A Performer architecture details

We define the Performer architecture formally as follows. \( X^{(\text{out})} = X^{(s)}W^{(\text{out})} + b^{(\text{out})} \) and for each \( 1 \leq r \leq s \):

\[
H^{(r-1)} = \text{LN}(\text{MultiHead-Att}(X^{(r-1)})) + X^{(r-1)},
\]

\[
X^{(r)} = \text{LN}(\text{FFN}(H^{(r-1)})) + H^{(r-1)}, \text{ where }
\]

\[
\text{MultiHead-Att}(X) = [H^{(1)} \ldots H^{(k)}],
\]

\[
\forall j \leq k : H^{(j)} = \text{Att}(\overline{X}W^{(j)}_Q, \overline{X}W^{(j)}_K, \overline{X}W^{(j)}_V),
\]

\[
\text{FFN}(H) = \text{GeLU}(HHW^{(1)} + b^{(1)})W^{(2)} + b^{(2)}.\]

Here \( k \) is the number of attention heads \((d_{\text{model}} = kd)\). \( W^{(\text{out})} \in \mathbb{R}^{d_{\text{model}} \times |\Sigma|} \), \( b^{(\text{out})} \in \mathbb{R}^{1 \times |\Sigma|} \), \( W^{(1)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}} \), \( b^{(1)} \in \mathbb{R}^{1 \times d_{\text{ff}}} \), \( W^{(2)} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}} \), \( b^{(2)} \in \mathbb{R}^{1 \times d_{\text{model}}} \), \( W_Q^{(j)}, W_K^{(j)}, W_V^{(j)} \in \mathbb{R}^{d_{\text{model}} \times d} \) are trainable parameters (separate for each instance of MultiHead-Att, FFN), “+” is broadcasted rowwise when biases are added and LN is layer normalization \([2]\), which is applied rowwise and depends on additional trainable parameters. GeLU denotes Gaussian error Linear Unit \([13]\), which is applied elementwise.

B Derivation of Gradient Expressions

\( \theta^{(n)} \) doesn’t affect terms \( \mathcal{L}^{(1)}(X^{(\text{out},1)}), \ldots, \mathcal{L}^{(n-1)}(X^{(\text{out},n)}) \), so corresponding gradients are zero:

\[
\nabla_{\theta^{(n)}} \mathcal{L} = \sum_{n'=n}^{N} \nabla_{\theta^{(n')}} \mathcal{L}^{(n'}(X^{(\text{out},n')}),
\]

Similarly, \( \mathcal{U}^{(n)} \) does not affect \( \mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(n)} \), so

\[
\mathcal{G}^{(n)} = \nabla_{\mathcal{U}^{(n)}} \mathcal{L} = \sum_{n'=n+1}^{N} \nabla_{\mathcal{U}^{(n')}} \mathcal{L}^{(n'}(X^{(\text{out},n')}),
\]

In particular,

\[
\mathcal{G}^{(N)} = \nabla_{\mathcal{U}^{(N)}} \mathcal{L} = 0_{r \times D_4}.
\]

For all \( 1 \leq n < n' \leq N \), \( \theta^{(n)} \) and \( \mathcal{U}^{(n-1)} \) affect \( \mathcal{L}^{(n')} \) only through \( \mathcal{U}^{(n)} \), so according to the chain rule

\[
\nabla_{\theta^{(n)}} \sum_{n'=n+1}^{N} \mathcal{L}^{(n')}(X^{(\text{out},n')}) = \sum_{r=1}^{s} \frac{\partial \mathcal{U}^{(n)}(\theta^{(n)})}{\partial \theta^{(n)}} \nabla_{\mathcal{U}^{(n)}} \sum_{n'=n+1}^{N} \mathcal{L}^{(n')}(X^{(\text{out},n')})
\]

\[
= \sum_{r=1}^{s} \frac{\partial \mathcal{U}^{(n)}(\theta^{(n)})}{\partial \theta^{(n)}} \nabla_{\mathcal{U}^{(n)}} \mathcal{L},
\]

\[
\forall 1 \leq r' \leq s : \nabla_{\mathcal{U}^{(n-1)}_{r'}} \sum_{n'=n+1}^{N} \mathcal{L}^{(n')}(X^{(\text{out},n')}) = \sum_{r=1}^{s} \frac{\partial \mathcal{U}^{(n)}(\mathcal{U}^{(n-1)}_{r'})}{\partial \mathcal{U}^{(n-1)}_{r'}} \nabla_{\mathcal{U}^{(n)}} \sum_{n'=n+1}^{N} \mathcal{L}^{(n')}(X^{(\text{out},n')})
\]

