An In-depth Investigation of Sparse Rate Reduction in Transformer-like Models

SCHOOL OF COMPUTING & DATA SCIENCE
The University of Hong Kong

Yunzhe Hu, Difan Zou, Dong Xu

NEURAL INFORMATION PROCESSING SYSTEMS

Background

- **Design Transformer-like Models via Sparse Rate Reduction**

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{d\times N}} R(\boldsymbol{Z}) - R^c(\boldsymbol{Z};\boldsymbol{U}) - \lambda\|\boldsymbol{Z}\|_0$$

$$R(\boldsymbol{Z}) \doteq \tfrac{1}{2}\log\det(\boldsymbol{I} + \tfrac{d}{N\epsilon^2}\boldsymbol{Z}^T\boldsymbol{Z}) \qquad \boldsymbol{U} = [\boldsymbol{U}_1,\dots,\boldsymbol{U}_K] \in \mathbb{R}^{d\times Kp}$$

$$R^c(\boldsymbol{Z};\boldsymbol{U}) \doteq \textstyle\sum_{k=1}^K R(\boldsymbol{U}_k^T\boldsymbol{Z})$$

- **Interpret Multi-head Self-attention as One Step Gradient Descent on $R^c(\boldsymbol{Z};\boldsymbol{U})$**

☐ Minimize $R^c(\boldsymbol{Z};\boldsymbol{U})$

$$\boldsymbol{Y}^\ell = \boldsymbol{Z}^{\ell-1} - \alpha\nabla R^c(\boldsymbol{Z}^{\ell-1};\boldsymbol{U}^\ell) \approx \boldsymbol{Z}^{\ell-1} + \alpha\gamma^2\,\mathrm{MSSA}(\boldsymbol{Z}^{\ell-1};\boldsymbol{U}^\ell)$$

$$\mathrm{MSSA}(\boldsymbol{Z};\boldsymbol{U}) = \sum_{k=1}^K \boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_k^T\boldsymbol{Z})^T(\boldsymbol{U}_k^T\boldsymbol{Z}))$$

$$= [\boldsymbol{U}_1,\dots,\boldsymbol{U}_K]\begin{bmatrix}\boldsymbol{U}_1^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_1^T\boldsymbol{Z})^T(\boldsymbol{U}_1^T\boldsymbol{Z}))\\ \vdots \\ \boldsymbol{U}_K^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_K^T\boldsymbol{Z})^T(\boldsymbol{U}_K^T\boldsymbol{Z}))\end{bmatrix}$$

☐ Minimize $\lambda\|\boldsymbol{Z}\|_0 - R(\boldsymbol{Z})$

$$\boldsymbol{Z}^\ell = \mathrm{ReLU}\left(\boldsymbol{Y}^\ell + \beta(\boldsymbol{D}^\ell)^T(\boldsymbol{Y}^\ell - \boldsymbol{D}^\ell\boldsymbol{Y}^\ell) - \beta\lambda\boldsymbol{1}\right) \qquad \boldsymbol{D}^\ell \in \mathbb{R}^{d\times d}$$

Problems

- Only focused on the basic implementation and lack of validating layer-wise optimization at both microscopic and macroscopic levels.
- Whether SRR can benefit generalization of its derived models in a causal and principled way is not clear.

Takeaways

- Alternative models exhibit similar behaviors, validating the optimization of SRR in the forward-pass.
- SRR could be guidance to design Transformer-like models, but not necessarily principled.
- SRR measure is a better indicator of generalization than flatness-based ones but non-trivial to be extended to standard Transformers.

Is Sparse Rate Reduction Optimized in Transformer-like Models?

- **Pitfalls in Deriving CRATE-C**

$$R^c(\boldsymbol{Z};\boldsymbol{U}) = \sum_{k=1}^K\sum_{i=1}^N \frac{1}{2}\log\lambda_i^k \geq \sum_{k=1}^K\sum_{i=1}^N \frac{1}{2}\left(\lambda_i^k - 1 - \frac{(\lambda_i^k-1)^2}{2}\right) = \sum_{k=1}^K \Bigg(\underbrace{\frac{\gamma}{2}\|\boldsymbol{U}_k^T\boldsymbol{Z}\|_F^2}_{\text{First-order term}} \underbrace{-\frac{\gamma^2}{4}\|(\boldsymbol{U}_k^T\boldsymbol{Z})^T\boldsymbol{U}_k^T\boldsymbol{Z}\|_F^2}_{\text{Second-order term}}\Bigg)$$

