Effect of learnable resizers on Vision Transformer (ViT) models

Varun Machhale Kumar [1]

Abstract

Transformer models have seen tremendous success in natural language processing tasks. Recently, there is a shift to apply these models for computer vision tasks which is currently dominated by convolutional architectures. However, Vision Transformer (ViT) models work best when pretrained on a large dataset and applied to smaller computer vision datasets. These pretrained models typically require our images to be resized to a fixed resolution. Traditionally, this is done using off-the-shelf image resizers like bilinear or bicubic interpolation frameworks. These resizers are observed to limit on-task performance. In this paper, we implement a vision transformer model with learnable resizer modules to perform image classification and experiment with two datasets to see whether the resizer modules improve performance. The implemented algorithm is tested on the beans dataset and CIFAR10 dataset (1k images).

1. Paper 1

1.1. Review of Paper 1

Dosovitskiy et al. ((Dosovitskiy et al., 2021)) propose to apply the Transformer architecture to solve problems in computer vision. For a long time, convolutional neural networks have been the state-of-the-art models for performing computer vision tasks like image classification. This paper demonstrates that instead of a CNN architecture a transformer model applied directly to a sequence of image divided into patches yields better performance results as they lack inductive biases and are permutation invariant.

For training, the input images are split into patches of fixed size, linearly embedded them, add positional encodings along with a classifcation token (CLS) for performing classification. The resulting sequence is fed into a transformer encoder model which consists of layers of multi-headed self-attention blocks and Multilayer perceptron (MLP) blocks. This model is trained on three large datasets: Imagenet (1.3M images), Imagenet (14M images) and a private dataset JFT (303M high-resolution images). The resulting pre-trained model is then evaluated on other benchmark datasets such as CIFAR10, CIFAR100, Oxford-IIIT Pets, Oxford Flowers-102.

The authors conclude that the Vision Transformer trained on JFT dataset outperforms all the state-of-the-art methods. Additonally, Vision Transformer models takes significantly less compute to pre-train on the smaller benchmark datasets compared to the state-of-the-art counterparts.

1.2. Critique for paper 1

1.2.1. Strengths

The authors show that by applying pure transformers from the NLP domain to image recognition tasks, they were able to match and even beat the state-of-the-art performance obtained from convolutional networks on several image datasets. Additionally, this requires less computational time and resources to pre-train on smaller datasets. Transformer models are found to be better than convolutional networks due to the lack of inductive biases and permutation invariancy. As these models are already trained on very large scale image recognition tasks, they can be easily fine-tuned on smaller image recognition tasks without the need to have large datasets.

1.2.2. Weaknesses

The paper illustrates that Vision Transformer models yields excellent results only when pre-trained on a very large dataset. The original paper train their model of three datasets Imagenet (1.3M images), its superset Imagenet (14M images) and a private dataset JFT (303M high-resolution images). They then transfer these model to other benchmark datasets such as CIFAR10/100, Oxford-IIIT Pets, Oxford Flowers-102 etc.

This creates a reliance on using very large datasets for obtaining comparable performance in other vision tasks. It also creates an unfair dependence on using transfer learn-

[1]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, USA. Correspondence to: Varun Machhale Kumar <kumar603@purdue.edu>.

Final term paper submission (ICML format) as part of the ECE570 Artificial Intelligence (Fall 2022) course taken under Prof. David Inouye at Purdue University

ing networks that are pre-trained on image classification datasets. The pretrained models trained on image classification problems would not necessarily adapt well to other tasks. This could severely limit the usage of ViT models in tasks where large-scale datasets are not available such as in medical domains.

1.2.3. QUESTIONS

Questions to author:

1) Can Vision transformer models be applied to other computer vision tasks like segmentation and object detection?
2) Is it possible to release the private JFT dataset publicly?

2. Paper 2

2.1. Review of paper 2

In the paper (Talebi & Milanfar, 2021), the authors develop a learnable resizer network to learn the optimal parameters required for resizing images.

