ToddlerNet : View Diversity vs Data Diversity

Kemmannu Vineet Venkatesh Rao
kemmann@umich.edu

Kshama Nitin Shah
kshama@umich.edu

Abstract

This paper explores the question of what is more important for representation learning - data diversity or view diversity. Toddlers constantly create learning experiences by actively manipulating objects and self-selecting object views for visual learning. With a limited variety of training data, toddlers are able to generalize and learn what objects are and their different categories. The data obtained from infants' everyday learning exhibit frequency distributions where a small number of categories are highly prevalent while most other categories are encountered infrequently. The learning data from their surrounding environments follows a long-tailed distribution. Nonetheless, the commonly occurring categories are observed from various viewpoints which encode multiple types of invariances to the objects which could promote generalization across all categories. Through this project, we were able to show that if the learning algorithm is trained in data statistics similar to what infants see in their everyday environment in their early developmental period, then the algorithm would be better able to recognize objects in different contexts and viewpoints. The main question we have addressed through our experiments is view diversity seems to be more important for learning. We also explore the most optimal learning objective for this view-diverse data and find that a supervised contrastive learning objective pre-trained on this view-diverse data achieves the best results. We plan to release our code and checkpoints here.

1. Introduction

Toddlers have a wide-ranging knowledge about the world around them.[3] They can discriminate between common categories, such as simple shapes and animal classes, even before learning to speak. The main question we are addressing is whether this early knowledge comes from a better learning objective or it emerges from having better inductive biases to learn from (modeled as having a better dataset).

In this paper, we want to address this question by leveraging two methodologies - a supervised learning objective using cross-entropy loss and a supervised contrastive learning objective using a contrastive loss. We pre-train these models and evaluate them using two datasets - one egocentric toy dataset namely ToyBox [6] dataset and a standard classification dataset namely CIFAR10. We trained both the supervised cross-entropy and supervised contrastive models on videos of toys that are commonly used by toddlers and CIFAR10, with the goal of extracting useful high-level visual representations. The acquired visual representations were then evaluated based on their ability to distinguish common visual categories in the child's environment, using only linear probes. Our results demonstrate, that supervised contrastive learning methods learn visual representations that generalize better as they minimize the intra-class variance along with maximizing the inter-class variance whereas supervised cross-entropy learning methods maximize the inter-class variance but ignore the intra-class variance.

2. Related Work

In this section, we discuss the relevant methods for Supervised Learning, Supervised Contrastive Learning, and Curriculum naturally developed by toddlers.

2.1. Supervised Learning

Supervised learning is a machine learning technique where a model is trained on labeled data to make predictions on new data. The model learns to map input data to output labels through a process of minimizing a predefined loss function.

2.2. Cross-Entropy Loss

Cross-entropy is a widely used loss function that leverages the concept of information theory entropy. It quantifies the disparity between two probability distributions over a specific set of events or random variables.

This loss function is versatile and can be utilized in both binary and multi-class classification tasks to evaluate the difference between the predicted and actual probability distributions.

One of the main shortcomings of this loss function is that it is not robust to noisy labels, has poor margins, and thus might lead to poor generalization performance.

2.3. Self-Supervised Contrastive Learning

The contrastive loss [5] uses cross-entropy loss to measure how well the model can classify the “future” representation amongst a set of unrelated “negative” samples. In self-supervised contrastive learning, a single positive is used for each anchor (augmented version of the same image) and contrasted against the rest of the examples as negatives in the batch as shown in 1.

2.4. Supervised Contrastive Learning

In supervised contrastive learning [1], clusters of points belonging to the same class are pulled together in the embedding space, while simultaneously pushing apart clusters of samples from different classes. The main idea of supervised contrastive learning is to consider many positives per anchor in addition to many negatives (as opposed to self-supervised contrastive learning which uses only a single positive example). These positives are drawn from samples of the same class as the anchor, rather than being data augmentations of the anchor, as done in self-supervised learning. Our methodology compares the supervised cross-entropy loss with the supervised contrastive loss.

