Laplacian Heads Improve Transformers by Smoothing Token Representations

Yuchong Zhang Vardan Papyan

University of Toronto · Vector Institute

NeurIPS 2026

Our Proposal

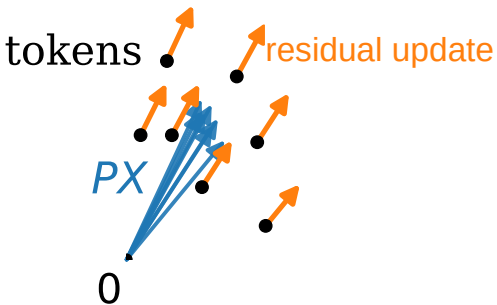
Attention	Laplacian
$P = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$	$I - P$
PV	$(I - P)V$

Forward pass:

$$X \leftarrow X + \sum_{i \in \mathcal{A}} P^{(i)} \,\mathrm{Norm}(X) W_{V_i} W_{o_i} + \sum_{i \in \mathcal{L}} (I - P^{(i)}) \,\mathrm{Norm}(X) W_{V_i} W_{o_i}.$$

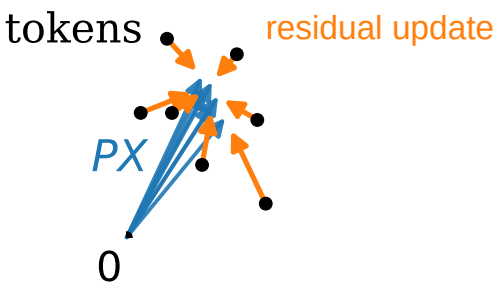
Updating Mean and Within-sequence Variance

$$X + P^{(i)}\,\mathrm{Norm}(X)W_{V_i}W_{o_i}$$



Attention

$$X + (I - P^{(i)})\,\mathrm{Norm}(X)W_{V_i}W_{o_i}$$



Laplacian

Attention-induced Token Graph

Tokens are vertices, and $P \in \mathbb{R}^{T \times T}$ is the edge weights.

$I - P$: the random-walk normalized graph Laplacian.

$x \in \mathbb{R}^{T}$: a signal over the tokens.

Diffusion over Graphs

The heat equation:

$$\frac{d}{dt}\,x(t) = -(I - P)\,x(t),$$

Euler discretization, $\Delta t = 1$:

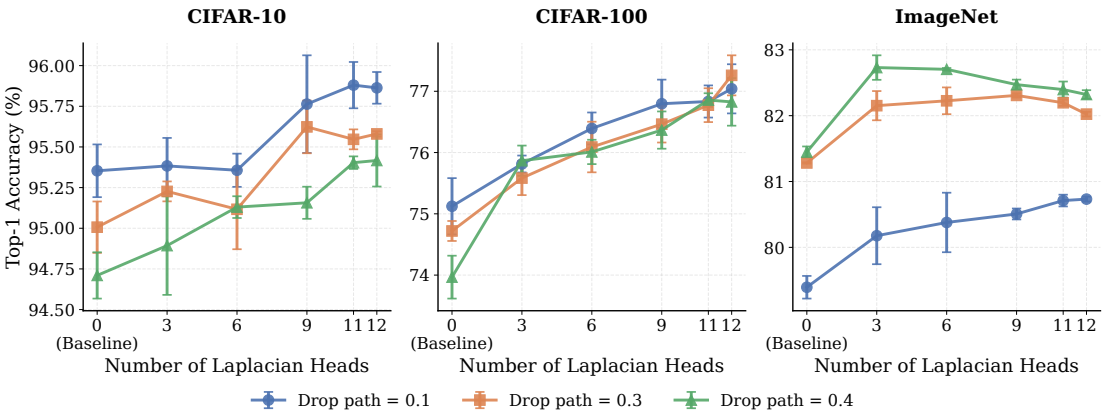
$$x_{t+1} = x_t - (I - P)x_t = Px_t.$$

Laplacian-head residual update:

