Evaluating ResNet-18

Authors: Ujjaini Das, Hyunsung Oh, Andrew Teoh

1 Background

We wish to evaluate ResNet-18 to figure out its structure and capabilities. Additionally, we would like to analyze the impact and inner workings of certain layers. We hope to learn more about ResNet as a model and thereby further understand convolutional neural networks as a whole. We will be using CIFAR-10 and ImageNet-V2 as our datasets of choice.

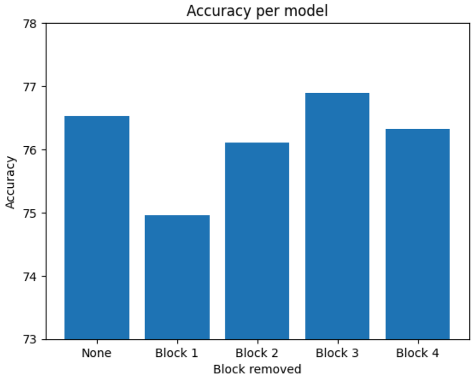
2.1 Experiment 1: Residual Block Removal

We will test a base model and 4 other models where one of the four residual blocks is removed. Then, we will evaluate the accuracy changes of the 4 modified models.

2.1.1 Reasoning and Hypothesis

We would like to see the effects of each residual basic block of ResNet-18 on accuracy. This will test how each block affects performance. We hypothesize that removing earlier blocks will reduce accuracy more since we think earlier blocks learn more fundamental features of the input image.

2.1.2 Figures



The base model is one the left (None), and the four other modified models are on the right.
Accuracies over 10 epochs.

2.1.3 Discussion

The results generally fit the hypothesis; however, when removing block 3, the accuracy increased a little bit. This could be because later layers learn more specific features, and turning them off could reduce overfitting, especially for a simpler and smaller dataset like CIFAR-10. Aside from this, the results indicate that earlier residual blocks provide a foundation for the model and likely learn more basic features, so removing them impacts accuracy more significantly. This can be said for other convolutional networks, as CNNs generally have a hierarchical structure for their convolutional layers (Dutta, 2024).

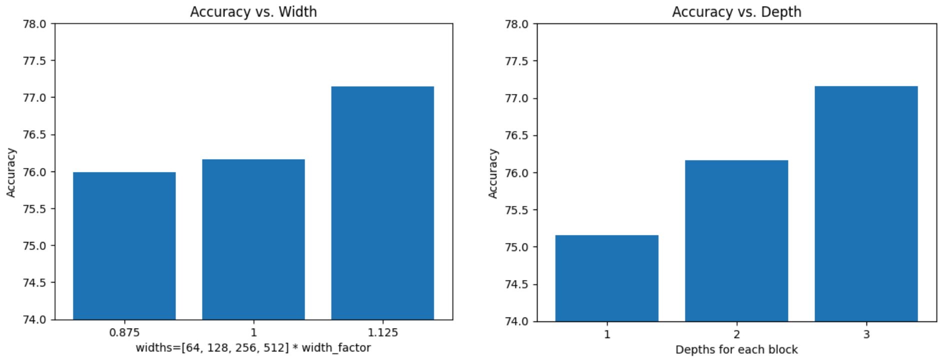
2.2 Experiment 2: Width and Depth vs. Accuracy

The experiment will involve a base model, a less wide model, a more wide model, a less deep model, and a more deep morel. We will test and compare accuracies. The modified width models will have each layers' widths multiplied by a width factor (i.e. 0.875 and 1.125). The modified depth models will have each layers' depth changed by 1 (i.e. 1 and 3).

2.2.1 Reasoning and Hypothesis

We would like to test how ResNet-18's performance would change depending on changes to width and depth. Our hypothesis is that, of course, as width and depth increase so will performance.

2.2.2 Figures



Accuracies over 8 epochs

2.2.3 Discussion

The results, as predicted, convey the positive correlation between width or depth and accuracy. The step size from 1 to 2 and 2 to 3 seems relatively constant in the Accuracy vs. Depth graph, whereas the step size from a width factor of 0.875 to 1 compared to that from a width factor of 1 to 1.125 in the Accuracy vs. Width graph indicates potentially exponential growth. This may mean that width's impact on performance is greater than that of depth's for ResNet-18. On the contrary, there has been research that proves depth plays a greater role in accuracy than width for RELU neural networks (Vardi et al., 2022), which ResNet-18 is.

