Appendix

MIMONets: Multiple-Input-Multiple-Output Neural Networks Exploiting Computation in Superposition

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A MIMONets Details

A.1 VSA representations and operations in MIMONets

There are numerous available options for binding and unbinding depending on the VSA models being used [1]. Table A1 summarizes the VSA representations and operations used in MIMOConv and MIMOFormer. MIMOConv relies on holographic reduced representation (HRR) [2] for binding, and matrix binding of additive terms (MBAT) [3] for unbinding. The binding and unbinding keys are real-valued and trainable. At initialization, each element in the D-dimensional key vector is drawn from an i.i.d. Gaussian distribution with zero mean and $1/D$ variance. Optionally, the binding keys can be frozen during training while maintaining a high accuracy (see Appendix E.5). MIMOConv's binding relies on our proposed position-wise HRR (PWHRR) binding, which maintains the image's local structure. The unbinding is implemented with MBAT, which computes the vector-matrix multiplication between the CNN's D_o-dimensional output feature vector and an unbinding matrix $\tilde{A}^{(i)} \in \mathbb{R}^{D_o \times D_o}$. The output dimension is D_o=640 in WideResNet-28-10. The MBAT unbinding provides a higher degree of freedom by having D_o^2 trainable parameters, whereas HRR's unbinding would generate a circulant matrix with D_o trainable parameters. However, it requires only 409,600 parameters per superposition channel, which is negligible compared to the remaining layers in MIMOConv, which have 36.54 M trainable parameters. Due to deep neural networks being highly nonlinear, we thus opt for this more flexible variant of unbinding by arbitrary linear transformations.

MIMOFormer uses the multiply-add-permute (MAP) [4] model, which uses bipolar keys and the element wise multiplication (Hadamard product) for binding and unbinding. The bipolar keys are drawn from a Rademacher distribution and are frozen during training and inference.

Table A1: Summary of VSA representations and operations used in MIMOConv and MIMOFormer.

			Binding		Unbinding	
	VSA framework	Key representation	Operation	Keys	Operation	Keys
MIMOConv	HRR/MBAT	real-valued	PWHRR	trainable/frozen	MBAT	trainable
MIMOFormer	MAP	bipolar	Hadamard product	frozen	Hadamard product	frozen

A.2 Illustration of the Blessing of Dimensionality

VSA builds upon the mathematical concept of the Blessing of Dimensionality. According to it, random vectors are (quasi-)orthogonal with high probability. Let us illustrate one version of it. Suppose independent random bipolar vectors $x, y \in \{-1, +1\}^D$ with i.i.d. Rademacher distributed components as used in MIMOFormer. It holds by Hoeffding's inequality (Appendix G)

$$\mathbb{P}\left(|\cos \measuredangle(x, y)| \geq \alpha\right) = \mathbb{P}\left(|\langle x, y\rangle| \geq \alpha D\right) \leq 2e^{-D\alpha^2/2} \qquad \forall \alpha \geq 0 \tag{1}$$

Similar bounds exist for Gaussian random vectors, which are used to generate keys for MIMOConv. Let us set some cutoff for interference events (IE), namely, we consider two vectors to interfere with each other if the angle between them is less than $70°$ (corresponding to $\alpha = \cos(70°)$), already $20°$ off from exact orthogonality. The bound demonstrates that the probability for two vectors (with i.i.d. Rademacher distributed components) to interfere (IE) is less than 0.785 for vectors in 16 dimensions. As such, high levels of interference could still occur frequently. In contrast, for 64 dimensions, the probability that two vectors interfere (IE) is already known to be less than 0.0474. Consequently, we are much more certain that interference occurs with low probability.

A.3 Noisy retrieval of values from a key-value superposition

Consider a superposition of N bound values $x^{(i)}$

$$s = a^{(1)} \odot x^{(1)} + a^{(2)} \odot x^{(2)} + \ldots + a^{(N)} \odot x^{(N)} \tag{2}$$

where the binding keys $a^{(i)}$ are (for instance) independent bipolar vectors of i.i.d. Rademacher entries. Unbinding with $a^{(k)}$ produces the signal of interest $x^{(k)}$ together with a noise vector orthogonal to it:

$$a^{(k)} \circledast s = a^{(k)} \circledast a^{(1)} \odot x^{(1)} + a^{(k)} \circledast a^{(2)} \odot x^{(2)} + \ldots + a^{(k)} \circledast a^{(N)} \odot x^{(N)} \tag{3}$$

$$= x^{(k)} + noise. \tag{4}$$

The noise vector stems from the approximate unbinding ($a^{(k)} \circledast a^{(k)} \odot x^{(k)} \approx x^{(k)}$) as well as randomized value vectors ($a^{(k)} \circledast a^{(j)} \odot x^{(j)}$). Importantly, $noise$ becomes orthogonal to $x^{(1)}$, hence distinguishable, with a growing embedding dimension according to the Blessing of Dimensionality. The effect of noise is mitigated after comparing against a dictionary of known outputs, based on a notion of inner product. Such a comparison naturally, but not exclusively, arises in classification tasks. Concretely, comparing against $a^{(k)} \odot \Omega$ for an Ω aligned with $x^{(k)}$ returns

$$\langle a^{(k)} \circledast s, \Omega \rangle = \langle s, a^{(k)} \odot \Omega \rangle \tag{5}$$

$$= \sum_{i=1}^{N} \langle a^{(i)} \odot x^{(i)}, a^{(k)} \odot \Omega \rangle \tag{6}$$

$$\approx \langle a^{(k)} \odot x^{(k)}, a^{(k)} \odot \Omega \rangle \tag{7}$$

$$= \langle x^{(k)}, \Omega \rangle \tag{8}$$

where we still assume bipolar binding and where the approximation relies on the Blessing of Dimensionality producing orthogonal vectors. For the VSA framework MAP, which uses bipolar keys of i.i.d. Rademacher entries, we provide a more precise formulation:

Theorem 1 (Dictionary Cleanup Noise)**.** *Let $\Omega \in \mathbb{R}^D$ be an element of a dictionary and consider the superposition of bound values $x^{(i)} \in \mathbb{R}^D$*

$$s = a^{(1)} \odot x^{(1)} + a^{(2)} \odot x^{(2)} + \ldots + a^{(N)} \odot x^{(N)} \tag{9}$$

where the binding keys $a^{(i)} \in \{-1, 1\}^D$ are independent bipolar vectors of i.i.d. Rademacher entries. It then holds

$$\mathbb{P}\left\{ \left\langle s, a^{(k)} \odot \Omega \right\rangle \notin [1-\alpha, 1+\alpha] \cdot \left\langle x^{(k)}, \Omega \right\rangle \right\} \leq 2 \exp\left(- \frac{\alpha^2 \left| \langle x^{(k)}, \Omega \rangle \right|^2}{2 \sum_{i \neq k} \left\| x^{(i)} \odot \Omega \right\|_2^2} \right) \tag{10}$$

with the exponent, given $x^{(i)}$ are all of roughly equal norm, according to Theorem 6 in Appendix G typically scaling as

$$-D\alpha^2 \cos^2(\measuredangle(x^{(k)}, \Omega)) \Big/ 2(N-1) \tag{11}$$

Thus, for $D \gg N$ comparison against elements from a dictionary allows faithful retrieval.

Proof. By the following equivalence relation

$$\left\langle s, a^{(k)} \odot \Omega \right\rangle \notin [1-\alpha, 1+\alpha] \cdot \left\langle x^{(k)}, \Omega \right\rangle \iff \left| \sum_{i \neq k} \left\langle a^{(i)} \odot x^{(i)}, a^{(k)} \odot \Omega \right\rangle \right| > \alpha \left| \left\langle x^{(k)}, \Omega \right\rangle \right| \tag{12}$$

it suffices to derive tail bounds on

$$\mathbb{P}\left\{ \left| \sum_{i \neq k} \left\langle a^{(i)} \odot x^{(i)}, a^{(k)} \odot \Omega \right\rangle \right| > \alpha \left| \left\langle x^{(k)}, \Omega \right\rangle \right| \right\}$$

$$= \mathbb{P}\left\{ \left| \sum_{i \neq k} \sum_{d} a_d^{(i)} a_d^{(k)} x_d^{(i)} \Omega_d \right| > \alpha \left| \left\langle x^{(k)}, \Omega \right\rangle \right| \right\} \tag{13}$$

Since $\{a_d^{(i)} \cdot a_d^{(k)}\}_{d, i \neq k}$ for k fixed is a set of i.i.d. Rademacher random variables, we are in a position to apply Hoeffding's inequality (see Appendix, Theorem 5) which gives

$$\mathbb{P}\left\{ \left| \sum_{i \neq k} \sum_{d} a_d^{(i)} a_d^{(k)} x_d^{(i)} \Omega_d \right| > \alpha \left| \left\langle x^{(k)}, \Omega \right\rangle \right| \right\} \leq 2 \exp\left(- \frac{\alpha^2 \left| \langle x^{(k)}, \Omega \rangle \right|^2}{2 \sum_{i \neq k} \sum_{d} \left| x_d^{(i)} \Omega_d \right|^2} \right) \tag{14}$$

□

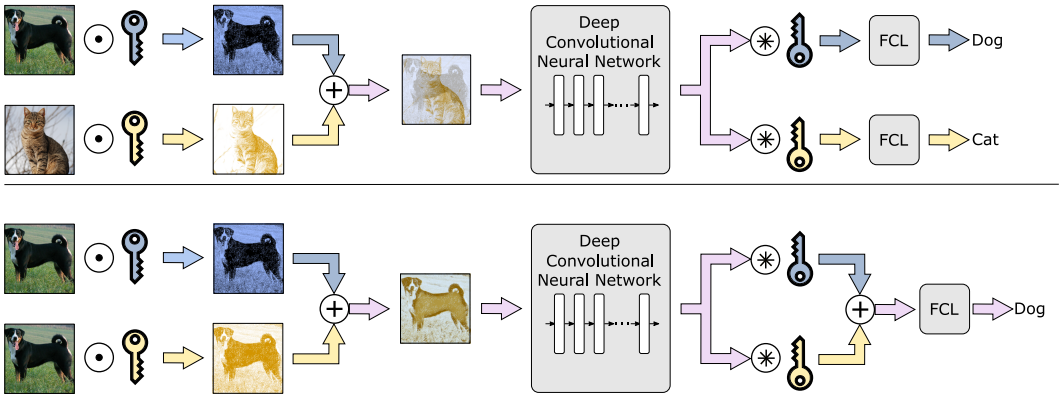


Figure A1: Depiction of a single trained MIMOConv performing dynamic inference. Instead of the fast N=2 mode (above) the same input can be inserted twice for the slow N=1 mode (below) effectively implementing an ensemble method. We can instantaneously switch between the modes.

A.4 Illustration of dynamic inference

To explore the idea of dynamic inference, suppose only two superposition channels are used with binding keys $a^{(1)}, a^{(2)}$ and unbinding keys $\tilde{a}^{(1)}, \tilde{a}^{(2)}$. We already know how the model performs standard computation in superposition (see Eq. (1) – (4) in the main text). Let us thus examine how a network with the same parameters can instead be used as an ensemble-method with higher accuracy, but lower throughput. A superposition is established of twice the same input x:

$$s = a^{(1)} \odot x + a^{(2)} \odot x \tag{15}$$

After applying the deep neural network f_θ to the superposition, we may unbind as

$$\tilde{a}^{(1)} \circledast f_\theta(s) \approx \tilde{a}^{(1)} \circledast f_\theta\left(a^{(1)} \odot x\right) + \tilde{a}^{(1)} \circledast f_\theta\left(a^{(2)} \odot x\right) \tag{16}$$

$$\approx f_\theta\left(x\right) + \tilde{a}^{(1)} \circledast f_\theta\left(a^{(2)} \odot x\right) \tag{17}$$

and

$$\tilde{a}^{(2)} \circledast f_\theta(s) \approx \tilde{a}^{(2)} \circledast f_\theta\left(a^{(1)} \odot x\right) + \tilde{a}^{(2)} \circledast f_\theta\left(a^{(2)} \odot x\right) \tag{18}$$

$$\approx \tilde{a}^{(2)} \circledast f_\theta\left(a^{(1)} \odot x\right) + f_\theta\left(x\right). \tag{19}$$

After averaging the two expressions, we get

$$\frac{1}{2}\left(\tilde{a}^{(1)} \circledast f_\theta(s) + \tilde{a}^{(2)} \circledast f_\theta(s)\right) \approx f_\theta\left(x\right) + noise \tag{20}$$

where $noise$ is a random noise vector and $f_\theta\left(x\right)$ is approximated as an average of two predictions. Owing to the introduction of stochasticity by the binding and unbinding process these predictions are decorrelated, i.e., each superposition channel is processed to some degree differently. See Figure A1 for an illustration.