\[
= \sum_{r=1}^{s} \frac{\partial \mathcal{U}^{(n)}(\theta^{(n)})}{\partial \mathcal{U}^{(n-1)}_{r'}} \nabla_{\mathcal{U}^{(n)}} \mathcal{L},
\]

where \( \frac{\partial}{\partial \square} \) denotes Jacobian matrices. Further, for all \( 1 \leq r \leq s \):

\[
\frac{\partial \mathcal{U}^{(n)}(\theta^{(n)})}{\partial \square} \nabla_{\mathcal{U}^{(n)}} \mathcal{L} = \nabla_{\square} \left( \mathcal{U}^{(n)}(\theta^{(n)}) \nabla_{\mathcal{U}^{(n)}} \mathcal{L} \right),
\]

where \( \square \in \{ \theta^{(n)} \cup \{ \mathcal{U}^{(n-1)}_{r'} \}_{1 \leq r' \leq s}, \langle \rangle \} \) denotes a stop-gradient operator, i.e. gradients are not propagated inside brackets and the argument is considered as constant.
We conclude that

\[ \nabla_{\theta(n)} \mathcal{L} = \nabla_{\theta(n)} \mathcal{L}^{(n)}(X^{(out,n)}) + \nabla_{\theta(n)} \sum_{n' = n+1}^{N} \mathcal{L}^{(n')}(X^{(out,n')}) = \nabla_{\theta(n)} \mathcal{L}^{(n)}(X^{(out,n)}) + \sum_{r=1}^{s} \partial U_r^{(n)} \top \times \nabla_{U_r} \mathcal{L} \]

\[ = \nabla_{\theta(n)} \left( \mathcal{L}^{(n)}(X^{(out,n)}) + \sum_{r=1}^{s} [U_r^{(n)}] \top \langle \langle \nabla_{U_r} \mathcal{L} \rangle \rangle \right) = \nabla_{\theta(n)} \Phi^{(n)}(\theta^{(n)}, U^{(n-1)}, \nabla_{U} \mathcal{L}) \]

\[ = \nabla_{\theta(n)} \Phi^{(n)}(\theta^{(n)}, U^{(n-1)}, g^{(n)}), \]

\[ \forall 1 \leq r' \leq s : \mathcal{G}_{r'}^{(n-1)} = \nabla_{U_{r'}^{(n-1)}} \mathcal{L} = \nabla_{U_{r'}^{(n-1)}} \mathcal{L}^{(n)}(X^{(out,n)}) + \nabla_{U_{r'}^{(n-1)}} \sum_{n' = n+1}^{N} \mathcal{L}^{(n')}(X^{(out,n')}) \]

\[ = \nabla_{U_{r'}^{(n-1)}} \mathcal{L}^{(n)}(X^{(out,n)}) + \sum_{r=1}^{s} \nabla_{U_{r'}^{(n-1)}} [U_r^{(n)}] \top \langle \langle \nabla_{U_r} \mathcal{L} \rangle \rangle \]

\[ = \nabla_{U_{r'}^{(n-1)}} \mathcal{L}^{(n)}(X^{(out,n)}) + \nabla_{U_{r'}^{(n-1)}} \Phi^{(n)}(\theta^{(n)}, U^{(n-1)}, \nabla_{U} \mathcal{L}) \]

\[ = \nabla_{U_{r'}^{(n-1)}} \Phi^{(n)}(\theta^{(n)}, U^{(n-1)}, g^{(n)}), \]

where the second chain of equalities is equivalent to (10).

C Efficient “Block” Computation of (23)

Denote \( \bar{Q} = (g(Q_i))_{i=1}^{L}, \bar{K} = (g(K_i))_{i=1}^{L}, N = (R_i \times Q_i)_{i=1}^{L}, D = (S_i^{\top} \bar{Q}_i)_{i=1}^{L} \). [17] propose the following algorithm for computation of (23). Initialize buffers \( \text{curR} = 0_{d \times M}, \text{curS} = 0_{M} \), iterate over \( l = 1, \ldots, L \) and compute

\[ \text{curR} := \text{curR} + V_l \times \bar{K}_l; \]
\[ \text{curS} := \text{curS} + \bar{K}_l; \]
\[ N_l := \text{curR} \times \bar{Q}_l; \]
\[ D_l := \text{curS}^{\top} \times \bar{Q}_l; \]
\[ Y_l := N_l/D_l. \]

