Untapped Potential of Data Augmentation: A Domain Generalization Viewpoint

Anonymous Authors[1]

Abstract

Data augmentation is a popular pre-processing trick to improve generalization accuracy. It is believed that by processing distorted inputs in tandem with the original instances, the model learns a more robust set of features which are shared across the inputs. However, it is unclear if that is the case. In this work, we take a Domain Generalization viewpoint of augmentation based methods. This new perspective allowed for probing overfit and delineating avenues for improvement. Our exploration with state-of-art augmentation method provides evidence that the learned representations are not as robust even towards distortions used during training. This suggests evidence for the untapped potential of augmented examples.

1. Introduction

Contemporary learning algorithms demonstrate strong performance, even surpassing humans at times, when training and testing under independent and identically distributed (iid) assumption. Notwithstanding performance under iid settings, they are far from human level robustness when evaluated under data shifts (Geirhos et al., 2018; Hendrycks & Dietterich, 2019; Mu & Gilmer, 2019). This problem is of central focus in learning distributionally robust models Hendrycks et al. (2019). While the related problem of robustness to imperceptible adversarial examples has received much larger interest Chakraborty et al. (2018); there has been an increasing push toward expanding the definition of robustness to include naturally occurring corruptions (Engstrom et al., 2019). This is especially so because best defenses against, the narrow focused, adversarial examples does much worse with robustness to natural corruptions.

There is a growing interest in building systems with better out-of-domain generalization performance also called Distributional Robustness (DR). DR is pursued under various fronts: (a) imposing inductive biases that penalize spurious correlations more (Wang et al., 2019) (b) through employing augmentation or optimization techniques to weed out known data-overfitting features of the dataset (Sagawa* et al., 2020; Geirhos et al., 2019) (c) general dataset-agnostic data augmentation (Hendrycks et al., 2019; Rusak et al., 2020). Arguably, general augmentation methods are more scalable. Robustness through general augmentation is the problem of our interest in this work.

Due to different manifestations of the DR problem, the research in this direction is somewhat fragmented. The objective of robustness to domain shifts is the common theme in domain generalization (Shankar et al., 2018; Carlucci et al., 2019; Piratla et al., 2020), distributional robustness (Hendrycks et al., 2019; Rusak et al., 2020), identifying and mitigating dataset biases (Gururangan et al., 2018). DR is a general version of the Domain Generalization (DG) problem. DG functions under the setting where the train data is drawn from multiple sources along with annotation of source id for each example with the objective of better generalization to unseen domains. Difference between DG and DR are superficial, the following assumptions of former are relaxed in the latter (1) the train data does not necessarily be pooled from multiple sources (2) annotation of the domain label per example could be missing. In this work, we borrow lessons from the DG line of work to emphasize the untapped potential of augmentations directed at improving distribution robustness.

Data augmentation technique is widely adopted for image preprocessing and has recently been shown to improve out-of-domain robustness (Hendrycks et al., 2019). It is, however, unclear how the augmented examples interact with the clean examples. Training under data augmentation resembles multi-source training of DG. An ideal DG algorithm exploits the train time domain variation so as to learn a hypothesis that is better equipped at generalizing to new domains. The Expected Risk Minimization (ERM) baseline on the other hand does not attend to the domain boundaries and yields bad domain-shift robustness owing to overfit on the seen domains. We want to draw attention to the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Do not distribute.

under-explored utility of domain generalization methods for better robustness. The prime focus of this work is to explore augmentation techniques in the context of out-of-domain robustness. Although some of our claims may also carry to generalization error, it is beyond the scope of this work.

In our exposition, we use domain, augmentation and input distribution interchangeably. Data augmentations are drawn from label consistent transformations of the clean examples, the introduced data-shift in the train data from augmentations is no different from what is usually referred to as the domain in the DG literature.

We make the following contributions.

- **Untapped potential**: We show that the standard augmentation (including state-of-art) methods under-utilize augmented examples by over-fitting on them.
- **Future direction**: We note that there is a broad scope for improving augmentations for even better robustness and conclude with a discussion of future line of work that exploit the observed patterns of overfit.

2. Untapped Potential of Augmentations

In this section, we provide evidence of augmentation overfit by systematically exploring a recent state-of-art augmentation method. We employ generalization measures from domain generalization literature to investigate overfit and to identify possible avenues for improvement. As a case study, we use models trained with *AugMix* (Hendrycks et al., 2019) when trained on CIFAR-10, CIFAR-100 and ImageNet across different network architectures.

