

# NiNformer: A Network in Network Transformer with Token Mixing Generated Gating Function

Abdullah Nazhat Abdullah<sup>1</sup>[0000-0002-1757-0785] and Tarkan Aydin<sup>2</sup>[0000-0002-2018-405X]

<sup>1</sup> Bahcesehir University,Turkiye  
nazhat.abdullah@bahcesehir.edu.tr

<sup>2</sup> Bahcesehir University,Turkiye  
tarkan.aydin@bahcesehir.edu.tr

**Abstract.** The Attention mechanism is the main component of the Transformer architecture, and since its introduction, it has led to significant advancements in Deep Learning that span many domains and multiple tasks. The Attention Mechanism was utilized in Computer Vision as the Vision Transformer ViT, and its usage has expanded into many tasks in the vision domain, such as classification, segmentation, object detection, and image generation. While this mechanism is very expressive and capable, it comes with the drawback of being computationally expensive and requiring datasets of considerable size for effective optimization. To address these shortcomings, many designs have been proposed in the literature to reduce the computational burden and alleviate the data size requirements. Examples of such attempts in the vision domain are the MLP-Mixer, the Conv-Mixer, the Perciver-IO, and many more. This paper introduces a new computational block as an alternative to the standard ViT block that reduces the compute burdens by replacing the normal Attention layers with a Network in Network structure that enhances the static approach of the MLP-Mixer with a dynamic system of learning an element-wise gating function by a token mixing process. Extensive experimentation shows that the proposed design provides better performance than the baseline architectures on multiple datasets applied in the image classification task of the vision domain.

**Keywords:** Deep Learning · Computer Vision · Transformers

## 1 Introduction

Since the advent of the Transformer architecture [1] and the introduction of the Attention mechanism as its main computational component within the context of Natural Language Processing (NLP), large advancements have been achieved not only in language-related tasks but across all aspects related to the research and application of Machine Learning (ML). Transformers changed the landscape of NLP with the adoption of their architecture in designing highly successful and capable Large Language Models (LLM) [2] such as GPT [3], LLaMA [4], Falcon [5]

and Mistral [6]. The Computer Vision (CV) domain also experienced rapid adoption of Transformer architectures. Vision-specific implementations such as ViT [7], MLP-Mixer [8], Conv-Mixer [9], and Swin Transformer [10] were introduced, along with many application-oriented designs that utilize such architectures, such as Detection Transformer (DETR) [11], Perceiver-IO [12], Unified-IO [13], DINO [14], and Segment Anything Model (SAM) [15]. Given such wide adoption on many tasks and modalities, the need for more efficient implementations of Transformer has increased in importance, and in that direction, significant research attempts have been introduced, such as Linformer [16], FNet [17], Local-ViT [18], Max-ViT [19], and Nystromformer [20]. The MLP-Mixer implementation of Transformers is of interest as it introduces a unique design choice that targets a more efficient architecture with emphasis on the process of "token mixing." This mixing is applied in two stages [8]; the first stage applies the mixing within the input token representations, while the second stage applies the mixing process between corresponding positions in each token. A drawback of the MLP-Mixer design is that the mentioned mixing processes are performed with static weight matrices, which limits the capabilities of the architecture in comparison to the traditional Transformers that utilize the dynamic process of the Softmax Attention mechanism. At the same time, the traditional Transformer architecture has its own drawback of quadratic complexity in input size [21], which imposes a considerable cost in both training and inference when selecting the architecture. It is notable that in the literature there is a lack of a design that adopts the efficiency measures introduced in the MLP-Mixer model while also maintaining a dynamic information filtering mechanism, as with the traditional Transformer design. In this paper, we introduce a newly formulated computational block that can be used as a core process in constructing Transformer architectures that blends both efficient elementary operations and dynamic information filtering. The new proposal utilizes the MLP-Mixer token mixing to learn a generator of dynamic per input gating function that selectively filters the input representation tokens that are then passed to the per token MLP stage as in traditional Transformers, which results in a block that contains two levels of processing, an inner and an outer, hence the chosen name for the proposal as a Network in Network Transformer, or (NiNformer). In this work, the newly proposed architecture was trained and its performance evaluated with respect to multiple baselines that represent different architectural directions and a variety of design choices. The comparison was conducted on three datasets, and the experimentation was performed in an equalized setting with the same computational resources to ensure a fair evaluation. From the experiments conducted, it was observed that the NiNformer architecture was the most performing, and the obtained results verified the validity and capability of the underlying assumptions employed in our proposed computational block.

