A Survey of Algorithms for CT Based Baggage Screening

Kraenion Labs, January 2020
Latest version: Web PDF

1 Introduction

The area of Machine Learning (ML) based threat object detection and modified object detection based on luggage CT scans is in its infancy and the research literature on the topic is sparse. However, a large fraction of the research in Computer Vision / Machine Learning (CVML) is portable across application domains provided suitable labeled data is available. There is a body of research from domains with similar challenges that can be adapted for passenger property screening. ML research for detecting various medical conditions using clinical CT scans has much in common with threat detection. There is a large body of research that has been developed for medical and industrial 2D X-ray analysis that can be extended to the 3D case. Many image based ML methods are agnostic to the type of sensor used, so research related to MRI scans can often be modified and applied to CT scans. Lastly, some research from the industrial inspection and Radiographic Non-Destructive Testing (NDT) fields is applicable.

In this survey, we consider existing research from the following broad categories: a) Research that explicitly targets security screening using 3D or 2D. b) Research from medical and industrial domains that can be adapted for passenger property screening. c) General ML and image processing techniques that can be adapted for passenger property screening.

2 Background Material

For the sake of completeness, we provide a brief introduction to ML as applied to CT scans before moving on to the research literature.

2.1 CT Scan Data

Structure and Visualization: A CT scan image is a 3D array of pixels / voxels where each voxel may be either single channel or multi-channel in the case of dual-energy or multi-energy CT. The 3D array is a stack of 2D image slices. Computer graphics rendering techniques commonly used for games and engineering work renders 3D *surfaces* using triangle meshes. In that case visualization techniques are concerned with the surface appearance of objects in the scene. Unlike 2D photographs and 3D graphics, CT scans must be volume rendered so that we can see internal details of objects, not just their surface. Volume rendering is an advanced computer graphics technique in which each voxel is given a partly transparent color. The left side of Figure 1 shows a representative CT image stack consisting of slices. The right side of the figure shows a volumetric rendering of the contents of a backpack.

Dynamic Range: X-ray imagery has high dynamic range, so details are often not easily visible when viewed on a computer display. Details can be made visible using statistical image processing techniques. The left side of Figure 2 shows an image slice from a medical CT scan. Because of the high dynamic range, most details will not be visible to the human eye either on a display or in a print out. The right side shows the same image after applying statistical image processing to render details visible. When ML techniques are applied to CT scans, it is necessary to integrate such rendering methods both for annotating training images and for visualizing the output of the ML system. Most image labeling and visualization tools currently used for ML lack the necessary features.

