An ML application to automate an existing manual process through the use of aerial imagery. Numerous areas throughout the ABS will benefit from the development of this ML application.

Organisation:	ABS
Author(s):	Daniel Merkas and Debbie Goodwin
Date:	28/02/2020
Version:	1.1

1. Background and why and how this study was initiated

The Address Register (AR) Automated Image Recognition (AIR) model looks to reduce the amount of manual intervention required during regular AR maintenance processes. The current processes utilising available administrative data are able to automate the classification of approximately 68% of the 100,000 new addresses the AR receives from the Geocoded National Address File (GNAF) quarterly. The remaining 32,000 addresses have typically been desktop canvassed by an ABS analyst using online tools such as aerial/satellite imagery and search engine queries. Desktop canvassing is a resource intensive process where typically 1,000 addresses per week can be classified per canvasser. This results in approximately 32 weeks of desktop canvassing required for one desktop canvasser each quarter.

The model was conceived based on the hypothesis that since most of the desktop canvassing work was actually simple and repetitive and therefore could be automated through machine learning. We were able to take part in the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) Data Fellowship which run by their Data 61 unit. This gave us the opportunity to leverage their expertise in machine learning to build a working model that we now refer to as AIR.

When the aerial imagery available is combined with existing administrative data, we are able to classify approximately 96% of addresses which means the 32,000 addresses that previously required desktop canvassing can be reduced to almost 4,000 per quarter. Furthermore, to illustrate the speed of the model, it was able to classify 340,000 unique addresses in three days. This significant reduction means that those desktop canvassers can focus on the more complex addresses that administrative data and the AIR cannot resolve whilst also reducing resource requirements through fewer desktop canvassers.

AIR has been trained to classify aerial images into one of six classes: residential dwelling, under-construction, vacant land, commercial, high density, and poor geocodes. These six classes are able to classify addresses where no administrative data is available and further strengthen classifications when it is available. One issue with administrative data is that it tends to future predict whereas the AIR is able to observe the address at a recent point in time. AIR is able to see whether the dwelling is habitable or is still either vacant land or under-construction. This reduces the amount of over-coverage in the AR and therefore has significantly benefits survey and census operations.

Key benefits:

- Automated process can rapidly classify aerial imagery. The process was able to classify 340,000 unique addresses over three days. This is extremely fast when compared to a desktop canvasser which would have taken up to 340 weeks to complete (assuming that a desktop canvasser can, on average, classify 1,000 addresses per week).
- By allowing the model to classify the bulk of addresses that are simple, the desktop canvassers can spend more time on improving the quality of the AR by focusing on the more difficult address uses that the model is not capable of classifying, e.g. hotels.
- The results from AIR can be used in conjunction with other administrative data sources to strengthen confidence and quality in the Address Register.

2. Data

2.1 Input Data (short description)

The input dataset consists of aerial images which are downloaded at a resolution of 150x150 pixels which covers an actual area of approximately 35m^2. The current training dataset consists of 6,000 images across six classes with 3,000 for training, 1,500 for validation, and 1,500 for testing. We aim to increase the number of training images when we progress to improving the quality of the model. The same image source was used previously by desktop canvassers who accessed the images through a browser to try and identify an address use.

2.2 Data Preparation

The Keras package allows for data augmentation by randomly applying a selection of changes to an image, such as skew, stretch, rotate, flip, etc. This helps with avoiding over-fitting.

2.3 Feature Selection

The images were manually assigned a class depending on which address use they described by the data scientist who developed the model. The six classes included: residential, under-construction, vacant land, commercial, high-density residential, and roads (to detect poor geocodes).

