Witnessing atrocities: quantifying villages destruction in Darfur with crowdsourcing and transfer learning

Julien Cornebise[*]
University College London

Daniel Worrall[*]
Philips Lab
U. of Amsterdam

Micah Farfour **Milena Marin**
Amnesty International

Abstract

We report the first, to the best of our knowledge, hand-in-hand collaboration between human rights activists and machine learners, leveraging crowdsourcing and transfer learning to automatically analyze satellite imagery, at country-wide scale, for conflict reporting. This will allow for large-scale quantitative evidence gathering, beyond closed borders, usable in conjunction with survivors testimonials to demand justice and accountability.

1 Introduction

Amnesty International is a world's leading human rights organization, campaigning against injustice and inequality everywhere. For more than a decade it has gathered evidence of the ongoing destruction of civilian villages in Darfur by the Sudanese government, see Fig. 1 (top row), starting with its Eyes On Darfur campaign in 2007, and leading to the publication of a scathing evidence report in 2016 [1]. Yet this was the work of a small team of highly-trained expert analysts, on a very precise sub-area. Its Decode Darfur campaign, launched late 2016, has a double goal: use crowdsourcing to assess this conflict at a larger scale, to bring to international light the brutal impact of this conflict; and foster public engagement by involving the public beyond simply clicking a petition.

A chance encounter between the authors led to extend it beyond crowdsourcing. This seems to be among the first use of machine learning for Human Rights led by a Non-Governmental Organization. This partnership between domain experts and technical experts, with a long-view incorporating both human skills, machine learning, and engineering, can become a blueprint for social impact via concrete tools that will empower activists to scale up their results. For example, like here, gathering large scale quantitative evidence beyond closed borders, to demand justice and accountability.

[*] Equal contribution. Contact: julien@cornebise.com. Work conducted while DW was at UCL.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

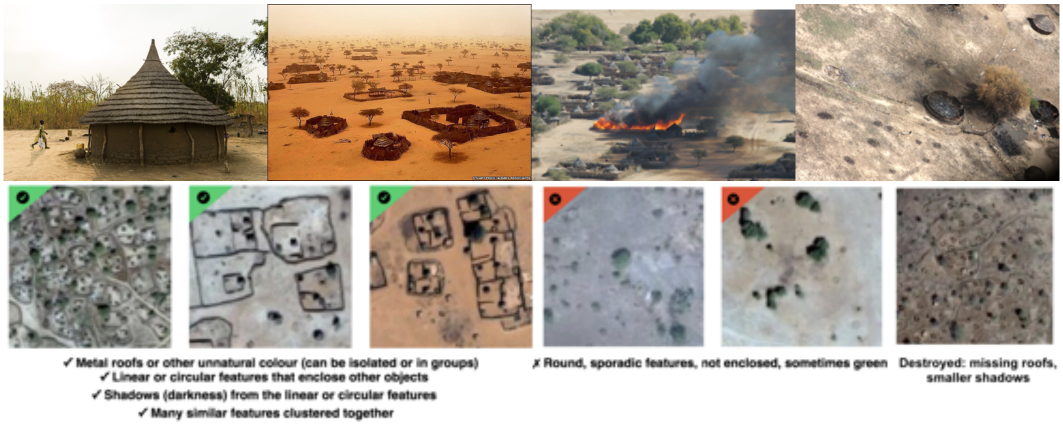

✓ Metal roofs or other unnatural colour (can be isolated or in groups)
✓ Linear or circular features that enclose other objects
✓ Shadows (darkness) from the linear or circular features
✓ Many similar features clustered together
✗ Round, sporadic features, not enclosed, sometimes green
Destroyed: missing roofs, smaller shadows

Figure 1: TOP, Left to Right: i) A typical tukul, made of straw, mud and wood. ii) A village of tukuls in the desert, surrounded by straw fences; iii) Ongoing arson; iv) Remnants of a burned tukul, with roof missing and walls burned to half-height. BOTTOM: Tutorial provided to the labelers.

