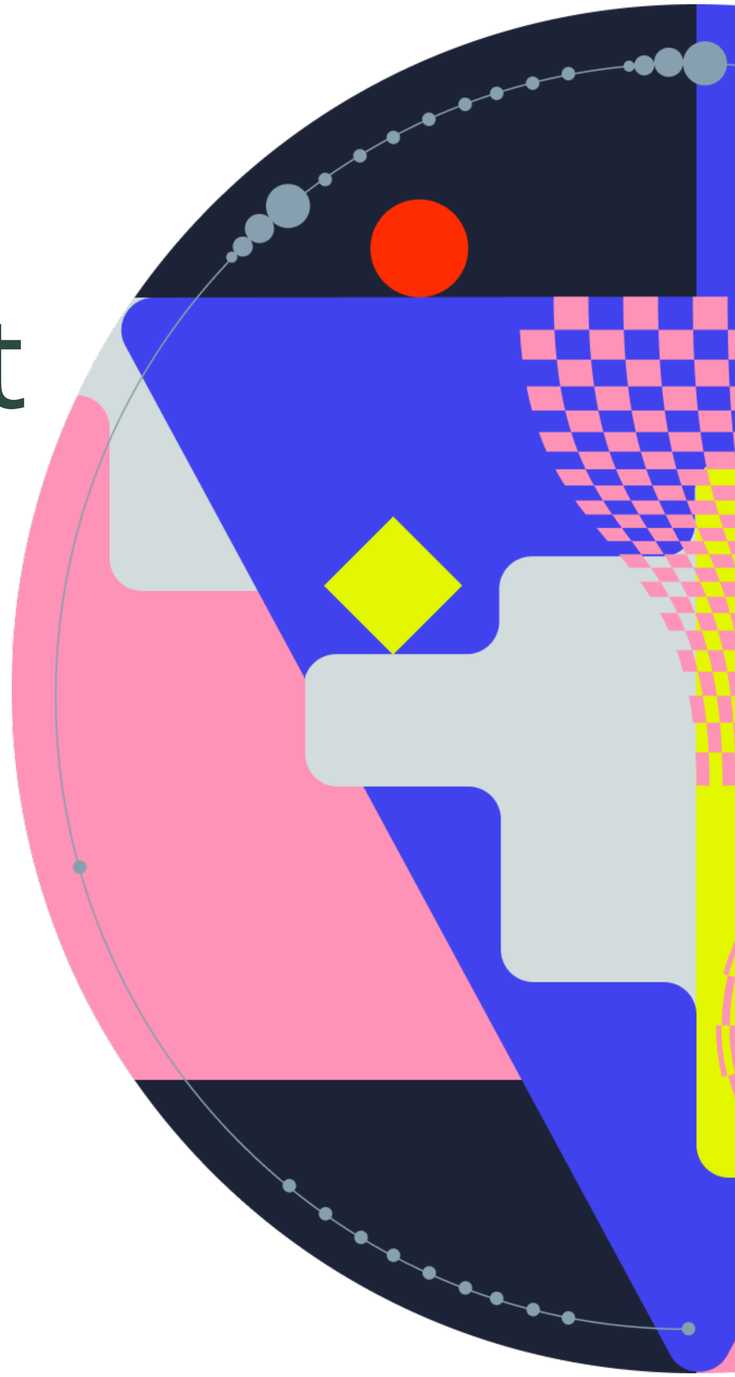
School of Data Analysis

SVD & ASVD for Transformer Models Compression

Arsenii Rybakov
Daniil Maslov
Ruben Safaryan

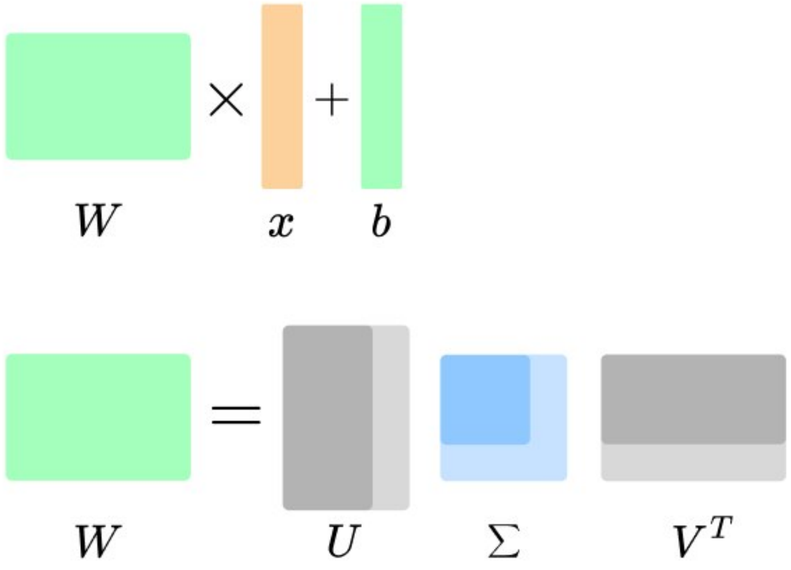
Motivation

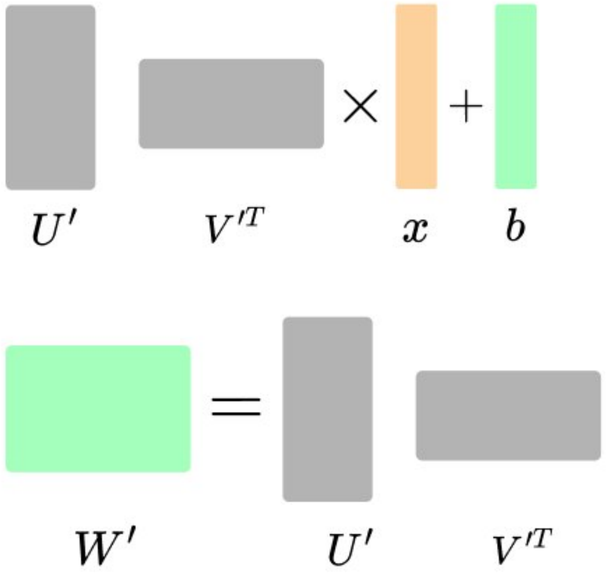
- Lightning LLM development
- Huge size of Transformers
- Sustainability/Economic



Brief Introduction into the *SVD*

SVD





ASVD

