**Enhancing Image Classification Through Non-Local Neural Networks**

**Introduction**

In recent years, deep learning advancements have significantly improved computer vision, particularly in image classification. Despite the success of convolutional neural networks (CNNs), their limitations in capturing long-range dependencies and global context hinder their efficacy in discerning intricate patterns, especially in tasks like facial emotion detection. This project explores the application of CNNs along with non-local neural networks (built from scratch based on well-established frameworks) to address these challenges in image classification. Non-local neural networks, inspired by self-attention mechanisms, provide a unique solution by allowing pixels to interact globally across spatial dimensions. The study focuses on emotion detection using facial features, leveraging the Facial Expression Recognition (FER 2013) dataset. Motivated by the need for improved sensitivity in emotion recognition, this research aims to evaluate the effectiveness of non-local neural networks in enhancing image classification precision, emphasizing their potential in the domain of facial emotion recognition.

Note: It is important to clarify that this work is designed to solely demonstrate the efficacy of non-local neural networks and does not seek to achieve state-of-the-art results.

**Framework and Tools**

* Deep Learning Framework: PyTorch
* Environment: Jupyter Notebook
* Base CNN Architecture: ResNet18
* Additional Architecture: Non-Local Neural Networks

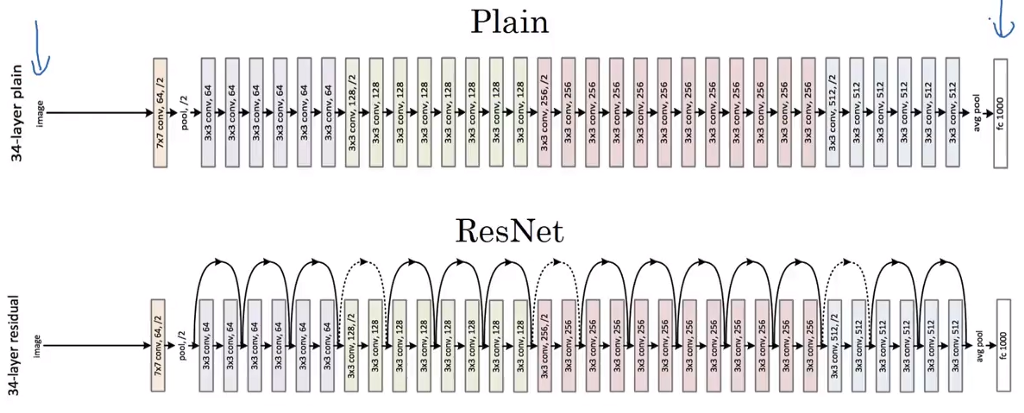
**About the dataset**

The FER 2013 dataset is a fundamental resource in the field of emotion recognition from facial expressions. It consists of 35,887 grayscale images of resolution 48 x 48 pixels, each labelled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, or a neutral state. The dataset provides a diverse set of annotated facial expressions, drawn from various sources such as movies, online databases, and natural settings. These annotations serve as a valuable ground truth for training and evaluating models. Overall, the FER 2013 dataset is widely used in research to develop and benchmark facial emotion recognition algorithms.