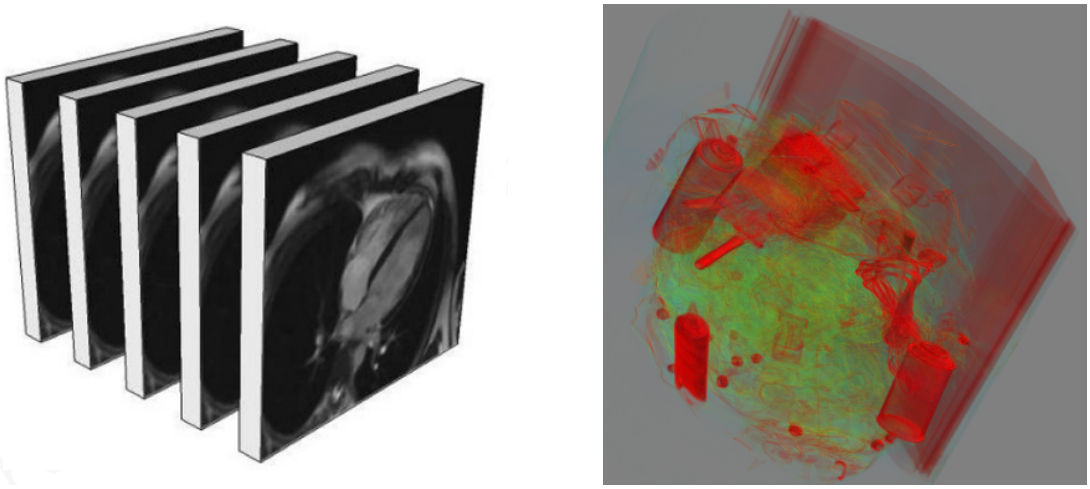

Figure 1: Left: CT Image Slices Right: Volume Rendered CT Image of a Backpack

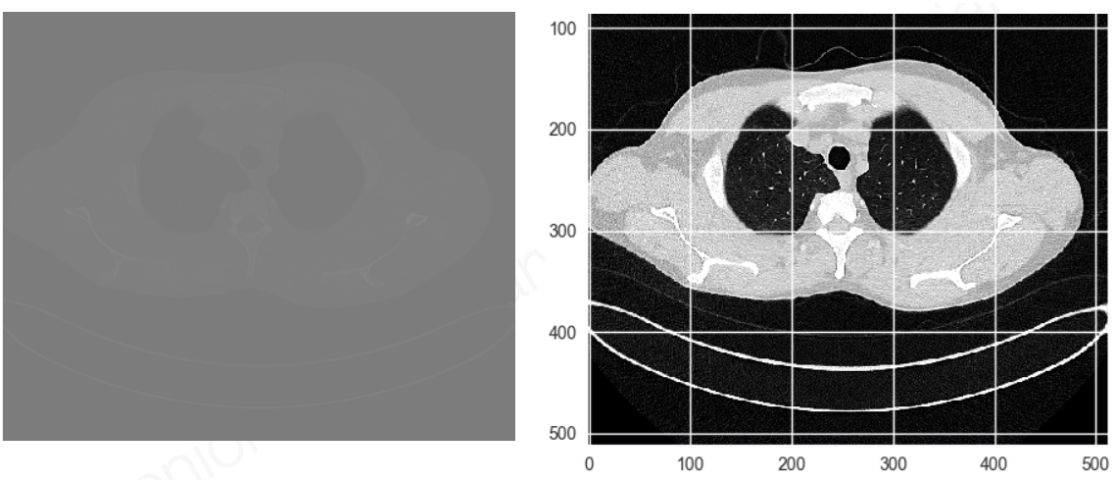

Figure 2: CT / X-ray Dynamic Range Processing

2.2 Convolutional Neural Networks for Volumetric Data

Convolutional Neural Networks (CNN) commonly used in CVML owe their name to the fact that the dominant operation is the application of a convolutional filter to the image data. Typically, the initial convolutional layers are followed by associative layers. For classification problems, the last layer is a softmax. When presented with an input image, the network produces a probability distribution over classes as the softmax layer output. The set of layers together define a receptive field on the input image. When an image larger than the receptive field needs to be analyzed, the CNN may be slid over the image with some stride and the process needs to be repeated on scaled versions of the image.

An alternative to the sliding window approach is to use a neural network to find regions and scales of interest within a large image and then apply a classifier network just to those areas. The region proposal network and the classifier network are often combined into one single network in approaches such as Faster-RCNN. [Ren et al., 2015] An alternative approach that works well for semantic segmentation and similar tasks is to consider all layers of the network to be convolutional. That approach is called a Fully Convolutional Network (FCN or FCNN) [Shelhamer et al., 2017]. The architectural simplicity of FCN makes it a particularly promising approach for analyzing high pixel count data such as baggage CT scans that include amorphous regions of liquids and powders and flexible objects such as cables and wire.

Convolutional networks for 3D and higher dimensional images have no natural diagrammatic representation since the convolutional filters will have more than 3 dimensions. Just to explain concepts, consider the example CNN in Figure 3 designed for use with 2D gray scale images.

