Project Report - ECE 176

Abhishikta Panja
ECE
A69027311

Geetika Agrawal
ECE
A59026563

Abstract

Our project explores the potential of integrating self-attention mechanisms into neural network models for image recognition, focusing on two distinct types: pairwise and patchwise self-attention. We develop a neural network model that leverages self-attention that significantly improves model performance to achieve higher accuracy in image recognition tasks, while also being less computationally demanding compared to a baseline model (SAN10[1]).

1 Introduction

In the dynamic field of computer vision, the continuous pursuit for models with an enhanced understanding of complex image data has prompted a shift towards novel methodologies beyond the traditional bounds of convolutional neural networks (ConvNets). This project focuses on the adoption and integration of self-attention mechanisms—pairwise and patchwise self-attention—into neural network architectures tailored for image recognition. By leveraging self-attention's ability to dynamically prioritize different segments of input data, we aim to unlock new levels of detail and context in image analysis. Patchwise self-attention, on the other hand, takes a broader approach than pairwise self-attention. It enables the model to analyze the image in segments or patches, understanding each patch's context and how it contributes to the overall image. This holistic view is beneficial for recognizing objects and their surroundings within an image, facilitating a more comprehensive analysis than what is possible through focusing on individual pixels or localized feature pairs.

These innovative approaches are designed not only to enhance the precision of image recognition tasks but also to optimize computational efficiency. Through this dual focus, the project endeavors to achieve a balance between high performance and reduced computational demand.

To validate the efficacy of these self-attention mechanisms, we integrate them into various ConvNet architectures and undertake a rigorous evaluation using the CIFAR-10 dataset. We propose a neural network architecture hat minimizes the number of convolutional layers, thereby enhancing computational efficiency and achieving higher accuracy relative to our baseline model, the SAN10[1].

2 Related Work

Our research primarily intersects with recent advancements in applying self-attention mechanisms to computer vision, significantly diverging from the traditional reliance on convolutional networks (ConvNets). Among the numerous contributions to this field, the works of Hu et al. and Ramachandran et al. stand out for their innovative approach to self-attention, specifically their exploration of localized self-attention, which is directly relevant to our project.

Hu et al[3]. introduced a novel method that confines the self-attention mechanism to small, localized patches within an image (e.g., 7×7 pixel areas). This approach significantly reduces the computational and memory overhead typically associated with global self-attention, making it more feasible for

application in high-resolution layers and throughout the network. Their work demonstrated that such localized attention could effectively capture relevant features within a constrained area, maintaining or even enhancing the model's performance on various tasks.

Ramachandran et al[2]. expanded on this concept by exploring different self-attention formulations that could be applied more broadly and efficiently across the network. They experimented with self-attention mechanisms that adapt to different channels, offering a more nuanced approach to feature extraction and processing. This adaptation allows the model to focus on more relevant features in each channel, potentially improving accuracy and efficiency.

Our project builds upon these foundational studies by exploring an even broader variety of self-attention mechanisms, including vector attention, which allows for channel-specific adaptations. We also introduce a new family of patchwise attention operators that serve as a more flexible and efficient alternative to conventional convolutional layers. By integrating these advanced self-attention techniques, our project aims to push the boundaries of what's possible in computer vision, seeking to achieve higher performance with more efficient computational resource use. These related studies provide a crucial context for our innovations, highlighting the potential of localized and adapted self-attention mechanisms to transform the field.

3 Method

3.1 Pairwise Self-attention

Our approach explores two distinct types of self-attention mechanisms. The first type is pairwise self-attention, which is formulated as:

$$y_i = \sum_{j \in R(i)} \alpha(x_i, x_j) \odot \beta(x_j), \tag{1}$$

where $\odot$ denotes the Hadamard product, i is the spatial index of the feature vector x_i, $R(i)$ is the local footprint location in the feature map, and α represents the adaptive weight vectors derived through the relation function δ, which accounts for the interaction of feature vector pairs. The transformation β produces the transformed feature vectors which are aggregated to construct the new feature y_i.

The adaptive weight vectors $\alpha(x_i, x_j)$ are given by:

$$\alpha(x_i, x_j) = \gamma(\delta(x_i, x_j)) \tag{2}$$

The function γ acts on the output of the relation function δ that encapsulates the interaction of x_i and x_j and converts it to a scalar weight that can be utilized in combination with the transformed features $\beta(x_j)$.