This way, the 3D tensor \( R \in \mathbb{R}^{L \times d \times M} \) is not stored in memory explicitly, resulting in \( O(L) \) time and \( O(L(d + M) + d M) \) memory complexity. In order to have the same memory consumption during back-propagation, [17] propose the following routine. Keep buffers \( \text{curR}, \text{curS} \) as the result of forward pass, and initialize gradient buffers \( \text{gradR} = 0_{d \times M}, \text{gradS} = 0_{M} \). Assuming that \( \nabla_N \mathcal{L} \in \mathbb{R}^{L \times d}, \nabla_D \mathcal{L} \in \mathbb{R}^{L} \) are computed using automatic differentiation, iterate in a backward direction \( l = L, \ldots, 1 \) and compute

\[ \nabla_{\bar{Q}_l} \mathcal{L} := (\nabla_{D_l} \mathcal{L}) \cdot \text{curS} + \text{curR}^{\top} \times \nabla_{N_l, \mathcal{L}} \]
\[ \text{curR} := \text{curR} - V_l \times \bar{K}_l; \]
\[ \text{curS} := \text{curS} - \bar{K}_l; \]
\[ \text{gradR} := \text{gradR} + (\nabla_{N_l} \mathcal{L}) \times \bar{Q}_l; \]
\[ \text{gradS} := \text{gradS} + (\nabla_{D_l} \mathcal{L}) \cdot \bar{Q}_l; \]
\[ \nabla_{V_l} \mathcal{L} := \text{gradR} \times \bar{K}_l; \]
\[ \nabla_{\bar{K}_l} \mathcal{L} := \text{gradS} \times V_l. \]
In practice, the described algorithm works slowly when implemented in pure PyTorch, because \( l \) is iterated one-by-one. [17] use low-level CUDA extensions to make the algorithm practical. Instead, we propose a “block” version, when we iterate through blocks of \( l \) of a small size \( C \) (we use \( C = 64 \)).

In each block we use explicit prefix sums on inputs of length \( C \) to find \( Y_{l:t+c−1} \), using the maintained front \( \text{curR}, \text{curS} \). The formal algorithm is as follows. Initialize buffers \( \text{curR} = 0_{d \times M}, \text{curS} = 0_{M} \). For simplicity assuming that \( C \) divides \( L \) (extension for an opposite case is straightforward), iterate over \( l = 1, C + 1, \ldots, L − C + 1 \) and compute

\[
\begin{align*}
\text{blockR} &:= \text{PS}((V_{t+t′−1} \times \mathbf{K}_{/l−1}^T l′=1)); \\
\text{blockR} &:= (\text{curR} + \text{blockR} l′)^{C}_{l′=1}; \\
\text{blockS} &:= \text{PS}((\mathbf{K}_{/l−1}^T l′=1)); \\
\text{blockS} &:= (\text{curS} + \text{blockS} l′)^{C}_{l′=1}; \\
\text{curR} &:= \text{blockR}_{C}; \\
\text{curS} &:= \text{blockS}_{C}; \\
N_{l:t+c−1} &:= (\text{blockR} l′ \times \mathbf{Q}_{/l−1})^{C}_{l′=1}; \\
D_{l:t+c−1} &:= (\text{blockS} l′ \times \mathbf{Q}_{/l−1})^{C}_{l′=1}; \\
Y_{l:t+c−1} &:= (N_{l+t′−1}/D_{l+t′−1})^{C}_{l′=1}.
\end{align*}
\]

In the “block” version, the number of outer sequential iterations is reduced to \( L/C \), resulting in \( O((L/C) \log C) \) parallel time complexity, when the logarithmic parallel algorithm is used to compute prefix sums [16,17]. In our experiments, we use \texttt{torch} \_\texttt{cunsum} to compute [16,17], which works fast in practice. The memory complexity of the algorithm is \( O(L(d + M) + CdM) \), where the second term is for storing \( \text{blockR} \). Assuming that \( C \) is a small constant (\( C = O(1) \)), we conclude that the “block” version has \( O(L(d + M) + dM) \) memory and \( O(L) \) time complexity — same as the algorithm of [17]. As for hidden constants in complexity estimates, the constant inside \( O(L) \) time complexity is reduced at the cost of increasing constant of the “small” \( dM \) term in the memory complexity (when \( d, M \ll L \)), making the “block” iterative algorithm a practical choice for computing (1).