$$\mathbf{W}_k^\star = \arg\min_{\mathbf{W}_k} \|\mathbf{W}_k\mathbf{X} - \mathbf{W}\mathbf{X}\|_F^2$$

Explored setups

1 Google/ViT-Base-16-224 + Attention Layers SVD/ASVD compression

2 Google/ViT-Base-16-224 + All Linear Layers SVD compression + FT

3 Google/bert-large-uncased + Attention Layers SVD/ASVD compression

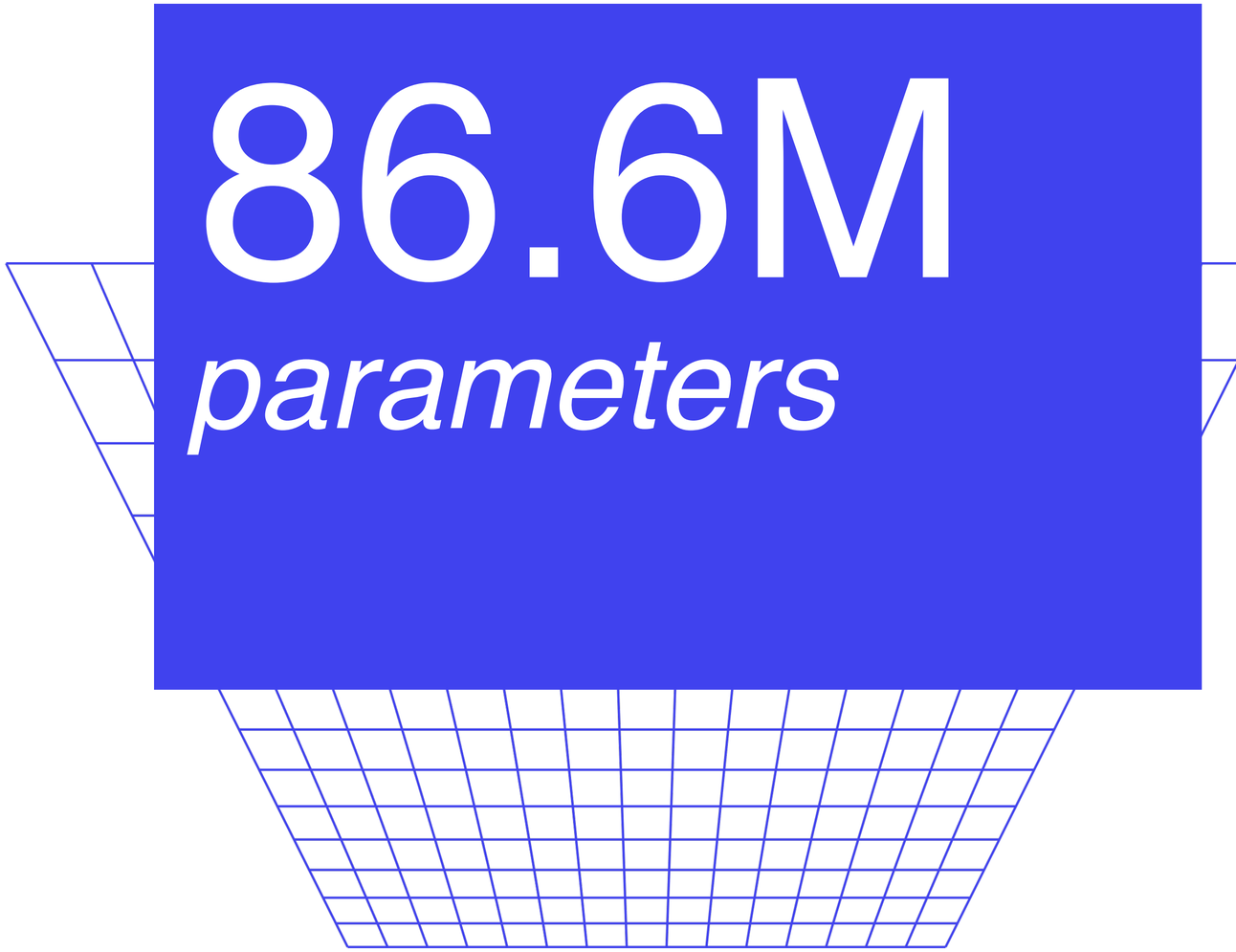
4 Google/bert-large-uncased + All Linear Layers SVD compression + FT

