Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICASSP.2019.8682602

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
CAPSULE-FORENSICS: USING CAPSULE NETWORKS TO DETECT FORGED IMAGES AND VIDEOS

Huy H. Nguyen*, Junichi Yamagishi*†‡, and Isao Echizen*†

*SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
†National Institute of Informatics, Tokyo, Japan
‡The University of Edinburgh, Edinburgh, UK
Email: {nhhuy, jyamagis, iechizen}@nii.ac.jp

ABSTRACT

Recent advances in media generation techniques have made it easier for attackers to create forged images and videos. State-of-the-art methods enable the real-time creation of a forged version of a single video obtained from a social network. Although numerous methods have been developed for detecting forged images and videos, they are generally targeted at certain domains and quickly become obsolete as new kinds of attacks appear. The method introduced in this paper uses a capsule network to detect various kinds of spoofs, from replay attacks using printed images or recorded videos to computer-generated videos using deep convolutional neural networks. It extends the application of capsule networks beyond their original intention to the solving of inverse graphics problems.

Index Terms— computer-generated video, replay attack, forgery detection, capsule network

1. INTRODUCTION

Forged images and videos can be used to bypass facial authentication and to create fake news media. The quality of manipulated images and videos has seen significant improvement with the development of advanced network architectures and the use of large amounts of training data. This has dramatically simplified the creation of facial forgeries. Nowadays, the only thing needed to create a forged facial image is simply a short video of the target person [1, 2] or an ID photo [3, 4]. The techniques developed by Chung et al. [4] and Suwajanakorn et al. [5] can improve the ability of attackers to learn the mapping between speech and lip motion, enabling the creation of fully synthesized audio-video data for any person. In this age of social networks serving as major sources of information, fake news with manipulated multimedia can quickly spread and have significant effects. The deepfake phenomenon [6] is a good example of this threat—any person with a personal computer can create videos incorporating the facial image of any celebrity by using a human image synthesis technique based on artificial intelligence.

Several countermeasures have been proposed to deal with manipulated images and videos. However, most of them are aimed at particular types of attacks. For example, local binary pattern (LBP)-based methods [7, 8] are effective against replay attacks in which the attacker places a printed photo or displays a video on a screen in front of the camera. However, the eyes-focused method designed to detect a deepfake forgery [9] can fail with the replay attack when the video displayed is of the actual target person. Other methods have more generalized ability; for instance, Fridrich and Kodovsky’s method [10] can be applied for both steganalysis and detecting facial reenactment videos. However, its performance on secondary tasks is limited in comparison with task-specific methods like that of Rossler et al. [11]. Moreover, while some methods can detect a single forged image [11, 12, 13], others require video input [9].

This paper presents a method that uses a capsule network to detect forged images and videos in a wide range of forgery scenarios, including replay attack detection and (both fully and partially) computer-generated image/video detection. This is pioneering work in the use of capsule networks [14, 15, 16], which were originally designed for computer vision problems, to solve digital forensics problems. A comprehensive survey of state-of-the-art related work and intensive comparisons using four major datasets demonstrated the superior performance of the proposed method.

2. RELATED WORK

In this section, we group forgery detection approaches into replay attack detection and computer-generated image/video detection on the basis of the features used and their target. Note that some approaches are two-fold while others are applicable only to certain types of attacks. We also provide some basic information about capsule networks and the dynamic routing algorithm that made this kind of network practical.
2.1. Replay Attack Detection

Prior to the current deep learning era, LBP methods were the primary defense against replay attacks [7, 8]. The method introduced by Kim et al. [17], which is based on local patterns of the diffusion speed (local speed patterns), achieves higher accuracy than that of LBP-methods. Now, with the introduction of deep learning, the ability to detect replay attacks has been greatly improved. The method of Yang et al. [18] uses a support vector machine to classify features extracted by a pre-trained convolutional neural network (CNN). That of Menotti et al. [19] uses a similar procedure but optimizes the filters in an available high-performance CNN architecture. The method of Alotaibi and Mahmood [20] uses non-linear diffusion based on an additive operator splitting scheme in their own CNN. The recently introduced method of Ito et al. [21] leverages a pre-trained CNN and utilizes the whole image instead of only the extracted face region.

