electronics

MDPI

Article

CNN-Based Ternary Classification for Image Steganalysis

Sanghoon Kang [1], Hanhoon Park [1,*] and Jong-Il Park [2]

[1] Department of Electronic Engineering, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea; totohoon01@naver.com

[2] Department of Computer Science, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul 04763, Korea; jipark@hanyang.ac.kr

* Correspondence: hanhoon.park@pknu.ac.kr; Tel.: +82-51-629-6225

Received: 11 September 2019; Accepted: 23 October 2019; Published: 26 October 2019

Abstract: This study proposes a convolutional neural network (CNN)-based steganalytic method that allows ternary classification to simultaneously identify WOW and UNIWARD, which are representative adaptive image steganographic algorithms. WOW and UNIWARD have very similar message embedding methods in terms of measuring and minimizing the degree of distortion of images caused by message embedding. This similarity between WOW and UNIWARD makes it difficult to distinguish between both algorithms even in a CNN-based classifier. Our experiments particularly show that WOW and UNIWARD cannot be distinguished by simply combining binary CNN-based classifiers learned to separately identify both algorithms. Therefore, to identify and classify WOW and UNIWARD, WOW and UNIWARD must be learned at the same time using a single CNN-based classifier designed for ternary classification. This study proposes a method for ternary classification that learns and classifies cover, WOW stego, and UNIWARD stego images using a single CNN-based classifier. A CNN structure and a preprocessing filter are also proposed to effectively classify/identify WOW and UNIWARD. Experiments using BOSSBase 1.01 database images confirmed that the proposed method could make a ternary classification with an accuracy of approximately 72%.

Keywords: image steganalysis; WOW; UNIWARD; ternary classification; convolutional neural network (CNN)

1. Introduction

Interest in information security technologies, such as image steganography/steganalysis, has significantly grown because of the universalization of digital multimedia and communication. Image steganography is a technique in which a secret message is embedded into an image, called *cover image*, and the message-embedded image, called *stego image*, is transmitted through a public channel without gaining the attention of a third party, thereby implementing covert communication. The image steganalysis is the reverse process of image steganography, which aims to determine whether or not the image to be tested contains a secret message and then finds out the hidden message.

The performance of image steganographic methods depends on two conflicting parameters: embedding capacity, which represents how many messages we can hide, and the image quality after embedding, which is closely related to message concealment. Therefore, most image steganographic methods have achieved a high embedding capacity at the expense of low image quality after embedding, and vice versa.

Early image steganographic methods include the least significant bit (LSB) substitution method [1], which replaces the least significant bits of image pixels by secret messages, and the pixel value

differencing (PVD) methods [2–4] that determine the amount of secret messages to be embedded in proportion to the difference between adjacent pixels. These early image steganographic methods sequentially embed secret messages into all pixels of an image, although they have been recently extended to embed messages in randomly selected pixels using pseudo-random generators for secure message hiding [5,6].

Sequentially embedding secret messages into all pixels of an image is well known to change the statistical characteristics of the image. In Figure 1, the solid line refers to a probability density function (PDF) of the differences between two adjacent pixels on a cover image. The dot lines refer to the different PDFs on the stego images created by different image steganographic methods. The PDF of the LSB stego image is significantly different from that of the cover image in the section where the differences are small. This statistical difference is easily detected by statistical attacks, such as the RS analysis in [7]. Thus, image steganographic methods have come to consider more how not to be detected by steganalytic attacks than how many messages to embed. To avoid statistical attacks, image steganographic methods began to consider where the message would be embedded. Methods such as HUGO [8], WOW [9], and UNIWARD [10] tried to embed a message into only pixels with a small distortion, mainly on image edges, by analyzing the distortion caused by embedding a message into each pixel. For example, HUGO measured the embedding distortion by reverse-engineering the processes of the subtractive pixel adjacency matrix (SPAM) [11], a steganalytic method that calculated a co-occurrence matrix for the differences of the adjacent pixels in eight directions of vertical, horizontal, and diagonal to analyze the statistical changes in the pixel values caused by the message embedding. HUGO could reduce the probability of being detected by the SPAM by 1/7.

