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

Jacob Levine **Colin Lu** **Matthew Tang**

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 (after tuning), links theory for distribution shift with the empirically-effective technique of data augmentation.

1 Introduction

1.1 Motivation

In computer vision and many other domains, data augmentation is a widely used and often highly effective approach for improving generalization error, which involves taking a training set and applying transformations to the inputs in a way which should not change the predicted class, and adding those transformed inputs as "augmented" data. In computer vision, this includes transformations such as flipping, blurring, and cropping. This is frequently framed as encouraging the model to learn certain "invariances" in the data (e.g. the model should not care if a car is facing left or facing right, and should still identify it as a car); from another perspective, data augmentation's effectiveness comes from increasing the number of "independent"/IID data points which helps generalization bounds.

On the other hand, we note that augmenting data also voluntarily incurs distribution shift between the augmented training set and the test set, which is strange at first glance as distribution shift is frequently harmful for performance. We are interested in exploring how this distribution shift relates to how effective an augmentation is at improving generalization accuracy of a trained classifier. We conjecture that certain augmentations are good because they either produce near IID data points without significant distribution shift, or they produce independent data points where the distribution shift is not harmful (why that would be the case, needs to be investigated). On the other hand, certain augmentations may be ineffective or even detrimental if the distribution shift is large and harmful, or simply if the augmentation fails to produce new "independent" data, e.g. in the limiting case where the augmentation is an identity transformation.

In particular, we hope to construct a measure of distribution shift which correctly captures *harmful* distribution shifts, while not considering benign distribution shifts as generating additional distance. We elaborate more on the desiderata in the methods section. One may additionally ask whether it is even meaningful to call this a measure of distribution shift when we disregard certain kinds of distribution shifts (in particular, by the standard of whether the shift is harmful or not for the sake of test accuracy on the augmented dataset) – is the optimal measure not just some complex predictor of test accuracy (e.g. in the most extreme case, training a model to predict the test accuracy and using this as the "measure")? However, we note that we wish to keep the measure reasonably similar to a metric over distributions, making at most a few simple, principled changes to adhere to the desiderata.

Preprint. Under review.

1.2 Related Work

There have been studies done on how image augmentation affects the model's robustness towards distribution shift and OOD generalization [2][4]. Additional studies have benchmarked the the effectiveness of different image augmentations on model performance [3]. There have also been image datasets designed to test a model's robustness to different distribution shifts, which were used to test new image augmentations [1]. 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.

2 Methodology

2.1 Background: Distance Metrics

While our method intends to allow any "pipeline" of operations that eventually lead to a good measure of distribution shift corresponding to augmentation effectiveness, one major component of this pipeline is the distance function we use. In particular, as we are comparing distributions (datasets), we need to use a notion of distance between two distributions. We consider the following choices (but omit their expressions), where we attempt to include as many candidates of distribution distances that we are aware of:

- KL divergence
- TV distance
- Wasserstein distance
- Jensen-Shannon divergence (and its square root, JS distance)
- Hellinger distance
- Lévy-Prokhorov metric

Among these options, we establish some criteria for determining which to use, although note that these criteria are heuristic:

1. Is a metric (loosely meaning it is non-negative, symmetric, and adheres to the triangle inequality)
2. Considers geometry of the space and 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)

The first criterion rules out KL divergence and Jensen-Shannon divergence, as they violate the triangle inequality (and KL divergence is asymmetric).

For the second criterion, Figure 1 illustrates an example of TV distance not considering geometry of the space. Hellinger distance and JS distance (as well as KL) suffer from this as well, noting that their expressions are dependent on integrals over the joint support of the two distributions, but consider each point in the support separately. Furthermore, this is not only suboptimal for heuristic reasons, but when applied on empirical distributions where datapoints are sampled from distributions having probability density on an infinite number of elements in the support (e.g. a typical "smooth" distribution with continuous support like Gaussian), two independently sampled distributions almost surely do not overlap, which makes such non-geometric metrics uninformative.

For the last criterion, we need a computationally tractable metric. Based on our brief initial investigations, Lévy-Prokhorov seems difficult to compute in high dimensions and with large numbers of data points. Wasserstein is tractable on empirical distributions in high dimensions, however, and furthermore can be computed using off-the-shelf minimum bipartite matching algorithms.

Thus, **Wasserstein distance** satisfies the criteria. Happily, it is also a common metric of choice for distribution distance.