2.5. Curriculum created by infants for statistical learning

Smith *et al.* [4] suggests that everyday world and objects of the infant serves as a training set for statistical learning that changes as the infant’s sensorimotor abilities develop. These changing environments could potentially form a curriculum that optimizes learning in different domains. For example, the infant sees one horse toy from multiple views and using multiple sensations. They do not see multiple colors or breeds of horses like commonly constructed image classification datasets used in deep learning today.

2.6. ToyBox Dataset

The Toybox dataset [6] comprises videos of 12 categories of toys, each with 30 different objects undergoing 10 transformations for approximately 20 seconds each. All 12 of these categories are among the most common early-learned nouns for typically developing children in the U.S. The dataset contains around 4,000 frames per category, totaling around 44,000 frames when sampled at 1 frame per second. The dataset was chosen for its developmentally realistic toys and early-learned basic-level categories in child development.Example images from the dataset can be found in Appendix A1.

3. Method

3.1. Problem Statement

In this study, we aim to investigate the optimal dataset strategy as well as optimal learning objective for training a model to identify objects. Specifically, if the task is of classifying dogs, we explore whether providing images of dogs with varying breeds, sizes, and colors or providing images of a single dog from different viewpoints is more effective [2]. Additionally, we also explore learning a general representation for downstream transfer using a supervised contrastive learning objective vs learning the task directly using a supervised cross-entropy learning objective. Although this may seem like a trade-off, it is worth noting that toddlers are capable of categorizing multiple objects during their developmental period when only the latter is available to them.

3.2. Training Details

3.3. Supervised Learning with Cross-Entropy Loss

The model used for training is ResNet-18. We used a cross-entropy loss to train the model for an image classification task. In specific, the cross-entropy loss uses labels and a softmax function to train the classifier. We used two datasets for training and evaluation. The first dataset was CIFAR10 and second dataset used was ToyBox dataset. We trained and evaluated the model on CIFAR10. We used the following transforms: Random Horizontal Flip, Random Resized Crop transform to transform the image to size 224. We trained and evaluated on the ToyBox dataset using a random train-test split of 90%-10%. We don’t use any transforms to train this model. We allow the multiple viewpoints to act as natural transforms of the dataset. A visual explanation of the pipeline can be seen in this figure 2.

3.4. Supervised Contrastive Learning

Given an input batch of data, we first apply data augmentation twice to obtain two copies of the batch. Both copies are forward propagated through the encoder network (ResNet-18) to obtain a 512-dimensional normalized embedding. During training, this representation is further propagated through a projection network that is discarded at inference time. The supervised contrastive loss is computed on the outputs of the projection network. To use the trained model for classification, we train a linear classifier on top of the frozen representations using a cross-entropy loss. A visual explanation of the pipeline can be seen in this figure 3. We used a ResNet18 as our encoder to learn representations along with an MLP as a projection head with an output dimension of 128. For the second stage (downstream transfer) we use the frozen encoder from the pre-training stage and a linear classifier with a single linear layer giving us probabilities for each loss. This linear classifier is trained with

