Author: Catherine Inness March 2020

Fairness in Machine Learning:
How Can a Model Trained on Aerial Imagery Contain Bias?

In the past forty years more than two million people have lost their lives in disasters caused by natural hazards[1]. Disaster Risk Management (DRM) is therefore vital to support the protection of people, property and infrastructure in the face of often unpredictable hazards.

With vast areas of land to consider in a wide range of potential disaster zones, the benefit of using artificial intelligence to survey geographic areas and aid decision-making in DRM is clear. The high accuracy level of state-of-the-art computer vision techniques means that machine learning carries the potential to speed up disaster response, reduce the monetary cost of analysis work, and perhaps detect vital information that humans could miss.

However, care must be taken to make sure the models are fair, for example to avoid providing protection to some sub-sections of the population over others, given equal need.

I'll explain how bias can occur in machine learning models based on aerial imagery, and explore some methods of reducing it.

The challenge of fairness:

Machine learning algorithms are optimised to **minimise their prediction error**, by discovering and exploiting patterns in data.

So far, so good. But in the pure pursuit of prediction accuracy, it's been shown that not everyone is being treated equally (for just one example, see the now well-known ProPublica investigation into the use of algorithms in the United States to rate a defendant's risk of committing a future crime).

It's commonly said that algorithms learn the inherent bias in the data that they are trained on: 'bias in, bias out'. It's clear how this could be true for case studies such as loan acceptance or fraud detection, where models are often trained on individual personal data, past outcomes or decision-making records from previous cases. But in the case of machine learning on aerial imagery for mapping, we're ultimately just working with pixels. How can a model learn a prejudice purely from pixels?

Our example cases:

Let's take two example case studies:

1. Using computer vision to map buildings from aerial imagery
2. Using computer vision to identify damage following a natural disaster

Let's assume that the outcome of both methods are used by authorities to inform decision-making on where future protective measures against natural disaster need to be deployed.

[1] https://www.worldbank.org/en/topic/disasterriskmanagement/overview

What we mean by fairness:

Most algorithms will make mistakes, just as humans do. A model might misclassify a roof as a field, or miss some damage to a road. The problem arises when the burden of model mistakes falls disproportionately on some people more than others.

One common definition of fairness in machine learning is the concept of **equalized odds** (Hardt, Price, & Srebro [1]), which is a form of classification parity that quantifies the extent to which **false-positive** and **false-negative** errors occur at the same rate for a **protected group** (a subgroup of the population that we suspect or fear is being treated unfairly by the model) versus the rest of the population.

For our disaster damage use case, false-negatives can be thought of as times where disaster damage is missed by the model, i.e. 'invisible' to the humans reviewing the output of the model and therefore ignored in decision-making on allocation of protective measures. If the model misses damage more often for one subgroup of the population than for others, it would fail the test of **equalized odds**.

A major limitation of the **equalized odds** measure is that we need to know our **protected groups** in advance so as to code them into the model, which often requires prior knowledge of the domain.

According to a 2018 World Bank report, *Building Back Better* [2], the poorest people are disproportionately impacted by disasters (the report urges us to consider the experience of a 1 dollar loss for a rich and a poor person). They are less likely to have savings or insurance and they are more likely to be living in a subsistence manner. This suggests that the pursuit of **equalized odds** with regard to income level may be a valuable aim for a machine learning model in this area.

It is important to note that the decision of which fairness measure to use, and how to implement it, is a policy and/or societal decision rather than a decision that solely falls within computer science. For example, consider the case where a model was found to place a slight **burden of error** on richer communities, by failing to identify natural disaster damage more often on the homes of richer communities than on poorer ones. If the real-world odds are already stacked against a poorer group, is it ethical to adjust the algorithm so that it performs better for the richer communities and brings the two groups in line with each other? This adjustment is what **equalized odds** would demand.

Similarly, achieving any fairness measure will often reduce the overall model accuracy level: we're no longer asking the model to just minimise it's prediction error, so we're 'distracting' it. Kearns and Roth [3] note the relevance of a **Pareto curve** here, typically used in the field of economics, to plot the range of reasonable combinations of the chosen fairness metric and error metric. It should then be a human activity to understand the trade-off and choose the appropriate point on the error-unfairness trade-off.

How bias can manifest:

Chouldechova and Roth [4] note three common causes of unfairness:

a) *Bias Encoded in Data*
In our example case, this could include:

Mistakes in labelling – Machine learning models for aerial imagery are commonly supervised learning models. Supervised models take as an input vast amounts of pre-labelled data. If

labelling errors occur in some areas more than others, or affect some building types more than others, then the burden of error won't be shared equally across the population.