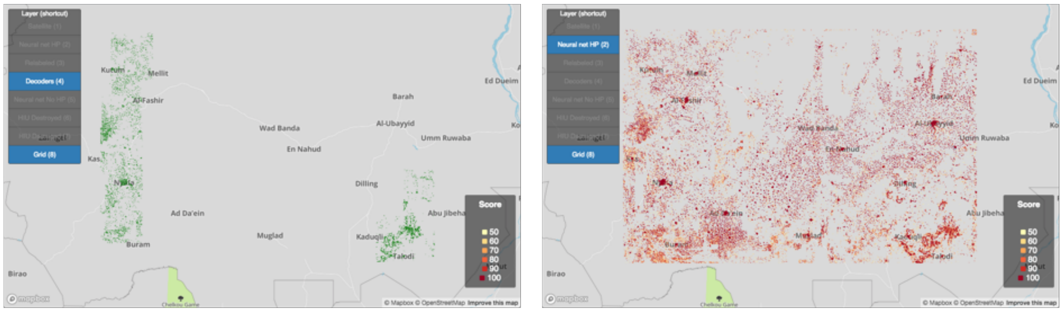


Figure 2: LEFT, in green: tiles labeled by the decoders as having human presence (majority vote). RIGHT, tiles with human presence as inferred by our neural net – color indicate classifier's score, from 0.5 (yellow) to 1 (red). Screenshots from our visualization web interface.

Most Darfur villages are not on a map. We used public micro-tasking and readily available satellite imagery to locate villages over time in selected areas – see Fig. 1 (bottom row). We then trained a classifier on this crowdsourced data to scan for other human habitations at country-wide scale, half a million square kilometers, see Fig. 2. This first success allowed a more narrow collection by experts focusing exclusively on destroyed villages, which we then again generalized by training a classifier.

The use of satellite imagery for aid and humanitarian purposes is growing. Project Sentinel was already focusing on Darfur, albeit without automation. Its postmortem [2] illustrates why technology alone is not the solution to geopolitical problems, and why domain expertise is required for effective influence. The Harvard Humanitarian Initiative also studied Tukul destruction [3], even starting work on an automated Tukul detector [4, p.15], sadly without a clear conclusion ever coming to light. In parallel, deep learning starts being used for accurate exploitation of satellite information at large scale. E.g. [5] refines existing poverty mapping efforts. The ideas are definitely in the air. We demonstrate here that their success can come through the combination of domain experts, machine learning experts, and proper engineering. Citizen science efforts such as pioneered by Galaxy Zoo [6], Cancer Research UK [7], or the Zooniverse platform [e.g. 8] can also be incorporated.

2 Finding human presence

Crowdsourcing to find human structures: a deluge of goodwill. The labeling of tiles by crowdsourcing was designed before the machine learning component. As we will see, it was particularly successful, thanks to the extra care put into the user experience: cross-platform gamified interface, and, most importantly, a one-click way to seek advice on any image from a very active forum. This "closing of the loop" provided feedback to the graders and was backed by a sustained initial effort from immediate answers by the expert[2]. By naturally creating a series of the most confusing examples clarified, the effort became self-sustaining when now-experienced graders started answering the new graders. It also created a sense of community, reinforcing the massive engagement.

Over little more than three weeks, $28,600$ volunteers connected at least once, from 147 countries, most from Sweden and Netherlands where Amnesty ran an advertisement campaign. Their 13 million grades covered 2.6 million satellite tiles of approximately 100x100 meters each. $3,712$ graders made the effort of creating a named account, the most active of which single-handedly provided $305,811$ grades over 71 hours. The most patient grader spent 174 hours. These registered users spent on the platform the equivalent of 3.2 working years of a full-time grader.