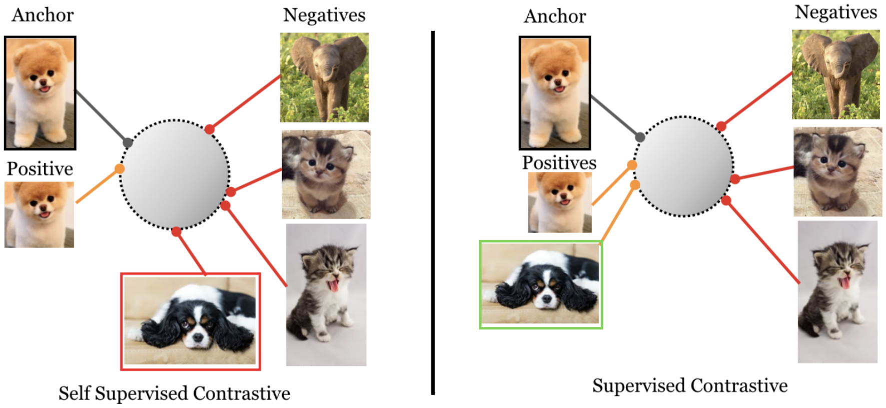


Figure 1: Self-Supervised Learning vs Supervised Contrastive Learning : The self-supervised contrastive loss, shown on uses a single positive for each anchor (an augmented version of the same image) and contrasts it against all other examples in the batch, which serve as negatives. On the other hand, the supervised contrastive loss contrasts all examples belonging to the same class against negatives from the remaining examples in the batch. By incorporating class label information, this approach yields an embedding space where elements belonging to the same class are more closely clustered together than in the self-supervised case.

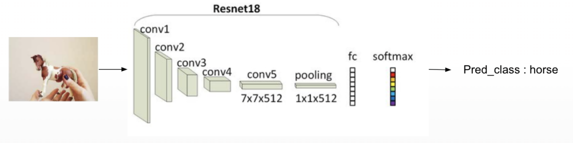


Figure 2: The cross entropy loss uses labels and a softmax loss to train a classifier

cross-entropy loss. During the first and second stages, for CIFAR-10 we use the following augmentations:

- RandomResizeCrop
- Horizontal Flip
- Color Jitter
- Random Grayscale

During the first and second stages, for ToyBox we only use the Random Resize Crop augmentation and once again allow the multiple viewpoint information to act as natural augmentations to learn viewpoint invariant representations.

4. Results and Discussion

4.1. Supervised Learning with Cross-Entropy Loss

After training the model on CIFAR-10 using a supervised cross-entropy loss, we find that this model achieves a test accuracy of 92.68%. After training on the ToyBox dataset using a supervised cross-entropy loss, we find that this model achieves a test accuracy of 97.56%. Since the view semantics of both datasets are completely different, there is no correct way to properly evaluate these two methods. Thus we move on to supervised contrastive-based learning to see the performance between these two.

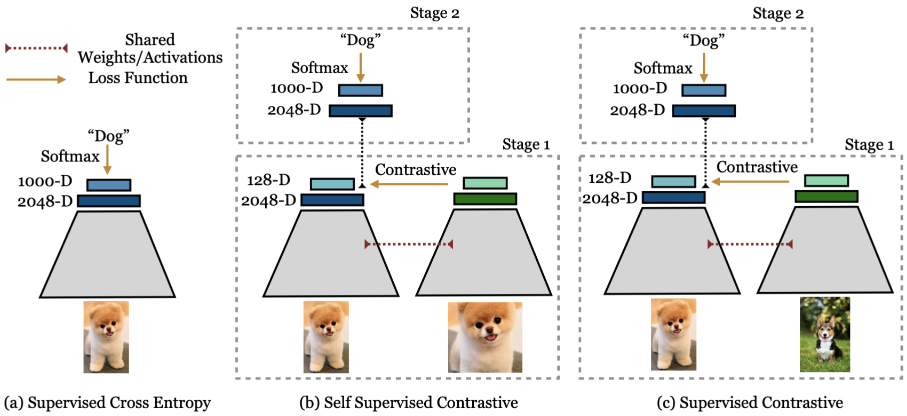
More information on training details such as loss curves and validation accuracy curves can be found in Appendix A2. The classification accuracy can be found in the table 1

Dataset	Model	Accuracy
CIFAR10	ResNet-18	92.68%
ToyBox	ResNet-18	97.56%

Table 1: Image Classification Task trained on the ToyBox Dataset using ResNet-18 performs well on the validation set as compared to the model trained on CIFAR10

