Analyzing the Effectiveness of Data Augmentation From the Perspective of Distribution Shift

Jacob Levine [1] Colin Lu [1] Matthew Tang [1]

[1]University of Illinois at Urbana-Champaign

Motivation

While a major benefit of data augmentation is **increasing the number of data points** yielding better bounds on generalization error, augmenting data also **incurs distribution shift** between the augmented training set and the test set, which can be harmful depending on the nature of the shift. We are interested in exploring how this **distribution shift relates to the "quality"** of augmentations, i.e. exploring what sorts of augmentations tend to be more effective in improving test performance, and whether they share any patterns in distribution shift.

Related Work

There have been studies done on how image augmentation affects the model's robustness towards distribution shift and OOD generalization [1][3]. Additional studies have benchmarked the the effectiveness of different image augmentations on model performance [2]. However, none of these previous works analyze the distribution shift caused by the data augmentation itself. We seek to discover whether the induced distribution shift by the image augmentation relates to the improvement of model performance.

Background: Augmentations and Distribution Shift

Data augmentation is a powerful method for improving generalization accuracy, which involves adding transformed data (e.g. applying flips, crops, etc.) to the dataset. A common perspective is that this makes the model robust to invariant properties, e.g. rescalings or horizontal flips should not change the class (depending on the domain).

However, from another perspective, it should be perplexing that augmentations are often so effective. Since they are not new IID data points, augmentations **induce distribution shift** between the test set and the (now-augmented) training set, which is harmful.

We resolve the discrepancy between these two perspectives by noting that augmentations can be beneficial due to increasing the number of data points (of varying informativeness). In the best case, if they correspond to invariances within the dataset, they may induce **minimal distribution shift** while producing **maximally informative** (i.e. effectively new IID) data points.

Background: Distribution Shift Metric

There are many options to measure distribution shift. To choose, we use three heuristic criteria:

1. Is a **metric** (excludes KL divergence)
2. **Considers geometry** of the space + is informative for empirical distributions in continuous space (excludes TV, JS distance, and Hellinger)
3. Is **easily computable** (excludes Lévy-Prokhorov, based on preliminary investigation)

Wasserstein distance satisfies the criteria, being easily computable in high dimensions for empirical distributions using off-the-shelf minimum bipartite matching algorithms. Happily, it is also a common metric of choice for distribution distance.

Wasserstein distance (for empirical distributions) is calculated as follows: Given two datasets P and Q each having n examples,

$$W_p(P,Q) = \inf_{\pi} \left(\sum_{i=1}^{n} \|P_i - \pi(P_i)\|_2^p \right)^{\frac{1}{p}}$$

where $\pi : P \to Q$ is a bijection. The most common choices for p yield W_1, "Earth Mover's distance", and W_2, our choice. We chose W_2 as it penalizes a pairing with a few far movements over one with many small movements, which we conjecture affects duplicate data more.

Abstract

We study the effect of data augmentation on the data distribution and quantify the relationship between the induced distribution shift and model performance. We leverage Wasserstein distance in CLIP and pixel space to measure distribution shift between augmented and original data, attempting to satisfy desiderata for a useful predictor of augmentation effectiveness. This method captures some of the desiderata, but needs tuning to be a useful predictor. Still, the perspective shows promise and, if effective, links theory for distribution shift with the empirically-effective technique of data augmentation.

Method: Desiderata for Distribution Shift Metric

We propose that a useful distribution shift metric for predicting the effectiveness of augmentations satisfies the following desiderata:

1. The minimum distance (on average) is obtained by two IID datasets (i.e. no shift)
2. Adding uninformative/non-IID data (e.g. copies or near-identity augmentations) increases the distance, counteracting claimed test error improvement from increased "IID" dataset size
3. Augmented data reflecting invariances (i.e. are effectively IID) has a small distance, and augmented data discarding information has a large distance

We empirically verified these properties for W_2 on a multivariate normal Gaussian setting, duplicating a small dataset and adding "augmentations" in the form of zero-mean Gaussians.

Method: Experiment Settings

For these experiments, we use CIFAR-10 for a lightweight dataset, and choose the following augmentations: **identity** (i.e. copy), Gaussian **blur**, random **crop**, **horizontal flip**, random **rotation**, random **shift + scale**, **grayscale**, and random **brightness and saturation** adjustment. To construct an augmented dataset, we sample n points + apply the augmentation to each, yielding $2n$ points.

