Dealing with Overfitting in the Context of Liveness Detection using FeatherNets with RGB image

Miguel Leão
miguel.leao@isr.uc.pt

Institute of Systems and Robotics
University of Coimbra
Portugal

Nuno Gonçalves
nunogon@deec.uc.pt

Portuguese Mint and Official Printing Office
(Imprensa Nacional-Casa da Moeda SA)
Portugal

Abstract

With the increased use of machine learning for liveness detection solutions comes some shortcomings like overfitting, where the model adapts perfectly to the training set, becoming unusable when used with the testing set, defeating the purpose of machine learning. This paper proposes how to approach overfitting without altering the model used by focusing on the input and output information of the model. The input approach focuses on the information obtained from the different modalities present in the datasets used, as well as how varied the information of these datasets is, not only in number of spoof types but as the ambient conditions when the videos were captured. The output approaches were focused on both the loss function, which has an effect on the actual "learning", used on the model which is calculated from the model's output and is then propagated backwards, and the interpretation of said output to define what predictions are considered as bonafide or spoof. Throughout this work, we were able to reduce the overfitting effect with a difference between the best epoch and the average of the last fifty epochs from 36.57% to 3.63%.

1 Introduction

With the rise of facial recognition technology in day-to-day applications, such as mobile payments, comes a concern for the security of these systems. To counteract these security vulnerabilities, which present themselves as Presentation Attacks (PA), the development of Presentation Attack Detection (PAD) or liveness detection has become a requisite of modern facial recognition systems.

Most PAD methods are developed with machine learning, with the data that is fed to these models being presented in multiple modalities, namely colour, depth and infrared. While the results obtained from the high end models using all of this information, it is unrealistic that these systems can be easily transitioned to real life use. This can be due to either the devices not being able to capture the different modalities of images or that they do not have the computation power required to run the model. While costly, it is these complexities that reduce or even remove the presence of overfitting in these models, which in the case of liveness detection can be attributed to certain factors like the binary nature of the problem itself: "bonafide or spoof?".

The overall objective is to reduce the requirements of liveness detection solutions, be it in computational requirements, monetary cost or information requirements in order to apply these solutions to the systems that would most benefit from them. This work does that by reducing the effect of overfitting in the results while restricting itself to the use of only colour images as its data.

2 Literature Review

Since the question of liveness detection can be put bluntly as "bonafide or spoof" the first **machine learning** solutions employ binary cross-entropy loss as the sole learning supervision for the network [7]. However due to its simplicity, the models are prone to overfitting since they can easily focus their learning in arbitrary features, not relevant to the liveness detection problem. While the use of different loss functions has been employed [6] by interpreting the issue in other ways, another solution was to aid the loss function using pixel-wise supervision.

Pixel-wise supervision can be made by using previous knowledge of liveness detection, and applying it to the model. For example, the use of pseudo depth maps [8] based on the knowledge that, two dimensional attacks (print and replay) will display a "flat" depth map can be used to aid the model. By the same logic, binary mask labels [5] or reflection maps [3] have been used.

The previously mentioned approaches are all based on colour inputs (RGB, YCbCr or HSV) and it is the modality most commonly used. However, thanks to the development in sensors, it is possible to retrieve datasets using other modalities like depth, infra-red or thermal images. The models can then use a singular type of modality, or use the information available from several modalities all at once.

One such work is **FeatherNets** developed by Zhang et al. [9] in the interest of adapting the current deep learning approaches to liveness detection, which are usually very heavy in both computation requirements and data storage, to use in mobile or embedded devices which are incapable of meeting these requirements. To solve this problem, they propose a network "as light as a feather" that using depth information is able to achieve ACER of 0.00168, with only 0.35 million parameters and 83 million flops down from the baseline using ResNet18 [2] with an ACER of 0.05 with 11.18 million parameters and 1800 million flops. This network was chosen since its lightweight nature is in line with the overall objective of our work.

3 Approach

For the most part, the work conducted for this paper follows the methods presented by the authors of FeatherNets, adding the use of the WMCA [1] dataset and resorting to the use of colour (RGB) information instead of the original use of depth information.

The two datasets used are **CASIA-SURF**, developed by Zhang et al. [10], with 21,000 videos of 1,000 individuals captured with an Intel Real Sense 3000 camera providing not only RGB images but also depth and infrared images, and **WMCA**, developed by George et al. [1], being quite smaller than the previous dataset with 1,679 videos of 72 individuals, which are divided in 347 bonafide cases and 1,332 spoofs. The datasets were both captured using the Intel Real Sense 3000 camera, removing questions regarding camera quality to the discussion. The interest is in how the larger dataset presents less variety in attacks, with only print attacks varying in their positioning and the smaller dataset having more types of attacks captured in more varied conditions.

FeatherNets' structure is based on a main block, a down sampling block and then a streaming module that substitutes the fully connected layer as to reduce overfitting. The main block is based on the "MobileNet v2" model proposed by Sandler et al. [4] which employs the use of depth wise convolution as well as inverted Rectified Linear Unit (ReLU) blocks to improve the computation requirements associated with the computer vision tasks. The main block is then followed by one of two down sampling blocks, creating the distinction between FeatherNetA and FeatherNetB. FeatherNetA's downsampler is the simpler of the two having a singular branch of the depth wise convolution/inverted ReLU combination while increasing the stride of the convolution to 2 thus reducing the dimensions of the input to 12.5% of the original size. FeatherNetB's downsampler has also a first branch equal to FeatherNetA but adds a parallel secondary branch with average pooling to better learn more diverse features.