* 1, 3 — zero-learning setups

Google/ViT-B-16-224

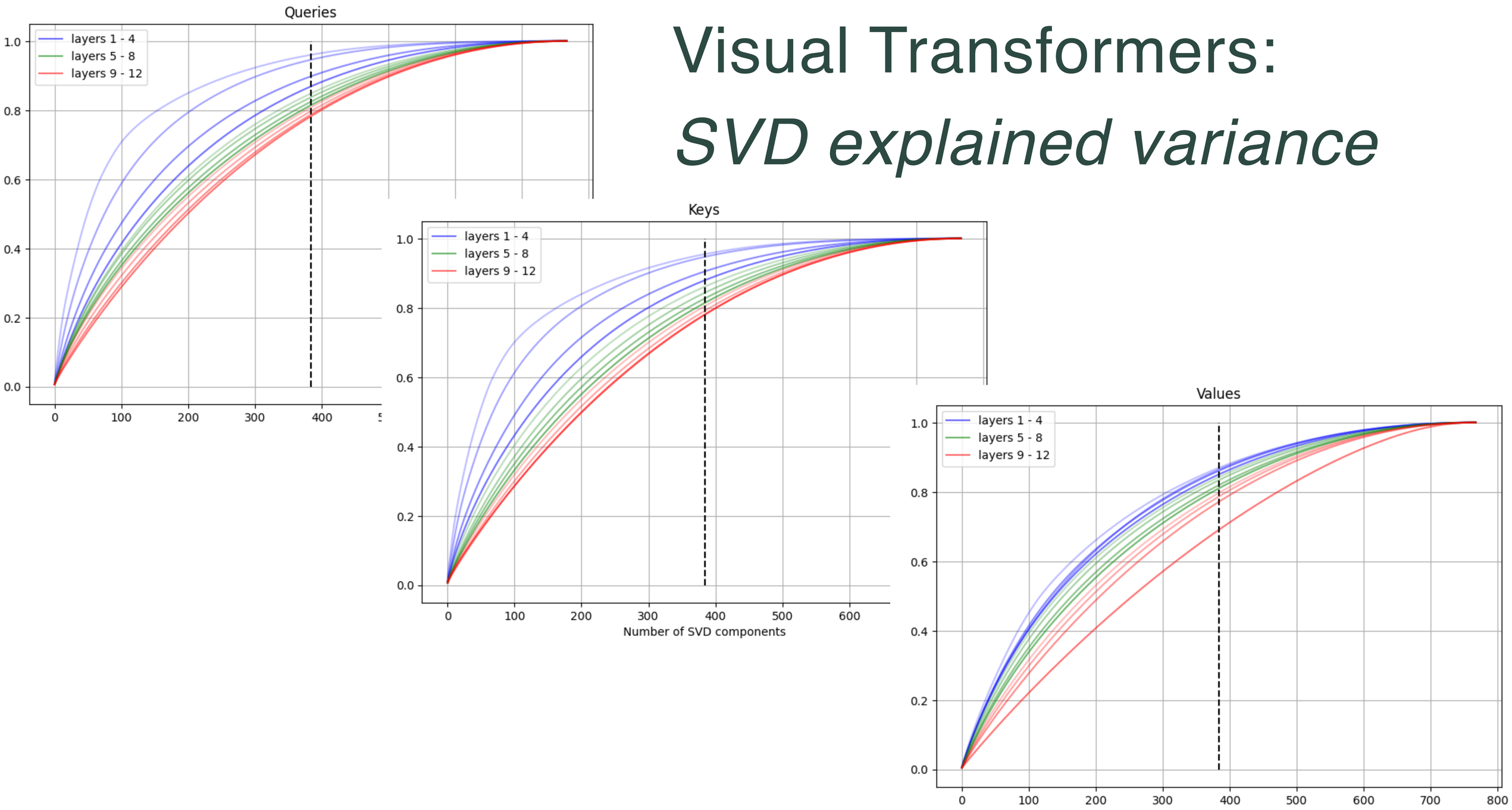
1 Family of Visual Transformers Introduced in 2021 by Dosovitskiy et al.

2 SOTA—level result for the ImageNet-1K image classification



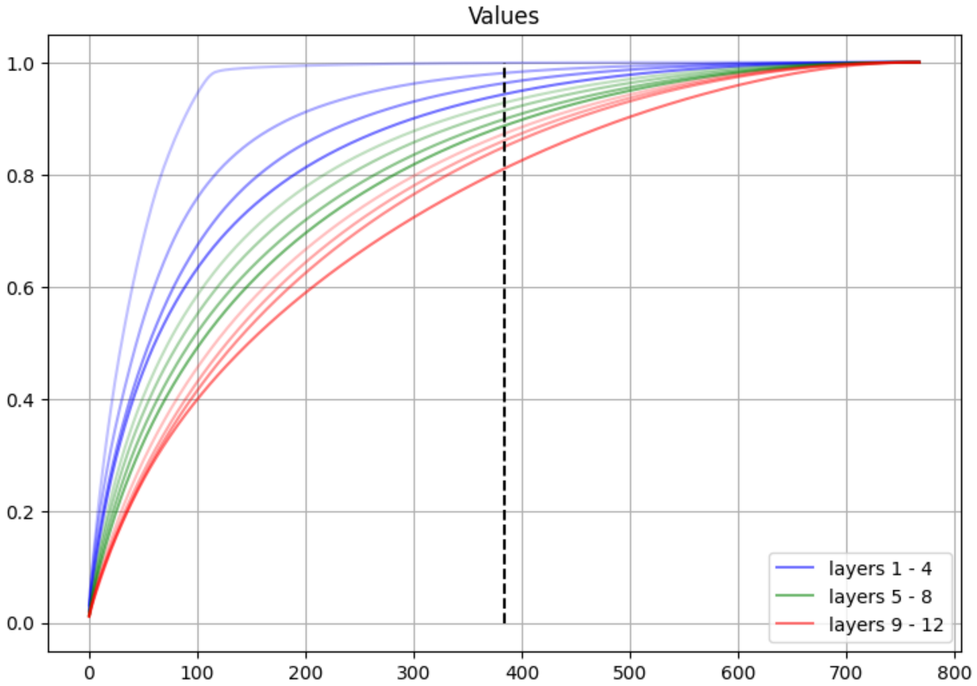
Visual Transformers:

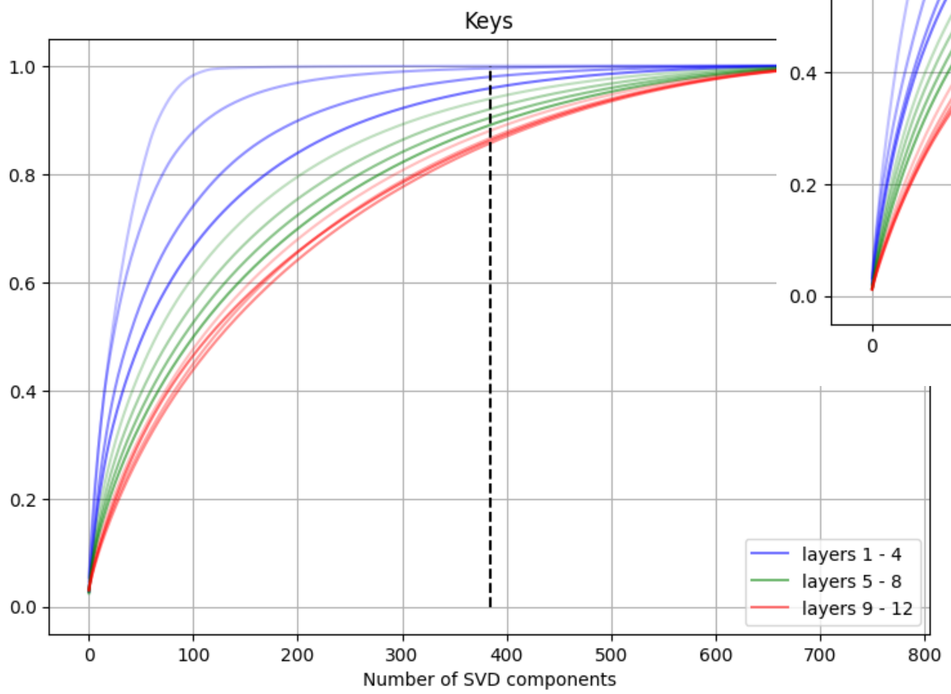
SVD explained variance

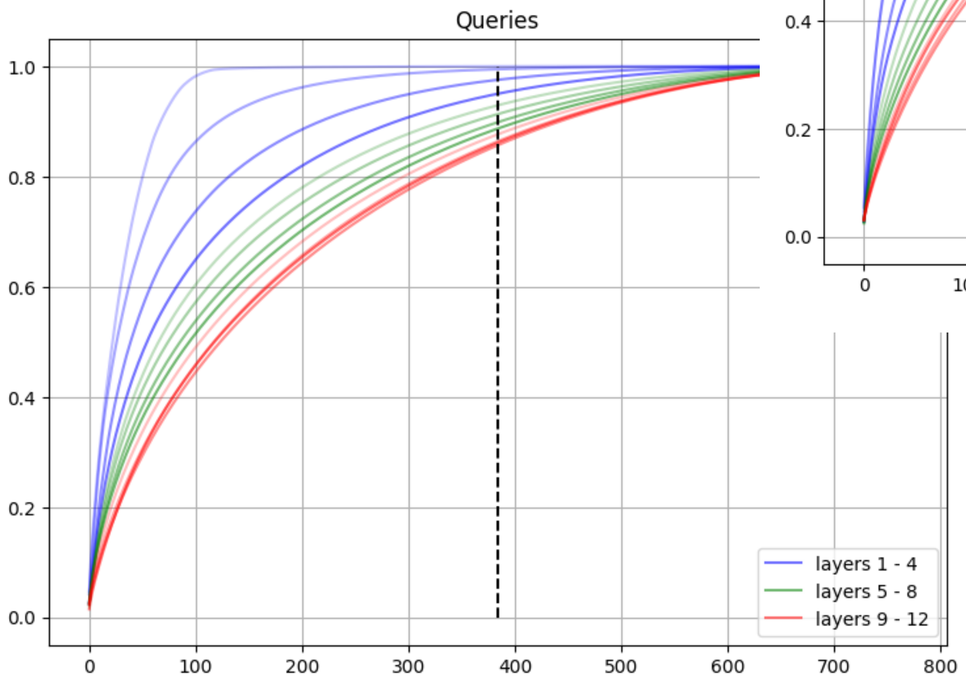


Visual Transformers:

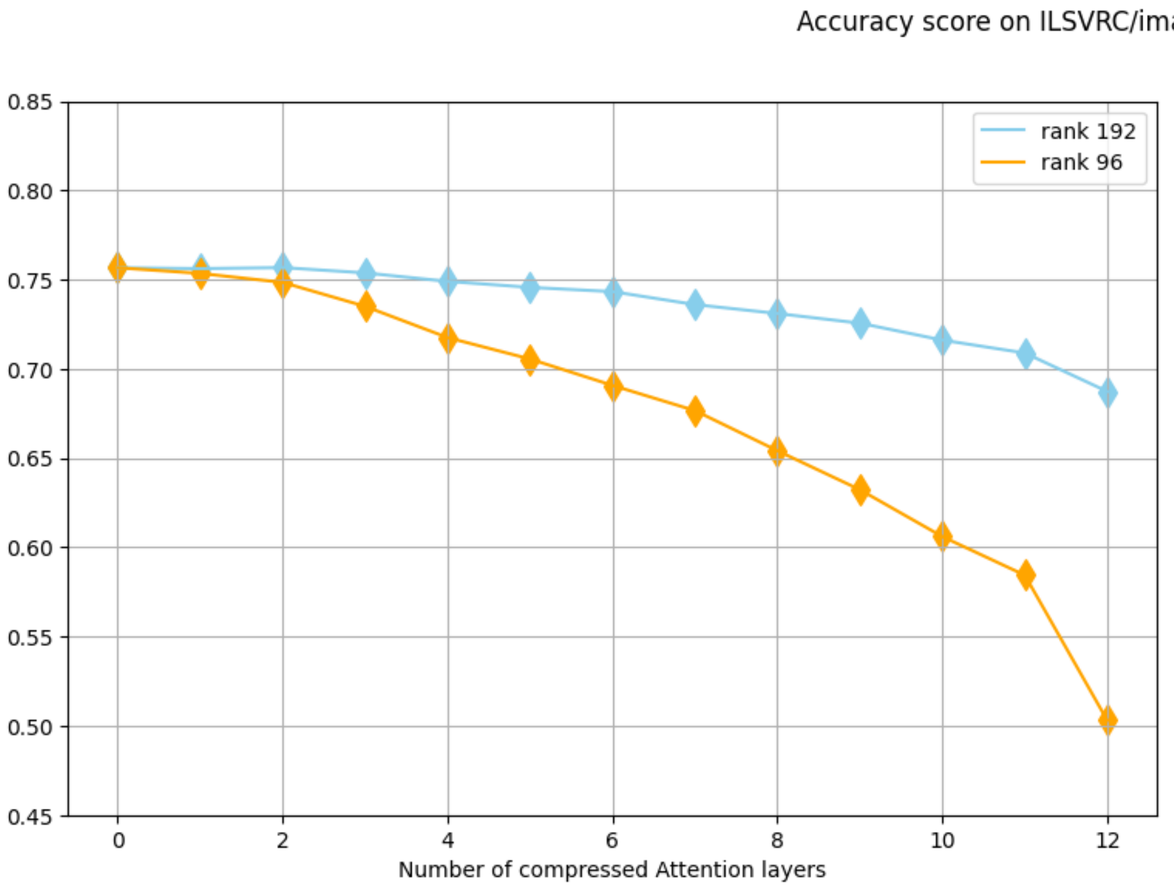
ASVD explained variance

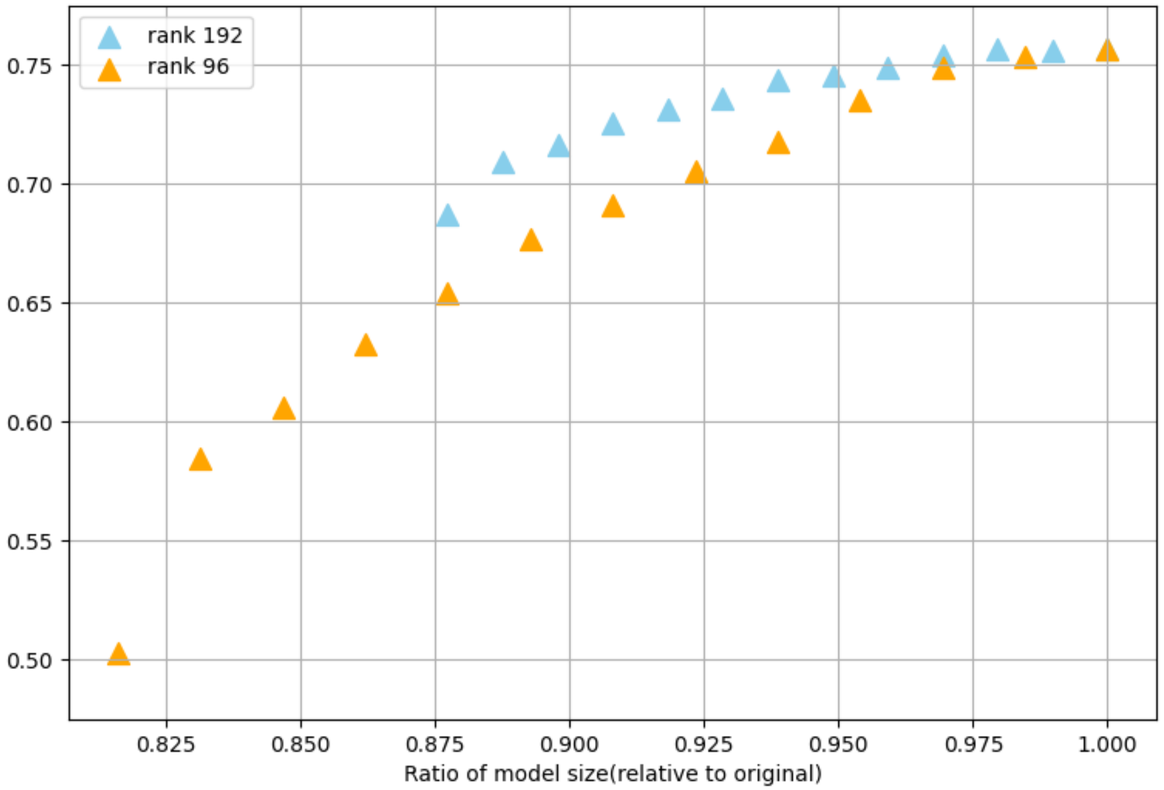




