

Beyond Skip Connection: Pooling and Unpooling Design for Elimination Singularities

Chengkun Sun¹, Jinqian Pan¹, Zhuoli Jin³, Russell Stevens Terry², Jiang Bian¹, Jie Xu¹

¹Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA

²Department of Urology, University of Florida, Gainesville, FL 32611, USA

³Department of Statistics and Applied Probability, University of California Santa Barbara, CA 93106-3110, USA
{sun.chengkun,jinqianpan,bianjiang,xujie}@ufl.edu, zhuoli.jin@pstat.ucsb.edu, russell.terry@urology.ufl.edu

Abstract

Training deep Convolutional Neural Networks (CNNs) presents unique challenges, including the pervasive issue of elimination singularities—consistent deactivation of nodes leading to degenerate manifolds within the loss landscape. These singularities impede efficient learning by disrupting feature propagation. To mitigate this, we introduce **Pool Skip**, an architectural enhancement that strategically combines a Max Pooling, a Max Unpooling, a 3×3 convolution, and a skip connection. This configuration helps stabilize the training process and maintain feature integrity across layers. We also propose the **Weight Inertia** hypothesis, which underpins the development of Pool Skip, providing theoretical insights into mitigating degradation caused by elimination singularities through dimensional and affine compensation. We evaluate our method on a variety of benchmarks, focusing on both 2D natural and 3D medical imaging applications, including tasks such as classification and segmentation. Our findings highlight Pool Skip’s effectiveness in facilitating more robust CNN training and improving model performance.

Code — <https://github.com/sunck1/PoolSkip>

Introduction

Convolutional Neural Networks (CNNs) are pivotal in advancing the field of deep learning, especially in image processing (Jiao and Zhao 2019; Razzak, Naz, and Zaib 2018; Miao et al. 2019, 2018). However, as these networks increase in depth to enhance learning capacity, they often encounter a notable degradation in performance (He et al. 2016a). This degradation manifests as a saturation point in accuracy improvements, followed by a decline, a phenomenon primarily driven by optimization challenges including vanishing gradients (He et al. 2016a). The introduction of Residual Networks (ResNets) with skip connections marked a significant advancement in mitigating these issues by preserving gradient flow during deep network training (He et al. 2016a; Orhan and Pitkow 2017).

Despite these advancements, very deep networks, such as ResNets with upwards of 1,000 layers, still face pronounced degradation issues (He et al. 2016a). A critical aspect of this problem is the elimination singularity (ES)—stages in the

training process where neurons consistently deactivate, producing zero outputs and creating ineffective nodes within the network (Orhan and Pitkow 2017; Qiao et al. 2019). This condition not only disrupts effective gradient flow but also significantly compromises the network’s learning capability. ES often results from zero inputs or zero-weight configurations in convolution layers, which are frequently observed due to the tendency of training processes to drive weights towards zero, contributing to excessively sparse weight matrices (Orhan and Pitkow 2017; Huang et al. 2020). Additionally, the widely used Rectified Linear Unit (ReLU) activation function exacerbates these issues by zeroing out all negative inputs (Qiao et al. 2019; Lu et al. 2019). This phenomenon, known as Dying ReLU, causes neurons to remain inactive across different data points, effectively silencing them and further complicating the training of deep networks (Qiao et al. 2019; Lu et al. 2019).

To address these persistent challenges, we developed **Pool Skip**, a novel architectural module that strategically incorporates Max Pooling, Max Unpooling, and a 3×3 convolution linked by a skip connection. This design is specifically engineered to counteract elimination singularities by enhancing neuron activity and preserving the integrity of feature transmission across network layers. Our approach not only aims to stabilize the learning process but also to enhance the robustness of feature extraction and representation in deep networks. The key contributions of our work are summarized below:

- We propose the **Weight Inertia** hypothesis to explain how persistent zero-weight conditions can induce network degradation. Based on this theory, we developed the Pool Skip module, which is positioned between convolutional layers and the ReLU function to help mitigate the risks associated with elimination singularities. We also provide mathematical proofs that demonstrate how Pool Skip’s affine compensation and dimensional Compensation optimize gradient fluctuations during the backpropagation process, thus addressing the degradation problem at a fundamental level.
- We evaluated the proposed Pool Skip module across various deep learning models and datasets, including well-known natural image classification benchmarks (e.g., CIFAR-10 and CIFAR-100), segmentation tasks (e.g., Pascal VOC and Cityscapes), and medical imaging chal-

lenges (e.g., BTCV, AMOS). Our findings validate the effectiveness of Pool Skip in reducing elimination singularities and demonstrate its capacity to enhance both the generalization and performance of models.

Related Work

Pooling Operations in CNNs

Max Pooling (Ranzato et al. 2007), a staple in CNN architectures, segments convolutional output into typically non-overlapping patches, outputting the maximum value from each to reduce feature map size (Dumoulin and Visin 2016; Gholamalinezhad and Khosravi 2020). This not only yields robustness against local transformations but also leverages the benefits of sparse coding (Boureau, Ponce, and LeCun 2010; Boureau et al. 2011; Ranzato et al. 2007). Its efficacy is well-documented in prominent CNN architectures like VGG (Simonyan and Zisserman 2014), YOLO (Redmon et al. 2016), and UNet (Ronneberger, Fischer, and Brox 2015), and is essential in numerous attention mechanisms, such as CBAM (Woo et al. 2018), for highlighting salient regions. Despite these advantages, Ruderman et al. (Ruderman et al. 2018) indicate that networks can maintain deformation stability without pooling, primarily through the smoothness of learned filters. Furthermore, Springenberg et al. (Springenberg et al. 2014) suggest that convolutional strides could replace Max Pooling, as evidenced in architectures like nnUNet (Isensee et al. 2021).

Max Unpooling, designed to reverse Max Pooling effects by restoring maximum values to their original locations and padding zeros elsewhere, complements this by allowing CNNs to learn mid and high-level features (Zeiler and Fergus 2014; Zeiler, Taylor, and Fergus 2011). However, the traditional “encoder and decoder” architecture, foundational to many modern CNNs like UNet (Ronneberger, Fischer, and Brox 2015), rarely adopts Max Unpooling due to concerns that zero filling can disrupt semantic consistency in smooth areas (Liu et al. 2023).

Our work reevaluates the conventional combination of Max Pooling and Max Unpooling, arguing that its effective utilization can still substantially benefit CNNs by focusing on significant features. Moreover, the finite number of layers in common encoder and decoder architectures limits the use of Max Pooling in deeply nested CNNs, posing challenges for deep-level information recognition.

Skip Connection and Batch Normalization

The concept of ES was first proposed by Wei et al. (Wei et al. 2008), to describe the issue of zero weights in the output of convolutional layers, a phenomenon commonly referred to as “weight vanishing” (Wei et al. 2008). This issue is particularly concerning because these zero weights do not contribute to the model’s calculations, leading to inefficiencies in learning processes. ES is deeply associated with slow learning dynamics and unusual correlations between generalization and training errors and presents a significant challenge in training deep neural networks effectively (Amari, Park, and Ozeki 2006).

To mitigate ES, two primary strategies have emerged: normalization, specifically Batch Normalization (BN) (Ioffe and Szegedy 2015), and skip connections (He et al. 2016a,b). BN helps maintain a stable distribution of activation values throughout training, while skip connections effectively increase network depth by preventing the elimination of singularities, ensuring that even with zero incoming or outgoing weights, certain layers maintain unit activation (Ioffe and Szegedy 2015; He et al. 2016a). This allows for the generation of non-zero features, making previously non-identifiable neurons identifiable, thereby addressing ES challenges (Orhan and Pitkow 2017). Despite these advancements, the degradation problem persists in extremely deep networks, even with the implementation of skip connections (He et al. 2016a). Further analysis by He et al. (He et al. 2016b) on various ResNet-1001 components—such as constant scaling, exclusive gating, shortcut-only gating, conv shortcut, and dropout shortcut—reveals that the degradation issue in the original ResNet block not only remains but is also exacerbated.

Pool Skip Mechanism

In this section, we introduce the Pool Skip mechanism, beginning with the Weight Inertia hypothesis that motivates its development. We then provide theoretical insights into how this hypothesis helps mitigate degradation caused by elimination singularities through dimensional and affine compensation.

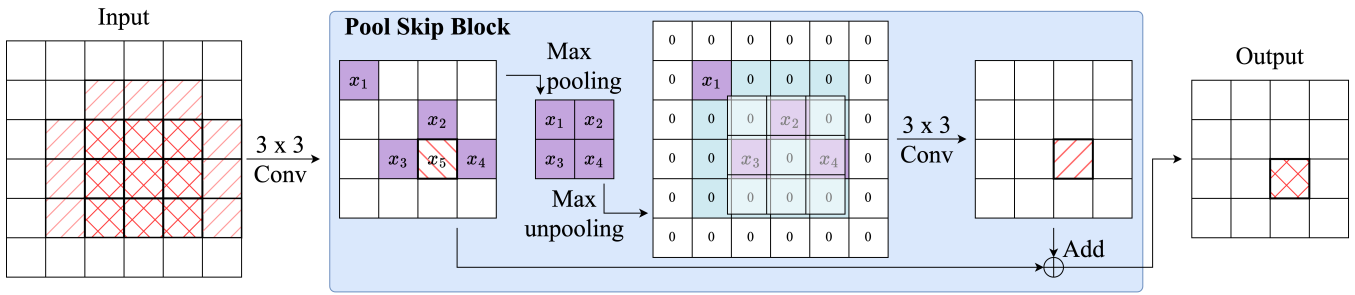
Weight Inertia Hypothesis

In the context of back-propagation (Rumelhart, Hinton, and Williams 1986), the process is defined by several essential components. L denotes the loss function, W represents the weights, Y refers to the output feature maps, X to the input feature maps, while c_{in} and c_{out} indicate the input and output channels, respectively. The $*$ is used for the convolution operation. The operation of back-propagation is captured as follows:

$$\begin{aligned} \frac{\partial L}{\partial W_{c_{in},c_{out}}} &= \frac{\partial L}{\partial Y_{c_{out}}} \cdot \frac{\partial Y_{c_{out}}}{\partial W_{c_{in},c_{out}}} \\ &= \frac{\partial L}{\partial Y_{c_{out}}} \cdot \frac{\partial \text{ReLU}(X_{c_{in}} * W_{c_{in},c_{out}})}{\partial W_{c_{in},c_{out}}}, \\ \frac{\partial L}{\partial X_{c_{in}}} &= \frac{\partial L}{\partial Y_{c_{out}}} \cdot \frac{\partial Y_{c_{out}}}{\partial X_{c_{in}}} \\ &= \frac{\partial L}{\partial Y_{c_{out}}} \cdot \frac{\partial \text{ReLU}(X_{c_{in}} * W_{c_{in},c_{out}})}{\partial X_{c_{in}}}. \end{aligned} \quad (1)$$

The activation function employed in this context is ReLU. During the convolution process, when the output $X_{c_{in}} * W_{c_{in},c_{out}}$ is less than or equal to zero, ReLU sets both derivatives, $\frac{\partial L}{\partial W_{c_{in},c_{out}}}$ and $\frac{\partial L}{\partial X_{c_{in}}}$ to zero, in accordance with its operational rules. Or if the weights themselves ($W_{c_{in},c_{out}}$) are zero, it would still result in zero gradients.

According to the standard gradient descent update rule, the weights are adjusted by subtracting a portion of the gradient from the current weight values: $\hat{W}_{c_{in},c_{out}} =$



Note:

1. Assume x_1, x_2, x_3, x_4 are maximum elements.
2. \square from unpooling, and \square from padding.
3. Areas affecting: With Pool Skip: $\square + \square$; Without Pool Skip: $\square \cdot (\square + \square) = \square$

Figure 1: Schematic representation of the computational process of Pool Skip.

$W_{c_{in}, c_{out}} - \eta \frac{\partial L}{\partial W_{c_{in}, c_{out}}}$, where η represents the learning rate. When training utilizes a fixed input space (a consistent set of training samples) and employs a diminishing learning rate towards the end of the training cycle, the updates to the weights in both the current and previous layers become minimal. This minimal update results in inputs $X_{c_{in}}$ and output $X_{c_{in}} * W_{c_{in}, c_{out}}$ at the current layer becoming inert. Consequently, the outputs $Y_{c_{in}}$ of subsequent layers also remain unchanged. The worst case is a continuous negative or zero output.

This stagnation, which we term **Weight Inertia**, results from sparse weights (i.e. ES) and consistent non-positive outputs (i.e Dying ReLU), particularly prevalent in what we refer to as the degradation layer. This layer is marked by a continuous inability to update zero weights, leading to a limited number of effective weights and exacerbating the degradation problem. This forms a self-reinforcing cycle: as weights fail to update, the network’s ability to learn and adapt diminishes, deepening the degradation. To break this cycle, controlling the negative outputs, specifically $X_{c_{in}} * W_{c_{in}, c_{out}}$, is crucial. By effectively managing these outputs, it is possible to interrupt the cycle, prevent further degradation, and enhance the network’s overall learning capabilities.

Motivated by the weight inertia hypothesis, we design Pool Skip, which consists of a Max Pooling layer, a Max Unpooling layer, and a 3×3 convolutional layer, tightly interconnected with skip connections spanning from the beginning to the end of the module. Figure 1 illustrates the computational process of Pool Skip. Initially, the Max Pooling layer prioritizes important features, facilitating the extraction of key information crucial for subsequent processing. Subsequently, the Max Unpooling layer ensures that the feature size remains consistent, preserving the gradient propagation process established by max-pooling. This characteristic allows Pool Skip to be seamlessly integrated into convolutional kernels at any position within the network architecture. Moreover, by selectively zeroing out non-key positions, Pool Skip effectively controls the magnitude of the compensatory effect, further enhancing its utility in stabilizing the training process.

Affine and Dimensional Compensation

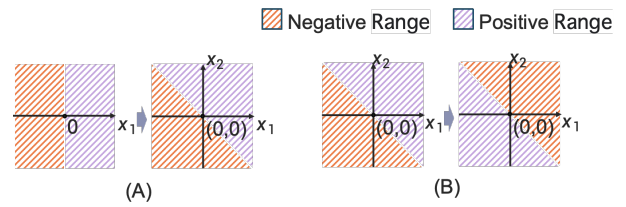


Figure 2: A simple example of dimensional and affine and compensation.

As discussed earlier in the weight inertia hypothesis, the output of a neural network is often predominantly determined by a linear combination of a few specific and influential input elements, despite the potential presence of numerous input elements. This selective influence suggests that modifying this fixed linear combination—either by activating input elements through changes in input dimensions or by adjusting the coefficients of the linear combination—can significantly impact the output. For convenience, we refer to these two adjustment mechanisms as dimensional compensation and affine compensation. A simple example illustrating this process is provided in Fig. 2. Initially, with one dimension x_1 , the region where $x_1 < 0$ is shown in orange on the left side of Fig. 2(A). Introducing a second dimension x_2 changes this to $x_1 + x_2 < 0$, shifting the orange area to the right side of Fig. 2(A). This demonstrates dimensional compensation. When the coefficient in front of x_1 and x_2 changes from 1 to -1, the shift in the orange area in Fig. 2(B) represents affine compensation. These compensations alter the negative region of the input space, thereby disrupting Weight Inertia.

Next, we will theoretically explain how Pool Skip introduces dimensional and affine compensation, subsequently affecting the output results. As depicted in Figure 1, we begin by establishing the input configuration based on the computational process of Pool Skip:

In the convolutional computation of a single layer, the output $y_{i,j}$ is derived from a linear combination of the input $x_{i,j}$ before the Pool Skip (Goodfellow, Bengio, and Courville

1. $X_{H \times W} = \{x_{i,j}\}_{H \times W}$: input matrix;
2. $W_{M \times M} = \{w_{i,j}\}_{M \times M}$: the convolutional kernel before Pool Skip. Assume W is $M \times M$ kernel, and M is an odd number;
3. $Y_{(H-M+1) \times (W-M+1)} = \{y_{i,j}\}_{(H-M+1) \times (W-M+1)}$: the output of first convolutional computation;
4. e : Max Pooling size which satisfies $e|H$, $e|W$, $e|H-M+1$ and $e|W-M+1$;
5. $A_{c \times d}$: the matrix obtained from max-pooling on Y . $c = (H-M+1)/e$ and $d = (W-M+1)/e$;
6. $\tilde{W}_{3 \times 3} = \{\tilde{w}_{i,j}\}_{3 \times 3}$: the convolutional kernel in the Pool Skip;
7. $O_{H-M+1, W-M+1} = \{o_{i,j}\}_{H-M+1, W-M+1}$: the output matrix.

2016):

$$y_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} w_{m,n} \times x_{i+m,j+n}. \quad (3)$$

It's important to note that the convolution kernel is not flipped in this context. Based on the Max-Pooling, we can divide the Y from previous computation by size $e \times e$ into $c \times d$ blocks. For each block of Y , indexed as $Y^{(u,v)}$, we have:

$$Y^{(u,v)} = \begin{pmatrix} y^{(ue,ve)} & \dots & y^{(ue,(v+1)e-1)} \\ \vdots & \ddots & \vdots \\ y^{((u+1)e-1,ve)} & \dots & y^{((u+1)e-1,(v+1)e-1)} \end{pmatrix} \quad (4)$$

$e \times e$,

where $u \in \{0, 1, \dots, c-1\}$, $v \in \{0, 1, \dots, d-1\}$. For each block $Y^{(u,v)}$, the maximum element is

$$\tilde{y}^{(u,v)} = \max_{i^{(u,v)}, j^{(u,v)} \in \{0, 1, \dots, e-1\}} Y_{i^{(u,v)}, j^{(u,v)}}^{(u,v)}, \quad (5)$$

with the corresponding $\tilde{i}_a^{(u,v)}$, $\tilde{j}_a^{(u,v)}$ as

$$(\tilde{i}_a^{(u,v)}, \tilde{j}_a^{(u,v)}) = \arg \max_{i^{(u,v)}, j^{(u,v)} \in \{0, 1, \dots, e-1\}} Y_{i^{(u,v)}, j^{(u,v)}}^{(u,v)}. \quad (6)$$

Therefore, we can write A as $A = \{\tilde{y}^{(u,v)}\}_{c \times d}$. And then via Max Unpooling and padding operations, we could get:

$$y'_{i,j} = \begin{cases} \tilde{y}_{(\tilde{i}_a^{(u,v)}, \tilde{j}_a^{(u,v)})}^{(u,v)}, & \text{if } e \bmod (i-1) = \tilde{i}_a^{(u,v)} \\ \text{and } e \bmod (j-1) = \tilde{j}_a^{(u,v)} & \text{in block } Y^{(u,v)} \\ 0, & \text{o.w.} \end{cases} \quad (7)$$

where $1 \leq i \leq H-M+1$ and $1 \leq j \leq W-M+1$. After passing the 3×3 convolutions, the result of the convolutional computation is shown as:

$$y_{out,i,j} = \sum_{s=0}^2 \sum_{t=0}^2 \tilde{w}_{s,t} \times y'_{i+s,j+t}. \quad (8)$$

After introducing the Pool Skip, the output $o_{i,j}$ is calculated by Eq. (2), where we denote set:

$$\begin{aligned} K_i &= \{(m, s) : ue + \tilde{i}_a^{(u,v)} + m \\ &\quad + s \in ([i, i+M] \cap [ue + \tilde{i}_a^{(u,v)}, ue + \tilde{i}_a^{(u,v)} + M + 2])\}, \\ L_j &= \{(n, t) : ve + \tilde{j}_a^{(u,v)} + n \\ &\quad + t \in ([j, j+M] \cap [ve + \tilde{j}_a^{(u,v)}, ve + \tilde{j}_a^{(u,v)} + M + 2])\}. \end{aligned} \quad (9)$$

Note that when $(m, s) \in K_i$, $ue + \tilde{i}_a^{(u,v)} + m + s = i + m$, and when $(n, t) \in L_j$, $ve + \tilde{j}_a^{(u,v)} + n + t = j + n$.

According to Eq. (2) (see detailed derivation in supplementary material), the original linear combination obtained two types of compensation. When the maximum value obtained after passing through the Pool Skip consists of the original linear combination elements x (from the input feature of the convolutional kernel before the Pool Skip), part of the x coefficients changed from $w_{m,n}$ to $(1 + \tilde{w}_{s,t}) \times w_{m,n}$, representing affine compensation. This change correlates closely with the weights of the convolutional kernel in Pool Skip. On the other hand, the remaining maximum values, which cannot be added to the original linear combination of x , expand the output dimensions (i.e. adding $\tilde{w}_{s,t} \times w_{m,n} \times x_{ue + \tilde{i}_a^{(u,v)} + m + s, ve + \tilde{j}_a^{(u,v)} + n + t}$), constituting dimensional compensation. This compensation also remains closely tied to the weights of the convolutional kernels in the Pool Skip.

These adjustments not only alter the contribution of input elements but also affect the negative range space of the output. This effectively breaks the constraints imposed by weight inertia, promoting diversity in output results and updating zero weights. Additionally, by adjusting the size of the Max Pooling and Max Unpooling kernels, we can control the number of maximum values, directly influencing the strength of the compensatory effects. Specifically, when the size of pooling kernels is 1, indicating only one convolution in the skip connection, every output element receives compensation. After receiving the compensatory effect, the original negative value range changes, allowing the original linear combination to output a non-zero effective value after ReLU. This activation enables neurons in the next layer to be activated during forward propagation and ensures that convolutional kernels with zero weights before ReLU receive gradient updates during backpropagation, thus alleviating the ES problem.

Experiments

We integrated the proposed Pool Skip into various deep networks and conducted comprehensive evaluations across common image tasks, including classification as well as natural image and medical image segmentation, utilizing diverse datasets for robust validation. All models were equipped with BN and ReLU or ReLU variations following the convolutions, ensuring a standardized architecture for comparison. Furthermore, all models were trained using a single NVIDIA A100 GPU with 80G of memory to maintain consistency in computational resources.

Final Output: $o_{i,j} = y_{i,j} + y_{out,i,j}$

$$= \begin{cases} \sum_{\mathbb{1}_{K_i}((m,s))=1 \text{ and } \mathbb{1}_{L_j}((n,t))=1} [(1 + \tilde{w}_{s,t}) \times w_{m,n} \times x_{i-1+m,j-1+n}] + \\ \sum_{\mathbb{1}_{K_i}((m,s)) \neq 1 \text{ or } \mathbb{1}_{L_j}((n,t)) \neq 1} (w_{m,n} \times x_{i-1+m,j-1+n} + \tilde{w}_{s,t} \times w_{m,n} \times x_{ue+\tilde{i}_a^{(u,v)}+m+s,ve+\tilde{j}_a^{(u,v)}+n+t}) \\ \text{if } e \bmod (i-1) = \tilde{i}_a^{(u,v)} \text{ and } e \bmod (j-1) = \tilde{j}_a^{(u,v)} \text{ in block } Y^{(u,v)}, \\ \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} w_{m,n} \times x_{i-1+m,j-1+n}, & \text{o.w.} \end{cases} \quad (2)$$

for all $i \in \{1, 2, \dots, H - M + 1\}$ and $j \in \{1, 2, \dots, W - M + 1\}$.

Image Classification

Datasets For the classification task, we utilized the **CIFAR** datasets (Krizhevsky, Hinton et al. 2009). CIFAR-10 comprises 60,000 color images categorized into 10 classes. CIFAR-100 consists of 60,000 images divided into 100 classes, with each class containing 600 images. The images are colored and share the same dimensions of 32×32 pixels.

Comparison Methods We evaluated the effectiveness of the Pool Skip across various CNN architectures, including MobileNet (Howard et al. 2017), GoogLeNet (Szegedy et al. 2015), VGG16 (Simonyan and Zisserman 2014), ResNet18 (He et al. 2016a), and ResNet34 (He et al. 2016a). To ensure robustness, each model was trained with 5 diverse seeds on the official training dataset, and the average and standard deviation of the Top-1 error from the final epoch were calculated on the official test dataset. Moreover, to assess the impact of the Pool Skip, it was implemented in each convolutional layer rather than solely in the first layer. All the training settings followed Devries and Taylor’s work (DeVries and Taylor 2017).

Experimental Results Our experimental findings, detailed in Table 1, showcase the performance enhancements observed across the CIFAR10 and CIFAR100 datasets. Notably, we observed moderate improvements ranging from 0.5 to 5.44 on CIFAR100 and from 0.03 to 2.74 on CIFAR10 in networks with fewer layers. However, the magnitude of improvement varied depending on the architecture of the network. Of particular significance was the notable enhancement observed for MobileNet upon the integration of the Pool Skip.

Natural Image Segmentation

Datasets For this task, we utilized **Cityscapes** (Cordts et al. 2016) and **PASCAL Visual Object Classes (VOC) Challenge (Pascal VOC)** (Everingham et al. 2010) datasets. Cityscapes offers a comprehensive insight into complex urban street scenes, comprising a diverse collection of stereo video sequences captured across streets in 50 different cities. Pascal VOC provides publicly accessible images and annotations, along with standardized evaluation software. For segmentation tasks, each test image requires predicting the object class of each pixel, with “background” designated

if the pixel does not belong to any of the twenty specified classes.

Model	CIFAR100	CIFAR10
	Top-1 error (%) ↓	Top-1 error (%) ↓
MobileNet	33.75 ± 0.24	9.21 ± 0.19
+ours	28.31 ± 0.23 -5.44	6.47 ± 0.20 -2.74
GoogleNet	22.95 ± 0.24	5.35 ± 0.19
+ours	22.36 ± 0.32 -0.59	5.19 ± 0.14 -0.16
VGG16	27.84 ± 0.38	6.24 ± 0.18
+ours	27.23 ± 0.21 -0.61	5.90 ± 0.24 -0.34
ResNet18	24.06 ± 0.18	5.17 ± 0.15
+ours	23.32 ± 0.14 -0.74	5.10 ± 0.14 -0.07
ResNet34	22.69 ± 0.18	4.89 ± 0.07
+ours	22.19 ± 0.22 -0.50	4.86 ± 0.06 -0.03

Table 1: Top-1 error rates (Mean ± Std) for image classification on CIFAR 100 and CIFAR 10 datasets.

Comparison Methods We evaluated the effectiveness of the Pool Skip on DeepLabv3+ models (Chen et al. 2017, 2018), utilizing ResNet101 (He et al. 2016a) and MobileNet-v2 (Sandler et al. 2018) backbones. The Pool Skip was exclusively employed in the convolution of the head block, the first convolution of the classifier, and all the atrous deconvolutions to validate its compatibility with atrous convolution. The models were trained using the official training data and default settings in (Chen et al. 2017, 2018), and the Intersection over Union (IoU) of the final-epoch model was recorded on the official validation data. Five seeds were selected to calculate the mean and standard deviation of the results.

Experimental Results As illustrated in Table 2, our experiments demonstrated a modest improvement in mIoU ranging from 0.16% to 0.53% for DeepLabv3+ models, considering the incorporation of only five layers of the Pool Skip (one in the Deeplab head, three in the atrous deconvolutional layers, and one in the classifier). This indicates the compatibility of the Pool Skip with atrous deconvolutions.

Medical Image Segmentation

Datasets We used abdominal multi-organ benchmarks for medical image segmentation, i.e., **AMOS** (Ji et al.

Model	Cityscapes	Pascal VOC
	mIoU (%) \uparrow	mIoU (%) \uparrow
DLP_MobileNet	71.72 \pm 0.49	66.40 \pm 0.37
+ours	71.96 \pm 0.35 +0.24	66.93 \pm 0.49 +0.53
DLP_ResNet101	75.59 \pm 0.30	74.83 \pm 0.56
+ours	75.89 \pm 0.10 +0.30	74.99 \pm 0.23 +0.16

Table 2: The mIoU (Mean \pm Std) results for natural image segmentation on the Cityscapes and Pascal VOC datasets. ‘‘DLP’’ denotes ‘‘DeepLabv3+’’.

2022) and Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) (BA et al. 2015) datasets. AMOS is a diverse clinical dataset offering 300 CT (Computed Tomography) scans and 60 MRI (Magnetic Resonance Imaging) scans with annotations. The public BTCV dataset consists of 30 abdominal CT scans sourced from patients with metastatic liver cancer or postoperative abdominal hernia.

Comparison Methods We evaluated the Pool Skip using nnUNet (Isensee et al. 2021) and V-Net (Milletari, Navab, and Ahmadi 2016). For nnUNet, our implementation closely follows the nnUNet framework (Isensee et al. 2021), covering data preprocessing, augmentation, model training, and inference. Scans and labels were resampled to the same spacing as recommended by nnUNet. We excluded nnUNet’s post-processing steps to focus on evaluating the model’s core segmentation performance. For a fair comparison, when reproducing nnUNet, we retained its default configuration. For V-Net (Milletari, Navab, and Ahmadi 2016), we adopted the preprocessing settings consistent with nnUNet.

For the BTCV dataset, 12 scans were assigned to the test set, and 18 to the training and validation set. From AMOS, 360 scans (containing CTs and MRIs) were divided into 240 for training and validation and 120 for testing, with a training-to-validation ratio of 4:1. We performed 5-fold cross-validation on all models, averaging their softmax outputs across folds to determine voxel probabilities. Our evaluation is based on the Dice Score (Milletari, Navab, and Ahmadi 2016), Normalized Surface Dice (NSD) (Nikolov et al. 2018), and mIoU metrics.

For the V-Net model, we only added the Pool Skip to the first two encoders due to the odd size of feature map outputs by the final encoder. As for the nnUNet model, we applied the Pool Skip in each convolutional layer on all encoders when training on the BTCV dataset. When training on the AMOS dataset, Pool Skips were employed on all encoders except for the final encoder.

Experimental Results As illustrated in Table 3, this Pool Skip architecture applies to networks for 3D medical imaging segmentation tasks. Enhancement on V-Net and nnUNet demonstrate the Pool Skip’s efficacy for complex image segmentation tasks.

Further Analysis

Efficacy of Pool Skip in Deep CNNs To assess the effectiveness of the Pool Skip in deep CNNs, we utilized

Model	BTCV		
	DICE (%) \uparrow	NSD (%) \uparrow	mIoU (%) \uparrow
V-Net	78.32	70.77	68.08
+ours	79.70 +1.38	72.35 +1.58	69.48 +1.40
nnUNet	81.52	76.10	72.14
+ours	82.47 +0.95	77.00 +0.90	73.08 +0.94
Model	AMOS		
	DICE (%) \uparrow	NSD (%) \uparrow	mIoU (%) \uparrow
V-Net	77.15	62.97	66.32
+ours	80.02 +2.87	67.32 +4.35	69.81 +3.49
nnUNet	89.75	85.58	83.13
+ours	89.78 +0.03	85.52 -0.06	83.13 +0

Table 3: Performance results of medical image segmentation on two datasets.

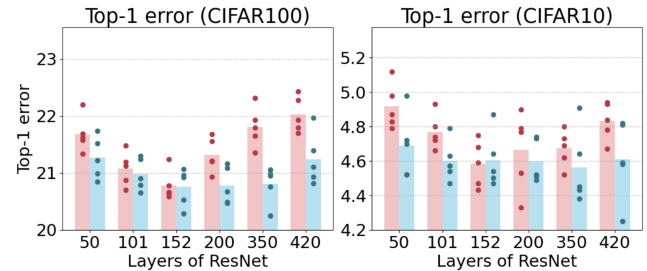


Figure 3: Top-1 error rates (%) of deep ResNet on CIFAR10 and CIFAR100 datasets. The pool kernel size is 4 for CIFAR100 experiments and 2 for CIFAR10 experiments.

ResNet (He et al. 2016a) as our baseline architecture. Our experiments covered a range of network depths, including 50, 101, 152, 200, 350, and 420 layers. For networks with fewer than 152 layers, we adhered to the original ResNet architecture as proposed by He et al. (He et al. 2016a). For deeper architectures (i.e., 200, 350, and 420 layers), we followed the architectural specifications outlined in a prior study (Bello et al. 2021). Training settings were consistent with those outlined in Devries and Taylor’s work (Devries and Taylor 2017).

The performance of the models across varying network depths is depicted in Figure 3. Initially, as the number of layers increases, the model’s performance improves before reaching a peak. The original ResNet achieves its best performance with 152 layers, achieving a top-1 error of 4.584 on CIFAR10 and 20.78 on CIFAR100. Subsequently, performance deteriorates rapidly. However, upon integrating the Pool Skip, performance improves, with a lower top-1 error of 4.562 on CIFAR10 with 350 layers and 20.752 on CIFAR100 with 152 layers. After stabilizing, performance begins to decline again. Nevertheless, there is a noticeable improvement in top-1 error (0.1-0.25 on CIFAR10 and 0.8-1 on CIFAR100) in deep networks. This underscores the efficacy of Pool Skip in enhancing the performance of deep CNNs. Thus, Pool Skip demonstrates promise in mitigating the degradation phenomenon that both BN and ResNet struggle to address, particularly with the increase in model depth.

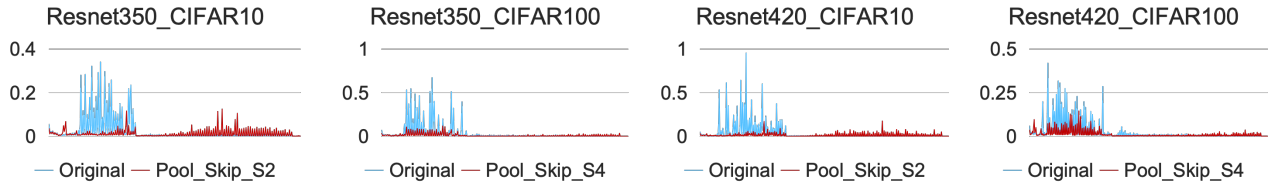


Figure 4: $\frac{l_2}{l_1}$ value quantitative comparison in ResNet350 and ResNet420 on CIFAR10 and CIFAR100 Datasets. The $\frac{l_2}{l_1}$ values were computed based on the output sequence of the network, with and without the incorporation of the Pool Skip. The plot highlights a moderate alleviation of the network degradation issue in shallow layers upon the integration of Pool Skip. Note: The horizontal axis represents the layers of the network along the output direction, from left to right. The “Pool_Skip_S4” means the size of Pool operation kernel is 4, “Pool_Skip_S2” does 2.

Efficacy of Mitigating Elimination Singularities (ES)

To assess weight sparsity, we calculated the $\frac{l_2}{l_1}$ ratio for each layer with and without the Pool Skip, as proposed by Hoyer et al. (Hoyer 2004). The higher the $\frac{l_2}{l_1}$ value is, the more zero the weights may contain (Wang et al. 2020; Hoyer 2004). Figure 4 illustrates the $\frac{l_2}{l_1}$ curve. Original ResNet350 and ResNet420 architectures suffered from severe degradation issues in shallow layers, despite the inclusion of batch normalization. However, integrating the Pool Skip noticeably alleviated this problem. Specifically, the $\frac{l_2}{l_1}$ ratios for ResNet350 decreased from a maximum of 0.7 to 0.1 on CIFAR100 and from 0.35 to 0.1 on CIFAR10. Similarly, for ResNet420, the ratios decreased from a maximum of 0.4 to 0.15 on CIFAR100 and from 1 to 0.15 on CIFAR10. This underscores the efficacy of the Pool Skip in mitigating elimination singularities, thereby enhancing model stability and maintaining feature integrity across layers.

Ablation Experiments We conducted ablation experiments with VGG16, as shown in Table 4, to assess the contribution of each block to the Pool Skip. Skip connections proved to be the most critical, with their removal causing a 13% increase in Top-1 error. Using only convolutions and skip connections increased the Top-1 error by 3%, while using only skip connection and pool design had minimal impact on performance.

VGG16	+{Pool, Skip}	+{Conv, Skip}	+{Pool, Conv}	+{Ours}
27.84	27.88	30.82	44.07	27.23

Table 4: Top-1 error rates (%) of ablation experiments on VGG16: “Pool” denotes Max Pooling and Max Unpooling, “Conv” represents 3×3 convolutions, and “Skip” indicates skip connections.

Discussion

While Pool Skip holds promise for mitigating the elimination singularities issue in convolutional kernels and has demonstrated effectiveness across extensive datasets and network architectures, the proposed structure still has some limitations. Firstly, since each convolutional kernel is followed by a 3×3 convolution, the overall number of parameters in the network increases significantly, thereby adding

a burden to both training and inference processes. Additionally, the sizes of Max Pooling kernels used in the experiments with deep ResNet are 2 and 4, while in compatibility experiments, it is 2. Additionally, the effectiveness of dimensional and affine compensations needs to be optimized through the adjustments of the pooling size in the Pool Skip.