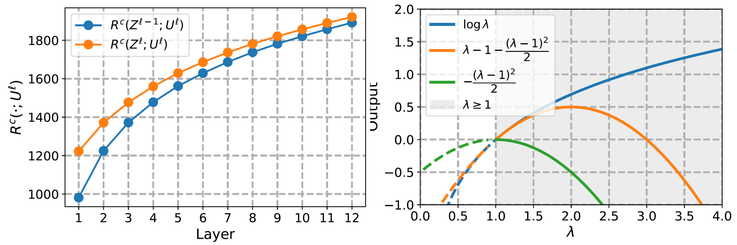


Figure 1: In an attention-only experiment, MSSA operator with skip connection implements an ascent method on $R^c(Z;U)$, opposed to its design purpose (*left*). This is due to an artifact in approximation with its second-order term. (*right*)

- **CRATE and Its Variants**

☐ CRATE-C

$$\boldsymbol{Z} + \alpha\gamma^2\sum_{k=1}^K \boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_k^T\boldsymbol{Z})^T(\boldsymbol{U}_k^T\boldsymbol{Z}))$$

☐ CRATE-N

$$\boldsymbol{Z} - \alpha\gamma^2\sum_{k=1}^K \boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_k^T\boldsymbol{Z})^T(\boldsymbol{U}_k^T\boldsymbol{Z}))$$

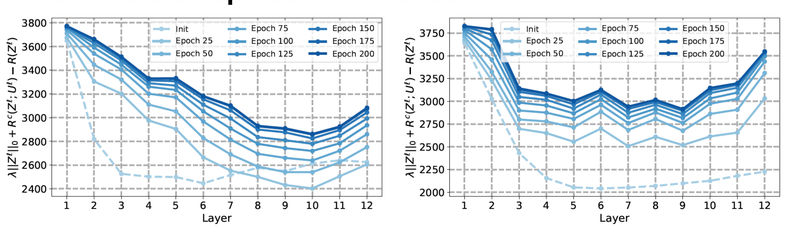
☐ CRATE-T

$$\boldsymbol{Z} + \alpha\gamma^2[\boldsymbol{U}_1,\dots,\boldsymbol{U}_K]^T\begin{bmatrix}\boldsymbol{U}_1^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_1^T\boldsymbol{Z})^T(\boldsymbol{U}_1^T\boldsymbol{Z}))\\ \vdots \\ \boldsymbol{U}_K^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_K^T\boldsymbol{Z})^T(\boldsymbol{U}_K^T\boldsymbol{Z}))\end{bmatrix}$$

☐ CRATE

$$\boldsymbol{Z} + \alpha\gamma^2\boldsymbol{W}\begin{bmatrix}\boldsymbol{U}_1^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_1^T\boldsymbol{Z})^T(\boldsymbol{U}_1^T\boldsymbol{Z}))\\ \vdots \\ \boldsymbol{U}_K^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_K^T\boldsymbol{Z})^T(\boldsymbol{U}_K^T\boldsymbol{Z}))\end{bmatrix}$$

- **Behaviors of Sparse Rate Reduction**



(a) Sparse rate reduction measure of CRATE-C

(b) Sparse rate reduction measure of CRATE-N

(c) Sparse rate reduction measure of CRATE-T

(d) Sparse rate reduction measure of CRATE

Figure 2: Sparse rate reduction measure of CRATE and its variants evaluated at different layers and epochs on CIFAR-10.

Whether Sparse Rate Reduction Benefits Generalization?

- **Sparse Rate Reduction as a Measure of Complexity**

$$\mu_{\mathrm{SRR}}(\boldsymbol{w};\boldsymbol{Z}) = \frac{1}{L}\sum_{\ell=1}^L \mu_{\mathrm{SRR}}^\ell(\boldsymbol{w}^\ell;\boldsymbol{Z}^\ell) = \frac{1}{L}\sum_{\ell=1}^L \left(\lambda\|\boldsymbol{Z}^\ell\|_0 + R^c(\boldsymbol{Z}^\ell;\boldsymbol{U}^\ell) - R(\boldsymbol{Z}^\ell)\right)$$

