Object Detection in Scientific Images (DRAFT)

Benjamin Killeen and Gordon Kindlmann

***Abstract*—Together with large training sets, Deep Neural Networks (DNNs) have enabled a wide range of advancements in Computer Vision tasks, especially with regard to object detection in real-world images. Object detection for scientific images, on the other hand, presents challenges unlike typical vision tasks. Scientific tasks exhibit *governing principles* in their datasets, which result from the phenomena under investigation. In this progress report, we describe our ongoing effort toward training DNNs which are agnostic to governing principles yet perform reliably on scarce data.**[1]

I. Introduction

Imaging systems are a vital component of many scientific experiments, used as the primary measuring device of properties like object location or orientation. In cases where no alternative measuring device exists, accurate analysis is of the utmost importance. However, such image analysis can prove labor-intensive. Traditional methods involve *ad hoc* solutions suited for one experimental setup even though more general object detection tasks have been well-studied in Computer Vision. In this report, we introduce an ongoing effort to generalize principles for image-data analysis across scientific tasks using DNNs, emphasizing the importance of careful data augmentation.

Deep Neural Networks[2] have shown remarkable success on real-world tasks [1], such as challenges for datasets like ImageNet [2] and COCO [3]. These and other data underlie supervised learning approaches, as in Fig. 1a, for increasingly complex tasks. Scientific experiments, on the other hand, have no dataset except what they generate; each experiment constitutes its own unique task, sometimes entirely disjoint from established datasets. The man-hour investment required for labeling scientific images motivates our approach, leveraging active learning [4], [5], [6] and data augmentation, [1], [7]. The focus of this report, moreover, is the application of data augmentation to reduce overfitting.

A. Governing Principles

Scientific experiments use images as a measuring technique, recovering properties like object position or orientation for further analysis. These quantities encode the underlying structure of the experiment, if any exists, which is under investigation. Images of a sphere in free-fall, for example, capture the laws of gravity, which any measurement *aiming to study gravity* ought to ignore for the sake of scientific rigor. Experiments that anticipate such *governing principles* in the data commit a logical fallacy by assuming the initial point. This observation is particularly important for training DNNs in scientific tasks. Like any statistical model, DNNS are prone to overfitting when underlying correlations are present. In order to address this issue, we contrast scientific tasks with traditional tasks in computer vision.

Fig. 1a outlines the training process for traditional vision tasks. Typically, such approaches leverage a large, fully labeled training set (X, Y), *e.g.* ImageNet, where X is the set of $M \times N$ images and $Y \subseteq \mathbb{R}^n$ is the label set. In any task, the labeling comprises a lower-dimensional representation of the data space, encoding valuable information from each example. n denotes the dimensionality of the labeling, *e.g.* the number of classes. Supervised learning aims to train a model (such as a DNN) $h : \mathbb{R}^{M\times N} \to \mathbb{R}^n$ 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.

For real-world tasks, sampling f in this manner is desirable, but in scientific tasks, governing principles make sampling from the original distribution impossible. In fact, for the purposes of training, we wish to sample from a distribution as uniform as possible in the label space $\mathbb{R}^n$. That is, we wish for the dataset X to correspond to labels Y which contain almost no structure, within reasonable boundaries. Fortunately, most scientific experiments have well-known boundaries incorporated into their design. A sphere in free-fall, for instance, might be bound to one-dimension by a track, even though its image-space position is described by two coordinates. Fig. 2a shows an experiment where dot markers are confined to a small circular region. These *label-space boundaries* are a low-dimensional representation of a data-space region $\chi \subseteq \mathbb{R}^{M\times N}$ such that the real distribution f obeys

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

Even with perfect knowledge of these boundaries, however, it is difficult or impossible to recover χ from the lower-dimensional label-space boundaries.

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

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

[2]Alternatively, deep convolutional neural networks.

