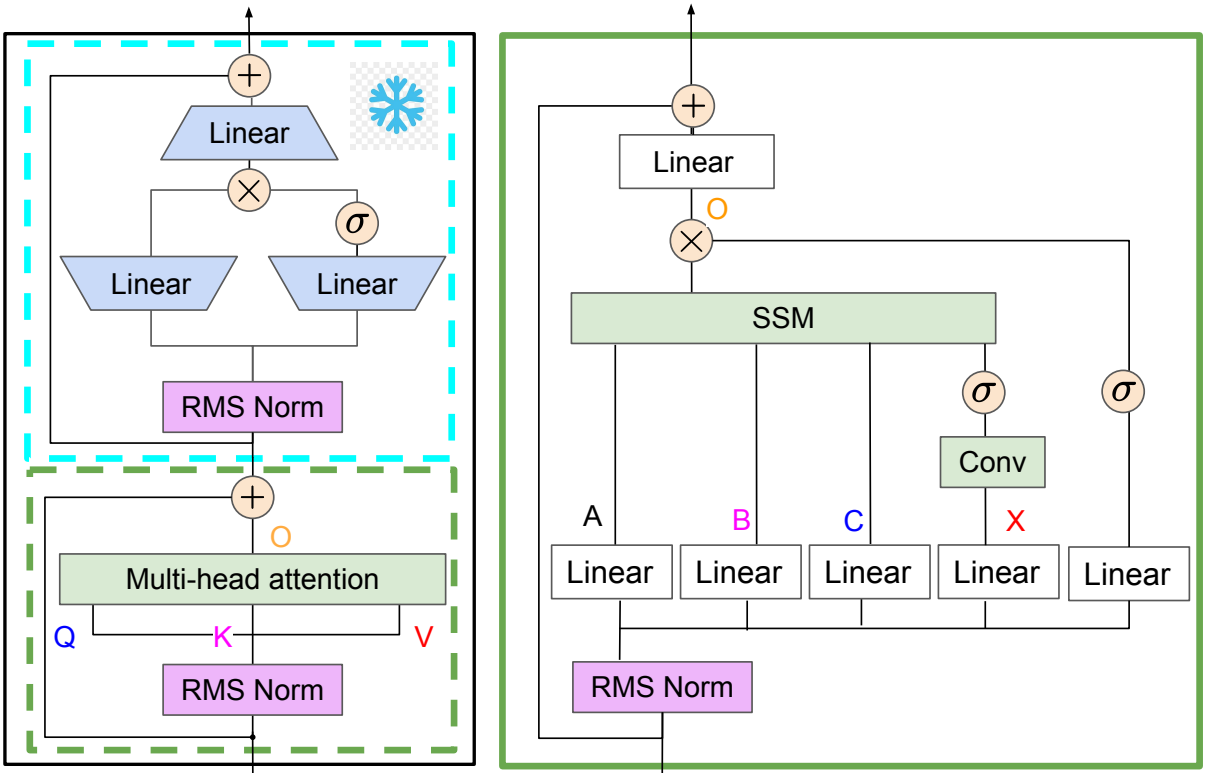

Figure 1: Transferring Transformer to Mamba. Weights in the same color are initialized from the transformer (Linear projections for Q, K, and V are initialized using linear projection for C, B, and X respectively). We replace individual attention blocks with Mamba blocks, and then finetune Mamba blocks while freezing the MLP blocks. Shapes are kept mainly the same. New parameters are introduced for the learned $\mathbf{A}$ and Δ parameters.

Derivations from softmax-based attention to linear RNNs:

Attention is computed in parallel for multiple differently parameterized heads $h \in \{1 \dots H\}$. Each head takes sequence $\boldsymbol{o}$ with hidden size D as an argument and computes,

$$\mathbf{Q}_t = \mathbf{W}^Q \boldsymbol{o}_t, \quad \mathbf{K}_t = \mathbf{W}^K \boldsymbol{o}_t, \quad \mathbf{V}_t = \mathbf{W}^V \boldsymbol{o}_t \quad \text{for all } t,$$

$$\alpha_1 \dots \alpha_T = \text{softmax}\Big(\frac{[m_{1,t}\mathbf{Q}_t^\top \mathbf{K}_1 \dots m_{T,t}\mathbf{Q}_t^\top \mathbf{K}_T]}{\sqrt{D}}\Big) \qquad \boldsymbol{y}_t = \sum_{s=1}^{t} \alpha_s \mathbf{V}_s$$

$$\text{where } \boldsymbol{o}_t \in \mathbb{R}^{D\times 1}, \quad \mathbf{W} \in \mathbb{R}^{N\times D} \quad \mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t \in \mathbb{R}^{N\times 1} \quad m_{s,t} = \mathbf{1}(s \le t)$$

Despite the superficially different form, there is a natural relationship between linear RNNs and attention. Linearizing the attention formula by removing the softmax yields:

$$\boldsymbol{y}_t = \sum_{s=1}^{t} m_{s,t}\alpha_s \mathbf{V}_s = \frac{1}{\sqrt{D}}\mathbf{Q}_t \sum_{s=1}^{t}(m_{s,t}\mathbf{K}_s^\top \mathbf{V}_s) = \frac{1}{\sqrt{D}}\mathbf{Q}_t \sum_{s=1}^{t} m_{s,t}\mathbf{K}_s^\top \mathbf{W}^V \boldsymbol{o}_s.$$

On the other side, expanding the recursive form of Mamba yields,

$$\boldsymbol{h}_t = \mathbf{A}_t\mathbf{A}_{t-1}\cdots\mathbf{A}_2\mathbf{B}_1\boldsymbol{x}_1 + \mathbf{A}_t\mathbf{A}_{t-1}\cdots\mathbf{A}_2\mathbf{B}_2\boldsymbol{x}_2 + \cdots + \mathbf{B}_t\boldsymbol{x}_t = \sum_{s=1}^{t} \mathbf{A}^\times_{t:s}\mathbf{B}_s\boldsymbol{x}_s$$

$$\boldsymbol{y}_t = \mathbf{C}_t \sum_{s=1}^{t} \mathbf{A}^\times_{t:s}\mathbf{B}_s\boldsymbol{x}_s$$

where $A^\times_{t:s} = \mathbf{A}_t\mathbf{A}_{t-1}\cdots\mathbf{A}_s$

If the two models have the same heads H and the head size N, we can set $\mathbf{B}_t = \mathbf{K}_t$, $\mathbf{C}_t = \mathbf{Q}_t$, $\boldsymbol{x}_t = \mathbf{W}^V\boldsymbol{o}_t$. This relationship motivates moving between attention and linear RNN representations.