Attention Residuals: A Memoir

Jianlin Su

2026-03-19

This article introduces our latest work, Attention Residuals (AttnRes); as the name suggests, it uses the idea of Attention to improve Residuals.

Many readers will have heard of the Pre Norm / Post Norm debate, but ultimately this is just "infighting" within Residuals itself, and many later variants of Normalization are the same. A more interesting change is HC, which began down the route of expanding the residual stream, but perhaps because of unstable results it did not attract much response. The later story is probably known to everyone: at the end of last year DeepSeek's mHC improved HC and verified its effectiveness on a larger scale.

Rather than further expanding the residual stream, we chose another radical route: directly performing Attention between layers to replace Residuals. Of course, getting the full pipeline to work involved many details and efforts; here we simply recall the related journey.

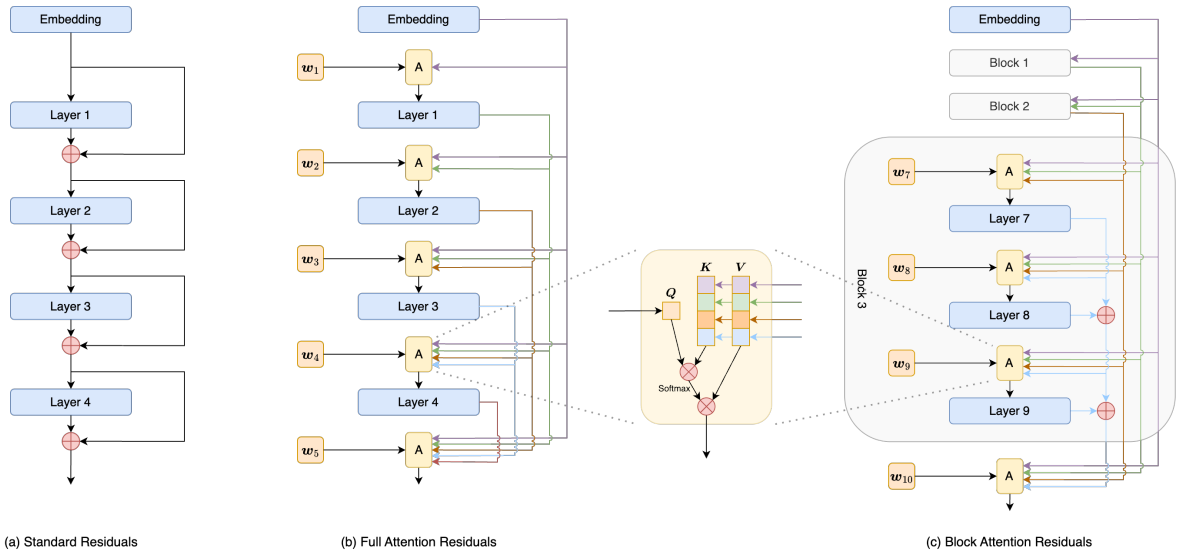

Figure 1: AttnRes schematic diagram

1 Inter-layer Attention

As usual, we start from Residuals, which everyone knows by heart. Its form is

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \boldsymbol{f}_t(\boldsymbol{x}_{t-1}) \tag{1}$$

Here we switch to another notation that lets us see something deeper. Let $\boldsymbol{y}_t = \boldsymbol{f}_t(\boldsymbol{x}_{t-1})$, then $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \boldsymbol{y}_t$, with the convention $\boldsymbol{y}_0 = \boldsymbol{x}_0$, so we easily obtain $\boldsymbol{x}_t = \boldsymbol{y}_0 + \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_t$, and it

can equivalently be written as

$$\boldsymbol{y}_{t+1} = \boldsymbol{f}_{t+1}(\boldsymbol{y}_0 + \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_t) \tag{2}$$

That is, from the $\boldsymbol{y}$ perspective, Residuals takes the equal-weight sum of $\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_t$ as the input to $\boldsymbol{f}_{t+1}$ to obtain $\boldsymbol{y}_{t+1}$. A natural generalization is to replace it with a weighted sum:

$$\boldsymbol{y}_{t+1} = \boldsymbol{f}_{t+1}\left(\sum_{s=0}^{t} a_{t+1,s}\boldsymbol{y}_s\right) \qquad \text{where} \qquad a_{t,s} \geq 0, \quad \sum_{s=0}^{t} a_{t+1,s} = 1 \tag{3}$$