How good is the crowd? Quality analysis: To assess the quality of the crowd's data, an expert analyst labeled $4,187$ tiles to serve as ground truth. Most tiles received 4 or 5 labels from independent graders – with a few spiking to 60 labels during server overload. Fig. 3 shows the remarkably good precision/recall curve of the crowd's vote compared to the expert.

Another frequent question in crowdsourcing is "How many graders do we need per image?" Studying precision and recall on the subsets of tiles graded by exactly $n = 1, 2, \ldots$ graders is too high-variance: most tiles had 4 graders. We instead use bootstrap: for $10,000$ samples, we resample each tile's

[2] This is a key difference with other, less successful Citizen Science engagements. A leading British charity (remaining unnamed here as it should not be discredited for trying to pioneer new methods) sadly had to close its Citizen Science unit due to poor crowd data quality; it did not have any such feedback.

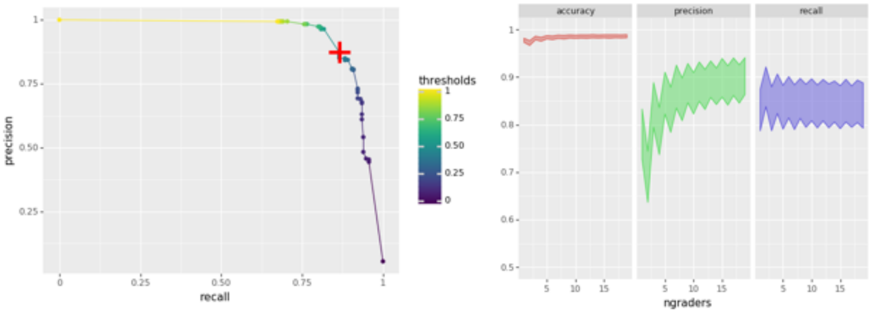
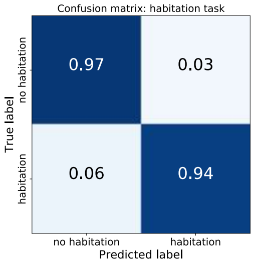


Figure 3: LEFT: Precision/recall curve for the crowd. Red cross marks simple majority vote, threshold of 0.5. RIGHT: number of graders per tile vs. precision and recall. Shaded area is 95% empirical highest density region. The accuracy is artificially high due to class imbalance. Precision increases but plateaus quickly. Recall seems independent of the number of graders, suggesting potential benefit for a more exhaustive tutorial. Odd/even oscillations show the influence of arbitrary tie-breaking in a highly class-imbalanced setting.

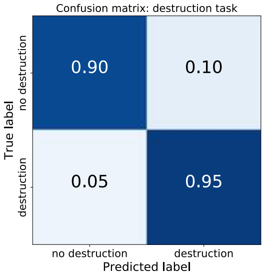
grades with replacement n times. Using majority voting (threshold 0.5) on each resulting bootstrapped sample, we obtain Fig. 3. Now that the partnership between domain experts and machine learners is established, the next study will benefit from such a study on a small pilot sample of tasks and graders.



Automation and results: To predict human presence, we finetuned an ImageNet-pretrained ResNet-50 [9] on a class-balanced subset of $132,000$ crowdsourced labels. We found that we did not need a deeper network than 50 layers. To combat overfitting, we only needed to use modest amounts of random flipping, rotation, jitter, and per-image zero centering of the pixel values. We did need to be careful, however, that our random rotations did not rotate small dwellings/tukuls out of the receptive field of the CNN. Symmetric padding followed by rotation then cropping back to the original size improved our eventual classed-balance test error from $\sim 6\%$ to $\sim 4\%$. Augmentation was performed online at each training step. Data ablation analyses showed that we only need $25,000$ labels for training, above which we found no further performance gains. Central to the success of finetuning on our small–medium dataset was a thorough hyper/meta-parameter search. We used 32 iterations of an off-the-shelf sequential Gaussian process-based Bayesian hyperparameter optimizer [10] in an outer loop, searching over batch size, image dimension, learning rate and its schedule, momentum, and jitter range.

