Multi-Label Image Classification with Deep Learning

Jan A. Marais

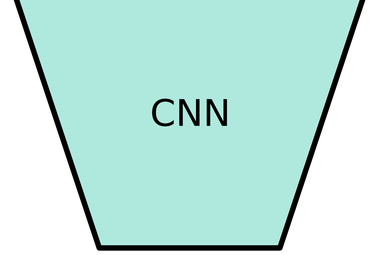
Stellenbosch University
Dept. Statistics & Actuarial Science
Supervisor: Dr. S. Bierman

maraisjandre9@gmail.com

Problem

1. Given a satellite **image**,

2. Apply a **Convolutional Neural Network** (CNN),



3. To obtain **multiplelabels,**

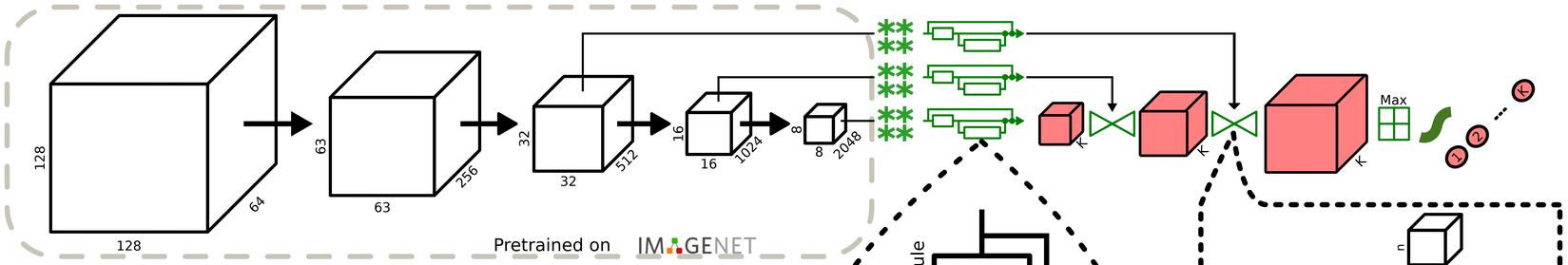
4. Which maximise the **F_2-score**,

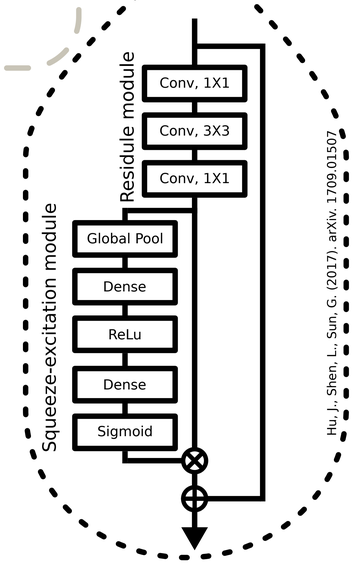
$$F_\beta = (1+\beta^2)\frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \beta = 2$$

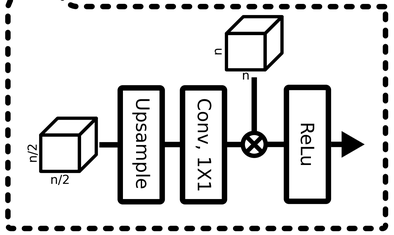
[1] Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. In CVPR, 2017.
[2] Zhu, F., Li, H., Ouyang, W., Yu, N. and Wang, X. (2017). Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. arXiv. 1702.05891.
[3] Wang, H.-D., Zhang, T. and Wu, J. (2017). The Monkeytyping Solution to the YouTube-8m Video Understanding Challenge. arXiv. 1706.05150.
[4] Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S. Deep Convolutional ranking for multilabel image annotation. (2013). arXiv. 1312.4894.
[5] Li, Y., Song, Y., Luo, J. Improving Pairwise Ranking for Multi-label Image Classification. In CoRR, 2017.

