Setting the Record Straight on Transformer Oversmoothing

Gbètondji J-S Dovonon[1] **Michael Bronstein**[2] **Matt J. Kusner**[1]
[1]Centre for AI, University College London
[2]Department of Computer Science, University of Oxford
{gbetondji.dovonon.22,m.kusner}@ucl.ac.uk

Abstract

Recent work has argued that Transformers are inherently low-pass filters that gradually oversmooth the inputs, limiting generalization, especially as model depth increases. How can Transformers achieve these successes given this shortcoming? In this work we show that in fact Transformers are not inherently low-pass filters. Instead, whether Transformers oversmooth or not depends on the eigenspectrum of their update equations. Further, depending on the task, smoothing does not harm generalization as model depth increases.

1 Introduction

The performance of transformer models Vaswani et al. (2023) can quickly saturate as model depth increases Kaplan et al. (2020); Wang et al. (2022). Recent work argues that it is because they are inherently low-pass filters Wang et al. (2022); Park & Kim (2022); Guo et al. (2023); Ali et al. (2023). In this work we show that in fact, *Transformer models are not inherently low-pass filters*. We make the following contributions: 1. we characterize the eigenspectrum of the Transformer update and its effect on oversmoothing as depth increases, generalising prior work Wang et al. (2022); Ali et al. (2023). 2. we detail how 'rank-collapse' Dong et al. (2021); Noci et al. (2022) will occur except in extremely rare cases, answering an open question about the role of the residual connection on rank-collapse. 3. we describe a simple way to reparameterize the Transformer to control the filtering behavior of the update. 4. we find that for certain tasks (e.g., image classification) *smoothing does not harm generalization, but improves it*. For other tasks (e.g., text generation) enforcing either smoothing or sharpening can hurt generalization. We derive a simple way to parameterize the weights of the Transformer update equations that allows for control over its spectrum

2 Background

The Transformer Update. At their core, Transformers are a linear combination of a set of 'heads'. Each head includes a self-attention function on the input $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{A} := \mathsf{Softmax}\Big(\frac{1}{\sqrt{k}}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top\Big)$, where the $\mathsf{Softmax}(\cdot)$ function is applied to each row individually. Further, $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d\times k}$ are learned query and key weight matrices. This 'attention map' $\mathbf{A}$ then transforms the input to produce the output of a single head $\mathbf{A}\mathbf{X}\mathbf{W}_V\mathbf{W}_{\mathsf{proj}}$, where $\mathbf{W}_V, \mathbf{W}_{\mathsf{proj}} \in \mathbb{R}^{d\times d}$ are learned value and projection weights. A residual connection is added to produce the output $\mathbf{X}_\ell$ of any layer ℓ: $\mathbf{X}_\ell := \mathbf{X}_{\ell-1} + \mathbf{A}_\ell\mathbf{X}_{\ell-1}\mathbf{W}_{V,\ell}\mathbf{W}_{\mathsf{proj},\ell}$, It is possible to introduce further complexity by learning additional heads (i.e., additional $\mathbf{A}, \mathbf{W}_V$) and summing all head outputs. For simplicity we will describe properties of the single-head Transformer.

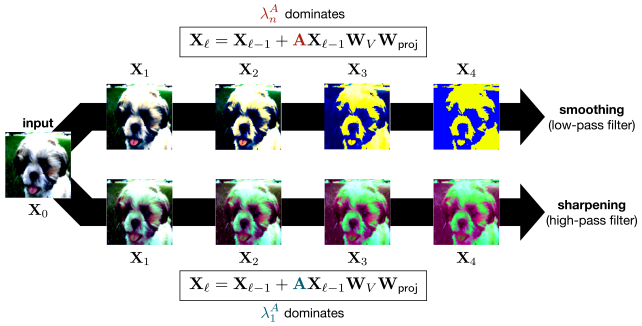

Figure 1: **Transformers can oversmooth, or not.** Evolution of the system $\mathbf{X}_\ell = \mathbf{X}_{\ell-1} + \mathbf{A}\mathbf{X}_{\ell-1}\mathbf{W}_V\mathbf{W}_{\text{proj}}$.

Oversmoothing via Low-Pass Filtering. There are many ways to measure oversmoothing, we opt here for the definition described in Wang et al. (2022) based on filtering, which we found the most intuitive. The overall idea is that we can view the layers of a deep learning model as a filtering operation that is applied repeatedly to $\mathbf{X}$. If the filtering operation is *low-pass*, it amplifies only the lowest frequency of $\mathbf{X}$, smoothing $\mathbf{X}$. On the other hand, a *high-pass* filter will amplify all other frequencies.

Specifically, let $\mathcal{F} : \mathbb{R}^{n\times d} \to \mathbb{C}^{n\times d}$ be the Discrete Fourier Transform (DFT). The DFT of $\mathbf{X}$ can be computed via matrix multiplication: $\mathcal{F}(\mathbf{X}) := \mathbf{F}\mathbf{X}$, where $\mathbf{F} \in \mathbb{C}^{n\times n}$ is equal to $\mathbf{F}_{k,l} := e^{2\pi \mathrm{i}(k-1)(l-1)}$ for all $k, l \in \{2, \dots, n\}$ (where $\mathrm{i} := \sqrt{-1}$), and is 1 otherwise (i.e., in the first row and column). Define the Low Frequency Component (LFC), also called the Direct Current, of $\mathbf{X}$ as $\text{LFC}[\mathbf{X}] := \mathbf{F}^{-1}\text{diag}([1, 0, \dots, 0])\mathbf{F}\mathbf{X} = (1/n)\mathbf{1}\mathbf{1}^\top\mathbf{X}$. Further, define the High Frequency Component (HFC), also called the Alternating Current, of $\mathbf{X}$ as $\text{HFC}[\mathbf{X}] := \mathbf{F}^{-1}\text{diag}([0, 1, \dots, 1])\mathbf{F}\mathbf{X} = (\mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^\top)\mathbf{X}$.

Definition 2.1 (Wang et al. (2022))**.** Given an endomorphism $f : \mathbb{R}^n \to \mathbb{R}^n$ where f^L denotes applying f repeatedly L times, f is a low-pass filter if and only if for all $\mathbf{X} \in \mathbb{R}^{n\times d}$

$$\lim_{L\to\infty} \frac{\|\text{HFC}[f^L(\mathbf{X})]\|_2}{\|\text{LFC}[f^L(\mathbf{X})]\|_2} = 0.$$

3 Do Transformers Always Oversmooth?

Preliminaries

We start by rewriting eq. (2) to simplify the analysis. Define the $\mathsf{vec}(\mathbf{M})$ operator as converting any matrix $\mathbf{M}$ to a vector $\mathbf{m}$ by stacking its columns. We can rewrite eq. (2) vectorized as follows

$$\mathsf{vec}(\mathbf{X}_\ell) = (\mathbf{I} + \underbrace{\mathbf{W}_{\text{proj}}^\top\mathbf{W}_V^\top}_{:=\mathbf{H}} \otimes \mathbf{A})\mathsf{vec}(\mathbf{X}_{\ell-1}). \tag{1}$$

This formulation is especially useful because $\mathsf{vec}(\mathbf{X}_L) = (\mathbf{I} + \mathbf{H} \otimes \mathbf{A})^L\mathsf{vec}(\mathbf{X})$. We now introduce an assumption on $\mathbf{A}$ that is also used in Wang et al. (2022).

Assumption 3.1 (Wang et al. (2022))**.** The attention matrix is positive, i.e., $\mathbf{A} > 0$, and invertible.

Note $\mathbf{A}$ is also right-stochastic, i.e., $\sum_j a_{i,j} = 1$. This combined with Assumption 3.1 immediately implies the following proposition.

Proposition 3.2 (Meyer & Stewart (2023))**.** *Given Assumption 3.1, all eigenvalues of* $\mathbf{A}$ *lie within* $(-1, 1]$*. There is one largest eigenvalue that is equal to* 1*, with corresponding unique eigenvector* $\mathbf{1}$*. No eigenvectors of* $\mathbf{A}$ *are equal to* 0*.*

The Eigenvalues

Consider the Transformer update with fixed $\mathbf{A} > 0, \mathbf{H} := \mathbf{W}_{\text{proj}}^\top \mathbf{W}_V^\top$, as described in eq. (1). Let $\{\lambda_i^A, \mathbf{v}_i^A\}_{i=1}^n$ and $\{\lambda_j^H, \mathbf{v}_j^H\}_{j=1}^d$ be the eigenvalue and eigenvectors of $\mathbf{A}$ and $\mathbf{H}$. Let the eigenvalues (and associated eigenvectors) be sorted as follows, $\lambda_1^A \leq \cdots \leq \lambda_n^A$ and $|1+\lambda_1^H| \leq \cdots \leq |1+\lambda_d^H|$. Let $\varphi_1^H, \ldots, \varphi_d^H$ be the phases of $\lambda_1^H, \ldots, \lambda_d^H$. As the number of layers L in the Transformer update eq. (1) increases, one eigenvalue of $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})^L$ will dominate the rest. Which eigenvalue(s) dominate will control the smoothing behavior of the Transformer.

Definition 3.3. At least one of the eigenvalues of $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$, i.e., $(1 + \lambda_j^H \lambda_i^A)$ has a larger magnitude than all others, i.e., there exists j^*, i^* (which may be a set of indices if there are ties) such that $|1 + \lambda_{j^*}^H \lambda_{i^*}^A| > |1 + \lambda_{j'}^H \lambda_{i'}^A|$ for all $j' \in \{1, \ldots, d\} \setminus j^*$ and $i' \in \{1, \ldots, n\} \setminus i^*$. These eigenvalues are called **dominating**.