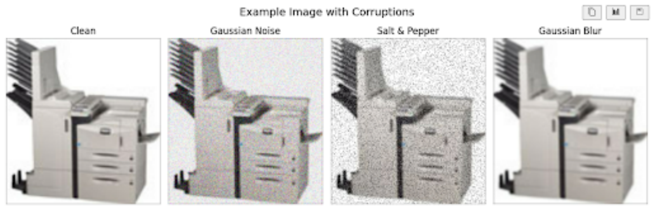
2.3 Experiment 3: Robustness to Noise

In this experiment, we evaluated the robustness of a pretrained ResNet-18 model on the ImageNet-V2 dataset. We first loaded a subset of 500 clean images from ImageNet-V2. To test robustness, we applied three types of controlled corruptions to the dataset: Gaussian noise, salt-and-pepper noise, and Gaussian blur. Each corrupted version of the dataset, along with the clean version, was processed using standard ImageNet normalization and evaluated separately. The dataset was not divided into training and testing subsets, as the pretrained model was evaluated directly on the full clean and corrupted test images to simulate real-world performance degradation under noise.

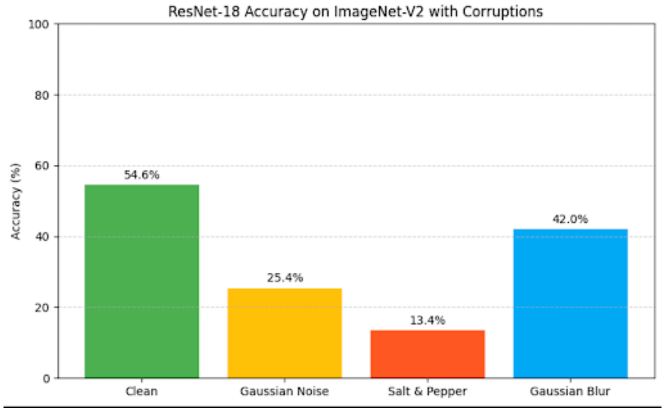
2.3.1 Reasoning and Hypothesis

Images are often affected by imperfections such as noise, blur, and other corruptions which can degrade model performance. Robustness testing evaluates how well a model that is trained on clean data can maintain its accuracy when subjected to these types of distortions. This is important because a model that performs well on clean data but fails under minor corruptions would be unreliable in practical applications. In this study, we hypothesize that adding noise and blur to test images will cause a noticeable drop in classification accuracy. Specifically, we expect that ResNet-18 will be more sensitive to salt-and-pepper noise than to Gaussian blur, given the abrupt pixel-level disruptions introduced by salt-and-pepper corruption.

2.3.2 Figures



Examples of image corruptions applied



Accuracy for each image group tested on ResNet-18

2.3.3 Discussion

Adding different types of noise to the ImageNet-V2 dataset causes a significant drop in accuracy for the ResNet-18 model. The clean dataset achieves an accuracy of 54.6% and performance declines with various corruptions. Gaussian noise reduces accuracy to 25.4%, indicating that the model struggles with random pixel variations that disrupt the image's fine details. Salt & pepper noise leads to an even sharper decline in accuracy (13.4%), as it introduces large areas of complete randomness in the image, which makes feature extraction more difficult. On the other hand, Gaussian blur, which smooths the image by averaging nearby pixel values, reduces accuracy to 42.0%, but the model still performs better compared to the other noise types. This suggests that while the sharp details are blurred, the general structure of the image remains intact, allowing the model to preserve more of its performance.

The differences in accuracy can be attributed to how each type of noise affects the image. Gaussian noise distorts the image in a more subtle, pixel-level manner, which makes it harder for the model to detect features accurately. Salt & pepper noise disrupts the image even more severely, leading to large areas of randomness that obscure key features. Gaussian blu preserves overall object shapes and context, enabling the model to maintain a higher accuracy than with the other types of noise.

2.4 Experiment 4: Class-specific Accuracy

We evaluated the pretrained ResNet-18 on 500 clean ImageNet-V2 images, tracking correct and total predictions per class to compute per-class accuracy. A confusion matrix was generated to visualize misclassifications, and the top and bottom five classes were identified and analyzed.