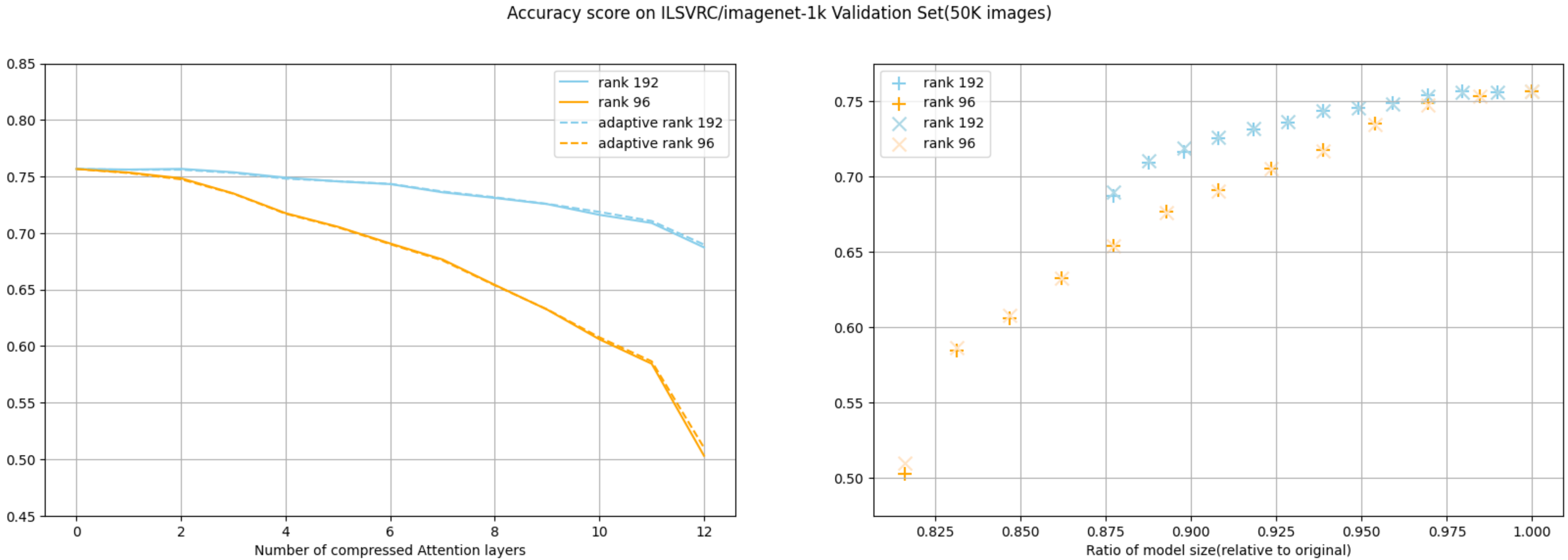


Basic SVD Compression for ViT-B-16-224 perf.





SVD/AVSD Compression for ViT-B-16-224 perf.



