A Novel Attention-Based Approach for Image Classification Using Pre-trained CNN model

1st Sriram Mandalika
Department of Computational Intelligence
SRM Institute of Science and Technology, Kattankulathur
Chennai, India
mc9991@srmist.edu.in

***Abstract*—This paper proposes a new approach to image classification by adding a custom attention module to a ResNet-18 backbone. We aim to improve accuracy and efficiency in image classification tasks with advanced attention mechanisms. Our work uniquely presents combining hierarchical attention, contextual relevance, temporal memory and dynamic neighbourhood adaptation in one attention module. This boosts the feature representation capability of the baseline ResNet-18 model to capture intricate patterns and spatial hierarchies in the image data. We used a pre-trained ResNet-18 to extract feature maps, from the *ImageNet* weights and then passed them through our proposed attention module. The relevance scores were calculated based on different attention mechanisms and enhanced the feature maps for classification tasks. We experimented on the CIFAR-100 and Stanford Cars datasets with data augmentation and normalization to improve model robustness and performance. The results show significant improvement in classification performance with the proposed attention module compared to the unsupervised baseline ResNet-18 performance. Specifically, the model with an attention module has higher training and test accuracy and lower loss. This shows more efficient learning and better generalization capabilities. Our approach not only improves the model accuracy but also provides a scalable solution for other computer vision tasks. Future work can extend this to other datasets and domains and potentially have a broader impact in fields that require detailed image analysis and classification. All the codes are available here:**

***Index Terms*—Spatial Features, Convolutional Neural Network (CNN), Image Classification, Feature Representation**

I. Introduction

Despite their emergence in recent years as state-of-the-art techniques for spatial attention mechanisms for computer vision applications, there have been multiple challenges to address: unstable training, high computational demand, as well as the need for large datasets to achieve competitive results [1]. Recent works have shown that training the CNN models on limited data or small datasets [2]–[6] can mitigate some of the mentioned issues for various vision-based tasks.

However, integrating attention mechanisms into CNNs for image classification remains challenging, particularly with limited data. Existing approaches often require extensive computational resources or fail to generalize well across different datasets [7], [8]. Researchers have worked towards these challenges and enhanced the models' applicability across diverse application scenarios.

To address this issue, a few other localization techniques were also utilized for the model to focus on a region of interest like GradCAM, *DINO*, TokenCut, *etc.,* [9]–[13].

This paper introduces a unified framework for dynamic spatial attention mechanism that can be applied to various computer vision applications commonly for image classification and object detection tasks. Our attention module combines hierarchical attention, contextual relevance, temporal memory, and dynamic neighbourhood adaptation to create a robust mechanism for focusing on significant features.

The main contributions of this paper are as follows:

- We design a custom attention module that effectively captures and integrates different types of attention schemes into a single framework.
- We demonstrate the integration of this module with the ResNet-18 model, showing significant improvements in classification accuracy and reduction in training loss.
- We provide a comprehensive analysis of the feature maps generated by our model, illustrating the enhanced focus on relevant image regions.
- We validate our approach on the CIFAR-100, Stanford cars and CIFAR-10 datasets, achieving a notable increase in both training and test accuracy compared to the baseline ResNet-18 model.

II. Related Works

Our work builds upon recent advances in attention modelling, as well as limited data pre-training.

Limited Data Pre-Training: Training deep learning models like CNN have first seen success in foundational computer vision tasks like image classification was presented in Krizhevsky *et al.* [14] discussing upon techniques for handling limited data through data augmentation and dropout to prevent overfitting. Real-time applications for using limited data have been practiced for a few reasons predominantly if that domain has very little labeled data such as remote sensing and medical imaging. In the case of medical imaging, creating vast datasets of human metabolic activity comes with various regulatory restrictions and privacy concerns about that data being misused.

Recent works on this, systematically review multiple data augmentation methods in medical imaging, including their detailed application in improving image classification

accuracy with limited training data using the CNN model [15], [16]. Understanding the demand in a few niche domains where there is little labelled data for training, researchers have also previously used Zero-shot, Single-shot and Few-shot learning methodologies involving very little data for training the CNN model as effectively and efficiently as possible. Palatucci *et al.* [17] were among earlier works presented that introduce the conceptual possibility of training a deep learning model without using any training samples, which laid the groundwork for further research on exemplar limited training of CNN model that was discussed in papers [18]–[23]. In contrast, our proposed mechanism is a unified framework for a spatial attention module that leverages the combined characteristics of solving computer vision problems.

