

VALUE RESIDUAL LEARNING FOR ALLEVIATING ATTENTION CONCENTRATION IN TRANSFORMERS

Zhanchao Zhou^{1,2} Tianyi Wu^{3†*} Zhiyun Jiang^{4†*} Zhenzhong Lan^{2,5◊}

¹Zhejiang University ²Westlake University

³University of Electronic Science and Technology of China

⁴China University of Mining and Technology

⁵Research Center for Industries of the Future, Westlake University

Abstract

Transformers can capture long-range dependencies using self-attention, allowing tokens to attend to all others directly. However, stacking multiple attention layers leads to attention concentration. One natural way to address this issue is to use cross-layer attention, allowing information from earlier layers to be directly accessible to later layers. However, this approach is computationally expensive. To address this problem, we propose Transformer with residual value (ResFormer) which approximates cross-layer attention through adding a residual connection from the values of the the first layer to all subsequent layers. Based on this method, one variant is the Transformer with single layer value (SVFormer), where all layers share the same value embedding from first layer, reducing the KV cache by nearly 50%. Comprehensive empirical evidence demonstrates that ResFormer mitigates attention concentration problem in deeper layers and enhances representation across most layers, outperforming the vanilla Transformer, DenseFormer, and NeuTRENO in training error as well as downstream tasks. SVFormer trains significantly faster than the vanilla Transformer and performs better than other methods like GQA and CLA, with performance influenced by sequence length and cumulative learning rate.¹

1 INTRODUCTION

The Transformer (Vaswani, 2017) model has become one of the leading architectures in recent years, excelling in both language modeling (Devlin, 2018; Lan, 2019; Brown, 2020) and computer vision tasks (Dosovitskiy, 2020). The discovery of scaling laws (Hoffmann et al., 2022; Kaplan et al., 2020) has driven the pursuit of larger Transformer models by increasing network depth and width.

Training large models presents significant challenges. Balancing the depth and width of a Transformer model within a fixed parameter budget is particularly difficult. While research indicates that deeper models generalize more compositionally than shallower ones (Petty et al., 2024), the training and deployment of deep models remain problematic. Although Transformers use residual connections (He et al., 2016) to address the vanishing gradient issue, training very deep Transformers is still challenging. For example, a 32-layer Vision Transformer (ViT) may perform worse than a 24-layer one (Zhou et al., 2021). This is mainly due to the smoothing mechanism of attention (Shi et al., 2022), which can lead to an over-smoothing effect (Nguyen et al., 2023) where the token representations become the same as the model’s depth increases.

Existing solutions to alleviate the over-smoothing problem in Transformer include adding extra regularizers (Nguyen et al., 2023; Shi et al., 2022) and optimizing the information flow within the model (Pagliardini et al., 2024). During the era of convolutional neural network architectures, Stochastic Depth (Huang et al., 2016) reduces the likelihood of over-smoothing by randomly dropping layers during training and DenseNet (Huang et al., 2017) mitigates the impact of over-smoothing by allowing each layer to directly access the hidden states of all preceding layers. Recently, DenseFormer

*Equal Contribution; †Work done during internship at Westlake University; ◊ Corresponding author.

¹Code is available at <https://github.com/Zcchill/Value-Residual-Learning>.

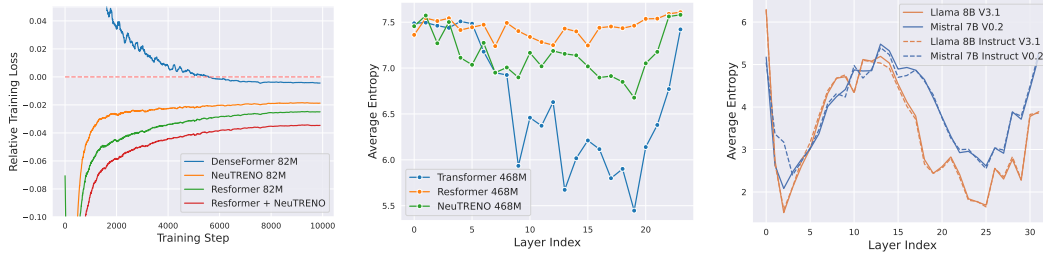


Figure 1: (Left) Illustration of the relative training loss (loss of target model - loss of vanilla Transformer) curve between different Transformer variants; model size is fixed to be 82M. (Middle) The average entropy of token importance across layers in ResFormer *vs.* the vanilla Transformer, where token importance is derived from the attention matrix. Lower entropy indicates more focused attention on specific tokens. More details can be found in Eqn. 11. (Right) The average entropy of token importance across layers in Llama (8B) (Dubey et al., 2024) and Mistral (7B) (Jiang et al., 2023).

(Pagliardini et al., 2024) adopts the idea of DenseNet when training Transformer. Additionally, NeuTRENO (Nguyen et al., 2023) alleviates over-smoothing through incorporating the difference between the value vectors of the first layer and the current layer to the attention output.

In this paper, we address the problem of multi-layer attention from a novel perspective. We introduce the phenomenon of attention concentration, which describes how a model’s attention increasingly focuses on fewer tokens as the network depth increases, as illustrated in Fig. 1 (Right). We quantify the degree of attention concentration using the entropy of the distribution of token importance, where lower entropy indicates a more pronounced concentration. In a deep model, the attention mechanism exhibits a pattern of “concentration - dispersion - concentration,” where the model could potentially lose useful information during the concentrated phase. See Fig. 15 for analysis of over-smoothing.

To alleviate the attention concentration from stacking multiple attention layers, an effective method is to use cross-layer attention on information from earlier layers. Given the high computational cost of cross-layer attention, we propose ResFormer as an approximation. ResFormer achieves a similar effect by applying a residual connection between the value vectors of the current layer and the first layer before the attention operation. Experimental results show that ResFormer mitigates the attention concentration effect in deeper layers, maintaining attention dispersion throughout the sequence, as shown in Fig. 1 (Middle).

