SAND2022-8694C

But it Looks so Real! Challenges in Training Models with Synthetic Data for International Safeguards

Zoe N. Gastelum, Timothy M. Shead, and Matthew R. Marshall

Sandia National Laboratories[i]

Albuquerque, New Mexico, USA

Abstract

The general unavailability of real international nuclear safeguards data for data science research and development projects has led many researchers in this domain to turn to synthetic and simulated datasets to develop and prove modeling concepts. In recent work, Sandia National Laboratories has developed a large, synthetic, machine learning-validated dataset of images of containers used to transport and store natural and low-enriched uranium hexafluoride – specifically 30B and 48-type cylinders. The dataset also includes synthetic images of distractor objects such as 55-gallon drums and propane tanks. The purpose of these synthetic images is to address the need for safeguards-relevant data to support computer vision research. In our validation process, we faced the canonical challenge of generalizing models trained on synthetic data to make predictions on real-world data. In this paper, we will describe the challenges and observations from our research training models on synthetic images to make predictions on real-world images. We will present our priorities in future research directions using our large, publicly available synthetic image dataset that has the potential to enhance the state of synthetic-to-real research and development.

Background

In recent years, there has been growing interest in assessing the potential of computer vision models to detect and categorize images relevant to international nuclear safeguards, including applications in open-source information collection and analysis (Feldman, et al. 2018) (Gastelum and Shead 2018), surveillance camera imagery review (Thomas, et al. 2021), and overhead imagery analysis (Rutkowski, Canty and Nielsen 2018). All of these projects faced challenges finding and labeling relevant dataset, and significant resources were expended to collect and label sufficient data. In some cases, costly partnerships with commercial entities were formed to facilitate access to data.

In response to a growing call for safeguards-relevant images to support development of computer vision models, our team is developing Limbo: an open-source dataset of one million safeguards-relevant synthetic images created from 3D computer models of objects in real-world and fully computer-generated environments, in which we control the virtual materials, conditions, lighting, position, and cameras. Our data development process is described in (Gastelum, Shead and Rushdi 2021). The subjects of our data are two types of containers used to store and transport uranium hexafluoride (UF_6) throughout the commercial nuclear fuel cycle. In particular, we focus on one 30-inch container design (30B) used to store and transport low enriched (up to five

percent Uranium-235) uranium (Figure 1), and several 48-type container designs used to store (48X, 48Y, and 48G) and transport (48X and 48Y) natural and depleted uranium (Figure 2). We selected these containers based on several desirable characteristics including their unclassified nature, allowing us to more easily share the data with researchers throughout the safeguards and computer vision research communities; their ubiquity throughout multiple stages of the nuclear fuel cycle, making them relevant for a broad set of research projects; their distinct appearances, allowing us to confirm container identities in real-world images with high certainty; and the relative availability of real world images with which we could validate our synthetic data.

Figure 1. 30B containers are used to store and transport low-enriched uranium. Credits: Katy Laffan / IAEA, October 2019, CC BY 2.0 (left); Los Alamos National Laboratory, 2012 (right).

Figure 2. 48-inch containers are used to store and transport natural and depleted UF_6. Our Limbo dataset includes three models: the 48Y (left) and 48X are used to store and transport UF_6, while 48G containers (right) are used exclusively for storage. Credits: IAEA, 2015 (left); United States Department of Energy, 2016 (right).

Validation

A key step in our data generation workflow is data validation. During data validation, we train computer vision models using the synthetic data, and test the models on real-world data. We

infer what the models are learning from the synthetic data using a variety of explanatory techniques applied to their outputs, adjust the content of our synthetic images to address any shortcomings, and repeat the process.

Our validation workflow is based on two types of computer vision models: first, we use *image classification* models, which are trained to predict which *class* from a set of mutually exclusive categories most strongly applies to an entire image. In past examples we classified images as "remote manipulator" or "not" (Figure 3), and in the current work we train classifiers to choose among classes that include the different types of UF_6 containers ("30B", "48X", "48Y", "48G") in addition to distractor classes such as "propane tank", "55-gallon drum", "beer keg", and so on. Second, we use *object detection* models, which can identify multiple *instances* of objects within a single image, producing a bounding box and predicted class for each instance.