Augmentation is a standard trick employed to ameliorate over-fitting, dominantly in image applications. In the extreme case of catastrophic over-fitting, the augmented examples cannot help generalization of the original examples. On the other extreme, in the ideal scenario, we expect the algorithm to draw what is common between the clean and augmented examples without having to employ any specific features for either clean or augmented data. Vanilla augmentation need not lead to the ideal scenario of learning common features between clean and augmented inputs. For example, Vasiljevic et al. (2016) report that train-time blur augmentations do not generalize to unseen blurs. Furthermore multiple DG studies (Motiian et al., 2017; Ghifary et al., 2015a) show that train data containing instances under multiple rotations does not generalize to unseen rotations. In practice, algorithms fall in between the two extremes of catastrophic over-fitting and perfect parameter sharing, demonstrating domain overfit of various degree.

We pose the question on how much feature sharing occurs between the clean and augmented examples with AugMix. We probe domain overfit using measures borrowed from the DG literature. In section 2.1, we probe how domain invariant are the representations obtained from various layers. Section 2.2 employs a recent common-specific decomposition strategy proposed in Piratla et al. (2020) to identify any overfitting components in the model weights. Finally in section 2.3, we make a more controlled evaluation of the generalization to augmentations of varying severity levels.

2.1. Domain Divergence Measure

Domain overfit can be qualitatively measured by looking at how transferable the parameters are between the train domains. The seminal paper on domain adaptation: Ben-David et al. (2006), proved an upper bound on generalization gap between any two domains in terms of a divergence measure between them. Equation 1 provides this measure for a given hypothesis class $\mathcal{H}$ and source and target distributions: $\mathcal{S}, \mathcal{T}$ with their respective populations: $n, n\prime$.

$$d_{\mathcal{H}}(\mathcal{S},\mathcal{T}) = 2(1 - min_{\eta\in\mathcal{H}}\{\frac{1}{n}\Sigma_{i=1}^{n} I[\eta(x_i)=0] + \frac{1}{n\prime}\Sigma_{i=n+1}^{n+n\prime} I[\eta(x_i)=1]\}) \quad (1)$$

Intuitively, the domain divergence would be low when the hypothesis class induced by the learned representations do not allow for domain prediction i.e. the representations should be domain invariant. Since it is hard to compute the divergence measure exactly, a proxy measure, accuracy of a trained discriminator proposed in Ganin et al. (2016), is adopted. We train a domain discriminator to discriminate augmented examples from clean examples. Higher the accuracy of the domain discriminator, greater is the scope of domain overfit.

We probe for domain invariance of the representation learned by AugMix on CIFAR and ImageNet datasets. We used representations from two different layers: the penultimate and antepenultimate layers, penultimate layer is the layer before the softmax layer. We use the representations obtained from these layers for clean (x_c) and augmented (x_a) examples along with their domain assignment: $\bigcup_i \{x_{ci}, 0\} \cup \{x_{si}, 1\}$ as the input data. A linear discriminator is then train on 40,000 examples with equal proportion of clean and augmented images. If the model learns generalizable common features, then information related to the augmentation's distortion should be minimal. On the other hand if the model relies on domain specific feature, that information will be present in the representation layers of the model. The same information can be used to correctly identify the domain of the input sample. As such the higher is the accuracy of a discriminator which can distinguish samples from the domains, the greater is the reliance of the model on non-robust domain specific features.

Table 1 shows the discriminator's performance for a range of models trained with AugMix. We also report discrimination accuracy on unseen test examples that are similarly collected as train. Since the discriminator is a simple linear classifier, any domain predictive capacity even under such favorable conditions is indicative of a strong overfit. Note how the domain predictive information is erased only in the penultimate layer. The prevalence of domain identifying information up until this layer is indicative of shallow parameter sharing between augmentations and clean examples. This highlights the need for measures that promote higher parameter sharing between augmentations and original instances.

2.2. Common vs Specialized Components of the Classifier

When training on multi-domain data, we desire to retain only the components of the classifier that rely on common features. To better understand the intuition, consider the illustrative example of water-bird vs land-bird classification (Sagawa* et al., 2020). The train data can be partitioned in to four groups: water-birds with water and land background, land birds with water and land background. We make the practical assumption that the train data is dominated with water-birds (land-birds) with water (land) background. Since it is arguably simpler to classify based on the background, the standard (such as ERM) algorithms pick on features (which is background here) that may not generalize to minority groups: water-birds (land-birds) on land (water). Piratla et al. (2020) also argues that the domain-specific components of the classifier as a cause for bad out-of-domain generalization error. More interestingly, their proposed solution of weeding out the domain specific component readily translated to better out-of-distribution robustness.