The occurrence of overfitting will be defined through the decrease of accuracy over the epochs, the larger the reduction, the more prevalent the overfitting. This can be simply read through the result tables presented throughout the document and is translated graphically in an increase of accuracy until it hits a peak (the highest accuracy score, considered then as the best epoch) and a subsequent decrease until a plateau is reached (here the model is no longer learning and is perfectly adapted to the training set).

4 Experiments and Results

Several experiments were conducted following the methods described in [9], first strictly to confirm that the model works as intended and then with only the colour image and varying certain parameters in the model. For the sake of brevity, these results showed that the dataset with more different types of attacks showed less overfitting than the dataset with only the print attacks, with the discrepancy between the best epoch and the average of the final 50 epochs being 1.00% and 36.57% respectively. This could be explained with the information that depth maps give in this context that can easily detect 2D types of attacks (print and replay) from 3D attacks (masks or mannequin heads for example) as can be seen in figure 1. This conclusion was further explored by adding 3D attack examples to the simpler dataset and observing that the overfitting would slightly reduce. The parameter variation, namely in the focal loss function, showed little effect.

Figure 1: Comparison between RGB and depth images of a print attack (left) and a bonafide face (right). Note that the depth images aren't of great quality, not being able to capture the eyes cut out of the print attack and not giving much detail to the bonafide case, but being possible to notice the differences. Images selected from the CASIA-SURF dataset [10].

The more notable results came from the tests involving a Precision-Recall (PR) curve, where in the threshold that defines the classification of bonafide or spoof was tuned to reach the best PR curve with Precision being the percentage of correctly predicted true cases among all predicted true cases and Recall the percentage of true values predicted as such. The approach is running the model at different thresholds between 1, where no image can be considered as bonafide and 0 where all predictions will be bonafide. Once all these values are obtained the points can be plotted in a graph and then a curve adjusted to them. From this curve a point can be picked out as what is considered ideal, in this case the closest point to what be considered perfect i.e. $(precision, recall) = (1, 1)$, however the threshold value needs to be inferred from where the ideal point stands in the graph. This ablation study was conducted using the CASIA-SURF dataset on FeatherNet A, reaching the curve presented in figure 2.

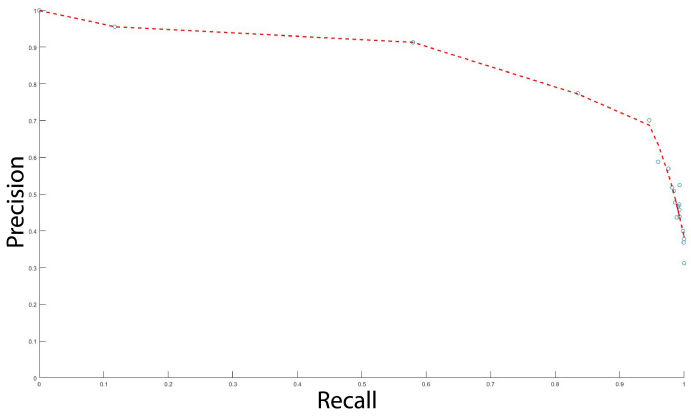

Figure 2: Precision-Recall curve. The curve was obtained using Matlab's polyfit() function. The threshold chosen was obtained by using Euclidean distance to find the closest point to the perfect (1,1) which resulted in point (0.8913,0.7828) which corresponds to a threshold value of roughly 0.9675.

With the "ideal" threshold calculated $threshold = 0.9675$, it is only a matter of repeating the initial experiments with the new threshold, which reduced the initial 36.57% discrepancy to only 3.63%.

5 Conclusion

This work showed the importance of a varied dataset and how these variations are able to compensate for loss of information associated with the multiple modalities an image can be presented with. From this loss of information, the overfitting effect present in the model became considerably noticeable with a difference between the best result, obtained at epoch 9 with an accuracy of 89.75%, and the average accuracy of the last fifty epoch's, equal to 36.57%. By adjusting the threshold that defined bonafide or spoof, this difference was reduced to 3.63%.

The results obtained during this work present possible considerations that could be helpful in the development of future solutions, both regarding the size, diversity and applicability of the datasets, as well as the modality given to the model.

References

[1] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, Andre Anjos, and Sebastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15, 2020.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Jun. 2016.

[3] Taewook Kim, YongHyun Kim, Inhan Kim, and Daijin Kim. Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 494–503, Oct. 2019.

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, Jun. 2018. doi: 10.1109/CVPR.2018.00474.

[5] Wenyun Sun, Yu Song, Changsheng Chen, Jiwu Huang, and Alex C. Kot. Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE Transactions on Information Forensics and Security*, 15:3181–3196, 2020. ISSN 1556-6013. doi: 10.1109/TIFS.2020.2985530.

[6] Xiang Xu, Yuanjun Xiong, and Wei Xia. On improving temporal consistency for online face liveness detection. *arXiv preprints arXiv:2006.06756*, Jun. 2020.

[7] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145, Nov. 2015.

[8] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5295–5305, Jun. 2020.

[9] Peng Zhang, Fuhao Zou, Zhiwen Wu, Nengli Dai, Skarpness Mark, Michael Fu, Juan Zhao, and Kai Li. Feathernets: Convolutional neural networks as light as feather for face anti-spoofing. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June, 2019.

[10] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z. Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, 2019.