Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Firearms On Twitter: A Novel Object Detection Pipeline

Ryan Harvey[1], **Rémi Lebret**[2], **Stéphane Massonnet**[2], **Karl Aberer**[2], **Gianluca Demartini**[1]

[1] The University of Queensland, Australia
[2] École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

Social media is an important source of real-time imagery concerning world events. One subset of social media posts which may be of particular interest are those featuring firearms. These posts can give insight into weapon movements, troop activity and civilian safety. Object detection tools offer important opportunities for insight into these images. Unfortunately, these images can be visually complex, poorly lit and generally challenging for object detection models. We present an analysis of existing gun detection datasets, and find that these datasets to not effectively address the challenge of gun detection on real-life images. Following this, we present a novel object detection pipeline. We train our pipeline on a number of datasets including one created for this investigation made up of Twitter images of the Russo-Ukrainian War. We compare the performance of our model as trained on the different datasets to baseline numbers provided by original authors as well as a YOLO v5 benchmark. We find that our model outperforms the state-of-the-art benchmarks on contextually rich, real-life-derived imagery of firearms.

Introduction

At present, Twitter features millions of images tagged as relevant to the Russo-Ukrainian war. Although only a very small portion of these images feature firearms, this small subset may give important information about the conflict. Traditional object detection methodology suggests a model ought to be trained on a representative dataset and then used to perform the task. Here, however this approach is rendered ineffective by two factors: the difficulty of firearm detection and the difference between the images in gun detection datasets and Twitter images. Firearm detection is challenging because guns are often intentionally obscured, held at various angles, camouflaged and include considerable intra-class variation. In contrast, the images in some firearm detection datasets are often unobscured, well lit, and display the subject firearm clearly in front of view. This means that object detection models trained on these pre-existing datasets tend to under-perform on real-world-derived social media images and it is not possible to address the current challenge using traditional methods. We address this challenge by proposing a novel few-shot object detection pipeline based on the VINVL visual language network (Zhang et al. 2021). Our approach takes advantage of a thoroughly pre-trained visual language model (VINVL) and uses a relatively simple dense network to transform visual language annotations into object detection predictions with a small number of fine-tuning samples. We find that our approach significantly outperforms a YOLO v5 pre-trained and fine-tuned comparison baseline model when considering images derived from Twitter. We also find that our approach outperforms the baseline set by the initial authors of the YouTube Gun Detection Dataset (Gu, Liao, and Qin 2022). We conclude that our approach works well on contextually complex firearm detection challenges.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Background and Related Work

Firearms feature heavily in any study of propaganda, military intelligence or armed violence. In this section we discuss prior work concerning firearms and social media. It has been found that Twitter can make a good source of continuous real time information on firearms activity in the US (Singh et al. 2022). This study examined the link between posting behavior and gun violence to see how the social media sphere was affected by and could be used to understand gun violence incidents. Another study showed that social media could be used as an effective means for safe firearm storage messaging, especially to younger generations (Lam et al. 2021). If we consider firearms in the context of war as opposed to city violence, we find a number of interesting perspectives. First, firearms have been studied in the context of propaganda (Klausen 2015; Gates and Podder 2015; Silvestri 2014). These studies examined the recruiting value of firearm imagery online, both for the US military and for Jihadists. They found that images of soldiers posing with their weapons were a powerful propaganda tool specifically targeted at people back home as opposed to those on the battlefield. In this way, we find imagery of war has an impact not just in the war zone but within home countries and foreign countries at peace. Within the war zone, the importance of firearms in social media is arguably even greater. A study of social media in the Russo-Ukrainian war found that online suppression of certain websites could be an effective way to curb misinformation and propaganda within a country at war (Golovchenko 2022). Finally, it is important to consider the impact of firearm imagery on the very civilians likely to

collect said imagery (Saugmann 2019). These images are a double edged sword in some cases, providing important intelligence for intelligence agencies, while putting the civilian population at serious risk of retaliation from those photographed.