Today, images used for machine learning tasks are downsampled using the standard bicubic or bilinear resizers. The input images are typically resized (usually generally downsampled) to a standard size (for eg. 224 x 224) to obtain same resolution for all the images in a mini-batch, adjust for memory limitations and training models faster. However, the authors show that these classical resizers hurt the performance of the trained networks and replacing them with a learned resizer module will help us obtain better classification performance.

The resizer network module is as shown in Fig. 1 (Source: (Talebi & Milanfar, 2021)). The authors achieve this by building a CNN model that includes convolutional layers, residual blocks, bilinear resizer modules and skip connections. The bilinear resizing modules allows the network to scale the input image to any required resolution.

The paper raises concerns regarding the "taken for granted" elements used in traditional architecture like resizing and helps us understand why its better for a model to learn preprocessing parameters directly from the data.

2.2. Critique for paper 2

2.2.1. STRENGTHS

The paper is able to convince through experimental results why a learned resizer model is better than the traditional resizers used in current CNN models. Since customised pre-processing methods such as resizing has not been explored extensively yet, the paper calls for the need for more research in this area.

2.2.2. WEAKNESSES

The paper illustrates that the perceptual quality in the intermediate images is irrelevant since a better classification performance was achieved. However, the paper provides no explanation for why the low visual quality actually helps their model to classify images better. There could be a possibility that the low perceptual quality actually leads to poorer performance due to lack of "good" features.

2.2.3. QUESTIONS

Question to author: Is there any specific reason why the resizing module used in the proposed model doesn't affect the overall performance?

3. Paper 3

3.1. Review of Paper 3

The authors in the paper (Zhao et al., 2020) compare the performance of self-attention based networks with their convolutional counterparts for image recognition tasks. For this they consider two variations of self-attention networks: pairwise self-attention and patchwise self-attention. They found that patchwise self-attention networks substantially outperform convolutional performances.

For comparison, they train their network on ImageNet dataset that contains 1.28 million images and 50,000 images in the validation dataset from 1000 classes.

Further the paper, conducts two experiments to assess the robustness of self-attention networks compared to convolutional networks. They do this by assessing the robustness to adversarial attacks and zero-shot testing on rotated and flipped images. The paper also finds that vector based self-attention substantially outperforms the dot-product attention (scalar attention).

3.2. Critique for paper 3

3.2.1. STRENGTHS

The authors show that replacing convolutions with self-attention networks showed significant performance gains. Patchwise self-attention networks work to generalize convolutions and can potentially provide better performance across all computer vision tasks.

3.2.2. WEAKNESSES

The original paper demonstrates the performance of self-attention based networks only in the context of image recognition tasks. More research needs to be conducted to explore the robustness of self-attention modules in other computer vision tasks.

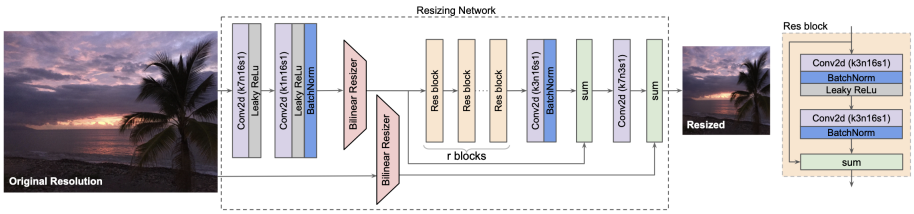

Figure 1. Resizer network module (Source: (Talebi & Milanfar, 2021)

3.2.3. QUESTIONS

Question to author: Is the model less affected by image specific inductive biases?

4. Implementation

4.1. Model description

For this paper, we implement a Vision Transformer model available in hugging face. In the preliminary implementation, I trained a model from scratch on the CIFAR10 dataset but did not see much improvement in accuracy. For the final implementation, I used a pretrained model, that was pretrained on ImageNet 21k with a patch size of 16 and input image size of 224.

