1 Bit ViT: Low-Bit Quantization Methods for Vision Transformers

Gautom Das, Nicholas Henderson,
Sean Im, Vincent La,
Ethan Lau

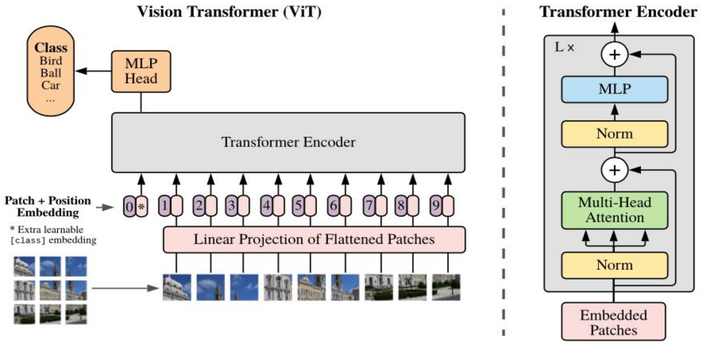
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **Efficiency:** Can we reduce the size of the model while maintaining performance?
- **Accessibility:** Vision Transformers are powerful and should be available on a wider array of devices.

Method

We took the well known ViT architecture from Dosovitskiy et. Al (2021) and attempted to quantize the linear layers of this architecture. Quantization is a method that maps a weights to more efficient smaller types: models are often trained as 16 and 32 bit floating numbers, then mathematically scaled down and normalized to a smaller number of bits.



ViT architecture from “An Image is Worth 16x16 Words” paper.

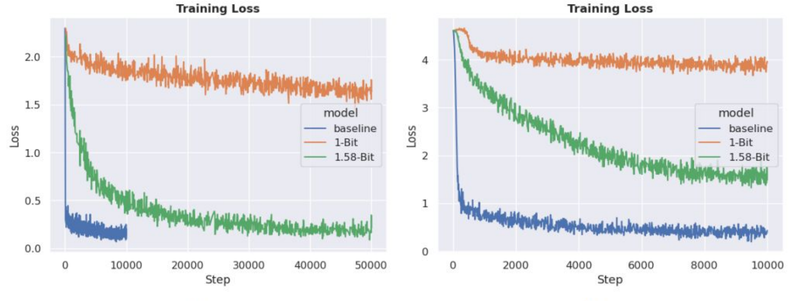
1-bit and 1.58-bit Quantization

1-bit and 1.58-bit quantization are new extremely low-bit quantization methods developed by Ma and Wang, et. al (2023). 1-bit quantization restricts our weights to be either -1 or +1, whereas 1.58-bit quantization allows them to be -1, 0, or +1.

Training and Testing

We trained and tested on both CIFAR-10 anad CIFAR-100, to compare how changing the number of classes affects the ViTs ability to generalize. We also used attention maps to help visualize the differences.

Results



Training loss plots for (a) CIFAR-10 and (b) CIFAR-100 fine-tuning runs

1-Bit Model Performance: The 1-bit model underperformed significantly compared to the baseline, with a marked drop in accuracy. This model struggled with generalization and demonstrated considerable confusion between classes.

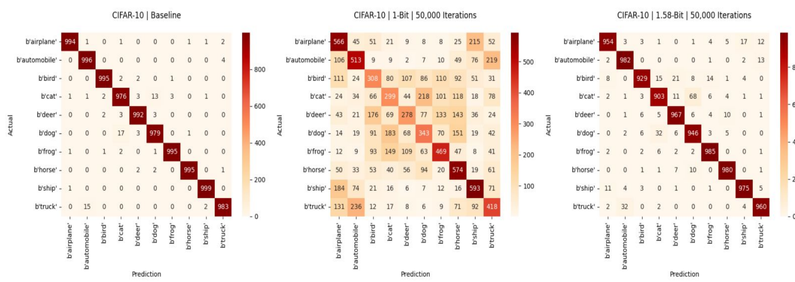
1.58-Bit Model Performance: The 1.58-bit model showed improved performance over the 1-bit model, with accuracy approaching that of the baseline While it still exhibited some confusion between classes, it retained a higher degree of accuracy and lower training losses than the 1-bit model.

