BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Face Anti-Spoofing with Out-of-distribution Detection

Bc. Petr Češka*

Abstract

This paper aims to improve Vision Transformer-based face anti-spoofing model's ability to detect unknown attacks. It uses out-of-distribution (OOD) detection to filter out images that are too different from the model's training dataset, referred to as in-distribution (ID) data. This is done by extracting image features from one of the last model layer and using various metrics to separate them. The paper investigates how well different metrics identify outliers and how using them to filter data affects the model accuracy. Using Relative Mahalanobis distance we can distinguish ID from OOD images with a 98% accuracy. Omitting OOD images that shouldn't be classified can provide an extra layer of security for critical applications against unknown face spoofing attacks.

*xceska05@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Face anti-spoofing is always evolving field, because it needs to react on newly developed attacks. But models usually identify only the attack that were used to train them. Out-of-distribution (OOD) detection is one of the approaches how to deal with the spoof images that weren't present in the training dataset. Unlike other methods OOD doesn't need samples from all attacks it detects so it is ideal for spoofs that are not common or yet to be created. Thus the aim is to improve accuracy of the model on datasets containing both seen and unseen attack.

In this paper OOD detection is used for analyzing image representations already processed by an anti-spoofing models and deciding if it is similar to training data or not. With this knowledge it can be decided, whether to reject the image or to mark it for human inspection.

For purpose of this paper four datasets shown on Figures 1-4 are considered. MSU-MFSD, Replay-Attack, CASIA-FASD and OULU-NPU are all datasets used for anti-spoofing, but the conditions during taking the photos, the devices that were used to take the photos and the people in the photos are different. That helps model to learn to adapt and it gives the room for testing.

2. Proposed method

The proposed method uses two sets of features, features after ViT (FV) and features after projection (FP). Both are exported from a Face Anti-Spoofing model with Language-Image Pretraining (FLIP) [1] which is based on Visual Transformer (ViT). There are three versions of FLIP model.

FLIP-V is a ViT with classification head. There FV are vectors outputed by ViT before embedding layer and FP are vectors before classification layer. FLIP-IT and FLIP-MCL are comparing image embedding with text embedding to classify the input. There FV are vectors outputs of ViT and FP are representations projected into 512 dimensional space shared for image and text embeddings. How models work, where features are exported and how OOD detection fits into the workflow is shown on Figures 5, 6.

Features and logits are processed by ten OOD methods similarly to [2]. Each method returns a score for all images. A decision boundary is set based on the scores. In this case threshold is where False Accept Rate and False Reject Rate are equal aka equal error rate.

2.1 Relative Mahalanobis Distance

In most cases Relative Mahalanobis Distance (RMD) [3] is the best performing metric. The core computation is the same as Mahalanobis distance (MD) 1 using mean μ and covariance matrix S of each

class. But RMD also takes into account position of the entire training dataset and tries to eliminate its influence 3.

$$MD_c(x) = \sqrt{(x-\mu_c)^T S_c^{-1}(x-\mu_c)} \quad (1)$$

$$MD_0(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \quad (2)$$

$$C_{RMD}(x) = -\min_c\{MD_c(x) - MD_0(x)\} \quad (3)$$

The final score C_{RMD} is equal to distance to the closest class. There is six classes - real and spoof for each training dataset. This helps by about 1% in comparison with using just two classes - real and spoof.

3. Evaluation

For evaluation four datasets were used MSU-MFSD, Replay-Attack, CASIA-FASD and OULU-NPU. Each test consists of four subtests. By using always three out of four datasets as in-distribution (ID) data and one dataset as OOD data.

Evaluation of the process is done in three steps. Evaluation of the models itself, evaluation of the OOD detection methods and influence of using the OOD detection for pruning the data before classifying by the model.

Apart from reported Evaluations other tests were performed. Influence of cropping the faces, altering OOD detection method's inner setting and pruning consequences to more extent are reported in the whole Diploma thesis.

3.1 Evaluation of FLIP models

LIP-V, FLIP-IT and FLIP-MCL were trained with all dataset serving as OOD dataset one by one. All training presets were ran five times and averaged for robustness. Achieved performance is shown in Figure 7 where its compared to other anti-spoofing models tested with the same presets. The reported performance was matched and, apart from one preset, FLIP models reaches the best values. That is why this paper is focused on these particular models.

3.2 OOD detection methods

Evaluation of OOD detection methods is done by using features (FV or FP) extracted from model to compute score. Scores are then used for computing area under its receiver operating characteristic curve (AUROC). This value is taken as a good representation of overall performance. In table Figure 8 best achieved AUROC values are shown. Type shows which features were used to reach this AUROC. The characters in brackets show whether the value was achieved with best model (B), last model (L) and whether the exported scores had to be inverted (N). Model that had the highest accuracy on validation dataset during training is denoted as best model and the one reached in last iteration of training is denoted as last model.

3.3 Pruning data

Based on scores computed in previous section, testing data are pruned. Model's performance is then tested on both original and pruned data. The improvement in accuracy is shown in Figure 9. The change was always positive to a greater or lesser extent.

4. Conclusion

OOD detection was successful with best AUROC of 0.9721, 0.9765 and 0.9568 on models FLIP-V, FLIP-IT and FLIP-MCL respectively. Models accuracy after pruning of detected OOD samples was increased by 0.97 % in average. That means that proposed method fulfilled its goal to help models detect unknown attacks by using OOD detection.

5. acknowledgements

I would like to thank my supervisor Ing. Jakub Špaňhel for his help and guidance and for providing me with all the training datasets.

References

[1] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19685–19696, October 2023.

[2] Jingyao Li, Pengguang Chen, Shaozuo Yu, Zexin He, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need, 2023.

[3] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection, 2021.