The result of this initial training round was an accuracy of 96% on a balanced dataset, see confusion matrix above. This was deemed sufficient as a first model, but could be refined by multiscale features accounting for classification of neighbouring tiles. We also performed interpretable error detection through saliency maps, as described in Appendix A.

3 Detecting destruction



Targeted expert labeling: To detect destruction, we undertook a much smaller and more targeted labeling effort, using only three experts guided by a previous study of past damages[11] and the custom-designed visualization/labeling tool of Section 4. Collecting purely for machine learning purpose, they only labeled tiles containing destroyed villages and neighbouring tiles without destruction but on similar terrain, leaving all other tiles unlabeled, to make for a balanced dataset by design. They gathered 8841 training samples, 1104 validation samples and 1108 test samples.

Automation and results: We cast the detection of sites of destruction into a multi-task binary classification problem: `destruction` vs. `no destruction`, and `habitation` vs. `no habitation`. We finetuned the weights of the habitation detection network. The final layer was split into two for each of the two classification tasks. For extra data efficiency on the small training set, we made aggressive use of geometric data augmentation: reflections, isotropic scaling, shear, stretching, rotation, jitter, and elastic deformations [12]. Elastic deformations

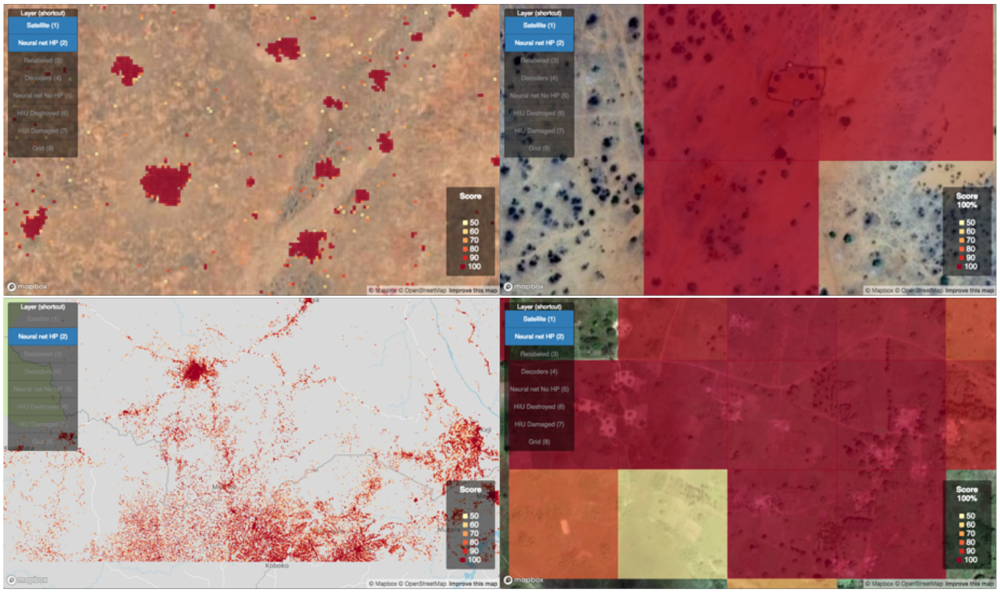


Figure 4: Screenshots from our visualization tool. TOP: Left: typical village patterns. Right: isolated tukuls. BOTTOM: Left: classification at the border between South Sudan, DRC, and Uganda. The widespread detection in Uganda (bottom of the images) suggested at first a poor performance of the network. It actually turned out to be an accurate detection of a very different urbanization pattern: a dense presence of very small but omnipresent clusters of tukuls, all along the network of roads crisscrossing this fertile region (Right). Color code: cf Fig. 2.

slowed training a lot but stabilized performance and prevented overfitting. The rest of the pipeline is unchanged. The initial performance is very promising, displaying class-balanced accuracies in the low to mid 90s, see confusion matrix above.