Though the purpose of our current work is not to train computer vision models, the validation workflow ensures that computer vision models can learn from our data, identifying issues with the synthetic data in the process. Based on this workflow, along with prior work training models on synthetic data to test real images previously published in (Gastelum, Shead and Higgins 2020), there are several interesting observations that we believe are relevant for future users of our data and other synthetic datasets. Those observations are:

1. Image backgrounds have an unexpectedly large impact on model performance.
2. Negative examples are more effective when they include distractors.
3. Object configuration and positioning influence identification.
4. Computer vision models generally are learning the wrong lessons from training data.

Each of these are discussed in detail in the following sections. Note that for all comparisons of model performance discussed below, we trained two sets of ten models and compared average performance metrics from each.

Observation 1: Image backgrounds have an unexpectedly large impact on model performance.

In ground-based visual imagery such as photographs available in open sources, it is common to think of an image as having a *foreground subject* (or *subjects*) and a *background*. Whether working with image classifiers or object detectors, it is well known that backgrounds can provide context that influences model predictions. Depending on circumstance, that influence can be positive or negative, reinforcing either correct or incorrect predictions. In extreme cases, a model may make predictions driven almost entirely by the contents of the background, as in the case of a model that misclassifies dogs as wolves when the background contains snow (Ribeiro, Singh and Guestrin 2016). To overcome these tendencies, a typical solution is to train with more data, in the hopes that increased background variance will force models weight foreground features more heavily for predictions.

Synthetic data provides a powerful tool for exploring the impact of backgrounds on predictions. As described in (Gastelum, Shead and Higgins 2020), we conducted a series of experiments in which we trained ResNet image classifiers (He, et al. 2015) using synthetic images of remote

manipulator arms and tested them on real-world images of industrial environments with-and-without manipulator arms. Each synthetic training image was rendered using a 3D model of a manipulator arm, in a randomly chosen pose, against a randomly chosen background. In our first round of experiments, our backgrounds were 360-degree panoramic high dynamic range (HDR) images of industrial environments. These images provided both the background and the lighting for the manipulator arm model, leading to extremely realistic shading and perspective. Because the backgrounds were chosen at random, they were – by definition – decorrelated with the foreground manipulators, so that they should have had little impact on the trained model results. Yet when we substituted a set of randomly chosen, non-panoramic background images, also from industrial environments, also decorrelated, and trained with the same set of foregrounds, the model accuracy decreased. To reiterate: the foreground manipulators were the same in both cases, and only the backgrounds changed. This was extremely surprising, and it bears repeating that – unlike the case of the dogs and snow – the randomly chosen backgrounds in this case should not have had any impact at all on trained model performance since they were not correlated with the ground-truth in any way.

We ultimately repeated these experiments with a variety of backgrounds, including blank backgrounds populated with black, white, the mean color from ImageNet (medium brown), and a never-repeated low frequency noise texture (see Figure 3). In all cases, training images with these backgrounds which, again, were completely decorrelated from the foreground ground truth, produced even lower-performing models.

The takeaway from this is that whether generating synthetic data or curating real-world data, *the choice of backgrounds is extremely important.* Sine the goal of our models is to identify objects of interest that are typically located in the foreground, this raises significant concerns and points to generally larger problems with computer vision models.

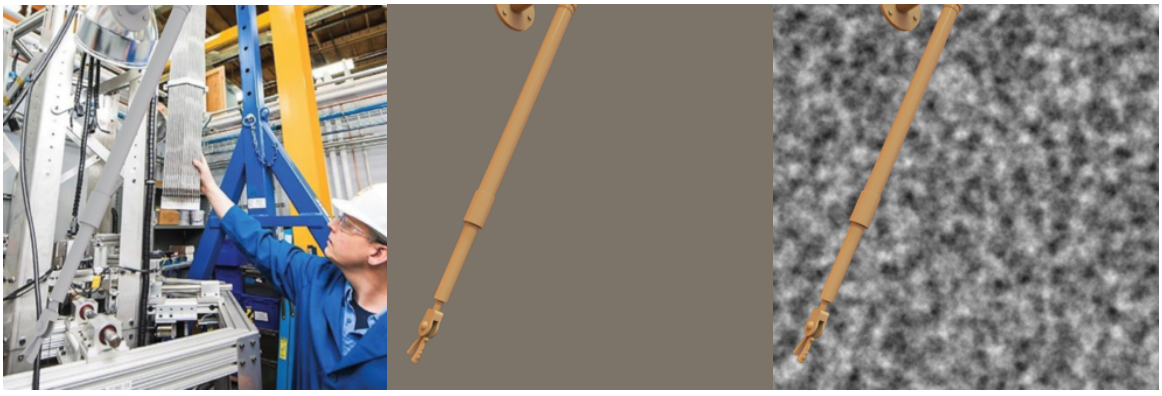

Figure 3. Synthetic remote manipulator arm against real-world industrial 2D images, random noise, and the mean color results from the ImageNet dataset as backgrounds. Industrial background image credit: TerraPower.