Since, the pretrained models expect the input image to be of fixed size (for example 224x224), we have to resize our images to the model requirements. For this I implemented the learnable resizer module given in (Talebi & Milanfar, 2021). I used the code available at thisthis Github repository and created a custom model with the resizer module on top that outputs a resized image to the Vision Transformer. Using the Vision Transformer model as baseline, I compare the performance effects after adding the learnable resizing network.

The pretrained Vision transformer model consisted of 85.8M parameters. The resizing network is a small module added on top of the ViT network which around 1k parameters and is learnt along with fine-tuning the entire network.

All the models were run on Google Colab Tesla T4 GPU and took between 30 - 45 min.

4.2. Dataset description

I ran the model on two datasets: CIFAR10 and Beans dataset. Both of these datasets were available in the hugging face built-in datasets.

1) CIFAR10: The original CIFAR10 dataset consists of 50000 32x32 images in the training set and 10000 images in the test set. Due to computational costs, I used a smaller subset of the CIFAR10 dataset consisting of 1000 images in training set and 1000 images in the test dataset. The training set was further split in to 900 training and 100 validation datasets (split = 0.1). The CIFAR10 dataset consists of 10 classes. In Fig.2, I have shown some examples of the dataset along with predicted output.

2) Beans dataset: The beans dataset consists of 1034 images each of 500x500 resolution in the training set and 128 images in the testing set. The training set was further split into 930 training images and 104 validation images (split = 0.1). The beans dataset maps each image into 3 classes (angular leaf spot, bean rust, healthy). In Fig. 3, I have shown some examples of the dataset along with predicted output.

4.3. Hyperparameters

We run each model, ViT and Resizer + ViT on each dataset for 50 epochs each. The optimizer used was AdamW and loss function used was multi-class binary-cross entropy loss. The initial learning rate was set to 2e-5 and weight decay is set to 0.05. The training batch size is 64 and test batch size is 32.