This is the germ of AttnRes. The above formula adds two extra constraints on $a_{t,s}$; let us discuss their necessity:

1. The constraint $a_{t,s} \geq 0$ ensures that the same $\boldsymbol{y}_s$ always contributes in the same direction to different layers, avoiding the inconsistency where one layer wants to increase $\boldsymbol{y}_s$ while another wants to shrink it, which is intuitively more friendly to model learning;
2. The $\boldsymbol{f}$ we use includes In Norm and will first apply RMSNorm to the input. Since $\text{RMSNorm}(\boldsymbol{x}) = \text{RMSNorm}(c\boldsymbol{x})$ holds identically for all $c > 0$, weighted averaging and weighted summation are completely equivalent, so the constraint $\sum_{s=0}^{t} a_{t,s} = 1$ does not reduce expressiveness.

2 Hyper-Connections

Before launching into AttnRes, let us briefly review HC (Hyper-Connections) and show that it too can be understood as inter-layer Attention, thereby demonstrating that inter-layer Attention is indeed a more fundamental route. HC changes Residuals to

$$\boldsymbol{X}_t = \boldsymbol{H}_t^{res}\boldsymbol{X}_{t-1} + \boldsymbol{H}_t^{post}\boldsymbol{f}_t(\boldsymbol{H}_t^{pre}\boldsymbol{X}_{t-1}) \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{k\times d}, \boldsymbol{H}^{res} \in \mathbb{R}^{k\times k}, \boldsymbol{H}^{pre} \in \mathbb{R}^{1\times k}, \boldsymbol{H}^{post} \in \mathbb{R}^{k\times 1}$, and the classic choice is $k = 4$. Simply put, the state variable is expanded k times; before input to $\boldsymbol{f}_t$ a matrix $\boldsymbol{H}_t^{pre}$ projects it back to one time; after output a matrix $\boldsymbol{H}_t^{post}$ expands it back to k times, and finally it is added to the $\boldsymbol{x}_{t-1}$ adjusted by $\boldsymbol{H}_t^{res}$. If we do not restrict the forms of $\boldsymbol{H}_t^{res}, \boldsymbol{H}_t^{pre}, \boldsymbol{H}_t^{post}$, then Post Norm, Highway, and others are all special cases of HC.

Similarly let $\boldsymbol{y}_t = \boldsymbol{f}_t(\boldsymbol{H}_t^{pre}\boldsymbol{X}_{t-1})$, then $\boldsymbol{X}_t = \boldsymbol{H}_t^{res}\boldsymbol{X}_{t-1} + \boldsymbol{H}_t^{post}\boldsymbol{y}_t$, with the convention $\boldsymbol{X}_0 = \boldsymbol{H}_0^{post}\boldsymbol{y}_0$, so it can also be expanded as $\boldsymbol{X}_t = \boldsymbol{H}_{t\leftarrow 1}^{res}\boldsymbol{H}_0^{post}\boldsymbol{y}_0 + \boldsymbol{H}_{t\leftarrow 2}^{res}\boldsymbol{H}_1^{post}\boldsymbol{y}_1 + \cdots + \boldsymbol{H}_{t\leftarrow t}^{res}\boldsymbol{H}_{t-1}^{post}\boldsymbol{y}_{t-1} + \boldsymbol{H}_t^{post}\boldsymbol{y}_t$, where $\boldsymbol{H}_{t\leftarrow s}^{res}$ is defined as $\boldsymbol{H}_t^{res}\boldsymbol{H}_{t-1}^{res}\cdots\boldsymbol{H}_{s+1}^{res}\boldsymbol{H}_s^{res}$. Further adopting the convention $\boldsymbol{H}_{t\leftarrow t+1}^{res} = \boldsymbol{I}$, we can write

$$\boldsymbol{y}_{t+1} = \boldsymbol{f}_{t+1}(\boldsymbol{H}_{t+1}^{pre}\boldsymbol{x}_t) = \boldsymbol{f}_{t+1}\Bigg(\sum_{s=0}^{t} \underbrace{\boldsymbol{H}_{t+1}^{pre}\boldsymbol{H}_{t\leftarrow s+1}^{res}\boldsymbol{H}_s^{post}}_{a_{t+1,s}}\boldsymbol{y}_s\Bigg) \tag{5}$$