2.2. Computer-Generated Image/Video Detection

There are several state-of-the-art methods for detecting images or videos generated by computer using, for example, a deepfake technique for face swapping [6], the Face2Face method for facial reenactment [1], or the deep video portraits technique [2] for the purpose of forgery. Fridrich and Kodovsky [10] proposed a hand-crafted-feature noise-based approach for steganalysis that can also be used for forgery detection. Cozzolino et al. [22] implemented a CNN version of this approach. Raghavendra et al. [23] described the special case of fine-tuning two available CNNs while Rossler et al. [11] used only one CNN. Bayar and Stamm [24], Rahmouni et al. [25], Afchar et al. [13], Quan et al. [26], and Li et al. [9] proposed their own networks. Li et al.’s network [9], for example, is video based and uses temporal information to detect eye blinking. We used a hybrid approach [12] incorporating part of a pre-trained VGG (Visual Geometry Group)-19 network [27] and a proposed CNN. Zhou et al. [28] proposed a two-stream network.

2.3. Capsule Networks

Hinton et al. [14] addressed the limitations of CNNs applied to inverse graphics tasks and laid the foundation for a more robust “capsule” architecture in 2011. However, this complex architecture could not be effectively implemented at the time due to the lack of an efficient algorithm and the limitations of computer hardware. Instead, easy-to-design easy-to-train CNNs became widely used. Now, with the introduction of the dynamic routing algorithm [15] and the expectation-maximization routing algorithm [16], capsule networks have been implemented with remarkable initial results. Two recent studies demonstrated that, with the agreement between capsules calculated by the dynamic routing algorithm, the hierarchical pose relationships between object parts can be well described. This has improved the accuracy of vision tasks. Application of a capsule network to the forensics task, the focus of this paper, is a challenging problem. However, the agreement between capsules achieved by using the dynamic routing algorithm could boost detection performance on complex and nearly flawless forged images and videos.

3. CAPSULE-FORENSICS

3.1. Overview

The proposed method (Fig. 1) works for both images and videos. For video input, the video is split into frames in the pre-processing phase. The classification results (posterior probabilities) are then acquired from the frames. The probabilities are averaged in the post-processing phase to get the final result. The remaining parts are constructed the same way as when the input is an image.

In the pre-processing phase, faces are detected and scaled to 128 × 128. Like we did in our previous work [12], we use part of the VGG-19 network [27] to extract the latent features, which are the inputs to the capsule network. Unlike we did in our previous work, we take the output of the third maxpooling layer instead of three outputs before the ReLU layers. We do this because we need to reduce the size of the inputs to the capsule network.

3.2. Capsule Design

The proposed network consists of three primary capsules and two output capsules, one for real and one for fake images (Fig. 2). The latent features extracted by part of the VGG-19 network [27] are the inputs, which are distributed to the three primary capsules (Fig. 3). As in our previous work [12], statistical pooling, which is important for forgery detection, is used. The outputs of the three capsules ($u_{ij}$) are dynamically...
Algorithm 1 Dynamic routing between capsules.

procedure ROUTING(\(u_{ij}, W, r\))
\[ W \leftarrow W + \text{rand(size}(W)) \]
\[ u_{ij} \leftarrow W_j \cdot \text{squash}(u_{ij}) \quad \forall W_j \in R^{m\times n} \]
for all input capsule \(i\) and all output capsules \(j\) do
\[ b_{ij} \leftarrow 0 \]
for \(r\) iterations do
for all input capsules \(i\) do \(c_i \leftarrow \text{softmax}(b_i)\)
for all output capsules \(j\) do \(s_j \leftarrow \sum_i c_i u_{ij}\)
for all output capsules \(j\) do \(v_j \leftarrow \text{squash}(s_j)\)
for all input capsules \(i\) and output capsules \(j\) do
\[ b_{ij} \leftarrow b_{ij} + u_{ij} \cdot v_j \]
return \(v_j\)

routed to the output capsules \((v_j)\) for \(r\) iterations using Algorithm 1. The network has approximate 2.8 million parameters, a relatively small number for such networks. We slightly improved the algorithm of Sabour et al. [15] by adding Gaussian random noise to the 3-D weight tensor \(W\) and applying one additional squash (equation 1) before routing by iterating. The added noise helps reduce over-fitting while the additional equation keeps the network more stable. The outputs of the primary and output capsules are illustrated in Fig. 4.