Table 1: CIFAR-10 Test Results for Different Models

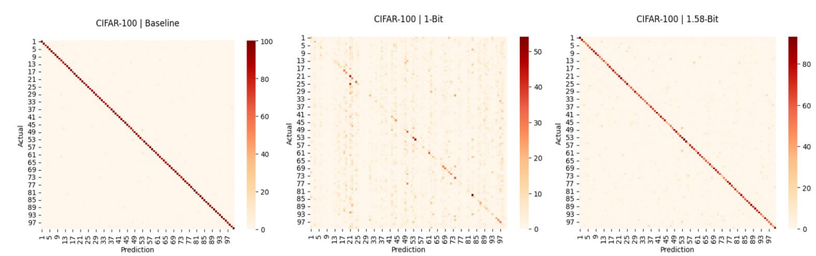
Metric	Baseline	1-Bit (50,000 Epochs)	1.58-Bit (50,000 Epochs)
Accuracy	0.9904	0.4361	0.9581
Precision	0.9904	0.4309	0.9583
Recall	0.9904	0.4361	0.9581
F1 Score	0.9904	0.4335	0.9582

Table 2: CIFAR-100 Test Results for Different Models

Metric	Baseline	1-Bit (10,000 Epochs)	1.58-Bit (10,000 Epochs)
Accuracy	0.9279	0.1096	0.6611
Precision	0.9286	0.0829	0.6640
Recall	0.9279	0.1096	0.6611
F1 Score	0.9282	0.0944	0.6625

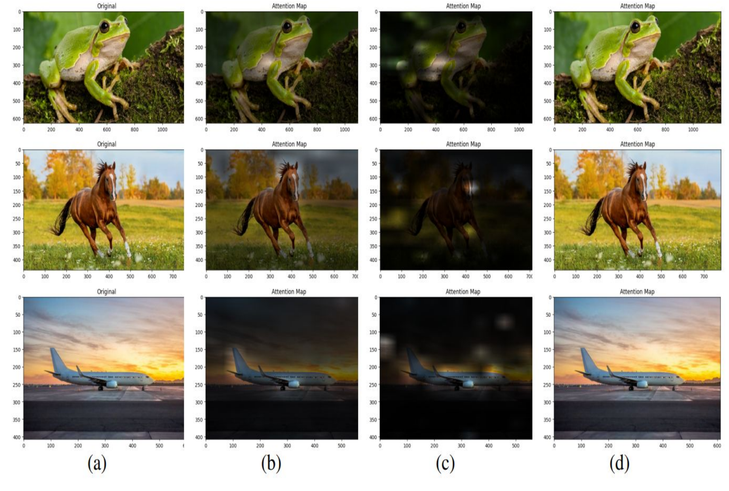


Confusion Matrices for CIFAR-10 Classification Task



Confusion Matrices for CIFAR-100 Classification Task

Attention Visualizations



Attention maps of example images. (a) Original, (b) Baseline attention, (c) 1.58-bit attention, (d) 1-bit attention

Discussion and Future Work

This study demonstrates the unique dynamics of different low-bit quantization and its implications on the capabilities of vision transformers. It highlights how, despite some accuracy issues the 1-bit model and the 1.58-bit model still demonstrate an ability to learn suggesting a novel relationship between quantization and comprehension of classes that diminishes. This was a domain previously unexplored with such low quantization levels and support for low-bit hardware support is still very much a work in progress. Future studies could explore the dynamics between quantization depth, model complexity, and performance, potentially revolutionizing the deployment of advanced models on resource-constrained devices.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., … Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2402.17764

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., … Wei, F. (2023). BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2310.11453

1 Bit ViT: Low-Bit Quantization Methods for Vision Transformers

Gautom Dos, Nicholas Henderson,
Sean Im, Vincent La,
Ethan Lau

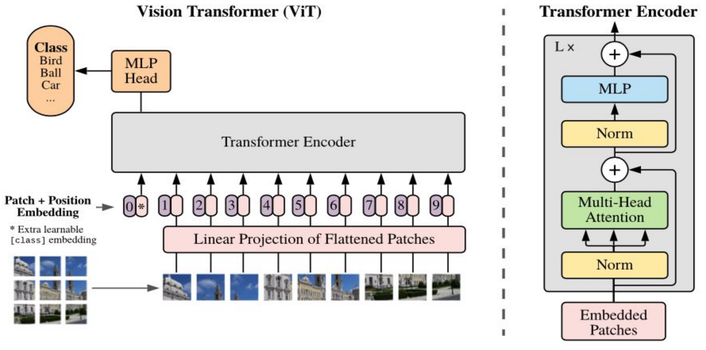
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **Efficiency:** Can we reduce the size of the model while maintaining performance?
- **Accessibility:** Vision Transformers are powerful and should be available on a wider array of devices