2.4.1 Reasoning and Hypothesis

We conducted the class-based accuracy experiment to better understand how the model's performance varies across different categories, rather than relying only on overall accuracy. This helps identify specific classes where the model excels or struggles, revealing potential biases or weaknesses. We hypothesize that the ResNet-18 model will perform well on more visually distinct classes, while it will struggle with classes that are visually similar or less common in the training data, leading to lower per-class accuracy and higher confusion rates.

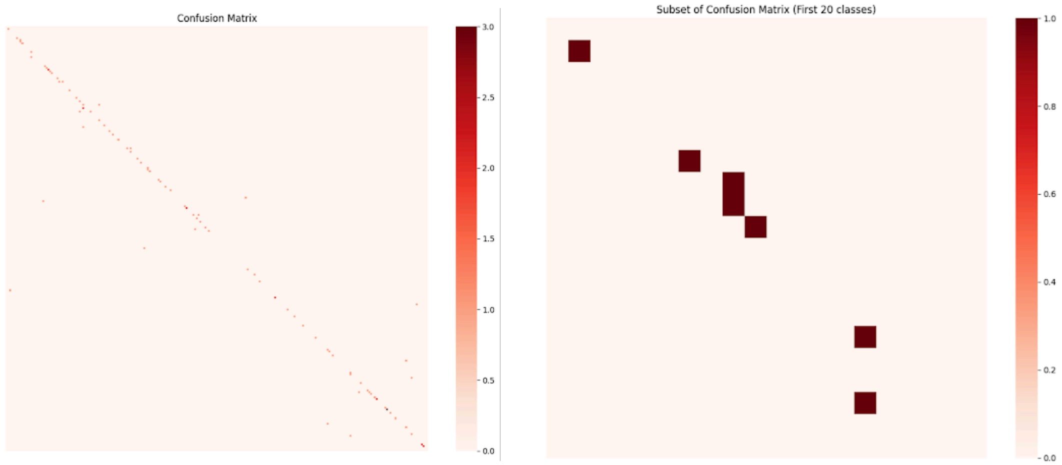
2.4.2 Figures

Top 5 Performing Classes:

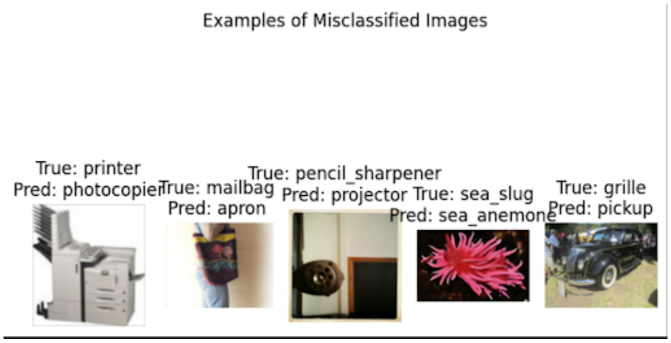
schooner: 100.00%
manhole_cover: 100.00%
cougar: 100.00%
ibex: 100.00%
marmot: 100.00%

Bottom 5 Performing Classes:

printer: 33.33%
wardrobe: 33.33%
perfume: 33.33%
car_mirror: 33.33%
trilobite: 33.33%



Confusion matrix for all classes (left) and first 20 classes (right).



2.4.3 Discussion

The confusion matrix shows the relationship between true labels (on the y-axis) and predicted labels (on the x-axis), where darker shaded squares along the diagonal represent correct predictions and off-diagonal squares indicate misclassifications between classes. There are noticeable deviations and off-diagonal entries, suggesting that the model struggles with certain classes.

The third image highlights a few misclassified examples, where the model's predictions differ from the true labels. For instance, "pencil_sharpener" was misclassified as "projector," and "grille" was predicted as "pickup." These errors suggest that the model may be confusing objects that share similar visual characteristics, such as items in similar contexts (e.g., "printer" vs. "photocopier" or "sea_slug" vs. "sea_anemone"). Overall, the confusion matrix and the misclassified examples suggest that while the model performs well on some classes, its accuracy can be significantly affected by class similarity and the inherent difficulty in distinguishing between certain categories.

2.5 Experiment 5: Feature Visualization

2.5.1 Reasoning and Hypothesis

To better understand how ResNet-18 builds feature representations, we used Grad-CAM to visualize which parts of an input image different layers attend to. Our hypothesis is that earlier layers will highlight broader, less specific regions corresponding to low-level features (like edges or textures), while deeper layers will localize more precisely onto objects or semantically meaningful parts of the image. We also expect that if the model misclassifies an image, Grad-CAM might still highlight relevant parts, revealing meaningful internal behavior even when predictions are wrong.

