Boundary-aware Augmentation for Object Detection in Scientific Images (DRAFT)

Benjamin Killeen and Gordon Kindlmann

***Abstract*—Recently, deep convolutional neural networks (DCNNs) have enabled remarkable progress in computer vision tasks. Given this success, we consider the application of DCNNs to image analysis in scientific experiments, which traditionally employ time-consuming, *ad hoc* solutions. DCNNs have the potential to accelerate this analysis through object detection, but the training process for DCNNs requires many labeled images. Crucially, the experiment must produce its own training set, and this circularity results in over-fitting which violates the scientific method. Unlike object detection for real-world tasks, scientific image analysis must be agnostic to non-uniformities in the natural world, such as natural laws or governing principles. In this report, we introduce boundary-aware augmentation (BAA) as an essential component of training DCNNs which agnostic to governing principles yet perform reliably on scarce data.**[1]

I. Introduction

Image analysis is a vital component of many scientific experiments, used to measure properties like object location or orientation. In cases where no alternative measuring technique exists, accurate analysis is of the utmost importance, but this accuracy often proves labor-intensive. Experimenters utilize *ad hoc* solutions suited for one experimental setup, even though recent advances in Computer Vision have yielded solutions for seemingly more general problems.

Many factors have contributed to the recent success of DCNNs. In 2012, Krizhevsky et al. trained a DCNN on the ImageNet dataset, achieving previously unseen classification accuracy, and showed that state-of-the-art GPUs could accelerate computation for these kinds of models [1], [2]. Further advances in computer power continue to motivate refinements to DCNN design, yielding even higher classification accuracies [3], [4], [5]. Meanwhile, datasets with more detailed labels, such as Microsoft COCO, facilitate supervised training for tasks like bounding-box detection and instance segmentation [6]. If sufficient training data exists, DCNNs seem ripe to dramatically accelerate scientific image analysis.

In general, however, each experiment must produce its own training data. Ronneberger et al. confront this scarcity for biomedical image analysis through data augmentation, previously introduced for vision tasks in general, and network design, introducing a novel DCNN architecture for semantic segmentation [1], [7]. Another approach to high-cost labeling is active learning, where a query strategy selects unlabeled examples which will most benefit training [8]. Active learning reduces the number of examples—and thus the man-hours—required for training, making it a vital part of any workflow with DCNNs for scientific image analysis. For data scarcity in general, these two approaches, active learning and simple data augmentation, are sufficient, but data from scientific experiments exhibits not just scarcity but also undesirable non-uniformity, resulting from general principles.

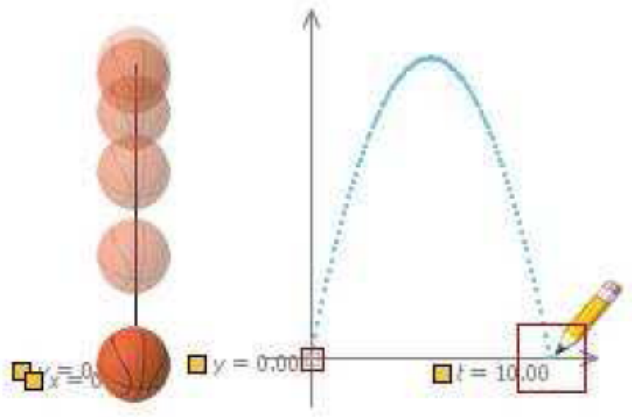


Fig. 1. An object in free fall [9]. (PLACEHOLDER IMAGE)

In order to illustrate non-uniformity from general principles, we consider the concrete example shown in Figure 1. Here, the experimenter wishes to study how objects fall, and to do this, he or she has captured images at regular intervals of a basketball in free-fall. Because objects accelerate downward at a constant rate of approximately 9.8 $\mathrm{m/s^2}$ [10], her data will consist of many more examples with the basketball in the top half of the image than the bottom. This uneven distribution comprises a "non-uniformity," and the natural law it results from is a *governing principle*. As a matter of course, the experimenter may have hypothesized that this principle exists, but according to the scientific method, should not incorporate her hypothesis into her measurement. This complication presents a challenge for training a DCNN—or any predictive model, for that matter—with data from the experiment itself. The distribution of y, shown in Figure 1, should not influence future measurements of y.