4.2. Supervised Contrastive Learning

After pre-training the model on CIFAR-10 using a supervised contrastive loss, and using it for a downstream classification task, we find that it achieves 94.64% accuracy within 100 epochs. After pre-training the model on ToyBox using a supervised contrastive loss, and using it for a downstream classification task, we find that it achieves 96.39% accuracy within 20 epochs. More information on training details such as loss curves can be found in Appendix A3. The classification accuracies can be found in the table **??**. We find that using the supervised-contrastive learning objective, pre-training on the ToyBox dataset leads the model to learn better and more generalized visual representations



(a) Supervised Cross Entropy (b) Self Supervised Contrastive (c) Supervised Contrastive

Figure 3: The cross-entropy loss (left) uses labels and a softmax loss to train a classifier; the self-supervised contrastive loss (middle) uses a contrastive loss and data augmentations to learn representations. The supervised contrastive loss (right) also learns representations using a contrastive loss but uses label information to sample positives in addition to augmentations of the same image. Both contrastive methods can have an optional second stage (downstream transfer) which trains a model on top of the learned representations.

Dataset	Model	Accuracy
CIFAR10	ResNet-18+Linear Classifier	94.64%
ToyBox	ResNet-18+Linear Classifier	96.39%

Table 2: Image Classification Task pre-trained on the ToyBox Dataset using ResNet-18 using supervised contrastive loss performs well on the validation set as compared to the model trained on CIFAR10

as shown by the classification accuracies achieved during the downstream image classification task.

4.3. Qualitative Comparison

Along with the quantitative evaluation of classification accuracies, we also evaluate the quality of representations learned by the various learning objectives pre-trained on the two datasets. We performed a t-SNE visualization of the learned embeddings on the two datasets - CIFAR10 and ToyBox. These visualizations can be viewed in figures 4 and 5

From these visualizations, it is apparent that both the CIFAR and ToyBox models, under the supervised objective, learn only the inter-class variance, and the ToyBox model fails to learn appropriate boundaries. In contrast, in the contrastive case, we can clearly observe the model attempting to vary the intra-class variance, as shown in the bottom-right figure (see Fig: 5).

Furthermore, the embeddings of the ToyBox are grouped according to more fine-grained viewpoints, which is a more nuanced categorization compared to just using CIFAR 10.

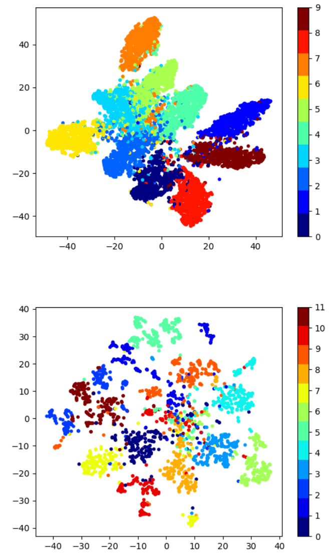


Figure 4: t-SNE visualizations of embeddings learned using supervised Cross Entropy loss trained on CIFAR10 (top) and ToyBox (bottom)

Thus, we provide sufficient evidence to support our hypothesis that when toddlers use their hands to explore objects, the representations they learn are much more fine-grained than those in traditional datasets used for training deep learning networks.

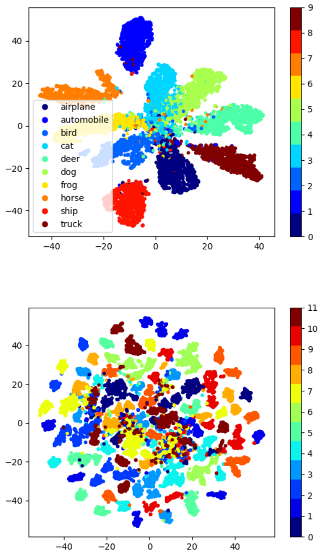


Figure 5: t-SNE visualizations of embeddings learned using supervised contrastive loss trained on CIFAR10 (top) and ToyBox (bottom)

5. Conclusion

In conclusion, we were able to show that pre-training on a multi-view dataset using a supervised contrastive learning objective helps yield more generalized visual representations that can be used for downstream tasks such as image classification. Supervised contrastive learning maximizes not only inter-class variance but is also able to learn intraclass variance between multiple instances of a class. We showed that pre-training on the ToyBox dataset consistently outperforms pre-training on CIFAR10 using both the supervised cross-entropy and supervised contrastive learning objective. We also find that using supervised contrastive learning yields better classification accuracies and leads to more generalizable visual representations than a supervised cross-entropy learning objective.

