Effect of Length of Skip-Connection of ResNet on its model complexity

Lokesh Das
Computer Science
University of Memphis
ldas@memphis.edu

Hunter Gore
Physics & Material Science
University of Memphis
hgore@memphis.edu

Bhavika Khare
Computer Science
University of Memphis
bbkhare@memphis.edu

Abstract

In this paper we try to find out the relationship between the skip-connection length in the ResNet-50 and the complexity of the ResNet-50 in terms of its Effective Model Complexity (EMC) (Nakkiran et al., 2020). we know that adding skip-connections to a convolutional neural netowrk (CNN) increases its complexity; however, finding a relationship between model complexity and skip-connection length can provide insight in to the optimal parameters in a residual network.

To study this relationship, we varied skip-lengths of the ResNet-50 among 5 different values and plotted the EMC versus skip-length graph for the 5 skip-lengths. This graph showed great variation but no predictable pattern, and we believe that a relationship can be gleaned upon further study of this problem using a larger dataset that helps us capture the truly large complexities of ResNet-50 models.

1 Introduction

Machine learning solutions such as deep learning models are ground-breaking for many applications due to their superb performances. However, the model complexity of deep learning is an important fundamental concern that has not yet been addressed sufficiently.

Model complexity defines how nonlinear and complex a function a given model with given parameters can learn or express. The Effective Model Complexity (EMC) is a special measure of model complexity that we use, introduced in (Nakkiran et al., 2020), that expresses the ability of a given procedure to learn a given distribution in the terms of the training error it yields with respect to dataset size. We also know that different factors, i.e., model architectures, data distribution, data complexity, dataset size, can affect the model complexity (Hu et al., 2021). This project studies one such factor and its influence on ResNet-50 EMC.

Although the depth of a learning network is crucial for specific tasks such as visual recognition, is better learning as simple as stacking more layers? This question motivated the advent of Residual Neural Networks or ResNets: CNNs where "skip-connections" allow propagating data to skip certain layers if it results in better results. In this project, we will investigate the effective model complexity of ResNet50 as a function of its skip-connection length.

1.1 Convolutional Neural Network

Convolutional Neural Networks (CNN) are powerful tools for image classification tasks. The fundamental idea is for each layer of the network to learn more detailed features of an image as more convolutional layers are added. Each convolutional layer passes a filter over subsections of an image to compress the image into a lower-dimensional array known as a feature map. The ’convolution’ is the operation of performing a dot product between the image subset and the filter, so the calculations required for CNNs are relatively simple matrix multiplication. CNN consists of 3 core elements: a convolutional, pooling, and fully-connected (FC) layer. The convolutional layer determines important features of the image, and the pooling layer, otherwise known as downsampling or subsampling layer, reduces the size of the input image to reduce the number of parameters. The most common pooling operations are max pooling, which selects the maximum pixel value from each group, thus reducing the size of images. The average pooling takes the average of each pixel of this group. The fully connected layer uses the features learned from the convolutional layer on the inputs to classify the features in each image.

1.2 Residual Neural Networks

Residual Networks (ResNets) are CNNs that can handle an exceptionally large number of layers and yet avoid the vanishing gradient and model performance degradation problem. They do so by reformulating the problem such that the stacked layers approximate the identity function with respect to the input instead of approximating the function itself. ResNets perform identity shortcuts if the input and output dimensions are the same. When the dimensions increase because of Convolutional-Batch Normalization operations, the skip-connection performs either identity mapping with extra zero entities padded to increase the dimension introducing no extra parameters or shortcut projections done by 1x1 convolution. The makers of the first ResNets, in (He et al., 2015), conjectured rightly that these identity connections are closer to zero and easier to learn than the original functions.

1.3 Skip-connections

Skip-connections(or shortcut-connections) are network connections that skip the one or more layers in a ResNet 'building block' as shown in Fig 1.A ResNet is comprised of several such blocks. The skip-connections directly map the input to the output, which is called identity mapping, and the outputs of the skip-connection are added to the output of the stacked layers in the building block as shown in figure 1. The skip-connection can allow deeper layers to act like shallower layers, and the network can pass all of its learned features to deeper layers. These connections add neither extra parameters nor computational complexity and cost us nothing. The only potential restriction they cause to the network is to have the same dimensions as the convolutional layers they skip.

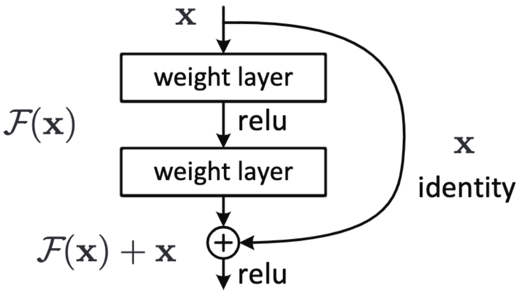
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1: Residual Learning: a building block(He et al., 2015)

Typically, skip connections are small in length, but their ideal length is not prescribed. We suppose there must be some optimal length between the two extremes of skip-connections that skip all the layers and skip-connections that skip none, and the search for this optimum motivates our efforts. How does skip-connection length affect the complexity? Quantifiably, how much more complex or robust is a network with skip connections than without? These are the questions explored here.

1.4 Effective Model Complexity (EMC)

As introduced by (Nakkiran et al., 2020), the EMC is a classifier complexity measure that defines model complexity as the maximum number of training samples on which it achieves close to zero training error. It is defined concerning given data distribution and a fixed error-tolerance threshold. The figure 3 shows how the value of a model's EMC is pinpointed on its error as a function of the training size for a given error-tolerance threshold e. We chose this measure of model complexity for our experiments due to its logical and empirical definition and unique properties: it can capture deep-learning phenomena, i.e., deep double descent. Thus, it appealed to us for its potential to get insightful results in the future.

Algorithm 1 EMC Calculation

```
$dataSize \leftarrow [10k, 20k, 30k, 40k, 50k]$
$no\_epochs \leftarrow 300$
$train\_error \leftarrow 0.0$
**while** $train\_error \leq 0.1$ **do**
    $train\_error \leftarrow Model.train(data\_subset)$    ▷ Model.train() returns training error after $n\_epochs$ epochs
**end while**
```

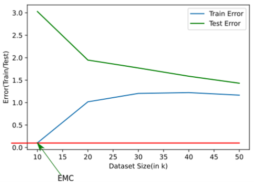

Figure 2: EMC calculation for our ResNet implementation and data distribution

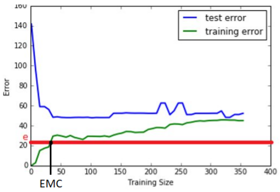

Figure 3: EMC calculation for any deep learner

2 Related Work

Deep Residual Networks were first published about in (He et al., 2015) for image classification tasks. The authors suggested that deeper networks contain shallower networks and so should have the shallow networks' solution space as a subset of their own solution spaces, and so they should perform at least as well as shallow networks. Their proposed methodology introduced skip-connections in convolutional neural networks and outperformed all contemporary networks designed for image classifications. This technique can handle the problem of performance degradation usually faced by deep networks. The question of how skip-connection length affects model complexity first came to us in connection with the Effective Model Complexity concept introduced in (Nakkiran et al., 2020). That paper made a fascinating study of the problem of overfitting in complex models. A similar study on how Convolution Networks are affected by increasing depth is performed by (Nichani et al., 2020). Note that increasing width is widely studied and known to decrease test error. They found that as the depth of ResNets with fixed skip connection lengths and convolutional neural tangent kernel networks increases, their test risk monotonically decreases, reaches a minimum, and increases in a U-shaped curve. In (Li and He, 2018), the authors study a closely related problem — they simplify the ResNet to understand how exactly skip connections solve the diminishing gradient problem. In this process, they also proposed an improved ResNet to redefine shortcut connections with an adjustable parameter 'k'. This parameter does not affect the skip-connection length, differentiating our research from theirs. This improved ResNet of theirs is comparable in computational complexity to the original but gives better results. In (Wójcik et al., 2020) and (O'Neill et al., 2020), they study methods of "compressing" deep learners, specifically ResNets, such that complexity goes down and performance stays as-is. Finally, (Allen-Zhu and Li, 2019) studies the causes of the incredible accuracy of deeper networks and proves a significant computation complexity advantage held by ResNets.

3 Datasets

In this experiment, we use two popular image datasets: CIFAR-10 and CIFAR-100. CIFAR-10 dataset has 60,000 32x32 color images of 10 classes, and each class contains 6000 images. CIFAR-10 dataset contains 50,000 training images, whereas the remaining 10,000 are for testing, and these 1000 images are selected randomly from each class. Like CIFAR-10, the CIFAR-100 dataset also contains 60,000 32x32 color images; however it has 100 classes and 600 images in each class. There are 500 training and 100 testing images per class. The table 1 shows the statistics of the datasets. For this preliminary experiment, we have used CIFAR-10 datasets by varying their sizes when needed for EMC calculation. The table 2 shows how we select appropriate-sized subsets from this dataset.

Dataset	**Classes**	**Train Size**	**Test Size**	**Total**
CIFAR-10	10	50,000	10,000	60,000

Table 1: Dataset statistics(Krizhevsky et al., 2009)

4 Methodology

ResNet skip-connection length is optimized differently in different networks, but prior research has not identified a way to find the optimal length. Our goal is to investigate how modulating the length of skip connections affects the Effective Model Complexity (Nakkiran et al., 2020) of ResNets. We assumed that networks that skip almost all layers or networks that skip almost no layers would have sub-optimal complexity. We expected that complexity would follow a U-shaped curve as the number of skip layers increased. Our method was to essentially determine the model complexity for each different skip-connection length. The number of convolutional layers per skip defines the length of the skip-connection. By measuring the EMC for skip-lengths 2, 3, 4, ... we were able to judge if there is an ideal number of skip-connections for image recognition tasks.

To determine the EMC of each model, we trained it on datasets of sizes 10k, 20k, 30k, 40k and 50k to find the largest dataset size where training error was below our fixed threshold.

5 Experiment

5.1 Implementing the ResNet-50

We implemented our version of the ResNet50 that allows for variation in skip-connection lengths, as existing ResNet libraries and APIs do not offer this functionality, and thus our implementation is from scratch. In particular, we implemented five different versions of ResNet with skip-connection of 2, 3, 4, 6, and 8 and trained them on the CIFAR-10 dataset. The figure 4 shows how ResNet looks with varying length of skip connections.

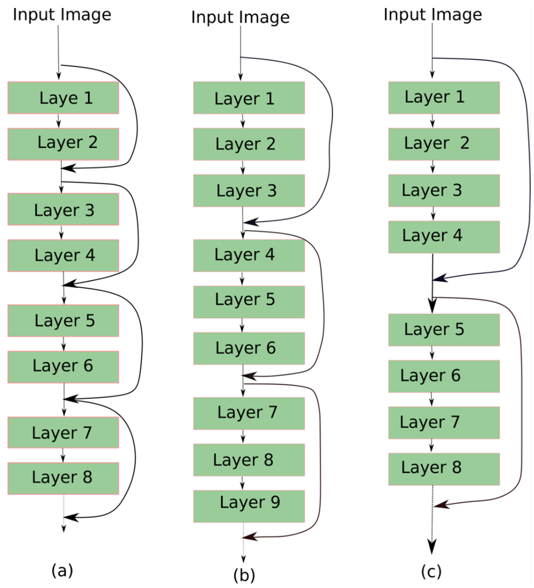

Figure 4: ResNets with varying lengths of skip connections.

5.2 EMC Calculation

We calculated the Effective Model Complexity (EMC) for our ResNet implementations using the algorithm 2 using different subsets of various sizes of original CIFAR-10 dataset(Krizhevsky et al., 2009). We made five subsets of sizes 10k, 20k, ..., 50k for preliminary results. As the graph in 2 shows, for the 10k-sized dataset, the training error was very close to 0.1; thus, the EMC of our ResNet — with skip-connection length unchanged from the default value — evaluated to the data size of 10,000 images and corresponding training and testing errors are listed in table 2. This code is reusable for future EMC calculations that will allow us to plot Complexity (EMC) against skip-connection length in a graph, thereby completing the experiment.

Algorithm 2 Algorithm to plot EMC against skip-connection length

```
$x \leftarrow 0$
$y \leftarrow 0$
$i \leftarrow 0$
**while** $i < 50$ **do**
    $i \leftarrow i + 1$
    $y \leftarrow ResNet50(i).EMC\_Calculation()$
    $plot(x, y)$
**end while**
```

5.3 Results

We ran five ResNet-50s with skip-lengths of 2, 3, 4, 6, and 8 on the CIFAR-10 dataset. For each network we measured the EMC, which gives us 5 data points in our final graph. In this experiment, we ran 25 ResNets using the Adam optimizer with the default learning rate (0.001). Other hyperparameters are (i) batch size of 1024, (ii) Batch Normalization which acclerates deep network training (Ioffe and Szegedy, 2015), (iii) ReLu activation throughout the network and softmax activation at the output layer, and (iv) categorical cross-entropy loss function. In addition, we used early stopping with patience=5, which monitors the training loss to determine the optimal number of epochs. We have used a slightly bigger batch size compared to the dataset size. This is because training a ResNet takes a massive amount of time due to the very deep nature of the model architecture, and we needed to train 25 models. Another practical reason is the limitation of computing resources we had available.

Dataset Size	Train Error	Test Error
10k	**0.1023**	**3.0292**
20k	1.0179	1.9476
30k	1.2045	1.7716
40k	1.2227	1.5870
50k	1.1657	**1.4319**

Table 2: Train/Test errors for different data sizes.