In this project, we consider Wasserstein distance on empirical distributions with the same number of data points n, though Wasserstein applies to arbitrary can be generalized (unsure about tractability).

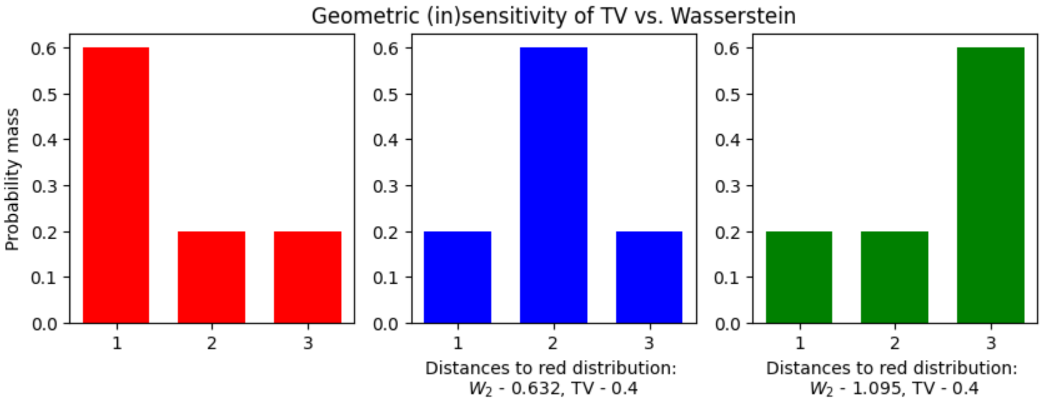


Figure 1: Wasserstein is an example of a metric which does consider the geometry of the space, as opposed to TV which does not. The blue distribution has more probability "closer" to red, whereas green's probability is "farther" away from red when compared to blue. Then red-blue W_2 is smaller than red-green W_2. On the other hand, TV is geometry-invariant and has the same TV distances for red-blue and red-green.

This is calculated as follows: given two empirical distributions P and Q consisting of samples $P_1, ..., P_n$ and $Q_1, ..., Q_n$ respectively,

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

where π is a permutation over $1, ..., n$. As noted, such a minimization problem is solvable by computing the pairwise distances between P_i and Q_j for all i, j and then computing a minimum bipartite matching. 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, though this is fairly arbitrary (it is possible other choices of p are preferred, e.g. $p = 1$).

2.2 Desiderata for Distribution Shift Metric

The above considerations were relevant for picking a possibly reasonable choice for measuring some notion of distance (i.e. shift) between two distributions. By the above criteria, we have established that our choice is tractable (Wasserstein on two empirical distributions with the same number of samples) and satisfies a few common sense assumptions (is a metric i.e. symmetric/triangle inequality, and takes into account the geometry of the space). However, this is by no means sufficient to ensure that our measure of distribution shift is useful for capturing "harmful" shifts as opposed to benign shifts.

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

1. **Minimum shift is IID**: if there is no distribution shift, i.e. we compare two IID datasets, we should have a smaller shift measure (on average) than any other two non-IID datasets.
2. **Penalize non-informative samples**: if we fail to penalize non-informative samples (e.g. augmentations which barely perturb the data and not in a useful way, or in the most extreme case, taking a copy i.e. identity), we will have the same distribution shift measure as IID, even though such copies are useless to improving performance. From another perspective, we claim to be adding new IID data and thus should have test error improvement, but they are not really IID, so we should balance it out (the shift should be large). Metrics on empirical distributions which don't take into account the sample size or the variance of the estimate may be vulnerable to this (since the empirical distribution does not change when duplicated).

3. **Good augmentations have small shift**: augmented data reflecting invariances are effectively IID (or close to it), and as such should have a small shift measure.
4. **Bad augmentations have large shift**: augmented data discarding useful information or interfering with classification should have a large shift measure, possibly more than non-informative augmentations.

