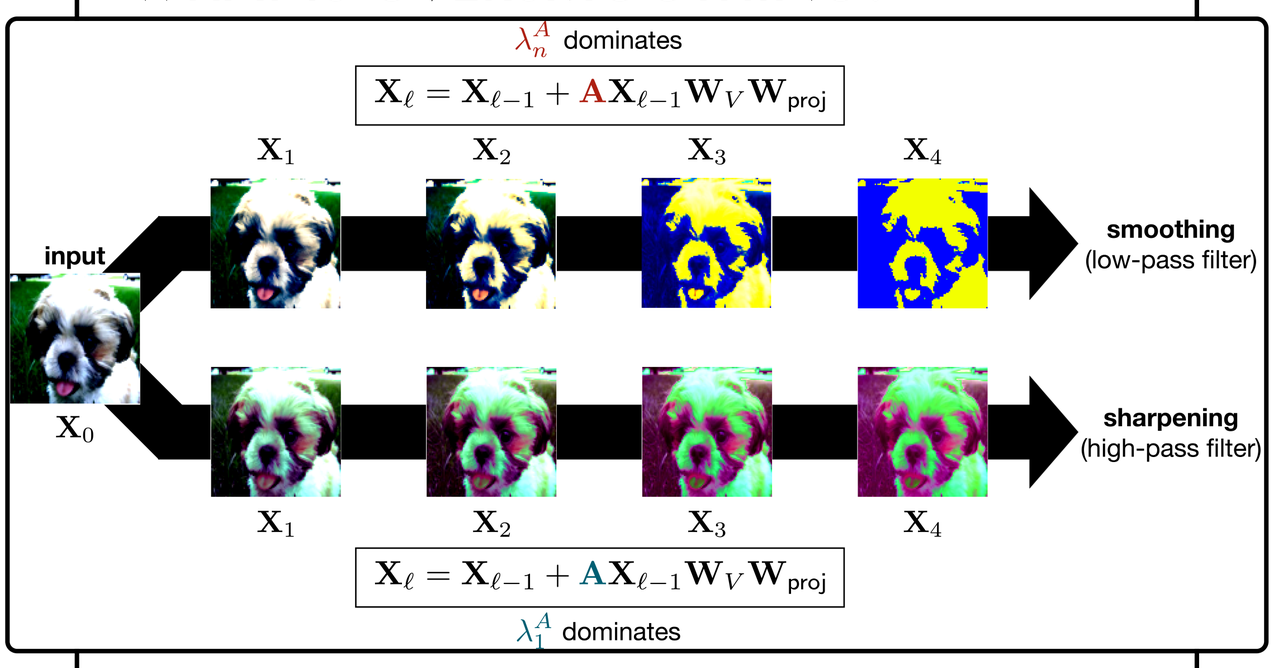
Setting the Record Straight on Transformer Oversmoothing

Gbetondji J-S Dovonon[1], Michael M Bronstein[2], Matt J Kusner[1]

[1]University College London
[2]University of Oxford

Recent work has argued that transformers are inherently low-pass filters. That is not always the case and we show why that matters for robustness and data efficiency.

WHAT IS OVERSMOOTHING?



Low Frequency component $\text{LFC}[\mathbf{X}] := (1/n)\mathbf{1}\mathbf{1}^\top\mathbf{X}.$

High Frequency component $\text{HFC}[\mathbf{X}] := (\mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^\top)\mathbf{X}.$

Oversmoothing happens when the LFC dominates the HFC

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

Issue: Smoothing reduces detail, making classification difficult.
How do vision transformers work if they oversmooth?
They learn not to.

We analyze a single self attention layer with a skip connection

$$\mathbf{A} := \text{Softmax}\Big(\frac{1}{\sqrt{k}}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top\Big)$$

$$\mathbf{X}_\ell := \mathbf{X}_{\ell-1} + \mathbf{A}_\ell\mathbf{X}_{\ell-1}\mathbf{W}_{V,\ell}\mathbf{W}_{\text{proj},\ell}$$

We can rewrite it in vector form using the vec operator and Kronecker product

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

MAIN THEORETICAL RESULTS

$\lambda_1^A \le \cdots \le \lambda_n^A$ and $\lambda_1^H \le \cdots \le \lambda_d^H$ the eigenvalues of A and H

$(1 + \lambda_j^H\lambda_n^A)$ dominates → over smoothing

$(1 + \lambda_j^H\lambda_1^A)$ dominates → no over smoothing

$\lambda_j^H \in [-1, 0), \forall j$ guarantees we avoid over smoothing

TRANSFORMERS LEARN TO AVOID OVERSMOOTHING

	Random Init.	ImageNet		CIFAR100		
Model	ViT-B/16	ViT-B/16	DeiT-B/16	ViT-Ti	ViT-Ti$^+$	**ViT-Ti$^-$**
$(1 + \lambda_j^H\lambda_n^A)$	100%	70.27%	57.82%	98.92%	100%	0%
$(1 + \lambda_j^H\lambda_1^A)$	0%	29.73%	42.18%	1.08%	0%	100%

Table 1: **Distribution of dominating eigenvalues.** We compare different models trained (or not) on ImageNet and CIFAR100 and count the percentage of cases where the dominating eigenvalue is $(1 + \lambda_j^H\lambda_n^A)$ or $(1 + \lambda_j^H\lambda_1^A)$.