During inference, deep networks require substantial *KV* cache, severely impacting model deployment (Xiao et al., 2023). Existing *KV*-efficient methods often process keys and values simultaneously. Building on ResFormer, we decouple the value from the attention operation and propose a new kind of Transformer with single layer value (SVFormer). In SVFormer, the queries and keys of all layers share the value from the first layer. Consequently, SVFormer can save nearly half of the *KV* cache. Additionally, SVFormer is orthogonal to the classical method GQA, and they can be used concurrently. Experimental results show that SVFormer trains significantly faster than the vanilla Transformer. Its performance mainly depends on factors like training sequence length and cumulative learning rate.

We experiment on a 20B SlimPajama sub-sampled dataset, using settings similar to popular large language models (Wei et al., 2023; Dubey et al., 2024; Kaplan et al., 2020). We compare different models by their relative training curves against the vanilla Transformer. Results show that ResFormer outperforms the vanilla Transformer, DenseFormer, and NeuTRENO across all settings. Furthermore, SVFormer trains faster than the vanilla Transformer and performs better when the sequence length is longer.

2 RELATED WORK

2.1 SHORTCUT CONNECTIONS FOR BETTER INFORMATION FLOW

Deep learning models often consist of multiple layers, posing a challenge to minimize information loss during transmission. ResNet (He et al., 2016) mitigates the vanishing gradient problem with

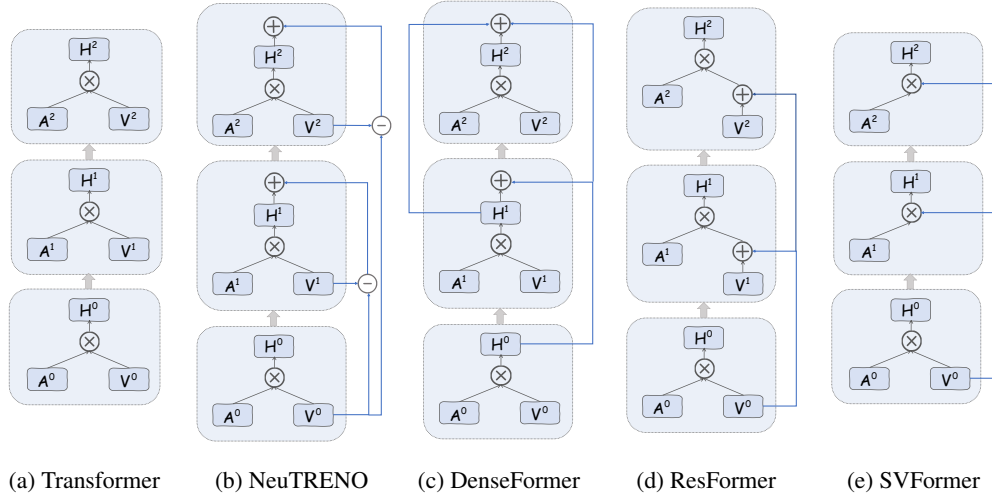


Figure 2: Simplified illustration of the vanilla Transformer, NeuTRENO, DenseFormer, ResFormer, and SVFormer, with only three-layer structures and no operations other than attention. A^i , V^i , and H^i denote the attention matrix, value vectors, and attention outputs at the i -th layer, respectively. \oplus , \ominus , and \otimes represent standard matrix addition, subtraction, and multiplication, respectively.

identity connections. Stochastic Depth (Huang et al., 2016) enhances training by randomly dropping layers. DenseNet (Huang et al., 2017) allows subsequent layers to directly access the hidden states of all preceding layers. These two methods further enhance the information flow after ResNet.

Related research indicates that for advanced Transformer architectures, although increasing depth continues to yield performance improvements in language modeling tasks, the gains become less significant with further increases (Petty et al., 2024). Furthermore, Zhou et al. (2021) illustrates that a 32-layer ViT underperforms a 24-layer ViT. Depth-Wise Attention (ElNokrashy et al., 2024) allows each query to access the key and value at the same position from previous layers through an attention-like mechanism before the output layer. DenseFormer (Pagliardini et al., 2024) integrates weighted fusion of outputs from all preceding layers after each layer. To explore why increasing depth in Transformers does not yield expected gains, Wang et al. (2022) finds that self-attention acts as a low-pass filter, smoothing token representations in ViTs. Additionally, Shi et al. (2022) investigates over-smoothing from a graph perspective in BERT-based language modeling tasks. NeuTRENO (Nguyen et al., 2023) adds the difference between the value vectors of the first and current layers to each layer’s attention output and significantly alleviates the over-smoothing problem.

In contrast to these methods, ResFormer accesses and integrates information from previous layers prior to the attention operation, as illustrated in Fig. 2. Moreover, it does not require the selection or tuning of additional hyperparameters.

2.2 KV CACHE COMPRESSING

The KV cache is a key factor limiting the efficiency of long-text model inference. Research in this area can be broadly classified into Transformer-based methods, which target redundant information in Transformer models, and non-Transformer methods, which mainly addresses the quadratic time complexity of attention with respect to sequence length.

For non-Transformer methods, Mamba (Gu & Dao, 2023) and RWKV (Peng et al., 2023) are two popular works. They replace the original softmax-based attention with SSM (Gu et al., 2021) and AFT (Zhai et al., 2021) mechanisms, respectively. Besides, several approaches have been proposed to enhance models’ ability to process long texts while reducing the reliance on KV cache. Dai (2019) advocates segmenting long texts into smaller parts for attention computation. Building upon this, Munkhdalai et al. (2024) proposes using a fixed-size memory matrix for storing and retrieving historical information.