2.3 Definition of Distribution Shift Measure

We define the distribution shift of an augmentation by sampling n IID data points from a dataset, applying the augmentation to each, yielding a total of $2n$ points in the augmented dataset A_{aug}, then comparing A against a baseline dataset B consisting of another $2n$ IID data points with the W_2 metric, i.e. $W_2(A_{aug}, B)$. For interpretability, we subtract the IID shift $W_2(A_{IID}, B)$ and divide by the shifted copy $W_2(A_{copy}, B) - W_2(A_{IID}, B)$ such that the distribution shift measure for an augment is $S_{aug} = \frac{W_2(A_{aug},B) - W_2(A_{IID},B)}{W_2(A_{copy},B) - W_2(A_{IID},B)}$, where A_{IID} uses the IID "augment" (i.e. just add an additional n data points such that A and B are both have $2n$ IID points), and A_{copy} uses the "copy" i.e. identity augment. Note that S_{aug} is a random variable and must be evaluated by choosing some n, and ideally sampling the datasets A and B multiple times to achieve an estimate with lower noise. If the desiderata are satisfied, then $S_{aug} \geq 0$ (if estimated exactly) and has units in terms of how far away the augmentation is from IID data compared to an uninformative, i.e. identity, augmentation.

Note as an extension of this, we may use different representation spaces to estimate distribution shift other than the raw/original representation, which affect both A and B. For example, in image datasets, we can evaluate the measure on pixel space, or use a model's latent space (e.g. CLIP).

2.4 Sanity Check on Gaussian Data + Augmentations

While we note above that we allow ourselves to make changes other than just the use of an existing metric + some cosmetic normalization (if it is helpful/necessary), it is all the better if we discover that an existing, common metric actually satisfies our desiderata, as it is both simple and immediately useful for understanding augmentations from the lens of distribution shift.

We run a basic experiment on multivariate Gaussian data to sanity check whether our proposed shift measure (effectively W_2 but normalized) satisfies these conditions, with positive results. To do this, we created simulated datasets of equal size, A and B, by sampling zero-mean Gaussian vectors, with each data point sampled IID from our multivariate Gaussian.

We choose three augmentations: new IID points (not really an augmentation), copy/identity, and resampling each coordinate with probability 0.5 (representing a helpful augmentation), forming the datasets A_{IID}, A_{copy}, and A_{coord} respectively, where as noted above A is formed half with original data points (here, sampled from a multivariate Gaussian with $d = 1000$) and half with augmented data (additional IID points if augmentation is IID). We ran this sampling scheme for 100 samples, yielding 95% confidence intervals for $W_2(A_{IID}, B), W_2(A_{copy}, B)$, and $W_2(A_{coord}, B)$ (we opt not to normalize these to show the values for the IID and identity augmentations). We have that $W_2(A_{IID}, B) = 1343.00 \pm 0.19, W_2(A_{copy}, B) = 1346.56 \pm 0.35, W_2(A_{coord}, B) = 1343.16 \pm 0.24$, satisfying the desiderata that the IID has the least distance, the uninformative augmentation (identity) has greater distance, and the helpful augmentation (near-IID) has distance very similar to IID, and significantly less than the uninformative augmentation. Note that while the differences are small relative to the scale (i.e. 3 vs. 1350), the differences are still large relative to the variations, which are surprisingly controlled (with $n = 100$ samples, roughly 3 vs. 0.3). We do not test the last desideratum but suspect this is trivial.

3 Experiments

3.1 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. Note that we compare each augmented dataset to a disjoint section of train data. 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, and take 10 samples. This allows more accurate estimations, as well as allowing the calculation of variance to obtain 95% confidence intervals for the estimated shift values.

3.1.1 Computation of Variances

Note that we estimate the standard deviations of our W_2 estimators (i.e. over 10 samples, the mean of $W_2(A_{aug}, B)$) as the standard deviation of the samples divided by $\sqrt{n-1}$, i.e. the unbiased estimator of the standard error. However, the measure is a function of random variables, namely one of the form $\frac{X-Z}{Y-Z}$. We assume that each of the Wasserstein distances are independent of each other, which is reasonable as each of the datasets have different IID samples, and thus we know that $Var[X - Z] = Var[X] + Var[Z]$ and similar for $Var[Y - Z]$. Then, we have from [5] that the first-order approximation of variance for the quotient of two random variables $\frac{A}{B}$ is $\frac{\mu_A^2 Var[B]+\mu_B^2 Var[A]}{\mu_B^4}$. Then the variance of our shift measure is $Var[\frac{X-Z}{Y-Z}] = \frac{\mu_{X-Z}^2 Var[Y-Z]+\mu_{Y-Z}^2 Var[X-Z]}{\mu_{Y-Z}^4} = \frac{\mu_{X-Z}^2 (Var[Y]+Var[Z])+\mu_{Y-Z}^2 Var[X]+Var[Z]}{\mu_{Y-Z}^4}$.

Then, for our 95% confidence intervals, we write $\frac{X-Z}{Y-Z} \pm 2\sqrt{Var[\frac{X-Z}{Y-Z}]}$.