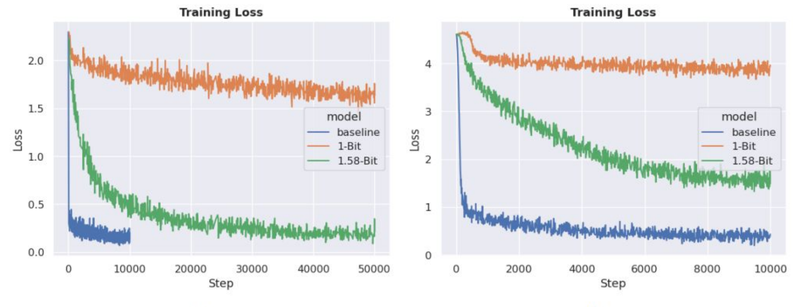
In this project, we apply low-bit quantization methods, made for LLMs, to vision transformers.

ViT Architecture



ViT architecture from “An Image is Worth 16x16 Words” paper.

Loss Plots for CIFAR-10 and CIFAR-100



Here, we see the training loss ViT on CIFAR-10 and CIFAR-100, highlighting the efficacy of quantization techniques. The baseline model pre-trained on ImageNet shows decreasing loss, reaching about 0.2 on CIFAR-10 and 0.5 on CIFAR-100 after 10,000 epochs.

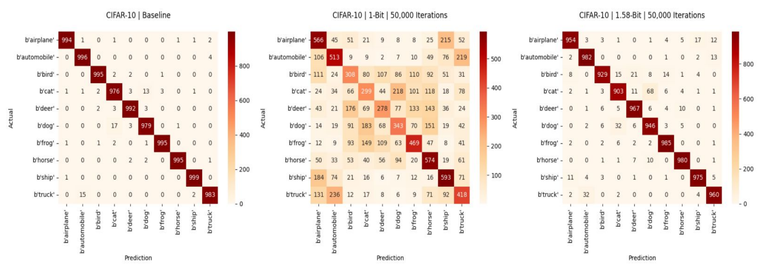
In stark contrast, the 1-bit model struggles significantly, with markedly higher losses and poor test accuracies of 43% for CIFAR-10 and 10% for CIFAR-100, indicating a substantial loss of generalization capability due to severe quantization. Conversely, the 1.58-bit model, despite a slower start, gradually matches the baseline performance, illustrating its potential to balance efficiency with minimal performance compromise. This supports the use of mid-level quantization as a viable strategy for enhancing neural network efficiency without drastically impacting model accuracy. Below are the test Results for each of the different models.

Table 1: CIFAR-10 Test Results for Different Models

Metric	Baseline	1-Bit (50,000 Epochs)	1.58-Bit (50,000 Epochs)
Accuracy	0.9904	0.4361	0.9581
Precision	0.9904	0.4309	0.9583
Recall	0.9904	0.4361	0.9581
F1 Score	0.9904	0.4335	0.9582

Table 2: CIFAR-100 Test Results for Different Models

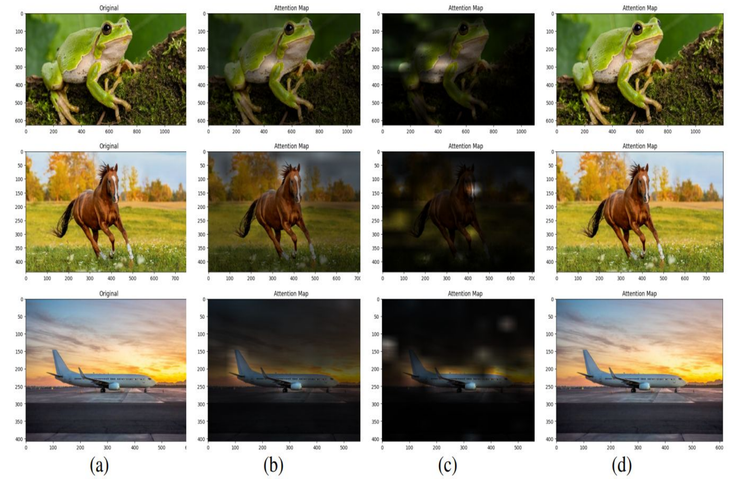
Metric	Baseline	1-Bit (10,000 Epochs)	1.58-Bit (10,000 Epochs)
Accuracy	0.9279	0.1096	0.6611
Precision	0.9286	0.0829	0.6640
Recall	0.9279	0.1096	0.6611
F1 Score	0.9282	0.0944	0.6625