Note that each $\boldsymbol{H}_{t+1}^{pre}\boldsymbol{H}_{t\leftarrow s+1}^{res}\boldsymbol{H}_s^{post}$ is a 1×1 matrix, equivalent to a scalar, so it too is of the inter-layer Attention form of Eq. (3). Readers familiar with linear attention should quickly understand this result: HC is essentially a DeltaNet "rotated by 90 degrees". In practice, the three $\boldsymbol{H}$ matrices are computed from simple linear layers with tanh activation, which causes the chained product $\boldsymbol{H}_{t\leftarrow s}^{res}$ to risk explosion or collapse, and also cannot guarantee the non-negativity of $a_{t+1,s}$.

Later mHC made improvements: first, all three $\boldsymbol{H}$ were changed to Sigmoid activation to guarantee $a_{t+1,s} \geq 0$; then it alternately normalized $\boldsymbol{H}_t^{res}$ to make it doubly stochastic, and by the closure of doubly stochastic matrices under multiplication it guaranteed the stability of $\boldsymbol{H}_{t\leftarrow s}^{res}$; finally experiments verified the effectiveness of these changes. However, some new experiments such as "Your deepseek mHC may not need 'm"' showed that directly setting $\boldsymbol{H}_t^{res}$ to the identity matrix is good enough.

3 Teamwork

Let us return to AttnRes. After realizing the feasibility of AttnRes, the next question is: what form should $a_{t+1,s}$ take? A very natural idea is to follow the standard Scaled Dot-Product Attention, but at the time the author wanted a quick first try, so chose a simpler form

$$a_{t+1,s} \propto \exp(\boldsymbol{w}_{t+1} \cdot \boldsymbol{y}_s) \tag{6}$$

where $\boldsymbol{w}_t$ is a learnable vector parameter, i.e., directly using a data-independent static vector as Q while K and V are both $\boldsymbol{y}_s$ to perform Softmax Attention. This is the first version of AttnRes. Surprisingly, with such a simple design the improvement over Residuals is already very significant!

When the author shared the preliminary experimental results of AttnRes within the group, @Zhang Yu and @Guang Yu showed great interest and joined in to start validation on larger-scale models, and the results were all encouraging. During this period, we also tried some more complex designs and found that most were not as good as this simple version; only adding an extra RMSNorm operation to K could obtain relatively stable gains, which constitutes the final form of AttnRes

$$a_{t+1,s} \propto \exp(\boldsymbol{w}_{t+1} \cdot \mathrm{RMSNorm}(\boldsymbol{y}_s)) \tag{7}$$

However, AttnRes is after all an intensive inter-layer connection scheme. Is training and inference feasible on K2 or even larger scales? Excitingly, @V via an elegant analysis first affirmed the feasibility of inference, and the "finishing touch" was precisely the convenient static-Q design! This allows us to compute the attention $a_{t,s}$ for $t > s$ immediately after computing $\boldsymbol{y}_s$, giving Infra enough room to maneuver.

But unfortunately, the training side, such as @Wang, after careful analysis judged that AttnRes was still not feasible in our current training environment (to put it bluntly, we are still poor), and needed a further scheme to reduce communication and memory; thus came the Block version below; correspondingly, the previous version is called the Full version.

4 Block Version

From Full AttnRes to Block AttnRes is analogous to the past process of linearizing quadratic Attention; various existing Efficient Attention ideas can be tried, such as SWA (Sliding Window Attention), which was our first attempt, but the actual effect was very poor, even worse than Residuals.

After reflection, the author thinks it can be understood this way: Residuals itself is already a very strong baseline, corresponding to the equal-weight sum of all state vectors. Any new design that wants to surpass it must at least formally be able to cover it. The Full AttnRes obviously satisfies this condition, but adding SWA does not, because it discards part of the state and cannot cover the special case of "equal-weight sum of all state vectors".

Thus we realized that for AttnRes, "compression" may be more effective than "sparsity", and the compression does not need to be too fine-grained; simple weighted summation may suffice. After

some conception and polishing, @Zhang Yu and @Guang Yu proposed the Block AttnRes design in the paper, which combines the ideas of block-wise processing and summation compression, achieving results close to the Full version.