2.5.2 Figures

We applied Grad-CAM to a pretrained ResNet-18 model trained on CIFAR-10 and visualized activation maps from four different layers: layer1, layer2, layer3, and layer4. We tested this with three classes: cat, dog, and truck.

Representative examples:

- **Layer 1**: Wide and diffused activation across the image; little focus on specific object regions.

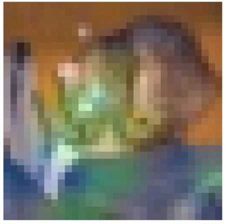


Figure 5.1: Grad-CAM visualization for Cat — Layer 1

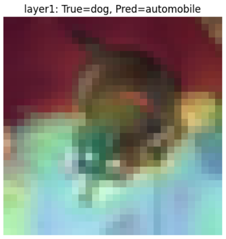


Figure 5.2: Grad-CAM visualization for Dog — Layer 1

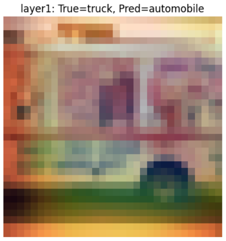


Figure 5.3: Grad-CAM visualization for Truck — Layer 1

- **Layer 2**: Similar to Layer 1, but slightly stronger activations around object boundaries.

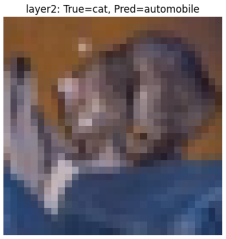


Figure 5.4: Grad-CAM visualization for Cat — Layer 2

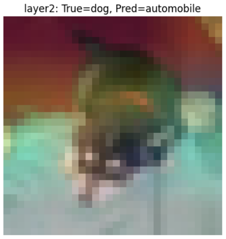


Figure 5.5: Grad-CAM visualization for Dog — Layer 2

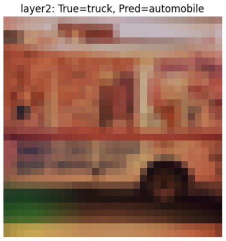


Figure 5.6: Grad-CAM visualization for Truck — Layer 2

- **Layer 3**: Activations begin to concentrate more around salient object parts (e.g., general object outline).

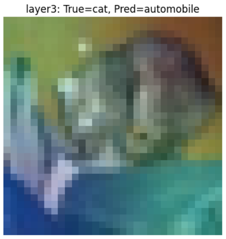


Figure 5.7: Grad-CAM visualization for Cat — Layer 3

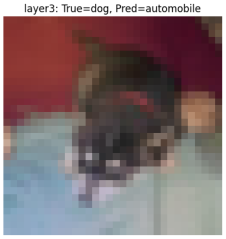


Figure 5.8: Grad-CAM visualization for Dog — Layer 3

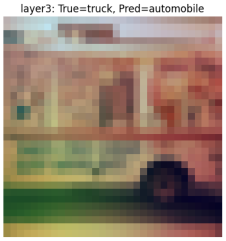


Figure 5.9: Grad-CAM visualization for Truck — Layer 3

- **Layer 4**: Activations still relatively broad, but slightly more intense on key object areas compared to earlier layers. Overall focus improvement is modest.

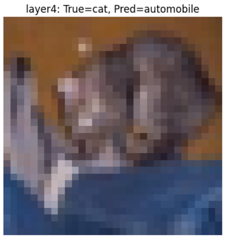


Figure 5.10: Grad-CAM visualization for Cat — Layer 4

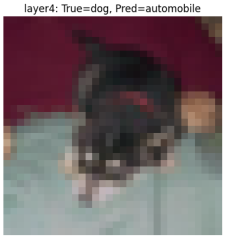


Figure 5.11: Grad-CAM visualization for Dog — Layer 4

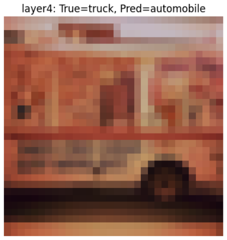


Figure 5.12: Grad-CAM visualization for Truck — Layer 4

2.5.3 Discussion

While Grad-CAM visualizations for different layers did show some differences, the change in localization sharpness across layers was modest. Early layers (Layer 1 and Layer 2) produced wide and diffuse activations over much of the image. As the depth increased (Layer 3 and Layer 4), activations became slightly more focused around salient object parts, but not dramatically. This limited differentiation is likely due to the relatively shallow depth of ResNet-18 and the low resolution of CIFAR-10 images, which make strong hierarchical feature abstraction less necessary. Overall, Grad-CAM still provided useful insights into the internal attention patterns of ResNet-18, though the layer-wise differences were subtler than initially expected.