Position Encoding : Position encoding plays a crucial role by processing vector pairs (x_i, x_j) independently from their spatial locations. This independence allows for the flexibility of positional context incorporation. A linear layer is used to transform normalized two-dimensional coordinates through a triangle linear layer, which then passes through a nonlinearity to produce the final position feature ϕ. The pairwise self-attention mechanism is enhanced by this position encoding:

$$P_i - P_j \to \gamma', \tag{3}$$

where γ' is the mapping that encodes the relative position information.

3.2 Patchwise Self-attention

The second type of self-attention we explore is patchwise self-attention, which operates on a broader scale:

$$y_i = \sum_{k \in R(i)} \alpha(x_{R(i)}) \odot \beta(x_k), \tag{4}$$

where $x_{R(i)}$ represents the aggregate of feature vectors within $R(i)$. The relation function δ and transformation β are similarly applied, albeit at a patch level. This is expressed as:

$$\alpha(x_{R(i)}) = \gamma(\delta(x_{R(i)})). \tag{5}$$

The relation function δ comprises several operations including summation, subtraction, concatenation, Hadamard product, and dot product, which provide versatility in capturing various types of interactions between feature vectors.

Function δ and Scalar Weights : Our methodology includes various operations within the relation function δ:

- Subtraction: $\delta_{sub}(x_i, x_j) = \phi(x_i) - \phi(x_j)$
- Hadamard product: $\delta_{had}(x_i, x_j) = \phi(x_i) \odot \phi(x_j)$

These operations are chosen to maintain dimensional consistency and to ensure that the transformation $\delta(x_i, x_j)$ preserves the rich information present in the interaction of feature pairs.

3.3 Baseline Model

The baseline model for CIFAR-10 image classification features an initial convolutional layer followed by batch normalization and ReLU activation. It utilizes max pooling to reduce dimensionality and integrates self-attention blocks (Figure 3 to capture global dependencies. After repeating a core set of layers, the architecture concludes with average pooling and a linear layer for final classification outputs. Figure 1 shows the architecture of our baseline model.

In the Self-Attention network (SAN) architecture, a Bottleneck layer is added to augment the existing convolution layers by performing more than simple feature transformation. It incorporates self-attention mechanisms through a Self-Attention Module (SAM) module, which allows the network to either perform a Pairwise or a Patchwise self-attention operation. The Bottleneck layer also helps in reducing the dimensionality of the feature maps in a controlled manner, using 1x1 convolutions. This keeps the computational load in check while still maintaining the richness of the feature representation. The residual connections in the Bottleneck layers further help in preventing the vanishing gradient problem, thereby allowing the network to benefit from deeper architectures.

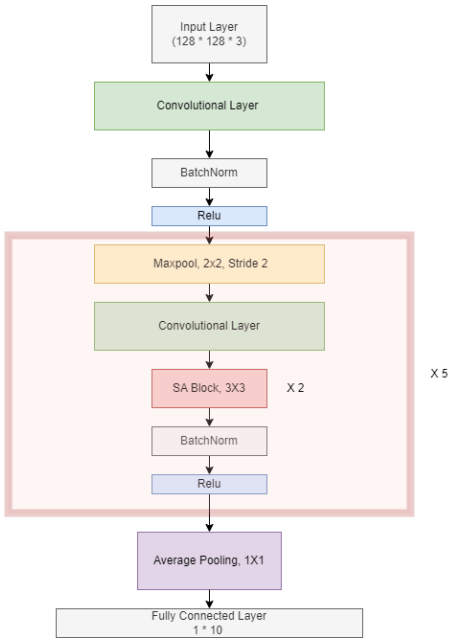


Figure 1: Baseline Model

3.4 Proposed Model

The new model depicted in Figure 2, utilizes a self-attention mechanism extensively within its architecture to improve image classification on the CIFAR-10 dataset. Starting with an input layer and an initial convolutional layer followed by batch normalization, the model introduces a consecutive

sequence of self-attention (SA) blocks that vary in layer depth and kernel sizes. These SA blocks are designed to process spatial relationships within the image data more effectively than traditional convolutional layers alone. Moreover, we have added a dropping rate into the Bottleneck layer of the model regularization effect, mitigating the risk of overfitting by randomly omitting units during training. Following the self-attention blocks, a ReLU activation function and average pooling are applied, culminating in a linear layer that maps the extracted features to the number of classes in the CIFAR-10 dataset for classification.