4 Toward impact: a practical web-based visualization tool

Motivation and engineering: To help domain experts applying their unique skills at larger scale using machine learning, we developed a web-based visualization (and relabeling) tool, stitching together all the imagery and overlaying labels and neural networks' output for interactive browsing. We used open-source libraries for cost-effectiveness: MapBox GL JS, Flask, jQuery, and Apache. Although still far from the dream end-goal, this took the analysis to a whole new level.

Urbanization patterns: the large-scale visualization of Fig. 2, allowed identification of villages the middle of the desert, systematically surrounded by a neighbouring orbit without any tukuls, then further isolated individual tukuls e.g. next to what seems to be cattle trails, see Fig. 4 (Top).

Generalization beyond Darfur: In order to visually assess the potential of the solution to generalize out of sample, we also applied our trained model on a very different region of East Africa, $1,600$ km further south, at the border between Uganda, Democratic Republic of the Congo, and South Sudan, see Fig. 4 (Bottom). Quantiative assessment of this generalization requires further expert labeling.

5 Discussion and future work

To our surprise, for the first time and against our ideals of reproducibility, ethical reasons prevent publication data *or* code without expert risk assessment. Doing either may indeed divulge the precise location of vulnerable villages. This touches upon a larger societal discussion of scientists' responsibility in the use of their tools – which such joint efforts with NGOs might help navigate.

We barely scratched the surface of what can be done. In particular, we still want to alleviate class imbalance, and report the number of destroyed villages or tukuls, e.g. using weak-supervision or counting [13].

Yet this shows the impact domain experts and a small technical team of just two can have. It bodes well for larger teams, such as now being built by one of the authors at Element AI: dedicated to making sure that the tools of the Machine Learning community empower those who are already fighting the good fights.

References

[1] Amnesty International. Scorched earth, poisoned air: Sudanese government forces ravage Jebel Marra, Darfur. Technical report, Amnesty International, September 2016. URL https://www.amnesty.org/en/documents/afr54/4877/2016/en/.

[2] Nathaniel A Raymond, Benjamin I Davies, Brittany L Card, Ziad Al Achkar, and Isaac L Baker. While we watched: Assessing the impact of the satellite sentinel project. *Georgetown Journal of International Affairs*, pages 185–191, 2013.

[3] Harvard Humanitarian Initiative. Satellite imagery interpretation guide: Intentional burning of tukuls. Technical report, Harvard University, 2015. URL https://hhi.harvard.edu/publications/satellite-imagery-interpretation-guide-intentional-burning-tukuls.

[4] Harvard Humanitarian Initiative. Annual report 2015. Technical report, Harvard University, 2015. URL https://hhi.harvard.edu/sites/default/files/hhi_year_end_report_2015_web.pdf.

[5] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353 (6301):790–794, 2016.

[6] M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9(1):010103, 2010. doi: 10.3847/AER2009036.

[7] Peter Smittenaar, Alexandra K Walker, Shaun McGill, Christiana Kartsonaki, Rupesh J Robinson-Vyas, Janette P McQuillan, Sarah Christie, Leslie Harris, Jonathan Lawson, Elizabeth Henderson, et al. Harnessing citizen science through mobile phone technology to screen for immunohistochemical biomarkers in bladder cancer. *British journal of cancer*, 119(2):220, 2018.

[8] Kevin Crowston, Carsten Østerlund, and Tae Kyoung Lee. Blending machine and human learning processes. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[10] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Zé Vinícius, cmmalone, Christopher Schröder, nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, rene rex, Kejia (KJ) Shi, Justus Schwabedal, carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, Loïc Estève, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch. scikit-optimize/scikit-optimize: v0.5.2, March 2018. URL https://doi.org/10.5281/zenodo.1207017.