A.5 Alternative throughput-increasing methods

Although not a focus of this work, we believe that computation in superposition can be combined with other throughput-increasing methods such as model-downsizing, quantization aware training, and pruning.

The Blessing of Dimensionality gives, in terms of dimensionality, an exponentially decreasing probability of interference for superpositions, even for (2-bit quantized) Rademachers. The extent to which these superpositions can be kept intact as linear layers act on them depends on the conditioning of the matrix (ideally nearly-isometric) not on the fidelity of its entries. As such we suspect that MIMOConv can be mixed with quantization, weight pruning, etc.

Regarding MIMOFormer, we can give quantitative insights. As is evident from Theorem 3, the error bounds have no dependence on the precision of projection weights, but depend only on the embedding dimensionality, the size of keys and queries, and the angles between them. Consequently, quantization, pruning, etc. are not in competition with our approach and can be easily combined.

Naturally, when combining different methods not only the gains but also the errors add up. However, with diminishing returns of each method we believe the combination of several to be most effective, especially given that our method is not competing with alternatives for the same resources of a model and allows it to conduct dynamic inference.

B MIMOConv Details

B.1 Inner-product preserving activation functions

Any inner-product preserving map is linear (see Appendix G Theorem 4). With activation functions being introduced to break the linearity of neural networks, they are innately at odds with inner-product preservation. According to [5], a trade-off can be reached between preserving inner products and introducing nonlinearities by replacing the ReLU activation function with shifted ReLU (sReLU)

$$sReLU_b(x) = ReLU(x - b) + b = \max(x, b), \tag{21}$$

where the trainable bias b determines the trade-off and is initialized to -1. However, in our experiments (see Appendix E.5), replacing sReLU with parametric ReLU (pReLU) [6], another activation function capable of choosing the extent of nonlinearity, defined as

$$pReLU_b(x) = ReLU(x) - b \cdot ReLU(-x) = max(x, 0) + b \cdot \ min(x, 0) \tag{22}$$

gives higher performance. The trainable parameter $b \in [-1, 1]$ controls the degree of linearity, where b=1 indicates fully linear behavior. It is initialized to b=0.5 at the beginning of training.

B.2 Details on isometric convolutional layers

As elaborated in the main text, we strive for inner-product preserving maps. With inner-product preserving maps being norm-preserving and by extension distance-preserving if linear, and with linear distance-preserving maps being norm preserving and according to the polarization identity also inner-product preserving, it holds that inner-product preserving maps are equivalent to linear isometries. Hence the name of the regularization term being *isometry regularization term*. The adopted regularization takes the form

$$L(W) = \tfrac{\gamma}{2} \|Conv(W, W) - \delta_{C_o}\|_F^2 , \qquad \delta_{C_o}[:, :, j, l] = I_{C_o \times C_o} \cdot \mathbb{1}_{j,l=\lfloor \frac{k}{2} \rfloor} \tag{23}$$

$$L(W^T) = \tfrac{\gamma}{2} \left\|Conv(W^T, W^T) - \delta_{C_i}\right\|_F^2 , \qquad \delta_{C_i}[:, :, j, l] = I_{C_i \times C_i} \cdot \mathbb{1}_{j,l=\lfloor \frac{k}{2} \rfloor} \tag{24}$$

where $W \in \mathbb{R}^{C_o \times C_i \times k \times k}$ contains the weights of a convolutional layer. C_o denotes the number of output feature maps, C_i the number of input feature maps, and k the (square) kernel size. W^T refers to a kernel with the first two dimensions of W transposed. In the notation of Einstein summations, the 2D convolution $Conv(U, V)$ evaluates to

$$O_{a,b,c,d} = U_{a,r,c+s,d+t} \cdot V_{b,r,s,t} \tag{25}$$

This is implemented in Pytorch by the usual zero-padded spatial 2D convolution taking an input in the first argument (with ranks: batch size, fmaps, height, width) and a convolutional kernel in the second argument (with ranks: output fmaps, input fmaps, kernel height, kernel width).

Unless the number of input fmaps and output fmaps coincide, only one of W and its adjoint W^T may be isometries. Thus we use $L(W)$ when $C_i > C_o$ and $L(W^T)$ otherwise.

For more information on why such a regularization term may help to preserve inner products, see [7] where it was first proposed.

B.3 Binding key regularization

We use a regularization term to keep the binding vectors ($a^{(i)}$) orthonormal:

$$L(a^{(1)}, \ldots, a^{(N)}) = \tfrac{\mu}{\binom{N}{2}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\tfrac{\langle a^{(i)}, a^{(j)} \rangle}{\left\|a^{(i)}\right\| \left\|a^{(j)}\right\|}\right)^2 + \tfrac{\mu}{N} \sum_{i=1}^{N} \left(\left\|a^{(i)}\right\| - 1\right)^2, \tag{26}$$

with hyperparameter μ. A grid search on the validation set found a value of μ=0.1 to give the best results. Alternatively, the binding keys may be frozen after random (Gaussian) initialization, guaranteeing orthogonality in the limit of high key dimension (see Appendix A.2).

C MIMOFormer Details

C.1 FAVOR+ in the Performer

Here, we revisit the Performer's FAVOR+ attention block [8] and in the next subsection we validate the use of the ReLU activation in the projection. FAVOR+ takes advantage of the fact that $a, b \mapsto \exp\Big(a^T b/\sqrt{D}\Big)$ is a kernel and can be represented as an explicit inner product (inverse kernel trick) in an infinite-dimensional space of transformed inputs. The mapping to this infinite-dimensional space is approximated with a randomized feature map $\phi : \mathbb{R}^D \to \mathbb{R}^R_+$ of finitely many entries. More explicitly, since

$$\mathbb{E}_{w\sim\mathcal{N}(0,I_D)}\left[\exp\Big(\tfrac{w_i^T q}{\sqrt[4]{D}} - \tfrac{\|q\|_2^2}{2\sqrt{D}}\Big)\cdot\exp\Big(\tfrac{w_i^T k}{\sqrt[4]{D}} - \tfrac{\|k\|_2^2}{2\sqrt{D}}\Big)\right] \tag{27}$$

$$= \exp\Big(\tfrac{-\|q\|_2^2-\|k\|_2^2}{2\sqrt{D}}\Big)\cdot\mathbb{E}_{w\sim\mathcal{N}(0,I_D)}[\exp\Big(w_i^T\tfrac{q+k}{\sqrt[4]{D}}\Big)] \tag{28}$$

$$= \exp\Big(\tfrac{-\|q\|_2^2-\|k\|_2^2}{2\sqrt{D}}\Big)\cdot\exp\Big(\tfrac{\|q+k\|_2^2}{2\sqrt{D}}\Big) = \exp\Big(q^T k/\sqrt{D}\Big), \tag{29}$$

drawing $w_1, \ldots, w_R \sim \mathcal{N}(0, I_D)$ i.i.d. induces a function $\phi : \mathbb{R}^D \to \mathbb{R}^R_+$ with components given by $\phi_i(x) = \exp\Big(\tfrac{w_i^T x}{\sqrt[4]{D}} - \tfrac{\|x\|_2^2}{2\sqrt{D}}\Big)/\sqrt{R}$ that, by the law of large number, approximates $\exp\Big(q^T k/\sqrt{D}\Big)$, i.e.

$$\langle\phi(k),\ \phi(q)\rangle \approx \exp\Big(q^T k/\sqrt{D}\Big). \tag{30}$$

Alternatively, by partitioning $w_1, \ldots, w_R$ into subsets of cardinality D and drawing each such subset from the orthogonal group before rescaling each w_i according to the χ_D-distribution it still holds $w_i \sim \mathcal{N}(0, I_D)$, but the entries are no longer independent. Using such orthogonal features, one obtains an unbiased estimate of $\exp\Big(q^T k/\sqrt{D}\Big)$ with lower variance than independently drawing $w_1, \ldots, w_R \sim \mathcal{N}(0, I_D)$, see [8].

The inverse kernel trick then allows FAVOR+ to take advantage of the associativity of matrix multiplication to give the following factored expression of dot-product attention:

$$o_i = \sum_{j=1}^{L} v_j \frac{\langle\phi(k_j),\phi(q_i)\rangle}{\sum_{j=1}^{L}\langle\phi(k_j),\phi(q_i)\rangle} = \frac{\sum_{j=1}^{L} v_j\left(\phi(k_j)^T\phi(q_i)\right)}{\underbrace{\sum_{j=1}^{L}\left(\phi(k_j)^T\phi(q_i)\right)}_{B_i}} = \frac{\overbrace{\sum_{j=1}^{L} v_j\,\phi(k_j)^T}^{A}\times\phi(q_i)}{\underbrace{\sum_{j=1}^{L}\phi(k_j)^T}_{C}\times\phi(q_i)}, \tag{31}$$

where A and C must only be computed once. With the computational complexity of evaluating ϕ being $\mathcal{O}(DR)$, this takes in both cases $\mathcal{O}(LDR)$. Computing $\phi(q_i)$ for all i requires another $\mathcal{O}(LDR)$. Finally, the matrix-vector product $\times$ for a given query position i takes $\mathcal{O}(DR)$ for the numerator and $\mathcal{O}(R)$ for the denominator. Thus, computing outputs o_i for all i takes only $\mathcal{O}(LDR)$, which can be a considerable improvement over the usual $\mathcal{O}(L^2 D)$ for long sequence lengths (L).

C.2 ReLU activation in FAVOR+S

To increase training stability, we use a ReLU-based projection $\phi_i(x) = ReLU(w_i^T x)/\sqrt{R\sqrt{D}}$ instead of the above-mentioned unbiased approximation of softmax through $\phi_i(x) = \exp\Big(\tfrac{w_i^T x}{\sqrt[4]{D}} - \tfrac{\|x\|_2^2}{2\sqrt{D}}\Big)/\sqrt{R}$. Such an approach was already mentioned in the Performer [8]. Here, we derive (new to our knowledge) a theoretical justification. As the following theorem expresses, the ReLU approximation leads to a (bi-)linear dependency on the norm of keys and queries while retain-

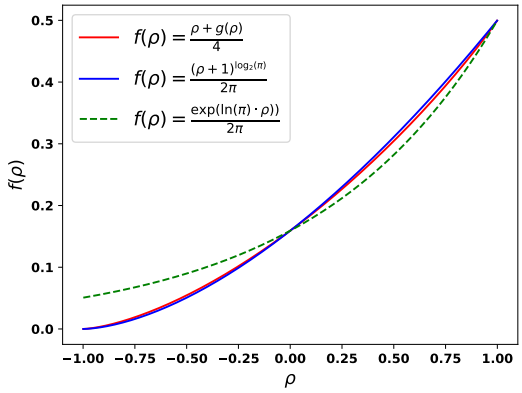


Figure A2: Approximation of $\mathbb{E}[ReLU(w^T x) \cdot ReLU(w^T y)]/(\|x\|\|y\|)$ (red line) with a polynomial (blue line). Softmax function is shown with green dashed line.

ing a roughly polynomial dependency (with exponent $\log_2(\pi) \approx 1.65$) on the query-key alignment $\rho = \langle x, y\rangle/(\|x\|\|y\|)$.

Theorem 2 (ReLU approximation)**.** *Let* $w \sim \mathcal{N}(0, I_D)$*. Then*

$$\mathbb{E}[ReLU(w^T x) \cdot ReLU(w^T y)] = \frac{\langle x, y\rangle + \|x\|\|y\| g(\frac{\langle x,y\rangle}{\|x\|\|y\|})}{4} \approx \|x\|\|y\| \frac{(\rho+1)^{\log_2(\pi)}}{2\pi} \tag{32}$$

for

$$g(\rho) = \frac{2}{\pi}\left[\sqrt{1-\rho^2} + |\rho| \arctan\left(\frac{|\rho|}{\sqrt{1-\rho^2}}\right)\right] \text{ with } \rho = \langle x, y\rangle/(\|x\|\|y\|) \tag{33}$$

Figure A2 illustrates the validity of interpreting $\mathbb{E}[ReLU(w^T x) \cdot ReLU(w^T y)]/(\|x\|\|y\|)$ as a polynomial. The general behaviour is also similar to the usual softmax function (green dashed line), although the norm of x, y no longer affecting the exponent gives greater stability.