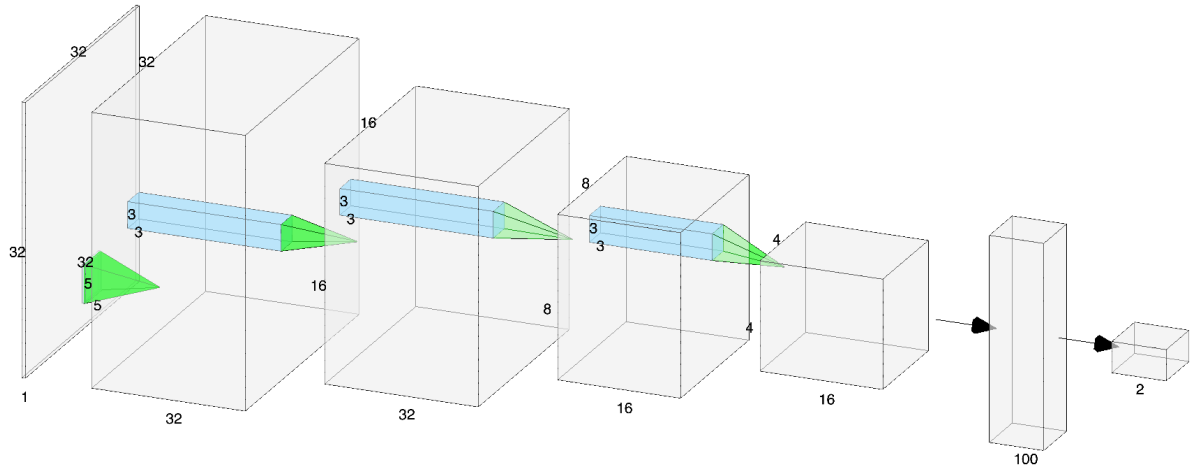

Figure 3: Example Convolutional Neural Network

Computational Complexity of 2D CNN: For a 2D CNN, the shape of the weight matrix of a single convolutional layer can be represented as $c_{out} \times c_{in} \times k_h \times k_w$ where c_{in} and c_{out} are the number of channels in the input and output images and the kernel looks at a window of $k_h \times k_w$ pixels. When this convolution is applied to an image I of size $c_{in} \times I_h \times I_w$, multiply-add operations happen at each pixel location. There are $c_{in} \times k_h \times k_w$ multiply-adds per output channel at each pixel location. Thus, the total computation per layer is $c_{out}c_{in}k_hk_wI_hI_w$ multiply-adds or $c_{out}c_{in}k^2I_hI_w$ for square filters. In the example in Figure 3 the image is $1 \times 32 \times 32$ and the first convolution kernel is $32 \times 1 \times 5 \times 5$ yielding an output volume of $32 \times 32 \times \times 32$ that is passed on to the next layer. This is the complexity of using an already trained network for inference. The complexity of inference is what determines the response time and throughput of a deployed system. Training using stochastic gradient descent is much more computationally demanding, but done off-line.

Convolution Lowering: The convolution itself happens as if the weights are organized into a 2D weight-matrix W of size $c_{out} \times (c_{in}k_hk_w)$. The matrix W is left multiplied against a column vector of size $c_{in}k_hk_w$ extracted from each pixel location in the image. This is called convolution lowering. Nvidia has published details of convolution lowering as implemented in cuDNN [Chetlur et al., 2014]. Nearly all neural network training software in 2020 are based directly on cuDNN or on reimplementations of cuDNN primitives.

Computational Complexity of 3D CNN: It is straight forward to extend CNNs to 3D volumetric images. The image is 4D with dimensions $c_{in} \times I_d \times I_h \times I_w$, i.e., I_d is the depth dimension and the image is a stack of $c_{in} \times I_h \times I_w$ 2D images. Convolution needs to be extended to the depth dimension in addition to the height and width dimensions as before. So the convolution kernel is now 5D with dimensions $c_{out} \times c_{in} \times k_d \times k_h \times k_w$ where k_d is the depth of the kernel. The output of a layer will then be a $4D$ (hyper) volume of $c_{out} \times I_d \times I_h \times I_w$. The number of multiply-add operations becomes $c_{out}c_{in}k_dk_hk_wI_dI_hI_w$ or for cubical filters $c_{out}c_{in}k^3I_dI_hI_w$.

2.3 Challenges for CNNs and Volumetric Data

While the volumetric extension of CNNs described in the previous section is quite simple from the mathematical point of view, it poses significant challenges in practice. To start with, the computation per layer

is a factor of kI_d larger. For a volumetric image with 1000 slices and a simple kernel of $k = 3$, each layer has a factor of 3000 more computation compared to 2D image analysis. Stochastic gradient descent used for training neural networks is usually applied to batches of images. Batch sizes of 50-100 are common. The factor of 1000 increase in image size when applied to a batch size of 50 is beyond the memory size of most year-2020 GPUs.

Neural network training software typically takes the data-flow graph describing the forward pass of a CNN and unrolls and transforms it to optimally use GPU compute resources. It is common to see GPU memory usage somewhere between 10 to 200 times the batch size. When GPU memory is inadequate for the volumetric case, all of the unrolling and optimization makes CNN training impossible requiring training strategies different from 2D images. The use of separable convolution can change the k^3 factor to $3k$. It is an area that has been studied for 2D images but is not widely used today. The use of training software that provides explicit control of memory buffers will be critical to training CNNs for volumetric data.

The locality of data, how it fits in the GPU caches and interacts with the memory hierarchy is a critical factor when dealing with GPU computing. The in-plane part of the volumetric convolution is identical to the 2D case and poses no difficulties to the memory system that is heavily optimized for that case. However, the convolution along the depth axis has a large stride equivalent to the size of each image slice and could cause severe performance degradation. It is possible to mitigate the slowdown by packing 3D data into specialized formats based on Morton / z-order space-filling curves [Nocentino and Rhodes, 2010]. Some GPUs have support for z-order indexing, however most CNN training software do not take advantage of such features.

As with all CVML applications, the processing of CT scans requires data-pipeline support for extract, transform, load, data annotation, model training, inference, visualization, reporting and analysis. Many ML libraries claim to support 3D images, but the support is often rudimentary, brittle and slow.

2.4 Applicability of 2D Algorithms

Using well understood projective geometry and computer vision techniques, 2D views may be synthesized from a 3D CT volume not only along the 3 axes of the volume, but also along arbitrary viewing directions. As such, 2D algorithms may be applied to 2D views extracted from CT scans. Just as region-proposal networks are incorporated into neural architectures such as RCNN, it might be useful to investigate view-proposal networks for CT scans. Such networks could propose interesting viewing directions, project the data and use a 2D network as the last stage of the detector.