\[
v_j = \text{squash}(s_j) = \frac{\|s_j\|^2}{1 + \|s_j\|^2 \|s_j\|} s_j \quad (1) \]

Unlike Sabour et al.’s work [15], we use the cross-entropy loss function:

\[
L = - (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (2)
\]

where \(y\) is the ground truth label and \(\hat{y}\) is the predicted label calculated using equation 3, in which \(m\) is the dimension of the output capsule \(v_j\).

\[
\hat{y} = \frac{1}{m} \sum_i \text{softmax} \left( \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} \right) \quad (3)
\]

Fig. 3. Detailed design of primary capsule. Upper numbers indicate number of filters (depth) while lower number indicate size of outputs of corresponding filters.

Fig. 4. Average results calculated by primary capsules and output capsules from real and fake images generated with Face2Face method [1]. Three primary capsules have significantly different reactions between real and fake inputs. Although their weights are also different, there is strong agreement in the output capsules.

The use of equation 3 instead of simply using the length of the output capsules [15] promotes separation between the two output capsules on each dimension.

4. EVALUATION

To evaluate the advantage of using random noise, we tested the proposed method with and without using random noise (Capsule-Forensics-Noise and Capsule-Forensics, respectively). The random noise was generated from a normal distribution \(N(0, 0.01)\) and was used in the training phase only. Two iterations \((r = 2)\) were used in the dynamic routing algorithm. We used the half total error rate (HTER) \(\frac{\text{FRR} + \text{FAR}}{2}\) and accuracy \(\frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}\) as metrics.

4.1. Replay Attack Detection

To determine the ability of the proposed method to detect replay attacks, we compared its performance with that of eight state-of-the-art detection methods on the well-known Idiap REPLAY-ATTACK dataset [7]. As shown in Table 1, the proposed method with random noise (Capsule-Forensics-Noise), as well as our previous method [12], had an HTER of zero.

Table 1. Half total error rate of state-of-the-art detection methods on REPLAY-ATTACK dataset [7].

<table>
<thead>
<tr>
<th>Method</th>
<th>HTER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chigovska et al. [7]</td>
<td>17.17</td>
</tr>
<tr>
<td>Pereira et al. [8]</td>
<td>08.51</td>
</tr>
<tr>
<td>Kim et al. [17]</td>
<td>12.50</td>
</tr>
<tr>
<td>Yang et al. [18]</td>
<td>02.30</td>
</tr>
<tr>
<td>Menotti et al. [19]</td>
<td>00.75</td>
</tr>
<tr>
<td>Alotabib et al. [20]</td>
<td>10.00</td>
</tr>
<tr>
<td>Ito et al. [21]</td>
<td>00.43</td>
</tr>
<tr>
<td>Nguyen et al. [12]</td>
<td><strong>00.00</strong></td>
</tr>
<tr>
<td>Capsule-Forensics</td>
<td>00.28</td>
</tr>
<tr>
<td>Capsule-Forensics-Noise</td>
<td><strong>00.00</strong></td>
</tr>
</tbody>
</table>
4.2. Face Swapping Detection

We determined the ability of our proposed method to detect face swapping using a deepfake technique on the deepfake dataset proposed by Afchar et al. [13] at both the frame and video levels. As shown in Tables 2 and 3, our proposed method with random noise (Capsule-Forensics-Noise) had the highest accuracy in both cases.

Table 2. Accuracy of face swapping detection at frame level on deepfake dataset [13].

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meso-4 [13]</td>
<td>89.10</td>
</tr>
<tr>
<td>MesoInception-4 [13]</td>
<td>91.70</td>
</tr>
<tr>
<td>Nguyen et al. [12]</td>
<td>92.36</td>
</tr>
<tr>
<td>Capsule-Forensics</td>
<td>94.47</td>
</tr>
<tr>
<td>Capsule-Forensics-Noise</td>
<td>95.93</td>
</tr>
</tbody>
</table>
7. PREFERENCES