Transformer-based methods can be categorized into three main groups. The first group consists of post-training methods like SnapKV (Li et al., 2024) and ThinK (Xu et al., 2024), which compress KV cache during inference based on attention matrices at token or hidden dimension levels. The second group focuses on quantization and adopts low-precision KV cache quantization rather than completely eliminating them (Hooper et al., 2024). The third group aims to maximize the efficiency of attention-based models via parameter or activation value sharing. The most representative works include Multi-Query Attention (Shazeer, 2019) and Grouped-Query Attention (Ainslie et al., 2023) which suggest to share key and value across a group of queries. MLKV (Zuhri et al., 2024) further suggest to share keys and values for queries across layers and MLA (Liu et al., 2024) introduces low-rank projection when processing keys and values. Besides, CLA (Brandon et al., 2024) and LISA (Mu et al., 2024) respectively point out that we can reuse keys, values, or the attention matrix across layers to reduce redundancy between layers. While these methods typically process both key and value simultaneously, SVFormer is the first approach to decouple value from query and key. It shares values across all layers and is compatible with other methods like GQA.

3 METHOD

3.1 MOTIVATION: INFORMATION TRANSFER VIA CROSS LAYER ATTENTION

Let $\mathbf{H}_n \in \mathbb{R}^{l \times d}$ be the input hidden state of the n -th layer, where l denotes the sequence length and d is the dimension size. In standard attention, the hidden state \mathbf{H}_n will be firstly projected into $\mathbf{Q}_n, \mathbf{K}_n, \mathbf{V}_n \in \mathbb{R}^{l \times d}$ through three linear projections $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d}$ respectively. For simplicity, we introduce dot-product attention of layer n as

$$\text{Attention}(\mathbf{Q}_n, \mathbf{K}_n, \mathbf{V}_n) = \text{Softmax}\left(\frac{\mathbf{Q}_n \mathbf{K}_n^T}{\sqrt{d}}\right) \mathbf{V}_n. \quad (1)$$

An ideal way to incorporate previous layers' information is cross layer attention. The attention mechanism naturally extracts relevant information from previous layers. If these layers contain low-quality information, the similarity between the current layer's query and the previous layers' keys will be low, thus minimizing negative impacts. Given $m < n$ and the information $(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m)$ of m -th layer, the cross layer mechanism calculates the attention output \mathbf{U}_n of n -th layer by the following attention formula:

$$\mathbf{U}_n = \text{Softmax}\left(\mathbf{Q}_n \text{Concat}(\mathbf{K}_n, \mathbf{K}_m)^T / \sqrt{d}\right) \text{Concat}(\mathbf{V}_n, \mathbf{V}_m). \quad (2)$$

In practice, cross-layer attention enhances feature fusion by allowing information to flow between layers, capturing both intra-layer and inter-layer dependencies. However, this approach introduces additional computational overhead due to the concatenation of keys and values from multiple layers. For example, in scenarios described by Eqn. 2, the overall computational complexity of the model nearly doubles compared with vanilla attention described in Eqn. 1.

3.2 EFFICIENT CROSS LAYER ATTENTION

To solve this problem, we propose to replace the \mathbf{K}_m with \mathbf{K}_n in Eqn. 2, as shown in Eqn. 3.

$$\mathbf{U}_n \approx \text{Softmax}\left(\mathbf{Q}_n \text{Concat}(\mathbf{K}_n, \mathbf{K}_n)^T / \sqrt{d}\right) \text{Concat}(\mathbf{V}_n, \mathbf{V}_m) \quad (3)$$

$$= \frac{1}{2} \text{Softmax}\left(\mathbf{Q}_n \mathbf{K}_n^T / \sqrt{d}\right) (\mathbf{V}_n + \mathbf{V}_m). \quad (4)$$

Utilizing the concept of block matrices, Eqn. 3 can be further simplified into Eqn. 4. This simplification converts the concatenation operation of the two value matrices into an addition operation. Compared to Eqn. 1, this new method only brings a minimal increase in computational complexity while still leveraging the information from the m -th layer in the n -th layer. Furthermore, Eqn. 4 can be generalized to incorporate cross-layer attention across all preceding $n - 1$ layers as follows:

$$\mathbf{U}_n \approx \frac{1}{n} \mathbf{A}_n \sum_{i=1}^n \mathbf{V}_i. \quad (5)$$

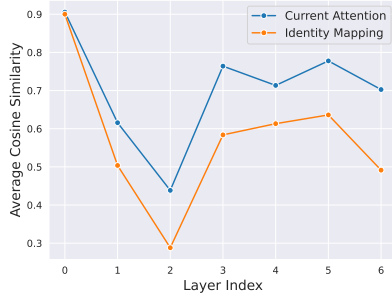


Figure 3: Average token similarity between the outputs of different mapping methods and that of Eqn. 2.

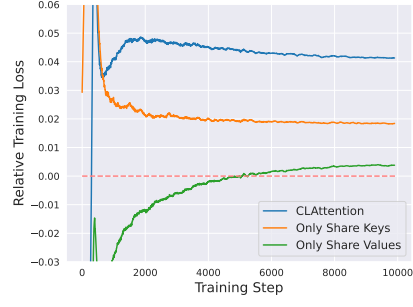


Figure 4: Ablation study on sharing keys or values in every two layers, with CLAttention denoting sharing both.

where \mathbf{A}_n denotes the original attention matrix for layer n . From the perspective of information propagation, model described by Eqn. 3 projects the historical values into the current layer’s embedding space using the current layer’s attention as a weight matrix. For example, a naive approach would be to perform identity mapping, as described by

$$\mathbf{U}_n = \mathbf{A}_n \mathbf{V}_n + \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{V}_i. \quad (6)$$

To evaluate the approximation effect of replacing the \mathbf{K}_m with \mathbf{K}_n , we randomly select 1,000 pre-training data samples. For each layer of a trained baseline model, assuming cross-layer attention defined by Eqn. 2 is required for each layer with respect to the previous one, we calculate the cosine similarity between the outputs from Eqn. 2 and Eqn. 3. We also calculate the cosine similarity between the outputs from Eqn. 2 and Eqn. 6 for comparison. Fig. 3 shows that using current attention as a mapping matrix provides a better approximation for cross-layer attention.