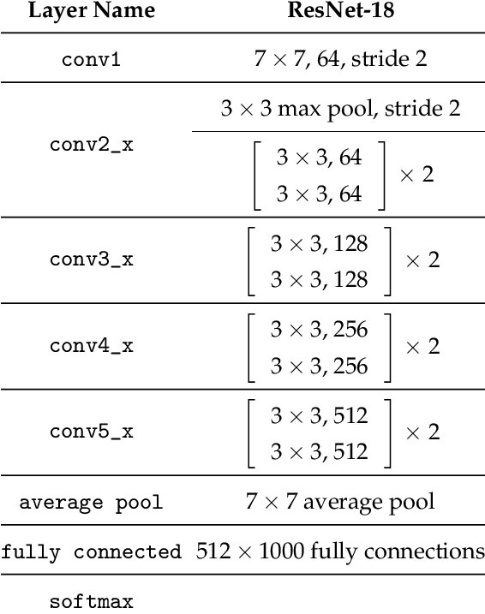
*Sample images from FER 2013*

**Methodology**

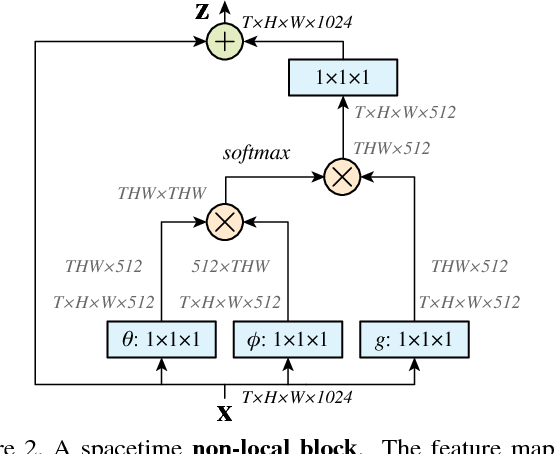
The methodology employed in this study involves a comparative analysis of the performance on the Facial Expression Recognition (FER) dataset between a base Convolutional Neural Network (CNN) model and an augmented version incorporating non-local layers. The chosen base CNN architecture is ResNet18, a well-established model recognized for its effectiveness in image classification tasks. The selection of ResNet18 is influenced by the paper on non-local neural networks where ResNet50 was employed for video frame recognition at a higher resolution of 224 x 224. Given the smaller input features in our case (48 x 48), a shallower architecture, ResNet18, is deemed appropriate for this project. This decision ensures alignment with the task requirements and computational efficiency while maintaining a principled approach to model selection.

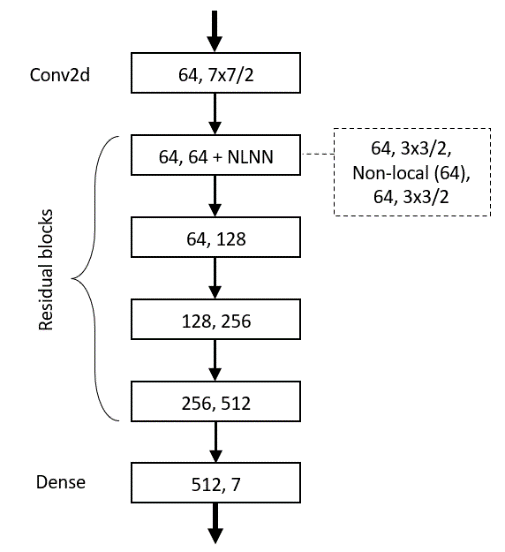
Residual Networks (ResNets) represent a groundbreaking advancement in deep neural network architectures, primarily designed to address the challenge of vanishing gradients during training. ResNets introduce the concept of residual learning, which involves the use of skip connections or shortcuts to jump over certain layers in a neural network. These skip connections facilitate the flow of gradients during backpropagation, enabling the training of very deep networks with hundreds of layers. ResNets have demonstrated remarkable performance in various computer vision tasks, showcasing their ability to capture intricate hierarchical features and accelerating convergence.

ResNet 18, a variant of the Residual Network architecture, leverages residual blocks as its foundational building blocks. These blocks incorporate two sequential 3x3 convolutional layers, each followed by batch normalization and a Rectified Linear Unit (ReLU) activation. This structure not only facilitates the extraction of intricate features but also mitigates the vanishing gradient problem, enabling the training of deep neural networks. The ReLU activation, applied element-wise after each convolution, introduces non-linearity critical for capturing complex patterns in image data. In contrast, deeper ResNet variants, such as ResNet 50 and ResNet 100, employ bottleneck residual blocks that consist of three sequential convolutional layers of kernel sizes 1x1, 3x3 and 1x1 respectively.

