Exploitation of Sensor Data using Artificial Intelligence for Battlefield Sensemaking

Valérie Lavigne valerie.lavigne@drdc-rddc.gc.ca
Marielle Mokhtari marielle.mokhtari@drdc-rddc.gc.ca
Mélanie Breton melanie.breton@drdc-rddc.gc.ca
Étienne Martineau etienne.martineau@drdc-rddc.gc.ca
Jonathan Fournier jonathan.fournier@drdc-rddc.gc.ca

Defence R&D Canada
2459, route de la Bravoure, Québec (QC)
CANADA

ABSTRACT

Although Artificial Intelligence (AI) techniques based on Machine Learning (ML) have shown a lot of progress in the recent years, leveraging these capabilities for military operations is still a challenging endeavour. At Defence Research and Development Canada (DRDC), we perform R&D activities aiming at leveraging ML for analyzing and making sense of a variety of sensor data. We use computer vision algorithms for analyzing Electro-Optical (EO) and Infrared (IR) Full Motion Video (FMV), 3D Light Detection and Ranging (LiDAR) and Wide Area Motion Imagery (WAMI) sensor data.

Academic AI research and datasets are often not adapted to military sensors and operational conditions. This leads to several challenges when transferring promising technologies to military applications. For instance, we need to handle noisy and blurry images, to process in real-time sensor modalities other than visible EO imagery, and to work with point clouds having different spatial resolutions.

In this paper, we discuss our work on creating military relevant labelled datasets, merging and aligning multiple training datasets, optimizing models for real-time analysis, and defining military relevant evaluation metrics.

GETTING LABELLED DATASETS FOR MODEL TRAINING

The first challenge that is faced when we want to train a model for object detection is the lack of labelled data. Academic models are often trained on large datasets such as ImageNet [1] and OpenImage [2], which both contain over 1 million images. For a dataset to be relevant for training our detection models, it should:

- Contain labelled instances of the objects of interest;
- Offer a perspective similar to the imagery it is intended for (ground-based vs. aerial perspective); and
- Have similar image characteristics (visible spectrum vs. infrared imagery, image degradation).

Large datasets that meet all of these requirements are not available. In order to create our training datasets, we decided to adopt a mixed approach. First, we combined multiple available open source datasets. Second, we performed labelling on our own sensor data to cover the visual aspects that were missing from the open source datasets.

Semantic Alignment for Leveraging Multiple Open Source Datasets

Our first approach was to leverage the open source datasets that are available. Merging multiple datasets comes with the following challenges:

- Multiple labelling formats are used by different datasets;
- Each dataset has its own label names and definitions; and finally,
- Some objects that are labelled in some datasets, but are present in other datasets but not labelled in the latter.

In order to handle the different file formats, we opted to create an intermediary labelling file format where we standardized all the labelling information, but kept the original label names. This allowed us to create configurations specific to each model to be trained. For each model, we specify which datasets should be included in the final training dataset and the desired label conversion to apply. This conversion states which labels from the original datasets should be kept and the mapping to our model object types. With this method, we can create a separate set of label files that is specific to each model. For now, the semantic alignment of the various labels is performed manually, but eventually, a measure of similarity based on detection confusion matrices and visual similarity [3] could be used to partially automate the label conversion process.

Labelling Sensor Data

The specific objects that we would like to detect with our models include military vehicles and equipment which are rarely present in open source labelled datasets. In addition, a vast majority of the datasets available offer ground-based visible imagery. Datasets that contain aerial imagery and other sensor modalities such as infrared can be found, but they are relatively modest and they usually do not combine all the aspects necessary to train the desired model. Although, it has been shown that Generative Adversarial Networks (GANs) can be trained to convert imagery between sensor modalities [4], this was performed from ground-based infrared to visible imagery and the performance is always lower than what we can obtain with a model directly trained on the source modality. Moreover, there is generally not enough labelled data to create GANs for each conversion that we need. This meant that we add to collect our own sensor data and label the objects of interest ourselves.

