Mathematical Exploration of SqueezeNet: CNN Architecture Design Strategies for Efficient Low-Cost Computing

Apurva Patel*, Harsh Benahalkar†, Devika Gumaste‡
Department of Electrical Engineering
Columbia University
New York, USA
Email: {*amp2365, †hb2776, ‡dg3370}@columbia.edu

Abstract

This paper provides a comprehensive quantitative analysis and mathematical exploration of various SqueezeNet architectures, emphasizing the enhancement of computational efficiency. SqueezeNet is renowned for its compact design and suitability for resource-constrained settings such as mobile and edge devices. Through meticulous investigation, we dissect the mathematical frameworks underlying architectural decisions across SqueezeNet variants, including vanilla configurations, those integrating residual connections, and those employing dense-sparse-dense training paradigms. Our inquiry elucidates the intricate trade-offs between model complexity, parameter reduction methodologies, and computational efficiency metrics. Leveraging advanced mathematical optimization techniques, we propose novel architectural design strategies tailored to further boost efficiency while preserving competitive performance benchmarks. Empirical validations corroborate the effectiveness of these strategies, showcasing tangible improvements in efficiency metrics without notable sacrifices in model accuracy. Overall, this study serves as a technical guide for researchers and practitioners engaged in the optimization of deep learning architectures for real-world deployment scenarios, particularly where computational efficiency is paramount. By rigorously dissecting the mathematical underpinnings and architectural nuances of SqueezeNet variants, this work contributes to the ongoing discourse on efficient neural network design for diverse deployment contexts.

Index Terms

SqueezeNet, parameter reduction, architecture design, mathematical analysis

I. Introduction

In the expansive domain of deep learning, Convolutional Neural Networks (CNNs) stand as one of the most influential innovations, particularly in tasks concerning image analysis, recognition, and understanding. These networks, inspired by the intricate organization of the visual cortex in biological organisms, revolutionized computer vision by mimicking the hierarchical processing of visual information.

CNNs operate by employing layers of learnable filters or kernels that convolve over input data, progressively extracting and learning hierarchical features. These features become increasingly abstract and complex as they traverse deeper layers of the network, ultimately enabling the network to discern intricate patterns and objects within images.

While CNNs have proven incredibly effective across a spectrum of applications, their widespread adoption on resource-constrained devices, such as mobile phones and embedded systems, has posed challenges. These devices often have limited computational power and memory, necessitating the development of architectures that strike a balance between accuracy and efficiency.

SqueezeNet emerges as a pioneering solution to this conundrum. Unlike traditional CNNs, which can be parameter-heavy, SqueezeNet achieves comparable performance while significantly reducing the number of parameters, thus rendering it lightweight and computationally efficient. This efficiency is pivotal for real-time applications and deployment on devices with restricted computational resources.

At the heart of SqueezeNet's efficiency lies its ingenious utilization of 1x1 convolutional filters. These filters, also known as pointwise convolutions, allow for dimensionality reduction by squeezing the input data channels before applying traditional convolutions. This reduction in channel dimensions drastically lowers the number of parameters while preserving crucial features, effectively compressing the model without sacrificing accuracy.

In this paper, we embark on an exhaustive exploration of SqueezeNet architectures, aiming to unravel the mathematical intricacies that underscore their design principles. We delve into various architectural variants of SqueezeNet, ranging from the original design to those enriched with residual connections and those trained using dense-sparse-dense methodologies.

Our objective is multifaceted. We seek to elucidate the delicate trade-offs between model complexity, parameter reduction techniques, and computational efficiency within SqueezeNet variants. Through meticulous analysis and empirical experimentation, we endeavor to provide comprehensive insights into the underlying mechanisms driving the efficiency of SqueezeNet.

These insights hold profound implications for both researchers and practitioners in the field of deep learning. They offer actionable guidance on how to optimize neural network architectures for real-world deployment scenarios, where computational efficiency is of paramount importance. By demystifying the mathematical underpinnings of SqueezeNet, we aspire to catalyze advancements in efficient neural network design tailored to the diverse demands of modern deployment environments.

II. Previous Related Work

A. Convolution Neural Networks

Convolutional Neural Networks (CNNs) have become the backbone of modern machine learning and artificial intelligence, particularly in the domain of computer vision. These deep neural networks are adept at automatically learning and extracting intricate patterns and features from visual data, making them indispensable tools for tasks such as image classification, object detection, and image segmentation.

At the core of CNNs lies a fundamental concept inspired by the architecture of the visual cortex in animals. The network consists of multiple layers, each comprising small, trainable filters or kernels that convolve across the input data. Through this convolutional process, the network extracts hierarchical representations of features, with lower layers detecting simple patterns like edges and textures, and deeper layers discerning more complex structures like shapes and objects.

The hierarchical and localized nature of convolution allows CNNs to efficiently capture spatial dependencies within the input data. Furthermore, CNNs often incorporate additional layers such as pooling layers to downsample feature maps and fully connected layers to make predictions based on the learned features.

One of the key strengths of CNNs lies in their ability to automatically learn hierarchical representations directly from raw data, without the need for handcrafted features. This end-to-end learning process enables CNNs to generalize well to unseen data and achieve state-of-the-art performance across a wide range of visual tasks.

Moreover, CNNs exhibit a degree of parameter sharing, meaning that the same set of filters is applied across the entire input data, leading to a significant reduction in the number of parameters compared to fully connected networks. This parameter efficiency, combined with their hierarchical feature learning capabilities, makes CNNs well-suited for large-scale datasets and real-world applications.

CNNs have revolutionized the field of computer vision by enabling the automatic extraction of meaningful features from raw data. Their hierarchical architecture, parameter efficiency, and end-to-end learning capabilities have propelled them to the forefront of machine learning research and applications, paving the way for advancements in various fields, from medical imaging to autonomous vehicles.

Wiener-Khinchin theorem (cross-correlation): In Convolutional Neural Networks (CNNs), convolution operations are often represented as cross-correlations between the input image (or feature map) and the filter (kernel or mask). This cross-correlation operation captures the similarity between the input and the filter at different spatial locations.

Let $f * g$ denote the cross-correlation of functions $f(t)$ and $g(t)$. Then

$$
\begin{aligned}
f * g &= \int_{-\infty}^{\infty} f^*(\tau) g(t+\tau) d\tau \qquad (1)\\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} F^*(\nu) e^{2\pi i \nu \tau} d\nu \right] \left[\int_{-\infty}^{\infty} G(\nu') e^{-2\pi i \nu' (t+\tau)} d\nu' \right] d\tau \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F^*(\nu) G(\nu') e^{-2\pi i \tau (\nu' - \nu)} e^{-2\pi i \nu' t} d\tau d\nu d\nu' \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F^*(\nu) G(\nu') e^{-2\pi i \nu' t} \left[\int_{-\infty}^{\infty} e^{-2\pi i \tau (\nu' - \nu)} d\tau \right] d\nu d\nu' \\
&= \int_{-\infty}^{\infty} F^*(\nu) G(\nu) e^{-2\pi i \nu t} d\nu \\
&= F\left[F^*(\nu) G(\nu) \right] \qquad (2)
\end{aligned}
$$

where F denotes the Fourier transform, z^* is the complex conjugate, and

$$
\begin{aligned}
f(t) &= \mathcal{F}_\nu \left[F(\nu) \right](t) = \int_{-\infty}^{\infty} F(\nu) e^{-2\pi i \nu t} d\nu \\
g(t) &= \mathcal{F}_\nu \left[G(\nu) \right](t) = \int_{-\infty}^{\infty} G(\nu) e^{-2\pi i \nu t} d\nu
\end{aligned}
$$

Applying a Fourier transform on each side gives the cross-correlation theorem,

$$
f * g = F\left[F^*(\nu) G(\nu) \right] \qquad (3)
$$

If $F = G$, then the cross-correlation theorem reduces to the Wiener-Khinchin theorem.

- **Mathematical Representation:** The cross-correlation operation, denoted as $f * g$, involves sliding the filter $g(t)$ across the input signal $f(t)$ and computing the integral of the product of the values at each overlapping point.
- **Relation to Fourier Transform:** According to the cross-correlation theorem, the cross-correlation operation in the spatial domain is equivalent to the multiplication of the Fourier transforms of the input and the filter in the frequency domain.
- **Connection to CNNs:** In CNNs, this theorem provides a deep insight into the computational efficiency of convolution operations. By performing convolution in the frequency domain using the Fourier transforms of the input and filter, convolutional layers can leverage fast Fourier transform (FFT) algorithms for efficient computation.
- **Implementation in CNNs:** CNN frameworks often exploit this relationship by converting convolution operations into frequency domain multiplications, especially when dealing with large filters or images. This approach can lead to significant computational savings, particularly in deep networks with numerous convolutional layers.
- **Generalization:**While the original theorem applies to continuous signals, its principles can be extended to discrete signals and images, which are commonly encountered in CNNs. The discrete Fourier transform (DFT) and its efficient implementation, the fast Fourier transform (FFT), play a crucial role in this context.

So, the cross-correlation theorem provides a theoretical foundation for understanding convolution operations in CNNs and guides the development of efficient algorithms for implementing convolutions, thereby contributing to the computational effectiveness of CNN architectures.

*1) **CNN architecture and working**:* Convolutional Neural Networks (CNNs) are a class of deep neural networks specifically designed for processing structured grid-like data, such as images. Figure 1 shows a typical layout of a CNN implementation. From a mathematical perspective, CNNs comprise several key components:

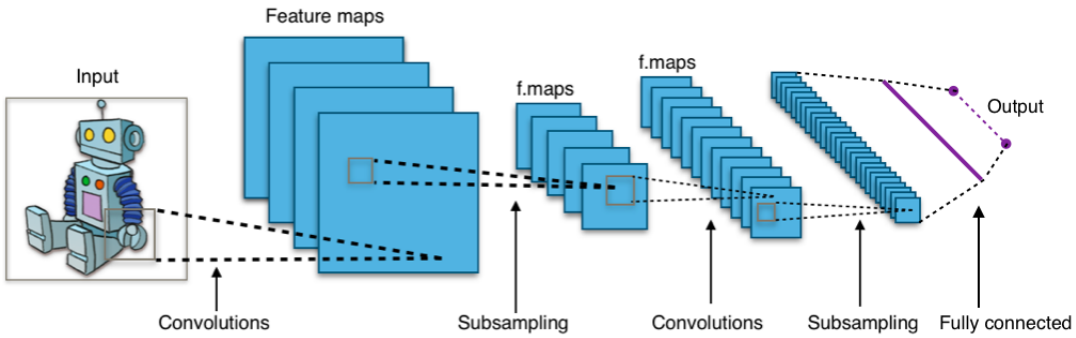


Fig. 1: Typical CNN architecture (figure adapted from Wikipedia [2])

1) **Convolutional Layers**: These layers apply convolution operations to the input data using learnable filters or kernels. Mathematically, a convolution operation can be represented as:

$$(I * K)(i,j) = \sum_{m=0}^{F-1} \sum_{n=0}^{F-1} I(i+m, j+n) \cdot K(m,n)$$

where I represents the input data, K represents the filter, and F is the size of the filter.

2) **Activation Functions (ReLU)**: After convolution, an activation function is typically applied element-wise to introduce non-linearity into the network. The Rectified Linear Unit (ReLU) is a commonly used activation function, defined as:

$$\text{ReLU}(x) = \max(0, x)$$

3) **Pooling Layers (Max Pooling)**: Pooling layers downsample the feature maps obtained from convolutional layers, reducing spatial dimensions while retaining important information. The max pooling operation selects the maximum value from each patch of the feature map and can be represented as:

$$\text{MaxPooling}(I)(i,j) = \max_{m,n} I(2i+m, 2j+n)$$

4) **Fully Connected Layers**: These layers receive flattened feature maps as input and perform high-level reasoning and decision-making. The output of a fully connected layer can be represented as:

$$\text{FC}(x) = Wx + b$$

where W represents the weight matrix, x represents the input vector, and b represents the bias vector.

These mathematical components work together to enable CNNs to automatically learn hierarchical representations of features directly from raw data, making them powerful tools for various machine learning and computer vision tasks.

B. Alexnet variant of CNN

Building on the design principle of CNN, Alexnet was developed. AlexNet, proposed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton et al. 2012 [10], was a significant advancement in the field of computer vision. It built upon the principles of Convolutional Neural Networks (CNNs) while introducing several key architectural innovations that contributed to its success.

It presents a fundamental, straightforward, and efficient Convolutional Neural Network (CNN) architecture. It predominantly comprises sequential stages, including convolutional layers, pooling layers, rectified linear unit (ReLU) layers, and fully connected layers. The architecture consists of 5 convolutional layers, followed by a pooling layer after the first four layers, and culminates in 3 fully-connected layers.

In AlexNet's architecture, convolutional kernels are learned through back-propagation optimization, optimizing the entire cost function with the stochastic gradient descent (SGD) algorithm. Convolutional layers process input feature maps using sliding convolutional kernels to produce convolved feature maps, while pooling layers aggregate information within specified neighborhood windows through operations like max pooling or average pooling.