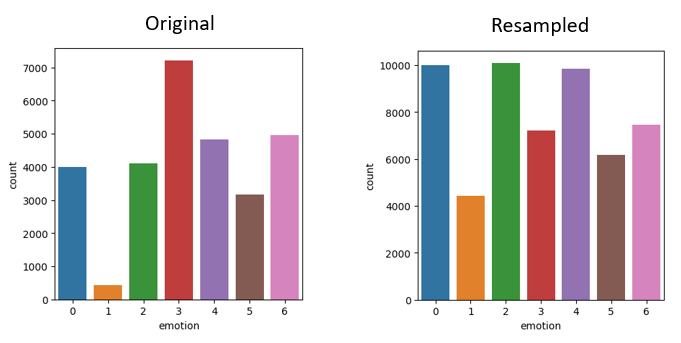


Non-Local Neural Networks (NLNNs) enrich neural network architectures by addressing the challenge of capturing long-range dependencies in data. Unlike traditional convolutional layers that rely on local context, NLNNs introduce a non-local operation that allows each pixel to interact with all other pixels in the image, regardless of spatial proximity. This mechanism is inspired by self-attention mechanisms discussed in Attention is all you need (the paper that inspired the GPT model as well). This enables the model to incorporate global contextual information into its predictions. The non-local operation involves pairwise interactions between pixels, with learnable parameters for adaptability during training.



In the paper introducing NLNNs, the experimental findings suggest that a higher number of non-local layers can be advantageous. However, in this study, a choice has been made to integrate only a single non-local layer within the initial residual block of the ResNet architecture. This decision is based upon the relatively small resolution of the input images. Given that ResNet inherently reduces the input dimensions through downsampling, placing the non-local layer at an early stage facilitates the extraction of more global information from the image before subsequent downsampling operations.

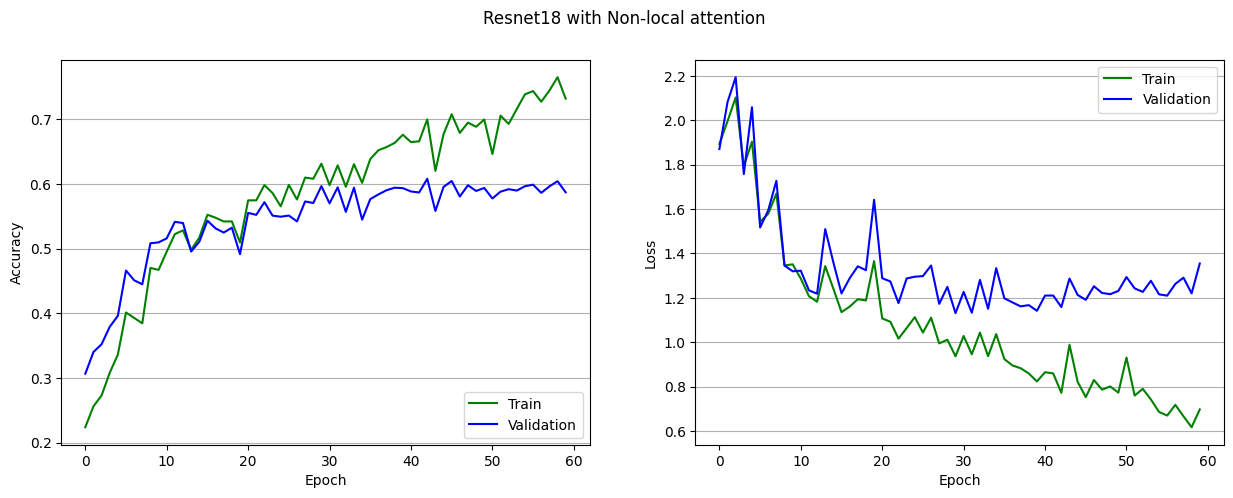
**Data preprocessing**

An initial examination of the label distribution demonstrates a notable imbalance in sample sizes across various emotion classes. To mitigate this imbalance, a fundamental technique known as data augmentation has been applied. The ImageDataGenerator library from Keras is employed for this purpose. Data augmentation plays a pivotal role in addressing class imbalances, contributing to the robustness of the model and enhancing its overall performance.

**Results**

Apart from the incorporation of the non-local layer, both models share identical hyperparameters. All sources of randomness, such as those pertaining to data augmentation and the initialization of weights and biases, have been seeded with the same value for both models. This uniformity in hyperparameters and random operations ensures a fair and comparable experimental setup.