The Features

Theorem 3.4. *As the number of total layers $L \to \infty$, the feature representation $\mathbf{X}_L$ converges. Which representation it converges to depends on the dominating eigenvalue, as given in Theorem B.6. If a single eigenvalue dominates, there are two cases:* ***(1)*** *If $(1 + \lambda_j^H \lambda_n^A)$ dominates then, $\mathbf{X}_L \to (1 + \lambda_j^H \lambda_n^A)^L s_{j,n} \mathbf{1} \mathbf{v}_j^{H\top}$,* ***(2)*** *If $(1 + \lambda_j^H \lambda_1^A)$ dominates then, $\mathbf{X}_L \to (1 + \lambda_j^H \lambda_1^A)^L s_{j,1} \mathbf{v}_1^A \mathbf{v}_j^{H\top}$ where $s_{j,i} := \langle \mathbf{v}^{Q^{-1}}{}_{j,i}, \mathsf{vec}(\mathbf{X}) \rangle$ and $\mathbf{v}^{Q^{-1}}{}_{j,i}$ is row ji in the matrix $\mathbf{Q}^{-1}$ (here $\mathbf{Q}$ is the matrix of eigenvectors of $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$). If multiple eigenvalues have the same dominating magnitude, the final representation $\mathbf{X}_L$ converges to the sum of the dominating terms.*

Filtering

Theorem 3.5. *For all $\mathbf{X} \in \mathbb{R}^{n \times d}$, as the number of total layers $L \to \infty$, if* ***(1)*** *$(1 + \lambda_j^H \lambda_n^A)$ dominates, $\lim_{L\to\infty} \frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} = 0$, and so $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$ acts as a low-pass filter, as in Definition 2.1. If* ***(2)*** *$(1 + \lambda_j^H \lambda_1^A)$ dominates, $\lim_{L\to\infty} \frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} \neq 0$, and so $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$ does not act as a low-pass filter. If* ***(3)*** *multiple eigenvalues have the same dominating magnitude, and there is at least one dominating eigenvalue $(1 + \lambda_j^H \lambda_i^A)$ where $\lambda_i^A \neq \lambda_n^A$, then eq. 7 holds and $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$ does not act as a low-pass filter.*

The proof is left to the Appendix. Wang et al. (2022) showed that if we just apply the self-attention matrix $\mathbf{A}$ alone to produce $\mathbf{X}_L$, i.e., $\mathbf{X}_L = \mathbf{A}^L \mathbf{X}$, then this model is always a low-pass filter, as defined in Definition 2.1. Theorem 3.5 shows that the residual connection and weights $\mathbf{H}$ can counteract this, so long as condition **(2)** or **(3)** holds.

4 A Reparameterization that Controls Filtering

Corollary 4.1. *If the eigenvalues of $\mathbf{H}$ fall within $[-1, 0)$, then at least one of $\{(1 + \lambda_d^H \lambda_1^A), (1 + \lambda_1^H \lambda_1^A)\}$ dominates. If the eigenvalues of $\mathbf{H}$ fall within $(0, \infty)$, then $(1 + \lambda_d^H \lambda_n^A)$ dominates.*

Corollary 4.1 states that so long as we can ensure the eigenvalues of $\mathbf{H}$ lie in $[-1, 0)$, the ratio of high frequencies over low frequencies (Definition 2.1) increases. See the Appendix for a proof. To ensure that the eigenvalues of $\mathbf{H}$ fall in these ranges, we propose to directly parameterize its eigendecomposition. Specifically, define $\mathbf{H}$ as $\mathbf{H} = \mathbf{V}_H \Lambda_H \mathbf{V}_H^{-1}$, where $\mathbf{V}_H$ is a full-rank matrix and Λ_H is diagonal. We learn parameters $\mathbf{V}_H$ by taking gradients in the standard way (i.e., directly and through the inversion). To learn the diagonal of

Λ_H, i.e., $\mathsf{diag}(\Lambda_H)$, we parameterize the sharpening model as $\mathsf{diag}(\Lambda_H) := \mathsf{clip}(\boldsymbol{\psi}, [-1, 0])$ (which we refer to in the following section using the superscript $^-$), where $\boldsymbol{\psi}$ are tunable parameters and $\mathsf{clip}(\boldsymbol{\psi}, [l, u]) := \min(\max(\boldsymbol{\psi}, l), u)$ forces all of $\boldsymbol{\psi}$ to lie in $[l, u]$. Similarly we parameterize the smoothing model as $\mathsf{diag}(\Lambda_H) := \mathsf{clip}(\boldsymbol{\psi}, [0, 1])$ (referred to with the superscript $^+$).[1] We also introduce a model that is neither dominated by smoothing nor sharpening dynamic, targeted at natural language tasks. This is done to allow the model's filtering behaviour to be dictated solely by the attention matrix. For this model, the diagonal matrix is $\mathsf{diag}(\Lambda_H) := [|\boldsymbol{\psi}|, -|\boldsymbol{\psi}|, \max(|\boldsymbol{\psi}|), -\max(|\boldsymbol{\psi}|) - \epsilon]$ where ϵ is a small learnable parameter. The addition of ϵ is done to balance out the fact that the eigenvalue with the largest norm of attention matrix is always 1 (we refer to this model with the superscript $^\sim$).

5 Results

We base our image classification experiments on the efficient ViT-Ti model Touvron et al. (2021a) our NLP experiments on Geiping & Goldstein (2023). We evaluate the vision models on CIFAR100 and the language models are pretrained on the Pile dataset Gao et al. (2020) and evaluated on SuperGLUE Wang et al. (2020). We then investigate (a) the distribution of dominating eigenvalues and (b) the filtering properties of both existing Transformer models and our proposed parameterizations (for all tasks: sharpening $^-$ and smoothing $^+$, and additionally for NLP: band-pass $^\sim$). We compare the image classification and text generation performance of these parameterizations and competing approaches. Crucially, even though our theoretical analysis applies for fixed attention $\mathbf{A}$ and weights $\mathbf{H}$, **we use existing model architectures throughout**, i.e., including different attention/weights each layer, multi-head attention, layer normalization (arranged in the pre-LN format Xiong et al. (2020)), and fully-connected layers.[2]

Existing models do not converge to low-pass filters. We investigate which eigenvalues dominate the Transformer updates of pre-trained and newly trained models in Table 1. To compute this we average over all heads and all layers for $\mathbf{H}$ and $\mathbf{A}$, and also over all batches for $\mathbf{A}$. We notice that even existing models (ViT-Ti Touvron et al. (2021a), and Crammed Bert Geiping & Goldstein (2023)) have a mixture of eigenvalues that lead to both sharpening and smoothing. Both FeatScale Wang et al. (2022) and Centered Attention Ali et al. (2023) seem to increase the proportion of eigenvalues that lead to oversmoothing. To further measure how much a model acts as a low-pass filter we compute the ratio of high frequencies over low frequencies (Definition 2.1), i.e., $\frac{\|\mathrm{HFC}[\mathbf{X}_\ell]\|_2}{\|\mathrm{LFC}[\mathbf{X}_\ell]\|_2}$ for each layer ℓ (and averaged over all batches) in vision Transformer models, in Figure 5. We observe that the average value of $\frac{\|\mathrm{HFC}[\mathbf{X}_\ell]\|_2}{\|\mathrm{LFC}[\mathbf{X}_\ell]\|_2}$ for ViT-Ti decreases as layers are increased but does not approach 0, instead roughly halving in value. This indicates

Reparameterization gives control over Transformer filtering behavior. We see that ViT-Ti$^-$ further sharpens while ViT-Ti$^+$ further smooths. This shows that even though our theory only applies to repeated attention $\mathbf{A}$ and weight $\mathbf{H}$ matrices, and even though we are only parameterizing $\mathbf{H}$ we can still control the sharpening/smoothing behavior of a model that includes different $\mathbf{A}, \mathbf{H}$ every layer, as well as multiple heads, layer normalization and fully-connected layers. For text generation, we see in Figure 6 that we have less control over filtering as the fully-connected layers seem to more strongly drive the filtering behavior.

Image classification. Figure 2 shows the train and test error of image classification on CIFAR100 for all models. We see that the smoothing model ViT-Ti$^+$ matches or outperforms the other models in test error. Parameterizing models to smooth improves generalization.

[1] While we could have allowed the smoothing model to use the space of positive reals via $\mathsf{diag}(\Lambda_H) := |\boldsymbol{\psi}|$, we found that restricting the space of allowed eigenvalues stabilized training.

[2] If a model has multiple heads we will define $\mathbf{W}_V = \mathbf{V}_H$ and $\mathbf{W}_{\mathsf{proj}} = \Lambda_H \mathbf{V}_H^\top$).

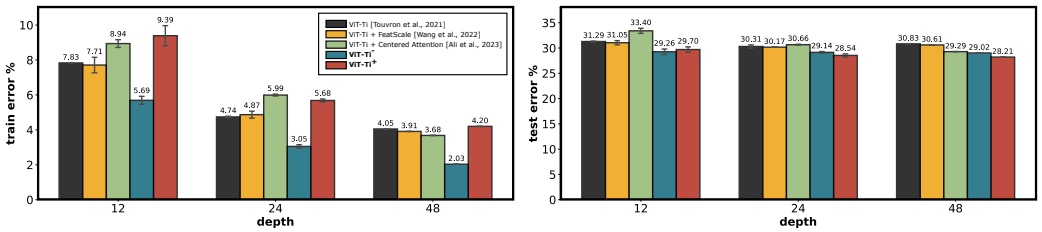

Figure 2: **Image classification.** The train and test error results on CIFAR100 for models with $\{12, 24, 48\}$ layers.