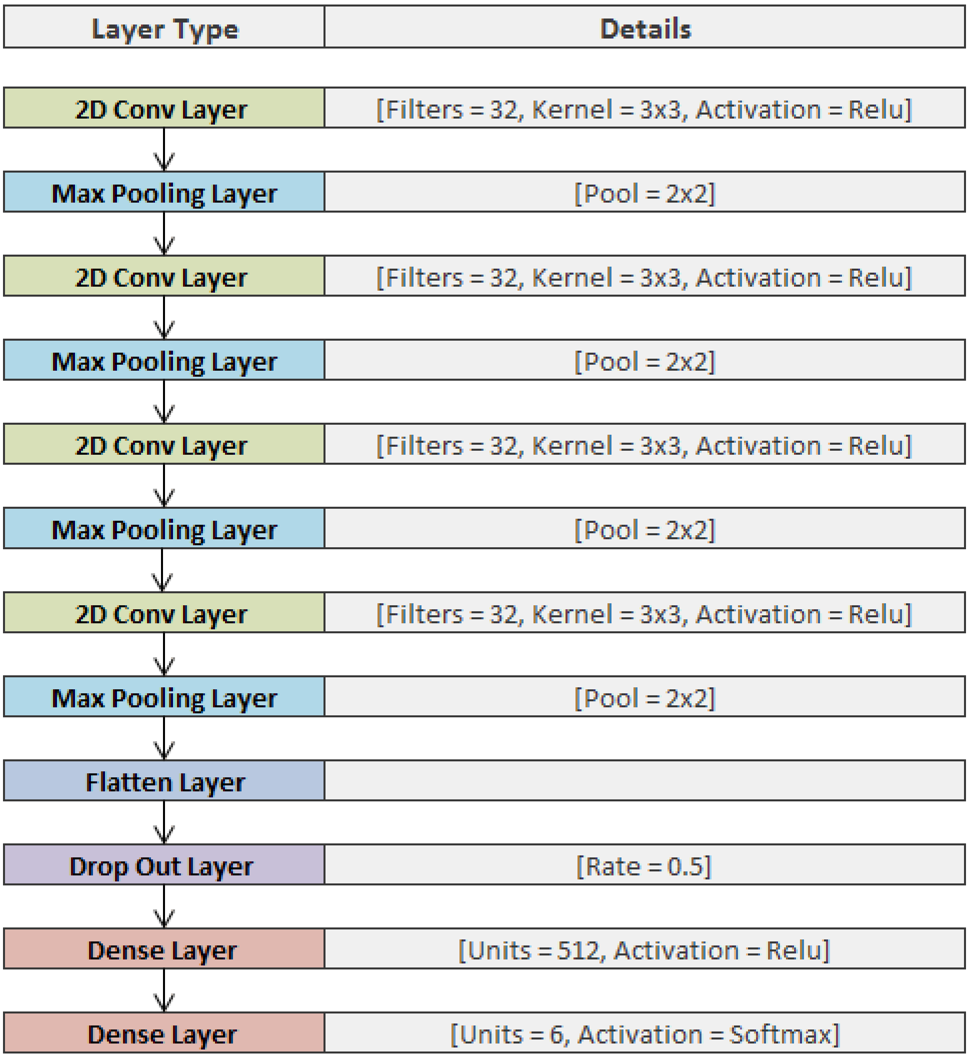
2.4 Output data

The model outputs a probability for each class where the sum of all equal one across the six classes as well as a classification which is essentially the class with the highest probability.

3. Machine Learning Solution

3.1 Models tried

The diagram below shows the layers used in the final model



Layer Type	Details
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
Flatten Layer	
Drop Out Layer	[Rate = 0.5]
Dense Layer	[Units = 512, Activation = Relu]
Dense Layer	[Units = 6, Activation = Softmax]

The model takes colour 150x150 pixel jpegs as inputs. This size was chosen to reduce resource use and minimise costs. Using the Nearmaps API, the images are downloaded with a Mercator zoom of 19 centered on the geocode. This results in an image that shows approximately a 35x35m area. This was chosen because it generally captures entire blocks to maximise the chance that the entire building will fall within the image. There is research being conducted to try and ensure the entire block is contained through the use of a bounding box based off geocodes.

During training, the inputs had data augmentation applied to reduce the chance of overfitting with a relatively small dataset. Only the training dataset has this applied, but not the validation or test datasets. It achieves this by not letting the model see the exact same image over and over again. Each of the following augmentations had a 20% chance of being applied: rotation, stretching vertically and horizontally, tilting, zooming, and flipping. Rectified Linear Unit (Relu) activation was used because is the recommended activation for this type of model. It defaults any negative values to zero.

The model contained four 2D convolutional layers with relu activation and a kernel size of 3x3. The kernel refers to the window that the algorithm uses to scan the image when it looks for features, such as textures or objects.

After each 2D convolutional layer, a max pooling layer was used to further reduce overfitting. The max pooling layer achieves this by finding the max value in a given shape, which for this model in 2x2. So, if a 2x2 grid had the values of 2, 3, 0, and 5, the 5 would now represent all four squares. This generalisation helps the model extract important features and reduces the chance that it focuses on too many minor features.

The flatten layer is applied to prepare the data output from the convolutional layers for the dense layers.

The dropout layer is another step used to reduce overfitting. A dropout value of 0.5 was applied which means that 50% of the information will be randomly removed. This means the model will rarely see the same image and increase its ability to classify a broader range of images in real world scenarios.

The next layer is a densely connected layer with relu activation. This type of layer uses all the available data and checks against it all the remaining data, hence being called densely connected. This layer reduces the data to 512 data points.