- **Correlation with Generalization**

☐ Kendall's coefficient

$$\mathcal{T} \triangleq \cup_{\boldsymbol{\theta}\in\Theta_1\times\cdots\times\Theta_n}\{(\mu(\boldsymbol{\theta}), g(\boldsymbol{\theta}))\},$$

$$\tau(\mathcal{T}) \triangleq \frac{1}{|\mathcal{T}|(|\mathcal{T}|-1)}\sum_{(\mu_1,g_1)\in\mathcal{T}}\sum_{(\mu_2,g_2)\in\mathcal{T}\setminus(\mu_1,g_1)} \mathrm{sign}\,(\mu_1-\mu_2)\,\mathrm{sign}\,(g_1-g_2)$$

☐ Results

Table 1: Correlation of complexity measures with generalization gap (width $d = 384$).

Complexity measures	Batch size	Learning rate	Dropout	Model type	Overall τ	Ψ
ℓ_2-norm	0.200	-0.333	-0.333	-0.429	-0.363	-0.224
ℓ_2-norm-init	0.200	-0.200	-0.333	-0.286	-0.290	-0.158
# params	0.000	0.000	0.000	-0.572	-0.351	-0.143
1/margin	-0.067	0.467	0.467	0.238	0.415	0.276
sum-of-spec	0.200	-0.333	-0.467	-0.381	-0.290	-0.245
prod-of-spec	0.200	-0.333	-0.467	-0.476	-0.338	-0.269
sum-of-spec/margin	0.333	-0.333	-0.467	-0.048	-0.230	-0.129
prod-of-spec/margin	0.333	-0.333	-0.467	-0.143	-0.260	-0.152
fro/spec	-0.200	0.333	0.467	-0.476	0.019	0.031
spec-init-main	0.333	-0.333	-0.467	-0.190	-0.273	-0.164
spec-orig-main	0.200	-0.333	-0.467	-0.095	-0.252	-0.174
sum-of-fro	0.200	-0.333	-0.333	-0.381	-0.325	-0.212
prod-of-fro	0.200	-0.333	-0.333	-0.429	-0.372	-0.224
sum-of-fro/margin	0.333	-0.200	-0.467	-0.048	-0.217	-0.095
prod-of-fro/margin	0.333	-0.200	-0.467	-0.143	-0.247	-0.119
fro-distance	0.200	-0.200	-0.333	-0.286	-0.290	-0.155
spec-distance	0.200	-0.200	-0.333	-0.286	-0.290	-0.155
param-norm	0.200	-0.333	-0.333	-0.429	-0.363	-0.224
path-norm	0.333	-0.600	-0.467	-0.286	-0.191	-0.255
pac-bayes-init	0.200	0.200	-0.600	0.238	0.015	-0.009
pac-bayes-orig	-0.200	0.333	0.467	0.381	0.333	0.245
$1/\sigma$ pac-bayes-flatness	-0.267	0.333	0.333	0.455	0.333	0.213
SRR	-0.067	0.467	0.333	0.714	**0.445**	**0.362**

Sparse Rate Reduction as Regularization

Incorporate SRR measure into training via explicit regularization.

$$\min_{\boldsymbol{w}} \mathcal{L}_{\mathrm{ce}}(\boldsymbol{w}) + \lambda\cdot\frac{1}{L}\sum_{\ell=1}^L \mu_{\mathrm{SRR}}^\ell(\boldsymbol{w}^\ell;\boldsymbol{Z}_{\mathrm{StopGrad}}^\ell)$$

Efficient implementation as per layer regularization already suffices to consistently boost performance.

Table 2: Top-1 accuracy for CRATE and its variants trained with or without SRR regularization on CIFAR-10/100 from scratch (width $d = 384$).

Models	CIFAR-10		CIFAR-100	
	cross-entropy	+ SRR regularization (L=12)	cross-entropy	+ SRR regularization (L=12)
CRATE-C	76.87	**77.61**	43.40	**44.53**
CRATE-N	81.52	**81.91**	55.11	**55.62**
CRATE-T	85.49	**85.52**	60.59	**60.69**
CRATE	86.67	**86.79**	62.40	**62.52**

References

- Yu et al. "White-box transformers via sparse rate reduction." NeurIPS 2023.
- Jiang et al. "Fantastic generalization measures and where to find them." ICLR 2020.
- Foret et al. "Sharpness-Aware Minimization for Efficiently Improving Generalization." ICLR 2021.