“Request for Plot”: is **softmax cross-normalization** fruitful in transformers?

Mishka (Michael Bukatin)

Dataflow Matrix Machines project

`https://github.com/anhinga`

I am looking for collaborators

"Request for Plot"

Replace

$$\text{Attention}(Q, K, V) = \text{softmax}(cKQ^{\mathsf{T}})V$$

with

$$\text{Attention}(Q, K, V) = \text{softmax}(cKQ^{\mathsf{T}})\text{softmax}^{\mathsf{T}}(V)$$

Does your favorite learning curve improve?

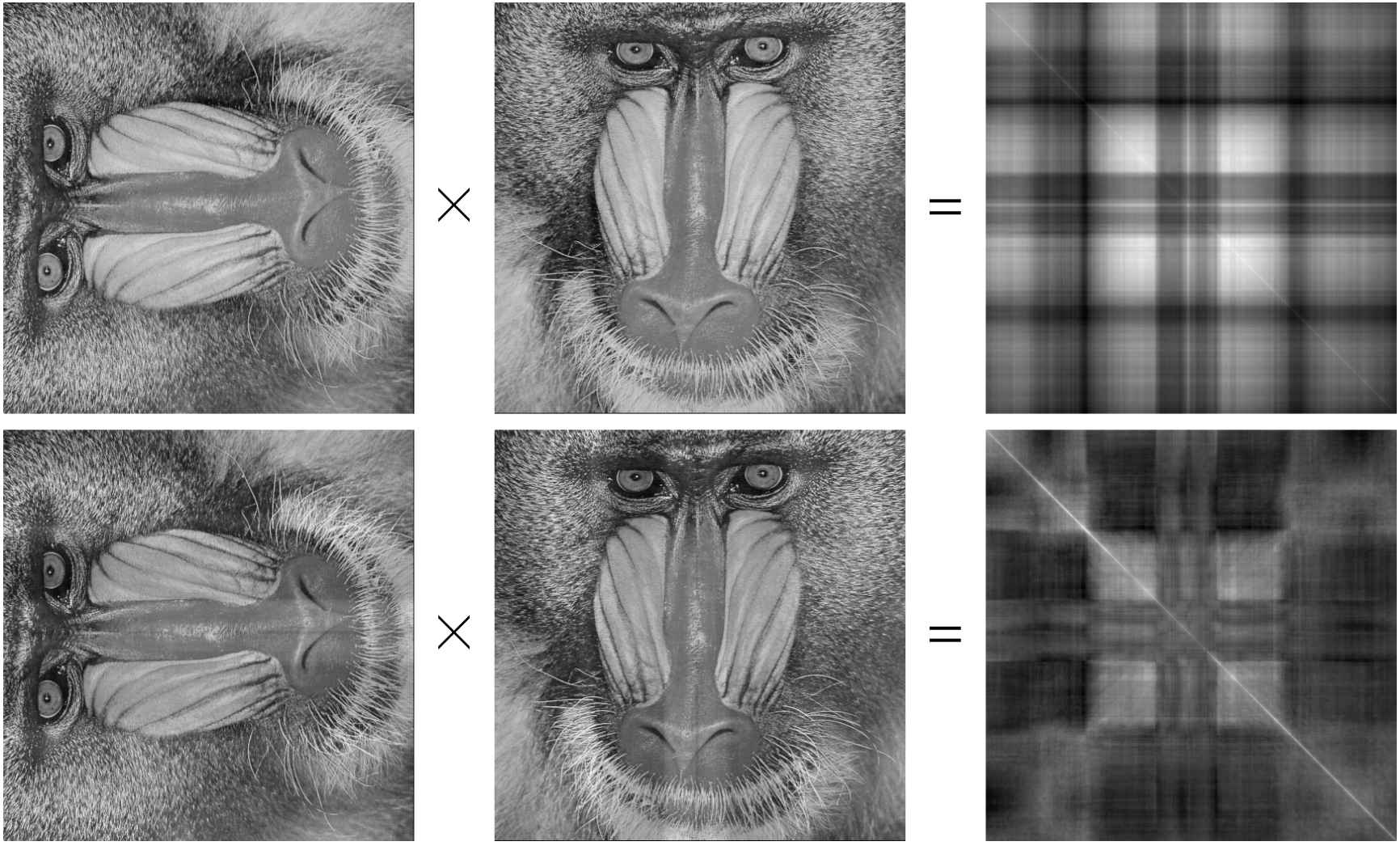
Why do I think this is a good shot?