Review of Firearm Detection Datasets

With the importance of firearms content online now established, we can turn our attention to the challenge of detection. (Debnath and Bhowmik 2021) presents a series of criteria which can be used to discuss this challenge. Their criteria are: Intra-class variation, scene complexity, occlusion, variety viewing angles, changing lighting and blur. Here, we discuss several different datasets using these criteria. We aim to give a representative cross-section of available data for a firearm detection endeavor. (Debnath and Bhowmik 2021; Yadav, Gupta, and Sharma 2022; Mahajan and Padha 2018).

The first dataset we discuss was originally presented by Pérez-Hernández et al. (2020), the small objects handled similarly to a weapon dataset or SOHA. The firearm images are part of a larger dataset featuring pistols, knives, smartphones, money, purses and cards. As shown in Figure 1a, these images appear relatively easy when considered using the criteria from (Debnath and Bhowmik 2021). Specifically, these firearms are clearly shown in view, they are generally well lit or lack complex shadows and importantly, this dataset features only pistols which dramatically reduces intra-class variation. We conclude that this dataset is unlikely to be of great value for a task involving social media images of war.

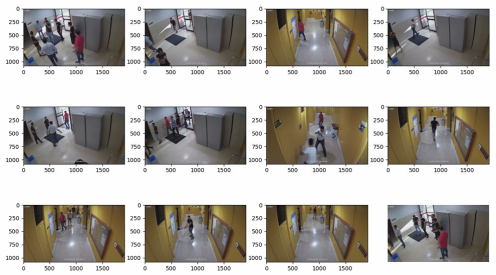
The second dataset we discuss here was assembled by Qi et al. (2021) and is by far the largest we consider. It features 51,882 images of firearms including images from movies, CCTV, and stock images. It includes images produced by the authors, as well as subsets from Lim et al. (2019), Olmos, Tabik, and Herrera (2018) and Sultani, Chen, and Shah (2018). Given the sheer number of images in this dataset, this dataset does feature a wide range of different lighting, occlusion and shadow effects. Mixed into this dataset are however many images with clear lighting, low occlusion and good framing. Furthermore, the datasets features CCTV style images which have low value for this task because the weapons are often completely concealed in CCTV footage. We conclude this dataset will have mixed relevance for a task focusing social media images of war.

The third dataset is the *YouTube Gun Detection Dataset (YGDD)* (Figure 1c) (Gu, Liao, and Qin 2022). The images in this dataset model well the intra-class variation challenge face in images of real-life war. Furthermore, the variety of different settings, both outside and inside give range to the lighting and shadow in the dataset. A drawback of this dataset however, is that due to the very nature of YouTube firearm content, the firearms are being demonstrated and discussed. In this way, guns are often clearly visible, and may not represent the occlusion or viewing angles typical of real-life images. We conclude this dataset has the best utility of the datasets discussed here when applied to real-life social media images of war.

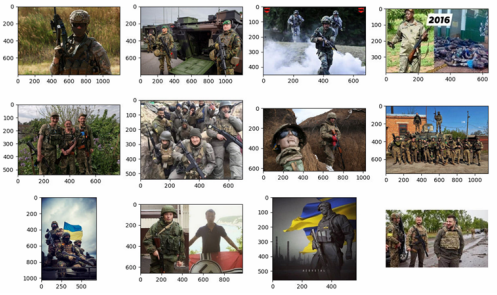
(a) SOHA (Pérez-Hernández et al. 2020)

(b) (Qi et al. 2021)

(c) YouTube GDD (Gu, Liao, and Qin 2022)



(d) MADS (González et al. 2020)



(e) Our Twitter Gun Dataset (TGDS)

Figure 1: Selection of images from the datasets

The final pre-existing dataset we consider is known as the *Mock Attack Dataset (MADS)* (Figure 1d), and was created to aid in the development of CCTV-based firearm detection algorithms (González et al. 2020). Due to the focus on CCTV, these images have very different characteristics to those on social media. The firearms in this dataset are often heavily concealed and tend to take up only a very small number of pixels. In this way, the detection problem becomes more one of inference where the presence of a firearm is detected through observing gestures, or attempting to overcome concealment. We conclude this dataset models significantly different challenges to those primarily relevant to social media images of war.