## 2 Related Work

The literature is rich with attempts to improve on the qualities and capabilities of the traditional Transformer architecture design [22],[23],[24],[25]. Guo et al. introduced Star Transformer [26], combining band attention and global attention. This formulation of the Transformer has a global node on which a band attention of width 3 is applied. Also, a shared global node connects a pair of non-adjacent nodes, while adjacent nodes are connected to each other. Beltagy et al. introduced Longformer[27], which also uses a combination of band attention and internal global-node attention. Classification tokens are selected as global nodes. The architecture substitutes the band attention heads in the upper layers with dilated window attention, thus increasing the receptive field without increasing computation. Kitaev et al. introduced Reformer [28] as a modified Transformer that employs locality-sensitive hashing (LSH). The LSH is used to select the key and value pairs for each query, therefore allowing each token to attend to tokens that exist in the same hashing bucket. BigBird architecture by Zaheer et al.[29] utilizes random attention to approximate full attention with a sparse encoder and sparse decoder, and it was shown by the analysis that this design can simulate any Turing Machine, explaining the capability of such architecture. Katharopoulos et al. proposed the Linear Transformer [30] with feature maps that target an approximation of the full Softmax Dot Product Attention and showed comparable performance in empirical tests. Wang et al. introduced Linformer [16], showing an approximation to the Attention mechanism by a low-rank matrix, thus lowering the computational requirement while maintaining comparable performance. Choromanski et al. proposed Performer [31], which uses random feature maps as an approximate to the Traditional Attention function. Wang et al. introduced the Cascade Transformer [32] By using a sliding window attention, the window size is exponentially increased when increasing the number of layers, leading to a reduction in complexity. Li et al. introduced the LogSparse Transformer [33] that facilitates long-term dependency on time series analysis by using Eponym attention. Qiu et al. introduced BlockBERT [34], which uses block-wise attention to split the input sequence into non-overlapping blocks. Tay et al. introduced the sparse Sinkhorn attention [35]. This mechanism is essentially block-wise attention, but the keys are sorted block-wise, therefore learning the permutations. Dai et al. proposed the Transformer-XL[36]. This design uses a recurrence between the windows that is segment-based. by storing the representations of the previous window and storing them in first-in, first-out memory (FIFO). After this step, the Transformer-XL applies attention to the sorted representations that have been stored in memory. Clustered Attention, proposed by Vyas et al. [37] clusters the quires, then calculates the attention distributions for cluster centroids. Zhang et al. proposed PoolingFormer [38], which utilizes a two-level attention, a sliding window attention, and a compressed memory attention. The compressed memory module is used after first applying the sliding window attention, then applying a compressed memory module for the purpose of increasing the receptive field. Liu et al. proposed Memory Compressed Attention (MCA) [39], which complements local attention with strided convolu-

tion, thus reducing the number of keys and values. This allows the architecture to process much longer sequences compared to traditional Transformers. Xiong et al. used the Nyström method to modify Transformer with the introduction of Nyströmformer [20]. This design selects landmark nodes by the process of strided average pooling and then processes these selected queries and keys with an approximation to attention by the Nyström method. Funnel Transformer [40] was proposed by Dai et al. by employing a funnel-like encoder that has a gradual reduction of the hidden sequence length using pooling along the sequence dimension; the proper length is then restored with an up-sampling process. Max-ViT [19] was introduced by Tu et al., which repeats the basic building block over multiple stages. The basic block consists of two aspects: blocked local attention and dilated global attention. Ho et al. proposed the Axial Transformer [41]. This architecture computes a sequence of attention functions with each one applied along a single axis of the input, reducing the computational cost. Swin Transformer [10] is an architecture proposed by Liu et al., and this design reduced the cost by splitting the image input into non-overlapping patches. These patches are then embedded as tokens for processing by Attention. FNets [17] was introduced by Lee-Thorp et al., and it proposes an attention-free Transformer architecture that substitutes the Softmax Dot Product function used in each The Fourier sublayer applies a 2D DFT to the embedded input in two steps: one 1D DFT along the sequence dimension and another 1D DFT along the hidden dimension. gMLP [42] was introduced by Liu et al., and this architecture is comprised of a series of blocks that are homogeneous in size and width. Each block layout is highly reminiscent of inverted bottlenecks. Another feature of this architecture compared to traditional Transformers is that it does not require position embeddings. Local-ViT [18] was introduced by Li et al. This architecture incorporates 2D depth-wise convolutions instead of the feed-forward network as in ViT. This design choice was inspired by the inverted residuals of MobileNets. Synthesizer [43] was proposed by Tay et al. as an architecture that learns synthetic attention weights and does not rely on interactions between tokens. The results showed competitive performance in relation to other linear transformer designs. Transformer iN Transformer (TNT) [44] was introduced by Han et al. This design treats the input images in a similar manner to a paragraph of text and divides them into several patches as “visual sentences” and then further divides them into sub-patches as “visual words.” With this hierarchical division, the architecture is divided into conventional transformer blocks for extracting features and attentions on the visual sentence level, and then a sub-transformer is introduced in order to extract the features of smaller visual words. De et al. proposed Hawk and Griffin models [45]; these are hybrid models combining gated linear recurrences and local attention with good extrapolation capabilities.

### 3 Methodology

#### 3.1 Baselines

For an extensive comparative analysis of capability, our proposed architecture is contrasted to multiple baseline architectures that represent a variety of functional principles. The ViT follows the principles of a traditional NLP Transformer, which represented the first iteration of designs that adopted such architecture. At its core, it relies on the Softmax Dot Product Attention, and as with NLP-oriented Transformers, the ViT also introduced the homogeneous layer structure. The following are the main equations for the ViT block:

$$\text{Attention}(Q,K,V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

$$Y(X) = \text{Attention}(\text{LayerNorm}(X)) + X \quad (2)$$

$$Z(Y) = \text{MLP}(\text{LayerNorm}(Y)) + Y \quad (3)$$

Below is the pseudocode for the ViT block:

---

**Algorithm 1** ViT

---

- 1: Input: Image  $I$ , number of classes  $C$ , patch size  $ps$ , embedding dimension  $d_{model}$ , number of Transformer blocks  $B$ , hidden dimension of MLP  $d_{mlp}$ , learning rate  $\eta$
  - 2: Output: Predicted class probabilities
  - 3: Steps:
    1. Divide  $I$  into patches of size  $ps \times ps$ .
    2. Flatten each patch and embed it into a  $d_{model}$ -dimensional vector using patch embedding layer.
    3. Concatenate the embedded patches into a sequence  $X$ .
    4. **for**  $i = 1$  to  $B$  **do**:
      - \* divide  $X$  into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply Attention.
      - \* Add the residual path.
      - \* divide Attention result into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply MLP block.
      - \* Add the residual path.
    5. Apply global average pooling to the output of the last Transformer block.
    6. Use a fully connected layer with  $C$  output units and softmax activation to obtain class probabilities.
    7. Train the model by minimizing the loss between predicted and true labels using gradient descent with learning rate  $\eta$ .
- 

The MLP-Mixer adopts the homogeneous layer structure as with the ViT but introduces efficiency-oriented computational operations of mixing (interacting)

the token representation with the application of MLP that are applied in two successive stages: first, an MLP mixing of per token representation, and second, a per position (channel) MLP mixing of representations in between the tokens.

The following is the main equation for the MLP-Mixer block:

$$Y(X) = \text{Transpose}(\text{MLP}(\text{Transpose}(\text{LayerNorm}(X)))) + X \quad (4)$$

$$Z(Y) = \text{MLP}(\text{LayerNorm}(Y)) + Y \quad (5)$$

Below is the pseudocode for the MLP-Mixer block:

---

**Algorithm 2** MLP-Mixer

---

- 1: Input: Image  $I$ , number of classes  $C$ , patch size  $ps$ , embedding dimension  $d_{model}$ , number of Transformer blocks  $B$ , hidden dimension of MLP  $d_{mlp}$ , learning rate  $\eta$
  - 2: Output: Predicted class probabilities
  - 3: Steps:
    1. Divide  $I$  into patches of size  $ps \times ps$ .
    2. Flatten each patch and embed it into a  $d_{model}$ -dimensional vector using patch embedding layer.
    3. Concatenate the embedded patches into a sequence  $X$ .
    4. **for**  $i = 1$  to  $B$  **do**:
      - \* divide  $X$  into residual and nonresidual paths.
      - \* normalize the nonresidual path and Transpose.
      - \* Apply MLP block.
      - \* Transpose.
      - \* Add the residual path.
      - \* divide result into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply MLP block.
      - \* Add the residual path.
    5. Apply global average pooling to the output of the last Transformer block.
    6. Use a fully connected layer with  $C$  output units and softmax activation to obtain class probabilities.
    7. Train the model by minimizing the loss between predicted and true labels using gradient descent with learning rate  $\eta$ .
- 

The Local-ViT adopts a conservative design choice to introduce a more lightweight variant of the original ViT by replacing the per-token MLP layer in the ViT block with convolutions.

The following are the main equations for the Local-ViT block:

$$Y(X) = \text{Attention}(\text{LayerNorm}(X)) + X \quad (6)$$

$$Z(Y) = \text{CONV}(\text{LayerNorm}(Y)) + Y \quad (7)$$

Below is the pseudocode for the Local-ViT block:

**Algorithm 3** Local-ViT

- 
- 1: Input: Image  $I$ , number of classes  $C$ , patch size  $ps$ , embedding dimension  $d_{model}$ , number of Transformer blocks  $B$ , hidden dimension of MLP  $d_{mlp}$ , learning rate  $\eta$
  - 2: Output: Predicted class probabilities
  - 3: Steps:
    1. Divide  $I$  into patches of size  $ps \times ps$ .
    2. Flatten each patch and embed it into a  $d_{model}$ -dimensional vector using patch embedding layer.
    3. Concatenate the embedded patches into a sequence  $X$ .
    4. **for**  $i = 1$  to  $B$  **do**:
      - \* divide  $X$  into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply Attention.
      - \* Add the residual path.
      - \* divide Attention result into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply CONV block.
      - \* Add the residual path.
    5. Apply global average pooling to the output of the last Transformer block.
    6. Use a fully connected layer with  $C$  output units and softmax activation to obtain class probabilities.
    7. Train the model by minimizing the loss between predicted and true labels using gradient descent with learning rate  $\eta$ .
- 

**3.2 Proposed Architecture**

The proposed computational block of this paper is comprised of two levels: an outer network that resembles a Transformer block by including a token-wise MLP, which provides the design with an optimization-driven token mapping capability. The token-wise MLP of the outer network is preceded in the proposed block by a substitute for the Attention mechanism, which has a gating function process on the outer network level that extends the concept of Gated Linear Unit (GLU) [46] by employing a network in network structure. In the proposed Gating-Unit, the gating signal is generated by a sub-unit in the inner network, where the inner sub-unit uses a token-mixing architecture of the MLP-Mixer. The proposed design significantly differs from TNT architecture [44] in that the two levels in our proposal are different in form and function, and both inner and outer levels apply their transformations to the input context as a whole, while the TNT architecture has two levels of the same traditional Attention mechanism that are applied on two separate scales, the visual word scale and the visual sentence scale within the input context. Such distinction of scales omits processing of the global correlations that may exist between parts of the context in the case of TNT, and our design utilizes the full context on both of its two levels to capture the global correlations of the input. In addition, the newly introduced gating mechanism has the advantage of using the non-dynamic, fixed-weight MLP-Mixer as an inner sub-unit to learn the interdependencies from the input representation, which is then used by the outer level as a dynamic gating signal that functions on an input by input basis to scale the values of its

linearly projected representation, thus facilitating further information processing by the outer level MLPs without the use of the Softmax Dot Product Attention employed in generic Transformer architectures.