Proof.

$$\mathbb{E}[ReLU(w^T x) \cdot ReLU(w^T y)] = \tfrac{1}{4}\mathbb{E}[(w^T x + \left|w^T x\right|)(w^T y + \left|w^T y\right|)] \tag{34}$$

$$= \frac{\mathbb{E}[w^T x \cdot w^T y] + \cancel{\mathbb{E}[w^T x \cdot \left|w^T y\right|]} + \cancel{\mathbb{E}[\left|w^T x\right| \cdot w^T y]} + \mathbb{E}[\left|w^T x \cdot w^T y\right|]}{4} \tag{35}$$

$$= \frac{\mathbb{E}[x^T w \cdot w^T y] + \mathbb{E}[\left|w^T x \cdot w^T y\right|]}{4} = \frac{x^T \mathbb{E}[w w^T] y + \mathbb{E}[\left|w^T x \cdot w^T y\right|]}{4} \tag{36}$$

$$= \frac{\langle x, y\rangle + \mathbb{E}[\left|w^T x \cdot w^T y\right|]}{4} \tag{37}$$

Let us now proceed to evaluate $\mathbb{E}[\left|w^T x \cdot w^T y\right|]$. First notice that $X = w^T x, Y = w^T y$ are jointly Gaussian with

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_X^2 & \rho\, \sigma_X \sigma_Y \\ \rho\, \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} \|x\|^2 & \langle x, y\rangle \\ \langle x, y\rangle & \|y\|^2 \end{bmatrix}\right). \tag{38}$$

where $\rho = \frac{Cov(X,Y)}{\sigma_X\, \sigma_Y} = \langle x, y\rangle/(\|x\|\|y\|)$ measures the alignment between x and y. Hence, by the innovations form for centered jointly Gaussians it holds

$$(X|Y = y) \sim \mathcal{N}(\mu_{x|y}(y), \sigma^2_{x|y}(y)) = \mathcal{N}(\rho\, \sigma_X \sigma_Y^{-1} y, (1-\rho^2)\sigma_X^2). \tag{39}$$

Thus, the expectation of the folded normal distribution is given by

$$\mathbb{E}[|X||Y = y] = \sigma_{x|y}(y)\sqrt{\tfrac{2}{\pi}} \exp\left(-\frac{\mu^2_{x|y}(y)}{2\sigma^2_{x|y}(y)}\right) + \mu_{x|y}(y)(1 - 2\Phi(-\tfrac{\mu_{x|y}(y)}{\sigma_{x|y}(y)})) \tag{40}$$

$$= \sqrt{1-\rho^2}\sigma_X \sqrt{\tfrac{2}{\pi}} \exp\left(-\frac{\rho^2 \sigma_X^2 \sigma_Y^{-2} y^2}{2(1-\rho^2)\sigma_X^2}\right) + \rho \sigma_X \sigma_Y^{-1} y (1 - 2\Phi(-\frac{\rho \sigma_X \sigma_Y^{-1} y}{\sqrt{1-\rho^2}\sigma_X})) \tag{41}$$

$$= \sqrt{1-\rho^2}\sigma_X\sqrt{\tfrac{2}{\pi}} \cdot \exp\Big(-\tfrac{\rho^2 y^2}{2(1-\rho^2)\sigma_Y^2}\Big) + |\rho|\sigma_X \cdot \sigma_Y^{-1} y(2\Phi(\tfrac{|\rho|\, y}{\sqrt{1-\rho^2}\sigma_Y}) - 1) \tag{42}$$

where $\Phi(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp\big(-t^2/2\big)dt$ denotes the cumulative distribution function of the standard normal distribution. Thus, we split the target expression in two:

$$\mathbb{E}[\big|w^T x \cdot w^T y\big|] = \mathbb{E}[|X||Y|] = \mathbb{E}[\mathbb{E}[|X| \cdot |Y| \;\big|\; Y]] = \mathbb{E}[|Y| \cdot \mathbb{E}[|X| \;\big|\; Y]] \tag{43}$$

$$= \sqrt{1-\rho^2}\sigma_X\sqrt{\tfrac{2}{\pi}} \cdot \mathbb{E}\left[|Y| \exp\Big(-\tfrac{\rho^2 Y^2}{2(1-\rho^2)\sigma_Y^2}\Big)\right] \tag{44}$$

$$+ |\rho|\sigma_X \cdot \mathbb{E}\left[|Y|\sigma_Y^{-1} Y(2\Phi(\tfrac{|\rho|\, Y}{\sqrt{1-\rho^2}\sigma_Y}) - 1)\right]. \tag{45}$$

For the first term, we get

$$\mathbb{E}\left[|Y| \exp\Big(-\tfrac{\rho^2 Y^2}{2(1-\rho^2)\sigma_Y^2}\Big)\right] = \int |y| \exp\Big(-\tfrac{\rho^2 y^2}{2(1-\rho^2)\sigma_Y^2}\Big) \tfrac{1}{\sqrt{2\pi}\sigma_Y} \exp\Big\{-\tfrac{y^2}{2\sigma_Y^2}\Big\} dy \tag{46}$$

$$= \sqrt{1-\rho^2} \int |y| \tfrac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_Y} \exp\Big(-\tfrac{y^2}{2(1-\rho^2)\sigma_Y^2}\Big) dy \tag{47}$$

$$= \sqrt{1-\rho^2}\, \mathbb{E}[|Z|] \;\text{ where } Z \sim \mathcal{N}(0, (1-\rho^2)\sigma_Y^2) \tag{48}$$

$$= \sqrt{1-\rho^2}\sqrt{1-\rho^2}\sigma_Y\sqrt{\tfrac{2}{\pi}} = (1-\rho^2)\sigma_Y\sqrt{\tfrac{2}{\pi}}, \tag{49}$$

using again the formula for the expectation of a folded normal distribution. The second term is a bit more involved to evaluate. We substitute $S = \frac{Y}{\sigma_Y}$, i.e. $S \sim \mathcal{N}(0,1)$:

$$\mathbb{E}[|Y| \tfrac{Y}{\sigma_Y}(2\Phi(\tfrac{|\rho|\, Y}{\sqrt{1-\rho^2}\sigma_Y}) - 1)] = \sigma_Y \mathbb{E}[|S| S(2\Phi(\tfrac{|\rho|\, S}{\sqrt{1-\rho^2}}) - 1)] \tag{50}$$

$$= \sigma_Y \mathbb{E}[S^2(2\Phi(\tfrac{|\rho|\, |S|}{\sqrt{1-\rho^2}}) - 1)] \tag{51}$$

$$= \sigma_Y \int p_s(s) s^2 \cdot (2\Phi(\tfrac{|\rho|\, |s|}{\sqrt{1-\rho^2}}) - 1) ds \tag{52}$$

$$= 2\sigma_Y \int p_s(s) s^2 \Phi(\tfrac{|\rho|\, |s|}{\sqrt{1-\rho^2}}) ds - \sigma_Y \tag{53}$$

$$= 4\sigma_Y \int_0^\infty s^2 p_s(s) \Phi(bs) ds - \sigma_Y \tag{54}$$

$$= 4\sigma_Y (\tfrac{1}{4} + \tfrac{1}{2\pi}(\tfrac{b}{1+b^2} + \arctan(b))) - \sigma_Y \tag{55}$$

$$= 4\sigma_Y (\tfrac{1}{4} + \tfrac{1}{2\pi}\left(\tfrac{b}{1+b^2} + \arctan(b)\right)) - \sigma_Y \tag{56}$$

$$= \tfrac{2\sigma_Y}{\pi}\left(|\rho|\sqrt{1-\rho^2} + \arctan\left(\tfrac{|\rho|}{\sqrt{1-\rho^2}}\right)\right). \tag{57}$$

where $b = \frac{|\rho|}{\sqrt{1-\rho^2}}$ and Owen's extensive list of integrals of Gaussian functions was employed. Therefore, we get

$$\mathbb{E}[\big|w^T x \cdot w^T y\big|] = \|x\| \cdot \|y\| \cdot \tfrac{2}{\pi}\left[(1-\rho^2)^{3/2} + \rho^2\sqrt{1-\rho^2} + |\rho| \arctan\left(\tfrac{|\rho|}{\sqrt{1-\rho^2}}\right)\right] \tag{58}$$

$$= \|x\| \cdot \|y\| \cdot \underbrace{\tfrac{2}{\pi}\left[\sqrt{1-\rho^2} + |\rho| \arctan\left(\tfrac{|\rho|}{\sqrt{1-\rho^2}}\right)\right]}_{g(\rho)}, \tag{59}$$

which yields for the full expression

$$\mathbb{E}[ReLU(w^T x) \cdot ReLU(w^T y)] = \frac{\langle x, y\rangle + \|x\| \cdot \|y\| \cdot g(\frac{\langle x,y\rangle}{\|x\| \cdot \|y\|})}{4}. \tag{60}$$

□

C.3 Attention normalization in FAVOR+S

We present the computation of the normalization term in FAVOR+S. Recall the usual FAVOR+ formulation from Appendix C.1

$$o_i = \overbrace{\sum_{j=1}^{L} v_j\ \phi(k_j)^T}^{A} \times \phi(q_i) \Big/ \underbrace{\overbrace{\sum_{j=1}^{L} \phi(k_j)^T}^{C} \times \phi(q_i)}_{B_i}. \tag{61}$$

As was remarked in Appendix C.1, construction of A, C, and $\phi(q_i)\ \forall i$ each requires $\mathcal{O}(LDR)$ computational complexity. Therefore, it is central to also compute the normalization factor B_i in superposition. We start with the construction of:

$$C_s^{(m)} = \underbrace{\left(\sum_{j=1}^{L} \phi(\sum_{w=1}^{N} k_j^{(m,w)})^T\right)}_{\text{construct } \forall m \text{ in } \mathcal{O}(LM(DR+ND))} \qquad Q_i^{(n)} = \underbrace{\phi(\sum_{t=1}^{M} q_i^{(t,n)})}_{\text{construct } \forall i,n \text{ in } \mathcal{O}(LN(DR+MD))} \tag{62}$$

Since $C_s^{(m)} \in \mathbb{R}^{1\times R}$, the evaluation of $\times$ for all query positions i and channels (m,n) is relatively inexpensive, demanding only $\mathcal{O}(LMNR)$ operations. According to the two approximations (P) from FAVOR+ and (H), which is explored thoroughly in Appendix D, it holds:

$$\begin{aligned}
B_i^{(m,n)} &= C_s^{(m)} \times Q_i^{(n)} && (63)\\
&= \sum_{j=1}^{L} \phi\left(\sum_{w=1}^{N} k_j^{(m,w)}\right)^T \phi\left(\sum_{t=1}^{M} q_i^{(t,n)}\right) && (64)\\
&\overset{P}{\approx} \sum_{j=1}^{L} \exp\left(\langle \sum_{w=1}^{N} k_j^{(m,w)},\ \sum_{t=1}^{M} q_i^{(t,n)} \rangle / \sqrt{D}\right) && (65)\\
&\overset{H}{\approx} \sum_{j=1}^{L} \exp\left(\langle k_j^{(m,n)},\ q_i^{(m,n)} \rangle / \sqrt{D}\right) && (66)
\end{aligned}$$

In total, while still setting M=N to balance the load, computing $B_i^{(m,n)}$ for all channels (m,n) and query positions i demands a runtime of $\mathcal{O}(LNDR + LN^2D + LN^2R)$, a significant improvement over $\mathcal{O}(LN^2DR)$ for computing B_i separately for each channel $(m,n) \in \{1,\dots,N\}^2$.

How the normalization is incorporated depends on the MIMOFormer instantiation. In the first case using superposition exclusively for the attention block (att.), the normalization scalar is directly applied on the output tokens after unbinding, i.e.,

$$o_i^{(m,n)} = \frac{S_i^{(n)} \odot \tilde{a}^{(m,n)}}{B_i^{(m,n)}}, \tag{67}$$

where $\tilde{a}^{(m,n)}$ is the unbinding key. In the second instantiation, where additionally the MLP computes in superposition (att.+MLP), we jointly normalize the output by the sum of all normalization scalars over m (where we enjoy additional computational savings by in fact already summing over m in the construction of $C_s = \sum_m C_s^{(m)}$):

$$\bar{S}_i^{(n)} = \frac{S_i^{(n)}}{\sum_{m=1}^{M} B_i^{(m,n)}}. \tag{68}$$

D Theoretical Basis for Noise Mitigation in FAVOR+S

As mentioned in the main text, our derivations in Section 4.1 rely on two estimates:

$$\phi(k)^T\phi(q) \overset{P}{\approx} \exp\Big(\langle k,q\rangle/\sqrt{D}\Big) \qquad \langle\sum_{w=1}^{N} k_j^{(u,w)},\ \sum_{t=1}^{M} q_i^{(t,n)}\rangle \overset{H}{\approx} \underbrace{\langle k_j^{(u,n)},\ q_i^{(u,n)}\rangle}_{\text{intended signal}} \tag{69}$$

The approximation P, which improves with increasing $R = \dim(\phi(\overline{q}_i))$, is due to FAVOR+ and is quantified in [8] whereas the approximation H follows from:

Inter-channel distortion. *The probability that inter-channel attention distorts the intended signal of the dot product by a factor outside* $[1-\alpha, 1+\alpha]$ *has various upper bounds, most notably decaying exponentially w.r.t.* $D\alpha^2\cos^2(\measuredangle(\overline{k}_j^{(u,n)}, \overline{q}_i^{(u,n)}))/(NM-1)^2$.