3.3 TRANSFORMER WITH RESIDUAL VALUE

Based on Eqn. 5, we propose a variant of Transformer with residual value (ResFormer) which only chooses first layer as the target of cross layer attention since the first layer contains all basic information of each token. The analysis of entropy in Fig. 1 (Right) supports this point, indicating that attention tends to be relatively dispersed across different tokens in the initial layers of the model. The attention mechanism of ResFormer can be formulated as

$$\mathbf{U}_n = \frac{1}{2} \mathbf{A}_n (\mathbf{V}_n + \mathbf{V}_1). \quad (7)$$

where $n \geq 2$ and standard attention is applied in the first layer. From the training perspective, it explicitly learns a residual mapping instead of directly learning the desired underlying mapping and that’s why we call it ResFormer.

3.4 A UNIFIED VIEW OF NEUTRENO AND DENSEFORMER

Using our framework, the NeuTRENO can be defined as

$$\mathbf{U}_n = (\mathbf{A}_n - \lambda \mathbf{J}) \mathbf{V}_n + \lambda \mathbf{V}_1. \quad (8)$$

where \mathbf{J} denotes a matrix of ones and λ is a hyper-parameter. It can be found that the term of $\lambda \mathbf{J}$ may have certain negative impact on the learning of original attention. If we ignore the attention output projections and the MLP layers, DenseFormer can also be modeled within our framework as

$$\mathbf{U}_n = \sum_{i=1}^n \alpha_i \mathbf{A}_i \mathbf{V}_i. \quad (9)$$

where $\{\alpha_i\}_{i=1}^n$ is a set of hyper-parameters. DenseFormer uses attention matrix of previous layer as the weight matrix of projecting values but this is not aligned with the concept of the efficient cross layer attention shown in Eqn. 3.

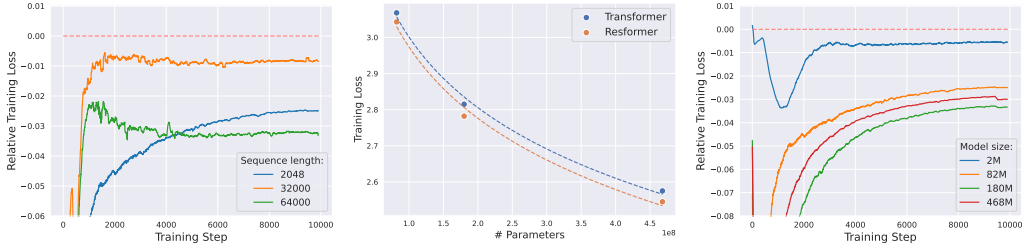


Figure 5: (Left) The relative training curve between a 82M ResFormer and Transformer across different training sequence lengths. (Middle) Average training loss for the final 50 steps across different model sizes and the corresponding fitted curves. (Right) The relative training curve across different model size for a fixed 2,048 training sequence length.

3.5 SVFORMER: SINGLE-LAYER VALUE FOR HALF KV CACHE

After ResFormer, a natural idea is whether we can remove the value vectors in each layer and have all layers share the value vectors from the first layer. We call this method SVFormer. Similar to ResFormer, SVFormer still adopts standard attention in the first layer and obtain the attention output \mathbf{U}_n for n -th layer where $n \geq 2$ through

$$\mathbf{U}_n = \mathbf{A}_n \mathbf{V}_1. \quad (10)$$

Compared to previous methods, SVFormer is the first method that decouple value vectors from attention. Its main advantage is that it only requires computing and storing the value vectors for the first layer, saving nearly half of the KV cache during inference. Similar methods like CLA reduce KV cache by sharing both of the key and value vectors every two layers, but the results in Fig. 4 show that sharing values has less negative impact on performance compared with sharing keys.

4 PRETRAIN EXPERIMENTS

4.1 SETTING

4.1.1 TRAINING DETAILS

Following Brandon et al. (2024), we choose the Llama-like architecture and SlimPajama (Soboleva et al., 2023) data for main experiments. Specifically, the architecture includes pre-normalization, SwiGLU activations (Shazeer, 2020), rotary position embedding (Su et al., 2024), and no dropout. For slimpajama, we randomly sample nearly 20B tokens according to the original data distribution of seven domains during training and adopt tokenizer used for “RedPajama-INCITE-7B-Base”. The details of training data can be found in Table 2 in Appendix.

Unless otherwise noted, we train all models using AdamW optimizer with 0.1 weight decay, $\beta_1 = 0.9$, $\beta_2 = 0.95$ and the max grad norm 1.0. The batch size is set to be around 2M tokens (Zhang et al., 2024) with a sequence length of 2,048 and the total steps is fixed 10,000 steps (Kaplan et al., 2020). We adopt linear learning rate warmup for the first 1,200 steps with the initial learning rate and the peak learning rate to be $1e-7$ and $6e-4$ respectively. The cosine decay schedule gradually decays to 10% of the peak learning rate by the end of training (Brandon et al., 2024; Zhou et al., 2024; Wei et al., 2023). See Table 4 and Table 3 in Appendix for more details.

All models are trained with 8 Nvidia A100 80G GPUs using mixed-precision training in FP16. We adopt deepspeed zero-2 optimizer and flash attention mechanism.