Labelling objects in images is a tedious and mind-numbing activity. Many open source datasets were created through crowdsourcing sites. We did not adopt this approach because some objects require military expertise to identify correctly.

As we were labelling video data, we speed up the labelling by copying the labels from the previous frame to the next, so that the user only needs to slightly adjust the boxes' localization and size. We were helped by two full-time military personnel for 4 weeks. During that time, around 40 000 objects bounding boxes were labelled in more than 120 000 images. For future labelling sessions, we intend to explore active learning to train a model in parallel with the labelling activities. This model could then suggest labels that users only need to confirm the localization, instead of having to draw all the object bounding boxes.

DEEP LEARNING PROCESSING PERFORMANCE

Gathering training, validation and test datasets, and training models are only the first steps. After that, we need to be able to apply those models in real time to sensor feeds. Both WAMI and FMV feeds are very large and require the development of optimization strategies, if we want the object detection, localization and classification

to happen in real time.

Model Architecture

Selecting the deep neural network architecture has a significant impact on model execution speed. Early on, we decided to use bounding boxes. The model architecture we selected was YOLOv3 [5]. Part of its high-speed performance comes from performing object detection, localization and classification in a single pass, whereas some other models do this in two passes. Some models that achieve better accuracy, but they cannot process information in real time. For most operational contexts, it is better to have imperfect timely detections rather than perfect detections too late.

Tiling Strategy and Object Tracking

To keep YOLOv3 fast, we used the medium detector resolution which is 416x416 pixels. The default approach is to reformat the input image to fit this smaller detector resolution. This causes small objects to sometimes become so small that they cannot be detected anymore, especially in an aerial operating context. In order to ensure that we exploited the full resolution of the sensor input, we opted for a tiling strategy, where we moved our model detection to different parts of the image for each frame. As we are analyzing video feeds, we can leverage temporal continuity between frames for the other parts of the image that are not been analyzed. Objects detected in the image are tracked between frames until the detection model comes back to the area where that object is and confirm its presence again. This strategy has allowed us to achieve real-time analysis (average of 25 fps) on a video feed on a single laptop equipped with a NVIDIA P5000 GPU card.

Detection and Tracking in WAMI Imagery

Images collected with a WAMI camera system proved to be challenging to process using available ML approaches. The frame rate of those systems is very low (e.g. 2-3 Hz) and the image resolution is generally in the order of 0.5 m to 1 m, depending of the selected altitude. In addition, images are generally quite large (e.g. 10,000 x 10,000 pixels) and they need to be processed using a tiling approach as described in the previous subsection.

Traditional ML object detection techniques have been proven to have a low performance with WAMI imagery. This is due to the fact that the number of pixels for each target is generally very low and the contrast of the target with respect to the background is poor. For this reason, hybrid approaches have started to emerge where motion and appearance change detection techniques are combined with a convolutional architecture to perform reliable detections [6]. This type of approach is what DRDC plans to implement into its current WAMI image processing chain.

It is envisioned that good quality datasets will be required to assess the performance of the implemented approaches. As of now, there are a limited number of annotated WAMI datasets openly available. To address this issue, it is planned to rely on manually annotated datasets collected by DRDC and its partners, as well as datasets generated by DRDC's airborne simulation software tools.

Point Cloud Classification

LiDAR laser scanners are the most common instruments used to collect geographic Point Cloud Data (PCD). Even if point cloud classification is an active research study area, the number of tools available for this classification is limited. Popular tools require a paying license in order to be used, and free of use tools are quite

rare. Also, these tools do not necessarily produce the best quality results and most of the time, the choice of labels are limited to ground, building and high vegetation.

Due to their different spatial resolution and number of points, Aerial Laser Scanned (ALS) and Terrestrial Laser Scanned (TLS) need to be classified separately. For example, objects of interest such as people, panel signs and bicycles cannot be extracted from ALS PCD, due to their small spatial resolution, but they can be extracted from TLS PCD. In ALS point clouds, the building roofs are being extracted, but in TLS it is often hard to extract roofs due to their field of view and range limits. ALS PCD can be classified using classic ML algorithms in less time and with fairly good results. On the other hand, TLS PCDs usually require more learning to segment labels and Deep Learning (DL) algorithms work best on this type of data.