We proceed to make this statement exact:

Theorem 3 (FAVOR+S Inter-Channel Noise)**.** *The probability that inter-channel attention distorts the true signal by a factor outside* $[1-\alpha, 1+\alpha]$ *shows the following tail bounds. Denote by*

$$P = \mathbb{P}\left\{\left\langle\sum_{w=1}^{N} k_j^{(u,w)},\ \sum_{t=1}^{M} q_i^{(t,n)}\right\rangle \Bigg/ \sqrt{D} \notin [1-\alpha,1+\alpha]\cdot\langle\overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle/\sqrt{D}\right\} \tag{70}$$

where

$$k_j^{(m,n)} := \overline{k}_j^{(m,n)} \odot a^{(m,n)} \qquad q_j^{(m,n)} := \overline{q}_j^{(m,n)} \odot a^{(m,n)} \tag{71}$$

with $a^{(m,n)}$ *being i.i.d. bipolar vectors of Rademachers and* (m,n) *denoting a channel. It holds*

$$P \leq \sum_{w=1}^{N}\sqrt{\sum_{t=1,\ldots,M}^{(u,w)\neq(t,n)} 1/\Xi_{(u,w)}^{(t,n)}} \qquad P \leq \sum_{t=1}^{M}\sqrt{\sum_{w=1,\ldots,N}^{(u,w)\neq(t,n)} 1/\Xi_{(u,w)}^{(t,n)}} \tag{72}$$

$$P \leq \sum_{\substack{w=1,\ldots,N\\ t=1,\ldots,M}}^{(u,w)\neq(t,n)} 1/\Xi_{(u,w)}^{(t,n)} \qquad P \leq 2\sum_{\substack{w=1,\ldots,N\\ t=1,\ldots,M}}^{(u,w)\neq(t,n)} \exp\left(-\frac{\Xi_{(u,w)}^{(t,n)}}{2(NM-1)^2}\right) \tag{73}$$

for

$$\Xi_{(u,w)}^{(t,n)} = \frac{\alpha^2\left|\langle\overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle\right|^2}{\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2} = \frac{\alpha^2\cos^2(\measuredangle(\overline{k}_j^{(u,n)},\overline{q}_i^{(u,n)}))\left\|\overline{k}_j^{(u,n)}\right\|_2^2\left\|\overline{q}_i^{(u,n)}\right\|_2^2}{\left\|\overline{k}_j^{(u,w)}\odot\overline{q}_i^{(t,n)}\right\|_2^2} \tag{74}$$

which for keys/queries of similar size according to Theorem 6 in Appendix G typically scales as

$$\Xi_{(u,w)}^{(t,n)} \sim D\,\alpha^2\cos^2(\measuredangle(\overline{k}_j^{(u,n)}, \overline{q}_i^{(u,n)})) \tag{75}$$

Proof. First notice that since

$$\sum_{\substack{w=1,\ldots,N\\ t=1,\ldots,M}} \left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle \notin [1-\alpha,1+\alpha]\cdot\langle\overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle \tag{76}$$

$$\iff \tag{77}$$

$$\left|\sum_{\substack{w=1,\ldots,N\\ t=1,\ldots,M}}^{(u,w)\neq(t,n)} \left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha\left|\langle\overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle\right|, \tag{78}$$

we may instead derive tail bounds on

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\ldots,N\\ t=1,\ldots,M}}^{(u,w)\neq(t,n)} \left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha\left|\langle\overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle\right|\right\}. \tag{79}$$

We shall derive the tail bounds for a threshold α and only replace it by $\alpha\left|\langle \overline{k}_j^{(u,n)},\ \overline{q}_i^{(u,n)}\rangle\right|$ in a final step. Applying Markov, the triangle inequality, and linearity of expectation gives

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\ t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right|>\alpha\right\} \tag{80}$$

$$\leq \mathbb{E}\left[\left|\sum_{\substack{w=1,\dots,N\\ t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right|\right]\Big/\alpha \tag{81}$$

$$\leq \sum_{w=1}^{N}\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right|\right]\Big/\alpha \tag{82}$$

$$=\sum_{w=1}^{N}\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\left\langle \overline{k}_j^{(u,w)}\odot a^{(u,w)},\ \overline{q}_i^{(t,n)}\odot a^{(t,n)}\right\rangle\right|\right]\Big/\alpha, \tag{83}$$

where $a^{(\cdot,\cdot)}$ denotes one of the (independent) binding vectors with entries given by independent Rademacher random variables and $\bar{k}_{(\cdot)}^{(\cdot,\cdot)}$, $\bar{q}_{(\cdot)}^{(\cdot,\cdot)}$ the unbound keys and queries respectively of a given channel and token position. For $\epsilon_p^{(t)}$ denoting independent Rademacher random variables we may simplify to

$$\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\left\langle \overline{k}_j^{(u,w)}\odot a^{(u,w)},\ \overline{q}_i^{(t,n)}\odot a^{(t,n)}\right\rangle\right|\right] \tag{84}$$

$$=\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p a_p^{(u,w)}\left(\overline{q}_i^{(t,n)}\right)_p a_p^{(t,n)}\right|\right] \tag{85}$$

$$=\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\epsilon_p^{(t)}\right|\right]. \tag{86}$$

The famous Khintchine inequality determines up to a constant the behavior of the expectation as

$$\tfrac{1}{\sqrt{2}}\sqrt{\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\right)^2}\leq\mathbb{E}\left[\left|\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\epsilon_p^{(t)}\right|\right] \tag{87}$$

$$\leq\sqrt{\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\right)^2}. \tag{88}$$

Hence, the Markov bound gives

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\ t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right|>\alpha\right\}\leq\sum_{w=1}^{N}\sqrt{\sum_{t=1,\dots,M}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2}\Big/\alpha \tag{89}$$

and (90)

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\ t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right|>\alpha\right\}\leq\sum_{t=1}^{M}\sqrt{\sum_{w=1,\dots,N}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2}\Big/\alpha \tag{91}$$

where the latter follows by symmetry. Alternatively, we could also apply Chebyshev to the problem, i.e.

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\, q_i^{(t,n)}\right\rangle\right|>\alpha\right\} \tag{92}$$

$$=\mathbb{P}\left\{\left(\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\, q_i^{(t,n)}\right\rangle\right)^2>\alpha^2\right\} \tag{93}$$

$$\leq\mathbb{E}\left[\left(\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\, q_i^{(t,n)}\right\rangle\right)^2\right]\Big/\alpha^2 \tag{94}$$

$$=\mathbb{E}\left[\left(\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle \overline{k}_j^{(u,w)}\odot a^{(u,w)},\, \overline{q}_i^{(t,n)}\odot a^{(t,n)}\right\rangle\right)^2\right]\Big/\alpha^2 \tag{95}$$

$$=\mathbb{E}\left[\left(\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p a_p^{(u,w)}\left(\overline{q}_i^{(t,n)}\right)_p a_p^{(t,n)}\right)^2\right]\Big/\alpha^2 \tag{96}$$

$$=\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\sum_{\substack{w'=1,\dots,N\\t'=1,\dots,M}}^{(u,w')\neq(t',n)}\sum_{p'=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\left(\overline{k}_j^{(u,w')}\right)_{p'}\left(\overline{q}_i^{(t',n)}\right)_{p'}\cdot \tag{97}$$
$$\cdot\,\mathbb{E}\left[a_p^{(u,w)}a_p^{(t,n)}a_{p'}^{(u,w')}a_{p'}^{(t',n)}\right]\Big/\alpha^2$$

Now, since for $(u,w)\neq(t,n)$, $(u,w')\neq(t',n)$ and $(t,w)\neq(t',w')$ the cardinality of $\{(u,w),(t,n),(u,w'),(t',n)\}$ is at least three, at least one entry has no duplicate. Let w.l.o.g. be that entry $a_p^{(u,w)}$. Then by independence, one has

$$\mathbb{E}\left[a_p^{(u,w)}a_p^{(t,n)}a_{p'}^{(u,w')}a_{p'}^{(t',n)}\right]=\mathbb{E}\left[a_p^{(u,w)}\right]\mathbb{E}\left[a_p^{(t,n)}a_{p'}^{(u,w')}a_{p'}^{(t',n)}\right]=0 \tag{98}$$

Consequently, all terms with $(t,w)\neq(t',w')$ vanish. Also, for $(u,w)\neq(t,n)$ and $p\neq p'$ all four entries are independent, i.e.

$$\mathbb{E}\left[a_p^{(u,w)}a_p^{(t,n)}a_{p'}^{(u,w)}a_p^{(t,n)}\right]=\mathbb{E}\left[a_p^{(u,w)}\right]\mathbb{E}\left[a_p^{(t,n)}\right]\mathbb{E}\left[a_p^{(u,w)}\right]\mathbb{E}\left[a_p^{(t,n)}\right]=0 \tag{99}$$

Hence, we have that all cross-terms vanish, which gives

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\, q_i^{(t,n)}\right\rangle\right|>\alpha\right\} \tag{100}$$

$$\leq\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2\mathbb{E}\left[\left(a_p^{(u,w)}\right)^2\left(a_p^{(t,n)}\right)^2\right]\Big/\alpha^2 \tag{101}$$

$$=\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2\Big/\alpha^2 \tag{102}$$

For higher orders than two, we can no longer rely on the absence of duplicates in the set of multiplied binding vectors. For example, a cross-term such as

$$\mathbb{E}\left[\underbrace{a_p^{(u,w)}a_p^{(t,n)}}_{1}\underbrace{a_{p'}^{(u,w')}a_{p'}^{(t',n)}}_{2}\underbrace{a_p^{(u,w)}a_p^{(t,n)}}_{3}\underbrace{a_{p'}^{(u,w')}a_{p'}^{(t',n)}}_{4}\right] = 1 \tag{103}$$

is non-vanishing even for $(t, w, p) \neq (t', w', p')$. Hence, for higher orders, we will have to work with

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha\right\} \leq \mathbb{P}\left\{\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left|\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha\right\} \tag{104}$$

$$\leq \mathbb{P}\left\{\exists(u,w)\neq(t,n) : \left|\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha/(NM-1)\right\} \tag{105}$$

$$\leq \sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\mathbb{P}\left\{\left|\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| > \alpha/(NM-1)\right\} \tag{106}$$

$$= \sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\mathbb{P}\left\{\left|\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p a_p^{(u,w)}\left(\overline{q}_i^{(t,n)}\right)_p a_p^{(t,n)}\right| > \alpha/(NM-1)\right\} \tag{107}$$

$$= \sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\mathbb{P}\left\{\left|\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p\left(\overline{q}_i^{(t,n)}\right)_p\epsilon_p\right| > \alpha/(NM-1)\right\} \tag{108}$$

for $\{\epsilon_p\}_{p=1}^D$ independent Rademacher random variables. Finally, we may apply Hoeffding's inequality (Theorem 5), which gives

$$\mathbb{P}\left\{\left|\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\left\langle k_j^{(u,w)},\ q_i^{(t,n)}\right\rangle\right| \geq \alpha\right\} \leq 2\sum_{\substack{w=1,\dots,N\\t=1,\dots,M}}^{(u,w)\neq(t,n)}\exp\left(-\frac{\alpha^2}{2(NM-1)^2\sum_{p=1}^{D}\left(\overline{k}_j^{(u,w)}\right)_p^2\left(\overline{q}_i^{(t,n)}\right)_p^2}\right) \tag{109}$$

□

E Experimental Setup and Ablation Study on MIMOConv

E.1 Experimental setup

Datasets

CIFAR10 and CIFAR100. The CIFAR10 [9] dataset contains 60,000 images, each of resolution 32×32, divided into 50,000 training and 10,000 test images. The images are grouped into ten classes, each with 6000 examples. CIFAR100 has the same number of images and resolutions, but contains 100 classes each with 600 examples.

MNIST. The MNIST dataset [10] provides grey-scale images, each of resolution 28×28, containing hand-written digits. The 60,000 training samples are divided into a training and validation set containing 55,000 and 5000 samples, respectively. Finally, the test accuracy is reported on the test set containing 10,000 samples.

SVHN. The street view house number (SVHN) dataset [11] provides cropped images of house number plates, each of resolution 32×32. As in MNIST, the task is to classify the printed digits (from 0 to 9). It contains 73,257 RGB images for training and 26,032 for testing.