For example, buildings can be missed entirely, perhaps because labels are generated from older building records or a human labeller didn't recognise the roof material (see Figure 1 for an example[2]). Or buildings surrounded by trees can be labelled incorrectly, perhaps because they are harder to spot (see Figure 2 for an example[2]).

A machine learning model is penalised for classifying buildings that are not labelled in the **ground-truth masks** that are provided in the training data, even if it is in fact a building. So if enough similar unlabelled examples exist in the labelled data, the model will learn to ignore buildings that look like these examples. As the model **generalises** to unseen data, it will continue to label buildings in the same way and propagate the error to new cases. In the extreme, buildings like the ones shown in Figure 1 and Figure 2 could end up 'invisible' to authorities and this could lead to the misallocation of protective measures.

Figure 1: Missed labelling on a house and a mis-labelled patch of grass (Dar es Salaam, Tanzania)

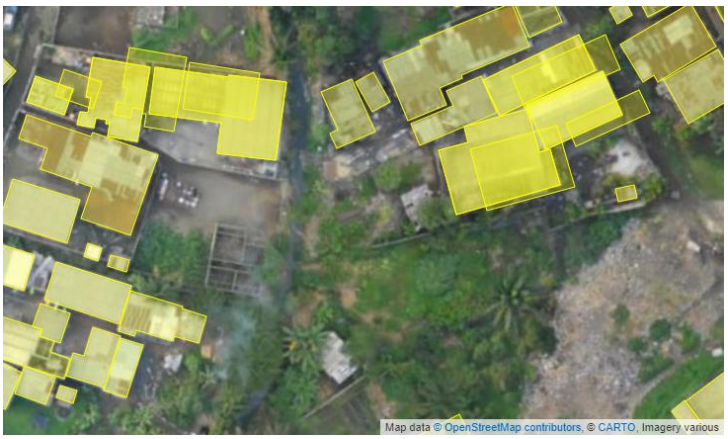

Figure 2: Missed labelling on buildings near trees (Pointe-Noire, Congo)

Differences in **data quality** – Imagery from different areas could have been taken in different weather conditions or with different photographic equipment, and some areas may have

[2] Images and labels from the Open Cities AI Challenge training dataset (https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/))

outdated or missing imagery. This could lead to the model failing to classify buildings in some areas compared to others.

b) *Minimizing Average Error Fits Majority Populations* – machine learning models use training data to 'learn' patterns, but **overfitting** (learning the training data 'by heart' and then being unable to generalize on unseen data) is a bad outcome. This means that the model has to make some generalizations.

For the example of damage detection, the model will recognise attributes that are linked to damage, for example that collapsed buildings do not cast a shadow, or that damaged buildings have uneven building boundary lines. It will then apply this 'knowledge' to the classification of buildings. (Note that this is hypothetical only - for most current state-of-the-art methods with many layers of learning, we don't actually know what attributes the model is using, and in some cases it may be using attributes that appear meaningless to the human eye)

Let's consider an area where the majority of people live in blocks of flats or houses along straight, paved roads, but that a small minority of low-income families live in corrugated metal roofed huts with narrow paths dividing them. If the majority live in the former type of home, with just a very small proportion in the latter, then the model can make mistakes on the smaller population without harming its overall accuracy as much as if it were to miss-classify homes of the majority type.

The above example model is trained only on image pixels. It has not been given explicit information on the income of the individuals involved, but the outcome can still be biased against low-income families. This is a simplified example. In reality the **protected groups** and associated attributes are far more subtle, increasing the risk that bias goes unnoticed.

c) *The Need to Explore* - we only have data from natural disasters that have already happened and for which we have image data. Ideally we'd be able to perform experiments and assess damage for different scenarios and for different groups of people. We'd also like to understand the future impact of the actions informed by the computer vision model.

If a major natural disaster were to happen tomorrow, and a computer vision algorithm were to be used to determine priority areas for immediate deployment of resources, the full impact on the communities (good and bad) of that chosen allocation of resources would not be clear for many months or years. The speed of algorithmic decision-making can be instantaneous, but the impact can persist for decades.

All disasters are different, and the likelihood is that each time an algorithm is employed for DRM, it is being applied to slightly different permutations of the problem it has been trained on. For example, a different geographic area, a different strength of hurricane or varying quality of post-disaster imagery. This makes it even more important that the measures of fairness and the degree of model explainability are robust and well understood, so that they generalise to unseen challenges as well as the model itself can.

What can be done to make models more fair:

The identification, reduction and removal of bias in machine learning is an area of ongoing research. Some examples of steps that can be taken include:

1. Improving the accuracy of the data labels across all sub-populations, as ultimately this is the only source of information for a supervised machine learning model. If the labels contain bias, even unintentional bias, then the algorithm will propagate it. Increased label accuracy can be achieved with measures such as:
    1. Improving the digitization of local building records
    2. Increasing diversity in the people selected to manually label data
    3. Looking for labellers from the mapped area who have expertise on the subject domain
    4. Asking for multiple people to label the same image and taking the majority vote.