Why this might be a good idea?

I experimented with interpreting monochrome images as matrices and multiplying those matrices via matrix product and checking whether results are visually interesting.

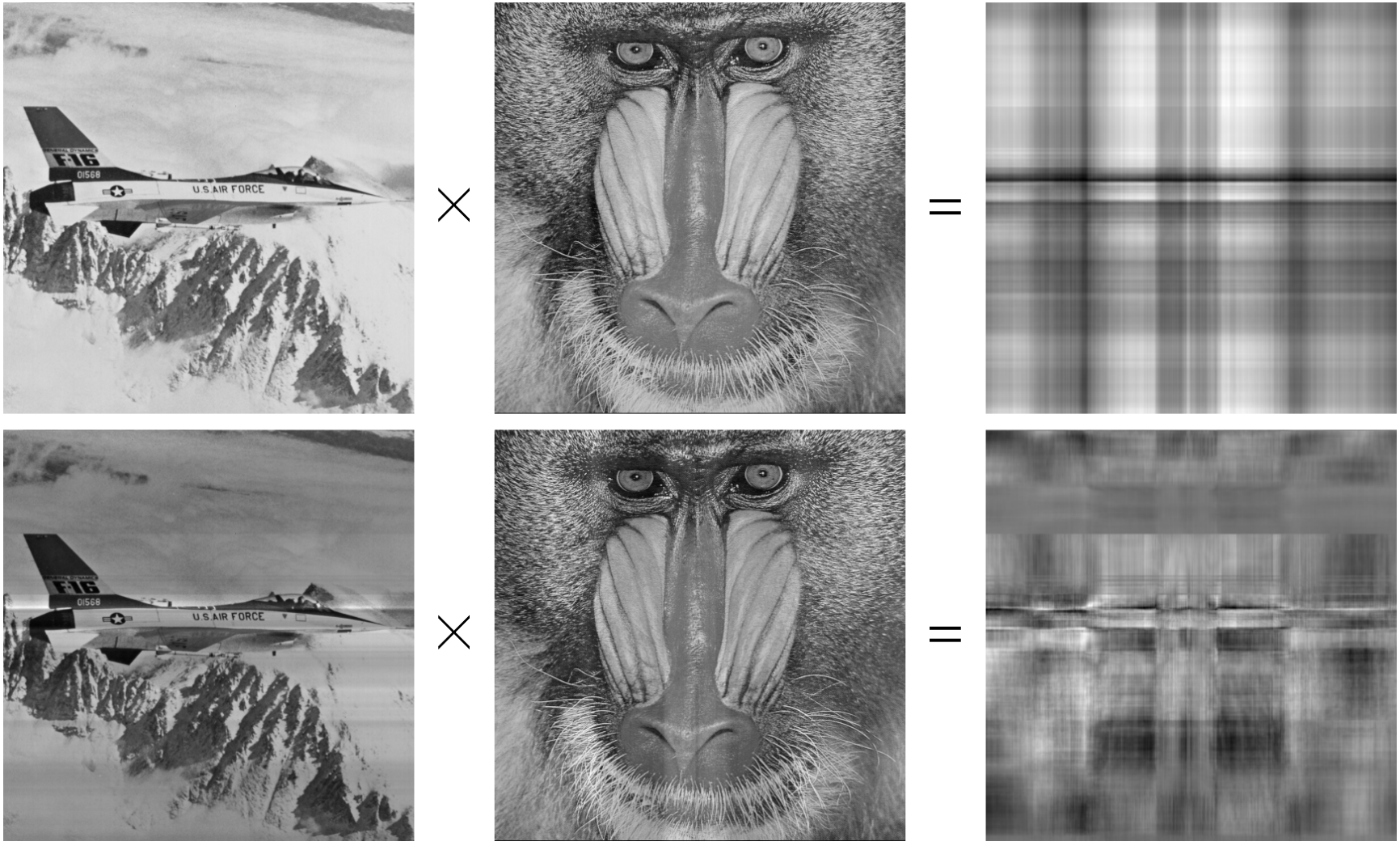
Multiplying monochrome images as matrices:
A*B and softmax