Training setups

The experiments are run on an NVIDIA A100 Tensor Core GPU with 80 GB memory and 8 CPU cores. All experiments are repeated five times with different random seeds. We report the mean and standard deviations of accuracy to account for variability in training. Overall, all MIMOConv experiments together required 3290 GPU hours when accumulating all the training runs required for generating the main results and the ablations study.

CIFAR10 and CIFAR100. In all experiments on CIFAR10 and CIFAR100, stochastic gradient descent (SGD) with momentum is used. Unless otherwise noted, we train for 1200 epochs using the OneCycleLR policy [12] with cosine annealing for two phases (30% increase, 70% decrease of learning rate). The initial learning rate is set to 0.008, the maximal learning rate to 0.2, and the final learning rate to 2e-5. Momentum is cycled inversely with base momentum set to 0.85 and maximal momentum set to 0.95. Due to overfitting, WideResNet28-10 shows higher test accuracy when trained with 200 epochs than 1200 epochs; hence, Table 1 in the main text shows WideResNet28-10's performance with 200 training epochs. For all parameters, except for the bias in shifted ReLU, the slope in parametric ReLU, and binding/unbinding keys, weight decay with value 1e-5 is applied.

The images are standardized on each color channel by subtracting the mean and dividing by the standard deviation. To augment data, the images are randomly flipped horizontally, and a random 32×32 crop is taken after zero padding the images on each side by four pixels. Furthermore, the data agnostic augmentation strategy mixup [13] is employed with parameter α=1, which is decisive for obtaining high accuracy.

The batch size is set to 128 elements per superposition channel (i.e., $128N$). Thus, after binding, a batch of 128 superpositions traverses the CNN. Increasing superposition channels would decrease the number of update steps per epoch. To correct for that, in each epoch, the dataset is traversed as often as the number of superposition channels used. While the results presented in this paper come from a train/test split of the datasets, the training dataset is split into a 90/10 train/validation split for all model design and hyperparameter choices. Furthermore, to decrease the degrees of freedom in the experiments, the remaining hyperparameters (learning rate, weight decay, mixup parameters) are tuned to yield good performance on the base model WideResNet28-10. Finally, to stabilize training, the average gradient norm of each epoch is recorded, and the batches of the subsequent epoch are discarded (without repetition) if their update gradient norm exceeds the recorded average of the last epoch by a factor of 10.

Training MIMOConv for 1200 epochs takes 11 hours independent of the number of superposition channels owing to the batch loading corrections that account for the same number of training steps.

MNIST. The experiments on MNIST use a similar setup to the ones on CIFAR: the same learning rate scheduler, batch size, weight decay, and mixup coefficients are used. In contrast to CIFAR, the

Table A2: Millions of multiply-accumulate (MMAC) operations per sample on CIFAR-100. Number in parenthesis shows the relative share of the overall complexity.

	First conv. layer	Binding	Rest of conv. layers	Unbinding	FCL	**Total**
WideResNet-28-10	0.49 (0.009%)	n.a.	5245 (99.99%)	n.a.	0.064 (0.001%)	**5251**
WideIsoNet-28-10	0.49 (0.009%)	n.a.	5245 (99.99%)	n.a.	0.064 (0.001%)	**5251**
MIMOConv (N=1)	1.97 (0.04%)	4.19 (0.08%)	5329 (99.88%)	0.41 (0.008%)	0.064 (0.001%)	**5335**
MIMOConv (N=2)	1.97 (0.07%)	4.19 (0.15%)	2664 (99.75%)	0.41 (0.015%)	0.064 (0.002%)	**2671**
MIMOConv (N=4)	1.97 (0.15%)	4.19 (0.31%)	1332 (99.50%)	0.41 (0.031%)	0.064 (0.005%)	**1339**

number of training epochs is reduced to 50. Moreover, the images are center-cropped to 20×20 pixels both in training and testing. A random horizontal flip serves as data augmentation during training.

We reduce the depth of MIMOConv from 28 to 10 layers, and the width factor from $10\times$ to $1\times$, hence we call it MIMOConv-10-1. Moreover, the initial width factor of the first convolutional layer is also set to $1\times$.

SVHN. In the SVHN experiments, we train the standard MIMOConv-28-10 architecture for 200 epochs. The remaining hyperparameters are kept the same as in CIFAR. During training, random crop with padding is used.

E.2 Computational complexity

Table A2 breaks down MIMOConv's computational benefits (in MMAC) on CIFAR100, as N is increased from 1 to 4. As shown, the integration of variable binding mechanisms via binding and unbinding operations is inconsequential, amounting to only between 0.008% (N=1) and 0.031% (N=4) of the total MACs for MIMOConv.

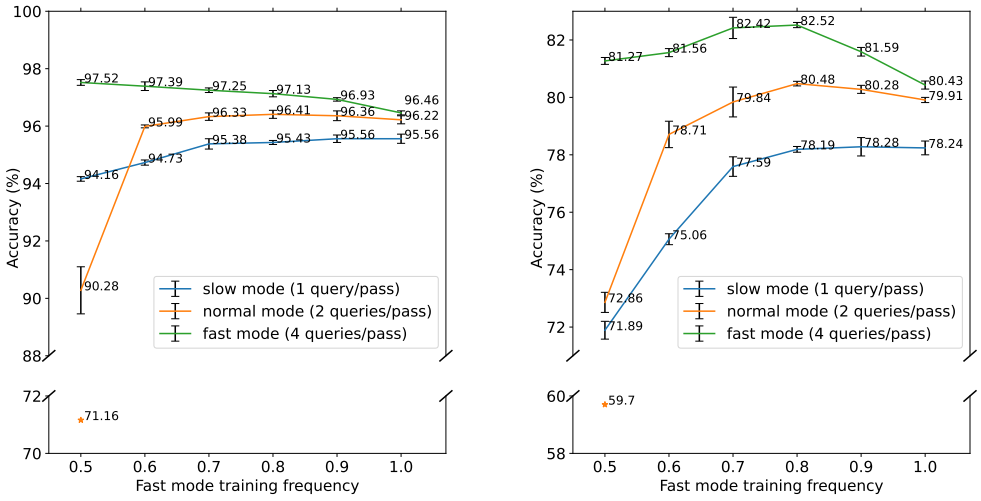
On MNIST, MIMOConv-10-1 has a computational cost of 5.10 MMAC per sample at N=1 superposition channels and manages to reduce the cost to 0.47 MMAC per sample at N=16, an effective reduction of $10.9\times$. The reduction is smaller than N due to the computationally dominating first convolutional layer, which is not operating in superposition. Yet, MIMOConv shows a notably higher reduction than the LeNet-like model (CNN+nonlinear ($8\times$)) from DataMUX [14], a key competitor, which requires 0.88 MMAC and 0.65 MMAC per sample at N=1 and N=16, respectively.

E.3 The effectiveness of position-wise binding (PWHRR) and isometry regularization

DataMux [14] also explored CNNs that compute in superposition and reported its findings on MNIST. Even with a trivial downsizing for fair comparison from a 28-layer very-wide ($10\times$) MIMOConv to a 10-layer narrow ($1\times$) MIMOConv, our method scales much better to high superposition channels (N) than DataMUX does. Indeed, our model shows an accuracy of 80.4% against their 52.9% in case of N=16 superposition channels (highest number of channels reported by DataMUX for vision tasks), despite being computationally cheaper (0.47 MMAC/s vs. 0.65 MMAC/s). DataMux's binding overhead results in a mere $1.35\times$ reduction in MACs compared to our $10.9\times$ as N goes from 1 to 16 demonstrating superior scaling of our method.

We attribute the improved performance to a set of innovations which we reiterate here: MIMOConv applies position-wise binding (PWHRR), thus retaining the locality property present in natural images and vital for CNNs, whilst as discussed by Murahari, Vishvak, et al. their primary binding does not. As a workaround they proposed binding via two layers CNNs each outputting 8 feature maps. The resulting (pixel-wise) superposition in a low-dimensional space (8-D) leads to high interference. Additionally to using an expensive binding mechanism, it also makes the first layer of the model 8 times as expensive no matter the number of superpositions. We are able to circumvent this issue by applying the first layer of the CNN *before* the pixel-wise binding, increasing the dimensionality of each pixel in an easy-to-understand manner.

Another difference to their work is our use of *isometric neural networks* to further reduce interference during the processing of superposed images.



(a) CIFAR10 (b) CIFAR 100

Figure A3: Dynamic inference with MIMOConv trained for 1200 epochs in slow and fast mode depending on the fast mode training frequency. Each model is evaluated in slow (1 input/pass), normal (2 inputs/pass), and fast mode (4 inputs/pass). We report the average accuracy and the standard deviation (error bars) over five runs with different seeds. Outliers in normal mode (indicated with $\star$) are not used for standard deviation computation.

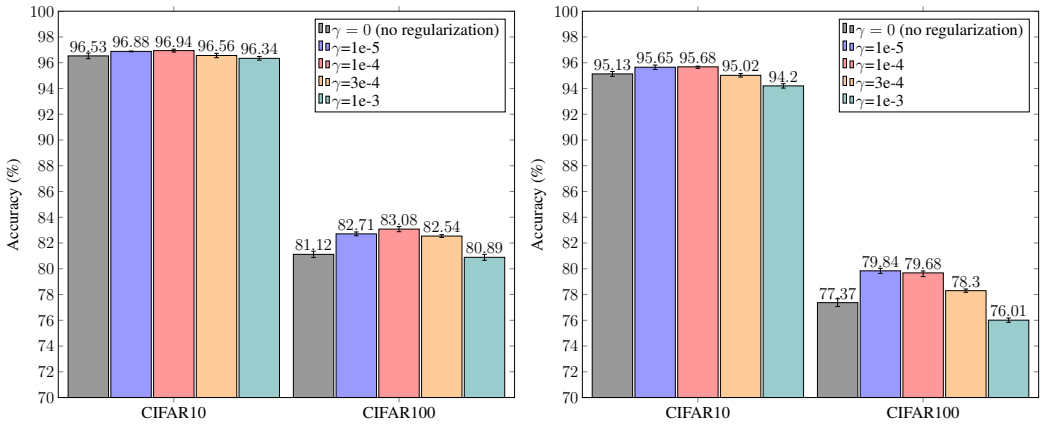
E.4 Dynamic inference

Dynamic inference enables the instantaneous on-demand partitioning of the superposition channels to select an operating point with a suitable speed/accuracy trade-off. Even though every MIMOConv can be configured to perform dynamic inference at any time, exposing MIMOConv to dynamic switching between different modes during training is beneficial. We set up a model with four channels and consider a fast (4 inputs/pass), normal (2 inputs/pass), and slow mode (1 input/pass). The fast mode maps each input to one channel; the normal mode distributes two inputs over pairs of channels; and the slow mode uses all channels for the same input. We then train the models for different frequencies in fast and slow modes. The normal mode is not used during training of the model. The potential switching between the modes happens between every batch.

Figure A3 shows the classification accuracy on CIFAR10 (a) and CIFAR100 (b) when using dynamic models trained with varying fast mode frequencies. The models, which are trained with a different fraction of inputs in fast mode, are evaluated in slow, normal, and fast modes. As is expected, increasing the fast mode training frequency is beneficial for both datasets when only looking at the fast mode inference. Conversely, the normal inference mode benefits from a mixture of fast and slow mode training, whereby a fast mode training frequency of 80% achieves the highest accuracy on both datasets. There is a notable volatility in the performance of normal mode at a fast mode training frequency of 0.5, which could be due to the optimization getting stuck in a local optimum exclusively learning a slow mode instance.

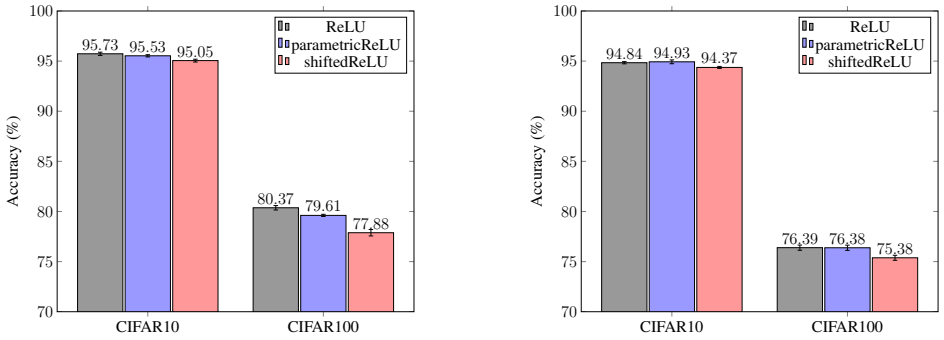
E.5 Ablation study on CIFAR10/100

Isometry regularization of CNN weights. We evaluate the impact of isometry regularization to the CNN weights by varying the orthogonal regularization coefficient (γ), described in Eq. (23) and Eq. (24). All models are trained for 200 epochs to reduce the training time. As can be observed in Figure A4, orthogonal regularization enables the network to perform notably better both for configurations with a single superposition channel and two superposition channels. However, a strong regularization hinders the ability of the network to adapt to the task. With two superposition channels,



(a) N=1 image superposed (b) N=2 images superposed

Figure A4: MIMOConv with varying orthogonal regularization coefficient (γ) for N=1 and N=2 superpositions. All models are trained for 200 epochs. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.