Unfortunately, a DCNN can incorporate that distribution into its parameters during learning. This makes it more likely to predict images with the basketball in the top half of the image than the bottom. It could also fail to capture any anomalies

Benjamin Killeen is an undergraduate with the Department of Computer Science, University of Chicago, Chicago, IL 60637, USA (email: killeen@uchicago.edu)

Gordon Kindlmann is with the Department of Computer Science, University of Chicago, Chicago, IL 60637, USA (email: glk@uchicago.edu)

[1]Code available at github.com/bendkill/artifice.

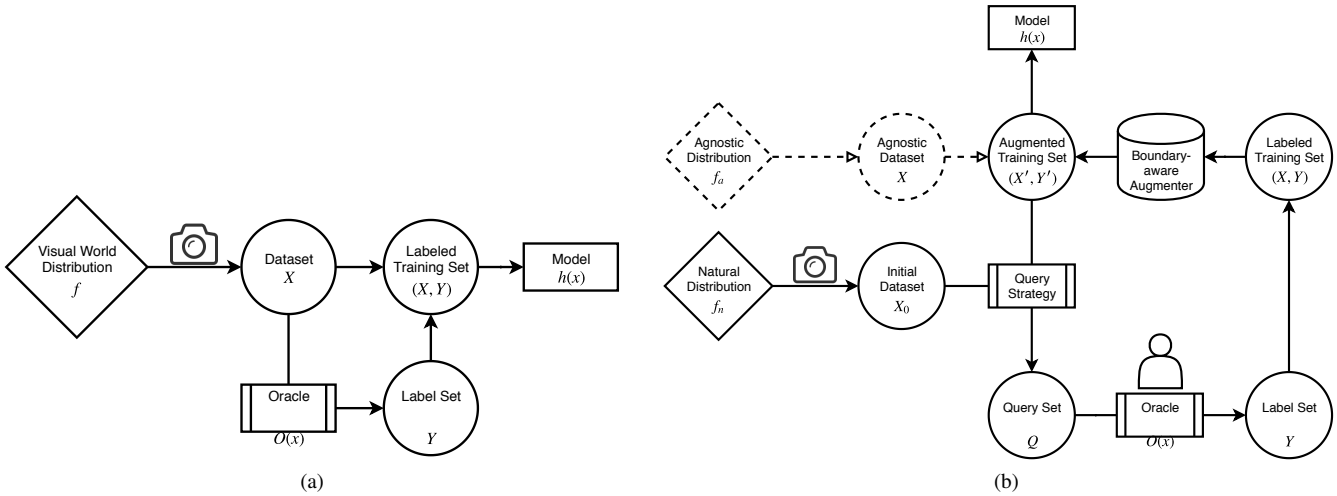


Fig. 2. During supervised learning for real-world vision tasks **(a)**, a camera samples a large dataset X from the "visual world" distribution f. For scientific tasks **(b)**, an experiment samples a small initial dataset X_0 from the natural distribution f_n, which includes the effects of any governing principles. Our proposed training method incorporates a boundary-aware augmenter A, which uses the training set (X, Y) to simulate drawing examples from the agnostic distribution f_a. Dashed lines indicate theoretical or simulated objects.

in the basketball's trajectory, if gravity didn't behave the way we expect. Ideally, therefore, we would sample the training set from a universe—so to speak—where gravity doesn't affect the basketball's position. The question immediately arises: how does the ball behave in this universe? Or rather, how *should* the ball behave? Because we wish to create a DCNN for object detection, we should sample position as uniformly as possible.

To formalize these ideas, we contrast between traditional and scientific vision tasks. Figure 2a outlines the training process for supervised learning. It begins with a large, ordered dataset $X \subseteq mathbbR^{M \times N}$ of $M \times N$ images, sampled from the "visual world." Next, an oracle $O : X \subseteq mathbbR^{M \times N} \to \mathbb{R}^n$ (usually a human) provides the label set $Y = \{O(x) : x \in X\}$, where n is the dimensionality of the labels, *e.g.* number of classes in a one-hot encoding. In any task, this label is a lower-dimensional representation of the data space, encoding valuable information from each example. Together, the data and label sets (X, Y) comprise the training set. Supervised learning aims to train a model (such as a DCNN) $h : \mathbb{R}^{M \times N} \to \mathbb{R}^n$ that can interpret the visual world. Crucial, the visual world is not a uniformly distributed set of images; we denote its probability distribution as $f : \mathbb{R}^{M \times N} \to \mathbb{R}$, from which every dataset draws examples as independently and identically distributed as possible (or attempts to).