`https://github.com/anhinga/JuliaCon2021-poster`



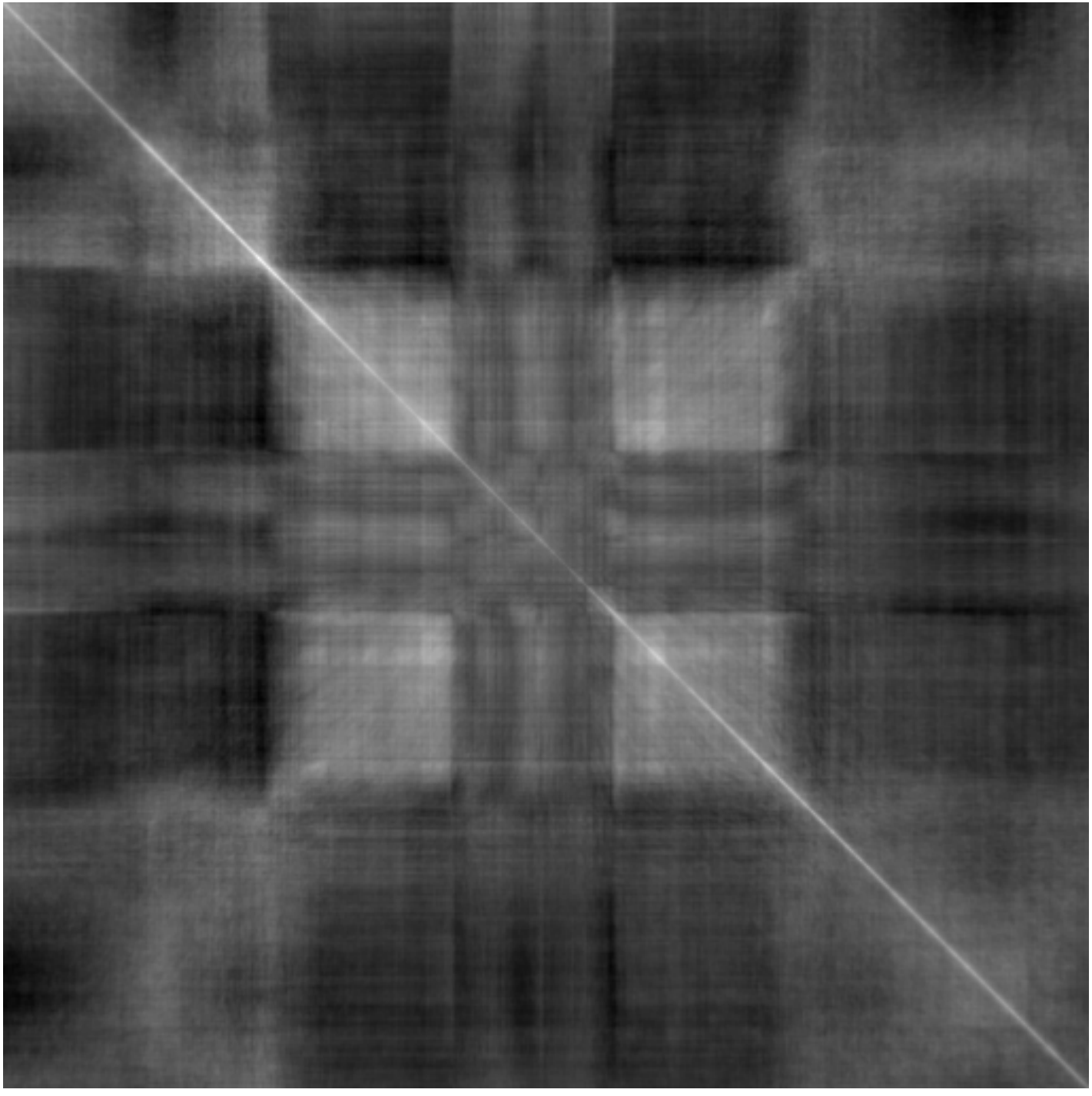
In Transformers people sometimes **softmax** rows of the left matrix: $\text{Attention}(Q, K, V) = \text{softmax}(cKQ^T)V$ from "Attention Is All You Need" (2017).