Solution

The proposed solution is a CNN architecture that follows the hourglass shape proven to be very effective in object detection and segmentation tasks. It can be built on top of any popular CNN for classification - we use ResNet 50. Its main advantage is that it can classify concepts relating to small parts of an image by utilising convolutional feature maps, , of multiple resolutions. The feature maps for each chosen scale is first transformed by a dilated convolution, , so that there is a channel for each label, while maintaining a larger field of perception. Then it is sent through Squeeze-Excitation (SE) residual module, , which is able to model interdependencies between channels - important to exploit spatial relations between labels. The output is merged with an upsampled version of a smaller, already processed, feature map in . The final feature maps are pooled globally, , and then transformed by a sigmoid activation, , so that the class scores, , are in [0,1] but not sum to 1. The network is trained by optimising the label-wise binary cross-entropy.







How to ...

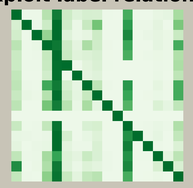
Detect small objects?

We make use of multiple feature maps of different scale to make predictions. Novel in multi-label image classification, popular in object detection and segmentation, e.g. [1]

Exploit spatial relations?

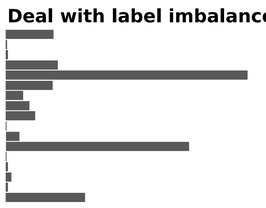
We use the SE-residual module on learned label attention maps. Inter-channel dependencies are learned, therefore spatial! Similar to, [2].

Exploit label relations?



Feature map output from SE-module is passed on to next scale prediction. Since, the maps are pixel-wise class scores, the model is effectively learning label relations. See Chaining, [3].

Deal with label imbalance?

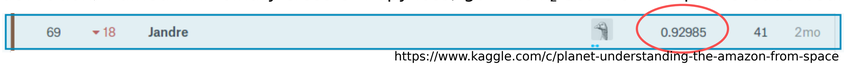


Assign higher weights to infrequent labels or use a more robust loss function. [4] used a pairwise rank loss.
We want to try stratified mini-batch sampling for better gradient approx. during training.
Another challenge is threshold selection. Either by grid search or by learning, as in [5].

Results

Results are work in progress. These are preliminary and baseline results.

On the *Planet: Understanding the Amazon from Space*, Kaggle competition our baseline model (ResNet50 with an added residual block with K filters, global average pooling and sigmoid activation, trained with binary cross-entropy loss) got this F_2-score on the public leaderboard:



https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

The winning score was 0.9333 (out of 938 competitors). Our results are acceptable since it is a single model solution, compared to large ensembles of models used by the winners.

... to come

> Implementation and evaluation of stratified mini-batch sampling.

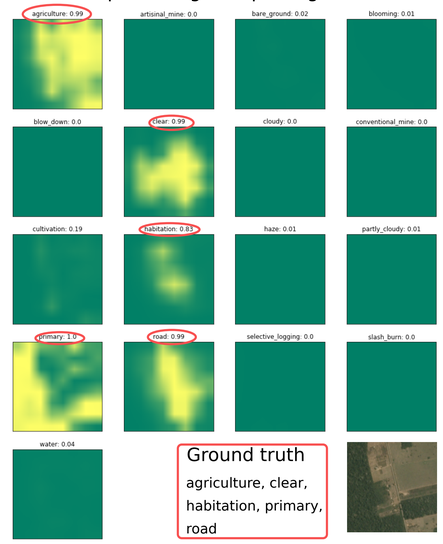
> Evaluate the effectiveness of adding lower level feature maps to make predictions.

> Evaluate proposed network on more datasets, such as MSCOCO and VOC2007*.

*requires big GPU.

Examples

The advantage of global pooling as prediction layer - get approximate localisation of labels for free. These are the most discriminative parts in an image and are the activations from the last feature map before global pooling.



Acknowledgements

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF..