Exploring the Effects of DoReFa-Net and MISH Activation on ResNet for CIFAR-10

Jayson Hao
Department of Computer Science
University of Toronto
40 St George St, Toronto, ON M5S 2E4
jayswagon0526@gmail.com

Elizabeth Li
Department of Computer Science
University of Toronto
40 St George St, Toronto, ON M5S 2E4
eliza.li@mail.utoronto.ca

Abstract

This paper presents a reproduction and extension of a landmark paper in the field of computer vision: ResNet. We'll reproduce ResNet-20 using the original architecture proposed by Kaiming He et al. [1]. We will extend this model using DoReFa-Net to view the effect of low bit-width representations and using the MISH activation function instead of the ReLU activation function. We evaluated the performance of these extended models on the CIFAR-10 dataset as this was the initial dataset used to train and test ResNet-20. Our reproduction was able to achieve similar accuracies as the original architecture. Our extensions found that only applying MISH activation resulted in increased accuracies with minimal impact on runtime. While applying DoReFa-Net resulted in lower accuracies, the resulting accuracy was still competitive and reduced runtime, as expected. However, combining the two methods lead to suboptimal results. It can be concluded that DoReFa is a suitable approach to utilize when the objective is to reduce memory and computational demands, whereas MISH is a favourable strategy when the objective is to improve accuracy with non-linear data, such as image classification or object detection.

1 Introduction

Image classification is a fundamental task in computer vision which involves assigning a label to an image based on its content. Recently, deep learning has provided huge advancements in the field of image classification and specifically deep neural networks have significantly improved the possible accuracy achieved. In particular, ResNet had an influential impact on the domain of image classification as it brought forth the introduction of residual learning and skip connections. We will look to further enhance using two methods: DoReFa-Net and MISH activation.

1.1 Related Works

In the sector of binary neural networks, the XNOR-Net architecture proposed by Rastegari et al. [4] gained well acclaimed attention due to its ability to perform accurate, yet efficient inference using binary weights and activations. XNOR-Net has been shown to achieve state-of-the-art results on CIFAR-10 and ImageNet datasets, making it a promising design choice for resource-restricted devices. XNOR-Net serves as the introduction to the idea of using low-bit quantization for weights and activations for the purpose of improving computational efficiency, while maintaining relatively competitive accuracies.

In the original paper where MISH was introduced [3], MISH was used to replace the ReLU activation function and compare their resulting accuracies. The results showed MISH outperforming ReLU in

Preprint. Under review.

experiments which investigated the effects of increasing the number of layers in the network and the effects of noisy inputs. Both of which resulted in MISH achieving better performance relative to other activations functions such as ReLU.

1.2 About ResNet

The main idea of ResNet is the skip connections. Skip connection is useful to overcome the problem of vanishing gradients in very deep neural networks. As the number of layers in a neural network increases, the gradients can become very small, making it difficult to train the network effectively. ResNet addresses this issue by introducing skip connections, which allow the network to learn residual functions instead of the actual mapping.

In our implementation, we will be using ResNet-20.

1.3 About DoReFa

DoReFa-Net is a method proposed by Shuchang Zhou et al. [5] where we quantize the weights and activations to low bit-width values using a quantization function that maps floating-point values to discrete values. Now, the computationally expensive part of computing forward pass can be done in a bit vector

$$\mathbf{x} \cdot \mathbf{y} = \text{bitcount}(\text{and}(\mathbf{x}, \mathbf{y})), x_i, y_i \in \{0, 1\} \forall i.$$

[5] Furthermore, DoReFa allows the bits to be less than 8, which was not achieved with previous works on binary neural networks. We can achieve an runtime of $\mathcal{O}(XY)$, which are propotional to the bitwidth of **x** and **y**.

These quantized weights and activations are then used in the forward and backward passes of the neural network. The gradients are computed with a straight-through estimator as the quantization function is non-differentiable and hence it is not possible to directly compute the gradients through the quantization function.

By using low-bit quantization, the memory and computational requirements of the neural network are greatly reduced. This can lead to faster inference times on hardware platforms such as CPUs and GPUs, and possible to run on lower capacity ones. Now, DoReFa networks can be more energy-efficient than traditional networks, which can be important for battery-powered devices.

Despite the reduced precision, the DoReFa method has been shown to maintain a competitive accuracy on a variety of tasks, including image classification and object detection.

1.4 About MISH

MISH activation was first introduced in Misra's paper [3] with its properties such as smoothness and non-monotonicity that other popular activation functions such as ReLU do not have. These properties allow for easier optimization in models using gradient-based optimization methods and is able to model more complex functions. The MISH activation function is defined as,

$$f(x) = x \tanh(softplus(x)) = x \tanh(\ln(1 + e^x))$$

Since the original implementation of ResNet and the ResNet implemented with DoReFa both currently use ReLU, we will attempt to replace it with MISH and examine the effect.

2 Method

1. Follow instructions on the previous work on DoReFa [5], and reproduce their result. We will fork their implementation on Github.
2. Fine-tune each of the DoReFa-Net ResNet models (there should be 25 models total) to optimize their performance by systematically trying different hyperparameters and training procedures.
3. Evaluate the performance using testing sets of CIFAR-10[2] and compare their performance with respect to the original ResNet.
4. Apply MISH to the original ResNet architecture and compare their run time and accuracy.

5. Apply MISH with each of the DoReFa-Net methods on the original ResNet architecture (there should be 25 models total, see Appendix) and compare their run time and accuracy.

3 Results

The following results are produced with `training_batch_size = 128`, `eval_batch_size = 100`, `epochs = 200`

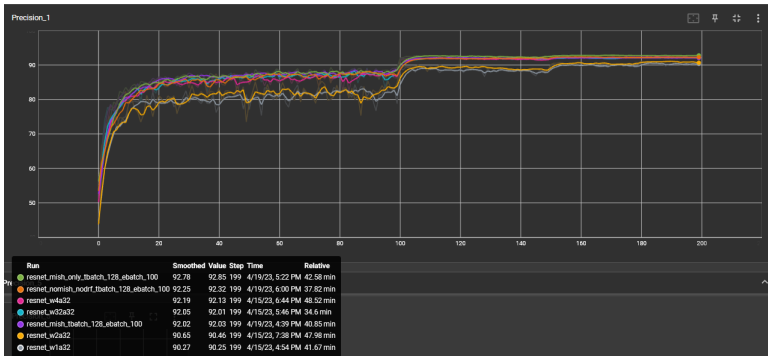

Figure 1: Accuracy per epochs (60% smoothing applied)

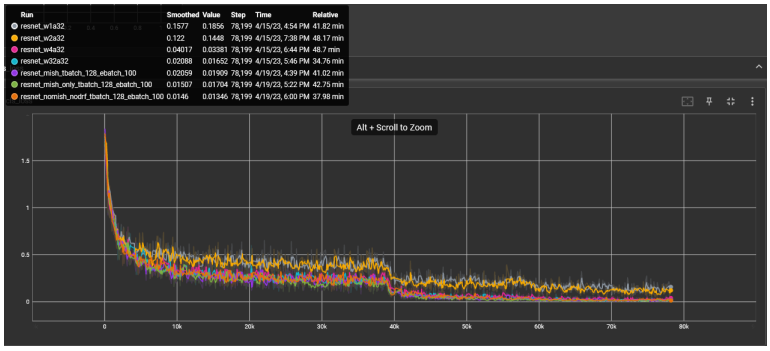

Figure 2: Loss value per steps (60% smoothing applied)

	Accuracy	Time (min)
ResNet	92.25	37.82
MISH	92.78	42.58
DoReFa with 1 w-bit 32 a-bits (w1a32)	90.27	41.67
DoReFa w2a32	90.65	47.98
DoReFa w4a32	92.19	48.52
DoReFa w32a32	92.05	34.6
DOReFa w32a32+MISH	92.02	40.85

Table 1: Final Accuracy with 60% smoothing.

4 Analysis and Discussion

In both sets of experiments, it can be seen that ResNet-20 with MISH activation resulted in the highest accuracy with minimal impact on time. ResNet-20 initially uses ReLU activation, but in replacing this with MISH, we were able to increase accuracy from 92.25% to 92.78%, a 0.53%

difference. While ReLU is widely used due to its ability to introduce non-linearity, MISH is also able to offer this benefit. Further, MISH provides benefits being a smooth, continuous, self-regularized, non-monotonic activation function[3]. These properties and the design of MISH eliminate the Dying ReLU phenomenon, avoids saturation and the undesired effects when performing gradient-based optimization[3]. As a result, MISH is able to achieve better performance than ReLU.

On the other hand, applying the DoReFa-Net method to ResNet-20 consistently resulted in lower accuracy, although was able to improve runtime. DoReFa achieves these reduced computational requirements through low-bit quantization, which in turn can negatively affect performance. Since low-bit quantization looses information from restricting the weights and activations to some fixed values, the model loses some of its ability to capture more complex relationships and thus achieve a lower accuracy. However, DoReFa-Net has been shown to still be able to achieve competitive accuracies through its straight-through estimator which helps minimize the effect of quantization on the training model. As a result of this, applying DoReFa-Net on ResNet does not significantly reduce the accuracy while improving runtime.