A NEW PARAMETERIZATION

We parameterize H as a symmetric eigendecomposition

$$\mathbf{H} = \mathbf{V}_H\Lambda_H\mathbf{V}_H^\top$$

$$[\mathbf{V}_H, \mathbf{R}] := \text{QR}(\Theta)\,\text{diag}(\Lambda_H) := \pm(\psi^2)$$

Learnable parameters

We refer to the model with negative eigenvalues Λ_H as ViT $^-$ and the one with positive eigenvalues as ViT $^+$

CIFAR100 RESULTS

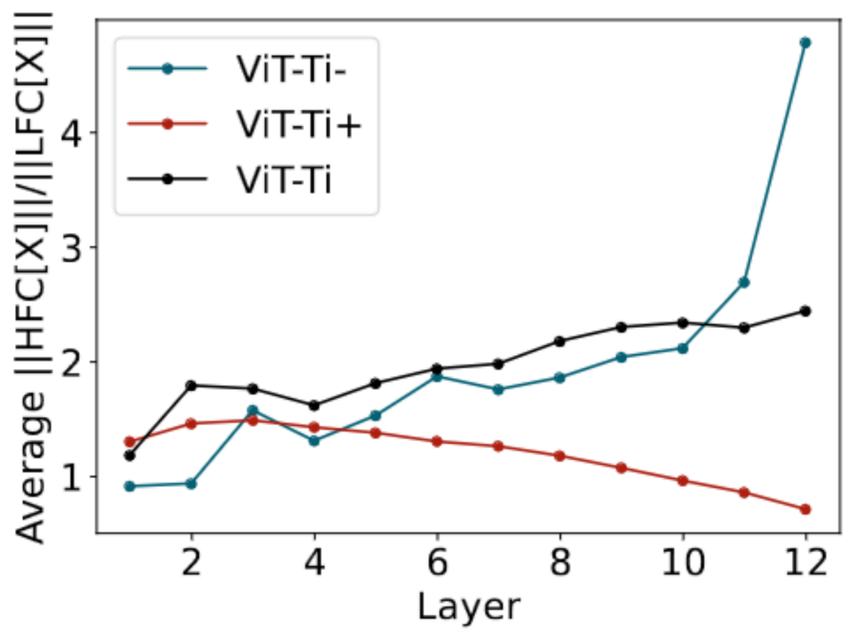
Data kept	100%	50%	10%
ViT-Ti	$\mathbf{66.78}_{\pm 0.1}$	$\mathbf{56.38}_{\pm 0.5}$	$32.47_{\pm 0.9}$
ViT-Ti$^+$	$64.34_{\pm 0.3}$	$55.21_{\pm 0.5}$	$28.67_{\pm 0.5}$
ViT-Ti$^-$	$\mathbf{66.62}_{\pm 0.3}$	$\mathbf{56.58}_{\pm 0.3}$	$\mathbf{33.99}_{\pm 0.4}$

Table 2: **Data efficiency.** Results for different percentages of the CIFAR100 used in training.

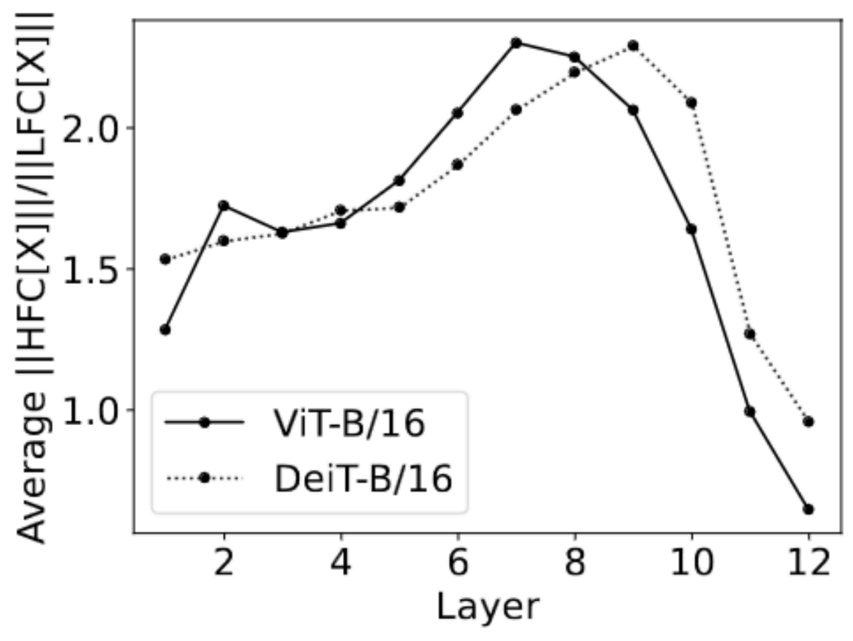
Corruption Intensity	0	1	2	3	4	5
ViT-Ti	$\mathbf{66.78}_{\pm 0.1}$	$\mathbf{61.59}_{\pm 0.2}$	$57.75_{\pm 0.2}$	$53.12_{\pm 0.1}$	$49.32_{\pm 0.1}$	$39.17_{\pm 0.2}$
ViT-Ti$^+$	$64.34_{\pm 0.3}$	$58.36_{\pm 0.4}$	$53.83_{\pm 0.5}$	$49.03_{\pm 0.6}$	$45.15_{\pm 0.6}$	$35.86_{\pm 0.6}$
ViT-Ti$^-$	$\mathbf{66.62}_{\pm 0.3}$	$\mathbf{61.62}_{\pm 0.2}$	$\mathbf{58.00}_{\pm 0.3}$	$\mathbf{53.56}_{\pm 0.2}$	$\mathbf{49.70}_{\pm 0.3}$	$\mathbf{40.07}_{\pm 0.2}$

Table 3: **Corruption Robustness.** Results for different corruption intensities on CIFAR100.

MEASURING OVERSMOOTHING



(b) CIFAR100 (trained)



(c) ImageNet (trained)