We further show how to back-propagate through (2,3) in \( O((L/C) \log C) \) time and \( O(L(d + M) + CdM) \) memory. Again, keep buffers \( \text{curR}, \text{curS} \) as the result of forward pass, and initialize gradient buffers \( \text{gradR} = 0_{d \times M}, \text{gradS} = 0_{M} \). Assuming that \( \nabla_{\mathbf{N}} \mathcal{L} \in \mathbb{R}^{L \times d}, \nabla_{\mathbf{D}} \mathcal{L} \in \mathbb{R}^{L} \) are computed using automatic differentiation, iterate in a backward direction \( l = L − C + 1, L − 2C + 1, \ldots, 1 \) and compute

\[
\begin{align*}
\text{curR} &:= \text{curR} − \sum_{l′=l}^{l+C−1} V_{l′} \times \mathbf{K}_{l′}^T; \\
\text{curS} &:= \text{curS} − \sum_{l′=l}^{l+C−1} \mathbf{K}_{l′}; \\
\text{blockR} &:= \text{PS}((V_{l+t′−1} \times \mathbf{K}_{/l−1}^T l′=1)); \\
\text{blockR} &:= (\text{curR} + \text{blockR} l′)^{C}_{l′=1}; \\
\text{blockS} &:= \text{PS}((\mathbf{K}_{/l−1}^T l′=1)); \\
\text{blockS} &:= (\text{curS} + \text{blockS} l′)^{C}_{l′=1}; \\
\nabla_{\mathbf{Q}_{/l−1}} \mathcal{L} &:= ((\nabla_{\mathbf{D}_{l+t′−1}} \mathcal{L}) \cdot \text{blockS} l′ + \text{curR} l′ \times \nabla_{\mathbf{N}_{l+t′−1}} \mathcal{L})^{C}_{l′=1}; \\
\text{gradR} &:= \text{gradR} + \sum_{l′=l}^{l+C−1} (\nabla_{\mathbf{N}_{l′}} \mathcal{L}) \times \mathbf{Q}_{l′}^T; \\
\text{gradS} &:= \text{gradS} + \sum_{l′=l}^{l+C−1} (\nabla_{\mathbf{D}_{l′}} \mathcal{L}) \cdot \mathbf{Q}_{l′}; \\
\text{blockgradR} &:= \text{PS}(((\nabla_{\mathbf{N}_{l+t′−1}} \mathcal{L}) \times \mathbf{Q}_{/l−1}^T l′=1)^{C}_{l′=1}); \\
\text{blockgradR} &:= (\text{gradR} − \text{blockgradR} l′)^{C}_{l′=1};
\end{align*}
\]
\begin{align*}
\text{blockgrad} S := & \text{PS}((\nabla D \cdot \tilde{Q}_{l+I-1}) \cdot (L_{l+I-1})_{l=1}^C) \\
\text{blockgrad} S := & (\text{grad} S - \text{grad} S_{l+I-1})_{l=1}^C \\
\nabla v_{l+I-1} := & (\text{blockgrad} R_{l+I-1} \times \tilde{K}_{l+I-1})_{l=1}^C \\
\nabla K_{l+I-1} := & (\text{blockgrad} R_{l+I-1}^T \times V_{l+I-1})_{l=1}^C.
\end{align*}

Finally, it’s easy to see how to use both one-to-one and “block” iterative computation as part of Algorithm 1 to compute the update (7-8). For that, when doing a forward computation for some \( n, r \), initialize \( \text{cur} R, \text{cur} S \) from corresponding subvectors of \( U(r-1,n-1) \), with the rest of the algorithm unchanged. Similarly, during a backward pass for some \( n, r \), initialize \( \text{grad} R, \text{grad} S \) from corresponding subvectors of \( G^{(n)} \) and leave the rest of the iterative back-propagation algorithm unchanged.

## D Additional Experimental Details

We use 200K, 100K, 200K SGD iterations in the Copying Task, Penn Treebank, Enwik8 setups respectively. We use Adam optimizer \([18]\) with \( \beta_1 = 0.9, \beta_2 = 0.999 \) (default configuration used in PyTorch). For the Copying task, we train with a learning rate \( 10^{-3} \) for 130K iterations and then decrease the learning rate to \( 10^{-4} \). We use a fixed learning rate of \( 10^{-4} \) and \( 2 \times 10^{-4} \) in Penn Treebank and Enwik8 experiments respectively.

Figure 5 is a version of Figure 2 for the configuration II. We draw the same conclusions to those reported in the main text. Figure 6 is a bigger version of Figure 4 from the main text. Figure 7 reports additional experimental results: Bits Per Character for the Copying Task and train set learning curves for Penn Treebank and Enwik8. Figure 8 is a version of Figure 4 showing a difference (–) between curves and the “Full” curve. We observe that memory-efficient algorithms result in a negligible loss of performance, which we attribute to numerical effects, accumulating with many thousands of iterations.
Figure 6: Bigger version of Figure 4.
Figure 7: Bits-per-character learning curve for the Copying task and train-set learning curves for language modelling on Penn Treebank and Enwik8 respectively.
Figure 8: A difference (−) between curves and the “Full” curve from Figure 4.