We first developed two classification models in order to automatically classify ALS point clouds in urban and rural areas, respectively. ALS classification can be done with standard ML algorithm such as Random Forest. Classification results on ALS point clouds show a 90% global accuracy for the following classes: ground, vegetation, roof, façade, and car.

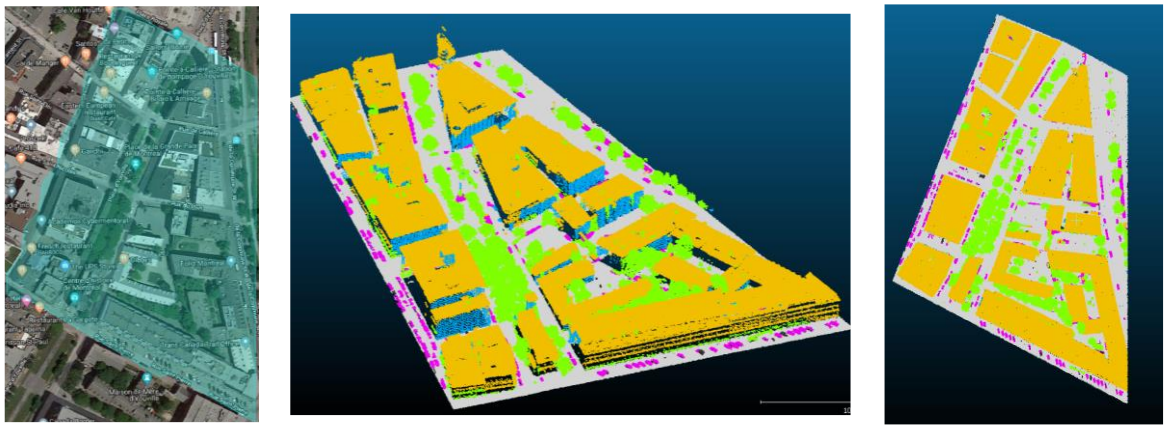

Figure 2: ALS point cloud classification in an urban area (Old Port Area, Montreal). Ground (grey), Vegetation (green), Roof (yellow), Façade (blue), Car (pink), and Unclassified (black)

Second, we investigated automatic classification for TLS point clouds using DL algorithms. Classification results on TLS point clouds show a 95% global accuracy for the following classes: man-made terrain, natural terrain, high vegetation, low vegetation, building, hardscape, scanning artefacts and car. We trained two models for automatic classification of TLS point clouds, one using the colours (Red-Green-Blue) and one without. Out of the two, the model with colour produced the best results.

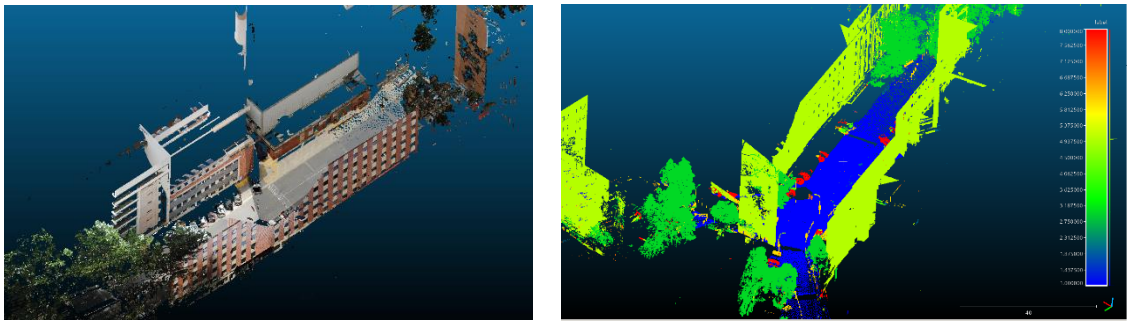

Figure 3: TLS point cloud classification (Adelaide Royal Hospital, Adelaide, Australia) – Man-Made Terrain (blue), Natural Terrain (teal), High Vegetation (dark green), Low Vegetation (green), Building (light green), Hardscape (yellow), Scanning Artefacts (orange) and Car (red).

