Best Practices to Train Deep Models on Imbalanced Datasets—A Case Study on Animal Detection in Aerial Imagery*

Benjamin Kellenberger[0000−0002−2902−2014], Diego Marcos[0000−0001−5607−4445], and Devis Tuia[0000−0003−0374−2459]

Wageningen University and Research, The Netherlands
{firstname.lastname}@wur.nl

Abstract. We introduce recommendations to train a Convolutional Neural Network for grid-based detection on a dataset that has a substantial class imbalance. These include curriculum learning, hard negative mining, a special border class, and more. We evaluate the recommendations on the problem of animal detection in aerial images, where we obtain an increase in precision from 9% to 40% at high recalls, compared to state-of-the-art.

Keywords: Deep Learning · Class Imbalance · Unmanned Aerial Vehicles.

1 Introduction

Convolutional Neural Networks (CNNs) [5] have led to tremendous accuracy increases in vision tasks like classification [2] and detection [9,8], in part due to the availability of large-scale datasets like ImageNet [11]. Many vision benchmarks feature a controlled situation, with all classes occurring in more or less similar frequencies. However, in practice this isn't always the case. For example, in animal censuses on images from Unmanned Aerial Vehicles (UAVs) [6], the vast majority of images is empty. As a consequence, training a deep model on such datasets like in a classical balanced setting might lead to unusable results.

In this paper, we present a collection of recommendations that allow training deep CNNs on heavily imbalanced datasets (Section 2), demonstrated with the application of big mammal detection in UAV imagery. We assess the contribution of each recommendation in a hold-one-out fashion and further compare a CNN trained with all of them to the current state-of-the-art (Section 4), where we manage to increase the precision from 9% to 40% for high target recalls. The paper is based on [3].

2 Proposed Training Practices

The following sections briefly address all the five recommendations that make training on an imbalanced dataset possible:

* Supported by the Swiss National Science Foundation (grant PZ00P2-136827).

Curriculum Learning For the first five training epochs, we sample the training images so that they always contain at least one animal. This is inspired by Curriculum Learning [1] and makes the CNN learn initial representations of *both* animals and background. This provides it with a better starting point for the imbalance problem later on.

Rotational Augmentation Due to the overhead perspective, we employ 90°-stop image rotations as augmentation. However, we empirically found it to be most effective at a late training stage (from epoch 300 on), where the CNN is starting to converge to a stable solution.

Hard Negative Mining After epoch 80 we expect the model to have roughly learned the animal and background appearances, and thus focus on reducing the number of false positives. To do so, we amplify the weights of the four most confidently predicted false alarms in every training image for the rest of the training schedule.

Border Class Due to the CNN's receptive field capturing spatial context, we frequently observed activations in the vicinity of the animals, leading to false alarms. To remedy this effect, we label the 8-neighborhood around true animal locations with a third class (denoted as "border"). This way, the CNN learns to treat the surroundings of the animals separately, providing only high confidence for an animal in its true center. At test time, we simply discard the border class by merging it with the background.

Class Weighting We balance the gradients during training with constant weights corresponding to the inverse class frequencies observed in the training set.

3 Experiments

3.1 The Kuzikus Dataset

We demonstrate our training recommendations on a dataset of UAV images over the Kuzikus game reserve, Namibia[1]. Kuzikus contains an estimated 3000 large mammals such as the Black Rhino, Zebras, Kudus and more, distributed over 103 km^2 [10]. The dataset was acquired in May 2014 by the SAVMAP Consortium[2], using a SenseFly eBee[3] with a Canon PowerShot S110 RGB camera as payload. The campaign yielded a total of 654 4000×3000 images, covering 13.38 km^2 with around 4 cm resolution. 1183 animals could be identified in a crowdsourcing campaign [7]. The data were then divided image-wise into 70% training, 10% validation and 20% test sets.

[1] http://kuzikus-namibia.de/xe_index.html
[2] http://lasig.epfl.ch/savmap
[3] https://www.sensefly.com

3.2 Model Setup

We employ a CNN that accepts an input image of 512×512 pixels and yields a 32×32 grid of class probability scores. We base it on a pre-trained ResNet-18 [2] and replace the last layer with two new ones that map the 512 activations to 1024, then to the 3 classes, respectively. We add a ReLU and dropout [12] with probability 0.5 in between for further regularization. The model is trained using the Adam optimizer [4] with weight decay and a gradually decreasing learning rate for a total of 400 epochs.