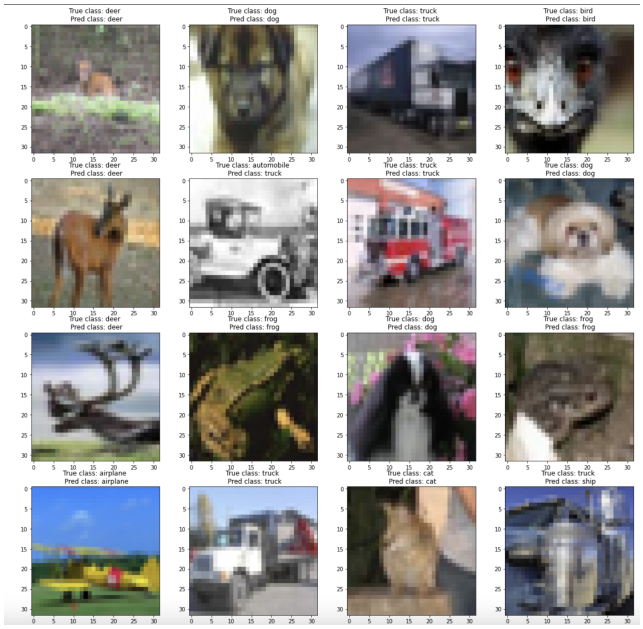

Figure 2. Sample Input and Output: CIFAR10

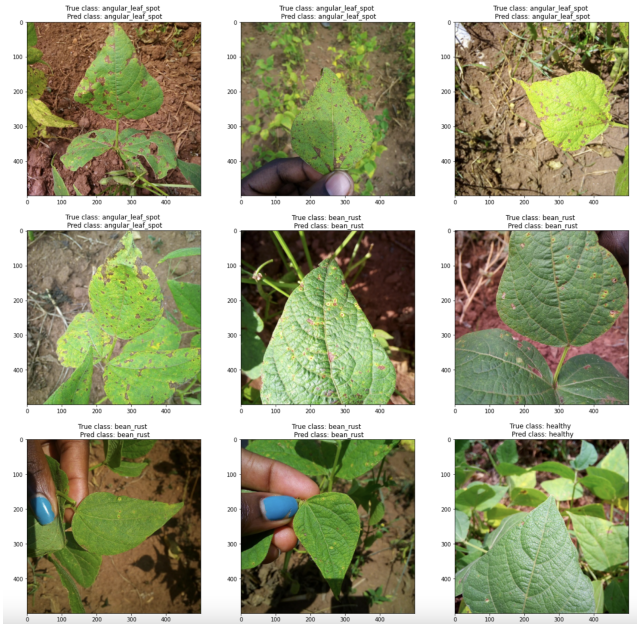

Figure 3. Sample Input and Output: Beans

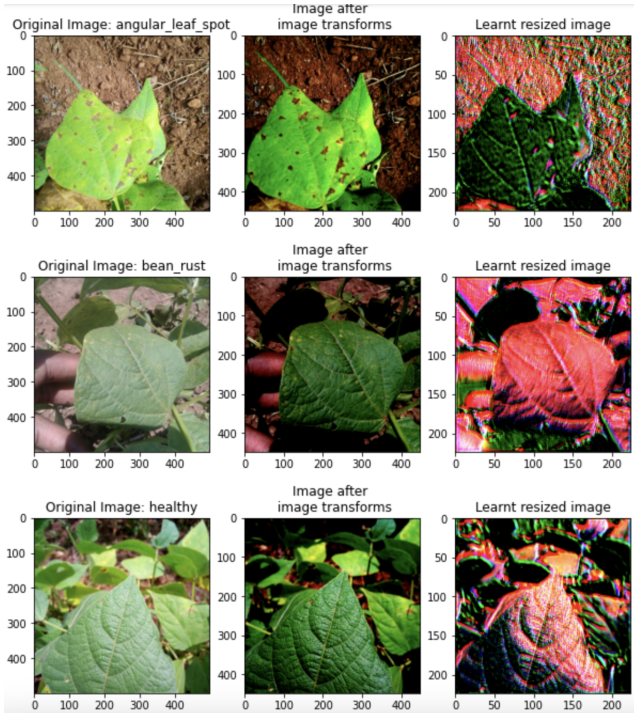

Figure 4. Resizer module images: Beans dataset

5. Evaluation

Since, this is an image classification problem with the CIFAR10 dataset having ten classes and the Beans dataset having 3 classes, the accuracy metric is used for evaluating the performance of our model.

The accuracies of each model is reported in Table 1. It can be observed from the results that ViT models were able to achieve better performance compared to their resizer module added counterparts.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

6. Results

6.1. Resizing network output

In Fig. 4 and 5, we display the original image, transformation and output of resizer network. As explained in (Talebi & Milanfar, 2021), the visual representation of the images are not appealing but instead supposed to be more useful to the network. In Fig. 5 the airplane image, it looks like parts of the image are removed and visually looks poor. However, the model classifies the image correctly as seen in Fig. 2. In the second image of Beans dataset, after resizing Fig. 4, the background and image look visually similar which might lead to poor model performance.

Model	Training Loss	Test Loss	Test accuracy	Train Runtime (GPU seconds)
ViT (CIFAR10)	0.4404	0.4378	**95.0**	1777.25
Resizer + ViT (CIFAR10)	0.6623	0.7106	86.0	2180.01
ViT (Beans)	0.0451	0.10977	**97.115**	2207.92
Resizer + ViT (Beans)	0.0206	0.3454	92.03	2721.09