3 Algorithmic Techniques

In this section we present a sampling of the algorithm research that uses CT/X-ray data for validation but does not involve any learning algorithm. Non-ML algorithmic techniques such are commonly used for denoising, image quality enhancement, segmentation and dataset and augmentation.

Applying Medical Segmentation Algorithms to Baggage: Megherbi et al. investigate applying 3D CT medical image segmentation methods to baggage and package security screening using CT imagery [Megherbi et al., 2013]. They present experimental results of 3D segmentation using Confidence Connected Region Growing, Fuzzy Connectedness, Watershed, and Fast Marching.

Segmentation Algorithm for Baggage CT: Grady et al. presented the Automatic Quality Measure (AQUA) and used it along with a segmentation technique for 3D CT images that makes no assumptions on the number or composition of objects in the image [Grady et al., 2012]. They use AQUA to measure the

segmentation confidence for any object from any segmentation method and use the confidence measure to choose between candidate splits and to optimize the segmentation parameters at runtime. The method was evaluated on 27 CT scans of luggage.

Reconstruction, Denoising, Segmentation: CT scans are characterized by low image resolutions, noise and clutter in comparison to state of the art photography. Mouton and Breckon present a survey of denoising and artifact algorithms for baggage CT scan images [Mouton and Breckon, 2015]. Topics addressed by the survey include denoising, metal artifact reduction, pre-reconstruction, post-reconstruction and iterative reconstruction algorithms for Dual-Energy CT (DECT) based aviation security-screening. Their survey also covers segmentation research with an emphasis on the ALERT dataset [Crawford et al., 2013]. They point out that there is limited research on addressing the problem of 3D object classification and the existing literature covers only the detection of one or two threat object classes, usually hand guns. They recommend that future research should focus on extending CT based object detection to a large number of classes.

Object Separation: Unlike photographs where the top most object occludes objects below it, X-ray images do not follow the concept of occlusion. SATISφ is a method for separating objects in a set of x-ray images using the property of additivity in log space [Heitz and Chechik, 2010]. The log-attenuation at a pixel is the sum of the log-attenuations of all objects that the corresponding x-ray passes through. It uses multiple projection views of the same scene from different angles to estimate the attenuation properties of objects in the scene. Such properties can be used to identify the material composition of objects, and are useful for ATD. The method is claimed to outperform standard image segmentation and reduce the error of material estimation. Even though the research targeted multi-view X-ray, the volumetric nature of CT scans allows an arbitrary number of projected views to be generated. So the research is relevant to CT scan based ATD.

X-ray Image Enhancement: Enhancing image quality can be help human screeners cope with challenges posed by low-resolution luggage X-ray imagery. A variety of enhancement algorithms have been studied in the literature. The ideal algorithm to use depends on the characteristics of the image to be enhanced. Researchers from Loughborough University, UK, developed a neural network that can be used for predicting, on a given test image, the best image enhancement algorithm for it as judged by human experts [Singh and Singh, 2005]. The research targeted 2D X-ray images, but the visualization problem for 3D CT scans could use similar enhancement techniques. High quality visualization and image enhancement is relevant not only to human screeners, but also to annotating large CT datasets and to transformations that can make information more accessible for representation learning by neural networks. We include this research under the non-deep learning category because even though the researchers developed a neural network, the ML model's output is a recommendation to use a particular non-ML image enhancement algorithm.

Dataset Augmentation: The difficulty of obtaining large datasets and the major imbalance in the occurrence of threat and benign objects complicates the training and testing of ML models for Automated Threat Detection (ATD). Joint work between researchers at Rapiscan Systems and the University College of London addressed this problem using a framework for Threat Image Projection (TIP) in cargo transmission X-ray imagery [Rogers et al., 2016]. They exploited the approximately multiplicative nature of X-ray imagery to extract a library of threat items. The TIP framework can project those library items into images of real cargo and generate a very large number of training examples by adding realistic, random, variation during projection. Variations include translations, magnification, rotations, noise, illumination, volume and density and obscuration. The end goal is to add robustness to ML algorithms by allowing them to learn representations that are invariant to such variations. In addition, it also enables a deeper understanding of algorithm performance by carefully controlling aspects of the test data such as threat position or obscuration. Even though TIP was developed in the context of 2D X-rays, the general principles

are applicable to dataset augmentation of CT scans.

4 Threat Object Detection without Deep Learning

CT scan datasets of baggage became available after deep learning rose into prominence through the ImageNet challenge from 2012 onwards. Before that period, Support Vector Machines (SVM) were the most significant approach for image classification. We expect most of the new research in CT based threat detection to use deep learning. We now present a sampling of early research that used approaches other than deep learning.