(a) N=2 and 200 training epochs (b) N=5 and 1200 training epochs

Figure A5: MIMOConv with different activation functions. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.

the performance difference is more striking, i.e., orthogonal regularization is more important. In all other experiments, an orthogonal regularization coefficient of γ=1e-4 was used.

Isometry at activation functions. We investigate the ReLU, shifted ReLU, and parametric ReLU activation functions to give the model more control over the extent of isometry. Each activation function owns separate, trainable parameters which are not shared between the feature maps and layers. Experimental results are shown in Figure A5. When using N=2 superposition channels and 200 training epochs, ReLU outperforms both parametric ReLU and shifted ReLU. For longer training times (1200 epochs) and more superposition channels (N=5), the network prefers parametric ReLU.

Surprisingly, in the case of low superposition counts, the model develops highly non-isometric parametric ReLU activation functions, as seen in Figure A6. With the convolutional layers being pushed toward isometry due to the isometry regularization term and residual skip connections increasing isometry further, the network seeks balance through strongly non-isometric activation functions. Nevertheless, increasing the number of images superposed incentivizes the network to learn isometric activation functions. It is unclear if the performance degradation induced by superpositions originates in interference or in the attempt of the network to reduce interference through isometry, and it is likely that some balance between the two is reached. Further research will be needed to gain more insight into the benefits and drawbacks of isometry. We employ parametric ReLU in all other experiments as its performance is comparable to ReLU, but it allows more degrees of freedom and hence could give additional performance benefits under different network configurations.

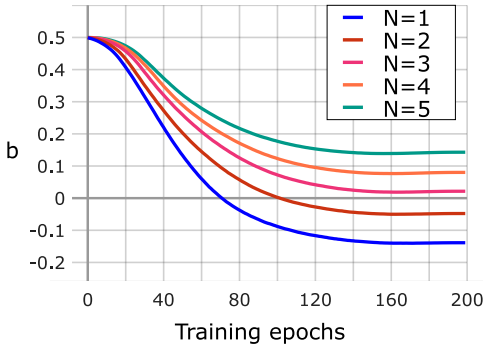


Figure A6: Average parametric ReLU parameter (b) during training on CIFAR10 for different number of superposition channels (N).

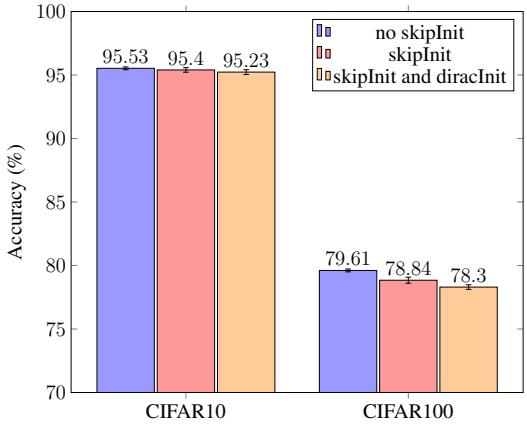


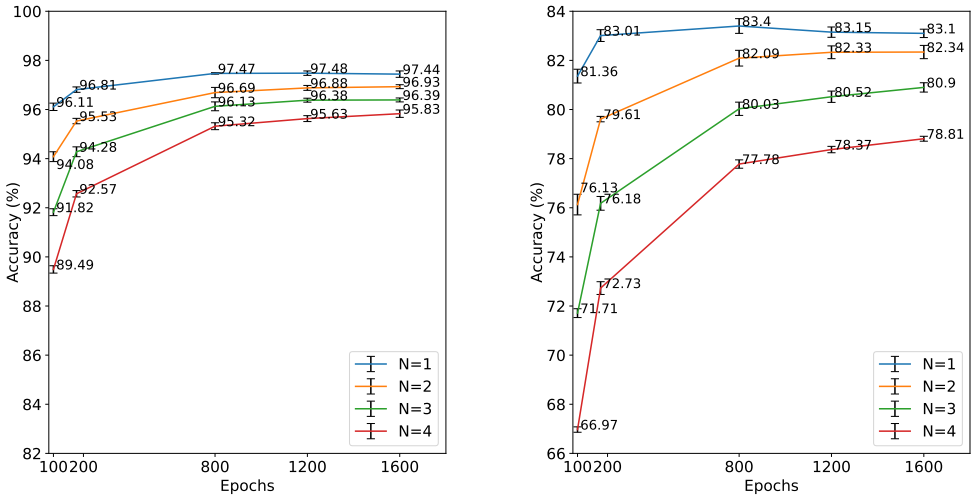
Figure A7: Effect of skipInit and diracInit on accuracy in superposition mode N=2, 200 epochs. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.

SkipInit and DiracInit. In [7], the benefit of skip initialization [15] via inducing maximum isometry at initialization was discussed. Furthermore, an initialization scheme for the convolutional layers as an identity, called *diracInit*, was promoted. In our experiments, both additions worsen the performance on CIFAR10 and CIFAR100, as seen in Figure A7.

Number of training epochs. Training a neural network model to simultaneously handle multiple images while providing high accuracy is a more difficult task than without superposition present. Figure A8 shows that the performance gap between single-image mode and multiple-image mode narrows when training for more epochs. For each epoch, the training set is passed through the network as many times as superposition channels were used: this corrects for the larger batch size used in superposition modes and to have roughly equal training time for each epoch.

Number of feature maps in the first layer. Compared to a standard ResNet, the WideResNet architecture [16] increases the number of feature maps of every layer by a *width factor*. The only exception is the first layer, which has 16 feature maps in WideResNet-28-10. However, binding takes place after the first layer in MIMOConv. In order to enter the regime of high dimensionality and to benefit from the Blessing of Dimensionality, we experiment with an additional parameter termed *initial width factor*, which increases the number of feature maps of the output of the first convolutional layer. For example, an initial width factor of 4 yields $4 \cdot 16$=64 feature maps after the first convolutional layer. The initial width factor can be configured independently from the general width factor.

Figure A9 shows the performance against variable initial width factors, while the general width factor is fixed to 10. Initial width factors 2 and 4 give satisfactory results, while factors 1 and 8 yield very unstable training (as indicated by the large variance). We attribute this to the large step either from 3



(a) CIFAR10 (b) CIFAR100

Figure A8: Accuracy of MIMOConv when trained with a variable number of epochs. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.

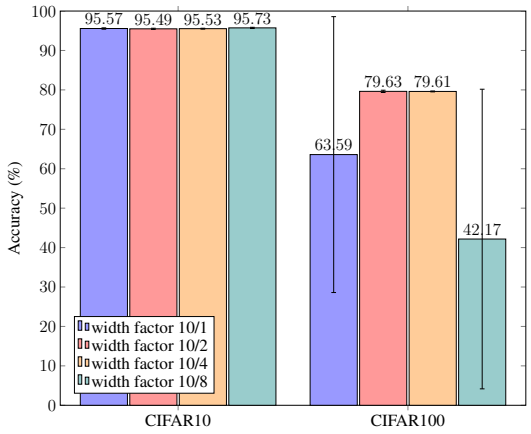


Figure A9: Initial width majorly affects stability in MIMOConv (N=2, 200 epochs). Configurations are labeled with *width factor/initial width factor*. Standard initial width factor of vanilla WideResNet would be 1. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.

feature maps to $16 \cdot 8$ in the first layer (initial width factor 8), or from $1 \cdot 16$ feature maps to $16 \cdot 10$ in the second layer (initial factor 1). An initial width factor of 4 strikes a good balance and provides stability; hence we will use this value for all other experiments.

MIMOConv width factor. WideResNet architectures identify the number of feature maps as the most crucial factor in determining the capacity of a model. At the same time, a higher number of feature maps means less interference between bound vectors. In Figure A10, the benefits of large width factors can be observed. Notice that width enters both the number of parameters and the computational complexity quadratically unless grouped convolutions are used, where the number of groups increases with the channel width. Not exploring grouped convolutions, a trade-off between performance and accuracy has to be struck. We shall go with a width factor of 10 for all other experiments.

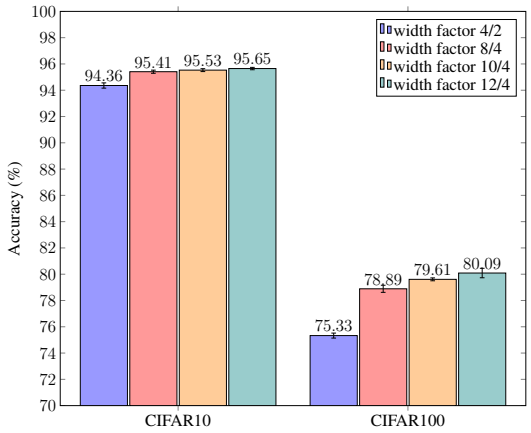
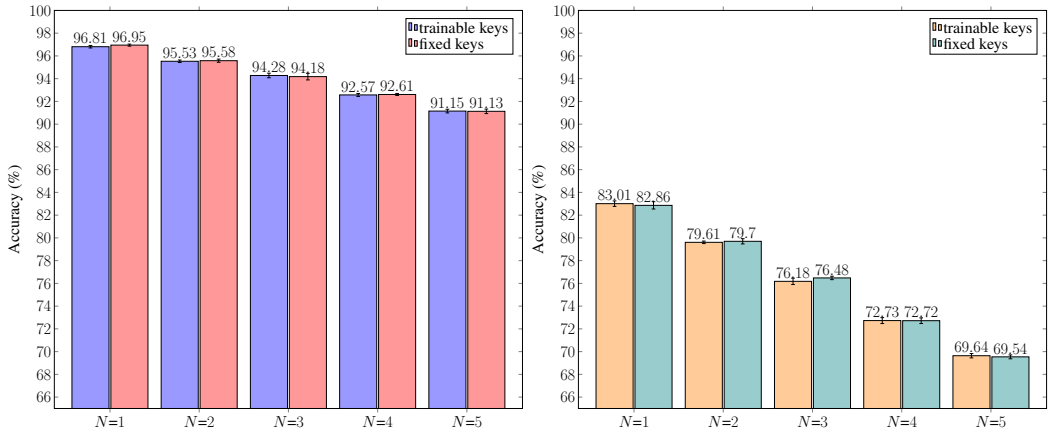


Figure A10: Performance effect of MIMOConv width factor (N=2, 200 epochs). Configurations are labeled with *width factor/inital width factor*. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.



(a) CIFAR10 (b) CIFAR100

Figure A11: MIMOConv with trainable or frozen binding keys for different numbers of superposition channels (N). Models are trained for 200 epochs. We report the average accuracy and the standard deviation (error bars) over five runs with different seeds.

Freezing of binding keys. We investigate, if the binding keys can be frozen during training. As can be observed in Figure A11, keys do not need to be trainable while still maintaining a high accuracy across a wide range of superposition channel counts. We note that unbinding keys are always left trainable.

F Experimental Setup and Evaluations on MIMOFormer

F.1 Experimental setup

Datasets

LRA. The long range arena (LRA) benchmark [17] is a suite of tasks consisting of sequences ranging from 1K to 16K tokens, covering a wide range of data types and modalities. Below, we list the LRA tasks used in this work:

- **ListOps**: This dataset [18] tests the capability of modeling hierarchically structured data comprised of long sequences with operators (max, mean, median, modulo sum). The sequence length is up to 2K, and the task is to perform a ten-way classification. ListOps is released under an MIT license.
- **Text**: The IMDb reviews dataset [19] is a document classification benchmark. The benchmark uses byte-level sequences of up to 4K, entailing a binary classification task.
- **Retrieval**: This task measures the model's capability of storing text document information into a compressed representation and matching or retrieving it against other documents. The ACL Anthology Network [20] is a byte/character level dataset with sequence lengths of up to 4K. It is a binary classification task. The ACL anthology corpus is released under the CC BY-NC 4.0 license.
- **Image**: In this task, RGB images of resolution 32×32 from the CIFAR10 [9] dataset are flattened and classified by the sequence classification model. As a result, this is a ten-way classification task with sequences of length 1024.
- **Pathfinder**: This task [21] requires a model to decide if two points (circles) are connected by a path of dashes on a black and white 2D image of dimension 32×32. The 2D images are flattened to sequences of length 1024, and classified by the sequence model as a binary classification. Pathfinder is released under MIT license.