For real-world tasks, sampling f in this manner is desirable, but in scientific tasks, f contains the effects of governing principles. Instead, we wish to sample a dataset X that correspond to labels Y as uniform as possible—within reasonable boundaries—in the label space $\mathbb{R}^n$. Incidentally, most scientific experiments have well-known *label-space boundaries* incorporated into their design. A ball on a ramp, for instance, is bound to one dimension, even though its image-space position is described by two coordinates. Figure 3a shows an experiment where dot markers are confined to a small circular region. The region $\psi \subseteq \mathbb{R}^n$ that these boundaries define corresponds to a higher-dimensional data-space region $\chi \subseteq \mathbb{R}^{M \times N}$ such that the real distribution f obeys

$$(\forall x \notin \chi)(f_n(x) = 0).$$

Note that χ is not necessarily the smallest such space; the experimenter simply determines reasonable boundaries for ψ such that the above condition is satisfied. Even with perfect knowledge of ψ, it would be impossible to recover χ. Nevertheless, we will attempt to do just that.

Although χ is largely a theoretical construct, it serves as a useful concept for training models which are agnostic to governing principles. Consider the uniform distribution across it, which we will refer to as the *agnostic distribution*:

$$f_a(x) = \begin{cases} \frac{1}{\mathrm{Vol}(\chi)} & x \in \chi \\ 0 & x \notin \chi \end{cases}$$

From the initial dataset X_0, then, we wish to generate and label an augmented dataset X that approximates an agnostic dataset $X_a \sim f_a$. For this purpose, we introduce *boundary-aware augmentation* (BAA), which utilizes the label-space region ψ and instance segmentations to approach X_a. We further denote the label-space uniform distribution

$$g_a(y) = \begin{cases} \frac{1}{\mathrm{Vol}(\psi)} & y \in \psi \\ 0 & y \notin \psi, \end{cases}$$

the lower-dimensional representation of f_a.

II. Method

With practical application in mind, we outline our method as it relates to an active learning strategy with an on-line oracle, *i.e.* a human. The query strategy is a vital part of any system utilizing our method, but its final form will depend on the exact details of our implementation. We rely on existing literature for inspiration, but for the time being, we may consider a query strategy which selects unlabeled images either at random or according to uncertainty [8], [11].

The function of our augmentation scheme, BAA, is more fundamental. Like any data augmentation scheme, it aims to mitigate the effects of data scarcity, which [1] achieves through image-global transformations such as flipping and brightness shifts. In order to address governing principles, however, our augmentation scheme must be able to generate examples x from arbitrary points y in label space $\mathbb{R}^n$. In principle, the label space can include a large number of object properties which the experimenter wishes to measure, including position in the image, apparent orientation, size, and any number of one-hot classifications. These apply to every object in the image, resulting in a label space that can grow relatively large, with independent boundaries on ψ for every coordinate. For the following explanation, we consider a label space of object positions, but maintain that the same principles apply for higher-dimensional label spaces.

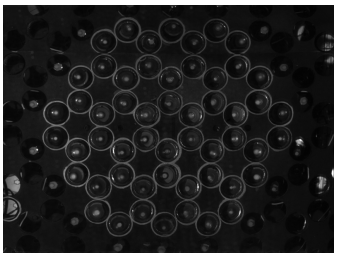
In order to freely manipulate examples in label-space, we introduce an intermediate *annotation space* $\mathbb{R}^{M\times N\times k}$ between the data and the labels, where k is the dimensionality of the annotation. Although the specific details of the annotation space are still under consideration, it includes an instance segmentation of the image x. With this information, BAA can extract the pixels belonging to every object in the image and translate them freely. Unlike image-global strategies, this method directly addresses the label-space representation of an example. It also raises two questions: what pixels should the augmenter use to replace the extracted object, and what new points in label-space should the augmenter introduce? The first question is a matter of practical importance, but the second directly relates to our primary goal.

There are many possible solutions to the problem of pixel-replacement. Most simply, one could use the mean value of the surrounding region, or else gaussian noise with the same mean and standard deviation. [12] describes a more nuanced approach that attempts to complete isophote lines arriving at the region's edge. [13] introduces Context Encoders: DCNNs that incorporate the entire image to inpaint a desired region. Any of these methods should prove effective for our purposes, although in many cases they may prove unnecessary. Many scientific tasks are the result of fixed-camera video data. If another example in the labeled dataset includes background pixels from the desired region, then the most effective approach would simply "transplant" these pixels, so to speak, from that example.