4.1.2 RELATIVE TRAINING LOSS CURVE ON SLIMPAJAMA

We trained all models for only one epoch on SlimPajama subsets, and primarily use training loss to compare different models. Furthermore, we use the relative training loss curve for better visualizing the difference among different models from the perspective of loss landscape. Specifically, for each method, we will subtract the smoothed training curve of the vanilla Transformer, obtained under the

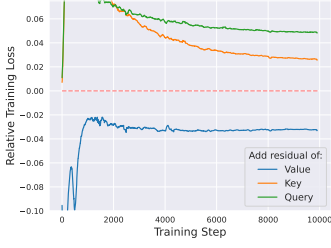


Figure 6: Ablation study of adding residual connection to queries or keys.

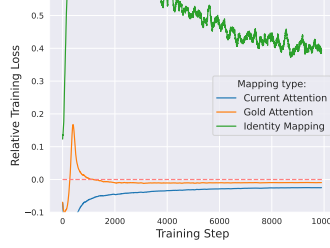


Figure 7: Ablation study of adding residual connection using different mapping matrix.

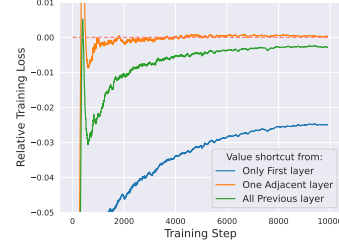


Figure 8: Ablation studies on which historical layer's value to include in residual connections.

same experimental settings, from the smoothed training curves of the method. The smoothing is done using a window size of 10 steps or 100 steps.

4.1.3 ENTROPY FOR ANALYZING ATTENTION CONCENTRATION EFFECTS

Given the attention matrix $\mathbf{A} \in \mathbb{R}^{l \times l}$ at one layer, we use entropy e to represent its concentration effect. To obtain entropy e , calculate the importance vector $\mathbf{a} = \frac{1}{l} \sum_{j=1}^l A_{ij}$ firstly where \mathbf{A} is a lower triangular matrix. The entropy can be formulated as follows:

$$e = - \sum_{i=1}^l a'_i \log a'_i. \quad (11)$$

where $a'_i = a_i / (\sum_{i=1}^l a_i)$ for $i = 1, 2, \dots, l$ and the higher the entropy e , the greater the degree of clustering in \mathbf{a} , i.e., attention matrix \mathbf{A} is more likely to focus on several specific tokens. 1,000 examples with a sequence length of 2,048 in the training dataset are selected for entropy analysis.

4.1.4 SPECTRAL DECOMPOSITION FOR ANALYZING REPRESENTATIONS

Spectral Decomposition is a classical method to analyze the representations of models. Zhu et al. (2021) suggests that the eigenvectors with larger eigenvalues are more transferable. Here we use spectral decomposition to analyze the feature space of value \mathbf{v} of one layer as following:

$$\frac{1}{l} \sum_{i=1}^l \mathbf{v}_i \mathbf{v}_i^T = \sum_{i=j}^d \mathbf{w}_j \lambda_j \mathbf{w}_j^T. \quad (12)$$

where \mathbf{w}_j is the j -th eigenvector with eigenvalue λ_j for $i = 1, 2, \dots, d$ and d is the dimensionality of the value's feature space.

4.2 RESFORMER vs. VANILLA TRANSFORMER

We trained ResFormer and vanilla Transformer with different model size on data with different sequence lengths. In Fig. 5, ResFormer consistently outperforms vanilla Attention throughout training for different models sizes. Additionally, the results in Fig. 1 (Left) illustrate that ResFormer outperforms DenseFormer and NeuTRENO. Furthermore, integrating ResFormer with NeuTRENO leads to additional performance improvements.

4.3 ABLATION STUDY OF RESIDUAL CONNECTION

In Eqn. 4, we employ residual connections for the values. We compare this approach with models that add residual connections to queries or keys. The results, shown in Fig. 6, indicate that only residual connections for values yield positive effects. One possible explanation is that attention mechanisms are sensitive to perturbations, and modifying queries or keys can significantly impact attention matrix.

Moreover, we compare with the models based on Eqn. 2 and Eqn. 6. From the cross layer attention perspective, the former uses a gold attention matrix in Eqn. 2 to map historical values, while the

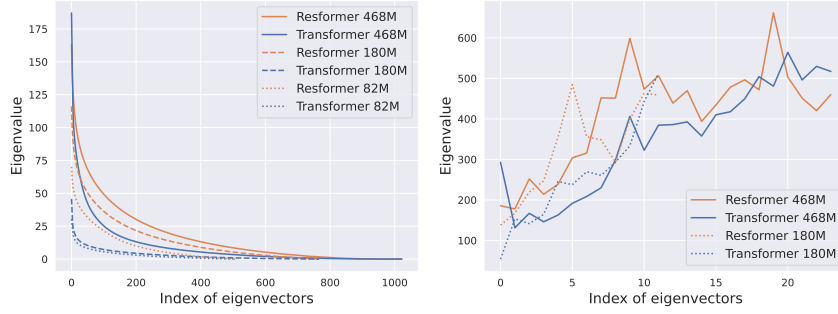


Figure 9: Left: Distribution of eigenvalues for the value vectors in the first layer of ResFormer and Transformer. Right: Maximum eigenvalue for each layer of ResFormer and Transformer.

latter uses an all-ones matrix in Eqn. 6. The results in Fig. 7 align with Fig. 3, showing that identity mapping causes significant perturbations, leading to poor performance. Interestingly, using current attention as the mapping matrix results in an even lower final loss than using gold attention.

When determining the mapping method and target value, it is crucial to consider which historical layers’ values should be included in the residual connection. Fig. 8 shows that each Transformer layer should add a shortcut to the first layer’s value rather than to the nearest preceding layer or all previous layers, highlighting the first-layer value’s critical importance. A potential explanation is that incorporating values from other layers may dilute the impact of the first-layer value.

