Detection of regions with the least impact on true and fake image classification through reinforcement learning

Wei Li¹, Peng Qiao and Yong Dou

Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha, 410005, China

LWei_gfkd@163.com, pengqiao@nudt.edu.com, yongdou@nudt.edu.com

¹wei.li@nudt.edu.cn

Abstract. With the development of artificial intelligence [1-3], convolutional neural networks (CNNs) have made significant progress in image generation and manipulation. The generated images using facial image synthesis methods, e.g., Faceswap [4], Deepfakes [5], Face2face [6], NeuralTextures [7], are very realistic and even difficult to distinguish by human eyes. It poses serious challenges to social and personal security. There are many methods that can detect synthetic facial images with an accuracy rate of over 99%. But the synthetic facial image classification model can only give a true or fake result, and it is impossible to know which part is the discordant area. Therefore, we propose and experimentally verify a method that using reinforcement learning to select the region in the image that has the least impact on the classification model.

1. Introduction

With the development of convolutional neural networks (CNNs), we have witnessed great progress in image generation and manipulation technologies. As the synthetic facial images become extremely realistic, it technically supports misbehaviour related to the facial image synthesis methods. For example, on the Internet, there are some false and inappropriate political speeches, e.g., Obama complains Trump, and pornographic videos, in which the faces in the videos are replaced with the faces of stars. Another example is Zao, an APP that can change faces between yourself and the people in the video with just a selfie in a few seconds.

In order to solve these security risks, many synthetic facial image classification models have appeared, which can effectively identify true and fake images. But they can only give a true or fake result, and cannot give specific classification details. Therefore, this paper intends to determine the discordant area in the image by selecting the area in the image that has the least impact on true and fake image classification.

Although deep convolutional neural networks (CNNs) have strong feature extraction capabilities, they still lack the ability to dynamically select regions. Reinforcement learning is good at action selection. Therefore, in this paper, we use reinforcement learning to select the unimportant area in the image. Firstly, according to the landmarks of the image, the image will be divided into 8 regions (left and right eyes, nose, mouth, left and right cheeks, left and right corners of the mouth). Secondly, we train a classification network, XceptionNet, which can effectively classify true and fake images. Thirdly, the unimportant areas in the images to be eliminated are selected through actions, and the
eliminated images are sent to XceptionNet for fine-tuning. If the classification accuracy does not decrease much, it means that the area has little effect on the classification.

The contributions of this paper are two-folds:

- Able to effectively select the area that has the least impact on true and fake image classification.
- Another application of reinforcement learning in computer vision.

2. Related Work

In this section, we give a brief review of forgery detection methods and application of reinforcement learning in computer vision.

2.1. Forgery Detection Methods

As shown in Fig.1, the images that generated using facial image synthesis methods are very realistic, and some images cannot even be distinguished by the human eyes. Therefore, some facial image classification methods were proposed, which can effectively identify true and fake images. In [8], Rossler et al. use XceptionNet architecture to detect real and synthesized images, and the accuracy is higher than other methods [9-11]. XceptionNet evolved directly from Inception v3, which introduces the structure of Residual learning. But it has fewer parameters than Inception V3 and is the classification network with the highest classification accuracy currently, which can reach more than 99%.

2.2. Application of reinforcement learning in computer vision

**Visual Tracking:** In [12], Sangdoon Yun et al. proposed a new type of tracking network, ADNet, which sequentially executes actions obtained through deep reinforcement learning to accurately track target. The tracker moved through the predicted action in the current state, and then predict the next action based on the moved position. The action space consists of eleven types of actions including translation moves, scale changes, and stopping action. Through the evaluation of the OTB dataset, the ADNet has been verified to achieve competitive performance, which is three times faster than state-of-the-art deep network-based tracker.
**Image Classification:** In [13], Pang proposed an image cropping model based on deep reinforcement learning to solve the problem of low image classification efficiency and accuracy, which is caused by the introduction of irrelevant noise and redundant information in the image feature extraction process. Regarding image cropping as a Markov model of a sequential decision-making process, the agent interacts with the cropping environment in reinforcement learning, and guides the agent to optimize actions by constructing a new reward function to find the best cropped image on the original image.

In [14], Yang proposed a self-reinforcing network algorithm. The algorithm has a feature decision intelligent system, which controls the classification time of each input image. For images that cannot be classified, it will select an image transformation and return the transformed image to the image classification network. The self-reinforcing network algorithm can reduce the original error rate of the image classification network by 18.82% through the decision of the feature decision intelligent system.