Confusion Matrices for CIFAR-10 Classification Task

These confusion matrices further illustrate the struggles of quantized models in image classification tasks. The 1-bit ViT, fine-tuned on CIFAR-10, still has a lot of confusion with various classes such as confusing similar animals like dogs, cats, and dears or vehicles like ships, trucks, and airplane. The 1.58-bit performed much closer to the baseline, full-precision ViT but still made mistakes between classifying similar animals and vehicles.

Attention Visualizations



Attention maps of example images. (a) Original, (b) Baseline attention, (c) 1.58-bit attention, (d) 1-bit attention

Results + Discussion

Our results demonstrate the feasibility of applying extreme low-bit quantization techniques, specifically 1-bit and 1.58-bit, to vision transformers, a domain previously unexplored with such low quantization levels. Across the board, the 1.58-bit quantized ViT was able to outperform the 1-bit quantized ViT for fine-tuned classification tasks with only a ~4% reduction in accuracy compared to the full-precision model on CIFAR-10. It’s interesting to note the differences in the visual attention map between the full-precision ViT and the 1.58-bit quantized ViT

Future studies could explore the dynamics between quantization depth, model complexity, and performance, potentially revolutionizing the deployment of advanced models on resource-constrained devices. Additionally, training low-bit precision ViTs from scratch could further evaluate the capabilities of these quantized models.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., … Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2402.17764

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., … Wei, F. (2023). BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2310.11453

1 Bit ViT: Low-Bit Quantization Methods for Vision Transformers

Gautom Dos, Nicholas Henderson,
Sean Im, Vincent La,
Ethan Lau

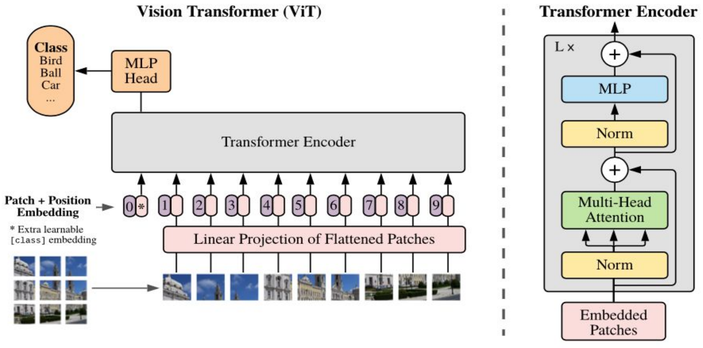
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **Efficiency:** Can we reduce the size of the model while maintaining performance?
- **Accessibility:** Vision Transformers are powerful and should be available on a wider array of devices

In this project, we apply low-bit quantization methods, made for LLMs, to vision transformers.

ViT Architecture



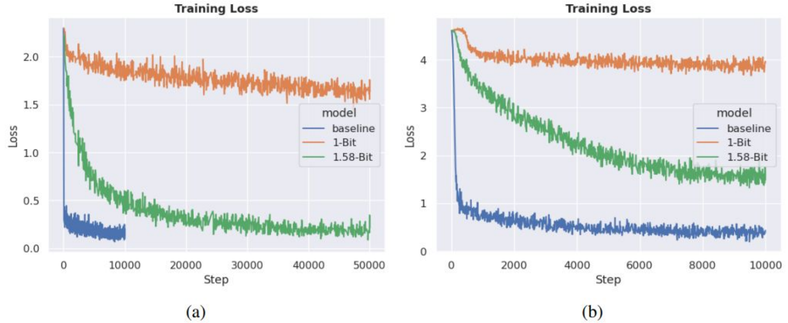
ViT architecture from “An Image is Worth 16x16 Words” paper.

1-bit and 1.58-bit Quantization

Quantization refers to mapping a set of continuous infinite values to a smaller set of discrete finite values. In the context of machine learning, we look to quantize the set of values that our weights can take on, allowing us to represent them in a lower-precision format for increased efficiency and decreased memory usage

1-bit quantization restricts our weights to be either -1 or +1, whereas 1.58-bit quantization allows them to be -1, 0, or +1.