The following equations describe the operation of the proposed block:

$$\text{Gating}(I) = (\text{MLPMixer}(I)) * \text{Linear}(I) \quad (8)$$

$$Y(X) = \text{Gating}(\text{LayerNorm}(X)) + X \quad (9)$$

$$Z(Y) = \text{MLP}(\text{LayerNorm}(Y)) + Y \quad (10)$$

The following is the pseudocode of the proposed design:

---

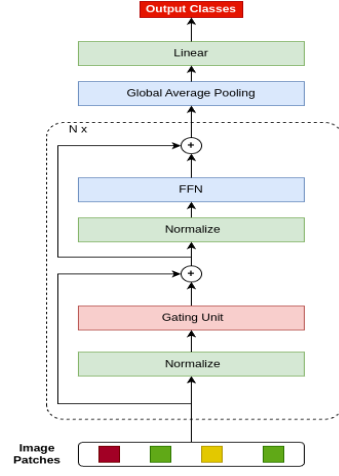
**Algorithm 4** NiNformer

---

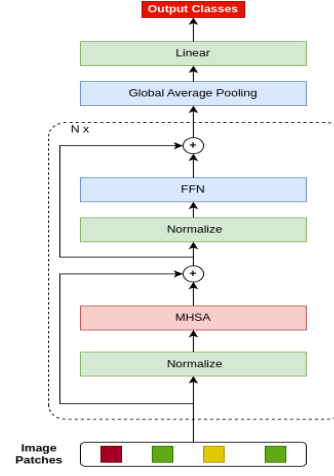
- 1: Input: Image  $I$ , number of classes  $C$ , patch size  $ps$ , embedding dimension  $d_{model}$ , number of Transformer blocks  $B$ , hidden dimension of MLP  $d_{mlp}$ , learning rate  $\eta$
  - 2: Output: Predicted class probabilities
  - 3: Steps:
    1. Divide  $I$  into patches of size  $ps \times ps$ .
    2. Flatten each patch and embed it into a  $d_{model}$ -dimensional vector using patch embedding layer.
    3. Concatenate the embedded patches into a sequence  $X$ .
    4. **for**  $i = 1$  to  $B$  **do**:
      - \* divide  $X$  into residual and nonresidual paths.
      - \* normalize the nonresidual path
      - \* generate the gating signal by the application of the MLP-Mixer sub-unit on the nonresidual path.
      - \* Apply the Gating by multiplying the linearly projected nonresidual path with the MLP-Mixer sub-unit output.
      - \* Add the residual path.
      - \* divide Gating result into residual and nonresidual paths.
      - \* normalize the nonresidual path and Apply MLP block.
      - \* Add the residual path.
    5. Apply global average pooling to the output of the last Transformer block.
    6. Use a fully connected layer with  $C$  output units and softmax activation to obtain class probabilities.
    7. Train the model by minimizing the loss between predicted and true labels using gradient descent with learning rate  $\eta$ .
- 

Figure 1 shows the NiNformer overall architecture in comparison to the ViT and Local-ViT architectures, while Figure 2 compares the proposed NiNformer block with that of ViT.