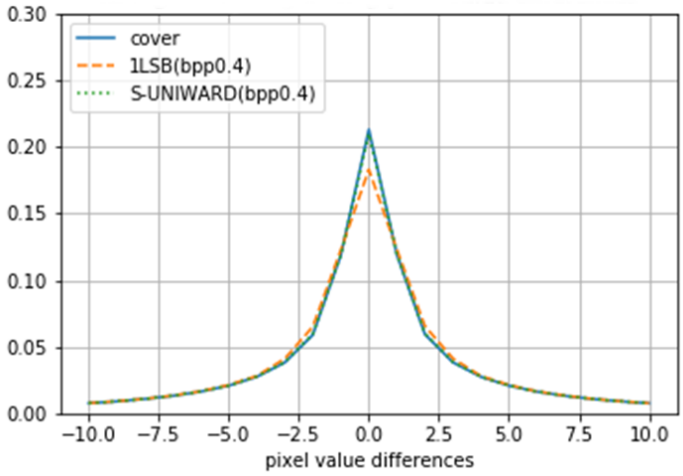


Figure 1. Probability density functions of the differences between the adjacent pixels on a cover image and its stego images.

The performance of image steganalysis in detecting image steganography has greatly improved with the development of image steganography to more covertly and skillfully hide a message. Image staganalytic methods generally try to extract traces of image steganography in the image by using high-pass filters (HPF) and identify images to which image steganography has been applied through classification. Early steganalytic methods extracted image features using manually designed HPFs (those features are called handcrafted features hereafter) and detected image steganography using classifiers based on machine learning algorithms, such as support vector machines (SVM) [12] and random forest [13]. A representative method using handcrafted features is the spatial rich model (SRM) [14].

With the great success of convolutional neural networks (CNN) in object detection and recognition [15,16], using CNNs for steganalysis has been actively investigated [17–27]. Unlike handcrafted feature-based methods, a CNN can automatically extract and learn the features that are

optimal or well suited for identifying steganographic methods. Therefore, CNN-based steganalytic methods have demonstrated a better performance compared to handcrafted feature-based methods.

However, most existing image steganalytic methods, regardless of whether or not CNNs are used, have focused on identifying whether or not a secret message is hidden in an image (i.e., the binary classification between a normal (or cover) image in which any message has not been embedded and a stego image in which a message has been embedded). Discriminating stego images created by different steganographic methods has been less considered; thus, the binary classifiers are not suitable for discriminating these stego images. Discriminating the stego images created by WOW and UNIWARD that embed a message in a similar and skillful manner is very difficult.

The classification of stego images created by different steganographic methods plays an important role in restoring embedded messages beyond judging whether or not a message is embedded. In this study, as the first step to restore messages embedded by steganographic methods, a CNN-based steganalytic method is proposed to classify the stego images created by different steganographic methods. The structure of a ternary classifier is specially designed to distinguish between the stego images created by WOW and UNIWARD and the normal images without messages. Through comparative experiments with the existing binary classifiers, the reason why multiple steganographic methods should be classified in a single ternary classifier, and various methods for improving the performance of the proposed ternary classifier are presented.

Compared to existing image steganalytic methods, the primary contributions of this study are as follows:

- a single framework is provided for identifying multiple steganographic methods;
- a CNN-based ternary classifier is proposed for image steganalysis; and
- effective methods for extending a CNN to discriminate similar WOW and UNIWARD stego images are proposed and evaluated.

This study is an extension of [28] and differs from the previous study in the following respect:

- a CNN-based ternary classifier with a new preprocessing filter is proposed;
- more details for designing it are provided; and
- the performance of the proposed classifier is intensively evaluated.

