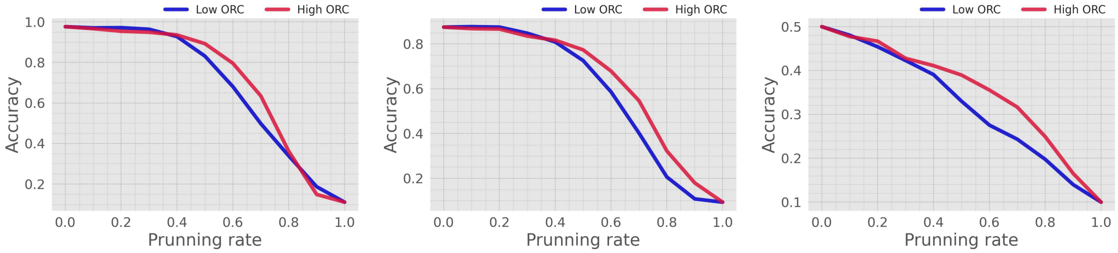

Figure 10: Effects of pruning rates and regimes on performance for an *L-ViT* model, evaluated over random subsets of the: MNIST (left), Fashion-MNIST (middle), and CIFAR10 (right) datasets. Pruning edges with lower curvature leads to greater performance impairment.

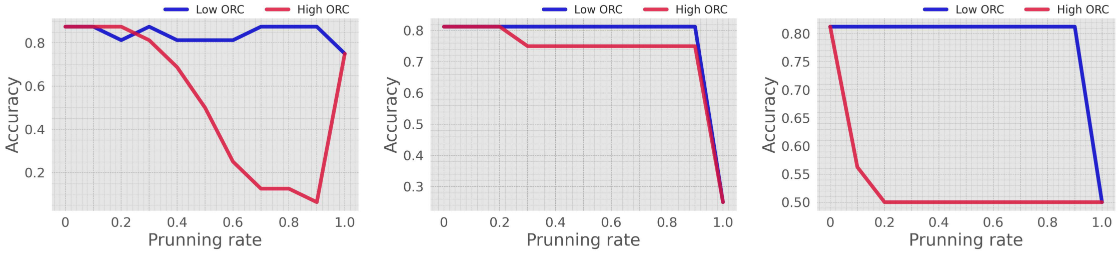

Figure 11: Effects of pruning rates and regimes on performance for a BERT-Tiny model, evaluated over random subsets of the: IMDb (left), SST-2 (middle), and Headlines (right) datasets. Pruning edges with lower curvature leads to smaller performance impairment.

implemented to stochastically remove edges across all attention maps at varying rates. In the testing phase, we assess the robustness of the trained models with unmanipulated attention, quantified by ρ as defined in Equation 13, over dataset subsets (see Appendix E.2). Informed by prior research on connection masking techniques (Luo et al., 2021; Zehui et al., 2019), we posit that its application will lead to enhanced model robustness. While the number of active attention connections may decrease, we anticipate this effect will not significantly impair overall performance.

Observing the ρ values at different Bernoulli masking rates on Figure 12, our experiments confirm that this robustness is indeed achieved. Noticeably, a trend toward increasing robustness is often observed when masking rates fall within the range of 0.2 to 0.5, indicating that masking at these rates promotes increased robustness. Furthermore, low performance degradation is observed even at very high masking rates. This phenomenon can be attributed to two counteracting effects. Firstly, the masking process reduces the number of active edges, potentially leading to a decrease in overall model capacity. Secondly, the remaining connections become less redundant due to the sparsity introduced by masking. This may allow individual connections to exert a more significant influence on the model's robustness. Meanwhile, as we increase the masking rate, we see in Figure 13 that the negative values of ORC diminish. Furthermore, the ORC distributions of the final attention layer exhibit a convergence towards modal values across all examined datasets, which suggests a diminishing extremity within these distributions. This provides another piece of evidence for our claims on the connection between ORC in the attention map and the robustness of the transformer.