For experiment 1, we train models on various above augmented datasets, and evaluate test error. For experiment 2, we evaluate the distribution shift between the augmented datasets and original, and see if shift corresponds to performance. To evaluate desiderata 1, we also compare against a fully IID baseline ($2n$ IID) for the augmented set. We also evaluate the metric in CLIP space and pixel space, taking 10 samples to obtain 95% confidence intervals for the estimated metric value.

Experiment 1 Results

After fine-tuning a ResNet-18 on train datasets with the number of base examples specified in the top row, we obtained the following validation losses. Note that this leverages older data, on a somewhat different setting (all data is augmented instead of 50-50 + resample augmentations every epoch), which we seek to rerun on the new setting, evaluating accuracy instead of loss.

Augmentation	1000	2000	4000	8000	16000	32000	50000
Identity	2.1	1.7	1.5	1.4	1.2	1.0	0.9
Blur	2.0	1.8	1.6	1.4	1.2	1.1	1.1
Bright. + Sat.	1.8	1.6	1.5	1.3	1.1	1.0	0.8
Crop	2.2	2.3	2.2	2.2	2.1	2.2	2.0
Horizontal Flip	1.8	1.8	1.4	1.2	1.1	0.9	0.8
Grayscale	2.1	1.6	1.5	1.4	1.2	1.0	0.9
Rotation	1.9	1.7	1.4	1.3	1.1	1.0	0.9
Shift + Scale	1.9	1.7	1.4	1.2	1.0	0.9	0.8

Table 1. Validation scores of the various augmentations

Experiment 2 Results and Analysis

We calculated the following W_2 distances between the augmented datasets and the original dataset, as described in the experiment settings section:

Augmentation	CLIP Space W_2	Pixel Space W_2
IID Standard	4.76 ± 0.08	386.35 ± 1.54
Identity	5.07 ± 0.06	396.15 ± 1.57
Blur	4.87 ± 0.11	384.80 ± 1.14
Bright. + Sat.	5.01 ± 0.07	393.93 ± 2.17
Crop	5.19 ± 0.12	377.68 ± 2.16s
Horizontal Flip	4.82 ± 0.07	389.13 ± 1.81
Grayscale	4.91 ± 0.08	388.09 ± 1.80
Rotate	10.23 ± 0.24	435.85 ± 1.90
Shift + Scale	6.86 ± 0.14	439.45 ± 2.25

Table 2. W_2 distances of the various augmentations from a reference dataset sampled IID from CIFAR-10

As expected, most of our augmentations yield W_2 distances between the IID standard and the identity baseline, and correspondingly yield test errors between the two. However, both blurring and cropping yield lower pixel-space W_2 values than entirely IID data; we conjecture that they create more "average" images in pixel space which W_2 prefers, which is likely addressable with W_1 or in semantic space (e.g. in CLIP space, their distances are larger).

Furthermore, some augmentations, despite training good models, yield large distances. This could either be the result of CLIP not being exposed to these transformations and using them as changes in semantic meaning, or a fundamental weakness in our metric's ability to characterize different augmentations. However (though not pictured here) we observed that the metric does correspond to the time the model spends to reach minimum validation error – rotate and shift/scale taking significantly longer than the others.

Future Work

Most notably, future work may address the **weakness of the existing metric** – penalizing augmentations which capture useful invariances but deviate from the input distribution. In particular, it should not penalize augmentations mapping the input distribution to regions unoccupied by data points of a *different* class, i.e. leveraging label information.

Another idea is to explore **different embedding spaces**. In particular, using a trained model's latent space to explain why some augmentations may work only for specific architectures. Ideally, this can also give insights as to the effectiveness of useful but heavily input-perturbing augmentations, which cannot be explained with the current metric.

A third point for future work is replacing the empirical distribution with a **Gaussian KDE-style PDF estimate** where each point contributes Gaussian probability around itself, with variance scaling with $\frac{1}{n}$. This further penalizes near-identity augmentations, and allows using geometry-ignoring metrics like TV distance, if desired. The main challenge seems to be approximate computation of the metric, as exact computation seems intractable in high dimensional space.

Lastly, as a brief idea, we would like to explore how to leverage a good distribution shift metric to **propose new augmentations**, rather than simply evaluating existing augmentations.

References

[1] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, Xiangyang Ji, and Antoni B. Chan. An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation, 2022.

[2] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning, 2022.

[3] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation, 2022.