Observation 2: Negative examples are more effective when they include distractors.

When classifying images as part of our Limbo validation workflow, we trained ResNet-50 (He, et al. 2015) and Inception V3 (Szegedy, et al. 2015) classification models on synthetic data and analyzed their predictions on real-world data to identify biases in our synthetic data that might lead to incorrect predictions. Using explanatory methods, including Integrated Gradients (Sundararajan, Taly and Yan 2017), Gradient SHAP (Lundberg and Lee 2017), and Occlusion (Zeiler and Fergus 2013), we used heat map visualizations of individual true-positive and false-positive predictions made by the models to infer their behavior.

In early experiments, we trained using positive examples of UF_6 containers viewed against real-world HDR backgrounds, and negative examples with just the backgrounds. Because this approach did not provide examples of cylindrical objects other than UF_6 containers, we saw that the model predictions were being driven by a set of generic cylindrical features, leading them to misclassify any real-world cylindrical object as a UF_6 container.

To address this, we introduced a wide variety of synthetic cylindrical distractors into our negative examples, including barrels, paint cans, gas cylinders, propane tanks, and water tanks. Subsequently, we saw large decreases in false positive rates, accompanied by smaller decreases in true positive rates as the models became more hesitant to classify any cylinder as a UF_6 container.

For object detection models including SSD (Liu, et al. 2015) and Faster RCNN (Ren, et al. 2015) we saw similar results, using the predicted locations of bounding boxes as an indication of features upon which the model was basing its prediction. From these experiments we noticed that we were missing one significant class of object that was present in our real-world data but not in our synthetic data: human beings, which were often incorrectly identified as UF_6 containers!

The takeaway here is that whether generating synthetic data or curating real world data, *the choice of distractors is extremely important*; this also points to deeper problems with computer vision models, since *it is impossible to generate or locate examples of every possible distractor*.

Observation 3: Object configuration and positioning influence identification.

Using the same Limbo data validation workflow described above, we trained image classification models on either individual containers, individual distractors, or blank backgrounds, and tested the models on our real-world data. Using the explainability methods described above, we found that when evaluating real world images of UF_6 cylinders organized in rows, the first few cylinders in a row contributed all of the salience for predictions, while the partially occluded containers in the remainder of the row contributed none. We saw similar behavior using the object detection models. Based on this observation, we generated new synthetic images in which our UF_6 containers appear in rows (see **Error! Reference source not found.**), and we are currently developing synthetic images that we hope will improve object detection performance in highly complex scenes.

However, like observation 2, the positioning and configuration of both relevant objects and distractors appears important whether generating synthetic data or curating real world data; yet *it*

is impossible to generate or locate examples of containers in every possible permutation of configuration or placement. Since a human with no prior training can generalize understanding from one individual container to two containers side-by-side, this is another example of limitations in computer vision models.

Observation 4: Computer vision models generally are learning the wrong lessons from training data.

In addition to the above, we have frequently observed that humans have no difficulty recognizing the subject(s) of our synthetic images. Anecdotally, people often assume that our synthetic images are real, see Figure 4. This is in stark contrast with the models, where training with synthetic data typically imposes a performance penalty when evaluating real images. This implies that the trained models are overly discriminating, making decisions based on features that humans are either unaware of or ignoring.

Machine learning researchers - particularly those who enjoy the luxury of easily obtainable data - might argue that this is simply because synthetic data isn't drawn from the same distribution as real data. While this is technically true – the synthetic data *is not* drawn from the same distribution – it ignores the larger problem that computer vision models and human are either using different features to make decisions, weighting those features differently, or some combination of the two, The rarity of real-world data for use in research and development on international nuclear safeguards problems compels us to approach the problem from a different perspective, to develop models to generalize better and behave more like people in their predictions.

Figure 4. Human observers have no difficulty recognizing synthetic UF_6 containers as UF_6 containers, and in some cases cannot tell whether the images are synthetic or not. The image on the left is a real-world image (Credit IAEA via Flickr, 2020), and the image on the right contains synthetic containers against a real-world background.