Hence, we can conclude that it matters not only the pre-training dataset used but also the learning objective for achieving strong generalization performance.

References

[1] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020.

[2] G. Murphy. *The Big Book of Concepts*. The MIT Press, 07 2002.

[3] A. E. Orhan, V. V. Gupta, and B. M. Lake. Self-supervised learning through the eyes of a child. *CoRR*, abs/2007.16189, 2020.

[4] L. B. Smith, S. Jayaraman, E. Clerkin, and C. Yu. The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.*, 22(4):325–336, Apr. 2018.

[5] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[6] X. Wang, T. Ma, J. Ainooson, S. Cha, X. Wang, A. Molla, and M. Kunda. Seeing neural networks through a box of toys: The toybox dataset of visual object transformations, 2018.

6. Appendix

6.1. A1

Figure 6 shows example images from each of the 12 classes in the dataset. Figure 7 shows examples from ToyBox Dataset with multiple views from a single object class.

Figure 8 shows examples from CIFAR Dataset and compares them to examples from the ToyBox dataset to visualize the semantic differences between the two datasets.

We chose to use CIFAR 10 as our baseline model because we discovered a high degree of class overlap between it and the ToyBox dataset, which we intended to use for our experiments. Despite the fact that the visual semantics of the two datasets are completely different, we decided to use CIFAR 10 to ensure a fair comparison between the models.

6.2. A2: Training Details of Supervised Learning Setup with Cross-Entropy Loss

The loss curves for the model trained on CIFAR10 and ToyBox using ResNet-18 is shown in figure 9.

The validation accuracy curves for the model trained on CIFAR10 and ToyBox using ResNet-18 is shown in figure 10.

6.3. A3: Training Details of Supervised Contrastive Learning

The loss curves for the model trained on CIFAR10 and ToyBox using ResNet-18 is shown in figure 11.

6.4. A4: Experiment Details

In our experiments, we followed a configuration-style training approach, which allowed us to easily modify and fine-tune various hyperparameters of the models. To implement this, we utilized open-source timm models, which provided us with a wide range of state-of-the-art pre-trained models to choose from. We would like to express our gratitude to the developers of Timm models for making their code publicly available.

To ensure that our experiments are reproducible, we saved all the log files generated during training and evaluation. These log files contain detailed information about the training process, including the loss, accuracy, and other metrics, which can be used to replicate our results.

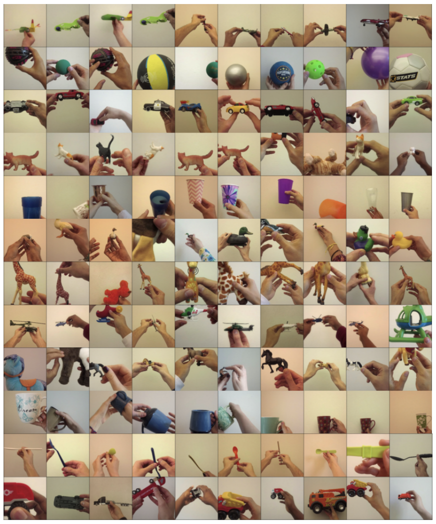


Figure 6: Example images from the Toybox dataset [6]. Each row shows 10 random images from a different class. From the top to the bottom row, the classes are: airplane, ball, car, cat, cup, duck, giraffe, helicopter, horse, mug, spoon, and truck.