$$X_{t+1} = X_t + \sum_i (I - P^{(i)})\,\mathrm{Norm}(X_t)W_{V_i}W_{o_i}.$$

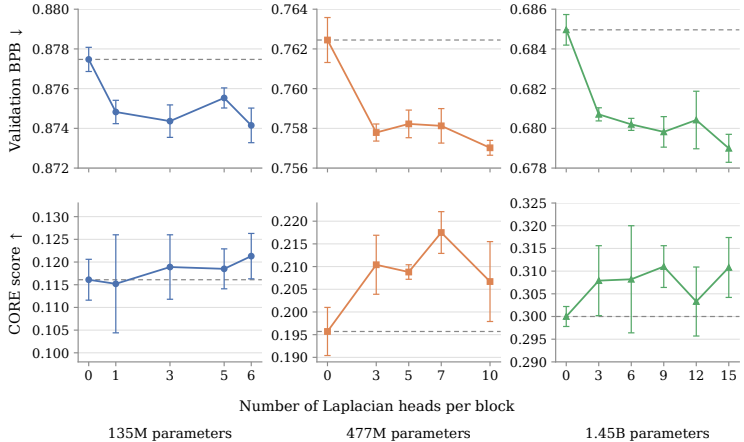
Laplacian heads improve performance of transformers on various tasks

Image Classification



Better top-1 accuracy for DeiT-3.

Language Model Pretraining



Lower validation BPB and better downstream performance.

DINO Vision SSL

	Number of Laplacian heads				
Task	0	1	3	5	6
ImageNet Top-1 (%)	72.08 ± 0.18	73.05 ± 0.27	**73.34 ± 0.20**	$\underline{73.12 \pm 0.13}$	72.94 ± 0.09
ADE20K mIoU (%)	30.27 ± 0.70	31.05 ± 0.57	**32.11 ± 0.54**	$\underline{31.83 \pm 0.82}$	30.04 ± 1.08

ViT-S trained on ImageNet. Better downstream performance for classification and semantic segmentation.

We study the effect of Laplacian heads on token representations

PCA of Token Representations



(a) Attention (b) Laplacian

ViT trained on CIFAR-10. Colors are classes.

Analysis of Variance (ANOVA)

$$\mu_{i,c} = \mathbb{E}_t[X_{t,i,c}], \qquad \mu_c = \mathbb{E}_i[\mu_{i,c}], \qquad \mu_G = \mathbb{E}_c[\mu_c].$$

$$\text{WithinSeqVar} = \mathbb{E}_{t,i,c}\left[\|X_{t,i,c} - \mu_{i,c}\|_2^2\right],$$

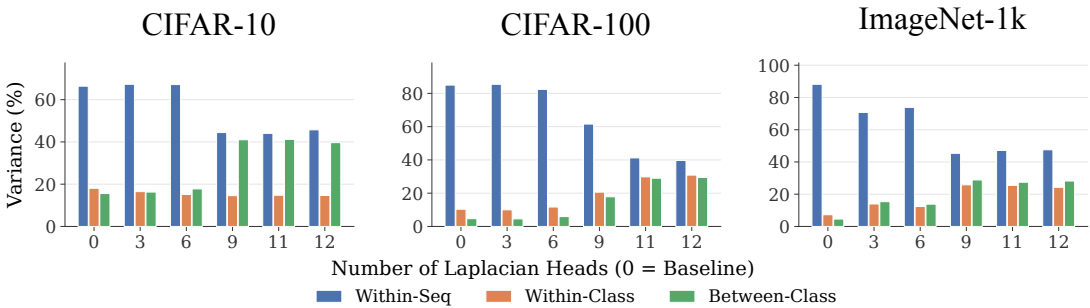
$$\text{WithinClassVar} = \mathbb{E}_{i,c}\left[\|\mu_{i,c} - \mu_c\|_2^2\right],$$