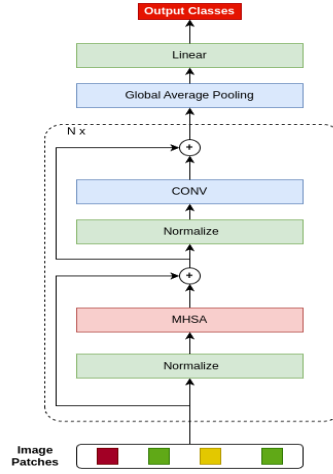




(a) NiNformer Architecture

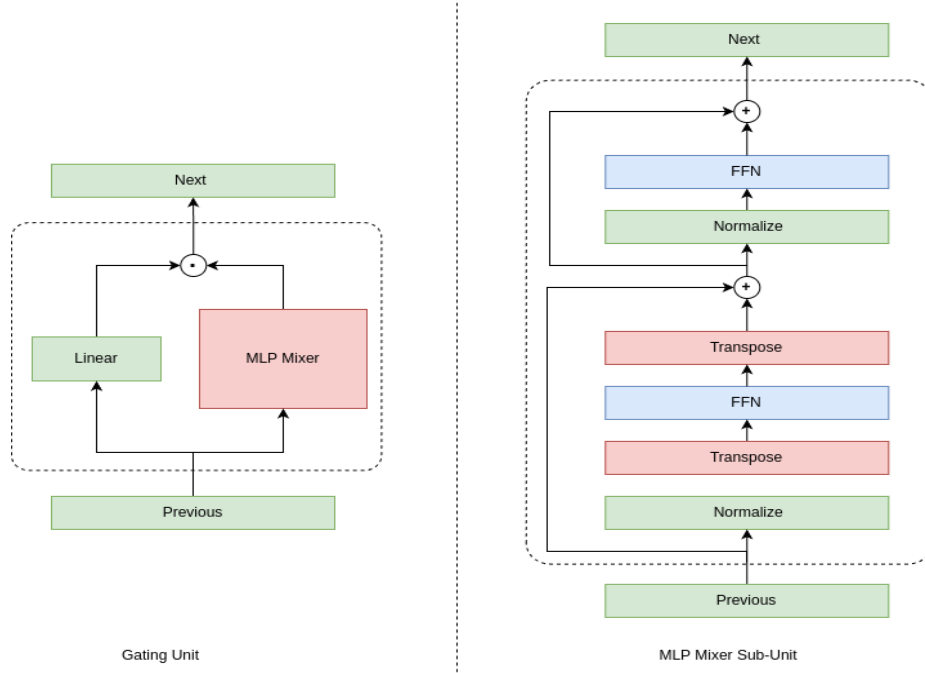


(b) ViT Architecture

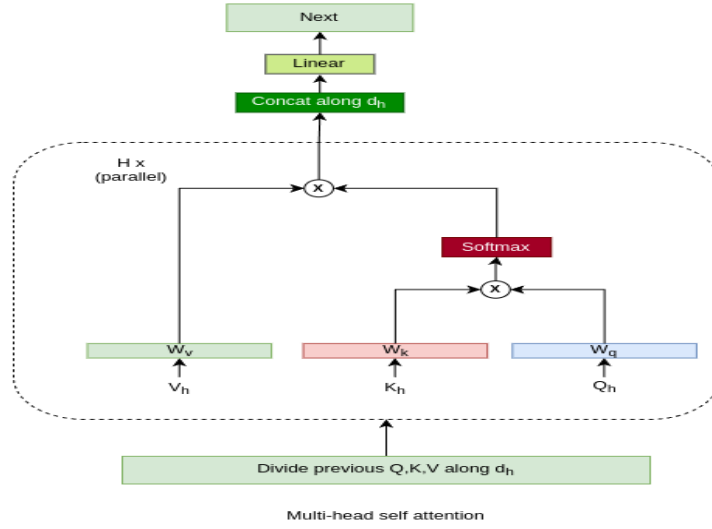


(c) Local-ViT Architecture

Fig. 1: Comparing NiNformer Architecture With ViT and Local-ViT



(a) NiNformer Gating-Unit and Mixer Sub-Unit



(b) Multi-Head Self Attention

Fig. 2: Comparing NiNformer Block with Attention Block

## 4 Experiments and Results

For the purposes of this work, three data sets have been selected: The CIFAR-10 [47] dataset consists of 60000 color images in 32 by 32 resolution provided for 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The CIFAR-100 [47] dataset consists of 60000 color images in 32 by 32 resolution; the number of classes is 100, resulting in 600 images per class. Similar to CIFAR-10, there are 50000 training images and 10000 test images. The MNIST [48] dataset consists of 70,000 grayscale images in 28 by 28 resolution. The number of classes is 10, as it is a dataset of handwritten numerical digits. There are 60000 training images and 10000 test images. The utilized software platform is Pytorch, version 1.13, supported by the CUDA toolkit, version 11.6.2. The version used for the Python programming language is 3.9. The available hardware system is a desktop personal computer with an Intel i9-9900k CPU, 32 Gigabytes of system RAM, an Nvidia RTX 2080ti GPU with 12 Gigabytes of RAM, and running an UBUNTU 20 LTS operating system. The implementation details of the selected Transformer architectures in this work are as follows: For the ViT architecture, the chosen patch size was 4 with a token dimension of 256, and the number of layers chosen was 4 with 4 Attention heads and an MLP dimension of 512. For the MLP-Mixer architecture, the chosen patch size was 4 with a token dimension of 256, and the number of layers chosen was 4 with a token-wise MLP dimension of 512 and a channel-wise MLP dimension of 512. For Local-ViT, the chosen patch size was 4 with a token dimension of 256, the number of layers chosen was 4, 4 Attention heads were selected, and the chosen channel dimension of the feedforward part was 512. For NiNformer, the chosen patch size was 4, the number of layers chosen was 4, the token dimension selected was 256, and the MLP dimension was 512 in the outer network. The inner sub-unit was designed with a token-wise MLP dimension of 512 and a channel-wise MLP dimension of 512. All models were fitted with a training loop comprised of 100 epochs with a batch size of 128. All experiments adopted the recommended learning rate for the Adam optimizer of 0.001 [49].

After performing the experimentation on the selected data sets applied to the baseline architectures and NiNformer architecture, the following results have been obtained:

Table 1: Experimentally obtained test accuracy metric on the utilized datasets

Dataset	Test accuracy (ViT) (%)	Test accuracy (MLP-Mixer)(%)	Test accuracy (LocalViT)(%)	Test accuracy (Ninformer) (%) ( <b>Ours</b> )
MNIST	97.12	97.73	97.79	<b>98.61</b>
CIFAR-10	65.74	70.12	77.71	<b>81.59</b>
CIFAR-100	34.87	39.16	41.61	<b>53.78</b>

Figures 3 and 4 show respectively, the accuracy and loss curves obtained on NiNformer for the CIFAR-10, CIFAR-100, and MNIST datasets.