4.4 DOWNSTREAM EVALUATIONS

We compare the different models on several classical reasoning tasks following (Zhang et al., 2024) in a zero-shot way. The tasks include Hellaswag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), WinoGrande (Sakaguchi et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018) and PIQA (Bisk et al., 2020). The results in Table 1 show that ResFormer (82M) achieves an average accuracy improvement of nearly 3% compared to the vanilla Transformer (82M).

Model	Max Length	HellaSwag	Obqa	WinoGrande	ARC-c	ARC-e	PIQA	Avg.
Transformer	2,048	0.263	0.142	0.492	0.199	0.331	0.572	0.333
ResFormer	2,048	0.273	0.148	0.512	0.182	0.414	0.604	0.355
Transformer	64,000	0.267	0.142	0.485	0.179	0.322	0.570	0.328
ResFormer	64,000	0.274	0.136	0.513	0.184	0.407	0.588	0.350

Table 1: Zero-shot accuracy on commonsense reasoning tasks.

4.5 VISUALIZATION OF RESFORMER

To figure out why ResFormer can achieve better performance on language modeling tasks than vanilla Transformer, we conduct visualization based on the eigenvalue decomposition discussed in Section 4.1.4. After sorting the eigenvalues in descending order, we compute the average eigenvalue for each layer across 1,000 randomly sampled pre-train data examples. The results in Fig. 9 indicate that the value vectors generated by most layers of the ResFormer exhibit stronger representational capacity compared to those of the vanilla Transformer. Besides, we analyze the attention concentration effects mentioned in Section 4.1.3 using the same batch of test data. Fig. 1 (Middle) illustrates that the clustering effect of attention increases significantly with the number of layers for the vanilla Transformer, whereas the clustering effect is relatively less pronounced for the ResFormer.

4.6 SVFORMER vs. GQA

In the Fig. 10, at a training sequence length of 64,000, SVFormer demonstrates lower final loss compared to existing KV-efficient methods such as CLA and GQA. Moreover, it can be used concurrently with GQA to enhance KV efficiency further. However, we observed that with a training

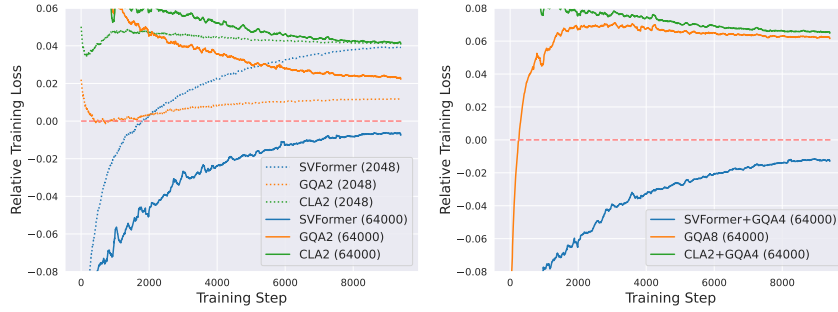


Figure 10: The relative training loss for SVFormer and other KV efficient model compared with vanilla attention. The numbers in parentheses represent the training sequence length. Left: Model with nearly $1/2$ KV cache. Right: Model with nearly $1/8$ KV cache.

sequence length of 2,048, SVFormer underperforms compared to GQA. The results indicate that sequence length significantly affects SVFormer’s performance. Thus, we conducted more comprehensive experiments on sequence length.

Results in Fig. 11 (Left) demonstrate that SVFormer will always be gradually surpassed by vanilla attention during training while its training speed is much faster than vanilla attention. However, as the training sequence length increases, the SVFormer model performs better. In this way, we focus on the critical point, defined as the number of training steps exceeded. Fig. 11 (Right) illustrates that the relationship between the critical point and sequence length exhibits an exponential trend. We argue that it’s due to the challenge deep models face in fully optimizing the increasingly larger first-layer value matrix as the training sequence length grows.

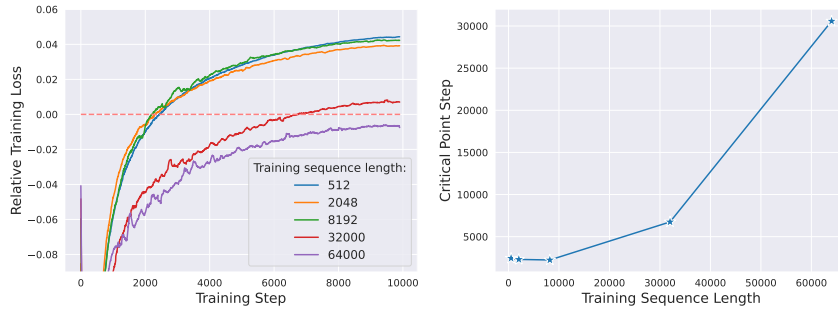


Figure 11: Left: The relative training loss for SVFormer under different sequence lengths with a fixed batch size of 2M tokens. Right: Analysis of critical point, and we predict it for length 64,000 using linear regression with the last 1,000 data points.

4.7 OTHER FACTORS INFLUENCING SVFORMER

Intuitively, the training effectiveness of SVFormer is influenced by factors such as the maximum learning rate, warmup steps, model size, and other factors beyond just the training sequence length. We conducted experiments to explore these relationships.

Based on the results shown in Fig. 12a and Fig. 12b, a smaller learning rate benefits SVFormer more, with warmup’s impact being comparatively small. This could be attributed to the model’s outcomes being closely tied to the total summed learning rate, which has weak connection with warmup steps (Kaplan et al., 2020). Moreover, larger models often require smaller learning rates to ensure training stability, making them more suitable for using SVFormer.

Except for the 2M model, Llama-like models ranging from 82M to 468M, as well as models with the GPT2 architecture, exhibit similar critical points and final losses (see Fig. 12c and Fig. 12d). This suggests that the difference between SVFormer and the vanilla Transformer is not sensitive to model size and architecture.