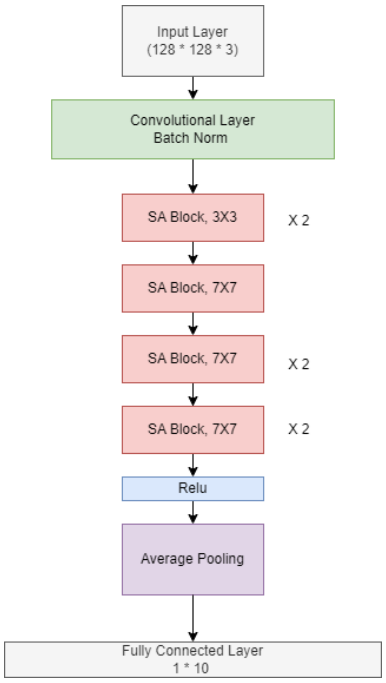
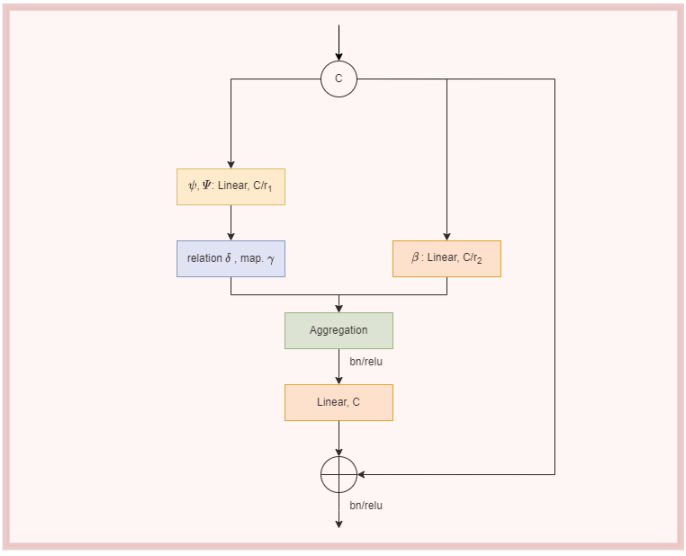


Figure 2: Proposed Model



δ

Figure 3: SA Block

Self-Attention (SA) (in Figure 3) refines feature processing by bifurcating the input tensor with channel dimensionality C into two pathways. One pathway computes attention weights α via the relations δ, ϕ, and ψ, subsequently passed through mapping γ. In parallel, the alternate pathway applies a linear transformation β that modifies and reduces the input features for efficient computation. The outputs of both pathways are amalgamated using the Hadamard product, followed by normalization and a ReLU nonlinearity. Finally, a subsequent linear layer reinstates the channel dimensionality back to C, resulting in an enriched, attention-enhanced feature representation.

4 Experiments

4.1 Dataset and Data Preprocessing

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another.

4.2 Data Preprocessing

For our test data, we employed a comprehensive data augmentation and preprocessing pipeline to enhance the generalization capabilities of our model for CIFAR-10 image classification. The pipeline initiates with a random crop and padding strategy to introduce translational variations, followed by a random horizontal flip to simulate the natural orientation diversity present in real-world scenarios. Further diversity in visual appearance is injected through color jittering, adjusting the brightness, contrast, and saturation levels, and random affine transformations, including slight rotations, translations, and scalings. These images are then converted to tensors and normalized with channel-wise mean and standard deviation values of 0.5.

4.3 Implementation

In our implementation setup, both the baseline and the proposed model were trained on the CIFAR dataset, utilizing two distinct types of self-attention mechanisms: pairwise and patchwise. Each model underwent a training regimen spanning 250 epochs, with multiprocessing techniques leveraged to expedite the training process on GPU hardware. We randomized the order of the training dataset and grouped the samples into batches of 512. This methodology guarantees that decisions regarding architecture and hyperparameters are independently made from the dataset used for baseline comparisons, ensuring a fair evaluation framework.

5 Results and Analysis

The outcomes for both the baseline and the newly proposed model are presented in Table 1. It is evident from the data that the proposed model significantly outperforms the baseline in terms of accuracy, across both pairwise and patchwise self-attention mechanisms.

Self Attention Type	Baseline Model	Proposed Model
Pairwise	53.58%	64.22%
Patchwise	59.73%	75.24%

Table 1: Accuracy Results

Patchwise self-attention typically outperforms pairwise due to its ability to capture broader contextual information within image segments, providing a more holistic understanding of spatial relationships. By aggregating features over larger patches, it efficiently encodes both local and global image structures, enhancing pattern recognition. Our proposed model achieves a higher accuracy for both pairwise and patchwise self-attention types. We observe that stacking multiple self-attention layers,

as opposed to alternating them with convolutional layers, results in a model that not only is less demanding in terms of computation but also presents us with a higher accuracy. Possible reasons for this improvement include:

- **Enhanced Feature Interaction**: Consecutive self-attention layers may allow for more complex interactions between features at different positions within the image, potentially leading to a richer hierarchical representation of the input data.