3. Methods

In this section, we describe the proposed area selection network, ASNet, which uses reinforcement learning to gradually select and eliminate unimportant areas in the image, as shown in Fig.2.

![Fig.2 Area selection network structure](image)

3.1. Face extraction and alignment

In order to eliminate the influence of factors such as face angle, we standardize all images through face alignment, making the subsequent training effect better. Given the distribution coordinates of the regular face, the image is aligned with the regular face through the extracted landmarks. Then the face image is extracted from the aligned image.

Aligning the face can not only improve the classification accuracy of the classification network, XceptionNet, but also make dividing face area more regular in the subsequent region deletion stage. As shown in Fig.2, an image is divided into 8 regions (left and right eyes, nose, mouth, left and right cheeks, left and right corners of the mouth).
3.2. Area selection network
Inspired by the AlphaGo Zero [15] network structure, the area selection network (ASNet) structure uses a residual structure. As shown in Fig. 3, it uses four residual blocks, that Block1 keeps the same number and size, Block2 has more layers and the size is halved. The input of the action network is batch-size images that are treated as a whole, and the output is an 8-dimensional vector that determines which area of the image is eliminated. When the selected area is still there, it will be eliminated, otherwise, the map of the area will be restored. This can avoid the impact of classification due to the common influence of different regions.

3.3. Reinforcement learning
In this paper, we use reinforcement learning to dynamically select the image areas that need to be eliminated at each step. These dynamic choices are not available in deep convolutional neural networks. In the model, the environment determines the reward value by fine-tuning the XceptionNet. If the classification accuracy of the XceptionNet drops below the threshold, a negative reward will be given, otherwise, a positive reward will be given.

For completeness, we provide the reinforcement learning algorithm below.

**Algorithm** Model-based reinforcement learning through model prediction to eliminate areas

1. gather dataset ($D_{fake}$) including true images and fake images generated by a certain facial image synthesis method
2. initialize empty dataset ($D_{RL}$), target network ($Q_{target}$) and evaluation network ($Q_{eval}$)
3. for episode=1 to max_episode do
4. initialize environment (Env) and get the initial state
5. for iter=1 to max_iter do
6. ASNet gets the action according to the current state ($S_t$), that is, the next area to be eliminated
7. Env gets the next state ($S_{t+1}$) based on the action ($A_t$), $S_{t+1} = Q_{eval}(A_t, S_t)$
8. use the next state to fine-tune XceptionNet for 3 epochs, and get the reward value ($R_t$)
9. ADNet learning, $Q_{target}(S_t) \leftarrow R_t + \gamma * Q_{eval}(S_{t+1})$
10. if the classification accuracy of XceptionNet is below the threshold
11. Break
12. end for
13. end for

4. Experiments
In this section, we first describe the training details, then discuss the experimental results.
4.1. Training Settings

4.1.1. Dataset
We conduct the following experiments using FaceForensics dataset [17], including real videos and their corresponding synthesized videos generated by four methods, namely Faceswap [4], Deepfakes [5], Face2face [6] and NeuralTextures [7]. In this paper, only the NeuralTextures dataset is used.

Firstly, we extract about 50 images from each video among the real and synthesized videos from NeuralTextures, so that each type of videos has 50,000+ images. Secondly, we use the face extraction and alignment method mentioned in section 3.1 to extract the facial images from the previous images. Thirdly, we use the images from the first 800 videos as the training dataset, and the images from the last 200 videos as the test dataset.

4.1.2. Training and fine-tuning XceptionNet
At the beginning, we train the XceptionNet that can accurately classify true and fake facial images. We transfer the XceptionNet pre-trained on ImageNet dataset to synthesized facial image classification on aligned facial images, which are from the NeuralTextures dataset in FaceForensics. We first train the last fully connected (FC) layer, which replaces the original one in XceptionNet for ImageNet classification task and is randomly initialized, for 3 epochs. After that, we train the entire network for another 10 epochs.

In the reinforcement learning stage, we load the previous XceptionNet model at each step, and then fine-tune it for 3 epochs.

For optimization, we use Adam optimizer [16] and the following hyperparameters in the two stages: the learning rate of parameters is set to 1e-6 for convolutional layer, and is set to 1e-5 for FC layer; $\beta_1=0.9$ and $\beta_2=0.999$; the batch size is set to 16.