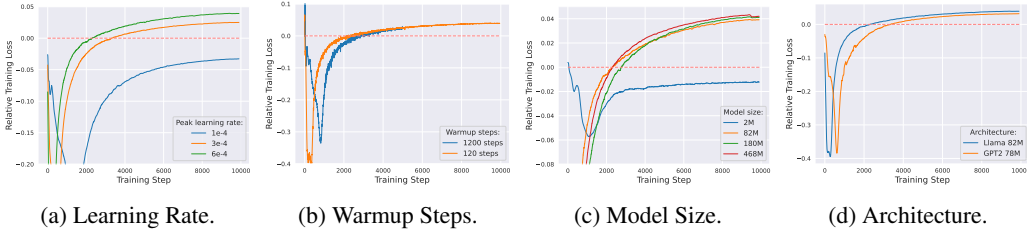


Figure 12: The relative training loss for SVFormer under different hyper-parameter setting.

4.8 ABLATION STUDY OF SVFORMER

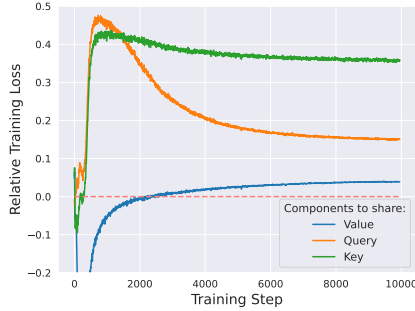


Figure 13: Ablation study of sharing first layer’s query(key) across all layers.

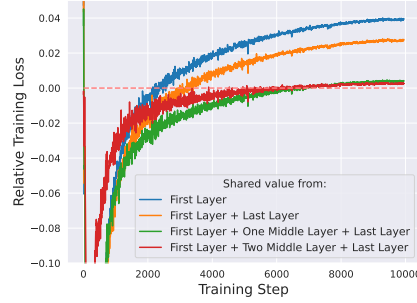


Figure 14: Ablation study on sharing values from different numbers of layers.

To better understand SVFormer, we conduct several ablation experiments. We first observe the effects of sharing the first layer’s queries or keys across all layers in Fig. 13, finding that this significantly impacts model performance, similar to the results in Fig. 4. Additionally, sharing the first layer’s values in a multi-layer network may reduce the network’s “effective depth.” By updating the shared values using intermediate layers as “anchors,” we find that increasing the number of “anchors” improves performance, as shown in Fig. 14.

5 CONCLUSION

In this paper, we propose the concept of attention concentration, a problem that arises from stacking multiple attention layers. From the perspective of cross-layer attention, we derive ResFormer, which adds a residual connection between the value vectors of the current layer and those of the first layer before the attention operation to alleviate attention concentration. Additionally, we introduce SVFormer, based on ResFormer, which reduces the KV cache by nearly half. We conducted comprehensive experiments on the language modeling task to validate the advantages of these two Transformer variants in different scenarios.

ACKNOWLEDGEMENT

This work was supported by the Research Center for Industries of the Future at Westlake University (Grant No. WU2023C017) and the Key Research.

ETHICS STATEMENT

On the one hand, the data employed in this paper is sourced from publicly available datasets provided by the company, which have undergone a certain level of filtering. On the other hand, the models trained in our study are solely utilized for experimental analysis and will not be publicly deployed.

REPRODUCIBILITY STATEMENT

We have detailed the complete experiments setup such as batch size, optimizer and learning rates in Section 4.1.1. Besides, we have released the source codes. These resources should be sufficient to reproduce the results of the paper.

REFERENCES

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/aaai.v34i05.6239. URL <https://doi.org/10.1609/aaai.v34i05.6239>.
- William Brandon, Mayank Mishra, Aniruddha Nrusimha, Rameswar Panda, and Jonathan Ragan Kelly. Reducing transformer key-value cache size with cross-layer attention. *arXiv preprint arXiv:2405.12981*, 2024.
- Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL <http://arxiv.org/abs/1803.05457>.
- Zihang Dai. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Muhammad ElNokrashy, Badr AlKhamissi, and Mona Diab. Depth-wise attention (dwatt): A layer fusion method for data-efficient classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4665–4674, 2024.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

-
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*, 2024.
- Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646–661. Springer, 2016.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Z Lan. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.
- Yongyu Mu, Yuzhang Wu, Yuchun Fan, Chenglong Wang, Hengyu Li, Qiaozhi He, Murun Yang, Tong Xiao, and Jingbo Zhu. Cross-layer attention sharing for large language models. *arXiv preprint arXiv:2408.01890*, 2024.
- Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.
- Tam Nguyen, Tan Nguyen, and Richard Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36:80233–80256, 2023.
- Matteo Pagliardini, Amirkeivan Mohtashami, Francois Fleuret, and Martin Jaggi. Denseformer: Enhancing information flow in transformers via depth weighted averaging. *arXiv preprint arXiv:2402.02622*, 2024.
- Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. RwkV: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.
- Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7232–7245, 2024.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

-
- Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.
- Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022.
- Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. <https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama>, 2023. URL <https://huggingface.co/datasets/cerebras/SlimPajama-627B>.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022.
- Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yuhui Xu, Zhanming Jie, Hanze Dong, Lei Wang, Xudong Lu, Aojun Zhou, Amrita Saha, Caiming Xiong, and Doyen Sahoo. Think: Thinner key cache by query-driven pruning. *arXiv preprint arXiv:2407.21018*, 2024.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL <https://doi.org/10.18653/v1/p19-1472>.
- Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer. *arXiv preprint arXiv:2105.14103*, 2021.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiao Chen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.
- Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021.
- Zayd Muhammad Kawakibi Zuhri, Muhammad Farid Adilazuarda, Ayu Purwarianti, and Alham Fikri Aji. Mlkv: Multi-layer key-value heads for memory efficient transformer decoding. *arXiv preprint arXiv:2406.09297*, 2024.

