SVD-Attention: Efficiently Analyzing Attention Blocks in Transformers

Debopam Sanyal[1], Hongjie Chen[2], Riya Elizabeth Kuriyan[1], Arth Tak[1], Divya Kothandaraman[2], Akshay Mehra[2], Joshua Kimball[2], Alexey Tumanov[1]

[1] Georgia Institute of Technology [2] Dolby Laboratories

EMNLP2026 BUDAPEST

tl;dr: **Which attention modes does real data use?**

The score combines singular values with token variation to locate unused rank and predict which layers tolerate truncation. No finetuning.

Attention has a rank budget

Each head compares queries and keys through their joint interaction matrix:

$$M_h \;=\; W_{q,h} W_{k,h}^{\top} \;\in\; \mathbb{R}^{d\times d},$$

Its rank is at most the head dimension.
A large singular value alone does not tell us whether tokens vary along that mode.

Redundancy depends on the task

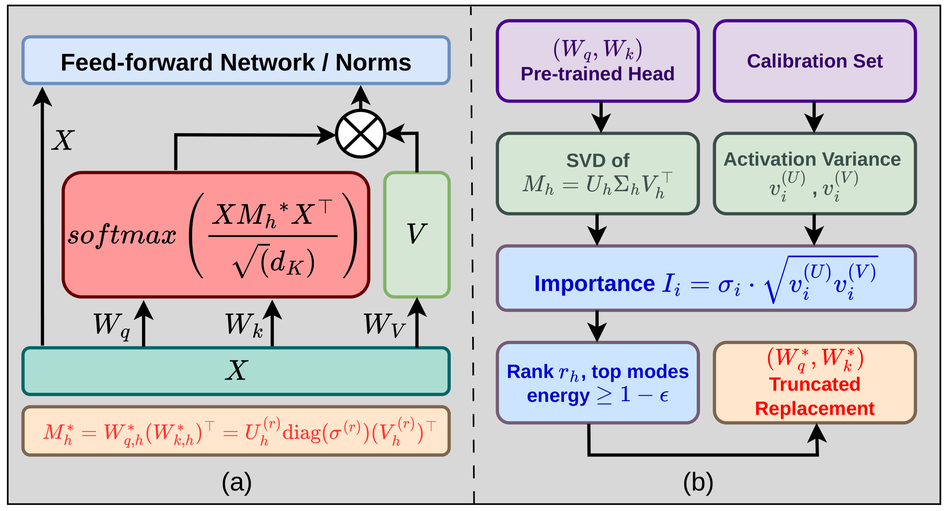
Task	Model	Flagged layers
Sentiment	DistilBERT	**0–2**
Image classification	ViT-B	**6–8**
Image classification	Swin-B	**4–7**
Generation	Llama-2-7B	**8–11**
Dense prediction	DPT-Large	**16–19**
Extractive QA	RoBERTa-base	**8–11**
Extractive QA	MiniLM	**9–11**

Classification and generation: earlier or middle layers.
Dense prediction and extractive QA: later layers.
Layer indices are zero-based.

Inference cost

At ε = 0.1: 0.8–1.5% fewer FLOPs alone, 8–11% with WANDA. Most combined savings come from WANDA. These are analytical FLOPs, not measured speedups.

The SVD-Attention diagnostic



The per-head QK interaction and the calibration-based diagnostic (Figure 1).

1. Decompose the joint QK matrix

Analyze the product of the pretrained query and key projections, rather than each projection separately.

$$M_h \;=\; U_h\, \Sigma_h\, V_h^{\top},$$

2. Score modes using calibration data

Measure token variance in the query and key singular bases. Combine it with each singular value:

$$\mathcal{I}_i \;=\; \sigma_i \cdot \sqrt{v_i^{(U)}\, v_i^{(V)}}.$$

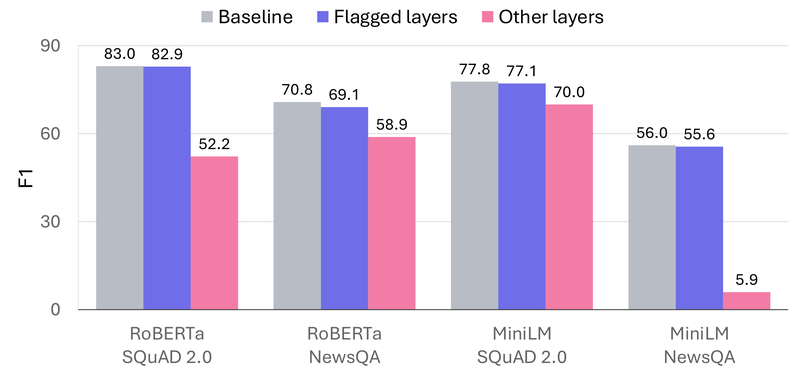
3. Select a rank for each head

Keep the fewest modes covering at least 1 − ε of total importance, then replace the Q/K projections.

$$W^*_{q,h} = U_h^{(r)}\, \mathrm{diag}(\sqrt{\sigma^{(r)}}), \quad W^*_{k,h} = V_h^{(r)}\, \mathrm{diag}(\sqrt{\sigma^{(r)}}).$$

Experimental results

(a) Does layer selection predict quality?



Same number of truncated layers at ε = 0.3.
Flagged layers lose at most 1.75 F1. Other layer groups lose up to 50 F1 (Table 3).

(b) Does data improve mode ranking?

RoBERTa-base, layers 8–11, ε = 0.3 (Table 1)

Dataset	σ only rank / F1	Data-aware rank / F1
SQuAD 2.0	38.6 / 63.08	**29.0 / 82.85**
NewsQA	38.6 / 59.33	**27.0 / 69.06**

The data-aware score retains fewer modes and preserves substantially more task quality.

(c) How much rank can we discard?

RoBERTa-base / SQuAD 2.0, layers 8–11 (Table 11)

	ε = 0.1	ε = 0.2	ε = 0.3
Mean rank / 64	49	38	**29**
F1	83.24	83.09	**82.85**

More than half the average head rank removed for a 0.16 F1 loss. Baseline: 83.01 F1.

Acknowledgements

This work was conducted in part during an internship at Dolby Laboratories.
Supported in part by CoCoSys, a JUMP 2.0 center in an SRC program sponsored by DARPA.