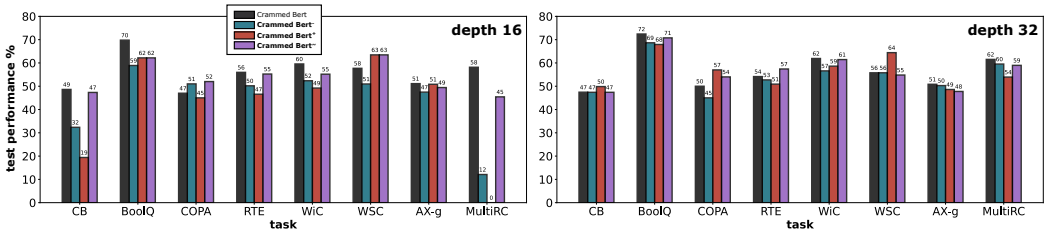

Figure 3: **Text generation.** The test performance results on SuperGlue tasks for models with $\{16, 32\}$ layers.

Text generation. Figure 6 details the performance (the higher the better) of Crammed Bert models on SuperGlue tasks (following the literature we report test F1 for the CB and MultiRC benchmarks, and test accuracy for the rest). We observe that both Crammed Bert$^-$ and Crammed Bert$^+$ largely harm the performance of the original model. Whereas the Crammed Bert$^\sim$ model can match or improve the performance in some cases by balancing sharpening and smoothing. Different from image classification, text generation seems to need a more complex filter.

6 Related Work

Oversmoothing is a concept that has been widely discussed in the graph neural network literature Rusch et al. (2023). Giovanni et al. (2023) prove that graph convolutions can enhance high frequencies. Their analysis inspires our work. For transformers, Zhou et al. (2021), Gong et al. (2021) and Raghu et al. (2021) found that feature similarity in vision Transformers increased with depth and Dong et al. (2021) show that self-attention layers converge doubly exponentially to a rank 1 matrix. Many works around this time found that it was possible to improve vision Transformers by replacing self-attention layers with convolutional layers (Han et al., 2021; Liu et al., 2021; Jiang et al., 2021; Touvron et al., 2021b; Yuan et al., 2021; Park & Kim, 2022). Other works introduced new layers to avoid oversmoothing (Wang et al., 2022; Guo et al., 2023; Ali et al., 2023; Choi et al., 2023). Oversmoothing also occurs in Transformer architectures for NLP Shi et al. (2022).

7 Discussion

In this paper, we presented a new analysis detailing how the eigenspectrum of attention and weight matrices impacts the final representation produced by the Transformer update, as depth is increased. Contrary to prior work, this analysis revealed that Transformers are not inherently low-pass filters. Empirically we show that existing Transformer models already have properties that partially prevent oversmoothing. We introduce a new parameterization for the Transformer weights that is guaranteed to avoid oversmoothing. This parameterization is prone to overfitting, while the the smoothing parameterization scales better with depth.

References

Ali, A., Galanti, T., and Wolf, L. Centered self-attention layers. *arXiv preprint arXiv:2306.01610*, 2023.

Brualdi, R. A. and Mellendorf, S. Regions in the complex plane containing the eigenvalues of a matrix. *The American mathematical monthly*, 101(10):975–985, 1994.

Choi, J., Wi, H., Kim, J., Shin, Y., Lee, K., Trask, N., and Park, N. Graph convolutions enrich the self-attention in transformers! *arXiv preprint arXiv:2312.04234*, 2023.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space, 2019.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Geiping, J. and Goldstein, T. Cramming: Training a language model on a single gpu in one day. In *International Conference on Machine Learning*, pp. 11117–11143. PMLR, 2023.

Giovanni, F. D., Rowbottom, J., Chamberlain, B. P., Markovich, T., and Bronstein, M. M. Understanding convolution on graphs via energies. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=v5ew3FPTgb`.

Gong, C., Wang, D., Li, M., Chandra, V., and Liu, Q. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*, 2021.

Guo, X., Wang, Y., Du, T., and Wang, Y. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

Jiang, Z., Hou, Q., Yuan, L., Zhou, D., Jin, X., Wang, A., and Feng, J. Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet. *arXiv preprint arXiv:2104.10858*, 3(6):7, 2021.

Kaddour, J., Key, O., Nawrot, P., Minervini, P., and Kusner, M. J. No train no gain: Revisiting efficient training algorithms for transformer-based language models, 2023.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Meyer, C. D. and Stewart, I. *Matrix analysis and applied linear algebra*. SIAM, 2023.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

Park, N. and Kim, S. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

Rusch, T. K., Bronstein, M. M., and Mishra, S. A survey on oversmoothing in graph neural networks, 2023.

Schacke, K. On the kronecker product. *Master's thesis, University of Waterloo*, 2004.

Shi, H., Gao, J., Xu, H., Liang, X., Li, Z., Kong, L., Lee, S., and Kwok, J. T. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention, 2021a.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 32–42, 2021b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020.

Wang, P., Zheng, W., Chen, T., and Wang, Z. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*, 2022.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2018.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., and Feng, J. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

Appendix

A Implementation Details

Initialization. We initialize $\mathbf{H} = \mathbf{V}_H \Lambda_H \mathbf{V}_H^{-1}$ to avoid too strong filtering behaviors and to mimic the initializations used in the ViT-Ti and Bert baselines. The baselines are initialized using He initialization He et al. (2015). We also initialize $\mathbf{V}_H$ with the same method. Randomly inialized matrices will typically have normally distributed eigenvalues centered at 0, so we initialize Λ_H using a normal distribution. Corollary 4.1 tells us that eigenvalues within the range [-1,0) avoid oversmoothing, however we observe that large eigenvalues lead to unstable training. To stabilize training, we initialize Λ_H with a normal distribution with mean 0 and standard deviation 0.1.

Image Classification: Training & Architecture Details. We base our image classification experiments on the ViT-Ti model Touvron et al. (2021a). Specifically, we train all models on CIFAR100 for 300 epochs using the cross-entropy loss and the AdamW optimizer Loshchilov & Hutter (2019). Our setup is the one used in Park & Kim (2022) which itself follows the DeiT training recipe Touvron et al. (2021a). We use a cosine annealing schedule with an initial learning rate of 1.25×10^{-4} and weight decay of 5×10^{-2}. We use a batch size of 96. We use data augmentation including RandAugment Cubuk et al. (2019), CutMix Yun et al. (2019), Mixup Zhang et al. (2018), and label smoothing Touvron et al. (2021a). The models were trained on two Nvidia RTX 2080 Ti GPUs.

Text Generation: Training & Architecture Details. We base our NLP experiments on Geiping & Goldstein (2023), using their code-base. Following this work we pre-train encoder-only 'Crammed' Bert models with a maximum budget of 24 hours. We use a masked language modeling objective and train on the Pile dataset Gao et al. (2020). The batch size is 8192 and the sequence length is 128. We evaluate models on SuperGLUE Wang et al. (2020) after fine-tuning for each task. In order to ensure a fair comparison, all models are trained on a reference system with an RTX 4090 GPU. We use mixed precision training with bfloat16 as we found it to be the most stable Kaddour et al. (2023).

B Proofs

Proposition B.1 (Meyer & Stewart (2023))**.** *Given Assumption 3.1, all eigenvalues of* $\mathbf{A}$ *lie within* $(-1, 1]$*. There is one largest eigenvalue that is equal to* 1*, with corresponding unique eigenvector* $\mathbf{1}$*. No eigenvectors of* $\mathbf{A}$ *are equal to* 0*.*

Proof. First, because $\mathbf{A}$ is positive, by the Perron-Frobenius Theorem Meyer & Stewart (2023) all eigenvalues of $\mathbf{A}$ are in $\mathbb{R}$ (and so there exist associated eigenvectors that are also in $\mathbb{R}$). Next, recall the definition of an eigenvalue λ and eigenvector $\mathbf{v}$: $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Let us write the equation for any row $i \in \{1, \dots, n\}$ explicitly:

$$a_{i1}v_1 + \cdots + a_{in}v_n = \lambda v_i.$$

Further let,

$$v_{\max} := \max\{|v_1|, \dots, |v_n|\} \tag{2}$$

Note that $v_{\max} > 0$, otherwise it is not a valid eigenvector. Further let $k_{\max}$ be the index of $\mathbf{v}$ corresponding to $v_{\max}$. Then we have,

$$\begin{aligned}
|\lambda| v_{\max} &= |a_{k_{\max}1}v_1 + \cdots + a_{k_{\max}n}v_n| \\
&\leq a_{k_{\max}1}|v_1| + \cdots + a_{k_{\max}n}|v_n| \\
&\leq a_{k_{\max}1}|v_{k_{\max}}| + \cdots + a_{k_{\max}n}|v_{k_{\max}}| \\
&= (a_{k_{\max}1} + \cdots + a_{k_{\max}n})|v_{k_{\max}}| = |v_{\max}|
\end{aligned}$$

The first inequality is given by the triangle inequality and because $a_{ij} > 0$. The second is given by the definition of $v_{\max}$ as the maximal element in $\mathbf{v}$. The final inequality is given by the definition of $\mathbf{A}$ in eq. (2) as right stochastic (i.e., all rows of $\mathbf{A}$ sum to 1) and because $|v_{k_{\max}}| = |v_{\max}|$. Next, note that because $v_{\max} > 0$, it must be that $\lambda \leq 1$. Finally, to show that the one largest eigenvalue is equal to 1, recall by the definition of $\mathbf{A}$ in eq. (2) that $\mathbf{A1} = \mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. So $\mathbf{1}$ is an eigenvector of $\mathbf{A}$, with eigenvalue $\lambda^* = 1$. Because $a_{ij} > 0$, and we showed above that all eigenvalues must lie in in $[-1, 1]$, by the Perron-Frobenius theorem Meyer & Stewart (2023) $\lambda^* = 1$ is the Perron root. This means that all other eigenvalues λ_i satisfy the following inequality $|\lambda_i| < \lambda^*$. Further $\mathbf{1}$ is the Perron eigenvector, and all other eigenvectors have at least one negative component, making $\mathbf{1}$ unique. Finally, because $\mathbf{A}$ is invertible, it cannot have any 0 eigenvalues Brualdi & Mellendorf (1994). □

Theorem B.2. *Consider the Transformer update with fixed* $\mathbf{A} > 0, \mathbf{H} := \mathbf{W}_{\text{proj}}^\top \mathbf{W}_V^\top$, *as described in eq. (1). Let* $\{\lambda_i^A, \mathbf{v}_i^A\}_{i=1}^n$ *and* $\{\lambda_j^H, \mathbf{v}_j^H\}_{j=1}^d$ *be the eigenvalue and eigenvectors of* $\mathbf{A}$ *and* $\mathbf{H}$. *Let the eigenvalues (and associated eigenvectors) be sorted as follows,* $\lambda_1^A \leq \cdots \leq \lambda_n^A$ *and* $|1 + \lambda_1^H| \leq \cdots \leq |1 + \lambda_d^H|$. *Let* $\varphi_1^H, \ldots, \varphi_d^H$ *be the phases of* $\lambda_1^H, \ldots, \lambda_d^H$. *As the number of layers* $L \to \infty$, *one eigenvalue dominates the rest (multiple dominate if there are ties):*

$$\begin{cases} \left.\begin{array}{ll} (1 + \lambda_d^H \lambda_n^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| > |1 + \lambda_d^H \lambda_1^A| \\ (1 + \lambda_1^H \lambda_1^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| < |1 + \lambda_d^H \lambda_1^A| \end{array}\right\} & \text{if } \lambda_1^A > 0 \\ \left.\begin{array}{ll} (1 + \lambda_d^H \lambda_n^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| > |1 + \lambda_k^H \lambda_1^A| \\ (1 + \lambda_k^H \lambda_1^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| < |1 + \lambda_k^H \lambda_1^A| \end{array}\right\} & \text{if } \lambda_1^A < 0, \varphi_d^H \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ \left.\begin{array}{ll} (1 + \lambda_d^H \lambda_n^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| > |1 + \lambda_d^H \lambda_1^A| \\ (1 + \lambda_d^H \lambda_1^A) & \text{if } |1 + \lambda_d^H \lambda_n^A| < |1 + \lambda_d^H \lambda_1^A| \end{array}\right\} & \text{if } \lambda_1^A < 0, \varphi_d^H \in (\frac{\pi}{2}, \pi] \cup [-\pi, -\frac{\pi}{2}) \end{cases}$$

where λ_k^H *is the eigenvalue with the largest index* k *such that* $\varphi_k^H \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$.

Proof. First, note that given the eigendecompositions of $\mathbf{H} := \mathbf{W}_{\text{proj}}^\top \mathbf{W}_V^\top$ and $\mathbf{A} > 0$, as $\{\lambda_i^A, \mathbf{v}_i^A\}_{i=1}^n$ and $\{\lambda_j^H, \mathbf{v}_j^H\}_{j=1}^d$, the eigenvalues and eigenvectors of $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$ are equal to $(1 + \lambda_j^H \lambda_i^A)$ and $\mathbf{v}_j^H \otimes \mathbf{v}_i^A$ for all $j \in \{1, ..., d\}$ and $i \in \{1, \ldots, n\}$ Schacke (2004) (Theorem 2.3). Recall that eigenvalues (and associated eigenvectors) are sorted in the following order $\lambda_1^A \leq \cdots \leq \lambda_n^A$ and $|1 + \lambda_1^H| \leq \cdots \leq |1 + \lambda_d^H|$. Now at least one of the eigenvalues of $(\mathbf{I} + \mathbf{H} \otimes \mathbf{A})$, i.e., $(1 + \lambda_j^H \lambda_i^A)$ has a larger magnitude than all others, i.e., there exists j^*, i^* (which may be a set of indices if there are ties) such that $|1 + \lambda_{j^*}^H \lambda_{i^*}^A| > |1 + \lambda_{j'}^H \lambda_{i'}^A|$ for all $j' \in \{1, \ldots, d\} \setminus j^*$ and $i' \in \{1, \ldots, n\} \setminus i^*$. As $L \to \infty$ the expression $(1 + \lambda_{j^*}^H \lambda_{i^*}^A)^L$ will dominate all eigenvalue expressions $(1 + \lambda_{j'}^H \lambda_{i'}^A)^L$ (again multiple will if there are ties). Our goal is to understand the identity of $\lambda_{j^*}^H \lambda_{i^*}^A$ for all possible values of λ_H, λ_A. For simplicity we will assume there are no ties, i.e., j^*, i^* each denote a single index. In this case we only need to consider strict inequalities of λ_H, λ_A (as equalities indicate that multiple eigenvalues dominate).

First recall that $\lambda_i^A \in (-1, 1]$ and $\lambda_n^A = 1$. A useful way to view selecting $\lambda_j^H \lambda_i^A$ to maximize $|1 + \lambda_j^H \lambda_i^A|$ is as maximizing distance to -1. If (i), $\lambda_1^A > 0$ then λ_1^A always shrinks λ_j^H to the origin. If $\varphi_j^H \in [-\pi/2, \pi/2]$ then this shrinking will always bring λ_j^H closer to -1. If instead $\varphi_j^H \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$ then this shrinking can bring λ_j^H farther from -1. The eigenvalue it can bring farthest from -1 is λ_1^H (as λ_1^H is already farthest from -1 given that $|1 + \lambda_1^H| \leq \cdots \leq |1 + \lambda_d^H|$). If this point is farther from -1 than $\lambda_d^H \lambda_n^A$, i.e., if $|1 + \lambda_1^H \lambda_1^A| > |1 + \lambda_d^H \lambda_n^A|$ then $(1 + \lambda_1^H \lambda_1^A)$ dominates. Otherwise, $(1 + \lambda_d^H \lambda_n^A)$ dominates. If instead (ii), $\lambda_1^A < 0$ then it is possible to 'flip' λ_j^H across the origin, and so the maximizer depends on φ_d^H. If a) $\varphi_d^H \in [-\pi/2, \pi/2]$ then let λ_k^H be the eigenvalue with the largest index k such that $\varphi_k^H \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$. It is possible that 'flipping' this eigenvalue across the origin makes it farther away than λ_d^H, i.e., $|1 + \lambda_k^H \lambda_1^A| > |1 + \lambda_d^H \lambda_n^A|$. In this case $(1 + \lambda_k^H \lambda_1^A)$ dominates, otherwise $(1 + \lambda_d^H \lambda_n^A)$ dominates. If instead b) $\varphi_d^H \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$ then

either $|1+\lambda_d^H\lambda_n^A| > |1+\lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \neq d$ and $i' \neq n$, and so $(1+\lambda_d^H\lambda_n^A)$ dominates, or 'flipping' λ_d^H increases its distance from -1, and so $|1+\lambda_d^H\lambda_1^A| > |1+\lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \neq d$ and $i' \neq n$, and so $(1+\lambda_d^H\lambda_1^A)$ dominates. □

Theorem B.3. *As the number of total layers $L \to \infty$, the feature representation $\mathbf{X}_L$ converges. Which representation it converges to depends on the dominating eigenvalue, as given in Theorem B.6. If a single eigenvalue dominates, there are two cases:* ***(1)*** *If $(1+\lambda_j^H\lambda_n^A)$ dominates then,*

$$\mathbf{X}_L \to (1+\lambda_j^H\lambda_n^A)^L s_{j,n} \mathbf{1} {\mathbf{v}_j^H}^\top, \tag{3}$$

(2) *If $(1+\lambda_j^H\lambda_1^A)$ dominates then,*

$$\mathbf{X}_L \to (1+\lambda_j^H\lambda_1^A)^L s_{j,1} \mathbf{v}_1^A {\mathbf{v}_j^H}^\top \tag{4}$$

where $s_{j,i} := \langle {\mathbf{v}^Q}^{-1}_{j,i}, \mathsf{vec}(\mathbf{X})\rangle$ and ${\mathbf{v}^Q}^{-1}_{j,i}$ is row ji in the matrix $\mathbf{Q}^{-1}$ (here $\mathbf{Q}$ is the matrix of eigenvectors of $(\mathbf{I}+\mathbf{H}\otimes\mathbf{A})$). If multiple eigenvalues have the same dominating magnitude, the final representation $\mathbf{X}_L$ converges to the sum of the dominating terms.