Graph-Cut Approach for CT Baggage Data: Martin et al. presented a learning-based framework for joint segmentation and identification of objects directly from volumetric DECT images, which is robust to streaks, noise and variability due to clutter [Martin et al., 2015]. They used the DHS ALERT dataset in conjunction with K-NN for learning [Crawford et al., 2013]. They segmented and identified a small set of objects of interest by learning appearance characteristics of threat objects from training images. Everything other than the set of objects is considered as background. They used data weighting to mitigate the effects of metal artifacts and approached the problem using efficient graph-cut algorithms for discrete optimization. Material labeling in the presence of metal, shading and streaking was achieved.

SVM and 3D Shape Descriptors: Megherbi et al. studied another SVM based system for detecting threats in CT based baggage imagery [Bouallagu et al., 2010]. They used SVM along with two popular 3D shape descriptors: 3D Zernike descriptors and the histogram of shape index. Their results indicate that the histogram of shape index descriptor performs better than 3D Zernike descriptors. The same group of researchers also did a comparison of multiple classification approaches for threat detection in CT based baggage screening [Megherbi et al., 2012]. They combined 3D medical image segmentation techniques with 3D shape classification and retrieval methods to compare five classification approaches: SVM, neural network, decision tree, boosted decision tree, and random forest.

SVM and Codebooks: An SVM classifier built on codebooks constructed via randomized clustering forests using a dense feature sampling strategy was explored by Mouton and colleagues [Mouton et al., 2014]. Their research claims high accuracy and fast processing time making it an attractive option for the automated classification of threats in security screening settings. Another SVM based classifier for detecting firearms in baggage security X-ray imagery was presented by Turcsany and colleagues [Turcsany et al., 2013]. They used a bag-of-words model in conjunction with SURF features. Their results indicate that class-specific clustering primes the feature space and simplifies the classification process. They concluded that the best classification results came from using SVM or random forest along with a feature descriptor.

Bag of Visual Words: Researchers at the University of Kaiserslautern explored the use of visual cortex inspired features for detecting illicit objects in X-ray images of luggage using Mutch and Lowe's SLF-HMAX and Pinto et al.'s V1-like features and compared them against SIFT and PHOW which are popular conventional feature descriptors [Schmidt-Hackenberg et al., 2012]. With a bag of visual words approach, the visual cortex inspired features outperformed the conventional features. The researchers hypothesized that the superiority of the visual cortex based features may be because they encode geometric information from textureless X-ray images better than conventional features.

5 Deep Learning for Baggage Screening

Transfer Learning with CNNs: To address the extreme scarcity of labeled X-ray images related to baggage security, Akcay et al. explored transfer learning using CNNs [Akcay et al., 2016]. They started with a CNN that was pre-trained for generalized image classification tasks where sufficient training data

exists and optimized it in a secondary process that targets X-ray baggage screening. As with other X-ray based efforts, they studied a simple 2 class problem first (fire-arm vs no fire-arm) but then extended it to 6 classes: firearm, firearm-components, knives, ceramic knives, camera and laptop. They applied CNNs to manually cropped baggage objects under the assumption a deployed detection solution would perform sliding window search through the whole baggage image as explained previously in Section 2.2. They reported that for the multi-class case, CNN with transfer learning achieves superior performance compared to prior work demonstrating the applicability of transfer learning and CNN to X-ray baggage imagery. In follow-on research, they investigated AlexNet features trained with an SVM classifier in addition to studying the applicability of multiple detection paradigms such as sliding window-based CNN, Faster RCNN, region-based fully convolutional networks (R-FCN), and YOLOv2 [Akcay et al., 2018].

Classification of Threat Objects: GDXray is an image dataset of 19,407 X-ray images of categories including baggage, natural objects, castings and welds [Mery et al., 2015]. While the number of images is very small compared to popular computer vision datasets such as ImageNet, it represents an initial attempt at a dataset of size suitable for deep learning. The creators of that dataset also published work on classifying three types threat objects: handguns, shuriken (ninja stars) and razor blades [Mery et al., 2017]. The objects were presented in isolation as cropped X-ray images. While there was significant variation in the pose of the object, the problem itself suffers from the criticism previously reported by Mouton: too few classes [Mouton and Breckon, 2015]. For comparison, as the very first exercise for students, many deep learning tutorials currently use the MNIST dataset where a neural network is trained to recognize one of the ten digits 0 to 9. The researchers compared ten approaches based on bag of words, sparse representations, deep learning, pattern recognition schemes etc. and concluded that methods based on visual vocabularies and deep features worked best.