A APPENDIX

A.1 TOKEN SIMILARITY ANALYSIS

Attention concentration tends to make embeddings of different tokens more similar, resulting in over-smoothing. The extent of over-smoothing can be assessed by calculating the average token similarity s of the hidden states using the following formula:

$$s = \frac{1}{l(l-1)} \sum_{i=1}^l \sum_{j=1, j \neq i}^l \text{Sim}(\mathbf{h}_i, \mathbf{h}_j). \quad (13)$$

where $\{\mathbf{h}_i\}_{i=1}^l$ is the hidden state of the i -th token and $\text{Sim}(\cdot)$ denotes the operation of cosine similarity. The results in Fig. 15 align with the results in Fig. 1. In the case of Llama and Mistral, the average token similarity demonstrates an “M”-shaped pattern with increasing network depth, while entropy follows a “W”-shaped pattern at corresponding positions. These trends suggest that attention concentration indeed leads to the phenomenon of over-smoothing.

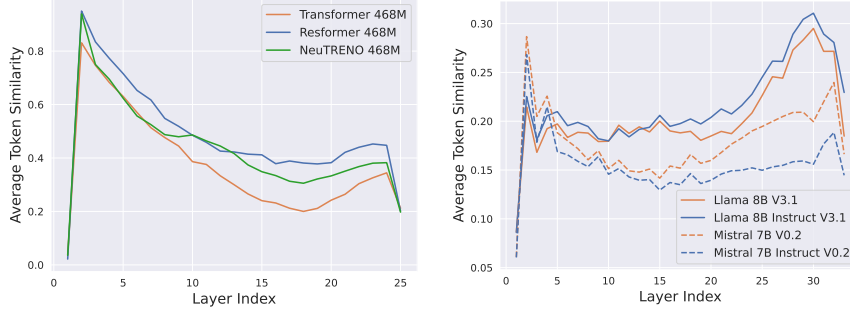


Figure 15: (Left) The average token similarity of hidden states across layers in ResFormer *vs.* the vanilla Transformer. (Right) The average token similarity of hidden states across layers in Llama (8B) (Dubey et al., 2024) and Mistral (7B) (Jiang et al., 2023).

A.2 PRE-TRAIN DATASET

Data source	proportions	Tokens
Commoncrawl	50%	10 B
C4	20%	4 B
GitHub	10%	2 B
Books	5%	1 B
ArXiv	5%	1 B
Wikipedia	5%	1 B
StackExchange	5%	1 B

Table 2: The details of pre-train dataset.

Based on the equation $D \geq 5000 \cdot N^{0.74}$ (Kaplan et al., 2020) where D is data size and N is the number of non-embedding parameters, we need to collect at least 17.5B for model has $N = 700\text{M}$ non-embedding parameters (corresponding to complete 1B model with 2,048 hidden size, 50,277 vocab size and 2,048 sequence length) to avoid over-fitting. Besides, Xie et al. (2024) indicates that the mixture proportions of pre-training data domains significantly affects the training results. In this way, we sampled 20B tokens data from original 627B data based on the original data proportions shown in the Table 2.

A.3 TRAINING DETAILS

Section 4.1.1 introduces the main experimental hyperparameters used in the paper. This section further details the training parameters for various model sizes and training sequence lengths. Table 4

Max Sequence Length	512	2,048	8,192	32,000	64,000
Total Batch Size	4,096	1,024	256	64	32
Per-GPU Batch Size	128	32	8	2	1
Gradient Accumulation Step			32		
GPUs			8		

Table 3: Training details for training dataset with different sequence length.

demonstrates the differences among models of various sizes. The configurations for the number of layers, attention heads, hidden dimensions, and FFN dimensions are based on Biderman et al. (2023). Additionally, the λ in Eqn. 8 is set to be 0.4 for NeuTRENO. Moreover, as reported in Table 3, the batch size that a single GPU can accommodate varies depending on the length of the training sequences. Note that the total number of tokens in each batch is consistently 2 million.

Model Size	2M	82M	180M	468M
Layers	4	8	12	24
Attention Heads	2	8	12	16
Hidden Dimension	16	512	768	1,024
FFN Dimension	56	1,792	2,688	3,584
Tie Word Embedding		False		
(Peak Learning Rate, Final Learning Rate)		$(6e-4, 6e-5)$		
Learning Rate Schedule		Cosine Decay		
Vocabulary Size		50,277		
Activation Function		SwiGLU		
Position Embedding		RoPE ($\theta = 10,000$)		
Batch Size		2M tokens		
Data Size		20B tokens		
(Warmup Steps, Training Steps)		(120, 10,000)		
Adam β		(0.9, 0.95)		
Dropout		0.0		
Weight Decay		0.1		

Table 4: Training details for models with different size.

A.4 VALIDATION LOSS ON SLIMPAJAMA

Section 4.1.2 introduces to use relative training loss as a main evaluation matrix. Table 5 reports the validation loss for different model on the whole validation split of slimpajama.

Model	Common Crawl	C4	Github	Stack Exchange	Wikipedia	Book	Arxiv	Avg.
Transformer (82M)	3.3595	3.5388	1.4247	2.3872	2.9047	3.3797	2.1779	2.7389
Transformer (180M)	3.0961	3.2834	1.2451	2.1651	2.5897	3.1309	2.0001	2.5015
Transformer (468M)	2.8514	3.0430	1.0908	1.9628	2.2821	2.8979	1.8362	2.2806
Resformer (82M)	3.3362	3.5191	1.3941	2.3592	2.8646	3.3572	2.1518	2.7117
Resformer (180M)	3.0631	3.2504	1.2200	2.1350	2.5435	3.0994	1.9732	2.4692
Resformer (468M)	2.8214	3.0115	1.0730	1.9388	2.2477	2.8696	1.8142	2.2537

Table 5: Validation loss on slimpajama.