Model	CIFAR100	CIFAR10
	Top-1 error (%) ↓	Top-1 error (%) ↓
ViT	25.37 ± 0.17	6.85 ± 0.23
+ours	25.07 ± 0.33 -0.30	7.07 ± 0.27 +0.22
CCT	19.16 ± 0.20	3.87 ± 0.19
+ours	18.61 ± 0.12 -0.55	3.89 ± 0.14 +0.02
CVT	22.92 ± 0.38	5.98 ± 0.19
+ours	22.80 ± 0.40 -0.12	6.36 ± 0.36 +0.38

Table 5: Top-1 error rates (Mean ± Std) of ViT-based models.

On the other hand, ViTs generally rely on one or more convolutional layers to generate image patches. To explore potential improvements in performance, we experimented with integrating the pool skip directly into the patch generation process. Our goal was to observe how this modification influences the overall effectiveness of the ViTs. We evaluated three ViT-based models including original ViT (Dosovitskiy et al. 2020), CCT (Hassani et al. 2021) and CVT (Wu et al. 2021) on the CIFAR10 and CIFAR100 datasets. According to Table 5, pool skip gains from 0.12 to 0.55 reduction in Top-1 error on CIFAR100, but from 0.02 to 0.38 deterioration. We believe that the observed performance changes are likely due to the pooling layer’s ability to extract key features, leading to performance improvements. However, the potential information loss caused by pooling may also contribute to performance deterioration. This also represents a limitation in the application of this structure.

Conclusion

In this paper, we proposed Pool Skip, a simple architectural enhancement addressing elimination singularities in deep CNN training. Based on the Weight Inertia hypothesis, it offers affine and dimensional compensation, stabilizing training. Experiments on various datasets and models confirm its effectiveness in optimizing CNNs.

References

- Amari, S.-i.; Park, H.; and Ozeki, T. 2006. Singularities affect dynamics of learning in neuromanifolds. *Neural computation*, 18(5): 1007–1065.
- BA, L.; Z, X.; JE, I.; M, S.; TR, L.; and A, K. 2015. Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge.
- Bello, I.; Fedus, W.; Du, X.; Cubuk, E. D.; Srinivas, A.; Lin, T.-Y.; Shlens, J.; and Zoph, B. 2021. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34: 22614–22627.
- Boureau, Y.-L.; Le Roux, N.; Bach, F.; Ponce, J.; and LeCun, Y. 2011. Ask the locals: multi-way local pooling for image recognition. In *2011 international conference on computer vision*, 2651–2658. IEEE.
- Boureau, Y.-L.; Ponce, J.; and LeCun, Y. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 111–118.
- Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.
- Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.
- DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dumoulin, V.; and Visin, F. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.
- Gholamalinezhad, H.; and Khosravi, H. 2020. Pooling methods in deep neural networks, a review. *arXiv preprint arXiv:2009.07485*.
- Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.
- Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; and Shi, H. 2021. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645. Springer.
- Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9).
- Huang, Z.; Shao, W.; Wang, X.; Lin, L.; and Luo, P. 2020. Convolution-weight-distribution assumption: Rethinking the criteria of channel pruning. *arXiv preprint arXiv:2004.11627*.
- Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.
- Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.
- Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhang, L.; Ma, W.; Wan, X.; et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732.
- Jiao, L.; and Zhao, J. 2019. A survey on the new generation of deep learning in image processing. *Ieee Access*, 7: 172231–172263.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Liu, W.; Lu, H.; Fu, H.; and Cao, Z. 2023. Learning to Upsample by Learning to Sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6027–6037.
- Lu, L.; Shin, Y.; Su, Y.; and Karniadakis, G. E. 2019. Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*.
- Miao, X.; Yuan, X.; Pu, Y.; and Athitsos, V. 2019. 1-net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4059–4069.
- Miao, X.; Zhen, X.; Liu, X.; Deng, C.; Athitsos, V.; and Huang, H. 2018. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5040–5049.
- Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical

- image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.
- Nikolov, S.; Blackwell, S.; Zverovitch, A.; Mendes, R.; Livne, M.; De Fauw, J.; Patel, Y.; Meyer, C.; Askham, H.; Romera-Paredes, B.; et al. 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*.
- Orhan, A. E.; and Pitkow, X. 2017. Skip connections eliminate singularities. *arXiv preprint arXiv:1701.09175*.
- Qiao, S.; Wang, H.; Liu, C.; Shen, W.; and Yuille, A. 2019. Rethinking normalization and elimination singularity in neural networks. *arXiv preprint arXiv:1911.09738*.
- Ranzato, M.; Boureau, Y.-L.; Cun, Y.; et al. 2007. Sparse feature learning for deep belief networks. *Advances in neural information processing systems*, 20.
- Razzak, M. I.; Naz, S.; and Zaib, A. 2018. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, 323–350.
- Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.
- Ruderman, A.; Rabinowitz, N. C.; Morcos, A. S.; and Zoran, D. 2018. Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs. *arXiv preprint arXiv:1804.04438*.
- Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.
- Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.
- Wang, C.; Yan, M.; Rahimi, Y.; and Lou, Y. 2020. Accelerated schemes for the L_1/L_2 minimization. *IEEE Transactions on Signal Processing*, 68: 2660–2669.
- Wei, H.; Zhang, J.; Cousseau, F.; Ozeki, T.; and Amari, S.-i. 2008. Dynamics of learning near singularities in layered networks. *Neural computation*, 20(3): 813–843.
- Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.
- Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22–31.
- Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833. Springer.
- Zeiler, M. D.; Taylor, G. W.; and Fergus, R. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision*, 2018–2025. IEEE.