Finally, there is another densely connected layer which reduces the data to 6 data points which correspond to the number of classes we have. The softmax activation is recommended for multiclass classifications.

3.2 Model(s) finally selected and the criterion

A script was created to automatically generate test results over all models created during training. Then the model with the best accuracy was chosen. More work on this will be done when we retrain the model to improve quality.

3.3 Hardware used

Processor: 2.4GHz Dual-Core Intel Core i5

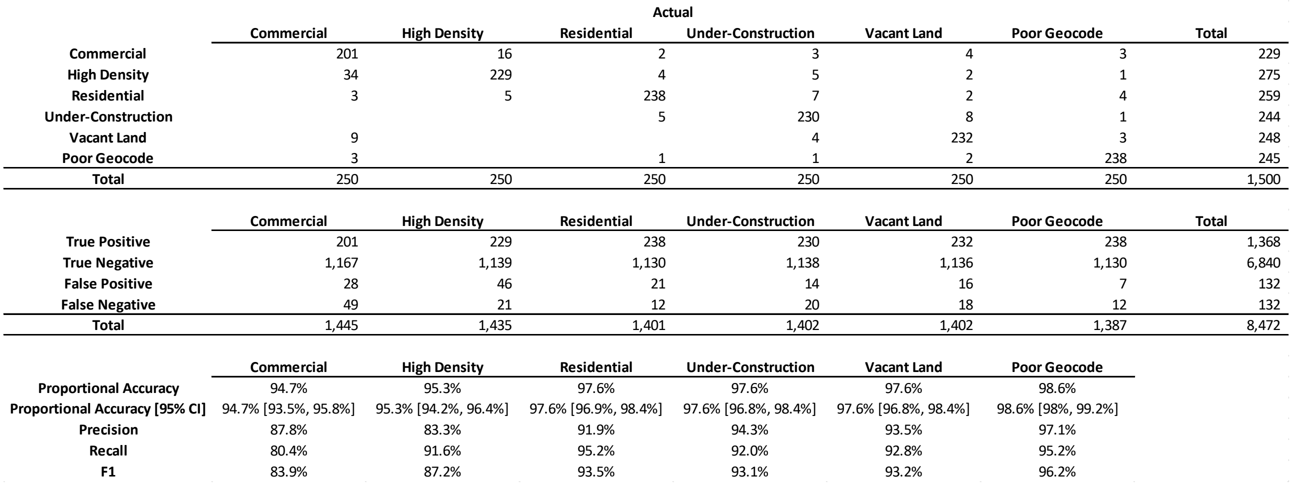
Memory: 8GB RAM

Graphics: Intel Iris 1536 MB

3.4 Runtime to train the model

Training took continued over around 150 epochs which took just over 17 hours for the final model.

4. Results



	Actual						
	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode	Total
Commercial	201	16	2	3	4	3	229
High Density	34	229	4	5	2	1	275
Residential	3	5	238	7	2	4	259
Under-Construction			5	230	8	1	244
Vacant Land	9			4	232	3	248
Poor Geocode	3		1	1	2	238	245
Total	250	250	250	250	250	250	1,500

	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode	Total
True Positive	201	229	238	230	232	238	1,368
True Negative	1,167	1,139	1,130	1,138	1,136	1,130	6,840
False Positive	28	46	21	14	16	7	132
False Negative	49	21	12	20	18	12	132
Total	1,445	1,435	1,401	1,402	1,402	1,387	8,472

	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode
Proportional Accuracy	94.7%	95.3%	97.6%	97.6%	97.6%	98.6%
Proportional Accuracy [95% CI]	94.7% [93.5%, 95.8%]	95.3% [94.2%, 96.4%]	97.6% [96.9%, 98.4%]	97.6% [96.8%, 98.4%]	97.6% [96.8%, 98.4%]	98.6% [98%, 99.2%]
Precision	87.8%	83.3%	91.9%	94.3%	93.5%	97.1%
Recall	80.4%	91.6%	95.2%	92.0%	92.8%	95.2%
F1	83.9%	87.2%	93.5%	93.1%	93.2%	96.2%

5. Code/programming language

The model was created and trained in RStudio using the Keras API with a Tensorflow (CPU version) backend that sits in an Anaconda environment.

6. Evolution of this study inside the organisation

Initially this project was created for the Address Register but downstream effects were always considered throughout. Since the Address Register forms the population frame for survey sampling it is important that it is of the highest quality and truly reflects the

Australian population and its housing stock. Consultation with Household Surveys and Census continue to ensure that their expectations are met.