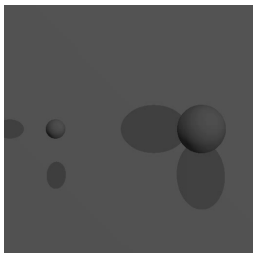
Once an object-region has been extracted, the question of placement depends on how uniformly Y covers ψ. By placing new examples properly (and potentially removing old ones) BAA should generate X' that more uniformly covers χ. This requires a comparison of the multivariate sampling Y with the multivariate distribution g_a, so we denote similarity function S to perform this comparison. For the time being, the choice of S is unclear, but possible choices include a coordinate-wise Kolmogorov-Smirnov test. Although this is not a general solution in the multivariate case, it is perfectly acceptable when ψ is bounded by affine subspaces. Additionally, Justel et al. describe a variation of the K-S test which can be efficiently approximated in high dimensions [14]. In any case, we wish to generate labels Y' such that

$$Y' = \underset{Y'|Y\subseteq Y'}{\operatorname{argmax}} S(Y', g_a).$$

After placing extracted objects, this yields an augmented dataset approximating $X_a \sim f_a$.



(a) Gyroscopic Model



(b) Coupled Spheres

Fig. 3. Example images from scientific experiments: **(a)** gyroscopic model for topological metamaterials [15]. Each dot is bound inside the circle surrounding it. **(b)** still frame from a simulated experiment of two spheres coupled by an invisible spring. Full video here.

III. Evaluation

In order to show our method's effectiveness, we develop several virtual experiments. These simulations offer several advantages over images from real experiments, such as in Figure 3a. First, we have perfect knowledge of the experiment's "Truth" $\tilde{Y}$, as opposed to imperfect measurements, or "ground truth," Y. In well established datasets, $\tilde{Y}$ and Y are nearly identical, but we must rely on one or a few human labelers for what should be unambiguous quantities. For testing, we calculate $\tilde{Y}$ from the known parameters of the simulation, and we emulate a human labeler by introducing small perturbations to $\tilde{Y}$, producing Y. Part of our goal is to train a DCNN that more closely predicts $\tilde{Y}$ than the labels Y. Simulated experiments allow us to test this performance.

Figure 3b shows one such experiment. In this case, two spheres with different masses rotate in free space, coupled by an invisible spring. The goal of the Coupled Spheres experiment is to recover physical properties of the spring using (x, y) positions of the two spheres. These physical properties comprise the general principles underlying the dataset. To evaluate our method's agnosticism toward general principles, we intend to train a DCNN on one experiment and evaluate its performance on experiments with different simulated springs. This simple example illustrates the general resilience that we wish to develop.

IV. Discussion

We have illustrated the problem of applying predictive models like DCNNs to scientific images without proper care. We have proposed a method, boundary-aware augmentation, that addresses this issue by considering a uniform distribution over the data region χ. Although this method is (in theory) successful in created an augmented dataset which is agnostic to general principles, the resulting dataset X is also agnostic to any other non-uniformities in X_0. We have also not addressed the reasonable constraint of temporal continuity. Instead, we

chose to consider isolated frames with an *arbitrary* rather than *temporal* ordering, but we could consider instead the uniform distribution of paths that an object could take from one frame to the next, within some reasonable maximum.

ACKNOWLEDGMENT

Thanks to Michael Maire for input and guidance, as well as William Irvine for access to image data from his laboratory.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," p. 8.

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sep. 2014, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *arXiv:1409.4842 [cs]*, Sep. 2014, arXiv: 1409.4842. [Online]. Available: http://arxiv.org/abs/1409.4842

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft COCO: Common Objects in Context," *arXiv:1405.0312 [cs]*, May 2014, arXiv: 1405.0312. [Online]. Available: http://arxiv.org/abs/1405.0312

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597

[8] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, Jun. 2012. [Online]. Available: https://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018

[9] Duquevieira, "Simple basket ball fall," Jan. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:Fall.png

[10] R. Munroe, "Wikipedian Protester." [Online]. Available: https://xkcd.com/285/

[11] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3162–3169.

[12] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image Inpainting," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424. [Online]. Available: http://dx.doi.org/10.1145/344779.344972

[13] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2536–2544. [Online]. Available: http://ieeexplore.ieee.org/document/7780647/

[14] A. Justel, D. Pea, and R. Zamar, "A multivariate Kolmogorov-Smirnov test of goodness of fit," *Statistics & Probability Letters*, vol. 35, no. 3, pp. 251–259, Oct. 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167715297000205

[15] L. M. Nash, D. Kleckner, A. Read, V. Vitelli, A. M. Turner, and W. T. M. Irvine, "Topological mechanics of gyroscopic metamaterials," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14 495–14 500, 2015. [Online]. Available: http://www.pnas.org/content/112/47/14495