The success of AlexNet can be attributed to practical strategies, notably the ReLU non-linearity layer and dropout regularization technique. ReLU, defined by the equation

$$f(x) = \max(x, 0),$$

and it acts as a half-wave rectifier function, accelerating the training phase and mitigating overfitting issue commonly seen in model with high number of parameters. Dropout regularization, a form of stochastic regularization, randomly sets a portion of input or hidden neurons to zero during training, reducing neuron co-adaptations, primarily applied in the fully connected layers of AlexNet to reduce over dependancy on features.

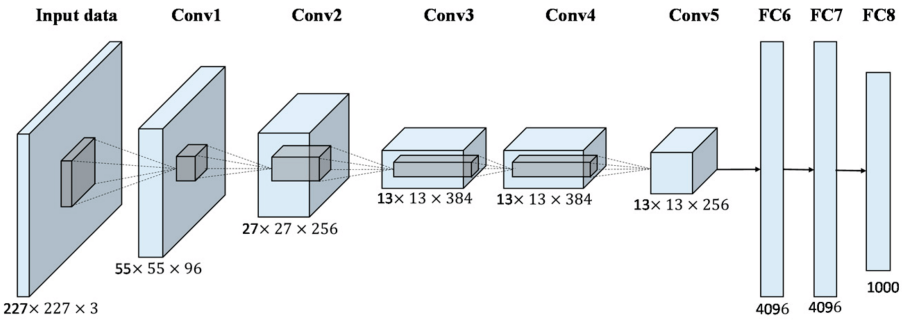


Fig. 2: Alexnet architecture (figure adapted from [5])

1) **Convolution Operation**:
   In AlexNet's convolutional layers, the convolution operation is applied to the input feature maps using learnable filters (convolutional kernels). Let $I \in \mathbb{R}^{H \times W \times 3}$ represent the input feature map, where H and W are the height and width, respectively, and 3 is the number of input channels corresponding to RGB. Let $K_1 \in \mathbb{R}^{11 \times 11 \times 3 \times 96}$ represent the filters in the first convolutional layer. The convolution operation at a specific spatial position (i, j, k) in the first layer can be represented as:

$$(I * K_1)(i, j, k) = \sum_{m=0}^{10} \sum_{n=0}^{10} \sum_{c=0}^{2} I(i + m, j + n, c) \cdot K_1(m, n, c, k)$$

   Here, k represents the index of the output channel, and the summation is performed over all elements of the filter and corresponding input feature map region.

2) **Max Pooling Operation**:
   After the convolutional layers, max pooling operations are applied to the feature maps. Let $I' \in \mathbb{R}^{H' \times W' \times C'}$ represent the output feature map after convolution, where H', W', and C' are the height, width, and number of channels, respectively. The max pooling operation with a pool size of 3×3 and stride of 2 can be represented as:

$$(I')_{\text{pooled}}(i, j, c) = \max_{m,n} I'(2i + m, 2j + n, c)$$

3) **Fully Connected Layer**:
   AlexNet includes three fully connected layers with ReLU activation functions. Let $x \in \mathbb{R}^N$ represent the input vector to the fully connected layer, where N is the number of neurons. The output of a fully connected layer with 4096 neurons and ReLU activation can be represented as:

$$\text{Output}(x) = \max(0, Wx + b)$$

   Here, W represents the weight matrix of dimensions $4096 \times N$, and b represents the bias vector.

At the end we get just over 62M parameters to train to acheive a decent level accuracy. It is very difficult to deploy Alexnet on the edge and real time computation was limited to hardware capabilities, and this paved the way for researchers to develop a different model architecture and acheive those deficits.

III. SqueezeNet Overview

From the detailed parameter analysis (see I in Appendix A for parameter calculation) of AlexNet, it's evident that the model requires a substantial number, $>$ 62M parameters, which can lead to computational overhead, especially in resource-constrained environments such as mobile devices or embedded systems.

SqueezeNet, proposed by Iandola et al. in 2016 [7], addresses this challenge by introducing a highly efficient convolutional neural network architecture that significantly reduces the number of parameters without sacrificing performance. The key idea behind SqueezeNet is to replace traditional 3×3 filters with 1×1 filters, which significantly reduces the number of parameters while still capturing meaningful information.

The primary objective behind the development of SqueezeNet was to create CNN architectures that could maintain competitive accuracy while significantly reducing the number of parameters. To achieve this, 3 strategies were employed:

1) **Strategy 1:** Replacing 3x3 filters with 1x1 filters helps reduce the number of parameters while maintaining accuracy.
2) **Strategy 2:** Decreasing the number of input channels to 3x3 filters is achieved through squeeze layers, further reducing computational complexity.
3) **Strategy 3:** Deferring down-sampling until later stages in the network results in larger activation maps, which can lead to higher classification accuracy.

A. Mathematical intuition / theorems involved

1) **Forward pass through one layer:** In a DCN with multiple layers, computation of the ith activation in layer $l+1$ of the DCN can be expressed as follows:

$$a_{(l+1),i} = \sum_{j=1}^{N} w_{(l+1),i,j} \cdot a_{(l),j} + b_{(l+1),i}$$

where (l) represents the lth layer, N represents the number of additions, $w_{(l+1),i,j}$ represents the weight, and $b_{(l+1),i}$ represents the bias.

2) **Universal Approximation Theorem:** It states that a feed-forward neural network with a single hidden layer containing a finite number of neurons can approximate any continuous function on a compact input domain to arbitrary accuracy, given a sufficiently large number of neurons and appropriate activation functions. Mathematically, it can be stated as follows:
Given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is continuous on a compact subset S of $\mathbb{R}^n$, and a non-constant, bounded, and continuous activation function $\sigma : \mathbb{R} \to \mathbb{R}$, there exists a single-hidden-layer feedforward neural network with sufficiently many neurons and appropriate weights and biases such that for any $\epsilon > 0$ and any x in S, the approximation error satisfies:

$$|f(x) - F(x)| < \epsilon$$

where $F(x)$ is the output of the neural network.

3) **Lipschitz Continuity of Neural Networks** It has been shown that neural networks, including CNNs, are Lipschitz continuous functions. This means that small changes in the input result in bounded changes in the output. Mathematically, this can be expressed as:

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

Where f is the neural network function, x and y are input vectors, $\|\cdot\|$ is a norm, and L is the Lipschitz constant. This property is important for the stability and robustness of neural networks.

4) **Gradient Descent:** Gradient descent is an optimization algorithm used to minimize a cost function $J(\theta)$ by iteratively updating the parameters θ in the direction of the negative gradient of the cost function. The update rule for gradient descent is given by:

$$\theta := \theta - \alpha \nabla J(\theta)$$

where α is the learning rate, and $\nabla J(\theta)$ is the gradient of the cost function with respect to the parameters θ.

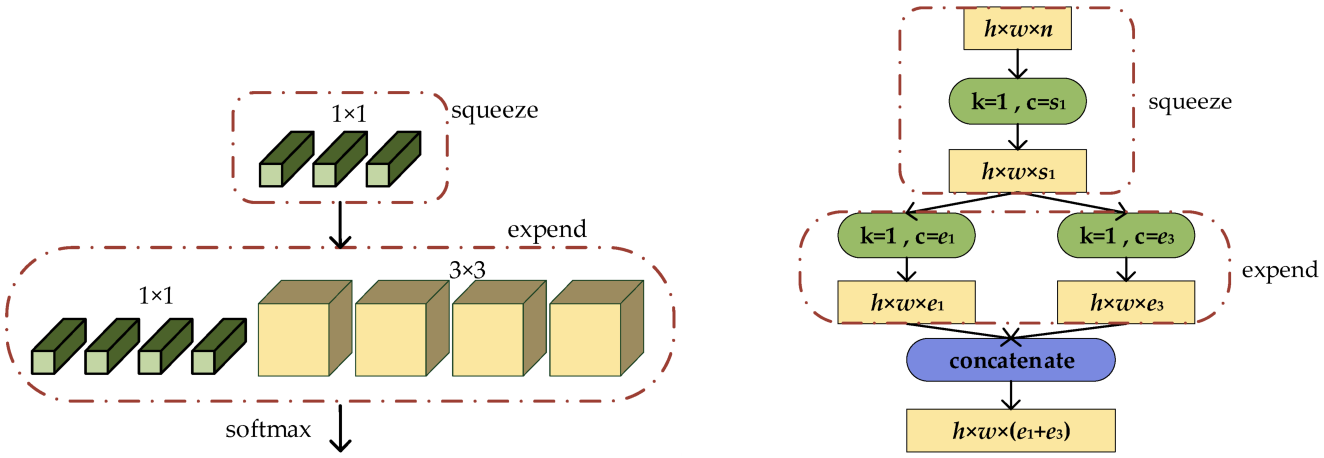
5) **Regularization:** Regularization techniques such as L1 and L2 regularization are used to prevent overfitting in neural networks. L2 regularization penalizes large weights by adding a term proportional to the square of the weights to the loss function, while L1 regularization encourages sparsity by adding a term proportional to the absolute value of the weights.
Mathematically, the regularized loss function $J_{\text{reg}}(\theta)$ is given by:

$$J_{\text{reg}}(\theta) = J(\theta) + \lambda \sum_{i=1}^{n} |\theta_i|^p$$

where $J(\theta)$ is the original loss function, λ is the regularization parameter, n is the number of parameters, and p is the regularization term (usually $p = 1$ for L1 regularization and $p = 2$ for L2 regularization).

B. Fire Modules

The core building blocks of SqueezeNet are Fire Modules. Each Fire Module consists of a squeeze layer followed by expand layers. The squeeze layer consists of 1×1 convolutional filters, which aim to compress the input channels (reduce the number of feature maps). The expand layers consist of a mix of 1×1 and 3×3 convolutional



(a) Fire module (b) Fire operation process

Fig. 3: Fire module in Squeezenet v1.0 and it's working (figures adapted from [12])

The squeeze layer reduces the number of input channels (e.g. from $128 \times 32 \times 32$ to $64 \times 32 \times 32$) to the 3×3 filters, while the expand layer helps in learning better representations by incorporating both 1×1 and 3×3 convolution filters, which aim to expand the feature maps in depth. The outputs of these layers are concatenated to enhance expressiveness.

In order to implement the Fire module, the authors utilized concatenation to connect layers with different filter resolutions (e.g., 1×1 and 3×3). This was done by implementing separate convolution layers for each filter resolution and concatenating their outputs.

In a precise way, let's denote the number of input channels to the squeeze layer as C_{squeeze}, the number of output channels from the squeeze layer as $C_{\text{squeeze\_out}}$, and the number of output channels from the expand layer as C_{expand}. Then, the total number of parameters in a Fire Module can be calculated as the sum of the parameters in the squeeze layer and the expand layer.

The squeeze layer performs 1×1 convolutions, resulting in $C_{\text{squeeze}} \times C_{\text{squeeze\_out}}$ parameters. The expand layer consists of a 1×1 convolution followed by a 3×3 convolution, resulting in $C_{\text{squeeze\_out}} \times (C_{\text{expand}} + 3 \times C_{\text{expand}})$ parameters.

By carefully selecting the number of output channels, SqueezeNet ensures that the total number of parameters in each Fire Module is minimized while preserving representational capacity.

C. SqueezeNet Architecture

SqueezeNet consists of multiple Fire Modules interleaved with max-pooling layers. These Fire Modules are designed to efficiently capture both spatial and temporal features across different scales in the input data. The max-pooling layers help reduce the spatial dimensions of the feature maps, further contributing to parameter reduction.

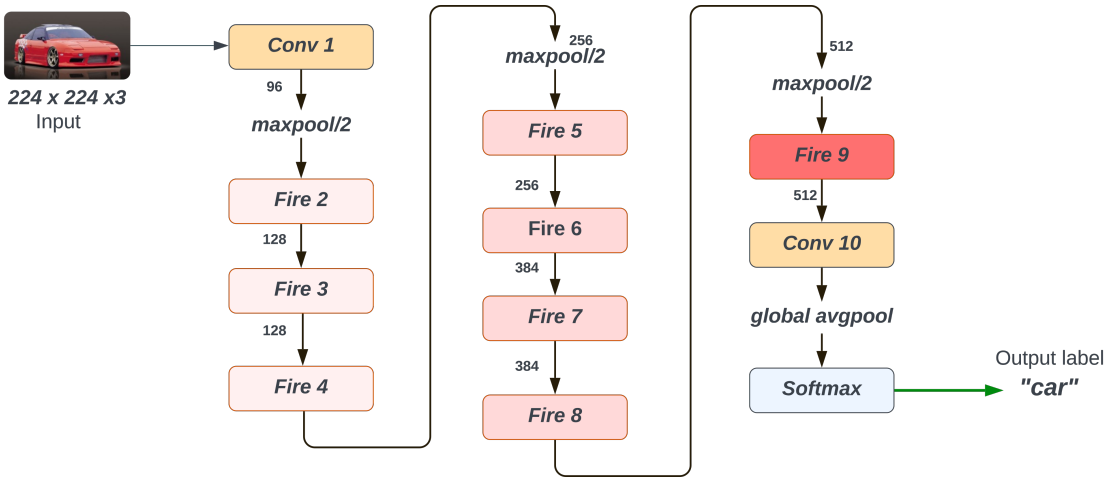


Fig. 4: Squeezenet v1.0 architecture (figure adapted from [1])

1) ***1x1 Convolutions****:* Mathematically, let's consider the operation of a 1×1 convolution on an input feature map X with dimensions $H \times W \times C_{\text{in}}$, where H is the height, W is the width, and C_{in} is the number of input channels. Let Y be the output feature map with dimensions $H \times W \times C_{\text{out}}$, where C_{out} is the number of output channels.

The output Y is computed as:

$$Y_{i,j,k} = \sum_{c=1}^{C_{\text{in}}} X_{i,j,c} \times W_{1,1,c,k} + b_k$$

Where:

- $X_{i,j,c}$ is the value of the input feature map at position (i, j) in the c-th channel.
- $W_{1,1,c,k}$ is the weight of the convolutional kernel at position $(1, 1)$ in the c-th input channel and k-th output channel.
- b_k is the bias term for the k-th output channel.

By applying 1×1 convolutions, SqueezeNet effectively reduces the number of parameters in the convolutional layers.

2) ***Bottleneck Design****:* The bottleneck design in SqueezeNet involves using 1×1 convolutions to reduce the number of input channels to the subsequent 3×3 convolutions in the expand layers. This design choice is motivated by the observation that a large number of input channels to a convolutional layer can significantly increase the number of parameters and computational cost without necessarily improving performance.

Mathematically, let's denote the number of input channels to the bottleneck layer as C_{in}, the number of output channels from the squeeze layer as C_{squeeze}, and the number of output channels from the expand layer as C_{expand}. The bottleneck layer consists of 1×1 convolutions, resulting in $C_{\text{in}} \times C_{\text{squeeze}}$ parameters. The expand layer then uses 1×1 convolutions to increase the number of channels to C_{expand}, followed by 3×3 convolutions.

By reducing the number of input channels to the 3×3 convolutions in the expand layer, the bottleneck design significantly reduces the number of parameters while still allowing the network to capture complex spatial features.

*3) **Global Average Pooling**:* Mathematically, global average pooling reduces the spatial dimensions of the feature maps to a single value per channel by computing the average of all values in each channel. Let's denote the input feature map to the global average pooling layer as X with dimensions $H \times W \times C$, where H is the height, W is the width, and C is the number of channels.

The output of global average pooling Y is computed as:

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c}$$

Where Y_c is the c-th channel of the output feature map.

This operation introduces no additional parameters, effectively reducing the overall parameter count in the network.

*4) **Model Compression Techniques**:* Model compression techniques such as network pruning and quantization involve identifying and removing redundant or less important parameters from the network while preserving performance.

- **Network Pruning (implementation below)**: In network pruning, parameters with low magnitudes are pruned or removed from the network. Mathematically, this can be achieved by setting parameters below a certain threshold to zero. Pruning reduces the parameter count and computational cost of the network without significantly affecting performance.
- **Quantization**: Quantization reduces the precision of weights and activations from floating-point numbers to lower bit-width integers or fixed-point numbers. This reduces the memory footprint and computational cost of the network, making it more efficient for deployment on resource-constrained devices.

These model compression techniques leverage mathematical concepts such as magnitude-based pruning and quantization to eliminate parameters that contribute minimally to the network's output, resulting in further parameter reduction.

Overall, SqueezeNet demonstrates that it is possible to achieve high performance on image classification tasks with a significantly smaller model size compared to traditional architectures like AlexNet. This makes SqueezeNet well-suited for deployment on resource-constrained devices where memory and computational resources are limited.

IV. SqueezeNet Design and Implementation

Squeezenet has a lot of variations/implementations considering it's popularity and model complexity. SqueezeNet, proposed by Iandola et al. in 2016 [7] implemented the first squeezenet architecture consisting of fire modules and pretrianed on the Imagenet dataset [3].

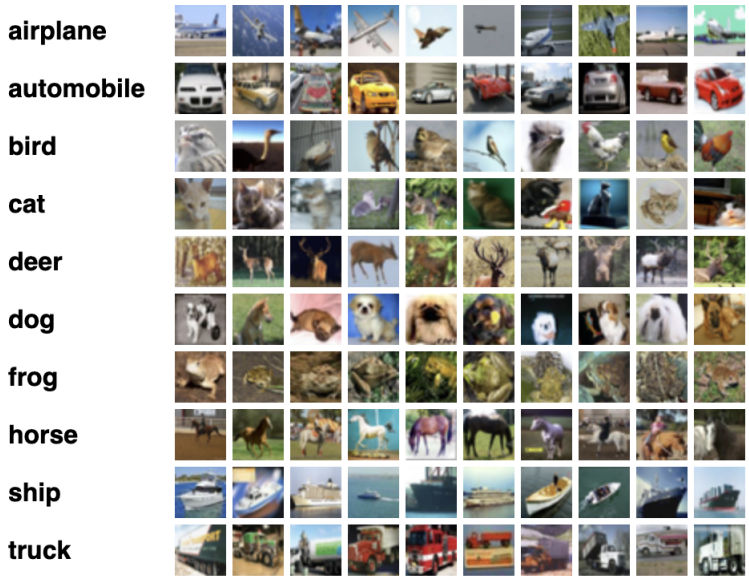


Fig. 5: Samples of all classes from the CIFAR-10 dataset (figure adapted from [9])

A. Vanilla SqueezeNet

SqueezeNet has been developed with efficiency in mind, aiming to achieve high-performance metrics while keeping the model and computation lightweight, so that accuracy is not sacrificed. The extensive use of novel fire modules which combine 1x1 and 3x3 convolutions, enable high accuracy while keeping the computation required in control. The breakdown of the functionality of the fire modules is as follows:

1) **Compression**: The fire module consists of 1x1 convolution layers. These layers first "squeeze" the input, effectively compressing the feature maps.
2) **Expansion**: Following the squeeze operation, the compressed features are passed through an expand operation. 1x1 and 3x3 convolutions are applied parallelly on the input. The purpose of using these layers is that the 1x1 convolution keeps the size constant and the 3x3 convolution captures spatial information.
3) **Concatenation**: The outputs of the 1x1 and 3x3 convolutions are concatenated along the channel dimension so that the output is now enriched spatially and channel-wise.
4) **Activation**: Finally, non-linearity is introduced by passing the output through an activation. the activation used in the fire module is Rectified Linear Unit (ReLU).

SqueezeNet has been implemented in 2 versions, SqueezeNet 1.0 and SqueezeNet 1.1. The major difference between the two versions is in the input layer and the conv layer where SqueezeNet 1.0 uses 96 channels with a kernel size of 7 and SqueezeNet 1.1 used 64 channels with a kernel size of 3.

SqueezeNet was originally trained on the ImageNet dataset using stochastic gradient descent with momentum. We applied standard data augmentation techniques such as random cropping and horizontal flipping to increase the diversity of the training data. However for our experiments, training the SqueezeNet model, using different Optimizers and Activation functions presented with an opportunity to explore the impact of these choices and the dataset, on the model performance. Optimizers selected include Adam (adam), Stochastic Gradient Descent (sgd), Perturbed Gradient Descent (pgd), Stochastic Gradient Descent with restarts (sgdr), and Accelerated Gradient Descent (agd). Activation functions selected for the classifier layers of the model include Rectified Linear Unit (relu), Leaky Rectified Linear Unit (leaky_relu), and Hyperbolic Tangent (tanh) activation functions. The choice of the dataset (CIFAR-10) for this purpose helps in analyzing the performance of all trainable layers in identifying global and local patterns in the image.

The activation functions used for experimenting with the vanilla architecture, and the intuition behind using them:

1) **TanH (Hyperbolic Tangent)**: This activation function zero-centers the input values by squashing them between -1 and 1. This property helps the network in learning symmetrical patterns and the smoothness of the gradient aids in gradient-based optimization techniques. Although it is very beneficial in gradient optimizations, it suffers from the problem of vanishing gradients, especially in deep networks. This activation function should also be used when the output range of the function is appropriate for the application.

$$\text{TanH}(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

2) **ReLU (Rectified Linear Unit)**: ReLU is a piecewise linear activation function that simply returns 0 for any negative inputs and the input value as it is for positive inputs. ReLU is extensively used as it is simple and computationally friendly. This also gets rid of the vanishing gradient problem seen with the Hyperbolic Tangent activation function. However they are prone to suffer from the dying ReLU problem, where neurons of the layer become inactive and don't learn if the output is zero consistently.

$$\text{ReLU}(x) = \max(0, x)$$

3) **Leaky ReLU**: Leaky ReLU is used to mitigate the problem of dying ReLU by allowing negative inputs to pass through. This creates the alpha parameter which needs to be optimized for the activation function to optimize properly. Hence this choice of the hyper-parameter alpha can affect network performance.

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{otherwise} \end{cases}$$

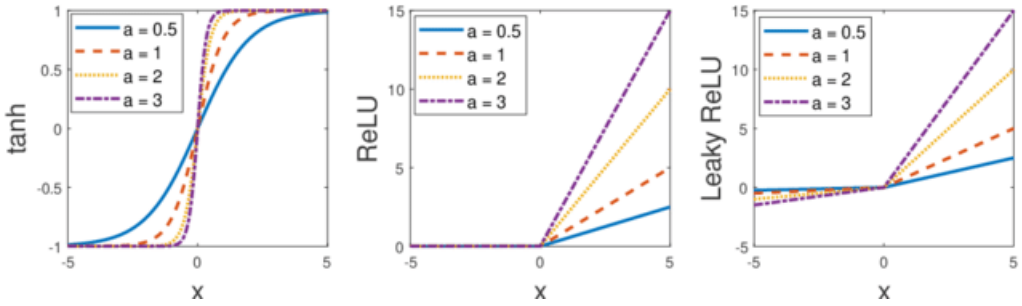


Fig. 6: Figure showing the graph of TanH, ReLU, and Leaky ReLU activation functions(figure adapted from [8])

The Optimizers used for experimenting with the vanilla architecture, and the intuition behind using them:

1) **Stochastic Gradient Descent (SGD)**: Stochastic Gradient Descent updates model weights in the direction of the minima. The factor by which the model weights are updated is called eta and it is constant. This is generally used for large datasets, and it is computationally friendly and efficient. However, this optimizer can cause oscillations if there is high variance in the weight updates.

$$\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t; x_i, y_i)$$

2) **Adam (Adaptive Moment Estimation)**: Adam combines the advantages of both AdaGrad and RMSProp. It dynamically adjusts the learning rate for each weight based on estimates of the first and second moments of the gradients. This makes Adam effective for optimizing non-convex objectives and noisy weights. And since Adam adjusts the weights according to their moments, the convergence is faster. However, this advantage comes at the cost of memory. Adam is also sensitive to hyperparameters, which makes it very tricky to use.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \epsilon} \cdot m_t$$

3) **Perturbed Gradient Descent**: Perturbed Gradient Descent deliberately introduces noise by adding a random noise value term to the weight during updates. This exploration of different directions in weights helps the model in escaping the local minima, and thus generalize better. Similar to Adam, this is also sensitive to the hyper-parameter, i.e. the pertubance value. The generation of additional perturbance also comes at the cost of computation and memory.

$$\theta_{t+1} = \theta_t - \eta(\nabla J(\theta_t) + \epsilon_t)$$

4) **Accelerated Gradient Descent**: This uses momentum to accelerate gradient descent updates, by accumulating past gradients to determine the direction and speed of parameter updates. This addition of momentum speeds up convergence and reduces oscillations. however, the momentum parameter makes this model sensitive to the hyperparameters, which when set incorrectly may cause the model to overshoot the minima, leading to instability.

$$v_{t+1} = \beta v_t + (1 - \beta) \nabla J(\theta_t)^2$$

5) **SGD with Restarts**: SGD with Restarts periodically restarts the learning rate schedule during training. This is primarily done to escape saddle points and local minima. Resetting the learning rate helps avoid convergence to sub-optimal solutions.

$$\eta_t = \eta_0 \cdot \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor}$$

B. SqueezeNet with Residual connections

SqueezeNet with residual connections enhances the original SqueezeNet architecture by incorporating residual connections between Fire modules. Residual connections allow for easier training of deeper networks by facilitating the flow of gradients through the network and mitigating the vanishing gradient problem.

The key mathematical concept behind residual connections is the Residual Learning Framework, which was introduced in the seminal paper "Deep Residual Learning for Image Recognition" by Kaiming He et al. [6].

Residual Learning Framework is to reformulate the underlying mapping as a residual mapping, rather than expecting the stacked layers to directly fit a desired underlying mapping. Specifically, if the desired underlying mapping is $H(x)$, the residual learning framework lets the layers fit another mapping $F(x) = H(x) - x$, where x is the input to the layer.

Mathematically, this can be expressed as:

$$H(x) = F(x) + x$$

The advantage of this formulation is that it makes the optimization process easier for the network. Instead of having the layers fit the complete mapping $H(x)$, they only need to fit the residual mapping $F(x)$, which is often simpler.

In the context of SqueezeNet, the residual connections are implemented between the "Fire" modules, as shown in the project code. These bypass connections, denoted as bypass_{23}, bypass_{45}, bypass_{67}, and bypass_{89}, enable the network to effectively increase its depth and capacity without significantly increasing the number of parameters, thereby improving the overall performance and efficiency of the model.

1) Architecture design:

- Each Fire module consists of a squeeze layer followed by expand layers, as in the original SqueezeNet architecture [7].
- Additionally, residual connections are introduced between Fire modules. These connections directly pass the output of one Fire module to the input of the next Fire module (see figure 7).
- The intuition behind residual connections is that they allow the network to learn residual functions, i.e., the difference between the input and output of a layer, which can be easier to optimize and backpropogate error/loss.

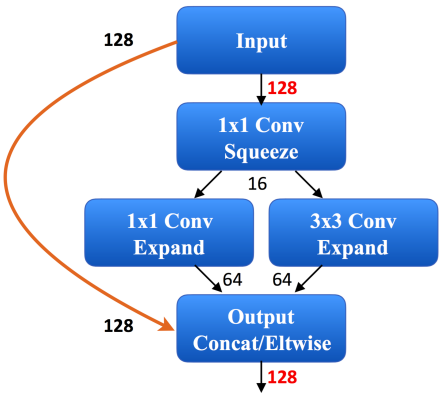


Fig. 7: Fire module with residual connections

Mathematically, let's denote the output of the i-th Fire module as y_i. The output of the i-th Fire module with residual connection y_i^{res} can be expressed as the sum of the output of the Fire module y_i and the input to the next Fire module x_{i+1}. This can be represented as:

$$y_i^{res} = y_i + x_{i+1}$$

The intuition goes as follows:

- By adding the residual connection, the network can learn to adjust the input to the next Fire module based on the difference between the desired output and the output of the current Fire module.
- This facilitates the training process, especially in deeper networks, by allowing gradients to flow more easily through the network.

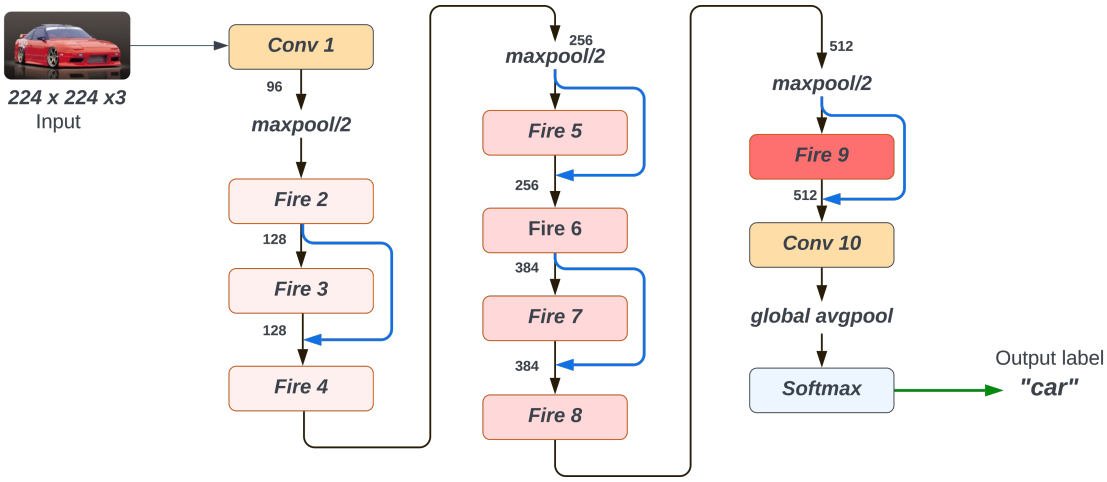


Fig. 8: Squeezenet architecture with residual connections

The loss function used in the SqueezeNet implementation is the standard Cross-Entropy Loss, which is commonly used for classification tasks. The mathematical formulation of the Cross-Entropy Loss is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(p_{i,j}) \tag{4}$$

Where:

- N is the number of samples in the batch
- C is the number of classes
- $y_{i,j}$ is the ground truth label (one-hot encoded) for the i-th sample and j-th class
- $p_{i,j}$ is the predicted probability for the i-th sample and j-th class

Backpropagation Equations

The backpropagation equations for the residual connections in our model can be derived using the chain rule. Let's denote the input to the l-th layer as x_l and the output as y_l. The residual connection is represented as $x_{l+1} = y_l + x_l$.

The gradients with respect to the weights in the l-th layer can be computed as:

$$\frac{\partial \mathcal{L}}{\partial W_l} = \frac{\partial \mathcal{L}}{\partial y_l}\frac{\partial y_l}{\partial x_l}\frac{\partial x_l}{\partial W_l}$$

And the gradients with respect to the input x_l can be computed as:

$$\frac{\partial \mathcal{L}}{\partial x_l} = \frac{\partial \mathcal{L}}{\partial y_l}\frac{\partial y_l}{\partial x_l} + \frac{\partial \mathcal{L}}{\partial x_{l+1}}\frac{\partial x_{l+1}}{\partial x_l}$$

Where:

- $\frac{\partial \mathcal{L}}{\partial y_l}$ is the gradient from the upper layers, propagated through the network
- $\frac{\partial y_l}{\partial x_l}$ is the gradient of the layer's activation function
- $\frac{\partial x_l}{\partial W_l}$ is the gradient of the layer's output with respect to its weights
- $\frac{\partial x_{l+1}}{\partial x_l}$ is the gradient of the residual connection, which is simply 1 (since $x_{l+1} = y_l + x_l$)

The key advantage of the residual connections is that they provide an alternative path for the gradients to flow during backpropagation (see figure 8 with blue lines as residual connections), which can help mitigate the vanishing gradient problem and improve the training of deep neural networks like SqueezeNet with a significant reduction in model training time.

C. SqueezeNet with DSD traing approach

- Introduction:
  Deep neural networks have demonstrated significant advancements across various fields such as computer vision, natural language processing, and speech recognition. The availability of more potent hardware facilitates the training of intricate DNN models with significant capacities. The benefit of complex models lies in their ability to expressively capture the intricate, nonlinear relationships between features and outputs. However, the drawback of such large models is their susceptibility to capturing noise rather than the intended patterns in the training data, leading to overfitting and high variance that does not generalize well to new datasets. Conversely, reducing model capacity may lead to underfitting and high bias, causing the machine learning system to overlook relevant feature-output relationships. Balancing bias and variance concurrently poses a challenge in optimization.
  To address this issue, researchers from Stanford, Nvidia, Baidu and Facebook propose a novel training strategy called dense-sparse-dense training flow (DSD). [4] This approach begins with training a dense model conventionally, followed by regularization with sparsity-constrained optimization, and finally, increasing model capacity by restoring and retraining pruned weights. At inference time, the resulting model from DSD retains the same architecture and dimensions as the original dense model, without incurring any additional overhead.

- Training Flow:
  DSD training methodology is a 3-step process: dense, sparse, re-dense. Each step is shown in Figure 5. The progression of weights is shown in Figure 6.
  - **Initial Dense Training:**
    First step in this method learns the connection weights and the importance using normal network training on the dense network. Unlike the conventional training, the goal of this step is not only to learn the values of the weights; but also which connections are important. The original paper uses the simple heuristic to quantify the importance of weights using their absolute values.
  - **Sparse Training:**
    The pruning step (S step) involves the removal of low-weight connections and the training of a sparse network. Uniform sparsity is applied across all layers, controlled by a single hyperparameter: the sparsity percentage, indicating the proportion of weights pruned to zero. For each layer W with N parameters, the parameters are sorted, and the k-th largest ($\lambda = S_k$) is selected as the threshold, where $k = N \times (1 - \text{sparsity})$. A binary mask is then generated to eliminate weights smaller than λ. This pruning of small weights is motivated by the Taylor expansion, as depicted by the loss function and its expansion in Equations (1) and (2). Minimizing the increase in loss during hard thresholding on weights requires minimizing the first and second terms in Equation (2). Since parameters are being nullified, ΔW_i simplifies to $W_i - 0 = W_i$. At the local minimum where $\frac{\partial \text{Loss}}{\partial W_i} \approx 0$ and $\frac{\partial^2 \text{Loss}}{\partial W_i^2} > 0$, only the second-order term matters. Given the computational expense of calculating second-order gradients and the squared nature of W_i, $|W_i|$ is used as the pruning metric. Smaller $|W_i|$ implies a lesser increase in the loss function.

    $$\text{Loss} = f(x, W1, W2, W3, \ldots)$$

    $$\Delta\text{Loss} = \frac{\partial \text{Loss}}{\partial W_i}\Delta W_i + \frac{1}{2}\frac{\partial^2 \text{Loss}}{\partial W_i^2}(\Delta W_i)^2 + \ldots$$

    Through retraining while adhering to the binary mask in each iteration, a transformation of a dense network into a sparse one with a predefined sparsity support is achieved, capable of fully recovering or potentially enhancing the original accuracy of the initial dense model under the sparsity constraint. The uniform sparsity across layers can be fine-tuned using validation. Generally, in experiments, a sparsity value between 25% and 50% yields satisfactory results.
  - **Final Dense Training:**
    The final D step involves the recovery of previously pruned connections, thereby restoring the network to a dense state. These connections, which were previously pruned, are initialized to zero, and the entire network is retrained with a learning rate reduced to 1/10 of the original value (given the sparse network's

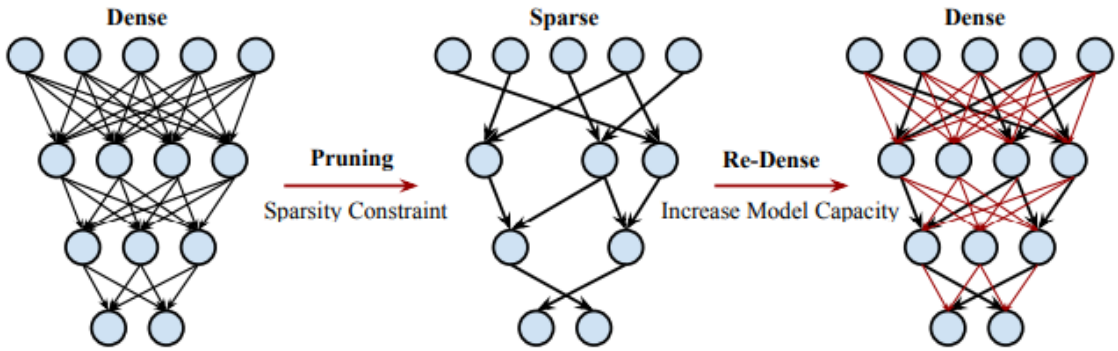


Fig. 9: Dense Sparse Dense Training Flow. Sparse training regularizes the model, final dense training restores the pruned weights (red), increasing the model capacity without over-fitting. [4]

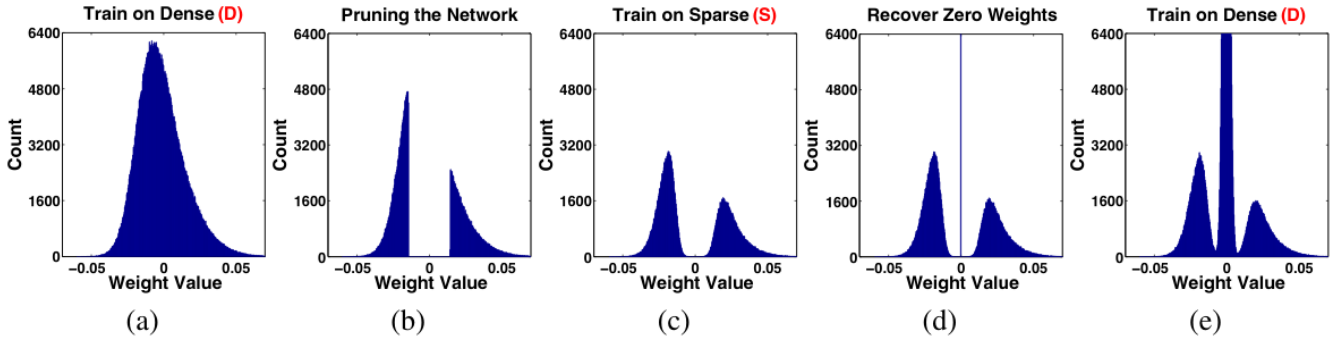


Fig. 10: Weight Distribution throughout the DSD training process. (a) Original Model (b) Pruned Model (c) After retraining the sparsity-constrained model (d) ignoring the sparsity constraints and recovering the zero weights (e) after re-training the dense network [4]

proximity to a favorable local minimum). Parameters such as dropout ratios and weight decay remain unchanged throughout this process. By reintroducing the pruned connections, the final D step increases the model capacity of the network, potentially leading to an improved local minimum compared to the sparse model resulting from the S step.

To visualize the DSD training flow, we observed the progression of weight distribution, as depicted in Figure 10. Initially, the weight distribution is centered around zero with quickly diminishing tails. Pruning, based on absolute value, results in the removal of a significant portion of the central region. During the subsequent retraining phase, the unpruned network parameters readjust, leading to a softened boundary and the emergence of a bimodal distribution, as illustrated in (c). As the re-dense training phase commences in (d), all previously pruned weights are reinstated and initialized to zero. Finally, in (e), the pruned weights undergo retraining alongside the unpruned weights, using consistent learning hyperparameters (e.g., weight decay, learning rate). A comparison between Figures (d) and (e) reveals that while the distribution of unpruned weights remains largely unchanged, the distribution of pruned weights extends further around zero, resulting in a reduction in the overall mean absolute value of the weight distribution. This phenomenon aligns with the notion that selecting the smallest vector to solve the learning problem suppresses irrelevant components of the weight vector.

- Related Work:
  While **Dropout and DropConnect** introduce random sparsity patterns at each stochastic gradient descent (SGD) iteration, DSD training adheres to a deterministic, data-driven sparsity pattern consistently throughout

sparse training. **Model distillation** offers an alternative avenue for improving neural network performance without architectural modifications. By transferring learned knowledge from a large model to a smaller, more deployable one, model distillation facilitates performance enhancements.DSD training and **model compression** share the common strategy of network pruning. However, while model compression primarily focuses on maintaining accuracy, DSD training advances further by significantly enhancing accuracy levels. Notably, DSD training achieves this without necessitating aggressive pruning; instead, a moderately pruned network (50%-60% sparse) can deliver robust performance. Conversely, model compression often requires aggressive pruning to achieve high compression rates.The theoretical underpinnings of **sparsity regularization and hard thresholding** have been extensively explored, particularly in the context of learning statistical models in high-dimensional spaces. Additionally, similar training strategies involving iterative hard thresholding and connection restoration have been proposed independently. The application of sparsity regularized optimization, notably in Compressed Sensing, further underscores its relevance in finding optimal solutions to inverse problems within highly underdetermined systems, leveraging the sparsity assumption.

V. Results and Experiments

A. *Vanilla SqueezeNet*

We experimented with 3 activation functions and 5 optimizers, totalling to 15 experiments. The model architecture was designed in PyTorch, and the re-trained weights were loaded into the model. Since we had to replace the activation function in the classification layer and train it on our own task, the final layers of the model were unfrozen to enable them to train. The CIFAR-10 dataset with 10 classes was used for training and validation of the models. The dataset for each experiment was augmented using random resize crops and normalizations. Every experiment was run for 10 epochs, with a CosineAnnealingLR applied on the optimizer. Metrics recorded include training accuracy, validation accuracy, training loss, validation loss, and epoch times. Apart from the aforementioned metrics, we wanted to see how the weights in the final layer are adjusted, so they were also recorded and visualized.

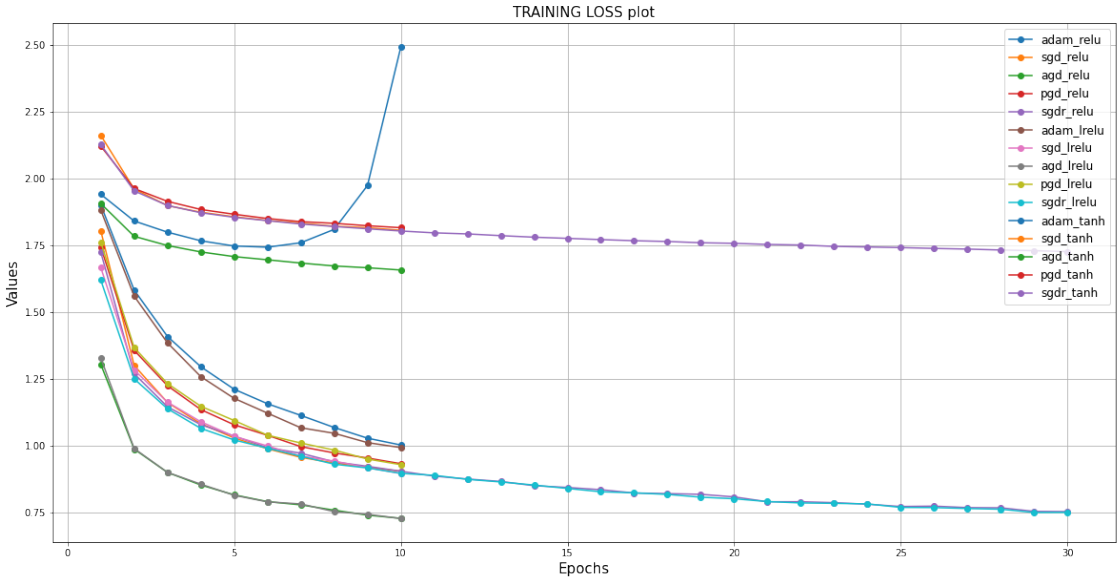


Fig. 11: Training loss values seen for all 15 experiments

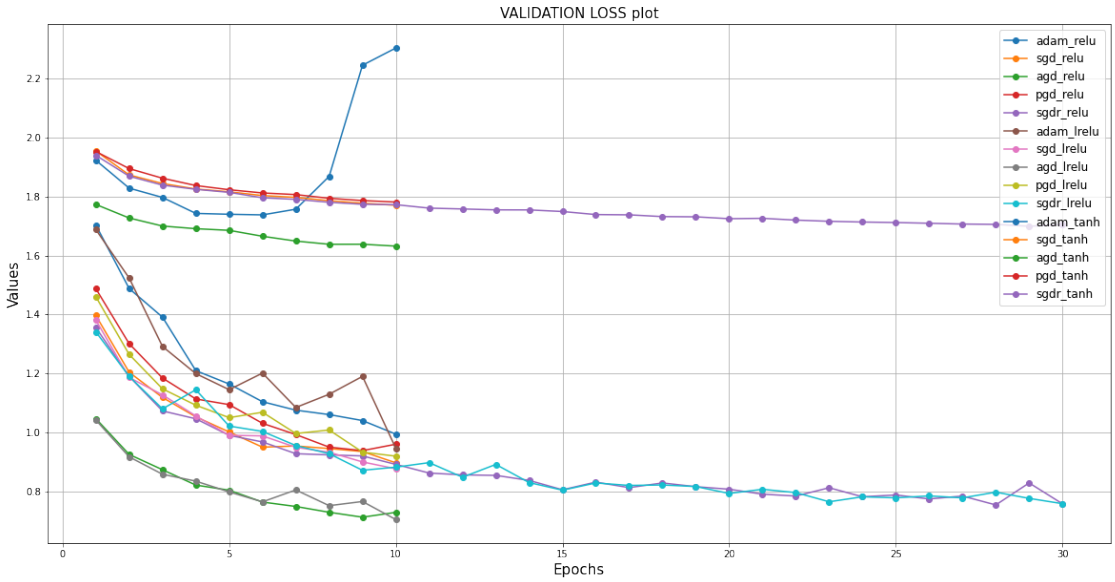


Fig. 12: Validation loss values seen for all 15 experiments

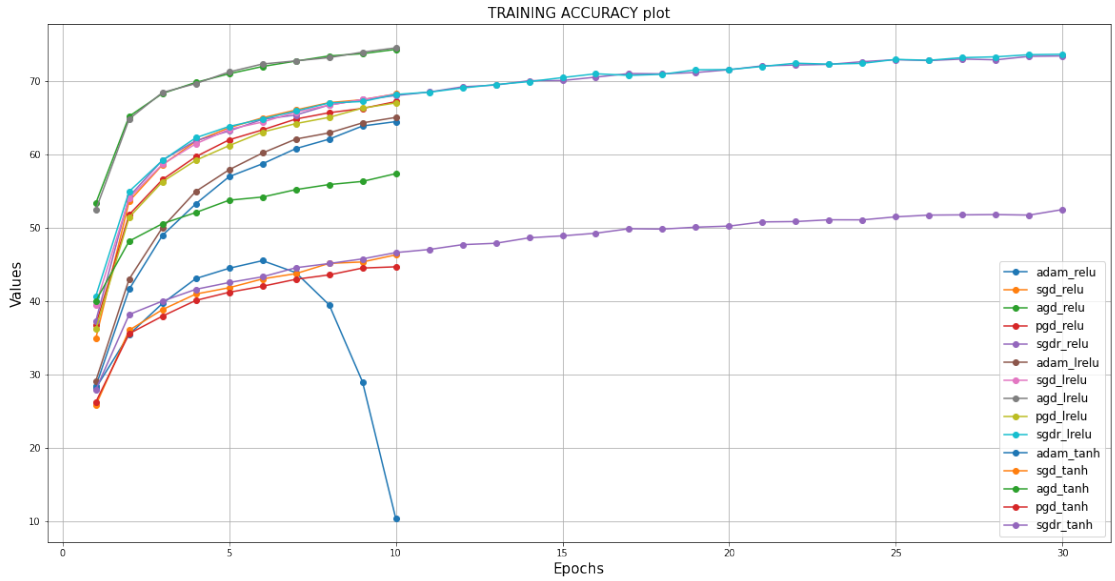


Fig. 13: Training accuracy values (%) seen for all 15 experiments

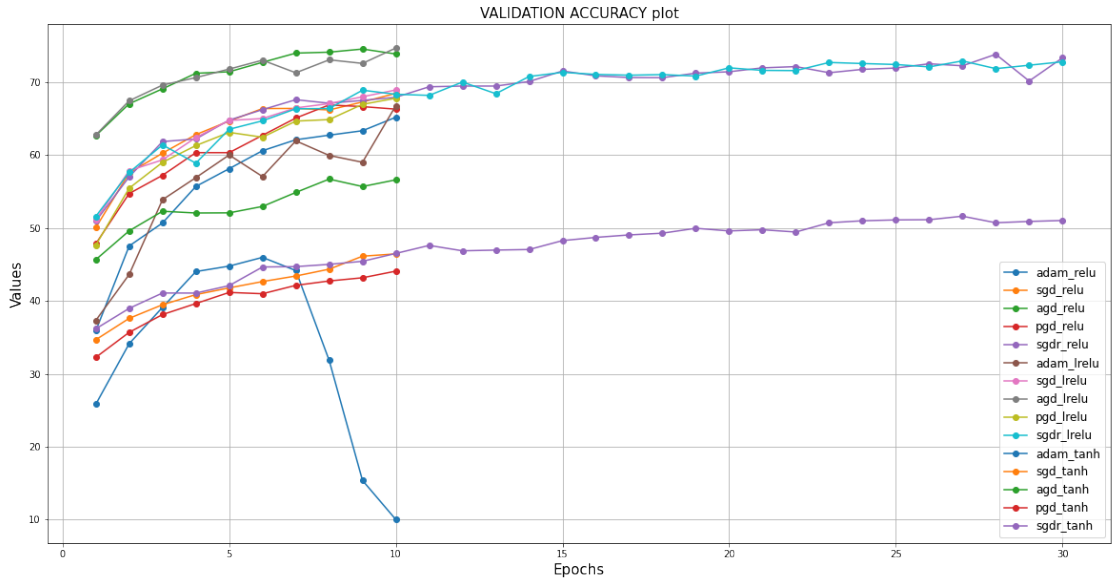


Fig. 14: Validation accuracy values (%) seen for all 15 experiments

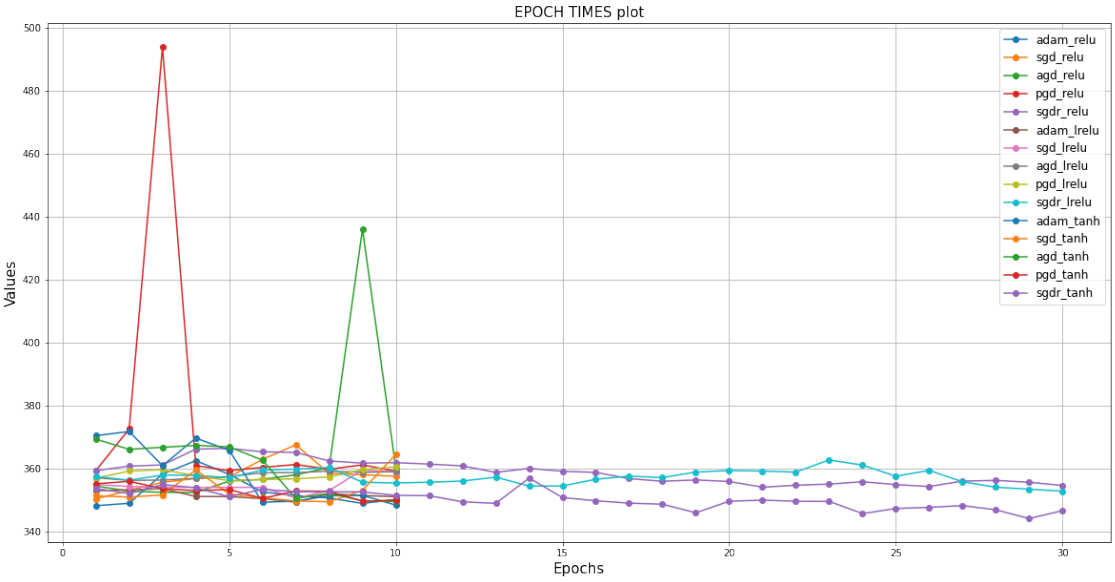


Fig. 15: Epoch times (seconds) for each cycle of training

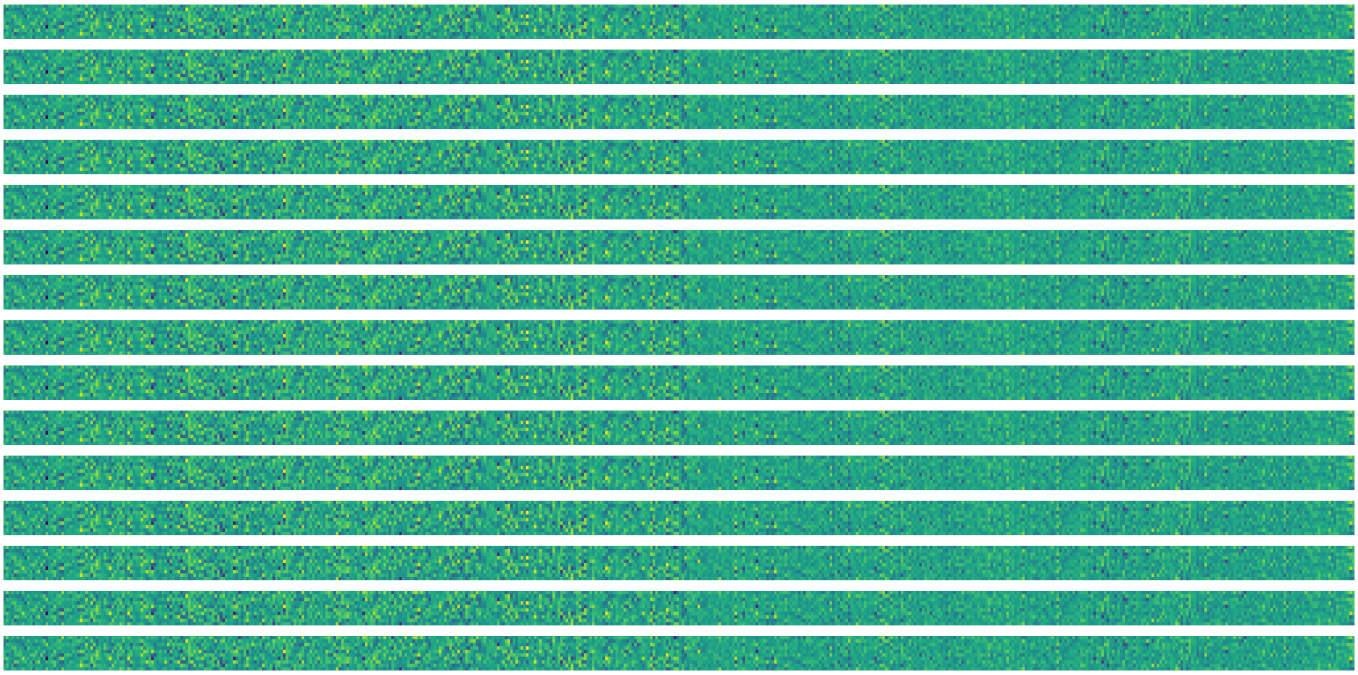


Fig. 16: Last layer classifier weights for all 15 experiments

B. SqueezeNet with Residual connections

This residual connection helps to alleviate the vanishing gradient problem that can occur in deep neural networks. By providing a direct path for the gradients to flow from the output to the earlier layers, the residual connections facilitate better gradient propagation during training.

Additionally, the residual connections allow the network to reuse and combine features from different layers, which can lead to more expressive and informative representations. This feature reuse can be particularly beneficial for small-scale models like SqueezeNet, where the network capacity is limited compared to larger models

1) ***Solving vanishing gradients problem in SqueezeNet****:* In our deep neural network with L (here 10) layers, where the input to the l-th layer is x_l and the output is y_l. In a traditional feedforward network, the mapping can be represented as:

$$y_l = F_l(x_l)$$

where F_l is the transformation performed by the l-th layer.

The vanishing gradient problem arises when the gradients of the loss function $\mathcal{L}$ with respect to the input x_l become exponentially small as l decreases. This can be expressed mathematically as:

$$\frac{\partial \mathcal{L}}{\partial x_l} = \frac{\partial \mathcal{L}}{\partial y_L} \frac{\partial y_L}{\partial y_{L-1}} \cdots \frac{\partial y_{l+1}}{\partial y_l} \frac{\partial y_l}{\partial x_l}$$

As the number of layers L increases, the product of the derivatives can become exponentially small, leading to the vanishing gradient problem.

In our case of a residual connection in squeezenet as implemented in figure 8, where the mapping is represented as:

$$y_l = F_l(x_l) + x_l$$

This is the key idea behind residual connections, where the network learns the residual mapping $F_l(x_l)$ instead of the complete mapping y_l.

The gradient of the loss function with respect to the input x_l can now be expressed as:

$$\frac{\partial \mathcal{L}}{\partial x_l} = \frac{\partial \mathcal{L}}{\partial y_L} \frac{\partial y_L}{\partial y_{L-1}} \cdots \frac{\partial y_{l+1}}{\partial y_l} \frac{\partial y_l}{\partial F_l(x_l)} \frac{\partial F_l(x_l)}{\partial x_l} + \frac{\partial \mathcal{L}}{\partial y_{l+1}} \frac{\partial y_{l+1}}{\partial x_l}$$

The key difference is the addition of the second term, $\frac{\partial \mathcal{L}}{\partial y_{l+1}} \frac{\partial y_{l+1}}{\partial x_l}$, which represents the gradient flowing through the residual connection, and the residual connection ensures that the gradients can be expressed as a sum of two terms, where one term is not subject to the exponential decay caused by the product of derivatives.

This residual connection provides an alternative path for the gradients to flow, bypassing the potentially vanishing gradients through the layers. As a result, the gradients can be maintained and effectively propagated to the earlier layers, mitigating the vanishing gradient problem.

*2) **Results for default configuration**:* The default configuration of squeezenet as mentioned in [7] is used to train a residual squeezenet architecture on Cifar-10 dataset [9] from scratch. Although the results obtained in figure 17 are nowhere close to the original paper which used Imagenet dataset [3] to pretrain the model, but it's important to note that how fast the model learns with fewer parameters (precisely 1.24M parameters (see II in Appendix B for parameter calculation in residual squeezeent) evern when started without any weight initialization.

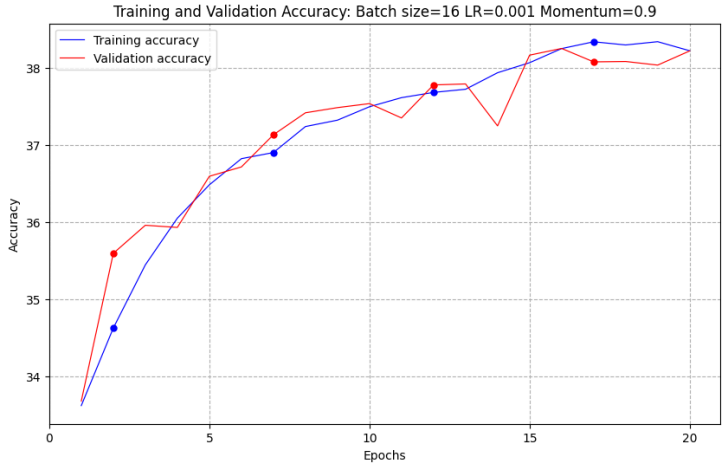


Fig. 17: Default configuration (Batch size: 32, Optimizer: SGD with 0.9 momentum & learning rate α: 0.001)

*3) **Experiments with different Batch size and Optimizers**:* Apart from the aforementioned configuration we experimented with 2 different batch sizes and 3 different optimizers. The choices of batch sizes were made by taking into consideration the GPU capabilities, estimated epoch time, memory and dataset limitations. The choices of batch sizes, 64 and 128, were made considering the GPU capabilities, estimated epoch time, memory, and dataset limitations. With the NVIDIA V100 GPU and 64GB memory, both batch sizes can be accommodated. A smaller batch size like 64 allows for faster updates to the model parameters but may result in noisy gradients. On the other hand, a larger batch size like 128 provides more stable updates but requires more memory.

***Optimizers**:*

1) **SGD with Nesterov Momentum**: This optimizer includes momentum to accelerate convergence. It helps SGD to converge faster and navigate through plateaus and local minima more efficiently.

$$\begin{aligned} &\text{Momentum term:} && v_{t+1} = \mu \cdot v_t - \alpha \cdot \nabla f(x_t + \mu \cdot v_t) \\ &\text{Update rule:} && x_{t+1} = x_t + v_{t+1} \end{aligned} \tag{1}$$

2) **Adam (Adaptive Moment Estimation)**: Adam is an adaptive learning rate optimization algorithm that computes individual adaptive learning rates for different parameters. It adapts the learning rate for each parameter, allowing for faster convergence and better performance on a wide range of deep learning tasks.

$$\begin{aligned}
&\text{Moment estimates:} && m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \\
& && v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \\
&\text{Bias correction:} && \hat{m}_t = \frac{m_t}{1-\beta_1^t} \\
& && \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\
&\text{Update rule:} && \theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}
\end{aligned} \tag{2}$$

3) **RMSProp (Root Mean Square Propagation)**: RMSProp is an adaptive learning rate optimization algorithm that divides the learning rate by an exponentially decaying average of squared gradients. It helps to adjust the learning rates dynamically based on the gradients, allowing for faster convergence and improved performance on non-stationary objectives.

$$\begin{aligned}
&\text{Exponentially decaying average:} && v_t = \beta \cdot v_{t-1} + (1-\beta) \cdot (g_t)^2 \\
&\text{Update rule:} && \theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} \cdot g_t
\end{aligned} \tag{3}$$

Here, α represents the learning rate, μ is the momentum parameter, β_1 and β_2 are the exponential decay rates for the moment estimates, ϵ is a small constant to prevent division by zero, g_t is the gradient at time step t, v_t represents the moving average of squared gradients, and θ_t represents the parameters at time step t.

***Mathematical Reasoning for Adam Optimizer**:* Adam optimizer combines the advantages of both AdaGrad and RMSProp. It computes adaptive learning rates for each parameter by considering both the first and second moments of the gradients.

The mathematical explanation for Adam optimizer is as follows:

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

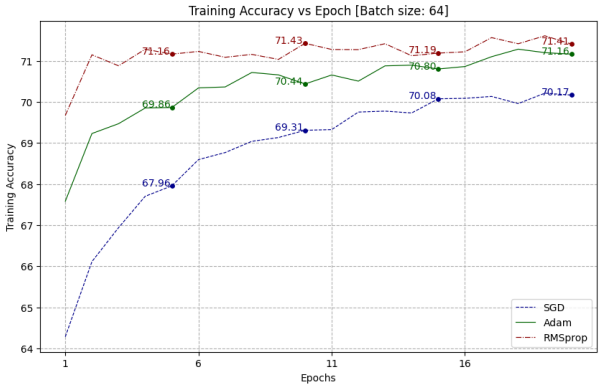
$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

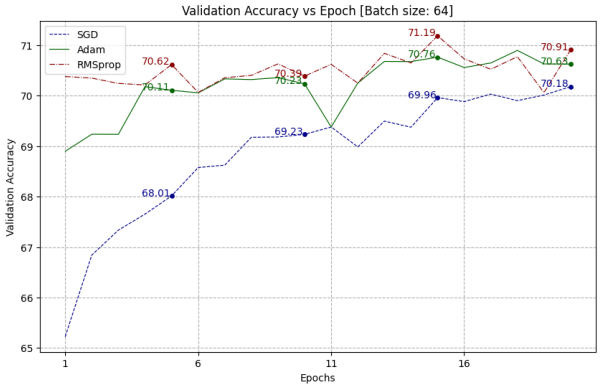
Here, m_t and v_t are estimates of the first and second moments of the gradients, β_1 and β_2 are the exponential decay rates for the moment estimates, α is the learning rate, ϵ is a small constant to prevent division by zero, and θ_t represents the parameters at time step t.

The adaptive learning rate computed by Adam optimizer allows for efficient updates of the model parameters, leading to faster convergence and improved generalization performance.