Novel Twitter Gun Dataset

As we have now seen, many existing datasets do not effectively model the challenges of social media images of war. We present here the novel Twitter Gun Dataset (TGDS) (Figure 1e). We argue that these images present an object detection challenge not typical in the other datasets. First, the firearms are often well occluded behind other elements of the image. Second, scenes vary greatly in lighting, distance to the target object and backdrop. Some images are taken in an urban setting and some are in a forest setting. Finally, we argue that because the firearms are not the primary subject of the images the detection challenge is greater than that of YGDD. It is for these reasons the the TGDS has been proposed as a useful tool in researching firearms in social media. The dataset was created in three steps. First the query *ukraine has:media* was performed on the Twitter Stream API. This resulted in a set of approximately 600,000 images. The query was performed in June 2022. Following this, duplicates were removed and obvious negatives such as internet memes were excluded. Finally, the images were manually filtered by the authors to find only those images featuring firearms. Finally, the authors drew bounding box annotations on all visible firearms in each image.

VINOD A Novel Object Detection Network

As we have shown, the problem of detecting firearms on social media images of war is a challenging one. The visual complexity of the images means the solution will need to work effectively on very complex images, while also training readily on a small number of images. We present the VINOD object detection pipeline. Our pipeline is built on the VINVL visual language model (Zhang et al. 2021). This model was created in an effort to improve the visual features generated by a visual language model as improvements in this area were demonstrated to improve visual language performance. The model is trained on four different object detection corpus datasets: COCO, OpenImagesV5 (OI), Objects365V1, and Visual Genome (VG). In total this constitutes 2.48 million images and 1848 classes in the training set. We chose this model for it's enormous corpus of pretraining, anticipating the model will perform well on unusual and challenging object classes such as firearms.

We prepare our pipeline in three phases. In phase one the ResNeXt-152 C4 architecture VINVL backbone breaks down every image into a number of sub-images demarcated by bounding boxes. Each of these sub-images is likely to contain a distinct object from the original image. Each of these sub-images is given an information rich vector by the pre-trained VINVL backbone. It is important to note that these vectors are rich in the sense that they contain a great deal of information about the subject image. In the original VINVL, these feature vectors were used as the basis for scene description and other visual language tasks.

In phase two, each proposed bounding box is compared to the target annotations on the parent image. Each proposed bounding box is given an intersection-over-union (IOU) score. In the case there are multiple firearms in the parent image, a given proposed bounding box will be allocated the highest IOU score given by the target bounding boxes. To clarify, a proposed bounding box is thus evaluated according to the firearm with which it best matches. Phase two concludes with a number of information rich vectors and their corresponding IOU scores according to the target firearms in the parent image. In many cases the IOU score will be 0, but other objects, such as gun barrels, scopes or hands will have intermediate IOU scores.

In phase 3 a dense neural network is trained to predict the IOU score using the rich vectors as input. In this way, the dense network uses the visual character as captured by the rich vector of a proposed bounding box/ sub-image to predict whether and to what extent that sub-image contains a firearm. The proposed VINOD pipeline is described in Figure 2.

This approach offers very efficient training of the dense network because for each parent image fed into the VINVL backbone, 20-30 bounding boxes and sub-images are proposed. Each one of these sub-images then passes through phase two, and is fed into the dense network. In this way, each human-given annotation of a parent image can deliver 20-30 training instances for the dense neural network. Furthermore, the approach takes full advantage of the information available by modelling the presence/absence of a firearm as a regression task. In this way, candidate bounding boxes/ sub-images do not need to be manually categorised into firearm or not firearm. Instead, they are characterised according to their IOU score. This means imperfect matches can be used in training the dense network. In contrast to established object detection methodologies such as YOLO, which fine-tune a pre-trained network by passing the bounding box, and image through the whole network, our approach only fine-tunes a smaller network at the end of the pipeline. This, in combination with the data efficiency measures outlined above means our approach is able to deliver benchmark-beating performance on very visually complex images with a very small training set.