When combining the two approaches, the result was quite promising as it had achieved both accuracy and computational efficiency similar to the original ResNet-20 model. This result is noteworthy as DoReFa typically lowers accuracy while improving computational efficiency, as shown in the results with only DoReFa applied to ResNet-20. On the other hand, MISH was able to increase accuracy with little impact on runtime. When applying only MISH, we achieved a runtime of 42.58 minutes and 92.78% accuracy and when only applying DoReFa, we achieved a runtime of 34.60 minutes and an accuracy of 92.05%. When we instead used a combination of both MISH and DoReFa, we achieved a runtime of 40.85 minutes and an accuracy of 90.02%. The effect of DoReFa is shown in the combination as both runtime and accuracy had decreased. However, these results seem promising as they are comparable to the performance of the original ResNet-20, meaning we likely have the presence of both MISH and DoReFa where the two offset each other's effects on accuracy.

5 Conclusion

In conclusion, our experiments demonstrate that replacing the ReLU activation function with MISH leads to an increase in accuracy while maintaining a minimal impact on runtime. MISH's properties enable it to capture more complex relationships, resulting in improved performance compared to ReLU.

Conversely, the application of the DoReFa-Net method on ResNet-20 leads to a trade-off between accuracy and computational efficiency. While DoReFa's low-bit quantization helps to reduce memory and computational demands, it can negatively impact the model's performance, as was found in our experiments. Nevertheless, DoReFa-Net's straight-through estimator helps mitigate the effects of quantization during training and thus is still able to maintain a competitive accuracy while reducing runtime.

The combination of the two approaches yielded promising results similar to that of the original ResNet-20 model. Our results imply that the adoption of MISH may help make up for the accuracy loss that DoReFa normally causes while increasing computational efficiency.

Future works can further explore other applications of DoReFa-Net and MISH activation on other architectures, such as AlexNet or DenseNet, as well as other domains aside from image recognition and classification.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[2] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.

[3] Diganta Misra. Mish: A self regularized non-monotonic activation function, 2020.

[4] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *CoRR*, abs/1603.05279, 2016.

[5] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients, 2018.

A Appendix

The following results are with 64 training batch sizes, 50 eval batch sizes, and 50 epochs.

A.1 Pure ResNet and Pure MISH

	Accuracy (%)	Time (min)
ResNet-20	85.7	15.24
MISH	86.51	13.51

Table 2: Final Accuracy with 60% smoothing.

A.2 ResNet with DoReFa

Accuracy (%):

	1 A-bits	2 A-bits	4 A-bits	8 A-bits	32 A-bits
1 W-bits	20.36	69.92	73.75	75.54	77.33
2 W-bits	33.87	73.49	74.5	74.6	79.54
4 W-bits	10.53	78.49	79.53	79.94	84.71
8 W-bits	10.12	76.75	77.12	80.43	85.91
32 W-bits	10.07	78.57	80.6	80.84	84.11

Table 3: ResNet with DoReFa implementation. Final Accuracy with 60% smoothing.

Time (minutes):

	1 A-bits	2 A-bits	4 A-bits	8 A-bits	32 A-bits
1 W-bits	22.22	22.71	23.26	22.7	18.77
2 W-bits	25.89	26.42	26.61	26.25	22.19
4 W-bits	25.68	26.27	26.29	26.03	22.51
8 W-bits	25.28	26.75	26.3	26.41	22.24
32 W-bits	18.3	19.02	19.26	17.68	13.41

Table 4: ResNet with DoReFa implementation. Total runtime in minutes.

A.3 ResNet with DoReFa and MISH

Accuracy (%):

	1 A-bits	2 A-bits	4 A-bits	8 A-bits	32 A-bits
1 W-bits	27.28	74.93	75.64	76.12	81.31
2 W-bits	14.43	74.08	75.89	77.09	80.67
4 W-bits	54.06	78.7	78.73	80.67	85.87
8 W-bits	47.11	80.68	80.98	81.97	85.7
32 W-bits	49.35	80.28	80.53	81.15	86.35

Table 5: ResNet with DoReFa and MISH. Final Accuracy with 60% smoothing.

Time (minutes):

	1 A-bits	2 A-bits	4 A-bits	8 A-bits	32 A-bits
1 W-bits	21.73	23.02	23.15	23.14	18.77
2 W-bits	26.11	26.82	27.02	26.96	22.63
4 W-bits	26.81	27.04	26.95	26.85	23.35
8 W-bits	26.04	27.12	26.92	27.35	22.84
32 W-bits	19.05	19.96	19.95	19.9	15.62

Table 6: ResNet with DoReFa and MISH. Total runtime in minutes.

B List of Contributions

Coding: Jayson and Elizabeth

Write-up: Jayson and Elizabeth

Running the code: my poor Intel i-8700, whatever 32GB memory I use, and NVIDIA GTX 1070 Ti.

Debugging: ChatGPT-4 Mar 23 Version.