Proof. Recall that the eigenvalues and eigenvectors of $(\mathbf{I}+\mathbf{H}\otimes\mathbf{A})$ are equal to $(1+\lambda_j^H\lambda_i^A)$ and $\mathbf{v}_j^H \otimes \mathbf{v}_i^A$ for all $j \in \{1,...,d\}$ and $i \in \{1,\ldots,n\}$. This means,

$$\mathsf{vec}(\mathbf{X}_L) = \sum_{i,j} (1+\lambda_j^H\lambda_i^A)^L \langle {\mathbf{v}^Q}^{-1}_{j,i}, \mathsf{vec}(\mathbf{X})\rangle (\mathbf{v}_j^H \otimes \mathbf{v}_i^A). \tag{5}$$

Recall that ${\mathbf{v}^Q}^{-1}_{j,i}$ is row ji in the matrix $\mathbf{Q}^{-1}$, where $\mathbf{Q}$ is the matrix of eigenvectors $\mathbf{v}_j^H \otimes \mathbf{v}_i^A$. Further recall that $\mathbf{v}_i^A = \mathbf{1}$. As described in Theorem B.6, as $L \to \infty$ at least one of the eigenvalues pairs $\lambda_j^H\lambda_i^A$ will dominate the expression $(1+\lambda_j^H\lambda_i^A)^L$, which causes $\mathsf{vec}(\mathbf{X}_L)$ to converge to the dominating term. Finally, we can rewrite, $\mathbf{v}_1 \otimes \mathbf{v}_2$ as $\mathsf{vec}(\mathbf{v}_2\mathbf{v}_1^\top)$. Now all non-scalar terms have $\mathsf{vec}(\cdot)$ applied, so we can remove this function everywhere to give the matrix form given in eq. (3) and eq. (4). □

Corollary B.4. *Let $\mathcal{E}$ be the set of pairs of indices (j,i) such that $|1+\lambda_j^H\lambda_i^A|$ is equal to the dominating eigenvalue magnitude. Define a* unique pair set *$\mathcal{U} \subseteq \mathcal{E}$, for which the following holds: $(j,i) \in \mathcal{U}$ iff $(j,i) \in \mathcal{E}$ and $(j,i'), (j',i) \notin \mathcal{U}$, for all $i' \in \{1,\ldots,n\} \setminus i$ and $j' \in \{1,\ldots,d\} \setminus j$. Define a* maximal *unique pair set $\mathcal{U}^*$ as $|\mathcal{U}^*| \geq |\mathcal{U}|$ for all unique pair sets $\mathcal{U}$. As $L \to \infty$, the rank of $\mathbf{X}_L$ converges to $|\mathcal{U}^*|$.*

Proof. First recall that the rank of a matrix $\mathbf{M}$ is the smallest number k such that $\mathbf{M}$ can be written as a sum of k rank-1 matrices. Next note that if we have $\mathcal{E} = \{(j,i),(j',i)\}$ for $j \neq j'$ then we have

$$\mathbf{X}_L \to a_{j,i}\mathbf{v}_i^A{\mathbf{v}_j^H}^\top + a_{j',i}\mathbf{v}_i^A{\mathbf{v}_{j'}^H}^\top = a_{j,i}\mathbf{v}_i^A({\mathbf{v}_j^H}^\top + \frac{a_{j',i}}{a_{j,i}}{\mathbf{v}_{j'}^H}^\top),$$

where $a_{j,i} := (1+\lambda_j^H\lambda_i^A)^L s_{j,i}$. This shows that $\mathbf{X}_L$ is rank 1, which agrees in this example with $|\mathcal{U}^*| = 1$ (the same holds for $\mathcal{E} = \{(j,i),(j,i')\}$). In general, whenever the same index appears in different pairs in $\mathcal{E}$, we can group all associated terms in the expression for $\mathbf{X}_L$ into a single rank-1 term. Therefore, an element in a unique pair set $\mathcal{U}$ corresponds to a grouped rank-1 term in the expression for $\mathbf{X}_L$. Every element in a maximal unique pair set $\mathcal{U}^*$ corresponds 1-to-1 to every grouped rank-1 term in the expression for $\mathbf{X}_L$. So we can write $\mathbf{X}_L$ as,

$$\mathbf{X}_L \to \sum_{(j,i)\in\mathcal{U}^*} a_{j,i}\mathbf{g}_i^A{\mathbf{g}_j^H}^\top, \tag{6}$$

where each $\mathbf{g}_i^A, \mathbf{g}_j^H$ are potentially grouped terms (i.e., linear combinations of $\mathbf{v}_i^A, \mathbf{v}_j^H$). Further, none of the elements of the above sum can be grouped to yield a sum with fewer rank-1 terms. Therefore, the rank of $\mathbf{X}_L$ approaches $|\mathcal{U}^*|$, and we are done. □

Corollary B.5. *If the eigenvalues of* $\mathbf{H}$ *fall within* $[-1, 0)$*, then at least one of* $\{(1+\lambda_d^H\lambda_1^A), (1+\lambda_1^H\lambda_1^A)\}$ *dominates. If the eigenvalues of* $\mathbf{H}$ *fall within* $(0, \infty)$*, then* $(1+\lambda_d^H\lambda_n^A)$ *dominates.*

Proof. Let $\lambda_1^H \leq \cdots \leq \lambda_d^H$. Again we can think of selecting $\lambda_j^H\lambda_i^A$ that maximizes $|1+\lambda_j^H\lambda_i^A|$ as maximizing the distance of $\lambda_j^H\lambda_i^A$ to -1. Consider the first case where $\lambda_1^H, \cdots, \lambda_d^H \in [-1, 0)$, and so λ_1^H is the closest eigenvalue to -1 and λ_d^H is the farthest. If $\lambda_1^A > 0$ then all λ^A can do is shrink λ^H to the origin, where λ_1^A shrinks λ^H the most. The closest eigenvalue to the origin is λ_d^H, and so $(1 + \lambda_d^H\lambda_1^A)$ dominates. If instead $\lambda_1^A < 0$, then we can 'flip' λ_j^H over the origin, making it farther from -1 than all other $\lambda_{j'}^H$. The eigenvalue that we can 'flip' the farthest from -1 is λ_1^H, and so $(1 + \lambda_1^H\lambda_1^A)$ dominates. If all eigenvalues of $\mathbf{H}$ are equal, then both $(1 + \lambda_d^H\lambda_1^A)$ and $(1 + \lambda_1^H\lambda_1^A)$ dominate. For the second case where $\lambda_1^H, \cdots, \lambda_d^H \in (0, \infty)$ the result follows directly from the first case in Theorem C.1, and so we are done. □

Corollary B.6. *If the phases of* $\lambda_1^H, \ldots, \lambda_d^H$ *all fall within specific ranges, the dominating eigenvalue conditions can be simplified as follows:*

$$\begin{cases} (1+\lambda_d^H\lambda_n^A) & & \\ (1+\lambda_d^H\lambda_n^A) & \text{if } |1+\lambda_d^H\lambda_n^A| > |1+\lambda_1^H\lambda_1^A| \\ (1+\lambda_1^H\lambda_1^A) & \text{if } |1+\lambda_d^H\lambda_n^A| < |1+\lambda_1^H\lambda_1^A| \end{cases} \Bigg\} \text{ if } \lambda_1^A > 0 \Bigg\} \quad \text{if } \varphi_j^H \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}],\ \forall j$$

$$\begin{cases} (1+\lambda_d^H\lambda_n^A) & \text{if } |1+\lambda_d^H\lambda_n^A| > |1+\lambda_d^H\lambda_1^A| \\ (1+\lambda_d^H\lambda_1^A) & \text{if } |1+\lambda_d^H\lambda_n^A| < |1+\lambda_d^H\lambda_1^A| \end{cases} \Bigg\} \text{ if } \lambda_1^A < 0 \Bigg\} \quad \text{if } \varphi_j^H \in (\tfrac{\pi}{2}, \pi] \cup [-\pi, -\tfrac{\pi}{2}),\ \forall j$$

Proof. The proof is similar to that of Theorem B.6 except here we consider special cases.

if $\varphi_j^H \in [-\pi/2, \pi/2]$ **for all** $j \in \{1, \ldots, d\}$**.** First recall that $\lambda_i^A \in (-1, 1]$ and $\lambda_n^A = 1$. As $\varphi_j^H \in [-\pi/2, \pi/2]$, we have that $|1 + \lambda_d^H\lambda_n^A| > |1 + \lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \in \{1, \ldots, d-1\}$ and $i' \in \{1, \ldots, n-1\}$. This is because, if (i), $\lambda_1^A > 0$, (and so $\lambda_i^A > 0$ for all i) then $\arg(\lambda_{j'}^H\lambda_{i'}^A) \in [-\pi/2, \pi/2]$, where $\arg(a)$ is the argument or 'phase' or $a \in \mathbb{C}$. This combined with the fact that $|\lambda_{j'}^H\lambda_{i'}^A| < |\lambda_d^H\lambda_n^A|$ means that $|1 + \lambda_{j'}^H\lambda_{i'}^A| < |1 + \lambda_d^H\lambda_n^A|$. This because for any two points a, a' where we have that $\arg(a), \arg(a') \in [-\pi/2, \pi/2]$ and $|a'| < |a|$, then it also holds that $|1 + a'| < |1 + a|$. Therefore $(1 + \lambda_d^H\lambda_n^A)$ dominates. If instead (ii), $\lambda_1^A < 0$ then for any negative eigenvalues $\lambda_{i^-}^A < 0$ we have that $\arg(\lambda_j^H\lambda_{i^-}^A) \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$ for all j. However, for each of these points we have that $|1 + \lambda_j^H\lambda_{i^-}^A| < |1 + \lambda_j^H\lambda_n^A|$. This is because for any $r \in (-1, 0)$ and point b where $\arg(b) \in [-\pi/2, \pi/2]$ we have that $|1 + r * b| < |1 + b|$. Further note that $|1 + \lambda_j^H\lambda_n^A| < |1 + \lambda_d^H\lambda_n^A|$ from our definitions: $\lambda_n^A = 1$ and $|1 + \lambda_1^H| < \cdots < |1 + \lambda_d^H|$. And so $(1 + \lambda_d^H\lambda_n^A)$ dominates. For the remaining positive eigenvalues $\lambda_{i^+}^A \geq 0$ we are in the same situation as (i), and so we are done.