Note that in general we could also compute the statistic $\frac{X-Z}{Y-Z}$ directly and then take the standard error, but we are repurposing existing data for which we only have the separate statistics for X, Y, Z (i.e. augmentations, IID, and copy mean and SE).

3.1.2 Note on Augmented Set with Non-Resampled IID

We received a thought-provoking suggestion during the TA evaluation at the poster session about using the same samples for the baseline B ($2n$ IID) and the IID portion of the augmented set A (n IID, or if the augment itself is additional IID, then $2n$ IID). Notably, this has the advantage of any metric being 0 for the IID augmentation for A (as A and B are the same datasets) and reducing the variance on the estimate in other cases (as we receive no shift from the IID portion of A). However, we realized after some thought that it violates desideratum 3 (small shift for good augmentations), so we decided in the end not to utilize this exact modification, and we stick with sampling a disjoint A and B for measuring distribution shift. However, factoring in ideas from both the above suggestion and the professor's feedback, we have added normalization to the distribution shift measure to improve interpretability.

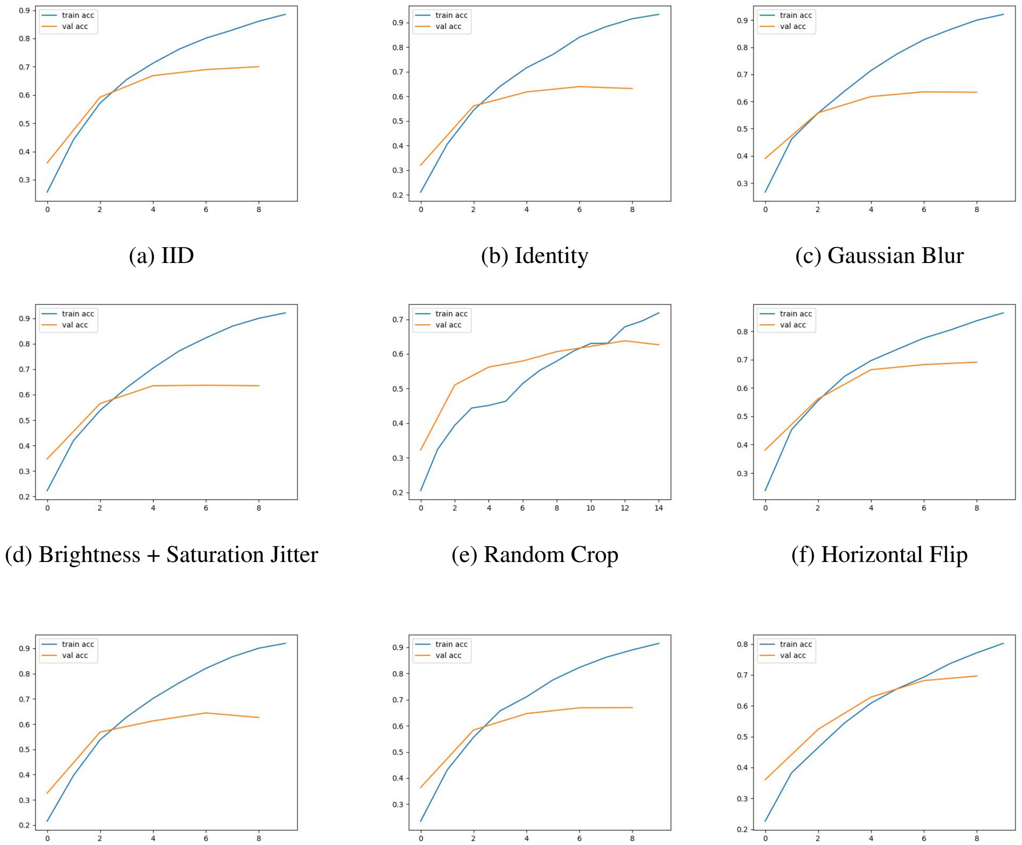
For the violation of desideratum 3, we have the following case. Let X and Y be sets of n points sampled IID from our underlying distribution. Under the proposed sampling scheme, our IID baseline (i.e. set A to $2n$ IID) would compare XY to XY, getting $W_2 = 0$. Our copy setting would compare XX to XY, and since we can match all points in A to themselves with 0 required distance, this divergence is at most $W_2(X, Y)$. Likewise, in a scenario where an augmentation creates near-IID points Z from existing IID points, we will get $A = XZ$, i.e. $W_2(XZ, XY) \approx W_2(XX, XY) = W_2(X, Y)$. Then note that instead of being close to 0 as is proper for a near-IID augmentation, this measure instead treats the augmentation as being effectively uninformative (same effect as copying the data).