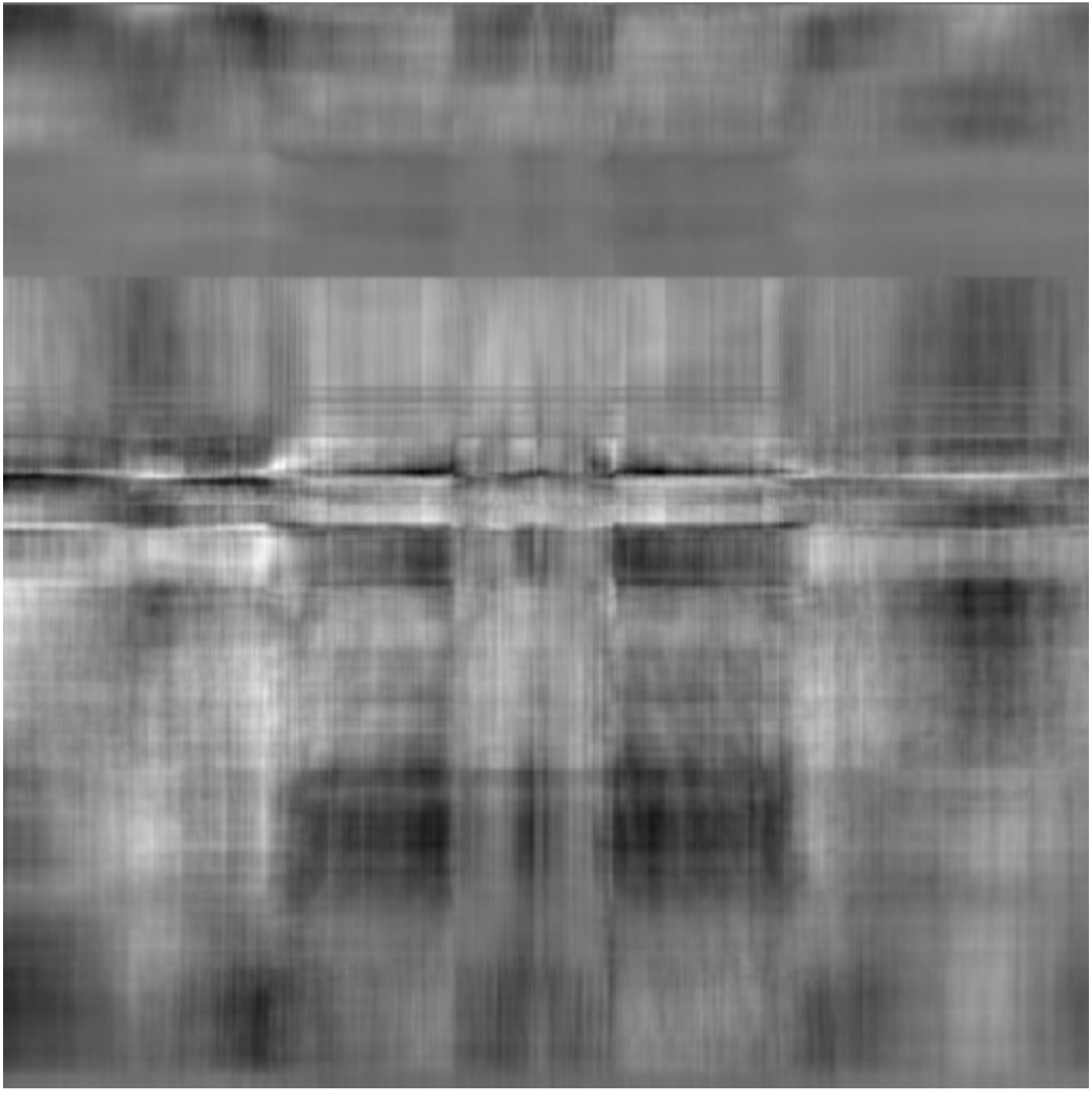
In the second example we **softmax** rows of the left matrix **and** columns of the right matrix resulting in products with richer, more fine-grained structure.



In Transformers people sometimes **softmax** rows of the left matrix: $\text{Attention}(Q, K, V) = \text{softmax}(cKQ^T)V$ from "Attention Is All You Need" (2017).

In the second example we **softmax** rows of the left matrix **and** columns of the right matrix resulting in products with richer, more fine-grained structure.





Cross-normalization

$$\text{softmax}^{\mathsf{T}}(V)$$

values are horizontal stripes

cross-norm: softmax vertical vectors

a vertical vector consists of i's coords of all values for a given i

Remix with classical Transformer attention

Various ways to remix, e.g.

$$\alpha * \text{softmax}^{\mathsf{T}}(V) + (1 - \alpha) * (V)$$

I originally thought about training Transformers from scratch when creating this RFP.

But if one starts with $\alpha = 0$ and increases α gradually, then one should be able to try fine-tuning a classical pretrained Transformer while deforming it from V towards $\text{softmax}^{\mathsf{T}}(V)$.

Contact

Open an issue in

`https://github.com/anhinga/JuliaCon2021-poster`

or send an e-mail to bukatin @ cs dot brandeis dot edu