The idea of Block AttnRes is roughly this: first, the Embedding layer is a separate Block, because by observing the Full version's Attention matrix (this is the benefit of the Attention concept, you can visualize the attention pattern at any time) we found that the model tends to give the Embedding layer considerable Attention, so it is necessary to keep Embedding independent; then every remaining m layers form a Block, intra-Block compression is done by summation, and inter-Block Attention is computed with the summed result as the unit.

Experiments show that simply fixing to about 8 Blocks can obtain most of AttnRes's benefits. After evaluation, both training and inference colleagues agreed that the extra overhead of Block AttnRes is very small; compared to its effect improvement, it is completely worthwhile (detailed analysis see @Wang and @V; if you want numbers, roughly within 5% overhead for 25% gain), so all colleagues made a full push to get it into the main branch. This was another fulfilling and pleasant experience, which we will not elaborate on.

5 Matrix Perspective

It is worth mentioning that we can also unify Residuals, HC/mHC, Full AttnRes, and Block AttnRes through the Attention matrix; this is also a rather interesting understanding perspective, as shown below. Here $\phi(\boldsymbol{q},\boldsymbol{k}) = \exp(\boldsymbol{q}\cdot\text{RMSNorm}(\boldsymbol{k}))$, the Block AttnRes version corresponds to $m=3$, and $\boldsymbol{y}_{s:t} = \sum_{i=s}^{t}\boldsymbol{y}_i$; this notation was also used in "Making Alchemy More Scientific (Part Four): New Identity, New Learning Rate".

5.1 Residuals

$$\boldsymbol{A} = \begin{pmatrix} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 1 & 1 & & & & \\ 1 & 1 & 1 & 1 & & & \\ 1 & 1 & 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

5.2 HC/mHC

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{H}_1^{pre}\boldsymbol{H}_0^{post} & & & & & & \\ \boldsymbol{H}_2^{pre}\boldsymbol{H}_{1\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_2^{pre}\boldsymbol{H}_1^{post} & & & & & \\ \boldsymbol{H}_3^{pre}\boldsymbol{H}_{2\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_3^{pre}\boldsymbol{H}_{2\leftarrow 2}^{res}\boldsymbol{H}_1^{post} & \boldsymbol{H}_3^{pre}\boldsymbol{H}_2^{post} & & & & \\ \boldsymbol{H}_4^{pre}\boldsymbol{H}_{3\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_4^{pre}\boldsymbol{H}_{3\leftarrow 2}^{res}\boldsymbol{H}_1^{post} & \boldsymbol{H}_4^{pre}\boldsymbol{H}_{3\leftarrow 3}^{res}\boldsymbol{H}_2^{post} & \boldsymbol{H}_4^{pre}\boldsymbol{H}_3^{post} & & & \\ \boldsymbol{H}_5^{pre}\boldsymbol{H}_{4\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_5^{pre}\boldsymbol{H}_{4\leftarrow 2}^{res}\boldsymbol{H}_1^{post} & \boldsymbol{H}_5^{pre}\boldsymbol{H}_{4\leftarrow 3}^{res}\boldsymbol{H}_2^{post} & \boldsymbol{H}_5^{pre}\boldsymbol{H}_{4\leftarrow 4}^{res}\boldsymbol{H}_3^{post} & \boldsymbol{H}_5^{pre}\boldsymbol{H}_4^{post} & & \\ \boldsymbol{H}_6^{pre}\boldsymbol{H}_{5\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_6^{pre}\boldsymbol{H}_{5\leftarrow 2}^{res}\boldsymbol{H}_1^{post} & \boldsymbol{H}_6^{pre}\boldsymbol{H}_{5\leftarrow 3}^{res}\boldsymbol{H}_2^{post} & \boldsymbol{H}_6^{pre}\boldsymbol{H}_{5\leftarrow 4}^{res}\boldsymbol{H}_3^{post} & \boldsymbol{H}_6^{pre}\boldsymbol{H}_{5\leftarrow 4}^{res}\boldsymbol{H}_4^{post} & \boldsymbol{H}_6^{pre}\boldsymbol{H} & \\ \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow 1}^{res}\boldsymbol{H}_0^{post} & \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow 2}^{res}\boldsymbol{H}_1^{post} & \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow 3}^{res}\boldsymbol{H}_2^{post} & \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow 4}^{res}\boldsymbol{H}_3^{post} & \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow 5}^{res}\boldsymbol{H}_4^{post} & \boldsymbol{H}_7^{pre}\boldsymbol{H}_{6\leftarrow} & \end{pmatrix}$$