EXPLOITING AI GENERATED INFORMATION

The exploitation of AI-generated content for sense making is highly application problem specific. We believe this capability can enhance situational awareness. In some cases, alerting strategies can be devised in order to keep an analyst informed of changes in a situation. Automated object counting may produce interesting information for intelligence analysis. Using automated detection and classification can allow operators and analysts to monitor multiple feeds at the same time. Analyzing and tagging large collection of video imagery can allow an analyst to find relevant imagery faster, and this is a tool that we intend to develop.

PERFORMANCE EVALUATION

The mean Average Precision (mAP) metrics [7] is well known in the machine learning community. It is the go-to metrics for evaluating object localization algorithms, and it works well for comparing new models with known models. However, selecting an algorithm to be deployed based on this sole criterion has several shortcomings.

First, the mAP usually evaluates localization-detection success on a specific dataset. It does not represent the detector's performance "in-the-wild", where the sensor might produce noisy, rotated, compressed, and/or blurry images. Also, it does not differentiate between the types of objects, although that could be an important feature if we want to favour performance on specific mission-relevant object categories. Finally, it penalizes highly an algorithm when an object is detected more than once. This hypothesis works well when the intended task relates to object counting within an image. However, if the task requires rare object detection or simply raising a flag when an anomaly is detected, the mAP might discard good algorithms.

For example, let's look at the model output pictured in Figure 4. The algorithm performs poorly at counting each instance of the vehicles class. However, if the task was to raise an alarm in real time, with a noisy and heavily compressed image, then this model might outperform other models.

Figure 4: The bounding box surround two vehicles at a time. For an object counting task perspective, this algorithm performs poorly, however, for a detection task, this algorithm performs correctly.

Problem Specific Metrics

We are in the process of developing specific metrics that include these three criteria: classification accuracy, localization accuracy, and scaling of the bounding box. To that, we add robustness to noise, sensor orientation, and image compression as well as inference time. In the context of Intelligence, Surveillance and Reconnaissance (ISR), these criteria should be weighted depending on the task.

CONCLUSION

In this paper, we discussed the challenges of applying ML-based AI to military problem domains and shared some of our lessons learned in working on different projects involving various types of imagery, as well as point cloud data. There is a need for developing common practices between NATO nations to ensure that we can share and leverage datasets, and that we have military relevant metrics for comparing model performances on common sets of operational problems.

ACKNOWLEDGEMENTS

The authors would like to thank Guillaume Gagné for his involvement in the labelling activity, Maxime Dionne for his technical advices, and Frédérick Lafrance for his work on the point cloud classification prototypes.

REFERENCES

[1] ImageNet Summary and Statistics, ImageNet. http://image-net.org/about-stats. Retrieved 4 September 2019.

[2] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., & Ferrari, V. (2018). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. ArXiv, abs/1811.00982.

[3] Brust, C., & Denzler, J. (2018). Not just a matter of semantics: the relationship between visual similarity and semantic similarity. ArXiv, abs/1811.07120.

[4] Berg, A., Ahlberg, J., & Felsberg, M. (2018). Generating Visible Spectrum Images from Thermal Infrared.

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1224-122409.

[5] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. ArXiv, abs/1804.02767.

[6] LaLonde R., Zhang D., Shah, M. Fully Convolutional Deep Neural Networks for Persistent Multi-Frame Multi-Object Detection in Wide Area Aerial Videos. ArXiv:1704.02694v2

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015), ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV), 115(3). Challenge website [image-net.org/challenges/LSVRC/2017].