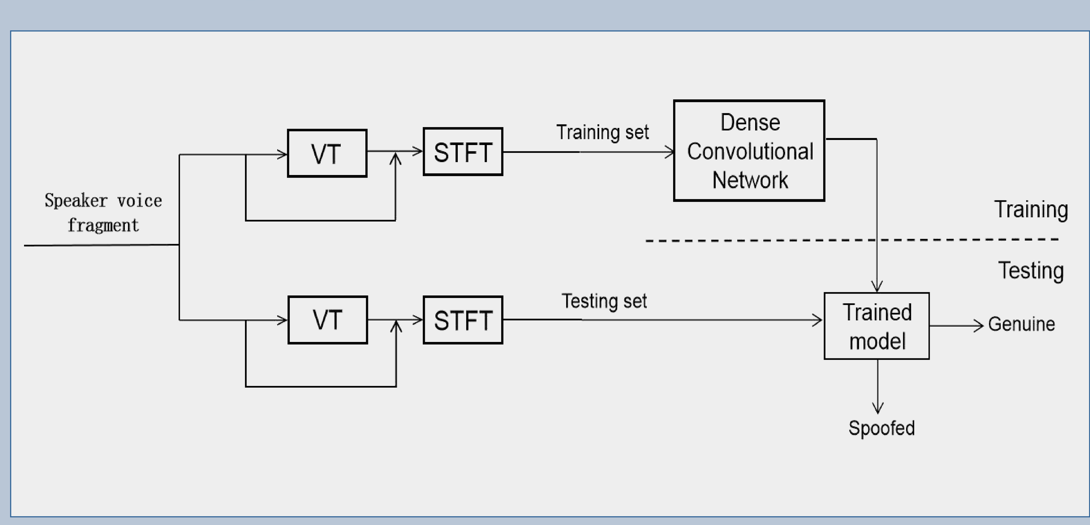
Detection of Voice Transformation Spoofing Based on Dense Convolutional Network

Yong Wang & Zhuoyi Su
Guangdong Polytechnic Normal University, GZ, CHN

Abstract

- Nowadays, speech spoofing is so common that it presents a great challenge to social security. Thus, it is of great significance to recognize a spoofed speech from a genuine one.
- Most of the current researches have focused on voice conversion (VC), synthesis and recapture which mimic a target speaker to break through ASV systems by increased false acceptance rates. However, there exists another type of spoofing, voice transformation (VT), that transforms a speech signal without a target in order 'not to be recognized' by increased false reject rates. VT has received much less attention. Thus, in this paper, we investigate the model of VT and propose a method using a very deep dense convolutional network with 135 layers to detect VT spoofed speeches from genuine speeches. The experimental results show that the average accuracies over intra-database and cross-database outperform the reported state-of-the-art methods.

The Experimental Process



Acknowledgement

- This work was supported by the National Natural Science Foundation of China (61672173), the Characteristic Innovation Project of Guangdong Province Ordinary University (2015KTSCX083), the Natural Science Foundation of Guangdong Province (2014A030313623) and the Guangzhou science and technology project (201803010081).

Methods

- In a conventional CNN, the output of the previous layer $X_{\ell-1}$ is transmitted to the next layer as input by a non-linear operation H_ℓ to get the output X_ℓ.

$$X_\ell = H_\ell(X_{\ell-1}) \qquad (1)$$

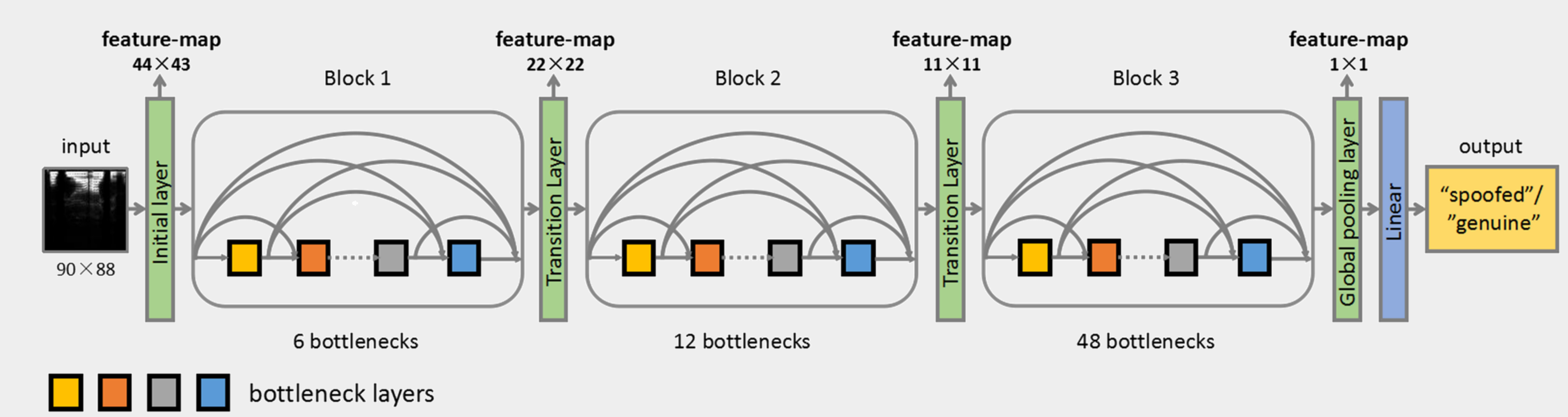
- It is difficult to train a conventional CNN as degradation occurs with the increment of layers. To have a good inhibitory effect on the degradation, Residual Networks (ResNets)[1], Highway Networks[2] and FractalNets[3] create short paths $X_{\ell-\alpha}$ from early layers to later layers as shown in Equ.(2).

$$X_\ell = H_\ell(X_{\ell-1}) + X_{\ell-\alpha} \qquad (2)$$

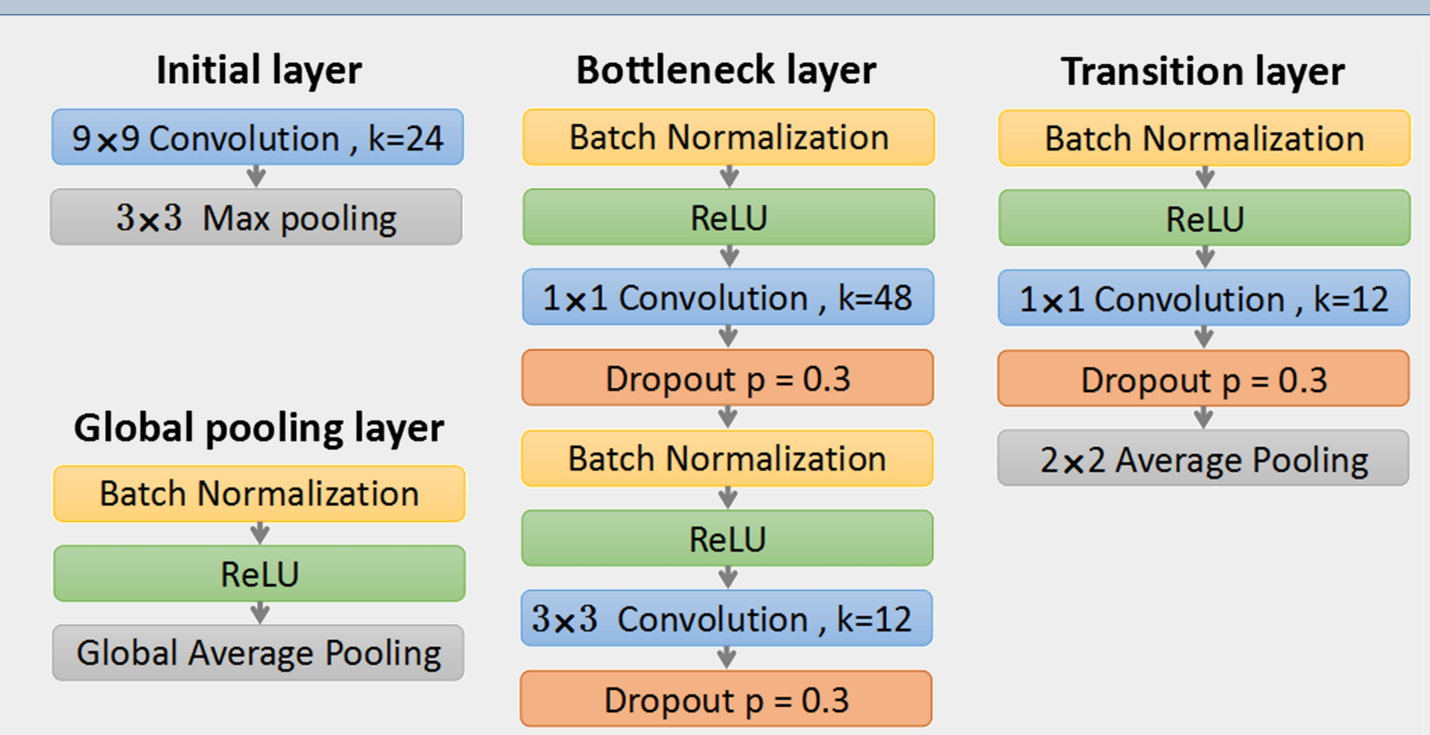
- However, recent research suggests that this type of connection leads to the fact that many layers contribute very little but occupy a large amount of computation [4]. Thus, an improved structure of ResNet named Dense Convolutional Network (DenseNet) was proposed to avoid this problem. In a DenseNet, any layer has direct connections to all subsequent layers, as shown in Equ.(3)

$$X_\ell = H_\ell([X_0, X_1, \ldots, X_{\ell-1}]) \qquad (3)$$

- where $X_0, X_1, X_{\ell-1}$ represent the output of the previous layer of layer ℓ and $[\ldots]$ on behalf of the concatenation operation. Furthermore, the output dimension of each layer has k feature maps, where k is usually set to a small value.



The architecture of the proposed network (135 layers)



- The inner structures of each kind of the layers. BatchNormalization, ReLU, Dropout and Pooling are ofthe operations before and after the convolution layer. k refers to the number of convolution kernels.

Results

a.The detection accuracy of intra-database evaluation

Training set	Testing set	Proposed Method	Liang's Method	Wu's Method
TIMIT_1	TIMIT_2	99.45%	96.52%	95.87%
NIST_1	NIST_2	98.04%	95.93%	94.56%
UME_1	UME_2	97.56%	94.85%	93.63%

b.The detection accuracy of cross-database evaluation

Case	Training set	Testing set	Proposed Method
Case1	TIMIT_1/NIST_1	UME_2	96.45%
Case2	NIST_1/UME_1	TIMIT_2	95.26%
Case3	TIMIT_1/UME_1	NIST_2	80.20%

Reference

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," pp. 770–778, 2015.
- [2] Rupesh Kumar Srivastava, Klaus Greff, and Jurgen Schmidhuber, "Training very deep networks," CoRR, vol. abs/1507.06228, 2015.
- [3] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," CoRR, vol. abs/1605.07648, 2016.
- [4] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger, "Deep networks with stochastic depth," pp. 646–661, 2016

Conclusion

- In this paper, a method based on dense convolutional network for detecting spoofed speech from genuine speech is presented. Deep features can be automatically extracted by the 135-layer DenseNet. It achieves computing efficiency by careful optimization of kernel reduction and by the employment of bottleneck layers. The experimental results indicates that it is superior to the state-of-the-art methods. The future work will focus on the extraction of deeper features to further improve the accuracy.

IEEE 2019 ICASSP