$$\text{BetweenClassVar} = \mathbb{E}_c\left[\|\mu_c - \mu_G\|_2^2\right],$$

$$\text{TotalVar} = \mathbb{E}_{t,i,c}\left[\|X_{t,i,c} - \mu_G\|_2^2\right].$$

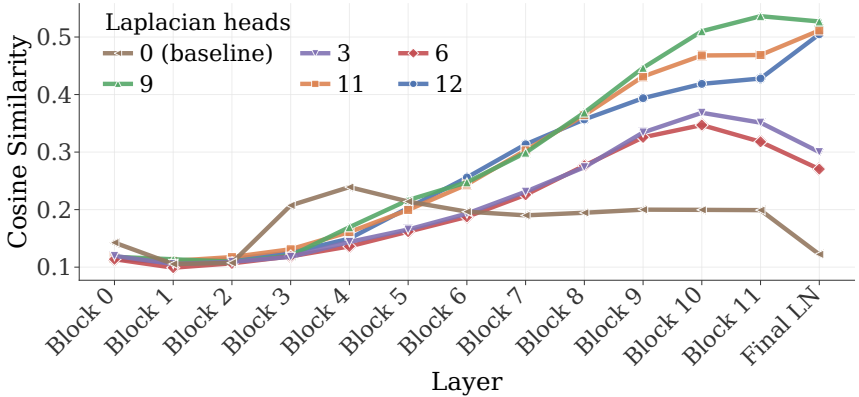
$$\text{TotalVar} = \text{BetweenClassVar} + \text{WithinClassVar} + \text{WithinSeqVar}.$$

Variance Moves from WithinSeq to BetweenClass



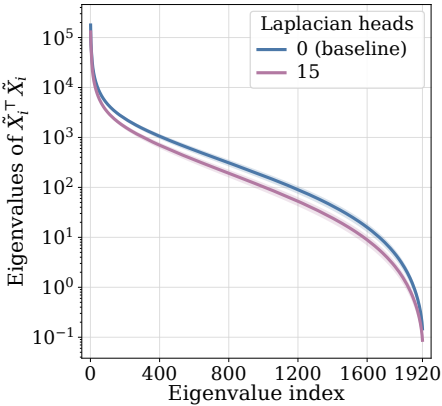
Each component is shown as a percentage of $\mathrm{TotalVar}$.

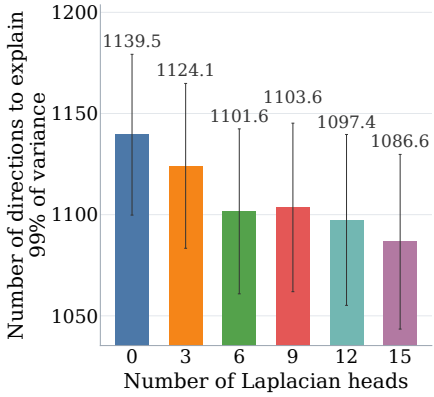
Laplacian Heads Collapse Tokens Across Depth



Average pairwise cosine similarity of tokens (DeiT-3 trained on ImageNet).

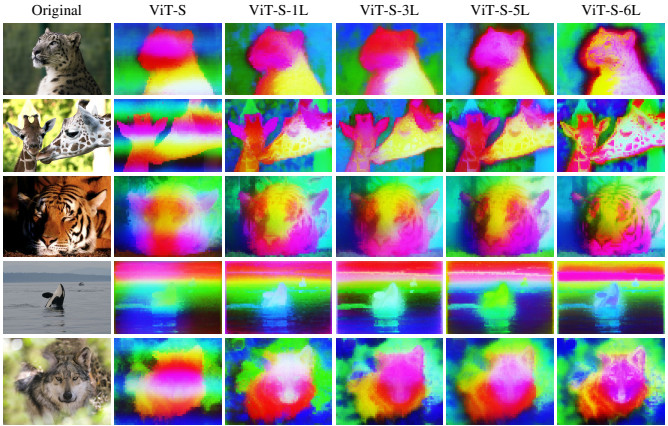
Faster-decaying Spectrum for LLM Token Representations





Fewer PCs explain 99% of token variance (1.45B model).

DINO: More Segmentation Structure in Top PCs



Token projections onto the top three principal components, mapped to RGB.

Conclusion

- Token smoothing is not inherently harmful.
- Laplacian heads improve performance across tasks through stronger smoothing.
- The geometry tokens converge towards depends on architecture and training objective.

Thank you! Visit our poster session:

Thu, Dec 10th 10:00 AM – 1:00 PM
Hall 1–4, ICC Sydney

Paper



Code