In order to study if AugMix suffers of a similar problem from domain specific components, we employ the common-specific decomposition on the classification parameters We obtain penultimate layer representations for a randomly sampled 20,000 examples of original and augmented images each. We then obtain optimal linear content-label. classifier individually for clean and augmented instances. These are denoted as w_o^*, w_a^* respectively. We are interested in decomposing these parameters in to a linear combination of common (w_c) and domain-varying (w_s) component accompanied by domain-specific combination parameter (γ_a, γ_s). This requires solving the following constrained problem shown in Equation block 2 [1]. .

[1]See theorem 1 of Piratla et al. (2020) for the decomposition algorithm

$$\begin{aligned} w_o^* &= w_c + \gamma_o w_s \\ w_a^* &= w_c + \gamma_a w_s \\ w_c &\perp w_s \end{aligned} \tag{2}$$

Note from the decomposition problem that (1) contribution of the common component w_c to each of w_o^*, w_a^* is the same, and (2) the contribution of specific component w_s varies. In the ideal case when the representation contains only features of consistent label correlations between domains, then the domain specific components ($\gamma_a w_s, \gamma_c w_s$) are diminutive compared to the common component (w_c). On the other hand when the representations contain features that favour only one of the two domains, it manifests in strong domain specific components.

In Table 2, we report the ratio of norms of specific and common components over a range of models trained with AugMix, expression for the reported measure shown below:

$$\frac{\|[\gamma_o w_s, \gamma_a w_s]\|}{\|[w_c, w_c]\|}$$

In the ideal case the ratio is expected to be very close to zero as tn specific components are negligible. However, for a range of AugMix trained models, the ratio is close to one implying that the specific components dominate. This strongly suggests scope for better robustness adding further to the case of untapped potential of augmentations.

2.3. Controlled Evaluation of Distributional Robustness

In this section, inspired from Geirhos et al. (2018); Vasiljevic et al. (2016), we make a controlled evaluation of the AugMix trained models in order to objectively measure domain sensitivity. AugMix allows for several knobs on the train time augmentations; Of our particular interest are (1) *mixing coefficient* that combines the augmented example with the original example (2) *severity level* of distortions for input transformation. We make a more modest evaluation on the test set using only the seen distortions but with differing severity and with or without mixing with clean examples.

Table 3 summarizes our findings. Without mixing means we evaluate on the augmented example directly. AugMix draws several samples from the convex combination of clean and distorted examples, and thereby we expect generalization to any convex combination of clean and augmented examples including either extremes. However, it is surprising that we found consistent drop in accuracy with the default severity level of 3 and when evaluated on an endpoint: distorted input. Also, we draw attention to the drop in accuracy when using severity level of 5 just outside of the train time value

	CIFAR-10		CIFAR-100		ImageNet
Layer \ Arch	AC	WRN	AC	WRN	ResNet-50
PL	50.2 (52.8)	51.9 (52.3)	52.3 (52.1)	51.0 (50.8)	54.5 (57.4)
APL	100 (100)	85.5 (91.8)	100 (100)	84.8 (86.6)	76.8 (84.0)

Table 1. Test and train domain discrimination accuracy (train accuracy shown in brackets) on CIFAR-10, CIFAR-100 and ImageNet. PL and APL stands for penultimate and antepenultimate layers. AC and WRN expand to AllConv and WideResNet respectively.

Dataset \ Arch	AC	WRN
CIFAR-10	0.6	3.3
CIFAR-100	0.8	1.4

Table 2. Ratio of norm of specific components to common components, smaller the better, for CIFAR-10, CIFAR-100 with AllConv (AC) and WideResNet (WRN) architecture.

of 3 [2]. These observations highlight the fragile robustness of AugMix.

	CIFAR-100		CIFAR-10	
	Mix	wo Mix	Mix	wo Mix
s=0	71.2		92.6	
s=3	69.9 (0.1)	65.4 (0.3)	91.8 (0.1)	88.9 (0.1)
s=5	66.8 (0.3)	61.4 (0.4)	90.1 (0.2)	87 (0.1)

Table 3. Classification accuracy of AugMix trained on CIFAR-100 and CIFAR-10 when evaluated on seen distortions of varying severity level (rows) and with or without mixing with clean example.

3. Related Work

Domain Generalization Domain generalization refers to zero-shot adaptation to examples from unseen new domains. Building on Ben-David et al. (2006) insight; a plethora of methods based on minimizing some form of domain divergence have been proposed (Ganin et al., 2016; Ghifary et al., 2015b). Other methods for domain generalization include parameter decomposition (Khosla et al., 2012; Piratla et al., 2020), domain adversarial augmentation (Shankar et al., 2018) and meta-learning (Balaji et al., 2018; Li et al., 2018).

Data Augmentation Researchers have developed various techniques to create the augmented data samples. These include random erasures (Zhong et al., 2020), random replacement (Takahashi et al., 2018), noise patching (Lopes et al., 2019). and image interpolation (Tokozume et al., 2018). Both the works of Madry et al. (2018) and Shankar et al. (2018) are versions of creating augmented examples using input gradient. Xie et al. (2019) employed data augmentation in semi-supervised teacher student framework.

[2]Augmix severity scale linearly from 0 to 10

4. Discussion

Vanilla training combining clean and distorted inputs are not necessarily enough to ensure robustness. Deep neural networks can learn unexpected properties from the the training distortions and overfit on them (Geirhos et al., 2019). (Hendrycks et al., 2019) proposed that by using stochastic methods to create a non-fixed number of distortions will mitigate the issue, by forcing the model to learn robust features.

Our experiments suggest that this mitigation is partial and overfitting is still an issue. The layer activations retain a significant information about whether the input is a clean or distorted input. This suggests that the model is not necessarily robust to out of domain distortion. This is highlighted by the deterioration of model performance on augmented inputs generated from slightly different distortion sampling parameters. Furthermore the model retains sensitivity to training parameters. The mixing operation used in Augmix would lead one to expect that the model is robust on the simplex between clean data and its augmentations. However contrary to expectations even on the training augmentations, one sees significant difference between different mixing patterns. The fact that the models are not as robust as believed, suggests there is still significant scope of improvement from the way augmentations are currently utilized.

We envision a future line of work targeting the overfit patterns which have been observed in this work to be of value. The presence of common and specific components in the representations can be mitigated by adopting methods from Piratla et al. (2020); Sanyal et al. (2020). Parameter sharing can be further promoted through a study of domain invariant networks (Ganin et al., 2016).

References

Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pp. 1006–1016, 2018.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, 2006. URL http://dl.acm.org/citation.cfm?id=2976456.2976474.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Engstrom, L., Gilmer, J., Goh, G., Hendrycks, D., Ilyas, A., Madry, A., Nakano, R., Nakkiran, P., Santurkar, S., Tran, B., Tsipras, D., and Wallace, E. A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, 2019. doi: 10.23915/distill.00019. https://distill.pub/2019/advex-bugs-discussion.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 7538–7550, 2018.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Ghifary, M., Bastiaan Kleijn, W., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015a.

Ghifary, M., Bastiaan Kleijn, W., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pp. 2551–2559, 2015b.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://www.aclweb.org/anthology/N18-2017.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A., and Torralba, A. Undoing the damage of dataset bias. In *ECCV*, pp. 158–171, 2012.

Li, D., Yang, Y., Song, Y., and Hospedales, T. M. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pp. 3490–3497, 2018.

Lopes, R. G., Yin, D., Poole, B., Gilmer, J., and Cubuk, E. D. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *CoRR*, abs/1906.02611, 2019. URL http://arxiv.org/abs/1906.02611.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

Mu, N. and Gilmer, J. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

Piratla, V., Netrapalli, P., and Sarawagi, S. Efficient domain generalization via common-specific low-rank decomposition. *arXiv preprint arXiv:2003.12815*, 2020.

Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. Increasing the robustness of dnns against image corruptions by playing the game of noise. *arXiv preprint arXiv:2001.06057*, 2020.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sanyal, A., Torr, P. H., and Dokania, P. K. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1enKkrFDB.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

Takahashi, R., Matsubara, T., and Uehara, K. Data augmentation using random image cropping and patching for deep cnns. *CoRR*, abs/1811.09030, 2018. URL http://arxiv.org/abs/1811.09030.

Tokozume, Y., Ushiku, Y., and Harada, T. Between-class learning for image classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5486–5494. IEEE Computer Society, 2018.

Vasiljevic, I., Chakrabarti, A., and Shakhnarovich, G. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xie, Q., Hovy, E. H., Luong, M., and Le, Q. V. Self-training with noisy student improves imagenet classification. *CoRR*, abs/1911.04252, 2019.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pp. 13001–13008. AAAI Press, 2020.