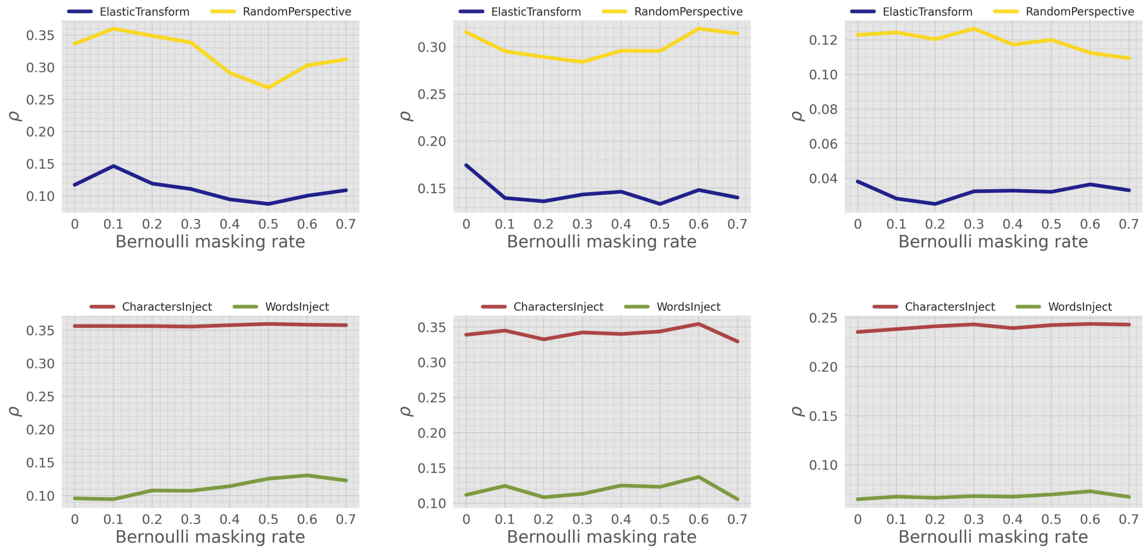

Figure 12: Perturbed test set accuracy rate degradation (ρ as defined in Equation 13) of *L-ViT* (top row) and BERT-Tiny (bottom row) models trained with Bernoulli masking. Results are evaluated over random subsets of the: (a) MNIST; (b) Fashion-MNIST; (c) CIFAR10; (d) IMDb; (e) SST-2; and (f) Headlines datasets. Masking rates affect model robustness in both vision and language models.

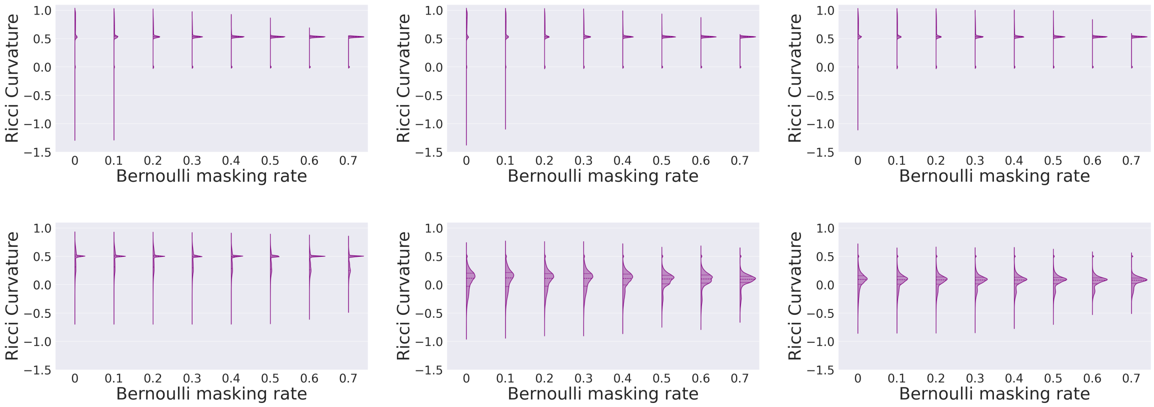

Figure 13: ORC distributions of the final layers of *L-ViT* (top row) and BERT-Tiny (bottom row) models trained with varying edge masking rates. Distributions are combined over random subsets of the: (a) MNIST; (b) Fashion-MNIST; (c) CIFAR10; (d) IMDb; (e) SST-2; and (f) Headlines datasets. Higher edge masking rates lead to curvature values convergence to a common modal value. Consequently, negative curvature values tend to diminish in both vision and language models.