ISSN: 2688-8653

MobileNetV2-based Face Anti-spoofing Algorithm

Tianqi Zhou*

School of Southwest Minzu University, Chengdu 610000, China

Abstract

For face Anti-spoofing tasks, with the booming development of deep learning, more face Anti-spoofing tasks have been completely shifted from traditional feature extraction to convolutional neural network framework through scholars' research today.In this paper, by improving mobilenetv2 and introducing an attention mechanism, we pre-trained shallow-medium-deep features and achieved good proformace on CASIA-MFSD, and Replay-Attack datasets, reducing the EER to 2.87%, and 1.94%, respectively.

Keywords

Convolutional Neural Networks; Mobilenetv2; CASIA-MFSD; Replay-Attack.

1. Introduction

1.1. Based on Traditional Methods

Face recognition technology has had many mature applications in our lives in recent years, such as face payment systems, access control gate systems, cell phone unlocking, smart homes, etc.Although the applications are very promising, the security issues that come along with them are worth studying, as criminals can break these systems based on deep learning by forging face information, which can cause unpredictable economic losses and negative social impacts.

Therefore, the research of face Anti-spoofing algorithms has been paid attention by a large number of scholars, however, the initial face Anti-spoofing algorithms did not make use of deep learning, and the early scholars extracted features manually to find the differences between living and non-living bodies, and then designed features by the differences, and finally sent them to the classifier to make decisions.

Wen D , Han H , Jain A K [1] et al, in 2015, proposed a single-frame input method designed with statistical features such as specular reflection, image quality distortion, and color to try to find the variability between living and non-living subjects.In 2016, Boulkenafet Z [2] proposed a concise approach to distinguish living from non-living bodies by distinguishing texture information other than in RGB space, using multi-level LBP features of HSV space faces + LPQ features of YCbCr space faces.It was demonstrated that features such as HSV can effectively distinguish between living and non-living individuals.

Samarth [3] have also proposed a multi-frame approach to capture the discrepancy between living and non-living subjects, and to extract optical flow histogram HOOF with dynamic texture LBP-TOP features by enhancing the face micro-motion through motion amplification with successive inputs of multiple frames.

In general, traditional machine learning-based in vivo detection algorithms focus on the design of texture features and the utilization of intrinsic properties in images and videos, and enhance the performance and robustness of the algorithms through multi-feature fusion and supplemented with other biological features as auxiliary information.

1.2. Based on Deep Learning Methods

After entering 2015, scholars have successively researched in the direction of deep learning. Xu Z [4] proposed a method based on CNN convolutional neural network to simulate LBP-TOP in

ISSN: 2688-8653

2016, which is not too effective. in 2017 Atoum Y [5] et al. proposed adding depth maps to the dataset to train the neural network, and although the results were improved, they still did not surpass the more established traditional methods in that year. Although CNN, FCN and other network structures have powerful feature extraction capabilities, they often face problems such as lack of data and overfitting in complex and changing real-world scenarios, so for a long-time detection algorithms based on deep learning frameworks are still not comparable to traditional algorithms. And the method proposed by Liu [6] et al. in 2018 using spatial and temporal auxiliary information surpasses the traditional algorithm in one fell swoop. They use spatial and temporal information as auxiliary information for supervision by combining CNN networks with rPPG with excellent performance. In the same year Yaojie Liu [7] proposed a DNN network by learning the depth map (Depth map) and proposed a method to align faces, which also achieved good results at that time. Xiao S [8] proposed a very good whole set of methods in the industry for practical application, directly put the live detection into the face detection framework such as SSD, MTCNN, by triple classification of background, real face and fake face directly, fast and effective.

Some scholars [9] also look for the differences between the live body as the original image and the non-live face as the original image after adding noise distortion to discriminate. In 2019 Liu Y [10] proposed a novel tree network to cope with these attacks considering the diversity and unknown nature of non-living attacks. In the same year Shao R [11] et al. designed a new loss function to discriminate non-live attacks.