Table 1. Comparison of model results

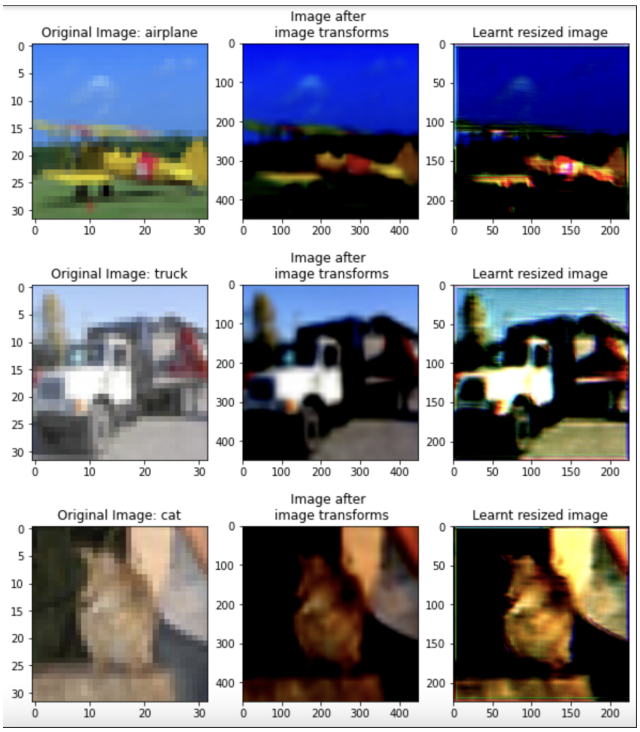

Figure 5. Resizer module images: CIFAR10 dataset

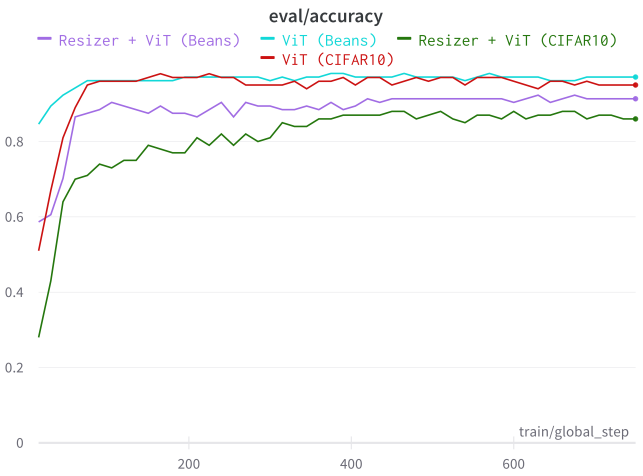

Figure 6. Evaluation accuracy

6.2. Model performance results

In Fig. 6, we plot the accuracies of our model as training proceeds (number of training steps). It can be seen that the models trained using the ViT model only consistently beat the ones trained on a combination of resizing module and ViT network.

In Fig. 7, we plot test loss curves for each trained model. Compared to the CIFAR10 dataset, the losses of Beans dataset quickly decrease and consistently have lower loss. As a result, the Beans dataset must be ”easier” to learn compared to the CIFAR10 dataset. One potential reason for this is that, the CIFAR10 dataset has 32x32 images which were upsampled to 224x224 (or 448x448) and that added a lot noisy pixels to each image making it a ”harder” dataset to learn.