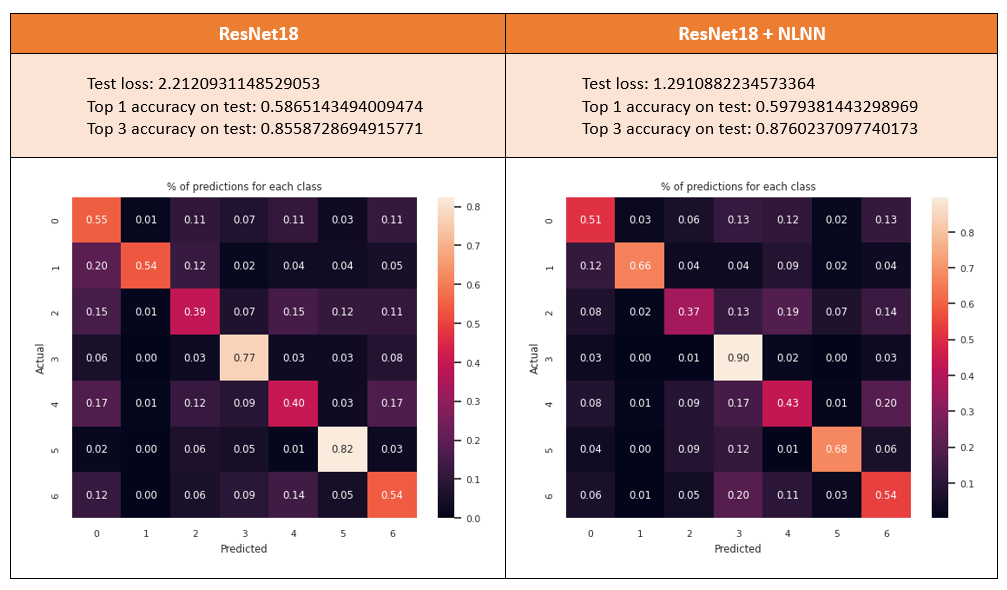
After training each model for 60 epochs, following are the results:

In the ResNet-only model, there is a rapid improvement in accuracy and a corresponding decline in loss for both training and validation data during the initial epochs. However, as training progresses, the model demonstrates signs of overfitting, with performance on training data surpassing that on validation data. The accuracy stabilizes around 55%, displaying a plateau effect, while the loss, reaching its minimum during the initial epochs, subsequently increases with the intensification of overfitting. It is evident that terminating the training process well before 60 epochs is necessary to obtain a model with optimal accuracy and minimal loss due to the observed overfitting phenomenon.

In the ResNet + NLNN model, there is a more gradual increase in accuracy and decrease in loss for both training and validation data compared to the ResNet-only model. Notably, the values for both datasets remain closely aligned, indicating a reduced tendency for overfitting. Despite the consistent application of regularization techniques in both models, the introduction of the non-local layer in the ResNet + NLNN model contributes an additional layer of regularization, improving generalization. As expected, the accuracy improvement rate diminishes with each epoch, but it avoids plateauing and continues to increase, although gradually, reaching 60% accuracy by the end of the training epochs. The loss, while exhibiting a rise after a few epochs, does so at a notably slower rate compared to the ResNet-only model, representing a more favourable outcome in terms of model performance and generalization.

Based on the observations, it is evident that the incorporation of the non-local layer enhances the model's training performance in several ways. This improvement is characterized by a more consistent and stable progression of accuracy and loss, a diminished tendency for overfitting, and consequently, superior values in performance metrics.

Below is a summary of the results obtained from both models on the test data



**Conclusion**

In conclusion, the study aimed to assess the performance enhancement resulting from the incorporation of non-local neural networks into a base Convolutional Neural Network (CNN) model. Choosing ResNet as the foundational architecture, known for its resilience against the vanishing gradient issue, ensured a more robust base performance compared to a vanilla sequential CNN. Both models, one utilizing the base ResNet architecture and the other integrated with a non-local layer, underwent training under identical conditions to establish a fair and direct comparison. Our findings reveal that the model augmented with the non-local layer outperformed the base ResNet model. This conclusion underscores the effectiveness of integrating non-local neural networks in refining the performance of the base CNN model, as evidenced by improvements in stability, lower chances of overfitting and overall superior performance.