The previously described detection methods all have a similar process, i.e., the first step feature extraction, the second step feature classification. Unlike these detection methods, in 2019 Nikitin [12] and other scholars used the same generative approach to improve the performance of the model. By synthesizing the overall picture of the 'non-living' body and increasing the number of negative samples, this method can theoretically achieve an exponential increase in the training set, so the method does not need to consider the problem of overfitting.

Deep learning based live detection algorithms focus on the use of data information, usually incorporating useful information such as RGB, depth and infrared of the image, by training a high-performance feature extraction network to distinguish between live and non-live bodies, while compared to traditional algorithms, deep learning methods increase the number of parameters while improving performance, which makes the computational complexity and cost an object to be considered in the implementation process.

This paper focuses on improving the mobilenetv2 model in two ways:

1.An attention mechanism combining spatial attention and channel attention was introduced and experimentally embedded in the network structure of MobilenetV2 [13], and a feature fusion mechanism was introduced to further improve the model accuracy by fusing shallow mid-level and deep-level features.

2.The depth map generated by PRnet [14] is used as a supervised label for pixel-level supervision, and the joint supervised model with cross entropy improves the model classification accuracy.

2. Methodology

2.1. Model Framework

The model built in this paper is based on MobilenetV2 as shown in Figure 1.

ISSN: 2688-8653

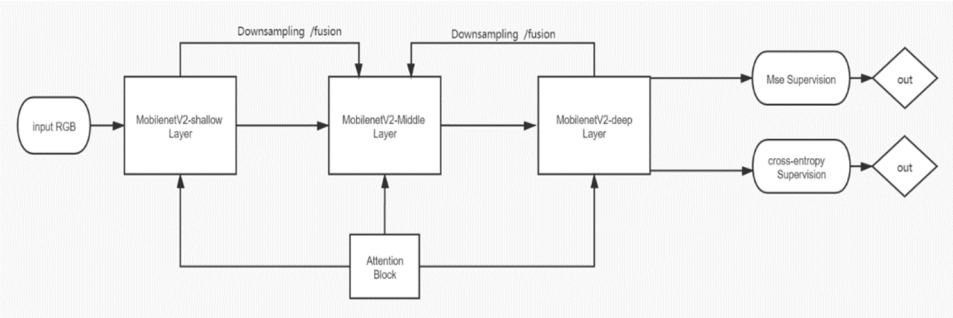

Figure 1. Based on the improved mobilenetv2 framework

Firstly, Thanks to MobilenetV2's small number of parameters, feature fusion and attention mechanisms are added without compromising accuracy. Through experiments we found that most scholars focus on deeper-level features in face Anti-spoofing tasks, yet shallow-level features are still important. According to our pre-training results, we found that the shallow features can better reflect some differences between living and non-living bodies, so we added Attention Block to the shallow features, and to the middle and deep features. The experiments show that the model accuracy has achieved better results after a single addition of Attention Block.

Secondly, this paper also fuses the MobilenetV2 shallow-medium-deep features and adds the parameter α to the feature fusion. The parameter α is generated by randomly generating a number of three respectively by feature fusion allowing the network to learn adaptively to the weights, which can also be referred to as the attention factor. Through experiments we found that the parameter α can improve part of the model performance.

2.2. Loss Function

In terms of loss function supervision, it is found that most scholars supervise the model by a single mse supervision or a single cross-entropy. However, for face Anti-spoofing tasks, pixel-level fine-grained supervision can better distinguish the difference between living and non-living subjects, so we use the depth map generated by PRnet as pixel-level supervision, with a feature map size of 32x32, so the final output of the network is 32x32, and a layer of 1x1 convolution is added afterwards as shown in Figure 2.