4.1.3. Action
As shown in Fig.3, the action is an 8-dimensional vector, select the area of the facial image to be processed. And each action can only process one area. When the selected area is still there, the area is eliminated, otherwise the area is restored. All images in the batch-size images take the same action, that is, eliminate the same area.

4.1.4. Reward
The reward value is determined by the number of the total elimination area after each action. When it is not the elimination action, the reward is +1, otherwise, the reward is -1. Since the image has a total of 8 areas, the total reward is between 0 and 8.

The termination of the game is controlled by the fine-tuned XceptionNet classification accuracy. When the classification accuracy falls below the threshold, which is set to 95% here, the game is over.

4.2. Results
After the training converges, 3 to 6 areas are generally eliminated, and the order of area elimination is 1-0-3-2-4-7. As shown in Fig.4. Because of the resolution, expression, angle, etc. of the facial images, the result of synthetic facial images is different. The more realistic the synthetic facial images are, the more areas can be eliminated, otherwise, the less areas that can be eliminated.

![Fig.4 Order of area elimination](image-url)
According to the order of area elimination, it can be seen that the less important areas for image classification, the earlier they are eliminated, and they will all be eliminated. The more important areas for image classification, the later they are eliminated, and even cannot be eliminated. Therefore, we rank the importance of the 8 areas of the image as follows for NeuralTextures dataset: mouth and left corner of the mouth; right corner of the mouth; right cheek; left cheek; nose; left eye; right eye. Therefore, the most discordant area is the mouth.

4.3. Loss
As shown in Fig.5, from the perspective of the loss function, it is also convergent. But there are serious shocks.

![Fig.5 The loss value trend chart](image)

The loss value in the figure can be as high as 100+, appearing in the first 100 iterations and not shown in the Fig.5. It also can be as low as 1e-2, but it will still oscillate to single digits later. There are two reasons for this situation: 1) in each epoch, it will take out a new batch-size images, that these images are quite different; 2) as mentioned in section 4.2, the synthesis effect of different images is inconsistent.

5. Conclusion
In this paper, we use reinforcement learning to select the areas that has the least impact on the classification of true and fake images. Based on the results and discussions presented above, the conclusions are obtained as below:

- For the classification of true and fake images, different areas have different effects on the classification. For the NeuralTextures dataset, the 8 areas of the image have the following order of influence on classification: mouth and left corner of the mouth; right corner of the mouth; right cheek; left cheek; nose; left eye; right eye. Therefore, the most discordant area is the mouth.
- The application of reinforcement learning in computer vision can also achieve good results, especially in dynamic selection.

Acknowledgments
This paper is supported by the National Natural Science Foundation of China (Grant No. 61902415).

References
[1] Yu, Ke, et al. "Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[2] Chen Huang, Simon Lucey, Deva Ramanan. "Learning Policies for Adaptive Tracking with
Deep Feature Cascades." ICCV (2017 Spotlight).

[3] Zhou, Kaiyang, and Yu Qiao. "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward." AAAI. 2018.

[4] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2018-10-29.

[5] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2018-10-29.

[6] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Demo of Face2Face: real-time face capture and reenactment of RGB videos. In ACM SIGGRAPH 2016 Emerging Technologies; 2016; pp. 1–2.

[7] Thies, J.; Zollhöfer, M.; Nießner, M. Deferred Neural Rendering: Image Synthesis using Neural Textures. arXiv preprint arXiv:1904.12356 2019.

[8] Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[9] Raghavendra, R.; Raja, K.B.; Venkatesh, S.; Busch, C. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017, pp. 1822–1830.

[10] Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017, pp. 1831–1839.

[11] Rahmouni, N.; Nozick, V.; Yamagishi, J.; Echizen, I. Distinguishing computer graphics from natural images using convolution neural networks. 2017 IEEE Workshop on Information Forensics and Security (WIFS). IEEE, 2017, pp. 1–6.

[12] Yun S, Choi J, Yoo Y, et al. Action-decision networks for visual tracking with deep reinforcement learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2711-2720.

[13] Pang Z. Image Classification Research Based on Fusion of Deep Reinforcement Learning and Relation Network [D]. Jiangxi University of Science and Technology, 2019. (in Chinese)

[14] Yang Z. Research on Applying Deep Reinforcement Learning in Image Based Control and Image Classification Tasks [D]. South China University of Technology, 2018. (in Chinese)

[15] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676):354-359.

[16] Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.

[17] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 2018.