AIR was presented widely to internal ABS areas that might have an interest in either using the process or the effects the process will have on their work. Additionally, the work to install the software for the first time has also meant others in the organisation are able to access these tools and start to build ML solutions for other problems.

7. Is it a proof of concept or is it already used in production?

This project has now moved to production but has taken many steps to gain acceptance within the organisation. Putting together a solid business case was crucial to getting the project off the ground and convincing to organisation to proceed with the development of AIR. Another important aspect was the choosing a solution that was as simple as possible while not trying to solve a problem that was too complex. This allowed AIR to quickly achieve its goals and show that it was effective early on.

Collaboration with various teams was crucial to AIR being cleared for production. Security was a major hurdle but by understanding their requirements we found a way to have the software deployed. It has also been cleared for production by our major downstream stakeholders.

7.1 What is now doable which was not doable before?

The most important benefit that AIR had provided is that we can now process a massive volume of work rapidly and confidently. This frees up our analysts to resolve the more complex issues that exist for which we didn’t have the resources to tackle before. Another important aspect is that we have more information that complements our existing processes and actually can help improve those results. The final aspect is that we now have the capability to create and deploy more ML solutions as application problems arise.

7.2 Is there already a roadmap/service journey available how to implement this?

Already in production.

7.3 Who are the stakeholders?

Stakeholder	**Engagement**
Census	We have had extensive presentations and meetings with the Census team. With the 2021 Census nearing, there is continuing need for close collaboration and engagement to ensure that AIR is meeting their expectations.
Household Surveys	We have an ongoing relationship with Household Surveys and any error in our data can have costly downstream effects. It is critical that AIR is achieving a high level of accuracy in a timely manner since we provide quarterly frames for survey sampling.
Building and Construction	There is interest from the Building and Construction Statistics team to use AIR to validate and complement their work. They collect building and demolition permits andn their work contributes to total Australian dwelling counts. We will continue to collaborate to ensure that AIR is fit for purpose.

7.4 Robustness

Currently each time AIR is run we can do visual checks on each classification. The software moves all the images into folders of which there is one for each class. Then the analyst can open the folder and check the thumbnails to ensure that the results are as expected. Any misclassifications can be identified and recorded so in future those examples can be used in the retraining of the model.

In the near future, we are looking to build the capability to retrain the model each quarter or as necessary. This will mean AIR will stay relevant and that it will be able to adapt to any changes in the target populations.

7.5 Fall Back

AIR is only one part of the process so if it fails then we can just use the systems that we used before. This would obviously increase the amount of work required and lower the quality of the Address Register because we don’t have the resources to do it all manually and therefore would have to prioritise work and delay resolving the remainder.

To help mitigate these risks, we have ensured that the knowledge of how to operate AIR is shared widely throughout the team and the documentation is clear and comprehensive. This means that if one instance of the software fails then we have other instances and operators that can continue the work.

-

8. Conclusions and lessons learned

Introducing AIR and ML to our work has proven that these tools and techniques can provide major benefits. Working in a risk-adverse environment is challenging but we were able to provide results and a solid business case to ensure that AIR was cleared for production. Putting a solid business case together and being able to produce tangible benefits has helped progress AIR to production.

9. Potential organisation risk if ML solution not implemented

There are two main risks if this ML solution was not implemented. The first is that we would not be able to realise the benefits of a reduced workload for less complex addresses. Manually classifying addresses is very resource intensive and our fiscal environment is challenging. By implementing AIR, we are able to achieve much more than before with the same amount of resources.

Secondly, we would not be able to build our skills internally so we can solve other problems that we face both within the immediate team and also the wider organisation.

10. Has there been collaboration with other NSIs, universities, etc?

This project started out of research conducted during a Data Fellowship run by Data 61 which is a part of the Commonwealth Scientific and Industrial Research Organisation (CSIRO). We were able to consult their data scientists to help build and refine the model.

11. Next Steps

The next steps include:

- Improving the quality of AIR by retraining the model and adjusting classes to ensure that they are more aligned stakeholder expectations, such as the under-construction class. This step includes developing a process to be able to retrain the regularly to ensure that new misclassifications can be incorporated into the learning.
- Research the use of land parcel polygons to remove bias from images. The current AIR model performs well by focusing on the parcel in the centre of the image but when fences are not present then issues can arise. The model can also be improved by varying the zoom of the aerial image to ensure that the entire parcel is captured with a minimal amount of non-relevant surrounding parcels in the image.
- Developing clearer quality metrics with collaboration with the Work Package (WP2) group on quality to instil confidence in our stakeholders.