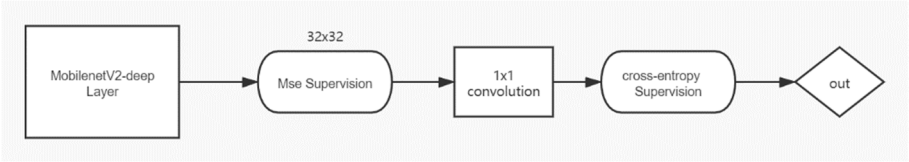

Figure 2. Loss function

We try not to take a fully connected layer in our experiments to ensure the lightweight of the number of parameters.

3. Datasets

3.1. CASIA-MFSD Datasets

CASIA - MFSD dataset. This dataset was released in 2012 Video dataset. There are 50 themes, 3 different scenarios, 150 live videos as well as 450 attack videos. All videos are positive face and can be classified according to the means of attack as print and replay. As shown in Table 1.

ISSN: 2688-8653

Table 1. CASIA-MFSD Datasets

Topics	50
Scenes	3
Live video	150
Fake video	450

3.2. Replay-Attack Datasets

Replay-Attack is a 2012 release of Video Dataset. Contains 50 themes, 1 scene, a total of 200 live videos and 1000 attack videos, all positive faces, with 1 print and 2 replay attacks. as shown in Table 2.

Table 2. Replay-Attack Datasets

Topics	50
Scenes	1
Live video	200
Fake video	1000

4. Evaluation Metrics

For the evaluation metrics we chose EER. Full name is equal error rate. There are two indicators for TPR and FPR respectively. As shown in Eq. 1 and Eq. 2

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

TPR is True Positive Rate, and FPR is False Positive Rate. The true Positive Rate is the ratio of positive cases to all positive cases as perceived by the learner, which is often referred to as the perfect or recall rate. The false positive rate is the rate at which the learner considers a positive case among all negative cases.

When FRR=FAR, it is the equal error rate EER.

5. Experiment and Results

5.1. Experimental Environment

For the experiments, our CPU is intel xeon silver 4110, graphics card is 3 RTX2080TI, training framework is pytorch framework, CUDA is selected from version 10.1, and the size of input image is set to 224x224 pixels size. The optimization method uses SGD, the learning rate is set to 1E-4, and Epoch is 100 rounds.

ISSN: 2688-8653

5.2. Experimental Results on CASIA-MFSD and Replay-Attack Datasets

In our experiments, we first intercepted the original videos of the two datasets into several frames by python scripts and saved them as images, and divided them into training set, validation set, and test set. In order to avoid overfitting, we do a test every 10 epoch in the experiment to evaluate the model effect and save the final weighting results, and evaluate the ROC curve as shown in Figure 3 and Figure 4.

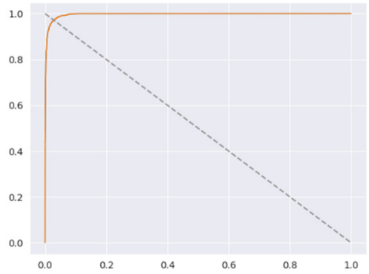

Figure 3. CASIA-MFSD ROC curve

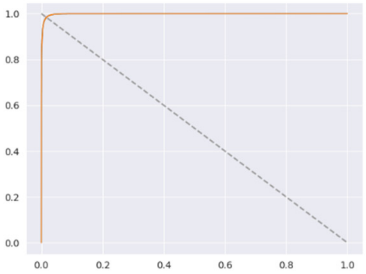

Figure 4. Replay-Attack ROC curve

As the ROC curve shows, the horizontal axis is FPR, the vertical axis is TPR, Model achieves good results.

The following figure shows a comparison with the common algorithms in recent years, as shown in Tables 3 and 4.

Table 3. Compare EER on CASIA-MFSD dataset of different methods

Algorithm	EER
Fine-tune VGG-Face [15]	5.2%
LSTM-CNN [5]	5.17%
Yang et al. [16]	4.92%
DPCNN [15]	4.5%
Siddiqui et al. [17]	3.14%
ours	2.87%

Table 4. Compare EER on Replay-Attack dataset of different methods

Algorithm	EER
Fine-tune VGG-Face [15]	8.4%
DPCNN [15]	2.9%
Yang et al. [16]	2.14%
ours	1.94%

6. Conclusion

In this paper, we improve the MobilenetV2 network by introducing an attention mechanism in it and fusing shallow, medium and deep features. The improvement improves the ability of network feature extraction. Also, the robustness of the model is improved by outputting two feature maps with simultaneous dichotomous supervision as well as pixel-level supervision. in our experiments, we reduced the EER to 2.87%, and 1.94% on the CASIA-MFSD, and Replay-Attack datasets, respectively. Experiments show that the improved model and the supervision

ISSN: 2688-8653

of the loss function in this paper can improve the effectiveness of the model, but there is still much room for improvement. it can be seen that a single input is not optimal for the face Anti-spoofing task, and how to solve the existing dataset is not large enough, realize the data increase, and improve the generalization ability of the model are some problems well worth studying in the future.

Acknowledgments

Special thanks to Southwest University for Nationalities for providing the experimental conditions and technical support for the work in this paper.

References

[1] D Wen, Han H , Jain A K . Face Spoof Detection with Image Distortion Analysis[J]. IEEE Transactions on Information Forensics & Security, 2015, 10(4):746-761.

[2] Boulkenafet Z , Komulainen J , Hadid A . Face Spoofing Detection Using Colour Texture Analysis[J]. IEEE Transactions on Information Forensics & Security, 2017, 11(8):1818-1830.

[3] Samarth Bharadwaj. Face Anti-spoofing via Motion Magnification and Multifeature Videolet Aggregation, 2014.

[4] Xu Z , Li S , Deng W . Learning temporal features using LSTM-CNN architecture for face Anti-spoofing [C]// Pattern Recognition. IEEE, 2016.

[5] Atoum Y , Liu Y , Jourabloo A , et al. Face Anti-spoofing using patch and depth-based CNNs[C]// The International Joint Conference on Biometrics (IJCB 2017). IEEE, 2018.

[6] Hernandez-Ortega J , Fierrez J , Morales A , et al. Time Analysis of Pulse-Based Face Anti-spoofing in Visible and NIR[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018.

[7] Yaojie Liu, Amin Jourabloo, Xiaoming Liu, Learning Deep Models for Face Anti-spoofing: Binary or Auxiliary Supervision ,CVPR2018.

[8] Xiao S , Xu Z, Liangji F, et al. Discriminative Representation Combinations for Accurate Face Spoofing Detection[J]. Pattern Recognition, 2018.

[9] Face De-Spoofing: Anti-spoofing via Noise Modeling, ECCV2018.

[10] Liu Y , Stehouwer J , Jourabloo A , et al. Deep Tree Learning for Zero-Shot Face Anti-spoofing[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[11] Shao R , Lan X , Li J , et al. Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[12] Nikitin M Y , Konushin V S , Konushin A S . Face Anti-spoofing with joint spoofing medium detection and eye blinking analysis[J]. Computer Optics, 2019, 43(4):618.

[13] Sandler M , Howard A , Zhu M , et al. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation[J]. 2018.

[14] Feng Y , Wu F , Shao X , et al. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network[J]. arXiv, 2018.

[15] Lei L , Feng X , Boulkenafet Z , et al. An original face Anti-spoofing approach using partial convolutional neural network[C]// 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2017.

[16] Yang J , Lei Z , Li S Z . Learn Convolutional Neural Network for Face Anti-spoofing[J]. Computer ence, 2014, 9218:373-384.

[17] Siddiqui T A , Bharadwaj S , Dhamecha T I , et al. Face Anti-spoofing with multifeature videolet aggregation [C]// 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2017.