Deep Learning for CT Reconstruction: Korean researchers have developed deep learning based 3D image reconstruction for a stationary CT using fixed X-ray sources and detectors developed for homeland and transportation security applications [Han et al., 2018]. Due to a limited number of projection views, analytic reconstruction algorithms produce streaking artifacts. They developed a novel image and sinogram domain deep learning architecture for 3D reconstruction from sparse view measurement and applied it to data from a prototype 9-view dual energy stationary CT carry-on baggage scanner developed by GEMSS Medical Systems, Korea. They report that the deep learning approach delivered high quality 3D reconstruction for threat detection. While this research does not directly target automated threat detection, it has applicability to enhancing image quality for human screeners and annotators as well as exposing information for better learning by detection networks.

A good general exposition of the use of deep learning for CT reconstruction has been published by Bazrafkan and colleagues [Bazrafkan et al., 2019]. They describe the impact of DNNs on the CT imaging in general and low dose CT reconstruction in particular. They caution that misleading industrial marketing claims are made about the use of DNNs in CT scanners, but neural networks are often only used in a secondary role to remove noise and artifacts from other reconstruction techniques.

6 Deep Learning Techniques for 3D Data from Depth Sensors

Volumetric CNNs are 3D convolutional neural networks designed for input data in the form of voxelized shapes. Pioneering work in volumetric CNNs include [Wu et al., 2015], [Maturana and Scherer, 2015] and [Qi et al., 2016]. A lot of general principles can be adapted from this field and applied to CT scan analysis. Robotics and autonomous driving are the dominant use cases for depth measurement using structured light and time of flight cameras, stereoscopy, LIDAR and RADAR. Data from such sensors can be in one of two forms: dense grid or point cloud. We will delve into the details of two representative

papers only, one for the dense grid case and one for point clouds.

Socher et al. from Stanford University developed a combination of convolutional and recursive neural networks (RNN) for classifying RGB-D images [Socher et al., 2012]. The initial CNN layer provides features used by multiple fixed-tree RNNs to create higher order features for classification. The RNNs learn hierarchical feature representations by applying the same neural network recursively in a tree structure. They demonstrate state of the art results on a 51-class dataset of household objects.

Qi and colleagues, also from Stanford University, presented deep learning techniques for 3D classification and segmentation of point cloud data [Qi et al., 2017]. Unlike most neural network researchers who convert irregular point cloud data into regular 3D voxel grids, their novel neural architecture named PointNet directly processes point clouds. PointNet effectively deals with unordered point clouds while providing a unified architecture for object classification, part segmentation and semantic parsing of scenes. The key insight that enables permutation invariance is to approximate a general function defined on a point set by using a symmetric function on transformed elements in the set. They provide experimental evidence to validate their network design using multiple 3D datasets.

As explained before, 3D CNNs have a scalability problem. Large windows needed to completely cover big objects could make the network computationally challenging. PointNet provides an alternative: the scale of the object does not matter at all. The scalability of PointNet depends on the number of points used as input, not on the physical / voxel volume spanned by the points. Even though CT scan data follows a 3D voxel grid, it may be useful to consider PointNet like architectures that consider a point cloud of critical features extracted from the CT volume.

7 Deep Learning Methods for Medical Data

ML research for detecting various medical conditions using clinical MRI and CT scans has much in common with threat detection. However, there are important caveats too. Anatomy is fixed and well understood - we have a known number of organs that appear in a known spatial layout. Anything deviating from the standard anatomical layout is an anomaly.

In contrast, a bag can be packed with an infinite variety of objects. Even if all the objects were known, they can be put together in different combinations. Context and clutter can make analysis extremely challenging. Object detection under such circumstances might seem like a lost cause, but one has to realize that a human screener has to deal with threat detection under severe time limits to keep baggage moving. Anything that can be done to aid visualization, partially automate the task and reduce the cognitive load of a human screener will benefit security.

NIH CT Dataset: Much of the deep learning work on detecting medical anomalies really only does 2D image analysis using individual slices of a CT volume. For instance, the National Institutes of Health (NIH) released the Deep Lesion CT dataset for detecting organ lesions using deep learning methods [Yan et al., 2018]. Though this dataset originates from clinical CT scans, it only provides 3 CT slices per lesion. Quoting from the paper: "we crop a patch with 50 mm padding around its bounding box. To encode 3D information, we use 3 neighboring slices (interpolated at 2 mm slice intervals) to compose a 3-channel image". Therefore, that dataset is not really volumetric. The slices are cropped from 512 $\times$ 512 pixel 12-bit images that together may be considered a small 3 channel image of much less size than the 3D CT scans of luggage. Despite not being volumetric, the NIH dataset will be foundational to developing deep learning techniques for medical science and will significantly benefit humankind.