Spatial Attention Mechanisms Attention modules have been around for a long time now, they were introduced by Xu, Ke *et al*, which discusses about using an RNN backbone performing image classification like tasks and a CNN model as a feature extractor. Where they initially conceptualized hard attention and soft attention [24]. This work was built upon the experimentations of [25]–[28] which were among the earliest works presented on visual descriptors and attention mechanisms. Recent works on strategy were proposed by Kakogeorgiou et al. [29], which integrates into a student-teacher knowledge distillation training process where the teacher network outputs an attention map that is used to mask the input image for the student network [30]. They had extended their experimentation on the masking approach to keep some of the attention patch maps visible, giving the model additional input that resembles "hints".

III. Methodology

This section introduces our proposed unified framework dealing with a self-adaptive attention mechanism based on probability distribution in the input data. Subsequently, we present our attention method.

A. Data Pre-processing

For all the datasets used, we employed several pre-processing methodologies to enhance model performance and generalization ability. We utilized random cropping and padding to generate *(32 × 32)* patches from the original images, which introduced variability and helped the model learn to generalize better on unseen images. Additionally, we applied random horizontal flipping with a probability of 50%, further augmenting the dataset by providing different orientations of the same image. Appropriate pre-processing techniques contribute to positive model training as limited images are used in every scenario.

Another critical component, data augmentation and batch normalisation played an important role in our training pipeline. input features are on a similar scale, which accelerates convergence during training. Additionally, we employed data augmentation techniques such as random rotations, scaling, and colour jittering to artificially expand the dataset's variability, enhancing the model's ability to generalize. Also, learning rate scheduling was used to further optimize the training loops/process, which involved adjusting the learning rate during the training phase to ensure better convergence and resulting in improved generalisation.

B. Deep Feature extraction using the pre-trained learning architecture

We employed ResNet-18 as the backbone model due to its optimal balance between complexity, performance and lightweight nature. Utilizing pre-trained weights on *ImageNet*, we leverage transfer learning to enhance initial performance and speed up convergence. The architecture of ResNet-18 was modified by removing the fully connected layers and repurposing it as a feature extractor that outputs feature maps. These feature maps, capture complex patterns and representations that serve as the foundation for subsequent processes and provide rich, detailed information for further processing. The robustness of ResNet-18 makes it adaptable to various datasets beyond ImageNet, ensuring broad applicability. Justifying our choice of use ResNet-18 is based on its superior performance and computational efficiency for this particular case.

The process of feature map generation in ResNet-18 involves transforming an input image through initial convolutional layers and residual blocks. This transformation captures various levels of spatial and semantic information. The following sections provide a mathematical representation of this process.

1) Initial Convolution and Activation: Given an input image X with dimensions $H \times W \times C$ (height, width, and number of channels), the initial convolutional layer applies a set of filters to produce the first set of feature maps. Let W_1 denote this initial set of filters.

$$F_1 = W_1 * X \tag{1}$$

where $*$ denotes the convolution operation, producing feature maps F_1 with dimensions $H' \times W' \times C_1$. The ReLU activation function is then applied:

$$F_1' = \text{ReLU}(F_1) = \max(0, F_1) \tag{2}$$

2) Residual Block: A residual block contains two convolutional layers and a skip connection. Let W_2 and W_3 represent the filters in the first and second convolutional layers of a block, respectively. For an input X to the block, the output $H(X)$ is given by:

$$F_2 = W_2 * X \tag{3}$$

$$F_2' = \text{ReLU}(F_2) \tag{4}$$

$$F_3 = W_3 * F_2' \tag{5}$$

The skip connection is added to the output of the second convolution:

$$H(X) = F_3 + X \tag{6}$$

3) Pooling Layer: Pooling layers reduce the spatial dimensions of the feature maps. For max pooling:

$$P_{i,j} = \max_{m,n} F_{i+m,j+n} \tag{7}$$

where P represents the pooled feature map, and (i, j) are spatial indices.

4) Global Average Pooling: At the end of the network, global average pooling is applied to the final feature maps to reduce each $H \times W \times C$ tensor to a $1 \times 1 \times C$ tensor:

$$G_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,c} \tag{8}$$

where G_c is the c-th channel of the globally averaged feature vector.