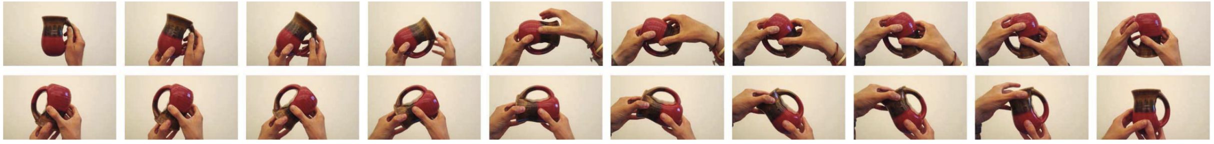


Figure 7: Frames from a video of a mug rotating around the Y + axis from the ToyBox dataset

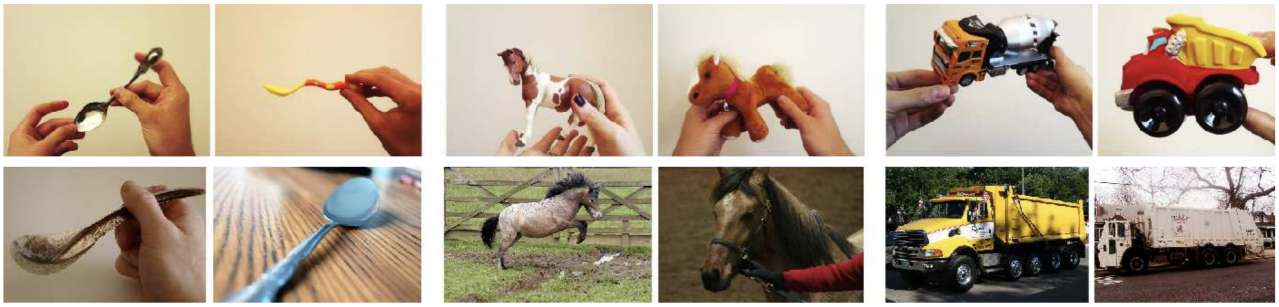


Figure 8: Comparison of images from ToyBox dataset (top) and CIFAR10 (bottom). For ToyBox dataset images, household objects (left) are real, functional objects, though they do come in "adult" and "kiddie" versions. Animals (center) and vehicles (right) are replicas, either "realistic" scale models or "cartoony" toy objects.

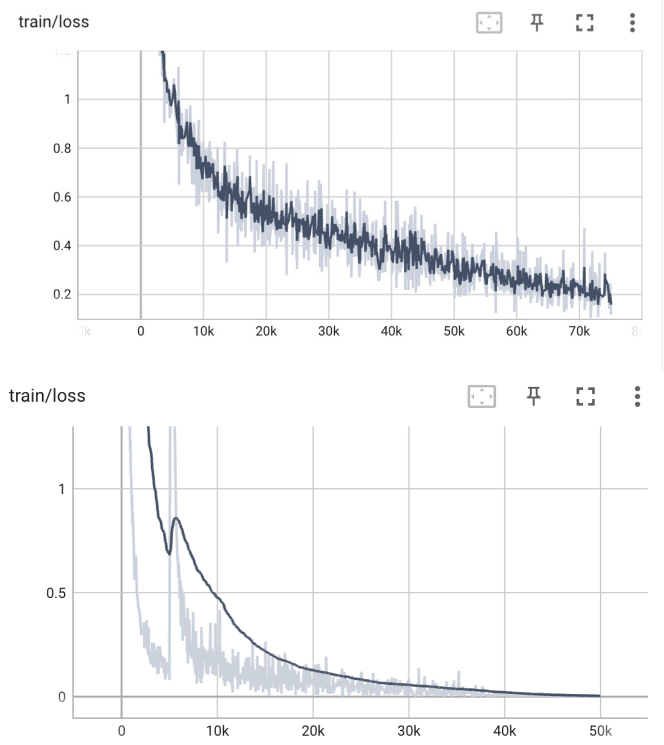


Figure 9: Training Loss Curve for ResNet18 trained on CIFAR10 (top) and ToyBox (bottom)