Training and Test Metrics



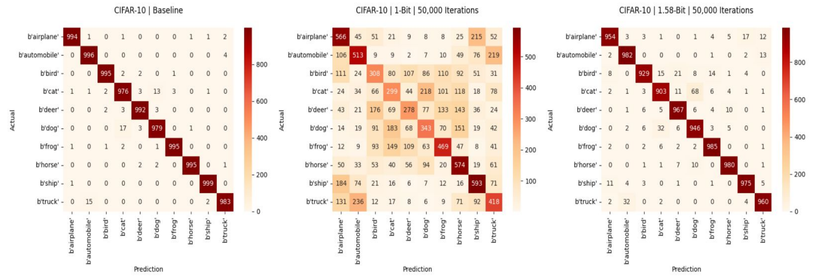
Training loss plots for (a) CIFAR-10 and (b) CIFAR-100 fine-tuning runs

Table 1: CIFAR-10 Test Results for Different Models

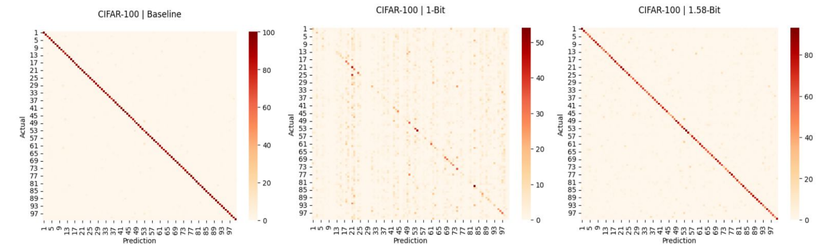
Metric	Baseline	1-Bit (50,000 Epochs)	1.58-Bit (50,000 Epochs)
Accuracy	0.9904	0.4361	0.9581
Precision	0.9904	0.4309	0.9583
Recall	0.9904	0.4361	0.9581
F1 Score	0.9904	0.4335	0.9582

Table 2: CIFAR-100 Test Results for Different Models

Metric	Baseline	1-Bit (10,000 Epochs)	1.58-Bit (10,000 Epochs)
Accuracy	0.9279	0.1096	0.6611
Precision	0.9286	0.0829	0.6640
Recall	0.9279	0.1096	0.6611
F1 Score	0.9282	0.0944	0.6625

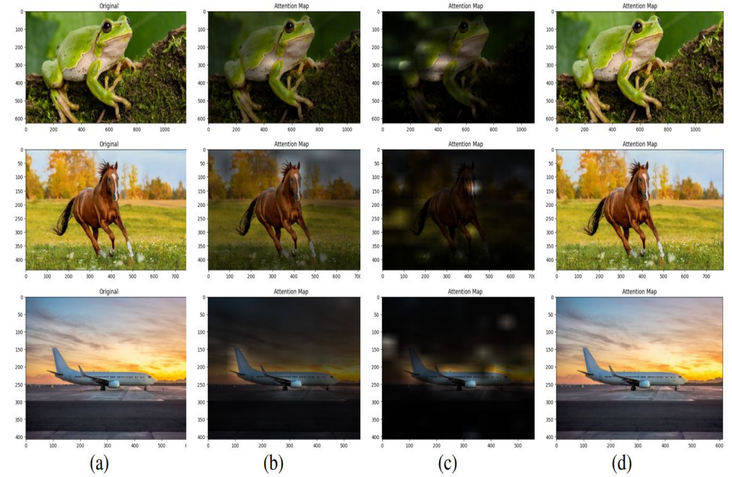


Confusion Matrices for CIFAR-10 Classification Task



Confusion Matrices for CIFAR-100 Classification Task

Attention Visualizations



Attention maps of example images. (a) Original, (b) Baseline attention, (c) 1.58-bit attention, (d) 1-bit attention

Discussion + Future Work

Our results demonstrate the feasibility of applying extreme low-bit quantization techniques, specifically 1-bit and 1.58-bit, to vision transformers, a domain previously unexplored with such low quantization levels. Future studies could explore the dynamics between quantization depth, model complexity, and performance, potentially revolutionizing the deployment of advanced models on resource-constrained devices.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., … Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2402.17764

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., … Wei, F. (2023). BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2310.11453