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* (GPs) 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 labeling 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 [3], such as challenges for datasets like ImageNet [4] and COCO [5]. 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 approaches such as active learning [6], [7], [8] and data augmentation, [3], [2]. The focus of this report, however, is the application of boundary-aware augmentation, which aims to reduce overfitting to governing principles (section I-A).

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* (GPs) 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 undesirable. 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 GP-agnostic models. 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 a GP-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

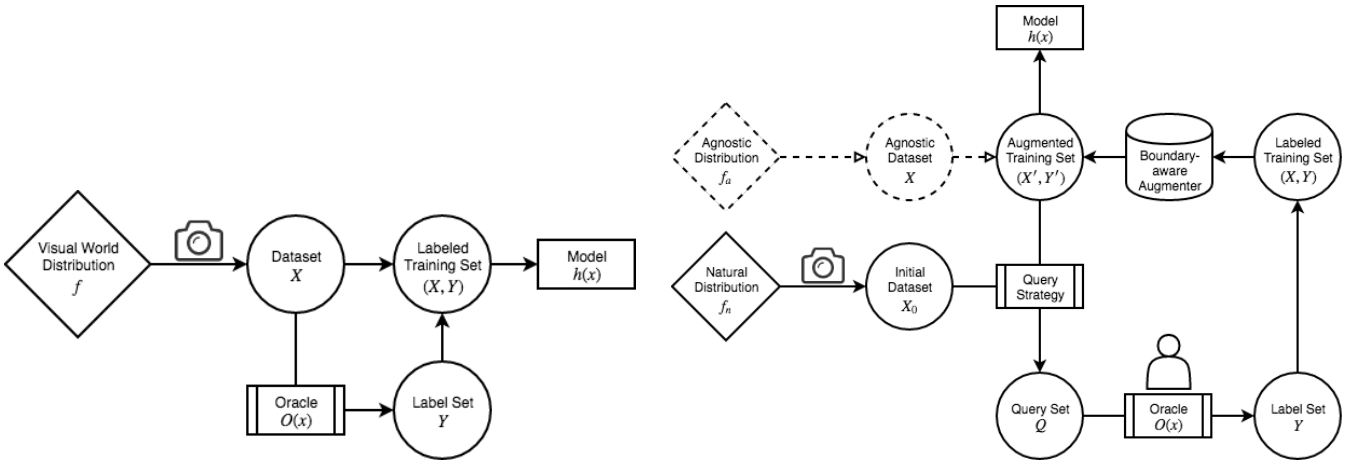


Fig. 1. During supervised learning for real-world vision tasks **(a)**, a camera samples a large dataset X from the "visual world" distribution f. For scientific tasks **(b)**, an experiment samples a small initial dataset X_0 from the natural distribution f_n, which includes the effects of any governing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

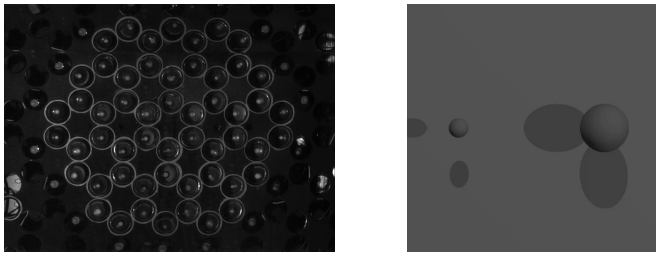
The problems of data scarcity and biases are well-considered in computer vision. [9] explores bias in popular datasets for computer vision by training DNNs on one dataset but testing them on another (employing a test-train split for fairness). Unsurprisingly, networks perform best on the dataset for 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.

[2] describes a novel segmentation architecture using up-convolutions which, paired with their data augmentation scheme, performs well on biomedical images taken from electron stacks. Like our approach, [2] 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

TODO: UPDATE METHOD

Fig. 1b shows our training scheme at a high-level. The details of this scheme are an area of active inquiry, with open questions pertaining to the *selector*, for active learning; the *augmenter*, which should incorporate the experiment's imposed constraints; and the model f, which employs one or more DNNs.



(a) Gyroscopic Model (b) Coupled Spheres

Fig. 2. Example images from scientific experiments: **(a)** gyroscopic model for topological metamaterials [10]. 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.

Because of the variety of scientific tasks, f must remain flexible with regard to its output. Toward this end, we envision a two-step procedure which (1) obtains an instance segmentation [2], [11] for objects of interest and (2) learns the target parameters (location, orientation, etc.) for each object. Whether these steps occur in an end-to-end fashion or are divided between several training steps is an open question. We do, however, consider (1) to be a vital step for the sake of maintaining generality. Even in scientific tasks where segmentation is unneeded, such as the Coupled Sphere experiment in III, obtaining pixel-level masks of each object enables more advanced data augmentation methods.

Because we wish to minimize the man-hours required for training as much as possible, we use an active learning scheme to select a query set $Q \subseteq X'$ which will most inform training. The application of active learning to semantic segmentation has received relatively little attention, with the exception of [12], and the added complication of an imperfect labeler remains an open question in the field [6]. We aim to address both issues with its *selector*.

Finally, the *augmenter* will incorporate both the semantic

segmentations obtained by f and the imposed constraints specified by the experimenter to improve training as much as possible. We hope to test many augmentation methods while keeping in mind that the image space of a scientific task is usually much more constrained than that of a real-world task. [3], for instance, uses sub-image extraction, flipping, and PCA analysis of RGB channels to augment ImageNet. These methods are not necessarily applicable to scientific tasks, where imposed constraints might invalidate a flipped image, for instance.

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] Bernardis, E., and Yu, S. X., 2010. "Finding dots: Segmentation as popping out regions from boundaries". In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 199–206.

[2] Ronneberger, O., Fischer, P., and Brox, T., 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *arXiv:1505.04597 [cs]*, May. arXiv: 1505.04597.

[3] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. "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., pp. 1097–1105.

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

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

[6] Settles, B., 2012. "Active Learning". *Synthesis Lectures on Artificial Intelligence and Machine Learning,* **6**(1), June, pp. 1–114.

[7] Kao, C.-C., Lee, T.-Y., Sen, P., and Liu, M.-Y., 2018. "Localization-Aware Active Learning for Object Detection". *arXiv:1801.05124 [cs]*, Jan. arXiv: 1801.05124.

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

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

[10] Nash, L. M., Kleckner, D., Read, A., Vitelli, V., Turner, A. M., and Irvine, W. T. M., 2015. "Topological mechanics of gyroscopic metamaterials". *Proceedings of the National Academy of Sciences,* **112**(47), pp. 14495–14500.

[11] Bai, M., and Urtasun, R., 2016. "Deep Watershed Transform for Instance Segmentation". *arXiv:1611.08303 [cs]*, Nov. arXiv: 1611.08303.

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