3.2 Measuring Augmentation Validation Accuracies

After fine-tuning a ResNet-18 on augmented datasets, we obtained the following validation accuracies and epochs until the best validation accuracy, i.e. "early-stopping point" (measured for section 3.3.1). Training curves for the train and test sets are included for each augmentation in Figure 2.

Augmentation	Val. Accuracy	Epochs to convergence
IID Standard	70.02	8
Identity	63.93	6
Blur	63.58	6
Bright. + Sat.	63.58	6
Crop	63.78	12
Horizontal Flip	69.13	8
Grayscale	64.43	6
Rotate	66.94	8
Shift + Scale	69.62	8

Table 1: Validation accuracies of the various augmentations, and epochs to convergence.



(a) IID

(b) Identity

(c) Gaussian Blur

(d) Brightness + Saturation Jitter

(e) Random Crop

(f) Horizontal Flip

(g) Grayscale

(h) Rotate

(i) Shift + Scale

Figure 2: Training curves for each augment, with train and validation accuracies. Epochs are on the x-axis, and accuracy is on the y-axis.

3.3 Distribution Shift Calculations

We calculated the following distribution shift values between the augmented datasets and the original dataset, as described in the experiment settings section. Note that in both cases, the identity W_2 is greater than the IID W_2, which is required for our distribution shift measure to be well-defined (otherwise the normalization denominator makes everything negated).