The remainder of this paper is organized as follows: Section 2 briefly reviews the conventional image steganographic and steganalytic methods; Section 3 explains the proposed steganalytic method; Section 4 experimentally evaluates its performance using images from a database available online; and Section 5 presents the conclusions and suggestions for future work.

2. Related Work

2.1. WOW and UNIWARD

WOW and UNIWARD calculate the degree of distortion when a message is embedded in an image, and then embed a small amount of message in regions where the distortion is small. We refer herein to such methods as adaptive steganographic methods. This makes it more difficult to detect hidden messages by embedding messages only in high-frequency regions with relatively little distortion and makes it possible to avoid steganalytic attacks using statistical analysis because the change in the statistical characteristics of the images caused by message embedding is very small (Figure 1).

Adaptive steganographic methods have suggested different approaches for quantifying the image distortion caused by message embedding. The image distortion function for WOW is defined as follows:

$$D(X, Y) = \rho_{ij}(X, Y_{ij})|X_{ij} - Y_{ij}|. \quad (1)$$

Here, X and Y are a cover and its stego images, respectively, and ρ is a function that examines the detectability in all neighboring directions of each pixel using the HPFs in Figure 2. Thus, a message is

not embedded if the detectability is high even in one direction. The message is embedded into the pixels for which the detectability is low in all directions.

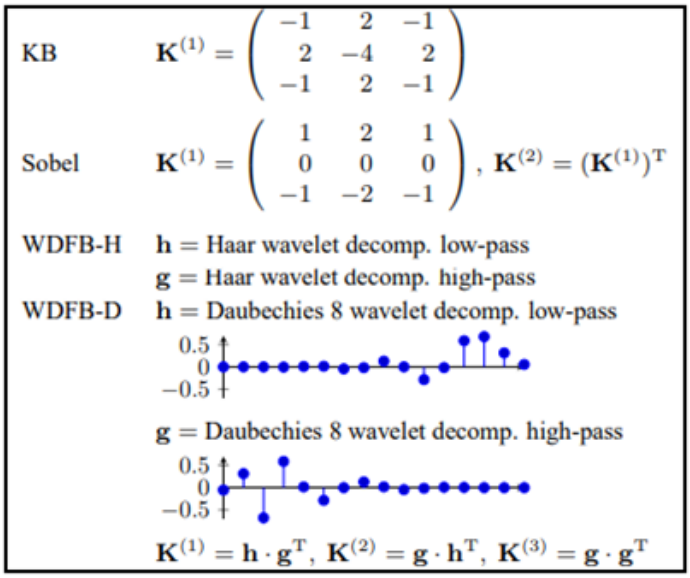


Figure 2. HPFs and wavelet filters used in WOW [9].

For UNIWARD, the residual images were calculated using the wavelet filters in Figure 2. The image distortion function is defined as follows by the sum of the absolute difference between the cover and the stego residual images:

$$D(X,Y) = \sum_{k=1}^{3} \sum_{u=1}^{n_1} \sum_{v=1}^{n_2} \frac{|W_{uv}^k(X) - W_{uv}^k(Y)|}{\sigma + |W_{uv}^{(k)}(X)|}. \quad (2)$$

Here, $W^{(k)}$ represents the residual image calculated using the kth filter; n_1 and n_2 are the image width and height, respectively, and σ is a constant stabilizing the numerical calculations.

Consequently, WOW and UNIWARD have different image distortion functions, but their approaches to embedding messages are very similar.

2.2. SRM

The SRM [14] is a handcrafted feature-based steganalytic method that uses various types of linear and nonlinear HPFs (Figure 3) to extract a number of meaningful features from the images. The features are then classified using an ensemble classifier (i.e., a random forest) that uses Fisher linear discriminants as the base classifiers.