The dense network has 2048 neurons on the input layer and two hidden layers of 1024 and 256 neurons. The output layer has a single neuron with a sigmoid function to assist the regression task. Two dropout layers are used with 10 percent dropout. These are between the first and second layers and also the second and third layers. All layers use a RELU activation function. MSE loss is used with SGD optimization. The learning rate is 0.01 with momentum 0.95. During

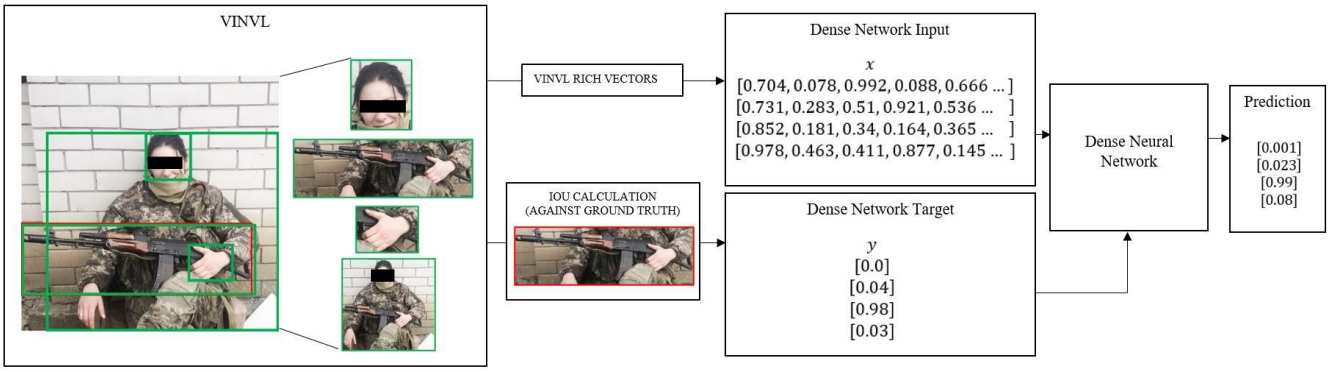


Figure 2: A schematic view of the proposed VINOD pipeline

training, the learning rate is decreased by 20 percent every 8 epochs. The dense network is trained over 20 epochs.

Demonstration and Results

We train and test our object detection pipeline on each of the datasets presented above. The training methodology followed the structure discussed above. The training set constituted 80 percent of the images in each respective set, with testing set making up the remaining 20 percent. All sets were randomly sampled. All scores are Average Precision over different IoU thresholds from 0.5 to 0.95, or otherwise called AP@[0.5:0.95] - the primary challenge metric of the COCO dataset.

Dataset	Train size	Publisher Benchmark	YOLO V5	VINOD
Qi et al.	41505		64.1	**81.9**
MADS	4000	14.6*	**40.9**	9.7
SOHA	2604		**82.8**	81.4
TGDS	400		52.6	**72.3**
YGDD	4000	52.1	54.3	**78.9**

Table 1: Comparison of VINOD performance based on AP@[0.5:0.95] scores. *Results for camera 5 only.

Our experimental results show that our VINOD method outperforms YOLO on challenging datasets. Specifically, we see that VINOD achieves an AP of 72.3 on the Twitter Gun Dataset, as opposed to a performance of 52.6 by YOLO v5 on the same data. This represents a 37 percent improvement over YOLO. VINOD delivers similarly excellent performance on the YGDD. VINOD also performs well on the Qi et al. dataset. The common factor between the three datasets on which VINOD outperformed YOLO is scene complexity and general image difficulty. As discussed above, each of these three datasets heavily feature images from real life and share the challenges of viewing angle, occlusion, lighting and blur. We have thus shown that VINOD is very well suited to complex object detection challenges. Furthermore, VINOD has been demonstrated to perform well on the very small TGDS dataset, showing it does not need a particularly large fine-tuning set to work well. VINOD does not outperform YOLO on all tests however, in the case of the MADS dataset, this is because the bounding-box creation step is not tuned to perform well on firearms with a very small number of pixels. In the case of the SOHA dataset, the difference in performance is very small. It is suspected that because the images in the SOHA dataset are not particularly challenging, YOLO v5 can be effectively trained without the need for the more complex VINOD.