[11] US Department of State - Humanitarian Information Unit. Darfur, sudan: Confirmed damaged and destroyed villages, february 2003 - august 2009. https://reliefweb.int/map/sudan/darfur-sudan-confirmed-damaged-and-destroyed-villages-february-2003-august-2009-includes, 2009.

[12] Patrice Y. Simard, David Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*, pages 958–962, 2003. doi: 10.1109/ICDAR.2003.1227801. URL https://doi.org/10.1109/ICDAR.2003.1227801.

[13] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 547–562, 2018.

[14] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Spinenet: Automatically pinpointing classification evidence in spinal mris. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, pages 166–175, 2016. doi: 10.1007/978-3-319-46723-8_20. URL https://doi.org/10.1007/978-3-319-46723-8_20.

A Interpretable error detection

A key objective of ours was to try and use the CNN to 'clean up' some of the noisy labels from the crowdsourcing. We did this in two ways. First, we would find the examples on which the CNN would disagree with a test label, then we would visualize why there was a disagreement. Either the label was incorrect, or the CNN. The visualization was conducted via a saliency map [14], as displayed in Fig. 5. The saliency map is simply the gradient of the output of the network $f(\mathbf{x})$ with respect to the input denoted $\mathbf{x}$. If we consider `habitation` vs. `no habitation`, then the output of the network was the `habitation` neuron. It has been shown that the visual quality of the saliency map can be improved by averaging the gradients over data augmented input samples, where $T_{\boldsymbol{\theta}}[\mathbf{x}]$ denotes an image $\mathbf{x}$ transformed with transformation parameters $\boldsymbol{\theta}$, drawn from some distribution $p(\boldsymbol{\theta})$. The gradient maps from each augmented image are aligned via the inverse transformation $T_{\boldsymbol{\theta}}^{-1}$ and averaged, providing the map:

$$\text{saliency map}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} T_{\boldsymbol{\theta}_i}^{-1}\left[\left|\nabla_{\mathbf{x}} f\left(T_{\boldsymbol{\theta}_i}\left[\mathbf{x}\right]\right)\right|\right], \qquad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}). \tag{1}$$

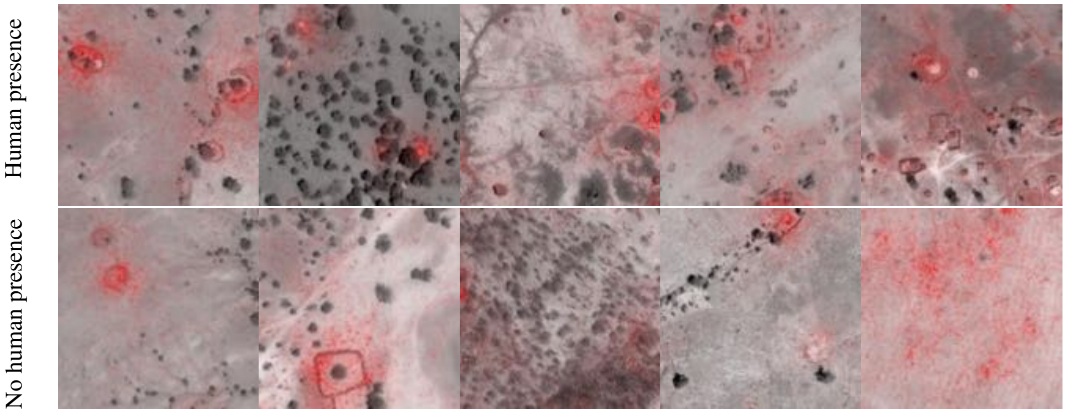


Figure 5: (Best seen on device in colour). Saliency maps: The left two images in each row are cherry-picked, the rest are randomly selected tiles for each class. TOP: tiles classified as having human presence. Notice how the saliency is concentrated about inhabited structures. BOTTOM: tiles classified as not having human presence. Most interesting is the first two images the saliency map, where the saliency highlights signs of habitation missed by the labelers. In the other three images, the saliency is evenly distributed or concentrated on ambiguous visual structures.