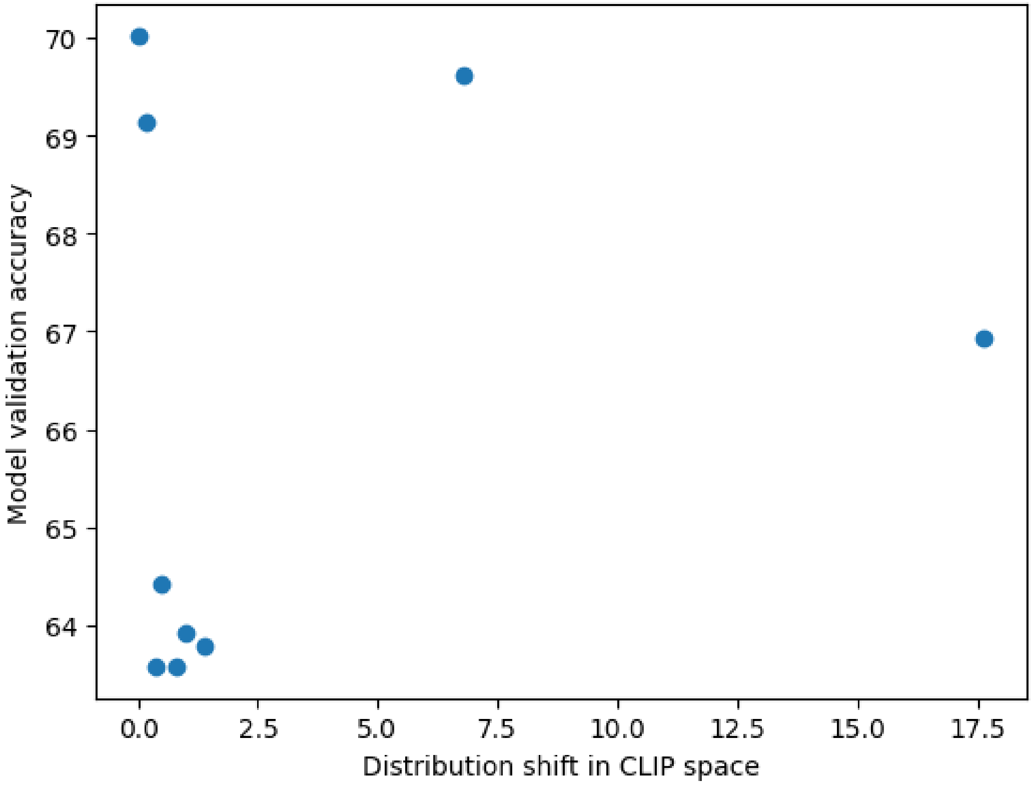


Figure 3: Validation accuracy vs. distribution shift in CLIP space

Augmentation	**CLIP Space S_{aug}**	**Pixel Space S_{aug}**
IID Standard	0 ± 0	0 ± 0
Identity	1 ± 0	1 ± 0
Blur	0.35 ± 0.45	-0.16 ± 0.20
Bright. + Sat.	0.81 ± 0.43	0.77 ± 0.32
Crop	1.39 ± 0.65	-0.88 ± 0.34
Horizontal Flip	0.19 ± 0.35	0.28 ± 0.25
Grayscale	0.48 ± 0.40	0.18 ± 0.24
Rotate	17.65 ± 5.75	5.05 ± 1.16
Shift + Scale	6.77 ± 2.25	5.42 ± 1.25

Table 2: Distribution shift values between the various augmentation distributions and a reference dataset (sampled IID from CIFAR-10).

In Figure 3, we plot the final accuracies of each model obtained in section 3.2, against the W2 distances obtained in section 3.3.

As expected, most of our augmentations yield W_2 distances between the IID standard and the identity baseline (in CLIP space; less so in pixel space), and correspondingly yield test accuracies between the two (or are similar to identity's accuracy). This yields a favorable result pointing towards the utility of the desiderata, as well as confirming W_2's ability to achieve the desiderata; Figure 3 shows a roughly inverse trend with the distribution shift measure and the final validation accuracy, ignoring the outliers (see below for discussion on these). One caveat is that the variations of the estimates are fairly high, though on expectation the predictive trend seems to hold (it is challenging to decrease the variation any more as it would take a computationally prohibitive number of samples at $n = 1000$, though perhaps this is possible to estimate with a smaller number of data points per sample).

We note that both blurring and cropping yield lower pixel-space W_2 values than entirely IID data, which breaks the first desideratum (IID is minimum shift) in pixel space; we conjecture that they create more "average" images in pixel space which W_2 prefers. This is addressed 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 seeing them as changes in semantic meaning, or a fundamental weakness in our metric's ability to characterize different augmentations. This seems to be the primary weakness in our approach.

3.3.1 Correcting a Mistake in the Poster

For completeness's sake, we remark that one observation which we included on the poster was that the distribution shift corresponds to the time required for the model to reach minimum validation error (i.e. "early stopping point"), but we now see that this trend disappears upon correcting the training setup.

In particular, rotate and shift/scale have the largest distribution shift values, and under the old training setup, took the most epochs to train until the early stopping point. However, we note that the old training setup deviated from this one in that the entire training set was based on augmentations (no original data) and at each epoch all augmentations were randomly resampled (the size of the augmented set is closer to en where e is the number of epochs), and we had erroneously used the augmented data for the test set as well. Under the corrected training setup, the augmentation taking the longest time until the early stopping point is crop, which does have a relatively larger shift value but is still small compared to rotate and shift/scale. Furthermore, rotate and shift/scale do not take longer than horizontal flip and IID standard, which don't have large shifts.

4 Discussion

In conclusion, we propose the perspective of analyzing the effectiveness of data augmentation through the lens of distribution shift, after noting the peculiarity of voluntarily inducing distribution shift (for the benefit of increasing the number of, hopefully, IID data points). We analyze distributional distance metrics and choose Wasserstein distance (in particular W_2) according to certain heuristics, and state desiderata for a measure of distribution shift which can usefully represent the effectiveness of augmentations. Through sanity checks on a synthetic Gaussian dataset, we determine that W_2, with some normalization and possibly comparing distributions in a latent space, is a reasonable candidate.

We run experiments on CIFAR-10 to determine the value of the distribution shift measure for different augmentations and across different latent spaces, comparing it to the actual effectiveness of the augmentation (measured by validation accuracy when trained on the augmented dataset), and conclude that certain desiderata do roughly hold (especially in CLIP space, less so in pixel space). However, in particular, we are incorrectly penalizing augmentations that are useful but deviate from the input distribution (such as rotation), and it could also be helpful to find ways to further penalize uninformative augmentations.

As the primary contributions, we propose the analysis of data augmentations through distribution shift, which may be promising for future work, and specify desiderata for what eventually. We explore one candidate which succeeds with some predictive power on CLIP space, and identify a major flaw which can be fixed (see below) to yield an improved and more predictive measure.

4.1 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] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, October 2021.

[2] 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.

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

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

[5] Douglas Zare. Variance of x/y (https://stats.stackexchange.com/questions/32659/variance-of-x-y), 2012.