- **Improved Contextual Information**: By focusing solely on self-attention, the network may become more adept at integrating contextual information across the entire image, rather than being limited to the receptive field of convolutional layers.

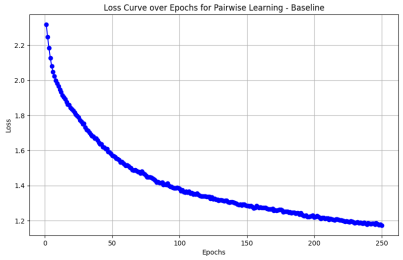
- **Depth Efficiency**: Self-attention can process all combinations of feature interactions regardless of their position, which can make deeper networks more parameter-efficient, especially if attention is computationally cheaper than the convolutions being replaced.

- **Model Capacity and Flexibility**: The new model may have increased capacity and flexibility in learning spatial hierarchies due to the dedicated stacking of self-attention layers, which could allow the network to adapt more effectively to the complexity of the dataset.

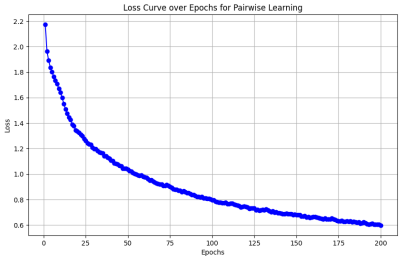
Inference from Loss Curves :

- **1) Baseline vs. Proposed Pairwise Learning:** The loss curve for the baseline pairwise learning model starts at a higher loss and a gradual decline, settling at a modest rate as it approaches the later epochs. In contrast, the proposed pairwise model initiates learning with a substantially lower initial loss. Throughout the training, it maintains a steep descent, converging to a notably lower loss than the baseline. This rapid reduction of loss suggests that the proposed model effectively captures features with fewer computational resources and iterations, implying a leaner and more computationally economical architecture that does not sacrifice learning depth or quality.

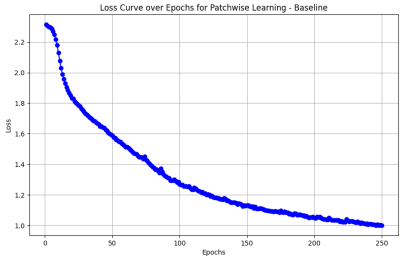
- **2) Baseline vs. Proposed Patchwise Learning:** For patchwise learning, the baseline model's loss curve indicates a traditional trajectory with an initial sharp drop. The proposed patchwise model, however, reveals a starkly steeper loss descent early in training, achieving a lower loss in fewer epochs. This quick convergence shows the proposed model's capacity to utilize information efficiently, translating into superior performance with less computational overhead.



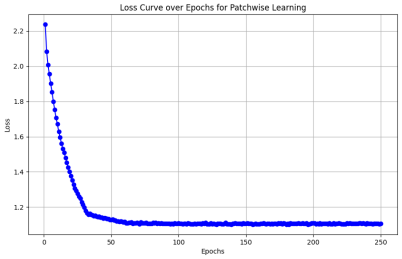
(a) Loss Curve for Pairwise SA (Baseline)



(b) Loss Curve for Pairwise SA (Proposed Model)



(c) Loss Curve for Patchwise SA (Baseline)



(d) Loss Curve for Patchwise SA (Proposed Model)

Figure 4: Loss Curves for Baseline and Proposed Models while Training

6 Conclusion

In conclusion, our project has successfully demonstrated the significant potential of integrating self-attention mechanisms, specifically pairwise and patchwise self-attention, into neural network models for image recognition tasks. By designing a neural architecture that emphasizes these self-attention blocks over traditional convolutional layers, we not only achieved a model that is computationally more efficient but also outperforms the baseline SAN10[1] model in terms of accuracy. The experimental results on the CIFAR-10 dataset clearly illustrate the superiority of our proposed model, showcasing enhanced feature interaction and a deeper understanding of contextual information within images.

7 Supplementary Material

1. Video Model : [**Video Link**]

2. Github Code : [**Code**]

3. Slides : [**Slides**]

8 References

[1] Zhao, H., Jia, J., Koltun, V. (Year). Exploring Self-attention for Image Recognition. In IEEE, 2020

[2] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In NeurIPS, 2019

[3]Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In ICCV, 2019.