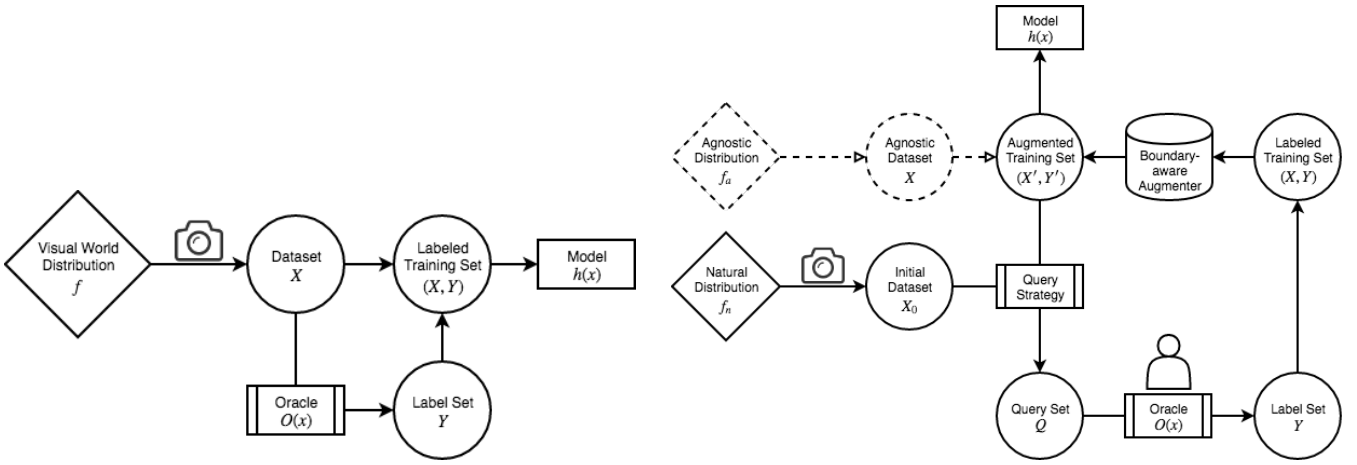


Fig. 1. During supervised learning for real-world vision tasks **(a)**, a camera samples a large dataset X from the "visual world" distribution f. For scientific tasks **(b)**, an experiment samples a small initial dataset X_0 from the natural distribution f_n, which includes the effects of any governing 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.

dataset $X_a \sim f_a$. In this report, we introduce a *Boundary-Aware Augmenter* (BAA) that utilizes label-space boundaries and instance segmentations to approach X_a.

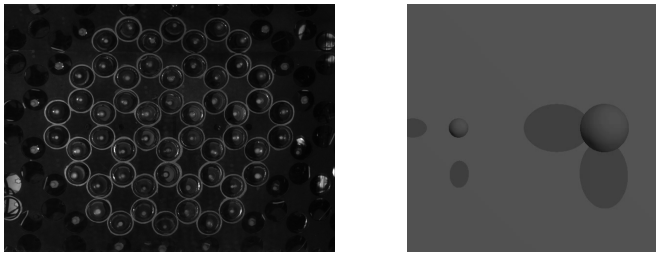
II. Related Work

The problems of data scarcity and biases are well-considered in computer vision. [8] explores bias in popular datasets for computer vision by training DNNs on one dataset but testing them on another (employing a test-train split for fairness). Unsurprisingly, networks perform best on the dataset for which they were trained, even though datasets like ImageNet aim to capture the unbiased visual world. Creating unbiased datasets for real-world vision tasks remains an open problem, one which may demand a reckoning for the community's focus on dataset performance scores.

[7] describes a novel segmentation architecture using up-convolutions which, paired with their data augmentation scheme, performs well on biomedical images from electron microscope stacks. Like our approach, [7] confronts data scarcity but also focuses on applications in biomedical imaging, where semantic segmentation is often sufficient. We aim to address scientific tasks more generally, using semantic segmentation as one component of a general detection system, capable of recovering location, orientation, or shape for multiple objects. Capturing these quantities with minimal human effort would accelerate a wide range of scientific experiments.

III. Method

Our method employs a combination of active learning and data augmentation to minimize the experimenter's labor-investment. Since this report focuses on minimizing the effect of governing principles on the dataset, we rely on existing literature [4], [10] for query selection strategies. In spite of this, the query strategy is a vital part of any practical application of our method, facilitating fast learning and ease of use.



(a) Gyroscopic Model (b) Coupled Spheres

Fig. 2. Example images from scientific experiments: **(a)** gyroscopic model for topological metamaterials [9]. 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.

The function of the 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, the BAA must be able to generate examples x from arbitrary points y in label space $\mathbb{R}^n$. In principle, the label space can include any number of object properties which the experimenter wishes to measure, including position in the image, apparent orientation, and size. These apply to every object in the image, resulting in a label-space that can grow relatively large, with independent boundary constraints for every coordinate constraining χ. For the following explanation, we consider a label space of object positions, such for the spheres in Fig. 2b, but maintain that the same principles apply for higher-dimensional label spaces.

In order to freely manipulate examples in label-space, we require an instance segmentation of each image. With this information, we 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, but 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. [11] describes a more nuanced approach that attempts to complete isophote lines arriving at the region's edge. [12] introduces Context Encoders: DNNs 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.

IV. Simulated Experiments

TODO: UPDATE SIMULATED EXPERIMENTS

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 Fig. 2a. 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 DNN with predictions $\hat{Y}$ that more closely approximate $\tilde{Y}$ than the labels Y. Simulated experiments allow us to test this performance.

Fig. 2b 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. Imposed constraints include the z-coordinate of each sphere, which is set to the image-plane, as well as each sphere's apparent size. Inherent constraints include the physical properties of the spring, *e.g.* the spring constant and relaxed length. Adverse noise, which in this case includes any shadows, lighting effects, and possible occlusion, also presents a challenge for detection.

To demonstrate our method's resilience to inherent constraints, we intend to train a DNN 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.

V. Conclusion

TODO: CONCLUSION

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] 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

[4] 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

[5] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "Localization-Aware Active Learning for Object Detection," *arXiv:1801.05124 [cs]*, Jan. 2018, arXiv: 1801.05124. [Online]. Available: http://arxiv.org/abs/1801.05124

[6] S. D. Jain and K. Grauman, "Active Image Segmentation Propagation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2864–2873.

[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] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, Jun. 2011, pp. 1521–1528.

[9] 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

[10] 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.

[11] 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

[12] 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/