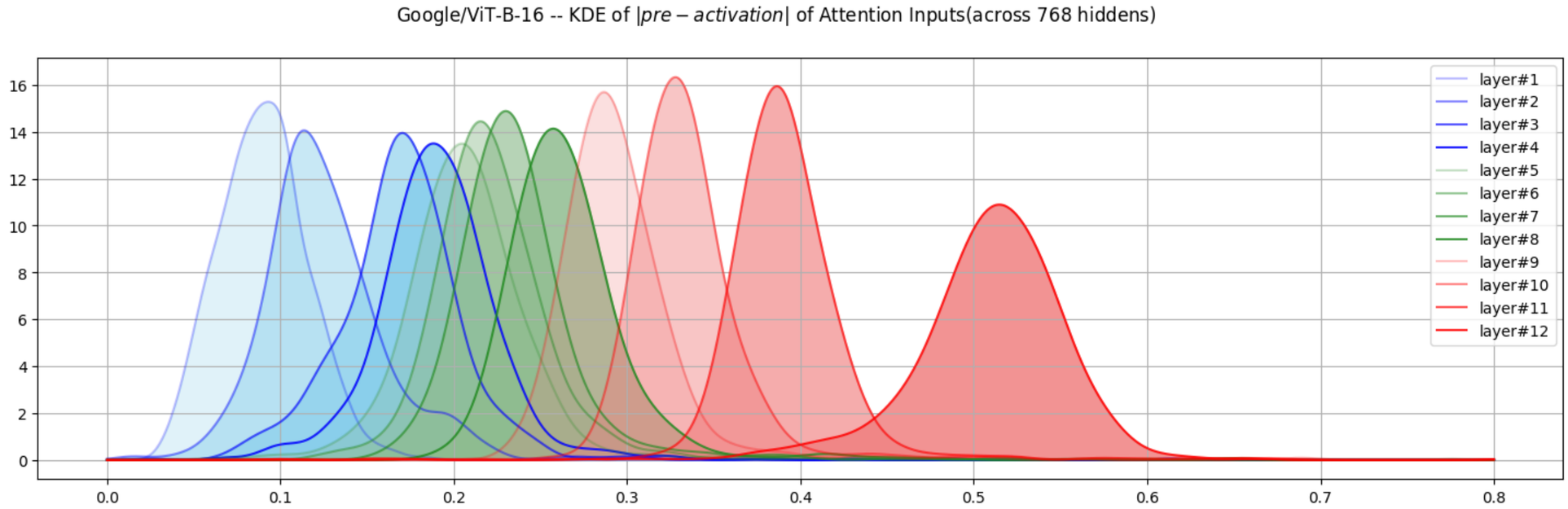
Unfortunately, ASVD doesn't work well

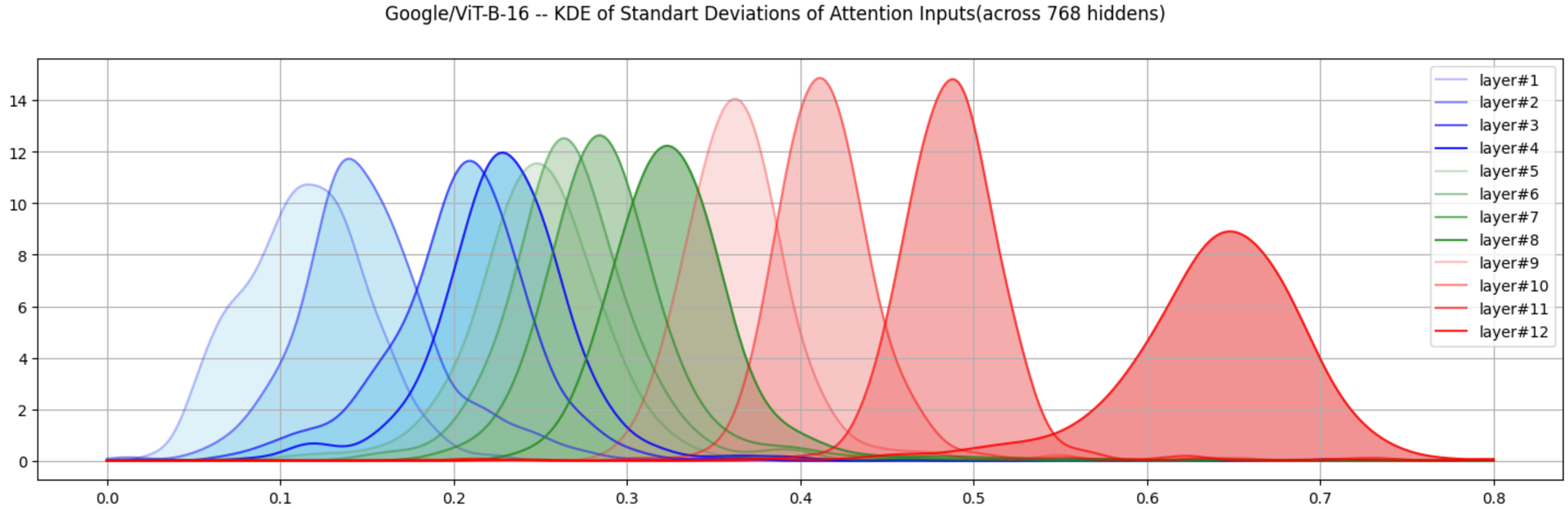
Possible reasons for inferior ASVD performance

- Small model that fully utilises its parameters potential

- LayerNorm included in Transformers in architecture

Proofs for the LayerNorm hypothesis





Google/ViT-B-16-224 — All Linear Layers SVD + FT

Model Type	Max rank	Share of params.	Acc. (no FT)	Acc(FT)
Original	768	100 %	0.86	—
Compressed—1	244	48 %	0.376	0.79
Compressed—2	192	33 %	0.14	0.79
Compressed—3	96	19 %	0.01	0.65