LRA further contains a more fine-grained version of Pathfinder, called Pathfinder-X, with 128×128 image resolution leading to a sequence length of 16K. However, all Transformer variants considered in [17] (including the Performer) failed to learn this task. Since this work demonstrates the MIMO-capability of self-attention models rather than increasing their sequence length, we do not consider Pathfinder-X.

Synthetic sequence benchmarks. We use two synthetic benchmarks [22] which measure the basic reasoning capability of neural sequence models.

- **Associative recall**: The task is to remember associations between pairs of tokens. For example, given a sequence of tokens *a 2 c 4 b 3 d 1*, if the model is prompted with *a*, the expected output is *2*, the token following *a* in the input sequence. If it were prompted with *b*, the correct output would be *3*, etc. Each sequence contains 40 characters, whereby a dictionary with 20 different characters is used.
- **Induction head**: The task is to recall content after a special token (e.g., $\vdash$). For example, the string $a\ d\ b \vdash g\ f \ldots\ h\ c \vdash$ would expect the response g. Each sequence contains 30 characters, whereby a dictionary with 20 different characters is used.

Both tasks provide a training set (5000 examples) and a test set (500 examples).

Training setup

As with MIMOConv, the experiments are performed on an NVIDIA A100 Tensor Core GPU with 80 GB memory and 8 CPU cores. All experiments are repeated five times with a different random seed. We report the mean and standard deviations of accuracy to account for variability in training. Overall, the training and evaluation of all MIMOFormer models, including the ablation of the number of training steps, consumed 3112 GPU hours.

Table A3: Architecture and training setup on LRA. L=number of layers; N_{head}=number of heads; D_{head}=head dimension; E=embedding dimension; D_{hidden}=hidden dimension in MLP; Bs=batch size; Lr=learning rate.

	Model					Training				
	L	N_{head}	D_{head}	E	D_{hidden}	Bs	Lr	Warmup steps	Train steps	Dropout
MIMOFormer										
Listops	6	8	64	512	2048	64	1e-4	1000	20,000	0.1
Text	6	8	64	512	2048	32	1e-4	8000	40,000	0.1
Retrieval	4	4	32	128	512	32	1e-4	800	60,000	0.1
Image	3	4	64	64	128	256	1e-4	175	70,000	0.1
Pathfinder	4	8	128	128	128	256	1e-4	312	124,800	0.1
DataMUX										
Listops	12	12	120	120	3072	48	2e-5	0	20,000	0.1

Table A4: Training time (hours) on the long range arena (LRA) using an NVIDIA A100 GPU.

	ListOps	Text	Retrieval	Image	Pathfinder	Total
Deep models	L=6, H=8	L=6, H=8	L=4, H=4	L=3, H=4	L=4, H=8	
Performer (reproduced)	$3.45^{\pm 0.00}$	$5.89^{\pm 0.00}$	$5.93^{\pm 0.01}$	$3.55^{\pm 0.02}$	$26.66^{\pm 0.02}$	$45.47^{\pm 0.04}$
MIMOFormer (N=2, att.)	$3.96^{\pm 0.03}$	$7.69^{\pm 0.01}$	$5.99^{\pm 0.02}$	$3.51^{\pm 0.01}$	$25.42^{\pm 0.17}$	$46.57^{\pm 0.20}$
MIMOFormer (N=2, att.+MLP)	$3.42^{\pm 0.03}$	$6.71^{\pm 0.26}$	$5.38^{\pm 0.01}$	$3.14^{\pm 0.01}$	$22.50^{\pm 0.18}$	$41.15^{\pm 0.31}$
MIMOFormer (N=4, att.)	$3.09^{\pm 0.01}$	$5.89^{\pm 0.01}$	$4.35^{\pm 0.02}$	$2.73^{\pm 0.29}$	$21.26^{\pm 0.21}$	$37.32^{\pm 0.27}$
MIMOFormer (N=4, att.+MLP)	$2.42^{\pm 0.26}$	$4.48^{\pm 0.26}$	$3.44^{\pm 0.01}$	$2.10^{\pm 0.01}$	$17.01^{\pm 0.37}$	$29.44^{\pm 0.31}$
Wide models	L=1, H=48	L=1, H=48	L=1, H=16	L=1, H=12	L=1, H=32	
Performer (reproduced)	$2.46^{\pm 0.05}$	$5.23^{\pm 0.01}$	$5.26^{\pm 0.01}$	$3.21^{\pm 0.39}$	$29.52^{\pm 0.03}$	$45.68^{\pm 0.38}$
MIMOFormer (N=2, att.)	$2.54^{\pm 0.00}$	$4.79^{\pm 0.01}$	$4.46^{\pm 0.01}$	$2.93^{\pm 0.01}$	$23.45^{\pm 0.01}$	$38.17^{\pm 0.02}$
MIMOFormer (N=2, att.+MLP)	$2.31^{\pm 0.02}$	$4.30^{\pm 0.00}$	$4.16^{\pm 0.00}$	$2.65^{\pm 0.00}$	$20.69^{\pm 0.06}$	$34.12^{\pm 0.07}$
MIMOFormer (N=4, att.)	$1.95^{\pm 0.08}$	$4.04^{\pm 0.15}$	$3.17^{\pm 0.14}$	$2.25^{\pm 0.10}$	$19.43^{\pm 0.37}$	$30.84^{\pm 0.66}$
MIMOFormer (N=4, att.+MLP)	$1.57^{\pm 0.08}$	$3.39^{\pm 0.00}$	$2.71^{\pm 0.12}$	$1.87^{\pm 0.08}$	$15.51^{\pm 0.04}$	$25.06^{\pm 0.16}$

LRA. Table A3 lists the deep MIMOFormer architecture and the training setup for each task in the LRA benchmark [17]. Both wide and deep models use the same training setup, but wide models shrink to a single layer $L = 1$ with inverse scaling in the number of heads N_{head}. MIMOFormer uses the same base architecture (number of heads, layers, dimensions, etc.) as proposed in the initial evaluation on LRA [17]. In addition, Table A3 also summarizes the model configuration and the settings used for the training of DataMux [14], our main competitor, on the Listops dataset. We base our DataMux model on the *Roberta* architecture as specified in [14]. We adjusted the number of heads, the number of layers, the embedding dimension, and the hidden dimension to approximately match the MIMOFormer's number of parameters. Before training on ListOps, the scaled-down DataMux model is first pre-trained on the "retrieval warm-up task" as outlined in [14].

The training setup and the evaluation setup for MIMOFormer is based on code provided by [23]. Training uses an Adam optimizer (β_1=0.9 and β_2=0.99) with a OneCycleLR policy [12] and additional warmup. All MIMOFormer configurations use the same number of training steps per task; we note however that configurations with many superposition channels converge more slowly and consequently might benefit from additional training steps. Dropout after the attention block is applied. Finally, the output tokens are fed through average pooling and classified with a task-specific readout mechanism.

MIMOFormer faces training issues when the number of superposition channels is high (e.g., N=4). To this end, we propose a curriculum learning strategy where the number of superposition is reduced to N'=$N/2$ at the beginning of the training. This warmup period is set to $1/6$th of the total number of training epochs. The overall training setup, including the learning rate scheduling, remains the same.

Table A4 shows the training time for the reported models. Since all MIMOFormer configurations use the same number of training steps, we observe a reduced training time using a large number of superposition channels. Contrary to MIMOConv, we do not repeatedly send batches through to ensure equal training time.

Table A5: Training for more steps improves MIMOFormer accuracy. We report the average test accuracy (%) on LRA over five runs with different seeds when training the model for $0.5\times$ / $1\times$ the training steps described in Table A3. MIMOFormer uses an equal number of query superpositions (N) and value-key tensor product superpositions (M), i.e., N=M. Computation in superposition is performed either in attention only (att.) or in both attention and MLP (att.+MLP). L is the number of layers, H the number of heads.

	ListOps	Text	Retrieval	Image	Pathfinder	Avg.
Deep models	L=6, H=8	L=6, H=8	L=4, H=4	L=3, H=4	L=4, H=8	
Transformer [24]	36.37	64.27	57.46	42.44	71.40	53.39
Performer [8]	18.01	65.40	53.82	42.77	77.05	51.41
Performer (reproduced)	37.93 / 38.94	65.45 / 65.70	81.37 / 81.58	40.04 / 40.14	73.01 / 73.82	59.56 / 60.04
MIMOFormer (N=2, att.)	38.07 / 38.08	64.47 / 65.00	77.16 / 79.37	37.33 / 38.21	68.19 / 72.36	57.04 / 58.61
MIMOFormer (N=2, att.+MLP)	37.28 / 37.65	64.30 / 64.39	73.33 / 76.02	31.62 / 33.85	56.31 / 67.98	52.57 / 55.98
MIMOFormer (N=4, att.)	31.39 / 37.22	64.73 / 64.59	57.67 / 60.99	27.48 / 28.16	49.86 / 55.50	46.23 / 49.29
MIMOFormer (N=4, att.+MLP)	17.91 / 17.74	53.97 / 60.71	66.24 / 72.20	23.30 / 24.01	50.26 / 50.33	42.33 / 45.00
Wide models	L=1, H=48	L=1, H=48	L=1, H=16	L=1, H=12	L=1, H=32	
Performer (reproduced)	39.13 / 39.40	65.73 / 65.73	83.20 / 83.67	41.53 / 41.67	73.88 / 74.11	60.70 / 60.93
MIMOFormer (N=2, att.)	38.31 / 38.90	65.40 / 65.39	78.71 / 81.27	39.98 / 40.25	71.97 / 73.51	58.87 / 59.86
MIMOFormer (N=2, att.+MLP)	37.76 / 37.59	64.73 / 64.64	75.26 / 78.30	35.14 / 36.69	67.60 / 68.22	56.10 / 57.09
MIMOFormer (N=4, att.)	36.97 / 37.71	64.61 / 64.22	71.50 / 74.99	31.13 / 35.43	67.56 / 69.52	54.35 / 56.37
MIMOFormer (N=4, att.+MLP)	17.41 / 18.52	64.24 / 63.53	68.91 / 74.30	24.21 / 26.54	53.36 / 56.33	45.63 / 47.84

Table A6: Billions of multiply-accumulate (GMAC) operations per sample on Text (subtask of LRA). Model configuration reads $L[\text{ayers}] = 6, N_{\text{head}} = 8, D_{\text{head}} = 64, E[\text{mbedding}] = 512, D_{\text{hidden}} = 2048$, and $R = 256$ where R determines the fidelity of the FAVOR+ attention approximation. Number in parenthesis shows the relative share of the overall complexity.

	K/Q/V Projections	Attention	Binding & Unbinding	MLPs	Readout Layer	**Total**
Transformer	19.34 (14.9%)	58.80 (45.3%)	n.a.	51.62 (39.8%)	0.001 (0.001%)	**129.8**
Performer	19.34 (21.4%)	19.58 (21.6%)	n.a.	51.62 (57.0%)	0.001 (0.001%)	**90.5**
MIMOFormer (N=2, att.)	19.34 (23.0%)	13.05 (15.5%)	0.050 (0.06%)	51.62 (61.4%)	0.001 (0.001%)	**84.1**
MIMOFormer (N=2, att.+MLP)	19.34 (35.2%)	9.80 (17.8%)	0.050 (0.09%)	25.81 (46.9%)	0.001 (0.002%)	**55.0**
MIMOFormer (N=4, att.)	19.34 (23.9%)	9.78 (12.1%)	0.050 (0.06%)	51.62 (63.9%)	0.001 (0.001%)	**80.8**
MIMOFormer (N=4, att.+MLP)	19.34 (52.0%)	4.90 (13.2%)	0.050 (0.14%)	12.90 (34.7%)	0.001 (0.003%)	**37.2**

Synthetic sequence benchmarks. We use a light-weight MIMOFormer with two layers, one head, an embedding dimension of 32, and a hidden dimension of 128. The model is trained with SGD for 400 epochs using a learning rate of 5e-4, batch size 32, and a weight decay of 0.1.

To configure DataMux for associative recall, we use the *SimpleLM* language model with 30.5K trainable parameters specified in the Safari repository [1] and insert multiplexing and demultiplexing layers at the input and the output of the model as specified in [14]. Before experimenting with N=2 channels, we first tested the setup with a single channel where DataMux reached 99% accuracy.

F.2 Number of training steps

In our standard training setup, we train the Performer and MIMOFormer models for a large number of training steps ($\approx 2\times$ of what was described in [23]). Here, we show the benefit of a longer training procedure. Table A5 compares the performance of the models when trained with $0.5\times$ or $1\times$ as many steps as the training setup reported in Table A3. Note that the test accuracies reported in Table 2 of the main text also used the standard training setup ($1\times$). The longer training procedure improves the Performer's accuracy marginally (0.23–0.48% gain). Conversely, MIMOFormer notably benefits from the longer training in both deep (1.57–3.06% gain) and wide models (0.99–2.21% gain).