if $\varphi_j^H \in (\pi/2, \pi] \cup [-\pi, -\pi/2)$ **for all** $j \in \{1, \ldots, d\}$**.** If (a), $\lambda_1^A > 0$ then either $|1+\lambda_d^H\lambda_n^A| > |1 + \lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \neq d$ and $i' \neq n$, and so $(1 + \lambda_d^H\lambda_n^A)$ dominates, or shrinking λ_1^H to the origin makes it the farthest from -1, i.e, $|1 + \lambda_1^H\lambda_1^A| > |1 + \lambda_{j'}^H\lambda_{i'}^A|$, and so $(1 + \lambda_1^H\lambda_1^A)$ dominates. If (b) $\lambda_1^A < 0$ then either $|1+\lambda_d^H\lambda_n^A| > |1+\lambda_{j'}^H\lambda_{i'}^A|$, and so $(1+\lambda_d^H\lambda_n^A)$ dominates, or 'flipping' λ_d^H across the origin makes it farthest from -1, i.e., $|1 + \lambda_d^H\lambda_1^A| > |1 + \lambda_{j'}^H\lambda_{i'}^A|$, and so $(1 + \lambda_d^H\lambda_1^A)$ dominates. □

Theorem B.3. *For all* $\mathbf{X} \in \mathbb{R}^{n\times d}$*, as the number of total layers* $L \to \infty$*, if* ***(1)*** $(1 + \lambda_j^H\lambda_n^A)$ *dominates,*

$$\lim_{L\to\infty} \frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} = 0, \tag{7}$$

and so $(\mathbf{I}+\mathbf{H}\otimes\mathbf{A})$ *acts as a low-pass filter, as in Definition 2.1. If* ***(2)*** $(1+\lambda_j^H\lambda_1^A)$ *dominates,*

$$\lim_{L\to\infty}\frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2}\neq 0, \tag{8}$$

and so $(\mathbf{I}+\mathbf{H}\otimes\mathbf{A})$ *does not act as a low-pass filter. If* ***(3)*** *multiple eigenvalues have the same dominating magnitude, and there is at least one dominating eigenvalue* $(1+\lambda_j^H\lambda_i^A)$ *where* $\lambda_i^A\neq\lambda_n^A$*, then eq. (8) holds and* $(\mathbf{I}+\mathbf{H}\otimes\mathbf{A})$ *does not act as a low-pass filter.*

Proof. If **(1)**, as $L\to\infty$ we have from Theorem 3.4 that,

$$\lim_{L\to\infty}\mathbf{X}_L=(1+\lambda_j^H\lambda_n^A)^L s_{j,n}\mathbf{1}\mathbf{v}_j^{H^\top}. \tag{9}$$

If we plug this into the expression in Definition 2.1 of a low-pass filter we get,

$$\begin{aligned}
\lim_{L\to\infty}\frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} &= \lim_{L\to\infty}\sqrt{\frac{\|\text{HFC}[\mathbf{X}_L]\|_2^2}{\|\mathbf{X}_L-\text{HFC}[\mathbf{X}_L]\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{X}_L\|_2^2}{\|\mathbf{X}_L-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{X}_L\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)(1+\lambda_j^H\lambda_n^A)^L s_{j,n}\mathbf{1}\mathbf{v}_j^{H^\top}\|_2^2}{\|\mathbf{X}_L-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)(1+\lambda_j^H\lambda_n^A)^L s_{j,n}\mathbf{1}\mathbf{v}_j^{H^\top}\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(1+\lambda_j^H\lambda_n^A)^L s_{j,n}(\mathbf{1}\mathbf{v}_j^{H^\top}-\mathbf{1}\mathbf{v}_j^{H^\top})\|_2^2}{\|\mathbf{X}_L-(1+\lambda_j^H\lambda_n^A)^L s_{j,n}(\mathbf{1}\mathbf{v}_j^{H^\top}-\mathbf{1}\mathbf{v}_j^{H^\top})\|_2^2}}\\
&= 0,
\end{aligned}$$

where (17) is due to the fact that $(1/n)\mathbf{1}\mathbf{1}^\top\mathbf{M}$ averages the columns of any matrix $\mathbf{M}\in\mathbb{R}^{n\times r}$. This means that $(1/n)\mathbf{1}\mathbf{1}^\top\mathbf{1}\mathbf{v}_j^{H^\top}=\mathbf{1}\mathbf{v}_j^{H^\top}$ as $\mathbf{1}\mathbf{v}_j^{H^\top}$ has identical values in each column.

If **(2)** we have from Theorem 3.4 that,

$$\lim_{L\to\infty}\mathbf{X}_L=(1+\lambda_j^H\lambda_1^A)^L s_{j,1}\mathbf{v}_1^A\mathbf{v}_j^{H^\top}. \tag{10}$$

Plugging this into Definition 2.1 we get,

$$\begin{aligned}
\lim_{L\to\infty}\frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} &= \lim_{L\to\infty}\sqrt{\frac{\|\text{HFC}[\mathbf{X}_L]\|_2^2}{\|\mathbf{X}_L-\text{HFC}[\mathbf{X}_L]\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{X}_L\|_2^2}{\|\mathbf{X}_L-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{X}_L\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)(1+\lambda_j^H\lambda_1^A)^L s_{j,1}\mathbf{v}_1^A\mathbf{v}_j^{H^\top}\|_2^2}{\|\mathbf{X}_L-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)(1+\lambda_j^H\lambda_1^A)^L s_{j,1}\mathbf{v}_1^A\mathbf{v}_j^{H^\top}\|_2^2}}\\
&= \lim_{L\to\infty}\sqrt{\frac{\|(1+\lambda_j^H\lambda_1^A)^L s_{j,1}(\mathbf{v}_1^A\mathbf{v}_j^{H^\top}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{v}_1^A\mathbf{v}_j^{H^\top})\|_2^2}{\|\mathbf{X}_L-(1+\lambda_j^H\lambda_1^A)^L s_{j,1}(\mathbf{v}_1^A\mathbf{v}_j^{H^\top}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{v}_1^A\mathbf{v}_j^{H^\top})\|_2^2}}\\
&\neq 0.
\end{aligned}$$

The final line holds because in general $(1/n)\mathbf{1}\mathbf{1}^\top\mathbf{v}_1^A\mathbf{v}_j^{H^\top}\neq\mathbf{v}_1^A\mathbf{v}_j^{H^\top}$, unless $\mathbf{v}_1^A=c\mathbf{1}$ for some $c\in\mathbb{R}$. However, this is impossible given Assumption 3.1, as the Perron-Frobenius Theorem states that there is only one eigenvector of $\mathbf{A}$ that has all positive real entries. As we know $\mathbf{v}_n^A=\mathbf{1}$, there is no other eigenvector of $\mathbf{A}$ such that $\mathbf{v}_i^A=c\mathbf{1}$. Therefore, $\lim_{L\to\infty}\frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2}>0$.

If instead, **(3)**, then by the definition of $\mathcal{E}$ in Corollary **??** we have that,

$$\lim_{L\to\infty} \mathbf{X}_L = \sum_{(j,i)\in\mathcal{E}} (1+\lambda_j^H\lambda_i^A)^L s_{j,i} \mathbf{v}_i^A \mathbf{v}_j^{H^\top}. \tag{11}$$

Therefore,

$$\lim_{L\to\infty} \frac{\|\text{HFC}[\mathbf{X}_L]\|_2}{\|\text{LFC}[\mathbf{X}_L]\|_2} = \lim_{L\to\infty} \sqrt{\frac{\|\sum_{(j,i)\in\mathcal{E}}(1+\lambda_j^H\lambda_j^A)^L s_{j,1}(\mathbf{v}_i^A\mathbf{v}_j^{H^\top} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{v}_i^A\mathbf{v}_j^{H^\top})\|_2^2}{\|\mathbf{X}_L - \sum_{(j,i)\in\mathcal{E}}(1+\lambda_j^H\lambda_j^A)^L s_{j,1}(\mathbf{v}_i^A\mathbf{v}_j^{H^\top} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{v}_i^A\mathbf{v}_j^{H^\top})\|_2^2}}$$
$$\neq 0.$$

The final line follows so long as $\mathbf{v}_i^A\mathbf{v}_j^{H^\top} \neq \mathbf{1}\mathbf{v}_j^{H^\top}$ for at least one $(j,i)\in\mathcal{E}$. If this is true then we have one term in the sums above for which $(1/n)\mathbf{1}\mathbf{1}^\top\mathbf{v}_i^A\mathbf{v}_j^{H^\top} \neq \mathbf{v}_i^A\mathbf{v}_j^{H^\top}$. This is because $\mathbf{v}_i^A \neq c\mathbf{1}$ (by Assumption 3.1 and the Perron-Frobenius Theorem, as described in the proof of condition **(2)**). As we know that there is at least one dominating eigenvalue $(1+\lambda_j^H\lambda_i^A)$ where $\lambda_i^A \neq \lambda_n^A$ (this is given in the Theorem statement), then $\mathbf{v}_i^A\mathbf{v}_j^{H^\top} \neq \mathbf{1}\mathbf{v}_j^{H^\top}$, and so we are done. □

C Additional Theorems

If $\lambda^H \in \mathbb{R}$. The following is a special case of Theorem B.6 where all eigenvalues of $\mathbf{H}$ are real.