5) Full Forward Pass: Combining these operations, the forward pass through ResNet-18 can be summarized as follows:

Step 1. Initial convolution and ReLU activation:

$$F_1' = \text{ReLU}(W_1 * X) \tag{9}$$

Step 2. Passing through each residual block:

$$H_1(X) = \text{Block1}(F_1') = \text{ReLU}(W_3 * \text{ReLU}(W_2 * F_1')) + F_1' \tag{10}$$

Step 3. Continuing through the network with more residual blocks.

Step 4. Applying global average pooling to the final feature maps:

$$G = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} H_{i,j,c} \tag{11}$$

The feature maps generated from ResNet-18 exhibit several key characteristics, they are as follows:

- The initial layers capture low-level features such as edges and textures, while deeper layers capture high-level semantic features.
- As the network's depth increases, the feature maps' spatial resolution decreases, leading to more abstract representations.
- Despite the reduction in spatial resolution, the deeper feature maps maintain information about the spatial locations of significant features.
- The feature maps in the final layers provide a robust representation for distinguishing between different classes, essential for tasks like classification.
- Through pooling and convolution operations, the network reduces redundancy, focusing on the most critical features for the given task.
- The feature maps incorporate contextual information from surrounding regions due to the receptive field expansion as the depth of the network increases.

C. Self-Adpating hierarchal spatial attention module

Our proposed attention module integrates several sophisticated mechanisms to enhance the model's ability to focus on the most informative regions of interest in the input feature maps to make appropriate intelligent decisions. This module begins with a hierarchal analysis, which operates at various scales on the feature maps. Initially, the input feature maps are processed through two *2D/3D* convolutional layers that progressively down-sample them, capturing multi-scale information. Sigmoid activations are applied to generate attention weights at each stage, which are then concatenated to form a comprehensive attention map that highlights features at varying granularity. Following this, the module undergoes contextual relevance analysis where the processing data is assessed using a single *2D/3D* convolutional layer that outputs a relevance score for each spatial location, emphasizing regions that are contextually significant for the task. This step ensures that the module not only focuses on prominent features but also considers their contextual importance within the entire feature map.

The relevance score is calculated and normalized between 0 and 1, let's denote the input feature map to the contextual relevance module as $\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. Here, C is the number of channels, H is the height, and W is the width of the feature map. As the contextual relevance is calculated using a convolutional layer with a single output channel. Let us consider $\mathbf{W}$ be the convolutional kernel and b be the bias term. The convolution operation can be represented as:

$$\mathbf{R}_{ij} = \sum_{c=1}^{C} \sum_{m=-k}^{k} \sum_{n=-k}^{n} \mathbf{W}_{c,m,n} \cdot \mathbf{X}_{c,i+m,j+n} + b \tag{12}$$

where:

- $\mathbf{R} \in \mathbb{R}^{1 \times H \times W}$ is the resulting relevance score map.
- k is the half-size of the convolutional kernel (e.g., for a 3×3 kernel, $k = 1$).
- i and j iterate over the spatial dimensions of the feature map H and W.

After computing the convolution, the relevance scores $\mathbf{R}$ are passed through a sigmoid activation function to normalize the scores:

$$\mathbf{R}_{\text{norm}} = \sigma(\mathbf{R}) = \frac{1}{1 + e^{-\mathbf{R}}} \tag{13}$$

Here, $\sigma(\mathbf{R})$ denotes the sigmoid function, which ensures that each value in $\mathbf{R}_{\text{norm}}$ lies within the interval $[0, 1]$.

The dynamic neighbourhood adaptation mechanism in our proposed module assesses the variation within local neighbourhoods of the feature maps. By applying a convolution followed by standard deviation computation, the module identifies regions with significant local variations. A sigmoid function is then applied to generate dynamic adaptation weights that highlight these regions.

Finally, the outputs from all the mentioned sub-parts of the proposed attention module are concatenated to form a

combined attention score. This score is processed through a final convolutional layer to produce the ultimate attention map. This attention map is then applied to the original feature maps, effectively weighting them according to the learned attention, which enhances the model's focus on the most relevant features while suppressing less important ones.

The proposed attention module allowed our model to dynamically adapt its focus, improving performance on general-purpose computer vision tasks such as image classification.

IV. Experimental Results

In this section, we present the experimental results obtained from our proposed attention module integrated with the pre-trained ResNet-18 backbone model that was trained on ImageNet [31] weights for the initial run. Our attention maps that are generated were experimented on commonly available benchmark datasets like CIFAR-100 [14], Stanford Cars dataset [32] and CIFAR-10 dataset [14].