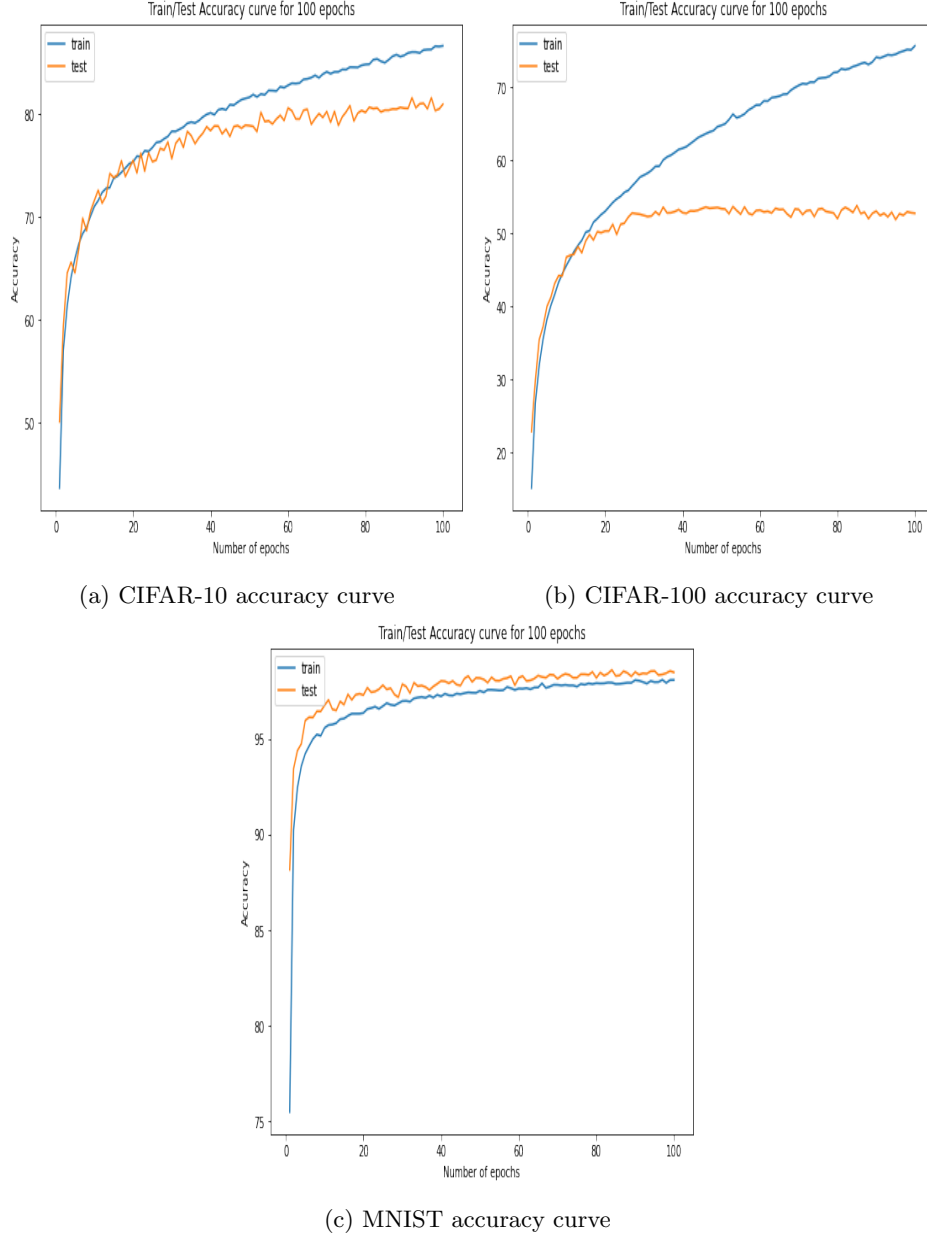


Fig. 3: Accuracy Curves for NiNformer Architecture

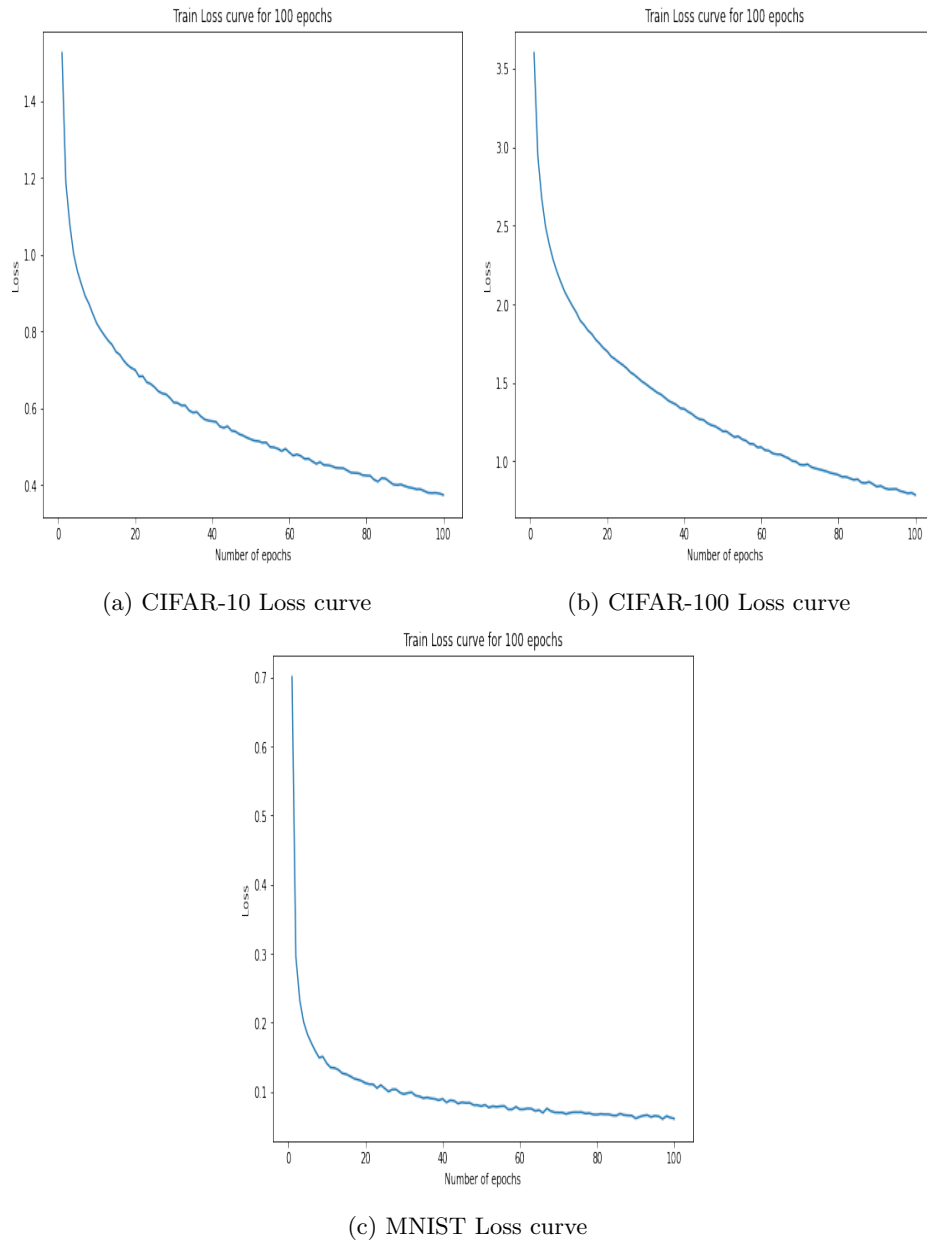


Fig. 4: Loss Curves for NiNformer Architecture

## 5 Conclusion

In this work, a newly designed Network in Network block was introduced that substitutes the Attention block used traditionally by Transformer architectures. By extending the token mixing approach presented in the MLP-Mixer to function as a gating signal generator and taking advantage of the gating mechanism to introduce dynamic behavior in the newly proposed block, thus enhancing the static weight approach of the MLP-Mixer by utilizing its layers as a sub-unit network within the outer network formulation, The experimental results show that in comparison to baseline architectures that were chosen from different families of Transformer formulations, our proposed block significantly outperforms, showing noticeable improvements on those baselines, specifically showing a great enhancement of accuracy compared to the standalone MLP-Mixer architecture that acts as a sub-unit, which validates that our proposal, with its dynamic gating of the upstream representation, properly enhances and circumvents the shortcoming of the static weight approach of the standalone MLP-Mixer while still providing more simplicity of operations in contrast to the vanilla ViT Transformer architecture. Future directions of this work are to investigate a multitude of sub-unit network selections, aiming for further enhancements and capabilities.

## References

1. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
2. Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. "Language Models are Few-Shot Learners." (2020).
3. Radford, Alec and Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training." (2018).
4. Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave and Guillaume Lample. "LLaMA: Open and Efficient Foundation Language Models." (2023).
5. Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei and Julien Launay. "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only." (2023).
6. Jiang, Albert Qiaochu, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix and William El Sayed. "Mistral 7B." (2023).

7. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” (2020).
8. Tolstikhin, Ilya O., Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic and Alexey Dosovitskiy. “MLP-Mixer: An all-MLP Architecture for Vision.” *Neural Information Processing Systems* (2021).
9. Trockman, Asher and J. Zico Kolter. “Patches Are All You Need?” *Trans. Mach. Learn. Res.* 2023 (2022).
10. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin and Baining Guo. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).
11. Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov and Sergey Zagoruyko. “End-to-End Object Detection with Transformers.” (2020).
12. Jaegle, Andrew, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Andrew Brock, Evan Shelhamer, Olivier J. H’enam, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals and João Carreira. “Perceiver IO: A General Architecture for Structured Inputs & Outputs.” (2021).
13. Lu, Jiasen, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi and Aniruddha Kembhavi. “Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks.” (2022).
14. Zhang, Hao, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming-shuan Ni and Heung-yeung Shum. “DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection.” (2022).
15. Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár and Ross B. Girshick. “Segment Anything.” *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).
16. Wang, Sinong, Belinda Z. Li, Madian Khabsa, Han Fang and Hao Ma. “Linformer: Self-Attention with Linear Complexity.” (2020).
17. Lee-Thorp, James, Joshua Ainslie, Ilya Eckstein and Santiago Ontañón. “FNet: Mixing Tokens with Fourier Transforms.” (2021).
18. Li, Yawei, K. Zhang, Jie Cao, Radu Timofte and Luc Van Gool. “LocalViT: Bringing Locality to Vision Transformers.” (2021).
19. Tu, Zhengzhong, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Conrad Bovik and Yinxiao Li. “MaxViT: Multi-Axis Vision Transformer.” *European Conference on Computer Vision* (2022).
20. Xiong, Yunyang, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Moo Fung, Yin Li and Vikas Singh. “Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention.” *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* 35 16 (2021).
21. Keles, Feyza Duman, Pruthuvi Maheshakya Wijewardena and Chinmay Hegde. “On The Computational Complexity of Self-Attention.” *International Conference on Algorithmic Learning Theory* (2022).
22. Lin, Tianyang, Yuxin Wang, Xiangyang Liu and Xipeng Qiu. “A Survey of Transformers.” *AI Open* 3 (2021).

23. Tay, Yi, Mostafa Dehghani, Dara Bahri and Donald Metzler. “Efficient Transformers: A Survey.” *ACM Computing Surveys* 55 (2020).
24. Fournier, Quentin, Gaétan Marceau Caron and Daniel Aloise. “A Practical Survey on Faster and Lighter Transformers.” *ACM Computing Surveys* 55 (2021).
25. Khan, Salman Hameed, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan and Mubarak Shah. “Transformers in Vision: A Survey.” *ACM Computing Surveys (CSUR)* 54 (2021).
26. Guo, Qipeng, Xipeng Qiu, Pengfei Liu, Yunfan Shao, X. Xue and Zheng Zhang. “Star-Transformer.” (2019).
27. Beltagy, Iz, Matthew E. Peters and Arman Cohan. “Longformer: The Long-Document Transformer.” (2020).
28. Kitaev, Nikita, Lukasz Kaiser and Anselm Levskaya. “Reformer: The Efficient Transformer.” (2020).
29. Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang and Amr Ahmed. “Big Bird: Transformers for Longer Sequences.” (2020).
30. Katharopoulos, Angelos, Apoorv Vyas, Nikolaos Pappas and Francois Fleuret. “Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention.” *International Conference on Machine Learning* (2020).
31. Choromanski, Krzysztof, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamás Szepesvári, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell and Adrian Weller. “Rethinking Attention with Performers.” (2020).
32. Chenguang Wang, Zihao Ye, Aston Zhang, Zheng Zhang, and Alexander J. Smola. 2020. “Transformer on a Diet.” (2020).
33. LI, SHIYANG, Xiaoyong Jin, Yao Xuan, Xiyu Zhou, Wenhui Chen, Yu-Xiang Wang and Xifeng Yan. “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting.” (2019).
34. Qiu, Jiezhong, Hao Ma, Omer Levy, Scott Yih, Sinong Wang and Jie Tang. “Block-wise Self-Attention for Long Document Understanding.” (2019).
35. Tay, Yi, Dara Bahri, Liu Yang, Donald Metzler and Da-Cheng Juan. “Sparse Sinkhorn Attention.” *International Conference on Machine Learning* (2020).
36. Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le and Ruslan Salakhutdinov. “Transformer-XL: Attentive Language Models beyond a Fixed-Length Context.” (2019).
37. Vyas, Apoorv, Angelos Katharopoulos and Francois Fleuret. “Fast Transformers with Clustered Attention.” (2020).
38. Zhang, Hang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan and Weizhu Chen. “Poolingformer: Long Document Modeling with Pooling Attention.” *International Conference on Machine Learning* (2021).
39. Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser and Noam M. Shazeer. “Generating Wikipedia by Summarizing Long Sequences.” (2018).
40. Dai, Zihang, Guokun Lai, Yiming Yang and Quoc V. Le. “Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing.” (2020).
41. Ho, Jonathan, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. “Axial attention in multidimensional transformers.” (2019).
42. Liu, Hanxiao, Zihang Dai, David So, and Quoc V. Le. “Pay attention to mpls.” *Advances in Neural Information Processing Systems* 34 (2021).



43. Tay, Yi, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao and Che Zheng. "Synthesizer: Rethinking Self-Attention for Transformer Models." International Conference on Machine Learning (2020).
44. Han, Kai, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. "Transformer in transformer." Advances in Neural Information Processing Systems 34 (2021).
45. De, Soham, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando de Freitas and Caglar Gulcehre. "Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models." (2024).
46. Dauphin, Yann, Angela Fan, Michael Auli and David Grangier. "Language Modeling with Gated Convolutional Networks." International Conference on Machine Learning (2016).
47. Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton. "Cifar-10 and cifar-100 datasets." URL: <https://www.cs.toronto.edu/kriz/cifar.html> 6, no. 1 (2009)
48. LeCun, Yann, Léon Bottou, Yoshua Bengio and Patrick Haffner. "Gradient-based learning applied to document recognition." Proc. IEEE 86 (1998)
49. Steiner, Andreas, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. "How to train your vit? data, augmentation, and regularization in vision transformers (2021).