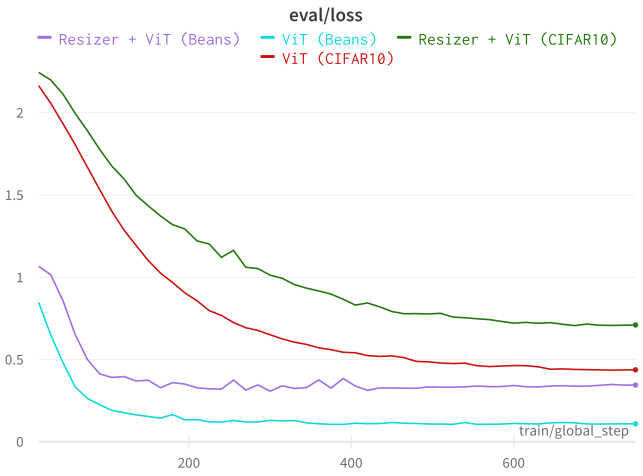

Figure 7. Test Loss curves

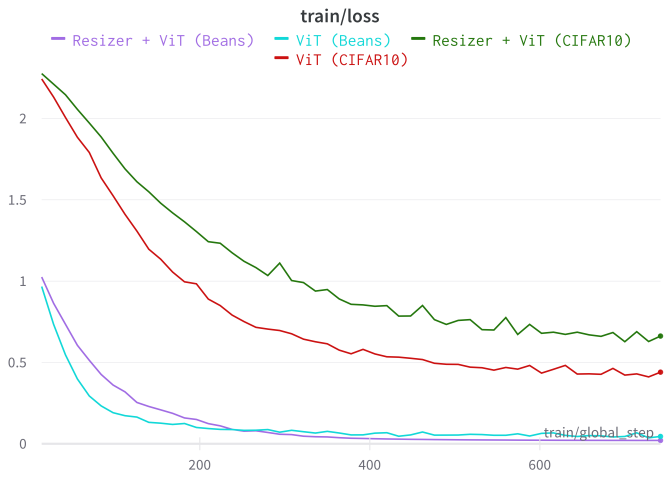

Figure 8. Train Loss curves

7. Discussion

In this paper, we analyze the effect of using a learnable resizer module on Vision transformer (ViT) models. The analysis is done by training the model with pretrained weights for ViT on the CIFAR10 (1k images) and Beans dataset.

From the results, it was observed that the resizer module negatively impacted the performance of the ViT model. One of the reasons, might be that the resized image domain doesn't accurately represent the original domain and creates a poor representation. In Fig. 4, it can be seen visually that the dataset might become poor due to the inaccurate transformation of the background of the image.

Vision Transformers can replace convolutional networks if sufficient datasets are available for getting pretrained models. However, the pretrained models should be task specific to allow for contextual information to be transferred. For this, more research is needed to explore its applications in other computer vision domains such as image segmentation and object detection.

Resizing modules doesn't seem to improve model performance and rather negatively affected vision transformer models. The original paper implementation performed analysis on a larger dataset which might have enabled the network to learn better resizing parameters. However, the paper brings into question the need to build a self-sufficient model that automatically learns data-specific parameters.

8. Future Work

For future studies, the following could be done to get a more robust comparison:

1) Training on the entire CIFAR10 dataset: For my implementation, I had used only 1k images. It might be possible that the 1k images I selected not represent the entire dataset.

2) Implement a deeper resizing module: In the current implementation only one residual block was used. This made the network lightweight and not add any additional computational costs.
As a result, the model might have underfit due to the lack of parameters to sufficiently represent the images. Future work could include a deeper resizer network with additional residual blocks aiming to capture better features required to resize the images.

3) Use a larger dataset that consists of higher resolution images. The CIFAR10 dataset contained only 32x32 images and upsampling them to 224x224 imaegs to learn resizing parameters might not be the best way. The Beans dataset contained 500x500 images, however, the dataset contained only 1k images for training. Instead using a larger dataset that contains higher resolution images might be beneficial.

4) Apply Vision Transformers (ViT) to image segmentation tasks and compare the performance with the current state-of-the-art methods. This will help understand the generalizability of transformers to tasks other than image recognition.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Talebi, H. and Milanfar, P. Learning to resize images for computer vision tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 497–506, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Talebi_Learning_To_Resize_Images_for_Computer_Vision_Tasks_ICCV_2021_paper.pdf.

Zhao, H., Jia, J., and Koltun, V. Exploring self-attention for image recognition. *CoRR*, abs/2004.13621, 2020. URL https://arxiv.org/abs/2004.13621.

