Comparative Analysis of Data Augmentation Techniques

Srikar Devulapalli

The two popular data augmentation techniques CutMix and CutOut, which are used to improve the generalization of models by introducing random variations to the training data are compared in this report in terms of their accuracy of model on the CIFAR-10, CIFAR-100 datasets.
CutMix is a data augmentation method that combines two images by randomly selecting a patch from one image and pasting it on the other.
Cutout randomly masks out a portion of an image to artificially introduce more diversity in the training data.

Methodology and choices made

1. Collect and prepare a dataset followed by exploration
    - CIFAR-10 and CIFAR-100 are datasets consisting of 60,000 32x32 color images in 10 and 100 classes, respectively, with 6000 images per class, and a test set of 10,000 images.
2. Perform data pre-processing
    - Split test and train data into images and labels
    - Convert the labels into one-hot encoded format for compatibility
    - Autotuning for optimizing data loading
    - Resizing and normalizing for maintaining consistency in data
    - Shuffling the data to improve the model's generalization
3. Define the type of data augmentation and apply the techniques on the dataset
4. Train a model using the augmented data
    - Defined a deep learning model using the ResNet architecture with 20 layers and a depth of 20
    - This is chosen because it is faster to train and less computationally expensive, while still maintaining good performance
5. Evaluate the model's performance using various metrics
6. Compare the model's performance with and without data augmentation

Results:

1. Comparison of model accuracy for cut-mix and cut-out algorithm across Cifar-10 and Cifar-100 datasets

    Findings:
    Cut-Mix comparatively has higher accuracy than cut-out as shown in Fig.1

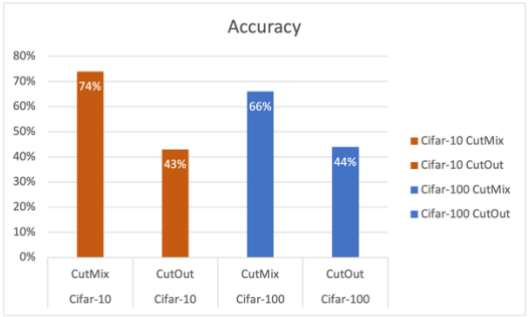

Fig.1: Accuracy comparison for cut-mix and cut-out augmentation techniques

Hypothesis:
CutMix is better than CutOut because it combines the features of two images, creating a more diverse dataset, while CutOut only masks a portion of one image, potentially losing important features.

2. Comparison of model accuracy for non-augmented data across Cifar-10 and Cifar-100 datasets

Findings:
Predictions on Cifar-10 comparatively has higher
accuracy than Cifar-100

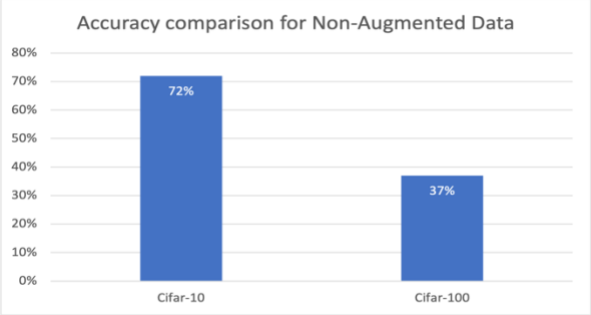

Fig.2: Model accuracy for non-augmented data across Cifar-10 and Cifar-100

3. A plot of Accuracy vs. the number of epochs for different algorithms

Findings:
The accuracy for Cut out Cifar-10 increases as number of epochs increase as shown in Fig.3

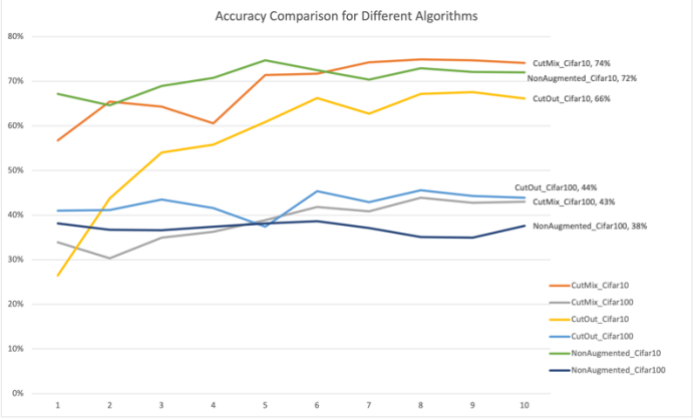

Fig.3: A plot of Accuracy vs. the number of epochs for different algorithms

Hypothesis:
As the model is trained over more epochs, it is exposed to
more training examples and thus can learn and generalize better which leads to increased accuracy.

Future Scope

- Investigating the effect of data augmentation on different types of data and models
- Introduce parallel processing and distributed computing to help increase the speed at which data augmentation is performed.
- Investigating the patterns of validation metrics (like accuracy) by varying the number of epochs and batch sizes

Notebooks:

1. https://colab.research.google.com/drive/1RNhpb7bczGAfd_dy7ZChB9bAABPp32Gk?authuser=2#scrollTo=Wn0A7WBP8evu
2. https://colab.research.google.com/drive/1RNhpb7bczGAfd_dy7ZChB9bAABPp32Gk?authuser=2#scrollTo=Wn0A7WBP8evu
3. https://colab.research.google.com/drive/14rTUEMr-VE-lCDIrYiiXSrxMP9dWXlT8?authuser=1#scrollTo=XwZBfOIPvuVt
4. https://colab.research.google.com/drive/14II7z5djh0u4XGApk2jZvld1ErELk0Mq?authuser=1#scrollTo=mcWd3jW2tOQ1
5. https://colab.research.google.com/drive/137pbTDeSBePP_dCTXNXd5Xh4vaXnMW2K?authuser=2#scrollTo=D0NGyTsX7YSi
6. https://colab.research.google.com/drive/1ryVdVhnNl447xqR29duqR-TgTZ-Q0gRE#scrollTo=D0NGyTsX7YSi

References:

1. https://keras.io/examples/vision/cutmix/#define-the-cutmix-data-augmentation-function
2. https://www.cs.toronto.edu/~kriz/cifar.html
3. https://jovian.ml/bountyhunter1999/cifar100-course-project/v/14?utm_source=embed
4. https://www.analyticsvidhya.com/blog/2021/08/how-to-code-your-resnet-from-scratch-in-tensorflow/