Theorem C.1 (eigenvalues $\lambda^H \in \mathbb{R}$)**.** *Consider the Transformer update with fixed $\mathbf{A} > 0, \mathbf{H} := \mathbf{W}_{\text{proj}}^\top \mathbf{W}_V^\top$, as described in eq. (1). Let $\{\lambda_i^A, \mathbf{v}_i^A\}_{i=1}^n$ and $\{\lambda_j^H, \mathbf{v}_j^H\}_{j=1}^d$ be the eigenvalue and eigenvectors of $\mathbf{A}$ and $\mathbf{H}$. Let the eigenvalues (and associated eigenvectors) be sorted in ascending order i.e., $\lambda_1^A \leq \cdots \leq \lambda_n^A$ and $\lambda_1^H \leq \cdots \leq \lambda_d^H$. Let the eigendecomposition of $(\mathbf{I} + \mathbf{W}_{\text{proj}}^\top \mathbf{W}_V^\top \otimes \mathbf{A})$ be $\mathbf{Q}\Lambda\mathbf{Q}^{-1}$, where $\Lambda_{ji} = (1+\lambda_j^H\lambda_i^A)$. As the number of total layers $L\to\infty$, one of four possible eigenvalues dominate the rest (multiple dominate if there are ties):*

$$\begin{cases}
(1+\lambda_d^H\lambda_n^A) & & & \text{if } \lambda_j^H > 0,\ \forall j \in \{1,\dots,d\} \\
(1+\lambda_d^H\lambda_n^A) & & \text{if } \lambda_d^H + 2 > |\lambda_1^H| & \\
(1+\lambda_1^H\lambda_1^A) & \text{if } |1+\lambda_1^H\lambda_1^A| > |1+\lambda_1^H\lambda_n^A| & & \text{if } \lambda_1^H < 0, \lambda_d^H > 0 \\
(1+\lambda_1^H\lambda_n^A) & \text{if } |1+\lambda_1^H\lambda_1^A| < |1+\lambda_1^H\lambda_n^A| & \text{if } \lambda_d^H + 2 < |\lambda_1^H| & \\
(1+\lambda_d^H\lambda_1^A) & \text{if } |1+\lambda_d^H\lambda_1^A| > |1+\lambda_1^H\lambda_n^A| & \text{if } \lambda_1^A > 0, \lambda_1^H > -2 & \\
(1+\lambda_1^H\lambda_n^A) & \text{if } |1+\lambda_d^H\lambda_1^A| < |1+\lambda_1^H\lambda_n^A| & & \\
(1+\lambda_1^H\lambda_n^A) & & \text{if } \lambda_1^A > 0, \lambda_1^H < -2 & \\
(1+\lambda_1^H\lambda_1^A) & & \text{if } \lambda_1^A < 0, \lambda_1^H > -2 & \text{if } \lambda_j^H < 0,\ \forall j \in \{1,\dots,d\} \\
(1+\lambda_1^H\lambda_1^A) & \text{if } |1+\lambda_1^H\lambda_1^A| > |1+\lambda_1^H\lambda_n^A| & & \\
(1+\lambda_1^H\lambda_n^A) & \text{if } |1+\lambda_1^H\lambda_1^A| < |1+\lambda_1^H\lambda_n^A| & \text{if } \lambda_1^A < 0, \lambda_1^H < -2 &
\end{cases}$$

Proof. Our goal is again to characterize the identity of $\lambda_{j^*}^H\lambda_{i^*}^A$ where $|1+\lambda_{j^*}^H\lambda_{i^*}^A| > |1+\lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \in \{1,\dots,d\}\setminus j^*$ and $i' \in \{1,\dots,n\}\setminus i^*$, for all ranges of λ_H, λ_A. This is because $(1+\lambda_{j^*}^H\lambda_{i^*}^A)^L$ will dominate as $L\to\infty$. We will again assume there are no ties i.e., j^*, i^* each denote a single index. Given this, we detail each case described in the theorem statement.

if $\lambda_j^H > 0$ for all $j \in \{1,\dots,d\}$. First recall that $\lambda_i^A \in (-1,1]$ and $\lambda_n^A = 1$. As $\lambda_j^H > 0$, we have that $|1+\lambda_d^H\lambda_n^A| > |1+\lambda_{j'}^H\lambda_{i'}^A|$ for all $j' \in \{1,\dots,d-1\}$ and $i' \in \{1,\dots,n-1\}$. This is because, by definition $\lambda_d^H\lambda_n^A > \lambda_{j'}^H\lambda_{i'}^A$. Further, $1+\lambda_d^H\lambda_n^A > |1+\lambda_{j'}^H\lambda_{i'}^A|$ as the largest $|1+\lambda_{j'}^H\lambda_{i'}^A|$ can be is either (i) $|1-\epsilon\lambda_d^H|$ for $0<\epsilon<1$ or (ii) $|1+\lambda_{d-1}^H\lambda_n^A|$ (i.e., in (i) λ_d^H

	CIFAR100					The Pile			
Model	ViT-Ti	FeatScale	Cent. Attn.	**ViT-Ti$^-$**	**ViT-Ti$^+$**	Cram. Bert	**Cram. Bert$^-$**	**Cram. Bert$^+$**	**Cram. Bert$^\sim$**
$(1+\lambda_j^H\lambda_n^A)$	73.99%	100%	88.51%	0%	100%	90%	0%	100%	100%
$(1+\lambda_j^H\lambda_1^A)$	26.01%	0%	11.49%	100%	0%	10%	100%	0%	0%

Table 1: **Distribution of dominating eigenvalues.** We compare different vision models trained on CIFAR100 and language models trained on The Pile, and count the percentage of cases where the dominating eigenvalue is $(1+\lambda_j^H\lambda_n^A)$ or $(1+\lambda_j^H\lambda_1^A)$.

is negated by λ_1^A and in (ii) λ_{d-1}^H is the next largest value of λ^H). For (i), it must be that $1+\lambda_d^H\lambda_n^A > |1-\epsilon\lambda_d^H|$ as $\lambda_d^H > 0$. For (ii) $\lambda_d^H > \lambda_{d-1}^H > 0$ (as we assume there are no ties), and so $|1+\lambda_d^H\lambda_n^A| > |1+\lambda_{d-1}^H\lambda_n^A|$. Therefore $(1+\lambda_d^H\lambda_n^A)$ dominates.

if $\lambda_1^H < 0, \lambda_d^H > 0$. Recall we can view selecting $\lambda_j^H\lambda_i^A$ to maximize $|1+\lambda_j^H\lambda_i^A|$ as maximizing distance to -1. In this condition the maximal $\lambda_{j^*}^H\lambda_{i^*}^A$ depends on whether λ_1^H or λ_d^H is farther away from -1. If λ_d^H is farther from -1, i.e., $|1+\lambda_d^H\lambda_n^A| > |1+\lambda_1^H\lambda_n^A|$ (which can be simplified to $\lambda_d^H + 2 > |\lambda_1^H|$) , then $|1+\lambda_d^H\lambda_n^A|$ is maximal because (a) any other $\lambda_{i'}^A$ will move λ_d^H closer to -1, and (b) any other $\lambda_{j'}^H$ is closer to -1. So $(1+\lambda_d^H\lambda_n^A)$ dominates. If λ_1^H is farther from -1, i.e., $|1+\lambda_d^H\lambda_n^A| < |1+\lambda_1^H\lambda_n^A|$ (which can be simplified to $\lambda_d^H + 2 < |\lambda_1^H|$) , then it depends on whether λ_1^A can push λ_1^H farther away from -1 than λ_1^H is itself (sidenote: this will only happen for $\lambda_1^A < 0$, when it can 'flip' λ_1^H across the origin, because by definition it has to beat $\lambda_d^H > 0$). If it can, i.e., $|1+\lambda_1^H\lambda_1^A| > |1+\lambda_1^H\lambda_n^A|$ then $(1+\lambda_1^H\lambda_1^A)$ dominates. Otherwise, $|1+\lambda_1^H\lambda_1^A| < |1+\lambda_1^H\lambda_n^A|$ and $(1+\lambda_1^H\lambda_n^A)$ dominates.