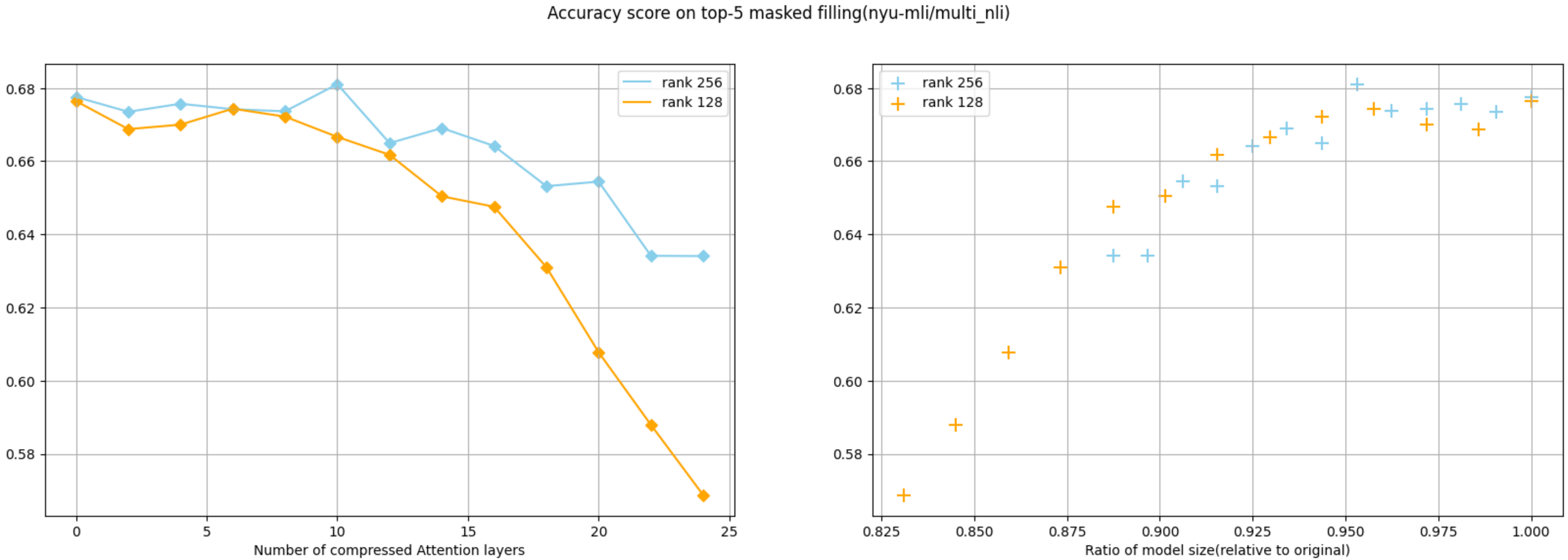
Google/Bert-large-uncased-whole-word-masking

1. Family of NLP models introduced by Google in 2018
2. SOTA — level on variety of NLP tasks
3. Evaluated on mask-filling task on the «*nyu-mll/multi_nli*» dataset

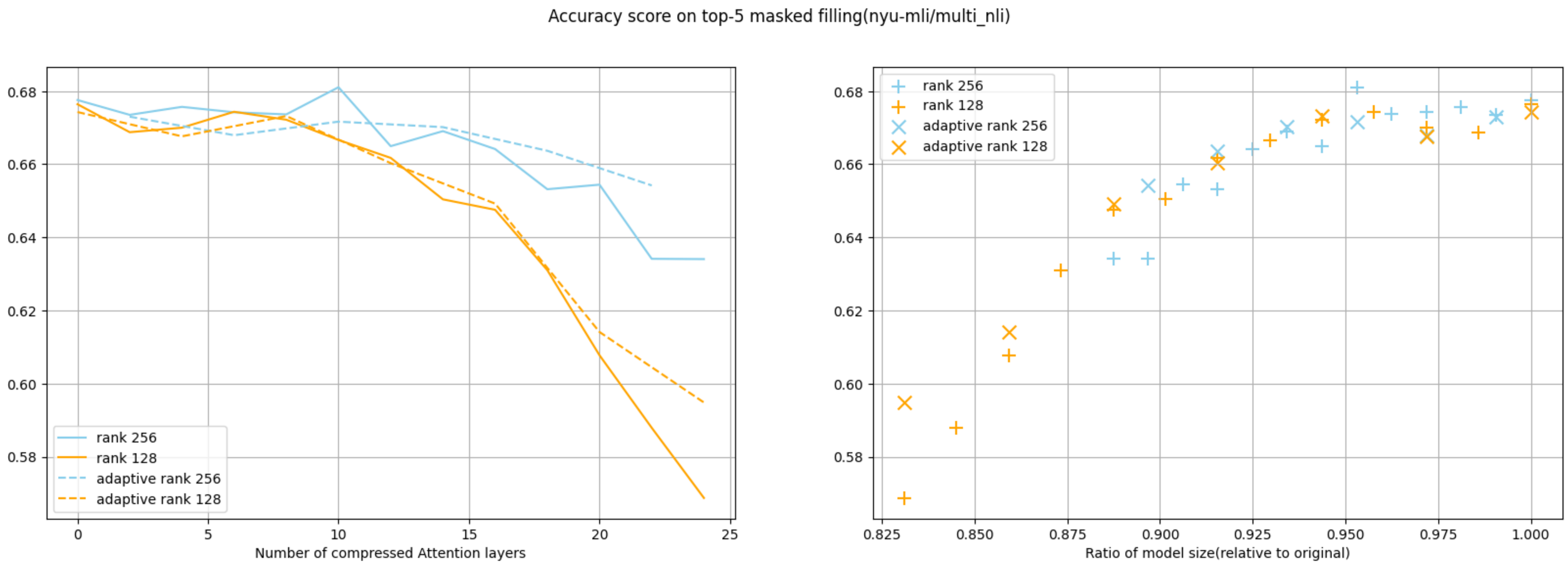
336M

parameters

bert-large compression perf.



bert-large compression perf.



School of Data Analysis

Thank you for your Attention!

It’s all we need!