Research Priorities

Our future research priorities are based upon the observation that computer vision models are learning the wrong features from training data. We see that computer vision models are overly influenced by image backgrounds, the presence and identity of distractors, and the placement and configuration of relevant and distractor containers. In each of these cases, what appear to be irrelevant differences in synthetic images are straining the performance of computer vision models when making inferences on real-world data.

Rather than considering these issues to be criticisms of synthetic data, we view them as research opportunities for the computer vision community. When people can make correct predictions from data and models cannot, we argue that the models are the problem, not the data. More pointedly, we believe that the computer vision community has been a victim of its own success – spectacular early improvements in model performance, driven by the development of CNNs, has stifled research into alternative features and decision-making, leading to the fastidiousness of current models. We do not believe that deeper networks or more elaborate training schedules, designed to extricate a few fractions of a percent in performance on popular sample datasets, are likely to overcome these limitations.

In large commercial applications, the supply of imagery is abundant and the consequences for missed detections are minor. Given the scarcity of real international safeguards data that can be used for research and development and the high consequences for missed indicators of nuclear proliferation activity, it is essential that we develop computer vision models that can make robust predictions from either very limited numbers of real training exemplars (in the single digits), or larger quantities of synthetic data. We challenge the computer vision research community – in collaboration with international nuclear safeguards and other nuclear nonproliferation experts – to reexamine basic assumptions about how to create robust computer vision models that can help ensure that global nuclear activities remain peaceful in use.

Acknowledgements and Information

This work was funded by the U.S. Department of Energy, National Nuclear Security Administration, Defense Nuclear Nonproliferation Research & Development program.

The Limbo data described in this paper are available via the Berkeley Data Cloud (BDC) Limbo library. More information about the data is available at https://limbo-ml.readthedocs.io/. Thanks to Lawrence Berkeley National Laboratory for hosting our data on BDC, and to Lawrence Livermore National Laboratory for providing initial seeds for our real-world image collection.

References

Feldman, Yana, Margaret Arno, Carmen Carrano, Brenda Ng, and Barry Chen. 2018. "Toward a Multi-Modal Deep Learning Retrieval System for Monitoring Nuclear Proliferation Activities." *Journal of Nuclear Materials Management* XLVI (3): 68-80.

Gastelum, Zoe N., and Timothy M. Shead. 2018. "Inferring the Operational Status of Nuclear Facilities with Convolutional Neural Networks to Support International Safeguards Verification." *Journal of Nuclear Materials Management* XVLI (3): 37-47.

Gastelum, Zoe N., Timothy M. Shead, and Michael Higgins. 2020. "Synthetic Training Images for Real-World Object Detection." *Proceedings of the Annual Meeting of the Institute of Nuclear Materials Management.* 1-10.

Gastelum, Zoe N., Timothy Shead, and Ahmad Rushdi. 2021. "A Large Safeguards-Informed Hybrid Imagery Dataset for Computer Vision Research and Development." *Joint Annual Meeting of the Institute of Nuclear Materials Management and the European Safeguards Research and Development Association.*

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." *CoRR* abs/1512.03385. http://arxiv.org/abs/1512.03385.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. "SSD: Single Shot MultiBox Detector." *CoRR* abs/1512.02325. http://arxiv.org/abs/1512.02325.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." Edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. *Advances in Neural Information Processing Systems.* Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Ren, Shaoqing, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *CoRR* abs/1506.01497. http://arxiv.org/abs/1506.01497.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier.* arXiv. https://arxiv.org/pdf/1602.04938v1.pdf.

Rutkowski, Joshua, Morton J. Canty, and Allan A. Nielsen. 2018. "Site Monitoring with Sentinel-1 Dual Polarization SAR Imagery Using Google Earth Engine." *Journal of Nuclear Materials Managament* XLVI (3): 48-59.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." *CoRR* abs/1703.01365. http://arxiv.org/abs/1703.01365.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. "Rethinking the Inception Architecture for Computer Vision." *CoRR* abs/1512.00567. http://arxiv.org/abs/1512.00567.

Thomas, M., S. Passerini, Y. Cui, J. Rutkowski, S. Yoo, Y. Lin, M. Smith, and M. Moeslinger. 2021. "Deep Learning Techniques to Increase Productivity of Safeguards Surveillance

Review." *Proceedings of the Institute of Nuclear Materials Management and European Safeguards Research & Development Association Joint Annual Meeting.* Virtual.

Zeiler, Matthew D., and Rob Fergus. 2013. "Visualizing and Understanding Convolutional Networks." *CoRR* abs/1311.2901. http://arxiv.org/abs/1311.2901.

---

[i] *Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.*