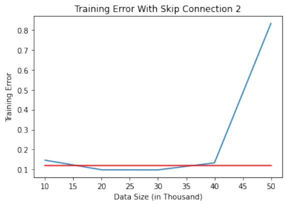

Figure 5: EMC with skip connection length 2

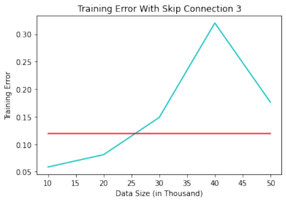

Figure 6: EMC with skip connection length 3

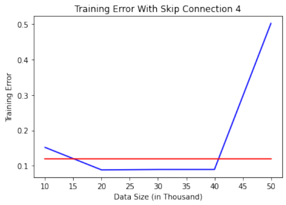

Figure 7: EMC with skip connection length 4

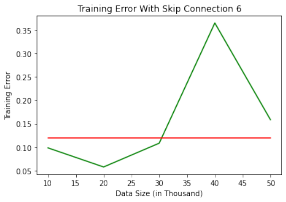

Figure 8: EMC with skip connection length 6

We set the interpolation threshold 0.12, unlike the results shown in figure 2, for computing the model complexity for each ResNet50 which is very intuitive as there is no principled way to

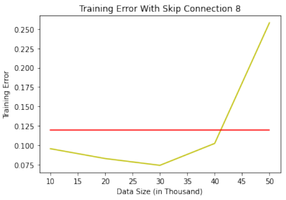

Figure 9: EMC with skip connection length 8

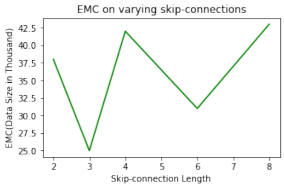

Figure 10: EMC as a function of skip-connection lengths

choose the error threshold and a formal specification never made for the value of the error threshold inb(Nakkiran et al., 2020). It can be shown from figure 5, 6, 7, 8, and 9 that the effective model complexity (EMC) changes with change in skip-length. As an example, the effective model complexity (EMC) of ResNet with skip-connection length 2 is at data size nearly 38k, whereas EMC of ResNet with skip-connection length 3 is at 25k of data-size. Figure 10 depicts the Effective Model Complexity (EMC) of ResNet-50 as a function of skip-connection length. Note again that the effective model complexity of a training procedure is the maximum number of samples on which the training error is less the threshold (Nakkiran et al., 2020). The result demonstrates that the effective model complexity actually changes as model architecture and data size change.

6 Conclusion

We have implemented five versions of ResNet50 with the skip-connection length of 2, 3, 4, 6, and 8 and trained them on different subsets of the CIFAR-10 dataset to verify the effect of length of skip-connection of ResNet50 on its model complexity. The model complexity changes with skip-connection lengths, but there does not appear to be a simple relationship between the variables. As skip connection length increases the EMC appears to increase, but not as expected. The model complexity appears to have a global minimum at short lengths, but has a secondary local minimum with longer lengths.

7 Contribution

7.1 Lokesh Das

- Wrote necessary code to preprocess the data to train ResNets
- Divided the datasets into different sizes to calculate the EMC
- Implemented the five versions of ResNet50 from scratch for skip-connection length 2,3,4,6, and 8
- Trained the ResNet50 model for each skip-connection length with different sizes of dataset
- Got all graphs and wrote result section
- Reviewed existing literature
- Wrote introduction, dataset, acknowledgement sections and rewrite the other sections

7.2 Hunter Gore

- Reviewed existing literature (related works)
- Wrote a significant part of the report
- Wrote introduction/background sections
- Wrote Problem Statement/Described Setup

7.3 Bhavika Khare

- Reviewed existing literature
- Designed the experiments for EMC calculation
- Wrote a significant part of the report
- Debugged the EMC code
- Attempted to automate the variation of skip-lengths in the code
- Wrote the abstract

Acknowledgement

Special thanks to the Professor Dr. Xiaolei Huang for his support and time during our project implementation, especially for his instructions to find out a way of building ResNet50 with variable lengths of skip-connections.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2019. What can resnet learn efficiently, going beyond kernels? *CoRR*, abs/1905.10337.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. 2021. Model complexity of deep learning: A survey. *arXiv preprint arXiv:2103.05127*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Baoqi Li and Yuyao He. 2018. An improved resnet based on the adjustable shortcut connections. *IEEE Access*, 6:18967–18974.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2020. Deep double descent: Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eshaan Nichani, Adityanarayanan Radhakrishnan, and Caroline Uhler. 2020. Do deeper convolutional networks perform better? *CoRR*, abs/2010.09610.

James O'Neill, Greg Ver Steeg, and Aram Galstyan. 2020. Compressing deep neural networks via layer fusion. *CoRR*, abs/2007.14917.

Bartosz Wójcik, Maciej Wolczyk, Klaudia Balazy, and Jacek Tabor. 2020. Finding the optimal network depth in classification tasks. *CoRR*, abs/2004.08172.