F.3 Computational complexity

As can be deduced from Table A6, the integration of variable binding mechanisms via binding and unbinding operations is inconsequential. It amounts to only between 0.06% and 0.14% of the computational complexity for MIMOFormer despite being performed at each attention layer. The

[1] https://github.com/HazyResearch/safari

K/Q/V projections make up a considerable part of the overall computational complexity; hence, computing them in superposition would further reduce the number of computes per input.

F.4 The importance of faithful attention scores

DataMux [14], claims to retain high performance for subsets of the GLUE [25] and CoNLL-2003 [26] benchmarks, despite using up to 40 inputs in superposition. However, as discussed in [27], none of the tasks reported on require attention layers at all. Indeed, DataMUX does not redesign its attention algorithm, but keeps a single scalar attention score $A_{i,j}$ for each pair of token positions, which effectively multiplies the (unnormalized) attention score of each (protected) superposition channel:

$$A_{i,j} = \exp\left(\langle \sum_{w=1}^{N} k_j^{(w)},\ \sum_{t=1}^{N} q_i^{(t)} \rangle / \sqrt{D}\right) \approx \exp\left(\sum_{w=1}^{N} \langle k_j^{(w)},\ q_i^{(w)} \rangle / \sqrt{D}\right) = \prod_{t=1}^{N} A_{i,j}^{(w)} \tag{110}$$

As our experiments confirm (see Section 5.2), on more nuanced tasks in NLP such as "associative recall" and "induction head", which require faithful attention, their method drops to 20.04% and 6.06% for N=2, while ours, at a score of 96.52% and 99.40% respectively, succeeds. Despite investing significant efforts in the training of DataMUX, it cannot perform on these synthetic tasks. This is in line with the findings of [22] which identifies the lack of attention as the reason that the Structured State Space Sequence (S4) model [28] is able to completely outperform state of the art in LRA [17], but is not competetive for large language models. In contrast to DataMUX, our work approximates true attention and our theoretical derivations show convergence to actual dot-product attention as the hidden dimension increases, giving us an even stronger case for applicability to large language models (for instance, GPT-3 uses embedding dimension 12,888, far exceeding the maximum of 512 we report on).

G Supporting Theorems

The theorems presented in this section are of general nature and stated for completeness.

Theorem 4. *Any inner-product preserving map $T : X \to Y$ between two inner-product spaces X, Y is linear*

Proof. Let $u, v \in X$ and $\lambda \in \mathbb{C}$. Then

$$\begin{aligned}
&\|T(\lambda u + v) - \lambda T u - Tv\|^2 = \langle T(\lambda u + v) - \lambda T u - Tv, T(\lambda u + v) - \lambda T u - Tv\rangle && (111)\\
&= \langle T(\lambda u + v), T(\lambda u + v)\rangle - 2\lambda\langle T(\lambda u + v), Tu\rangle - 2\langle T(\lambda u + v), Tv\rangle \\
&\quad + \lambda^2\langle Tu, Tu\rangle + 2\lambda\langle Tu, Tv\rangle + \langle Tv, Tv\rangle && (112)\\
&= \langle \lambda u + v, \lambda u + v\rangle - 2\lambda\langle \lambda u + v, u\rangle - 2\langle \lambda u + v, v\rangle + \lambda^2\langle u, u\rangle + 2\lambda\langle u, v\rangle + \langle v, v\rangle && (113)\\
&= 2\langle \lambda u + v, \lambda u + v\rangle - 2\lambda\langle \lambda u + v, u\rangle - 2\langle \lambda u + v, v\rangle && (114)\\
&= 2\langle \lambda u + v, \lambda u + v\rangle - 2\langle \lambda u + v, \lambda u + v\rangle && (115)\\
&= 0 && (116)
\end{aligned}$$

which implies

$$T(\lambda u + v) = \lambda T u + Tv \tag{117}$$

□

Theorem 5 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent bound random variables satisfying $|X_i| \leq a_i$ and $\mathbb{E}[X_i] = 0$. Then,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| > t\right\} \leq 2\exp\left(-\tfrac{t^2}{2\sum_{i=1}^n a_i^2}\right) \tag{118}$$

Proof. Following [29], we shall prove that

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left(-\tfrac{t^2}{2\sum_{i=1}^n a_i^2}\right) \tag{119}$$

from which by symmetry and union bound the statement follows. By Markov's inequality and for $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} &= \mathbb{P}\left\{\lambda\sum_{i=1}^n X_i > \lambda t\right\} = \mathbb{P}\left\{\exp\left(\lambda\sum_{i=1}^n X_i\right) > \exp(\lambda t)\right\} && (120)\\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] \Big/ \exp(\lambda t) = \exp(-\lambda t)\prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right] && (121)
\end{aligned}$$

where the last equality follows from independence of $\{X_i\}_{i=1}^n$. Because the function $x \mapsto \exp(\lambda x)$ is convex it holds

$$\exp(\lambda x) \leq \tfrac{a_i + x}{2a_i}\exp(\lambda a_i) + \tfrac{a_i - x}{2a_i}\exp(-\lambda a_i) \tag{122}$$

for all $x \in [-a_i, a_i]$. Thus, since $|X_i| \leq a_i$ we may use the above and that $\mathbb{E}[X_i] = 0$ to bound $\mathbb{E}\left[\exp(\lambda X_i)\right]$

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X_i)] &\leq \mathbb{E}\left[\tfrac{a_i + x}{2a_i}\exp(\lambda a_i) + \tfrac{a_i - x}{2a_i}\exp(-\lambda a_i)\right] = \tfrac{1}{2}(e^{\lambda a_i} + e^{-\lambda a_i}) = \cosh(\lambda a_i) && (123)\\
&= \sum_{n=0}^\infty \tfrac{(\lambda a_i)^{2n}}{(2n)!} \leq \sum_{n=0}^\infty \tfrac{(\lambda a_i)^{2n}}{2^n n!} = \exp\big((\lambda a_i)^2/2\big) && (124)
\end{aligned}$$

where the Taylor expansion of $\cosh(\cdot)$ and $\exp\big((\cdot)^2/2\big)$ were used and the penultimate step is given by $(2n)! \geq 2^n n!$. Hence, we get for any $\lambda > 0$

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp(-\lambda t)\prod_{i=1}^n \exp\big((\lambda a_i)^2/2\big) = \exp\left(-\lambda t + \tfrac{\lambda^2}{2}\sum_{i=1}^n a_i^2\right) \tag{125}$$

and in particular for $\lambda = t/(\sum_{i=1}^n a_i^2) > 0$ one gets

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left(-\tfrac{t^2}{2\sum_{i=1}^n a_i^2}\right) \tag{126}$$

□

Theorem 6 (On the Norm of Hadamard Products)**.** *Let $X \in S^{D-1}$ follow an arbitrary distribution and let $Y \in S^{D-1}$ be uniformly distributed and independent from X. Then*

$$\mathbb{E}\left[\|X \odot Y\|_2^2\right] = \tfrac{1}{D} \tag{127}$$

and

$$\mathbb{P}\left\{\|X \odot Y\|_2^2 \leq \tfrac{1+\beta}{D}\right\} \geq \tfrac{1}{1+1/\beta} \tag{128}$$

Proof. Since by definition $\|Y\|_2^2 = 1$, it follows by rotational symmetry and linearity of expectation that

$$\mathbb{E}\left[Y_q^2\right] = \sum_{p=1}^D \mathbb{E}\left[Y_p^2\right] \Big/ D = \mathbb{E}\left[\sum_{p=1}^D Y_p^2\right] \Big/ D = \tfrac{1}{D} \tag{129}$$

Also, by linearity of expectation and independence

$$\mathbb{E}\left[\|X \odot Y\|_2^2\right] = \mathbb{E}\left[\sum_{p=1}^D X_p^2 Y_p^2\right] = \sum_{p=1}^D \mathbb{E}\left[X_p^2\right]\mathbb{E}\left[Y_p^2\right] = \tfrac{1}{D}\sum_{p=1}^D \mathbb{E}\left[X_p^2\right] = \tfrac{1}{D}\mathbb{E}\left[\sum_{p=1}^D X_p^2\right] = \tfrac{1}{D} \tag{130}$$

Hence, we can apply Markov to get

$$\mathbb{P}\left\{\|X \odot Y\|_2^2 \leq \tfrac{1+\beta}{D}\right\} = 1 - \mathbb{P}\left\{\|X \odot Y\|_2^2 \geq \tfrac{1+\beta}{D}\right\} \geq 1 - \tfrac{D \cdot \mathbb{E}\left[\|X \odot Y\|_2^2\right]}{1+\beta} = 1 - \tfrac{1}{1+\beta} = \tfrac{1}{1+1/\beta} \tag{131}$$

□

H Limitations

MIMONets exploit the Blessing of Dimensionality, that with high probability exponentially many (in dimension D) vectors are almost orthogonal. Although the components of MIMONet are made near isometric through regularization, a certain number of (hidden) dimensions is still necessary. This naturally limits MIMONets to large (oftentimes over-parametrized) models or models employing low-rank decompositions.

The number of inputs that can be superposed without incurring heavy losses in accuracy is limited given a fixed neural network due to increasingly strong interference between the superposition channels.

The proposed superposition capable attention mechanism converges to faithful attention (without interference between channels) as the embedding dimension increases, but at the price of only a speedup of N when using N^2 superposition channels. Being built on linearized attention such as FAVOR+, it further inherits all their benefits (linear scaling) and drawbacks (limited parallelization and increased memory accesses for autoregressive training (see Section 3.1 in [30]). On the other hand, trivial superposition would yield a speedup of N^2 instead, but at the cost of blurring the attention scores with each token-token score summarizing attention in all superposition channels at once. Such models employing blurry attention are limited to application where imprecise "summarizing" information suffices.

References

[1] D. Kleyko, D. Rachkovskij, E. Osipov, and A. Rahimi, "A survey on hyperdimensional computing aka vector symbolic architectures, part I: Models and data transformations," *ACM Computing Surveys*, vol. 55, no. 6, 2022.

[2] T. A. Plate, "Holographic reduced representations," *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 623–641, 1995.

[3] S. I. Gallant and T. W. Okaywe, "Representing objects, relations, and sequences," *Neural Computation*, vol. 25, no. 8, pp. 2038–2078, 2013.

[4] R. W. Gayler, "Multiplicative binding, representation operators & analogy," in *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, 1998.

[5] W. Hu, L. Xiao, and J. Pennington, "Provable benefit of orthogonal initialization in optimizing deep linear networks," in *International Conference on Learning Representations*, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[7] H. Qi, C. You, X. Wang, Y. Ma, and J. Malik, "Deep isometric learning for visual recognition," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 7824–7835.

[8] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser *et al.*, "Rethinking attention with performers," in *International Conference on Learning Representations (ICLR)*, 2020.

[9] A. Krizhevsky, "Learning multiple layers of features from tiny images," *University of Toronto*, 2009.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] N. Yuval, "Reading digits in natural images with unsupervised feature learning," in *Proceedings of the NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[12] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006, 2019, pp. 369–386.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[14] V. Murahari, C. Jimenez, R. Yang, and K. Narasimhan, "DataMUX: Data multiplexing for neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 17 515–17 527, 2022.

[15] S. De and S. Smith, "Batch normalization biases residual blocks towards the identity function in deep networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 19 964–19 975, 2020.

[16] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, September 2016, pp. 87.1–87.12.

[17] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," in *International Conference on Learning Representations (ICLR)*, 2021.

[18] N. Nangia and S. R. Bowman, "Listops: A diagnostic dataset for latent tree learning," *arXiv preprint arXiv:1804.06028*, 2018.

[19] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[20] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, *Language Resources and Evaluation*, vol. 47, pp. 919–944, 2013.

[21] D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre, "Learning long-range spatial dependencies with horizontal gated recurrent units," *Advances in neural information processing systems*, vol. 31, 2018.

[22] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Re, "Hungry hungry hippos: Towards language modeling with state space models," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.

[23] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh, "Nyströmformer: A nyström-based algorithm for approximating self-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 138–14 148.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[25] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations (ICLR)*, 2019.

[26] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[27] M. Hassid, H. Peng, D. Rotem, J. Kasai, I. Montero, N. A. Smith, and R. Schwartz, "How much does attention actually attend? Questioning the importance of attention in pretrained transformers," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 1403–1416.

[28] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations (ICLR)*, 2022.

[29] A. S. Bandeira, *Mathematics of Data Science*, 2020, book draft version 0.1. [Online]. Available: https://people.math.ethz.ch/~abandeira/teaching.html

[30] W. Hua, Z. Dai, H. Liu, and Q. Le, "Transformer quality in linear time," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 9099–9117.