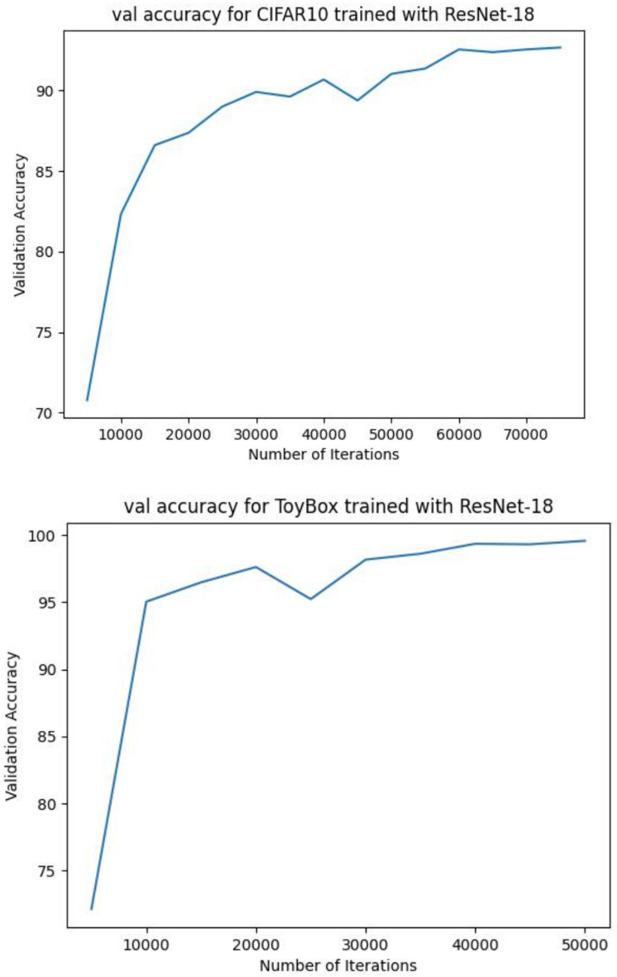


Figure 10: Validation Accuracy Curves for ResNet18 trained on CIFAR10 (left) and ToyBox (right))

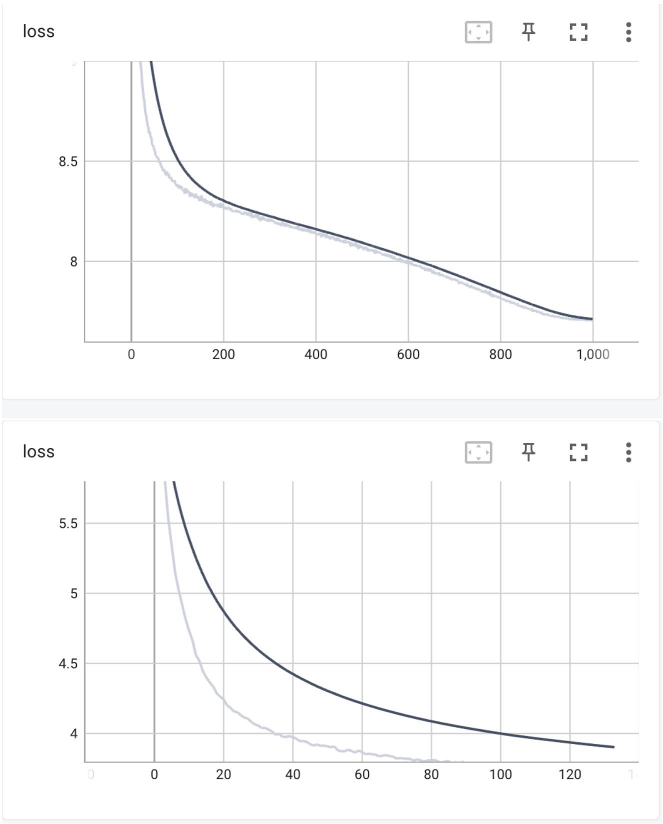


Figure 11: Training Loss Curve for ResNet18 trained on CIFAR10 (top) and ToyBox (bottom)