We assess all recommendations in a hold-one-out fashion, and further compare them to a full model and the current state-of-the-art on the dataset, which employs a classifier on proposals and hand-crafted features (see [10] for details).

4 Results and Discussion

Figure 1 shows the precision-recall curves for all the models.

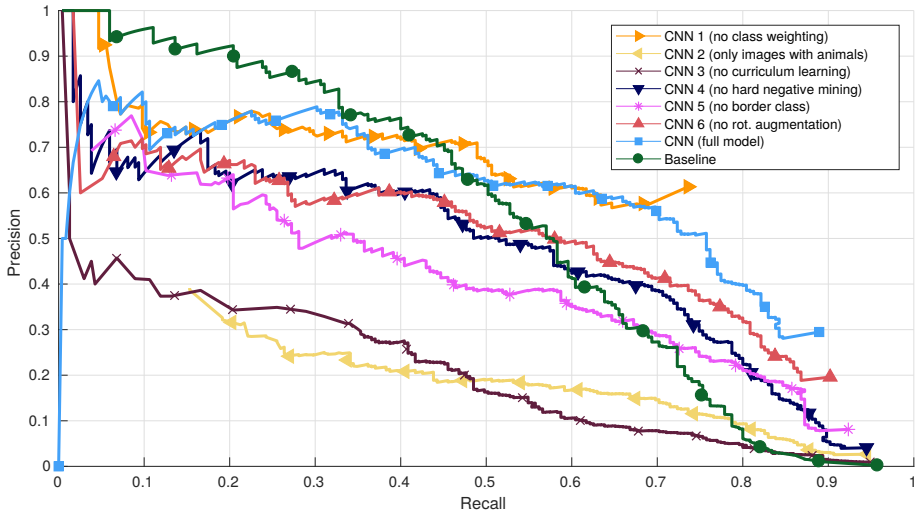

Fig. 1. Precision-recall curves based on the animal confidence scores for the hold-one-out CNNs (first six models), the full model and the baseline.

All recommendations boost precision, but with varying strengths. For example, disabling curriculum learning ("CNN 3") yields the worst precision at high recalls—too many background samples from the start seem to severely drown any signal from the few animals. Unsurprisingly, a model trained on only images that contain at least one animal ("CNN 2") is similarly bad: this way, the model only sees a portion of the background samples and yields too many false alarms. The full model provides the highest precision scores of up to 40% at high recalls of 80% and more. At this stage, the baseline reaches less than 10% precision, predicting false alarms virtually everywhere. In numbers, this means that for 80% recall our model predicts 447 false positives, while the baseline produces 2546 false alarms.

5 Conclusion

Many real-world computer vision problems are characterized by significant class imbalances, which in the worst case makes out-of-the-box applications of deep CNNs unfeasible. An example is the detection of large mammals in UAV images, out of which the majority is empty. In this paper, we presented a series of practices that enable training CNNs by limiting the risk of the background class drowning the few positives. We analyzed the contribution of each individual practice (curriculum learning, hard negative mining, etc.) and showed how a CNN, trained with all of them, yields a substantially higher precision if tuned for high recalls.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48. ACM, New York, NY, USA (2009)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
3. Kellenberger, B., Marcos, D., Tuia, D.: Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sensing of Environment (in revision)
4. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
6. Linchant, J., Lisein, J., Semeki, J., Lejeune, P., Vermeulen, C.: Are unmanned aircraft systems (uass) the future of wildlife monitoring? a review of accomplishments and challenges. Mammal Review **45**(4), 239–252 (2015)
7. Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., Briant, J., Millet, P., Reinhard, F., Parkan, M., Joost, S.: Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response. Big Data **4**(1), 47–59 (2016)
8. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
10. Rey, N., Volpi, M., Joost, S., Tuia, D.: Detecting animals in african savanna with uavs and the crowds. Remote Sensing of Environment **200**, 341–351 (2017)
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
12. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1), 1929–1958 (2014)