3 References

1. Dutta, S. (2024, October 6). *Understanding the convolutional layer in convolutional neural networks (cnns)*. Medium. https://medium.com/@sanjay_dutta/understanding-the-convolutional-layer-in-convolutional-neural-networks-cnns-ef4065a0e3ca
2. Vardi, G., Yehudai, G., & Shamir, O. (2022, June 1). *Width is less important than depth in Relu Neural Networks*. arXiv.org. https://arxiv.org/abs/2202.03841
3. He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. arXiv.org. https://arxiv.org/abs/1512.03385
4. Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* (Technical Report). University of Toronto.
5. Selvaraju, R.R., Cogswell, M., Das, A. et al. (2016). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. arXiv.org. https://arxiv.org/abs/1610.02391
6. Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). *Do ImageNet classifiers generalize to ImageNet?* In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:5389–5400.

4 Contributions

- Ujjaini Das - Experiments 3 & 4
- Hyunsung Oh - Experiment 5
- Andrew Teoh - Experiments 1 & 2
- ChatGPT - Model Card and Datasheets

5 Model Card

Model Overview

Name: ResNet-18
Architecture: Convolutional Neural Network (CNN) with residual connections (skip connections)
Developer: Kaiming He et al. (2015)
Purpose: Image classification on medium-scale datasets (e.g., CIFAR-10, ImageNet)

Intended Use

- **Primary Task:** Image classification
- **Dataset Used:** CIFAR-10 (10 classes, 32×32 color images)
- **Training Objective:** Minimize cross-entropy loss between predicted and true class labels.
- **Users:** Machine learning practitioners, researchers conducting model ablation or robustness studies.

Model Details

- **Input:** RGB images of size 32×32
- **Output:** Softmax probabilities over 10 classes
- **Architecture:**
  - 1 initial convolutional layer
  - 4 groups of residual blocks (2 blocks each)
  - Each block contains two 3×3 convolutions with a shortcut connection
  - Downsampling via strided convolutions between layers

    - Global average pooling before final fully connected (FC) layer

- **Number of Parameters:** ~11.7 million (for original ResNet-18)
- **Optimizer Used:** Adam (lr=0.001)
- **Loss Function:** Cross-Entropy Loss

Training & Evaluation

- **Training Dataset:** CIFAR-10 training set (50,000 images)
- **Test Dataset:** CIFAR-10 test set (10,000 images), ImageNet-V2 (500 images)
- **Evaluation Metrics:** Top-1 accuracy

Performance

- **Baseline ResNet-18 Test Accuracy:** ~90% on CIFAR-10 (with standard training)
- **Impact of Ablations:**
    - Removing early layers (e.g., layer1) leads to significant performance degradation.
    - Removing mid-to-deep layers (e.g., layer3) can sometimes slightly improve generalization on simple datasets like CIFAR-10.

Limitations

- **Dataset Specificity:** Optimized for small images (CIFAR-10); resizing needed for larger datasets (e.g., ImageNet).

- **Overparameterization Risk:** Potential overfitting when used on small datasets without regularization.
- **Adversarial Vulnerability:** Like most CNNs, ResNet-18 can be sensitive to small input perturbations.

6 Datasheet - CIFAR-10

Motivation

For what purpose was the dataset created?

- CIFAR-10 was created by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton to benchmark machine learning algorithms, particularly for supervised image classification tasks on small, natural images.

Who created the dataset and on behalf of which institution?

- Researchers at the University of Toronto.

What tasks is the dataset intended to support?

- Image classification, representation learning, and supervised learning experiments.

Composition

What are the instances in this dataset?

- 60,000 color images (32×32 pixels), divided into 10 object categories.

What is the label set?

- airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck

How many instances are there?

- 50,000 training images and 10,000 test images.

Does the dataset contain all instances from a particular population?

- No. It is a curated subset of the larger Tiny Images dataset.

Does the dataset contain sensitive information?

- No. All images are publicly available and do not contain personally identifiable information.

Collection Process

How was the data collected?

- Images were selected and categorized manually from the Tiny Images dataset.

Over what time frame was the data collected?

- Originally compiled around 2008–2009.

Was any preprocessing performed?

- Images were resized to 32×32 pixels. No other pre-processing (like normalization or color correction) was applied.

Preprocessing/cleaning/labeling

Was any preprocessing done?

- Minimal. Images were downscaled; no additional augmentation or filtering.

Who was responsible for labeling?

- Researchers manually categorized and verified labels.

Is the dataset self-contained, or does it link to external resources?

- It is self-contained.

Uses

What are the primary intended uses of the dataset?

- Benchmarking image classification models.
- Studying generalization, overfitting, and transfer learning on small images.

What are the known uses of the dataset?

- Model evaluation (e.g., ResNet, DenseNet, VGG)
- Curriculum learning
- Adversarial robustness experiments

Is there a repository that links to papers using CIFAR-10?

- No official repository, but CIFAR-10 is cited in thousands of papers.

Distribution

Will the dataset be freely available?

- Yes, it is available under the MIT license.

Where can it be accessed?

- CIFAR-10 official page

- Also available through machine learning libraries (e.g., PyTorch torchvision.datasets, TensorFlow tf.keras.datasets).

Are there any restrictions on use?

- No. The dataset is free for research and educational purposes.

Maintenance

Who maintains the dataset?

- There is no formal maintenance — it is static and unchanging.

Will the dataset be updated?

- No. CIFAR-10 is a finalized dataset.

Who should be contacted for questions?

- Originally: Alex Krizhevsky, University of Toronto (no ongoing support currently).

7 Datasheet - ImageNet-V2 (with added noise)

Motivation

For what purpose was the dataset created?

- ImageNet-V2 was created to evaluate and benchmark the performance of image classification models. The images were modified with various types of corruption or alterations.

Who created the dataset and on behalf of which institution?

- Researchers at Stanford University.
- We added noise/perturbations to a subset of 500 images.

What tasks is the dataset intended to support?

- Image classification, representation learning, and supervised learning experiments.

Composition

What are the instances in this dataset?

- ImageNet-V2 consists of 10,000 images across 1,000 object categories.

What is the label set?

- The label set consists of 1,000 classes, which include a wide range of objects such as animals, vehicles, instruments, and more.

How many instances are there?

- 10,000 test images

Does the dataset contain all instances from a particular population?

- No, it is a subset derived from the original ImageNet validation set.

Does the dataset contain sensitive information?

- No. All images are publicly available and do not contain personally identifiable information.

Collection Process

How was the data collected?

- Images were selected as a subset of the original ImageNet dataset. Disturbances (Gaussian noise, Gaussian blur, salt and pepper noise) were then added.

Over what time frame was the data collected?

- The dataset was released in 2021 as an extension of ImageNet for robustness testing purposes.

Was any preprocessing performed?

- Images were resized and standardized to a consistent format for testing. Corruptions such as noise and blur were applied to the images.

Preprocessing/cleaning/labeling

Was any preprocessing done?

- Yes, the images were resized and standardized to fit the input requirements of popular deep learning models. Corruptions like Gaussian noise, Gaussian blur, and others were applied to simulate real-world disturbances.

Who was responsible for labeling?

- Researchers manually categorized and verified labels.

Is the dataset self-contained, or does it link to external resources?

- It is self-contained.

Uses

What are the primary intended uses of the dataset?

- Benchmarking image classification models.
- Studying generalization, overfitting, and transfer learning on small images.

What are the known uses of the dataset?

- Model evaluation (e.g., ResNet, DenseNet, VGG)
- Curriculum learning
- Adversarial robustness experiments

Is there a repository that links to papers using CIFAR-10?

- No official repository, but ImageNet-V2 is cited in multiple academic papers.

Distribution

Will the dataset be freely available?

- Yes, it is available under the MIT license.

Where can it be accessed?

- ImageNet official page
- Also available through machine learning libraries (e.g., PyTorch torchvision.datasets).

Are there any restrictions on use?

- No. The dataset is free for research and educational purposes.

Maintenance

Who maintains the dataset?

- There is no formal maintenance — it is static and unchanging.

Will the dataset be updated?

- No. ImageNet-V2 is a finalized dataset.

Who should be contacted for questions?

- ImageNet team at Stanford University (Li et. al.)