Semi-supervised Learning and Fast Annotation: The real relevance of the NIH Deep Lesion work for CT based threat detection is in the methodology developed for creating the dataset. NIH has access to a vast, loosely-labeled, and largely untapped CT scan dataset stored in hospital Picture Archiving and Com-

munication Systems (PACS). In addition to patient images PACS contains radiological reports, markings, and measurements created by radiologists and doctors during clinical duties. The data is typically not in a form suitable for supervised ML training. NIH undertook the challenging task of developing methods to mine PACS radiology data for ML training. Likewise, security personnel screening baggage could mark CT images quickly during their work duties. Methods similar to those used by NIH could be developed to transform the data into a form suitable for ML training.

Triplet Networks and Similarity Relationships: To model similarity relationships between lesions, the researchers used multiple attributes annotated by clinicians and radiologists including lesion type, location and size. Those attributes are useful for ML training and they need very little manual annotation effort. Similar approaches may be used by human screeners to produce quick annotations that are useful for ML training without significantly impeding their normal workflow. The NIH researchers used a triplet network to learn embeddings along with a sequential sampling strategy to depict hierarchical similarity structure. A triplet network uses three samples to compute a loss function: an anchor sample, a hard positive sample and a hard negative sample [Schroff et al., 2015]. The network then learns an embedding that ensures that the anchor sample is closer to the hard positive than to the hard negative.

Fusion of 2D and 3D CNNs: Gao et al. from Middlesex University, London have published the results of their research that combines 2D and 3D CNNs to classify medical CT scans [Gao et al., 2017]. In their work, the dataset has more slices than in the Deep Lesion dataset. The original brain scans were collected at the Navy General Hospital, China in resolutions of 512×512 or 912×912 pixels with 16 or 33 slices and cropped and normalized into $200 \times 200 \times 20$ pixels. The categories include Alzheimer's disease, brain lesions and normal brains with symptoms of age. The 2D CNN is applied one slice at a time along the depth axis. The 3D network resembles our generalized description in 2.2. The convolution kernels have an added dimension compared to their 2D counterparts and 3D max pooling is used to reduce the size of the feature map in successive layers.

The 2D CNN is trained using 2D image tiles while the 3D CNN is trained on 3D box tiles with dimensions of $40 \times 40 \times 10$. Softmax layers are used individually at the output of each network to create class probability distributions and the outputs of the 2D and 3D models are fused linearly to yield the final output. The paper also provides a comparison against 2D and 3D versions of the SIFT and KAZE conventional feature detectors. The 2D/3D fused CNN outperformed the other architectures, but the performance of the 2D SIFT and KAZE variants were quite close to that of the completely neural architecture.

This paper serves as a good starting recipe for the design of neural networks for baggage screening. 3D tile based approaches as well as the fusion of 2D and 3D neural architectures provide a reasonable and computationally tractable initial approach. Unlike the medical case where lesions and disease characteristics are small, threat objects in the case of baggage screening may be much larger than the $40 \times 40 \times 10$ tile size used in the paper. Therefore, 3D versions of region proposal architectures or 3D anchor box designs extended from YOLO may be necessary.

8 Summary

Object detection and segmentation using baggage CT scans is computationally challenging on account of the high pixel count and the extreme memory and disk storage requirements. The infinite variety and combinations of objects that can be packed into cluttered bags makes the problem extremely difficult. Automating the monotonous aspects of the screening task and enhancing image and visualization quality will help in utilizing the cognitive resources of human screeners effectively. Research directly addressing ML for baggage screening is sparse. This is likely due to two reasons: a) Lack of easily available datasets. b) Limited market with many gate keepers including government and equipment makers. The only rea-

sonably large scale luggage dataset we are aware of is the ALERT CT dataset from Northeastern University [Crawford et al., 2013]. Even that DHS sponsored dataset requires an NDA and a payment [Crawford et al., 2013]. There is a reasonable body of research from the medical ML area that can be adapted to jump start CT based baggage screening.

References

[Akcay et al., 2016] Akcay, S., Kundegorski, M. E., Devereux, M., and Breckon, T. P. (2016). Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1057–1061.

[Akcay et al., 2018] Akcay, S., Kundegorski, M. E., Willcocks, C. G., and Breckon, T. P. (2018). Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE Transactions on Information Forensics and Security*, 13(9):2203–2215.

[Bazrafkan et al., 2019] Bazrafkan, S., Van Nieuwenhove, V., Soons, J., De Beenhouwer, J., and Sijbers, J. (2019). Deep Learning Based Computed Tomography Whys and Wherefores. *arXiv e-prints*, page arXiv:1904.03908.

[Bouallagu et al., 2010] Bouallagu, N. M., Flitton, G. T., and Breckon, T. P. (2010). A classifier based approach for the detection of potential threats in ct based baggage screening. *2010 IEEE International Conference on Image Processing*, pages 1833–1836.

[Chetlur et al., 2014] Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., and Shelhamer, E. (2014). cudnn: Efficient primitives for deep learning. *CoRR*, abs/1410.0759.