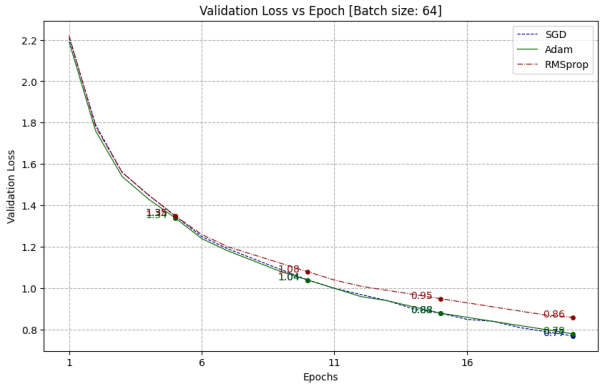
For Batch size: 64*:* The training and validation accuracy plots for each optimizer (refer to Figures 18a and 18b) reveal the fluctuations and inflection points across epochs. Additionally, Figure 18c illustrates the validation loss trends for a batch size of 64 across different optimizers. Notably, it's intriguing to note that similar loss values across optimizers don't necessarily imply identical validation or test accuracy. This disparity stems from the loss function's dependence on the optimizer output rather than the predicted label.



(a) Training accuracy vs. Epoch



(b) Validation accuracy vs. Epoch



(c) Validation Loss vs. Epoch

Fig. 18: Performance characteristics for batch size of 64

Time per epoch comparison:
The main advantage of residual connection is the reduced time per epoch to train the model. This is achieved by backpropogating the error via the residual connection and hence reducing the time to compute the chain rule otherwise which is encountered during vanilla backpropogation algorithm.
Figure 19 shows the time per epoch for a batch size of 64. It is important to note that the implementation is achieved in PyTorch which uses the parallel computing power of GPU and hence the sequence of instruction execution might determine the time for epoch for each optimizer. To better understand the time we can look at the average time for the entire experiment period; which implied that the average time per epoch follows this order: SGD $>$ RMSProp $>$ ADAM. This can be concluded by having a look at the update rule for each optimizers in equations 1, 2, 3; which outlines the total computations necessary for each optimizer and the difference with global minima. Essentially its a linear combination of no of computations and the "distance" from global minima for that optimization landscape.

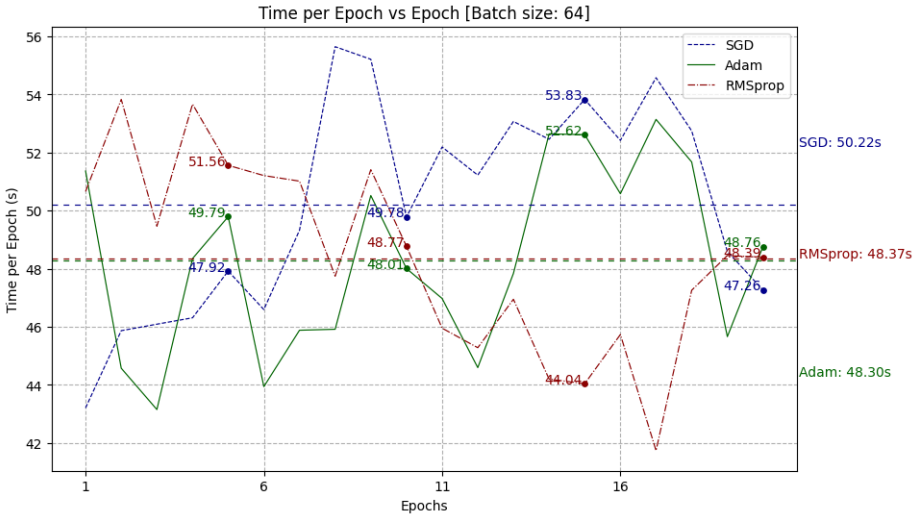


Fig. 19: Time per epoch(s) vs. Epoch for batch size of 64

Comparative study of LR Scheduler for Batch size: 128 *:* We experimented with *"ReduceLROnPlateau"* Learning rate scheduler. To illustrate the functionality of the `ReduceLROnPlateau` scheduler using mathematics, let's denote the validation loss at epoch t as val_loss_t, the learning rate at epoch t as lr_t, the patience as P, and the factor by which the learning rate is reduced as factor.

The scheduler monitors the validation loss over consecutive epochs. If the validation loss does not decrease for a specified number of epochs (patience), it reduces the learning rate by a factor. The equation to update the learning rate can be expressed as:

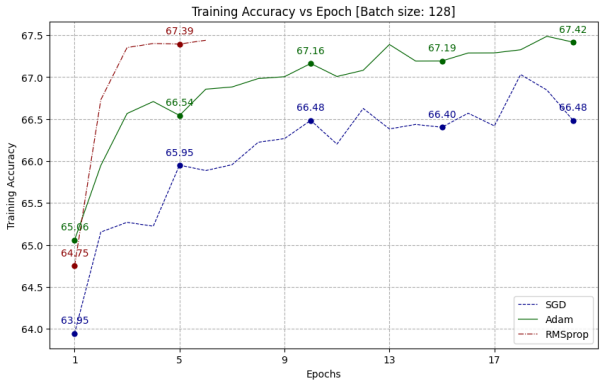
$$\text{new\_lr}_{t+1} = \begin{cases} \text{old\_lr}_t \times \text{factor} & \text{if } \text{val\_loss}_{t-P} \geq \text{val\_loss}_{t-P+1} \geq \ldots \geq \text{val\_loss}_{t-1} \geq \text{val\_loss}_t \\ \text{old\_lr}_t & \text{otherwise} \end{cases}$$

where:

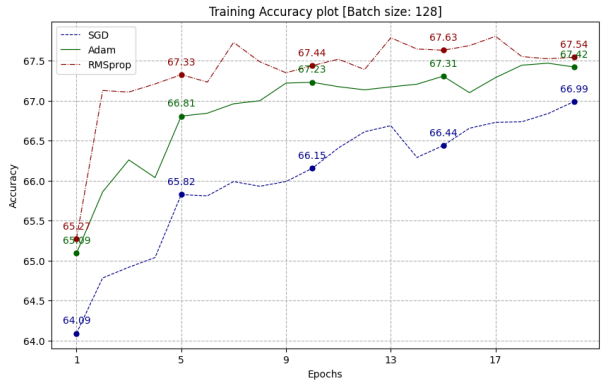
- new_lr_{t+1} is the new learning rate for epoch $t+1$.
- old_lr_t is the learning rate at epoch t.
- val_loss_{t-P} is the validation loss at epoch $t-P$.
- P is the patience, i.e., the number of epochs to wait before reducing the learning rate.
- factor is the factor by which the learning rate is reduced.

It demonstrates that the learning rate is reduced by the factor factor only if the validation loss does not improve (i.e., does not decrease) over the specified number of epochs (patience). Otherwise, the learning rate remains unchanged.

Figures 20a & 20b shows the difference in training accuracy when a higher batch size is given with and without a learning rate scheduler. In our experimentation we noticed a strange yet mathematically explainable event of RMSProp model training crashing despite numerous attempts. This can be comprehended by large already stored values in the memory and also the high batch size which fills up the memory and not allowing space allocation for incoming computation sequence and hence leading to a system crash.



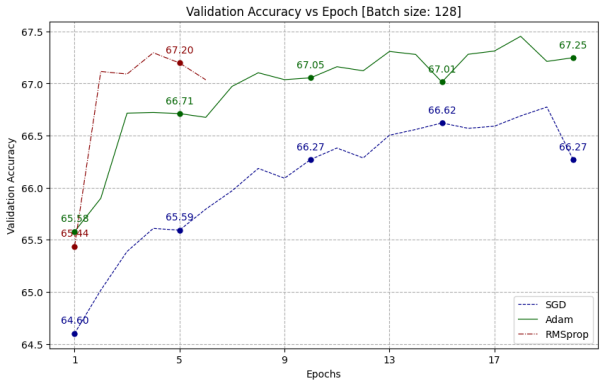
(a) Without learning rate scheduler



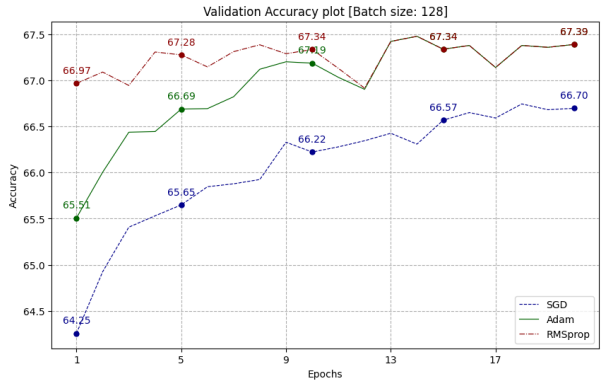
(b) With learning rate scheduler

Fig. 20: Training Accuracy comparison with 20 epochs

Validation accuracy is almost similar for all optimizers but we see a slight bump in values when a learning rate scheduler is employed. Figures 21a & 21b shows the validation accuracy metric plot for the model.



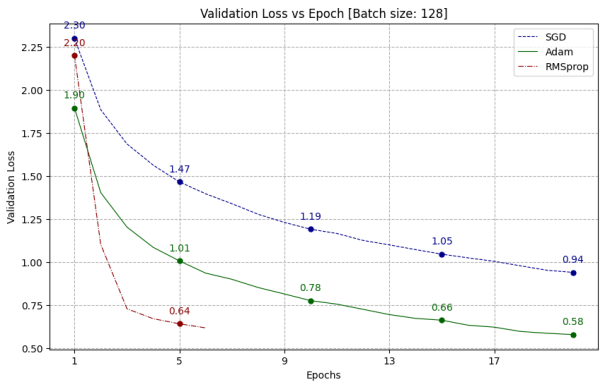
(a) Without learning rate scheduler



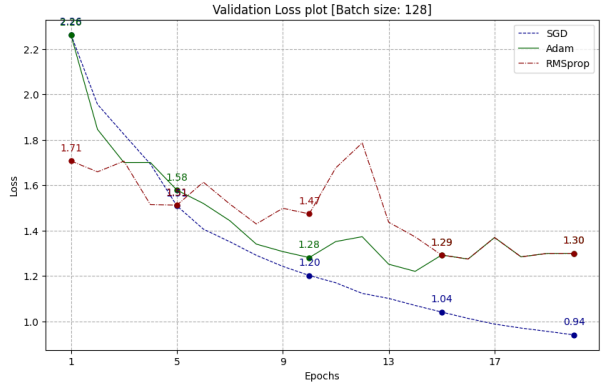
(b) With learning rate scheduler

Fig. 21: validation Accuracy comparison with 20 epochs

Through our experiments we found out that when a learning rate scheduler is employed we see a better overall convergence and this is because of the update steps in the optimizer governed by the scheduler's output.



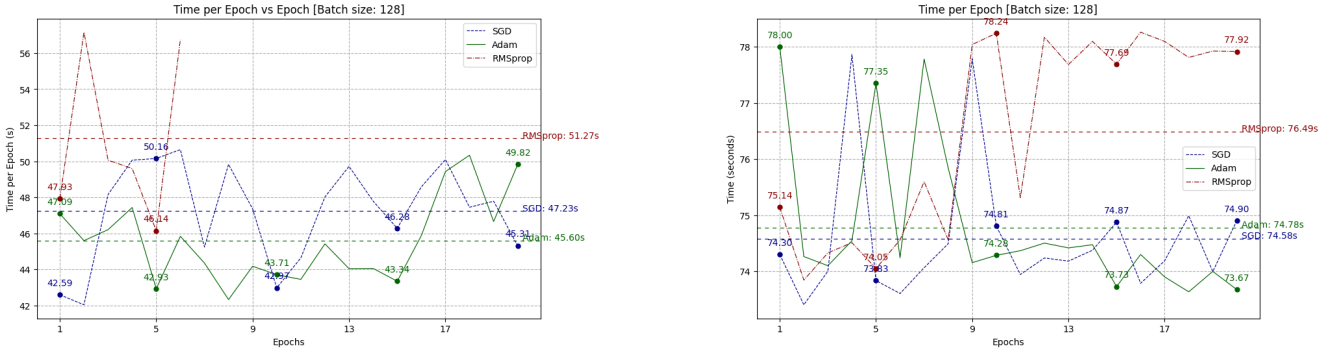
(a) Without learning rate scheduler



(b) With learning rate scheduler

Fig. 22: Validation loss comparison with 20 epochs

The main difference in performance is observed in training time per epoch. This additional time is accounted by the numerous extra computation steps required to find the optimal learning rate each time we start an epoch. Also those computations don't support GPU implementation and hence a CPU is used to perform that which increases its time for each epoch by a significant margin.