if $\lambda_j^H < 0$ for all $j \in \{1,\dots,d\}$. In this case we need to know if (a) $\lambda_1^A > 0$ or (b) $\lambda_1^A < 0$. If (a) then all $\lambda_j^A > 0$ and so we cannot 'flip' λ^H across the origin. Because of this, if $\lambda_1^H > -2$ then we have that $\lambda_j^H\lambda_i^A \in (-2, 0)$ for all j. Note that $|1+\lambda_j^H\lambda_i^A|$ is symmetric in this interval around -1 so whichever $\lambda_j^H\lambda_i^A$ is closest to the ends of the interval will maximize $|1+\lambda_j^H\lambda_i^A|$. Note that $\lambda_1^H\lambda_n^A$ will be closest to -2 and $\lambda_d^H\lambda_1^A$ will be closest to 0. Therefore if $|1+\lambda_d^H\lambda_1^A| > |1+\lambda_1^H\lambda_n^A|$ then $(1+\lambda_d^H\lambda_1^A)$ will dominate. If the opposite is true then $(1+\lambda_1^H\lambda_n^A)$ will dominate. If instead $\lambda_1^H < -2$ then $\lambda_1^H\lambda_n^A$ is farthest from -1 as $\lambda_j^H\lambda_i^A < 0$, so $(1+\lambda_1^H\lambda_n^A)$ dominates. If case (b) and we have that $\lambda_1^A < 0$ and $\lambda_1^H > -2$ then $\lambda_1^H\lambda_1^A > 0$. This means that $|1+\lambda_1^H\lambda_1^A| > |1+\lambda_j^H\lambda_i^A|$ because any 'flip' of λ_j^H across the origin by $\lambda_i^A < 0$ makes $\lambda_j^H\lambda_i^A > \lambda_j^H\lambda_{i'}^A$ where $\lambda_{i'}^A > 0$. The flip that is largest is $\lambda_1^H\lambda_1^A > \lambda_j^H\lambda_i^A$, by definition of λ_1^H, λ_1^A. So $(1+\lambda_1^H\lambda_1^A)$ dominates. If instead $\lambda_1^A < 0$ and $\lambda_1^H < -2$ Then it depends on whether λ_1^A can 'flip' λ_1^H farther from -1 than λ_1^H is itself. If it can, then $(1+\lambda_1^H\lambda_1^A)$ dominates, otherwise $(1+\lambda_1^H\lambda_n^A)$ dominates. (For completeness, note that $\max\{|1+\lambda_1^H\lambda_1^A|, |1+\lambda_1^H\lambda_n^A|\} > |1+\lambda_d^H\lambda_n^A|$ because either $\lambda_d^H\lambda_n^A < -2$ in which case $|1+\lambda_1^H\lambda_n^A| > |1+\lambda_d^H\lambda_n^A|$ or $\lambda_d^H\lambda_n^A \in (-2, 0)$ in which case $|1+\lambda_1^H\lambda_1^A| > |1+\lambda_d^H\lambda_n^A|$. Also note that $|1+\lambda_1^H\lambda_1^A| > |1+\lambda_d^H\lambda_1^A|$ as λ_d^H is closer to the origin than λ_1^H).

As these cases define a partition of λ^H and λ^A, we are done. □

D Additional Results

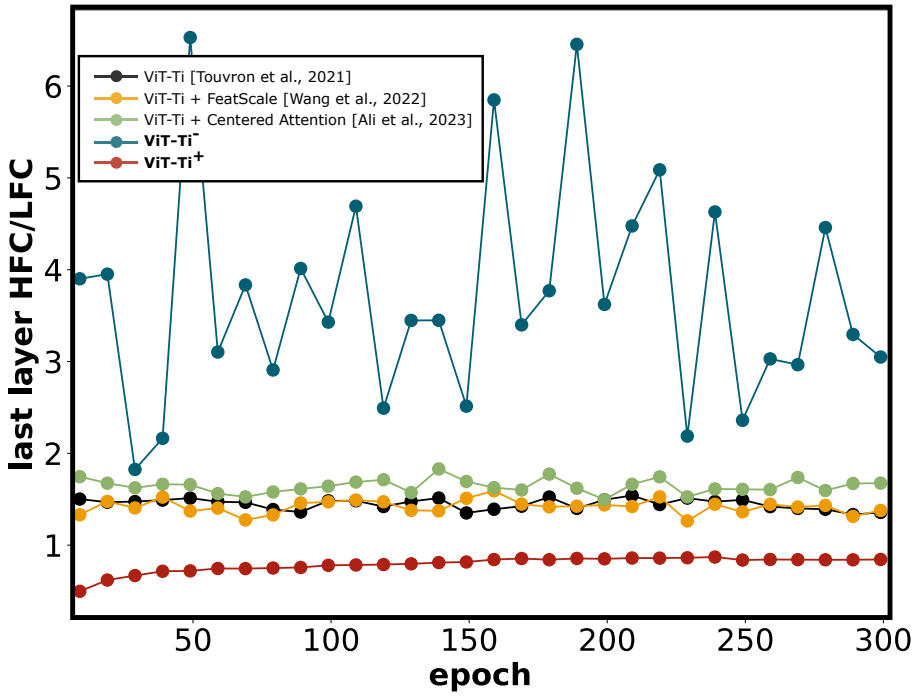

Figure 4: **Filtering during training.** $\frac{\|\mathrm{HFC}[\mathbf{X}_\ell]\|_2}{\|\mathrm{LFC}[\mathbf{X}_\ell]\|_2}$ of the last layer for CIFAR100 models during training.

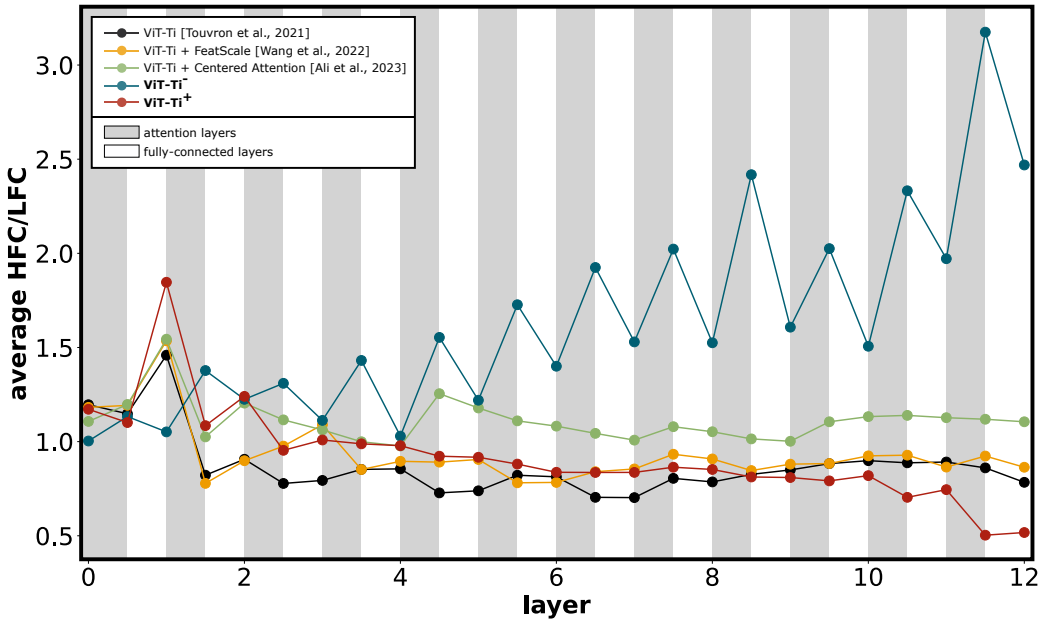

Figure 5: **Filtering, image classification.** $\frac{\|\mathrm{HFC}[\mathbf{X}_\ell]\|_2}{\|\mathrm{LFC}[\mathbf{X}_\ell]\|_2}$ for different models on CIFAR100.

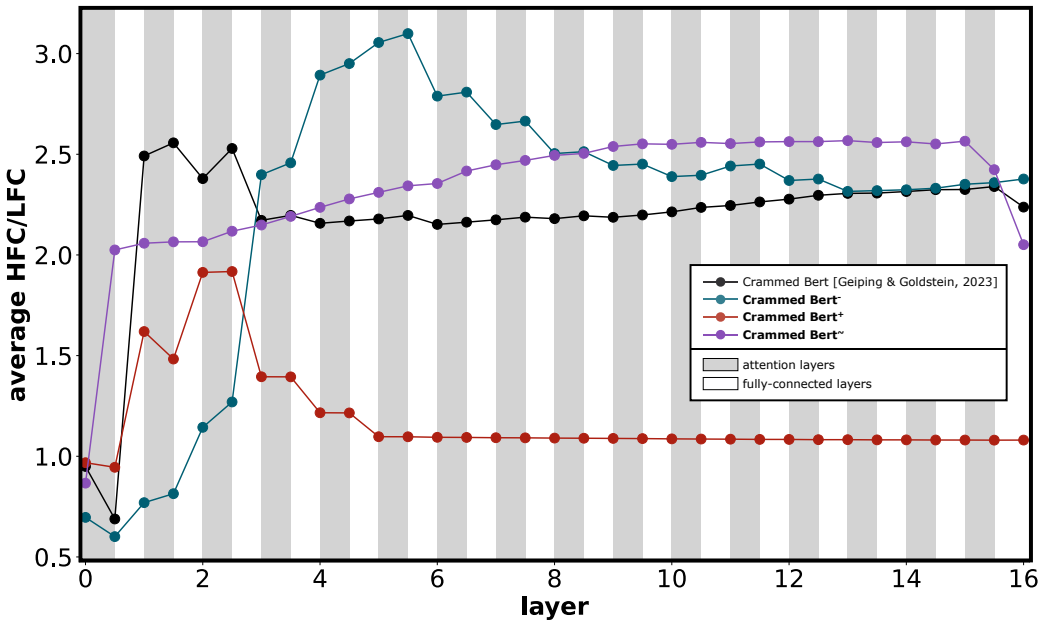

Figure 6: **Filtering, text generation.** $\frac{\|\mathrm{HFC}[\mathbf{X}_\ell]\|_2}{\|\mathrm{LFC}[\mathbf{X}_\ell]\|_2}$ for different models on The Pile.

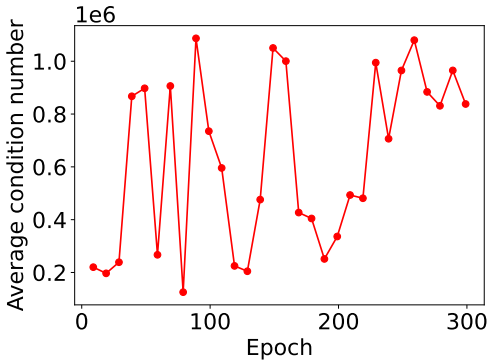

Figure 7: The average condition number of all $\mathbf{H}$ for ViT-Ti throughout training on CIFAR100.