[Crawford et al., 2013] Crawford, C., Martz, H., and Karl, W. C. (2013). Research and development of reconstruction advances in CT-based object detection systems - final report HSHQDC-12-J-00056. Dept. Homeland Security Center of Excellence, ALERT at Northeastern University.

[Gao et al., 2017] Gao, X. W., Hui, R., and Tian, Z. (2017). Classification of ct brain images based on deep learning networks. *Computer Methods and Programs in Biomedicine*, 138:49 – 56.

[Grady et al., 2012] Grady, L., Singh, V., Kohlberger, T., Alvino, C., and Bahlmann, C. (2012). Automatic segmentation of unknown objects, with application to baggage security. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision – ECCV 2012*, pages 430–444, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Han et al., 2018] Han, Y., Kang, J., and Ye, J. C. (2018). Deep learning reconstruction for 9-view dual energy CT baggage scanner. *CoRR*, abs/1801.01258.

[Heitz and Chechik, 2010] Heitz, G. and Chechik, G. (2010). Object separation in x-ray image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2093–2100.

[Martin et al., 2015] Martin, L., Tuysuzoglu, A., Karl, W. C., and Ishwar, P. (2015). Learning-based object identification and segmentation using dual-energy ct images for security. *IEEE Transactions on Image Processing*, 24(11):4069–4081.

[Maturana and Scherer, 2015] Maturana, D. and Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928.

[Megherbi et al., 2013] Megherbi, N., Breckon, T. P., and Flitton, G. T. (2013). Investigating existing medical CT segmentation techniques within automated baggage and package inspection. In Zamboni, R., Kajzar, F., Szep, A. A., Burgess, D., and Owen, G., editors, *Optics and Photonics for Counterterrorism, Crime Fighting and Defence IX; and Optical Materials and Biomaterials in Security and Defence Systems Technology X*, volume 8901, pages 198 – 206. International Society for Optics and Photonics, SPIE.

[Megherbi et al., 2012] Megherbi, N., Han, J., Breckon, T., and Flitton, G. (2012). A comparison of classification approaches for threat detection in ct based baggage screening. pages 3109–3112.

[Mery et al., 2015] Mery, D., Riffo, V., Zscherpel, U., Mondragon, G., Lillo, I., Zuccar, I., Lobel, H., and Carrasco, M. (2015). Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):1–12.

[Mery et al., 2017] Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J. M., and Banerjee, S. (2017). Modern computer vision techniques for x-ray testing in baggage inspection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):682–692.

[Mouton and Breckon, 2015] Mouton, A. and Breckon, T. P. (2015). A review of automated image understanding within 3d baggage computed tomography security screening. *Journal of X-ray science and technology*, 23 5:531–55.

[Mouton et al., 2014] Mouton, A., Breckon, T. P., Flitton, G. T., and Megherbi, N. (2014). 3d object classification in baggage computed tomography imagery using randomised clustering forests. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5202–5206.

[Nocentino and Rhodes, 2010] Nocentino, A. E. and Rhodes, P. J. (2010). Optimizing Memory Access on GPUs Using Morton Order Indexing. In *Proceedings of the 48th Annual Southeast Regional Conference*.

[Qi et al., 2017] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Qi et al., 2016] Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., and Guibas, L. (2016). Volumetric and multi-view cnns for object classification on 3d data. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

[Ren et al., 2015] Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497.

[Rogers et al., 2016] Rogers, T. W., Jaccard, N., Protonotarios, E. D., Ollier, J., Morton, E. J., and Griffin, L. D. (2016). Threat image projection (tip) into x-ray images of cargo containers for training humans and machines. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–7.

[Schmidt-Hackenberg et al., 2012] Schmidt-Hackenberg, L., Yousefi, M. R., and Breuel, T. M. (2012). Visual cortex inspired features for object detection in x-ray images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2573–2576.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Shelhamer et al., 2017] Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651.

[Singh and Singh, 2005] Singh, M. and Singh, S. (2005). Image enhancement optimization for hand-luggage screening at airports. In Singh, S., Singh, M., Apte, C., and Perner, P., editors, *Pattern Recognition and Image Analysis*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Socher et al., 2012] Socher, R., Huval, B., Bhat, B., Manning, C. D., and Ng, A. Y. (2012). Convolutional-Recursive Deep Learning for 3D Object Classification. In *Advances in Neural Information Processing Systems 25*.

[Turcsany et al., 2013] Turcsany, D., Mouton, A., and Breckon, T. P. (2013). Improving feature-based object recognition for x-ray baggage security screening using primed visualwords. In *2013 IEEE International Conference on Industrial Technology (ICIT)*, pages 1140–1145.

[Wu et al., 2015] Wu, Z., Song, S., Khosla, A., Zhang, L., Tang, X., and Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA.

[Yan et al., 2018] Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A. P., Bagheri, M., and Summers, R. M. (2018). Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.