(a) Without learning rate scheduler

(b) With learning rate scheduler

Fig. 23: Epoch time comparison with 20 epochs

Weights from convolution layer: After experimenting on each computationally and practically possible configuration we get the weights in convolution layers and other modules. These weights are 7×7 filters and we have displayed weights of one of the filters form *Conv1* layer in figure 24

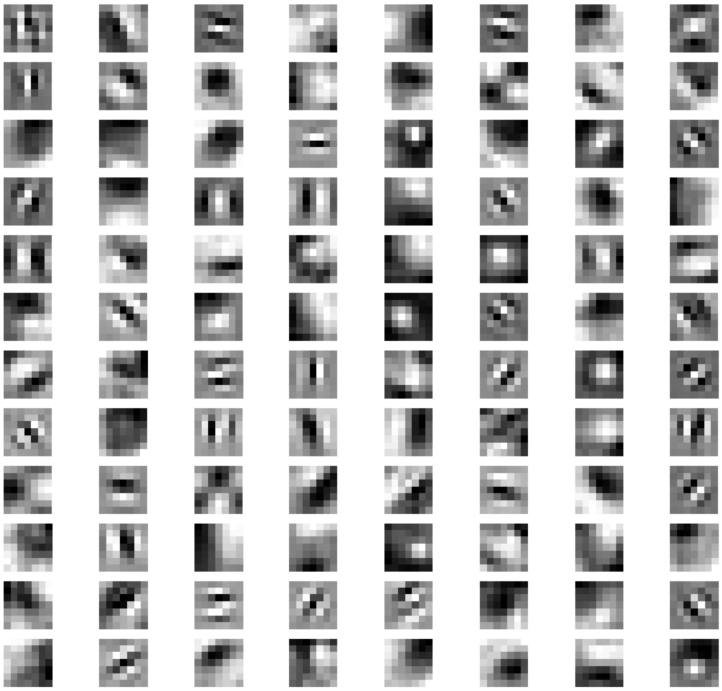


Fig. 24: Weights from Conv1 layer

These weights clearly are much more distributed than the original implementation all thanks to residual connections which helped backpropagate the error and not vanish it. The introduction of residual connections has proven to be instrumental in mitigating the issue of vanishing gradients during training. By facilitating the propagation of error signals through the network more effectively, residual connections have enabled the model to learn richer and more nuanced representations of the input data.

C. SqueezeNet with DSD traing approach

We implemented the Dense-Sparse-Dense (DSD) training approach from scratch for the SqueezeNet architecture on the CIFAR-10 dataset, leveraging insights from existing implementations such as the codebase for applying DSD to the LeNet architecture on the MNIST dataset. [11] Then, we conducted a series of experiments to investigate the impact of varying sparsity levels and dense-sparse-dense epoch distributions on the performance and efficiency of the trained SqueezeNet model.

- **Varying Sparsity Levels:** In our experimental investigation on varying sparsity levels, we explored the impact of different levels of sparsity on the performance of the SqueezeNet model trained using Dense-Sparse-Dense (DSD) training. By systematically adjusting the sparsity parameter, we conducted a series of experiments to evaluate the model's training and validation losses across a range of sparsity levels. Our results demonstrate a clear relationship between sparsity and model performance, with higher sparsity levels leading to increased training and validation losses. Furthermore, we observed that while sparser models tend to exhibit reduced computational complexity, there exists a trade-off between sparsity and model accuracy. These findings underscore the importance of carefully selecting the optimal sparsity level based on the desired balance between model size, computational efficiency, and performance metrics. Moreover, it is noteworthy that since our experiments were conducted using the SqueezeNet architecture, which, due to its design complexity, posed challenges when trained from scratch on the CIFAR-10 dataset. CIFAR-10, being a relatively small dataset compared to ImageNet, presented difficulties in achieving competitive results with SqueezeNet.
- **Different Epoch Distributions:**In our investigation of epoch distributions, we sought to understand the effect of varying the distribution of dense and sparse training epochs within the Dense-Sparse-Dense (DSD) training paradigm. By manipulating the number of dense and sparse epochs, we explored how the model's convergence behavior and final performance were influenced. Through a series of experiments with different epoch distributions, ranging from predominantly dense epochs to predominantly sparse epochs, we observed distinct patterns in the training and validation losses over the course of training. Specifically, we found that increasing the proportion of sparse epochs resulted in slower convergence rates and higher final losses, indicating the importance of maintaining a balance between dense and sparse training phases for effective model optimization. These findings provide valuable insights into the interplay between epoch distributions and training dynamics within the DSD framework, offering guidance for optimizing training strategies for neural network compression and efficiency.

VI. Conclusion & Discussions

In conclusion, our experiments into the vanilla SqueezeNet architecture and its variant incorporating residual connections and DSD training paradigm has provided valuable insights into the effectiveness of architectural enhancements in deep neural networks.

For the vanilla SqueezeNet, we observed its compact design and computational efficiency, making it suitable for resource-constrained environments like mobile and edge devices. Despite its simplicity, SqueezeNet achieved competitive performance on various tasks, demonstrating its potential for real-world deployment.

On the other hand, the integration of residual connections in SqueezeNet significantly enhanced its representational capacity. By enabling the network to learn residual mappings, these connections facilitated the training of deeper architectures while mitigating the vanishing gradient problem. Consequently, the SqueezeNet model with residual connections exhibited improved convergence properties and achieved higher accuracy on complex datasets.

The Dense-Sparse-Dense (DSD) training methodology, characterized by the iterative process of pruning and re-densing, is shown to significantly enhance optimization performance, particularly in the context of SqueezeNet. This approach effectively mitigates the challenge of navigating saddle points encountered during the optimization of deep neural networks by leveraging pruning to perturb the learning dynamics, facilitating movement away from these points. Similar to Simulated Annealing, DSD strategically deviates from converged solutions, inducing sparsity to enable escape from sub-optimal solutions. Furthermore, DSD fosters the attainment of superior minima by decreasing both training and validation losses, thus enhancing the robustness of SqueezeNet. The regularization inherent in sparse training shifts optimization to a smoother, lower-dimensional space, further augmenting the model's robustness against noise. Additionally, DSD offers robust re-initialization opportunities, mitigating the impact of weight initialization, and effectively breaking symmetry among hidden units, thereby reducing co-adaptation

risks. In conclusion, the application of DSD to SqueezeNet presents a compelling framework for neural network regularization, demonstrating remarkable optimization performance and underscoring its potential to significantly enhance the accuracy and robustness of the model.

In summary, our analysis underscores the importance of architectural innovations in shaping the performance and efficiency of deep neural networks. While the vanilla SqueezeNet offers a lightweight solution for deployment in resource-constrained scenarios, the incorporation of residual connections and DSD training methodology empowers the model to learn more expressive representations, leading to superior performance across various tasks.

VII. Future Work & Scope for Business

In the sense of future developments and business applications, SqueezeNet holds significant potential, particularly now that industries need models that are lightweight. For potential future applications in Deep learning, computational efficiency, and model size play pivotal roles. As the demand for such intelligent systems continues to surge exponentially across domains such as autonomous vehicles, IoT devices, and mobile applications, SqueezeNet's lightweight architecture presents a promising solution. Looking ahead, advancements in SqueezeNet, to make it even more efficient could focus on enhancing its adaptability to specialized hardware accelerators, optimizing its performance for low compute environments.

Moreover, the business scope for this model extends beyond traditional machine learning applications and reaches into sectors where real-time processing and minimal computational overhead are extremely crucial. Its competitive efficiency enables faster inference even on embedded devices with limited resources, making it an attractive choice for businesses seeking to deploy AI-powered solutions at scale. For example, in the health sector, SqueezeNet's ability to run efficiently on low and very-low power devices could facilitate the development of wearable technology such as health monitors, and assistive technologies for remote patient care. Similarly, in the automobile industry, its small footprint could enable integration of intelligent features into vehicles, enhancing driver safety without compromising performance. As SqueezeNet continues to evolve, its role in shaping the landscape of embedded AI applications is poised to expand, driving innovation and bringing new avenues for growth.

APPENDIX A
ALGORITHMS

Algorithm 1: Workflow of DSD training

Initialization: $W^{(0)}$ with $W^{(0)} \sim \mathcal{N}(0, 1)$
Output: $W^{(t)}$

Initial Dense Phase:
while not converged **do**
  $W^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla f(W^{(t-1)}; x^{(t-1)})$
  $t = t + 1$
end while

Sparse Phase:
Initialize the mask by sorting and keeping the Top-k weights.
$S = \text{sort}(|W^{(t-1)}|)$; $\lambda = S_k$
Mask $= \mathbb{1}(|W^{(t-1)|} > \lambda)$
while not converged **do**
  $W^{(t)} = W^{(t-1)} - \eta^{(t)} \nabla f(W^{(t-1)}; x^{(t-1)})$
  $W^{(t)} = W^{(t)} \odot \text{Mask}$
  $t = t + 1$
end while

Final Dense Phase:
while not converged **do**
  $W^{(t)} = W^{(t-1)} - \eta(t) \nabla f(W^{(t-1)}; x^{(t-1)})$
  $t = t + 1$
end while
Go to *Sparse Phase* for iterative DSD;

APPENDIX B
DERIVATION OF PARAMETERS IN ALEXNET

TABLE I: Parameter computation for AlexNet

Layer name/type	**Filter size**	**Bias**	**Total parameters**
First Convolutional Layer	$11 \times 11 \times 3 \times 96$	96	34,944
Second Convolutional Layer	$5 \times 5 \times 96 \times 256$	256	614,656
Third Convolutional Layer	$3 \times 3 \times 256 \times 384$	384	885,120
Fourth Convolutional Layer	$3 \times 3 \times 384 \times 384$	384	1,327,488
Fifth Convolutional Layer	$3 \times 3 \times 384 \times 256$	256	884,992
First Fully Connected Layer	$6 \times 6 \times 256 \times 4096$	4096	37,752,832
Second Fully Connected Layer	4096×4096	4096	16,781,312
Third Fully Connected Layer	4096×1000	1000	4,097,000
Total	-	-	62,380,346

Total Parameters: 62,380,346 (over 62 million parameters)

APPENDIX C
DERIVATION OF PARAMETERS IN SQUEEZENET

TABLE II: Parameter computation for SqueezeNet

Layer name/type	**Output size**	**Filter size/stride**	**Depth**	**s1x1**	**e1x1**	**e3x3**	**Total parameters (before pruning)**	**Total parameters (after pruning)**
Input image	224x224x3	-	-	-	-	-	-	-
conv1	111x111x96	7x7/2 (x96)	1	1	6-bit	14,208	14,208	
maxpool1	55x55x96	3x3/2	0	-	-	-	-	-
fire2	55x55x128	2	16	64	64	6-bit	11,920	5,746
fire3	55x55x128	2	16	64	64	6-bit	12,432	6,258
fire4	55x55x256	2	32	128	128	6-bit	45,344	20,646
maxpool4	27x27x256	3x3/2	0	-	-	-	-	-
fire5	27x27x256	2	32	128	128	6-bit	49,440	24,742
fire6	27x27x384	2	48	192	192	6-bit	104,880	44,700
fire7	27x27x384	2	48	192	192	6-bit	111,024	46,236
fire8	27x27x512	2	64	256	256	6-bit	188,992	77,581
maxpool8	13x12x512	3x3/2	0	-	-	-	-	-
fire9	13x13x512	2	64	256	256	6-bit	197,184	77,581
conv10	13x13x1000	1x1/1 (x1000)	1	1	6-bit	513,000	103,400	
avgpool10	1x1x1000	13x13/1	0	-	-	-	-	-
Total	-	-	-	-	-	-	1,248,424	421,098

Total Parameters: 1,248,424 (about 1.25 million parameters)

ACKNOWLEDGMENT

The entire project team would like to express our most sincere gratitude toward the professor, Dr. Predrag Jelenkovic who has been extremely supportive in helping us to complete this project, thereby improving our knowledge experience in the domain of Statistical Deep Learning.

REFERENCES

[1] Urja Banati et al. "Soft Biometrics and Deep Learning: Detecting Facial Soft Biometrics Features Using Ocular and Forehead Region for Masked Face Images". In: (Dec. 2021). DOI: 10.21203/rs.3.rs-1174842/v1.

[2] Wikimedia Commons. *File:Typical cnn.png — Wikimedia Commons, the free media repository*. https://commons.wikimedia.org/w/index.php?title=File:Typical_cnn.png&oldid=671717156. [Online; accessed 4-May-2024]. 2022.

[3] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: (2009), pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[4] Song Han et al. "DSD: Dense-Sparse-Dense Training for Deep Neural Networks". In: *Published as a conference paper at ICLR 2017* (2017). URL: https://arxiv.org/pdf/1607.04381.

[5] Xiaobing Han et al. "Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification". In: *Remote Sensing* 9.8 (2017). ISSN: 2072-4292. DOI: 10.3390/rs9080848. URL: https://www.mdpi.com/2072-4292/9/8/848.

[6] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[7] Forrest N. Iandola et al. *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size*. https://doi.org/10.48550/arXiv.1602.07360. 2016. arXiv: 1602.07360 [cs.CV].

[8] Ameya D. Jagtap. *Adaptive activation functions accelerate convergence in deep and physics-informed neural networks*. https://www.researchgate.net/figure/Left-to-right-Sigmoid-or-logistic-tanh-ReLU-and-Leaky-ReLU-activation-functions-for_fig2_333632820. [Online; accessed 4-May-2024]. 2019.

[9] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Tech. rep. Technical Report, University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/cifar.html.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[11] Ferdinand Mom. *DSD Training Repository*. https://github.com/3outeille/DSD-training/blob/master/src/mnist_dsd_pytorch.ipynb. 2021.

[12] Aili Wang et al. "A Dual Neural Architecture Combined SqueezeNet with OctConv for LiDAR Data Classification". In: *Sensors* 19.22 (2019), p. 4927. DOI: 10.3390/s19224927. URL: https://www.mdpi.com/1424-8220/19/22/4927.