Conclusion

We have presented an investigation into existing firearm detection datasets, and shown that images therein tend to lack the complexity of real-life images found on social media. Consequently, we have presented a novel object detection pipeline trained on images from Twitter. We have demonstrated this pipeline and shown it out-performs existing benchmarks on contextually complex images and small datasets.

Ethical Statement

It is important to discuss the potential broader impacts of the current work. On the positive side, the current work can help accelerate and support understandings of ongoing combat situations. More broadly, the techniques presented here can potentially be employed on any small dataset of objects poorly represented by existing object detection datasets. On the negative side, if the techniques fall into the wrong hands, they could aid bad actors in tracking down the civilians responsible for posting images of particular combatants. To limit potential negative impacts on social media users, the Twitter Gun Dataset will not be shared. While the images were collected anonymously it is possible to reverse image search or otherwise track an image back to an account or individual and we would prefer to avoid this possibility.

References

Debnath, R.; and Bhowmik, M. K. 2021. A comprehensive survey on computer vision based concepts, methodologies, analysis and applications for automatic gun/knife detection.

Journal of Visual Communication and Image Representation, 78: 103165.

Gates, S.; and Podder, S. 2015. Social media, recruitment, allegiance and the Islamic State. *Perspectives on Terrorism*, 9(4): 107–116.

Golovchenko, Y. 2022. Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media. *The Journal of Politics*, 84(2): 639–654.

González, J. L. S.; Zaccaro, C.; Álvarez-García, J. A.; Morillo, L. M. S.; and Caparrini, F. S. 2020. Real-time gun detection in CCTV: An open problem. *Neural networks*, 132: 297–308.

Gu, Y.; Liao, X.; and Qin, X. 2022. YouTube-GDD: A challenging gun detection dataset with rich contextual information. *arXiv preprint arXiv:2203.04129*.

Klausen, J. 2015. Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism*, 38(1): 1–22.

Lam, E.; Moreno, M.; Bennett, E.; Rowhani-Rahbar, A.; et al. 2021. Receptiveness and responsiveness toward using social media for safe firearm storage outreach: mixed methods study. *Journal of medical internet research*, 23(6): e24458.

Lim, J.; Al Jobayer, M. I.; Baskaran, V. M.; Lim, J. M.; Wong, K.; and See, J. 2019. Gun detection in surveillance videos using deep neural networks. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1998–2002. IEEE.

Mahajan, R.; and Padha, D. 2018. Detection of concealed weapons using image processing techniques: A review. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 375–378. IEEE.

Olmos, R.; Tabik, S.; and Herrera, F. 2018. Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275: 66–72.

Pérez-Hernández, F.; Tabik, S.; Lamas, A.; Olmos, R.; Fujita, H.; and Herrera, F. 2020. Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. *Knowledge-Based Systems*, 194: 105590.

Qi, D.; Tan, W.; Liu, Z.; Yao, Q.; and Liu, J. 2021. A Dataset and System for Real-Time Gun Detection in Surveillance Video Using Deep Learning. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 667–672. IEEE.

Saugmann, R. 2019. The civilian's visual security paradox: how open source intelligence practices create insecurity for civilians in warzones. *Intelligence and national security*, 34(3): 344–361.

Silvestri, L. 2014. Shiny happy people holding guns: 21st-century images of war. *Visual Communication Quarterly*, 21(2): 106–118.

Singh, L.; Gresenz, C. R.; Wang, Y.; Hu, S.; et al. 2022. Assessing social media data as a resource for firearm research: analysis of tweets pertaining to firearm deaths. *Journal of medical internet research*, 24(8): e38319.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Yadav, P.; Gupta, N.; and Sharma, P. K. 2022. A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods. *Expert Systems with Applications*, 118698.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.