2. Identifying **protected groups** (subgroups of the population that we suspect or fear are being treated unfairly by the model) and explicitly giving the model the aim of achieving fairness.

    One recent example is **Adversarial Debiasing** (e.g. Zhang et al. [5]), which is inspired by **Generative Adversarial Networks (GANs)**. GANs are commonly implemented by the creators of 'deepfake' videos; a generative network learns to make a prediction whilst the discriminative network tries to spot fakes. Both networks work together to create the best possible 'deepfake' videos. **Adversarial Debiasing** is a similar concept, but the role of the discriminator is to spot bias with regard to a protected group, rather than fakes. There are many other methods to explore, some of which are being implemented in code libraries for developers to include in their models.

3. Using methods for **Explainable AI** to better understand why a model has made predictions. For example, LIME [6], in which the authors used as an example an image classifier that is trained to distinguish between husky dogs and wolves. In one of the author's examples, the training data for wolves had snow in the background, whilst the husky photos did not. Their model presents to the user the specific pixels that caused the 'wolf' classification, and it was shown that the model was in fact basing the classification on the presence of snow, rather than on the features of the animal. In our example case, users could query the reasons behind a classification of 'damage' and 'not-damage', or 'building' and 'not building' in an attempt identify why the model is making mistakes and whether there are suggestions that any groups are subject to unfairness.

4. The creation of a **human feedback loop**, both to improve and augment the labelled data, but also to 'sign-off' the most impactful decisions. We don't wish to remove the sense of ownership and accountability from human decision makers; many factors impacting the effectiveness of DRM are only truly visible by humans who are based in or close to the affected communities. Even if a statistical measure of fairness is achieved, there is no substitute for experienced local knowledge in helping to assure true fairness. For example, regional human oversight would spot immediately if a community had been missed in the allocation of protective measures, or insufficient supplies were about to be sent due to a miscalculation of housing density.

    It may also be prudent to plan for a feedback loop in situations where protective measures are implemented and a further disaster occurs, i.e. using machine learning to help understand where the allocation of previous measures may have been unfair to protected groups.

    Where aerial imagery exists from both before and after a past natural disaster, research papers have attempted to build automatic damage detection models with some success. These types of **retrospective models** could be assessed for bias, or new models could be devised that explicitly target fairness. Taking this concept further, such models could be used

by regional centres with local expertise in a mock disaster recovery exercise, assessing the outcome for fairness and comparing to what was really deployed at the time of the disaster. The sharing of data across geographies and organisations would be an advantage here, using data and insights from areas of similar geographic make-up to obtain more imagery and understand past learnings.

Facial recognition software makes use of machine learning, and was subject to a backlash due to initial limitations in performance on different ethnicities. Given that this is another purely pixel-based application of machine learning, and is now in widespread use, it would be relevant to understand the fairness challenges and feedback loops experienced in this domain and look at applications for aerial imagery.

In summary:

This article only scratches the surface of the research and thought going into responsible AI at the moment. Machine learning can bring real benefit, for example by speeding up the response to disasters and reducing the cost of analysis work. However the effect of bias and unfairness in models can be significant, and it has the potential to exacerbate the gap between rich and poor communities if not fully understood and mitigated.

I hope I have exposed some of the ways that unfairness can be subtle but impactful, that individual personal data doesn't have to be involved for a model to encode unfairness, and that there are ways in which to measure and control bias such that the impact is reduced.

References:

[1] Hardt, M., Price, E., & Srebro, N. (2016), *Equality of Opportunity in Supervised Learning,* https://arxiv.org/pdf/1610.02413.pdf

[2] Hallegatte, S., Maruyama Rentschler, J., & Walsh, B. (2018), *Building Back Better: Achieving resilience through stronger, faster, and more include post-disaster reconstruction*, http://documents.worldbank.org/curated/en/420321528985115831/pdf/127215-REVISED-BuildingBackBetter-Web-July18Update.pdf

[3] Kearns, M., & Roth, A. (2020), *The ethical algorithm: the science of socially aware algorithm design*, Oxford University Press.

[4] Chouldechova, A., & Roth, A. (2018), *The Frontiers of Fairness in Machine Learning,* https://arxiv.org/pdf/1810.08810.pdf

[5] Zhang, B., Lemoine, B., & Mitchell, M. (2018), *Mitigating Unwanted Biases with Adversarial Learning*, https://arxiv.org/pdf/1801.07593.pdf

[6] Ribeiro, M., Singh, S., & Guestrin, C. (2016), *"Why Should I Trust You?" Explaining the Predictions of Any Classifier,* https://arxiv.org/pdf/1602.04938.pdf