5.3 Full AttnRes

$$\boldsymbol{A} = \begin{pmatrix} \phi(\boldsymbol{w}_1, \boldsymbol{y}_0) & & & & & & & \\ \phi(\boldsymbol{w}_2, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_2, \boldsymbol{y}_1) & & & & & & \\ \phi(\boldsymbol{w}_3, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_3, \boldsymbol{y}_1) & \phi(\boldsymbol{w}_3, \boldsymbol{y}_2) & & & & & \\ \phi(\boldsymbol{w}_4, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_1) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_2) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_3) & & & & \\ \phi(\boldsymbol{w}_5, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_1) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_2) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_3) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_4) & & & \\ \phi(\boldsymbol{w}_6, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_1) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_2) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_3) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_4) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_5) & & \\ \phi(\boldsymbol{w}_7, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_1) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_2) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_3) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_4) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_5) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_6) & \end{pmatrix}$$

5.4 Block AttnRes

$$\boldsymbol{A} = \left(\begin{array}{c|ccc|ccc} \phi(\boldsymbol{w}_1, \boldsymbol{y}_0) & & & & & & \\ \hline \phi(\boldsymbol{w}_2, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_2, \boldsymbol{y}_1) & & & & & \\ \phi(\boldsymbol{w}_3, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_3, \boldsymbol{y}_{1:2}) & \phi(\boldsymbol{w}_3, \boldsymbol{y}_{1:2}) & & & & \\ \phi(\boldsymbol{w}_4, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_4, \boldsymbol{y}_{1:3}) & & & \\ \hline \phi(\boldsymbol{w}_5, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_5, \boldsymbol{y}_4) & & \\ \phi(\boldsymbol{w}_6, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_{4:5}) & \phi(\boldsymbol{w}_6, \boldsymbol{y}_{4:5}) & \\ \phi(\boldsymbol{w}_7, \boldsymbol{y}_0) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{1:3}) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{4:6}) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{4:6}) & \phi(\boldsymbol{w}_7, \boldsymbol{y}_{4:6}) \end{array}\right)$$

6 Related Work

From the time we planned to do AttnRes, the author and many friends have been immersed in polishing, verifying, and accelerating. Some readers may know that the author's research style is to first do his best to deduce and solve, until encountering difficulties or completely solving the problem before looking for related literature. It happens that the author met a group of similar friends; it happens that this AttnRes exploration went relatively smoothly overall, so we only started to survey related literature when all tests were basically passed and we began to prepare the technical report.

But also because of this, "not checking is fine, checking is shocking": relevant work on Dense Connection and Depth Attention is already very extensive. In addition to the classic DenseNet, we also found DenseFormer, ANCRe, MUDDFormer, MRLA, Dreamer, etc.; even ELMo from before BERT partially applied a similar design. All of these are included in our references.

After the technical report was released, we gradually received comments from some readers pointing out some related works not yet included, such as SKNets, LIMe, DCA, etc. For this we apologize and thank them, and promise to try to supplement them in subsequent revisions. But whether readers or the authors themselves, please remain rational about this: literature review is not easy, some omissions are inevitable, and we pay high respect to all related works.

At the same time, we also call on everyone to pay more attention to the workload of AttnRes beyond the "Depth Attention" concept. We very much agree that today in 2026, "Depth Attention" or "Layer Attention" is an idea devoid of novelty, but how to use it for a model large enough, as a strong enough substitute for Residuals, while also meeting training and inference efficiency requirements, is not an easy task. To our knowledge, AttnRes is the first work to achieve this.

7 Summary

This article introduced our latest results on model architecture, Attention Residuals (AttnRes). It uses inter-layer Attention to replace plain Residuals, and through careful design makes it satisfy training and inference efficiency requirements, ultimately successfully scaling it to a sufficiently large model.