A. Performance Evaluation

We evaluate the model's performance through various metrics, including accuracy, loss value, and computational efficiency, comparing it to the baseline model with and without the attention mechanism. The CIFAR-100 dataset [14] that was considered consists of 100 classes, out of which for all the experimentations a subset of 10 images per class formulating the size of the training images of 1000 images and 5 images from every class were taken eventually forming a validation subset of 500 image subset for the validation process.

TABLE I: Results with the proposed spatial attention module on CIFAR-100 dataset

Benchmarks	**Avg. Accuracy**	**Avg. Loss**	**Epochs**
ResNet-18 Baseline model	78.5%	0.54	100
+ Proposed Attention Module	83.2%	0.42	100

B. Attention Mechanisms

Review recent advances in attention mechanisms, including self-attention, spatial attention, and channel attention [22]. [23] [] []

References

[1] M. H. M. Noor and A. O. Ige, "A survey on deep learning and state-of-the-art applications," *ArXiv*, vol. abs/2403.17561, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268691608

[2] M. Ariaei, H. Ghassemian, and M. Imani, "Highresolution remote sensing image classification with limited training data," *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, pp. 1–5, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269035318

[3] J.-H. Jeong, Y.-H. Song, I.-U. Kang, and J. yeol Ryu, "Addressing data scarcity and overcoming limitations of lightweight models for unmanned military image classification ai model using knowledge distillation," *Journal of the Korea Academia-Industrial cooperation Society*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269018999

[4] S. Piffer, L. Ubaldi, S. Tangaro, A. Retico, and C. Talamonti, "Tackling the small data problem in medical image classification with artificial intelligence: a systematic review," *Progress in Biomedical Engineering*, vol. 6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270171339

[5] Y. Ren, P. Jin, Y. Li, and K. Mao, "An efficient hyperspectral image classification method for limited training data," *IET Image Process.*, vol. 17, pp. 1709–1717, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256836973

[6] A. Gao, "More for less: Compact convolutional transformers enable robust medical image classification with limited data," *ArXiv*, vol. abs/2307.00213, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259317205

[7] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Zhang, Y. Lin, Y. Sun, T. He, R. M. Mueller, M. Li *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[8] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[9] S. Park and C. Wallraven, "Comparing facial expression recognition in humans and machines: Using cam, gradcam, and extremal perturbation," *ArXiv*, vol. abs/2110.04481, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238582701

[10] F. Li, H. Zhang, H.-S. Xu, S. Liu, L. Zhang, L. M. shuan Ni, and H. yeung Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3041–3050, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249395535

[11] C. Lu, H. Zhu, and P. Koniusz, "From saliency to dino: Saliency-guided vision transformer for few-shot keypoint detection," *ArXiv*, vol. abs/2304.03140, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257985125

[12] A. Wollek, R. Graf, S. eatka, N. Fink, T. Willem, B. O. Sabel, and T. Lasser, "Attention-based saliency maps improve interpretability of pneumothorax classification," *Radiology. Artificial intelligence*, vol. 5 2, p. e220187, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257289817

[13] Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley, and D. Vaufreydaz, "Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 15 790–15 801, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251979706

[14] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18268744

[15] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, p. 100340, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277244252400042X

[16] N. Ghasemi, J. Á. Justo, M. Celesti, L. Despoisse, and J. Nieke, "Onboard processing of hyperspectral imagery: Deep learning advancements, methodologies, challenges, and emerging trends," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269033109

[17] M. Palatucci, D. A. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Neural Information Processing Systems*, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:7490338

[18] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4166–4174, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:446581

[19] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:309759

[20] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[21] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Neural*

Information Processing Systems, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8909022

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. Kang, Y. Zhang, X. Liu, and Z. Cheng, "Hyperspectral image classification using spectral–spatial double-branch attention mechanism," *Remote Sensing*, vol. 16, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/1/193

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:1055111

[25] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1473–1482, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:9254582

[26] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *ArXiv*, vol. abs/1411.2539, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:7732372

[27] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv: Computer Vision and Pattern Recognition*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:3509328

[28] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:9026666

[29] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247627906

[30] L. Sick, D. Engel, P. Hermosilla, and T. Ropinski, "Attention-guided masked autoencoders for learning image representations," *ArXiv*, vol. abs/2402.15172, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267897808

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:57246310

[32] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:14342571