9. Appendix

As a final note, I have added some statistics for understanding the GPU, memory and runtime statistics. All these graphs were generated using the Weights and Biases platform (wandb.ai) that developers to visualize hyperparameters, system metrics and compare models metrics.

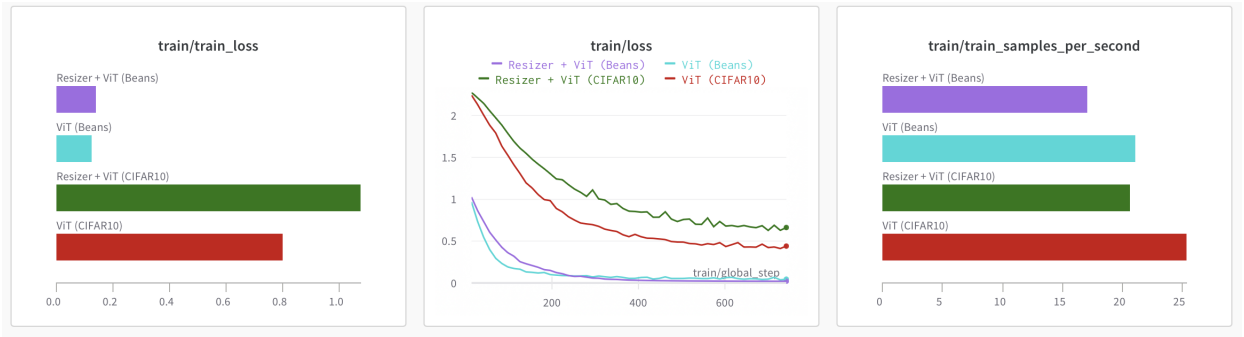

Figure 9. Train Graphs

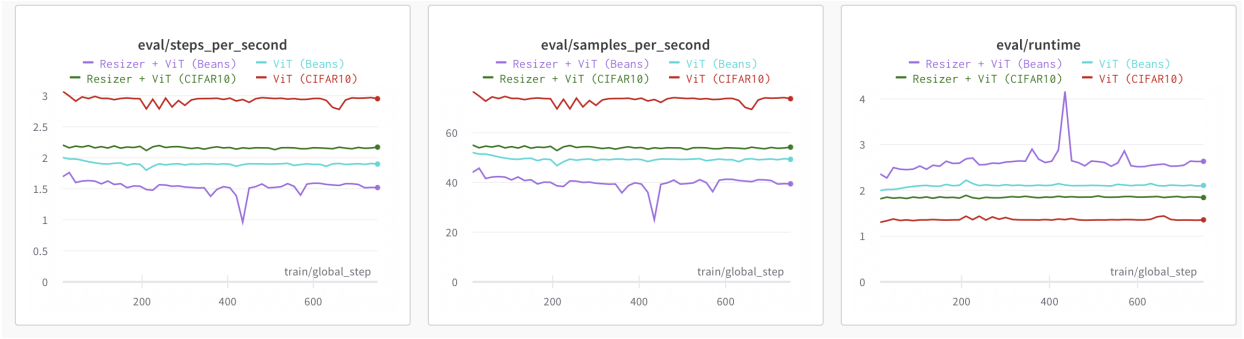

Figure 10. Runtime Statistics

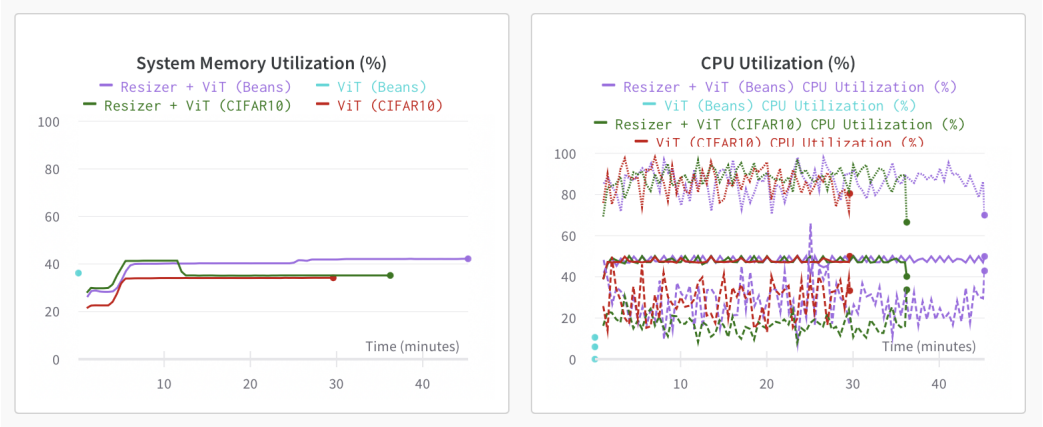

Figure 11. Memory Utilization Statistics

Figure 12. GPU Statistics

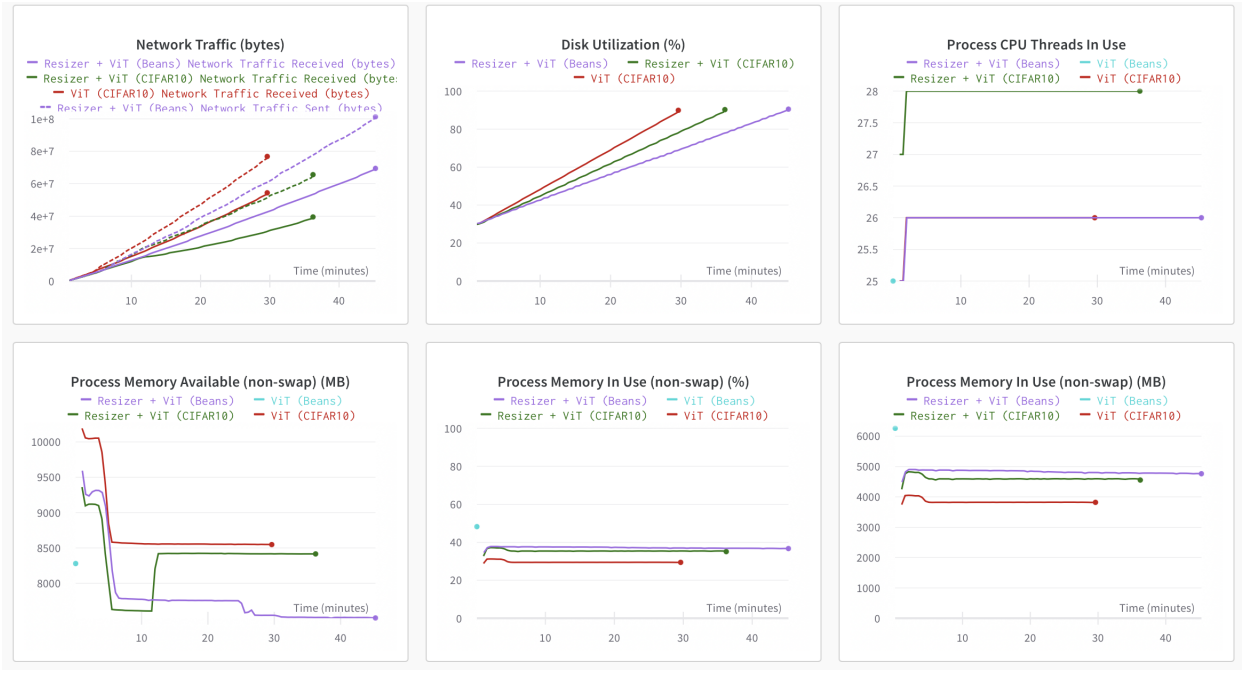

Figure 13. Memory Statistics