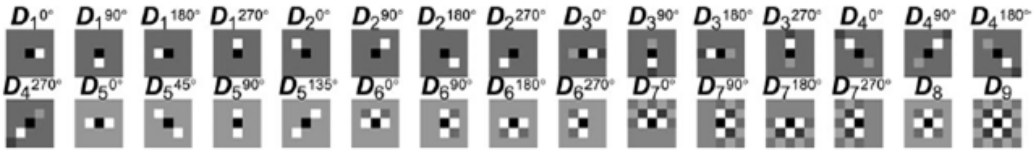


Figure 3. Thirty linear and nonlinear 5 × 5 SRM filters [19]. The filters are padded with zeros to obtain a unified size of 5 × 5.

The SRM was the most effective method used to detect image steganography before the CNN-based image steganalytic methods emerged. The SRM is highly accurate compared to

CNN-based methods. The method of extracting many features using various types of HPFs has also been widely used in CNN-based ones [19,20,25–27].

2.3. CNN-Based Image Steganalysis

CNNs can automatically extract the optimal features required to detect and recognize objects in images, and can classify features with high accuracy [15,16]. Therefore, studies using CNNs are greatly increasing in the image steganalysis field. However, unlike other deep learning problems, the CNN-based image steganalysis has a preprocessing process of applying HPFs to input images. This process enhances the pixel variation caused by embedding messages such that the CNN can detect it well while also removing the low-frequency area, where the messages are less likely to be embedded.

Xu and Wu proposed a simple yet effective initial CNN for image steganalysis [17]. They used a network comprising five convolutional layers and a single fully connected layer (Figure 4). They also used a 5 × 5 HPF in a preprocessing stage, generated eight feature maps in the first convolutional layer, and doubled the number of feature maps and halved the size of the feature maps in the subsequent convolutional layers. Each convolutional layer comprised the processes of convolution, batch normalization, activation, and pooling. They improved the steganalytic performance of the network by adding the absolute layer to the first convolutional layer and by using the tanh activation function in the first two convolutional layers. Yuan et al. used the same network structure as the initial CNN, but utilized three HPFs in a preprocessing stage [18].

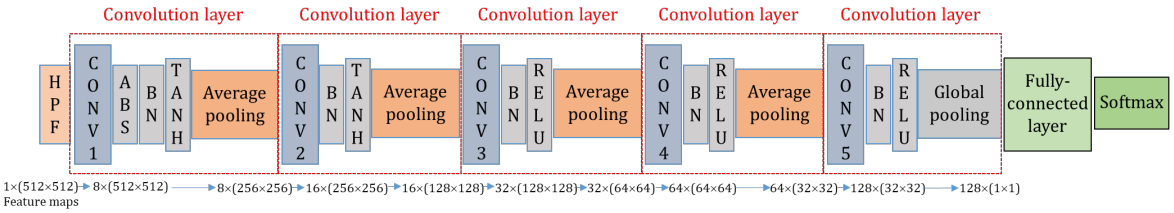


Figure 4. Initial CNN for image steganalysis [17]. The CNN extracts 128 1 × 1 feature maps from a 512 × 512 input image.

ReST-Net [19] uses three different filter sets, namely 16 simplified linear SRM, 14 nonlinear SRM, and 16 Gabor filters (Figures 3 and 5) in the preprocessing stage to extract much more features from the input images. In addition, ReST-Net constructs three subnetworks (Figure 6). After separately training the subnetworks using each preprocessing filter, it trains a new fully connected layer using transfer learning while fixing the parameters of three subnetworks.

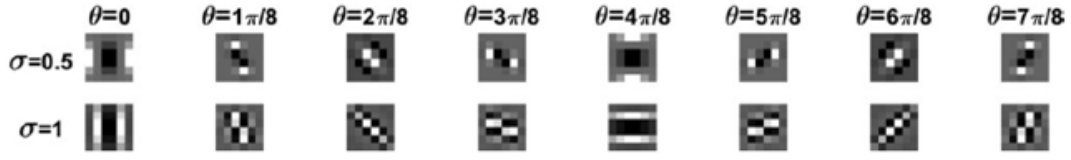


Figure 5. Sixteen 6 × 6 Gabor filters with different orientations and scales [19].

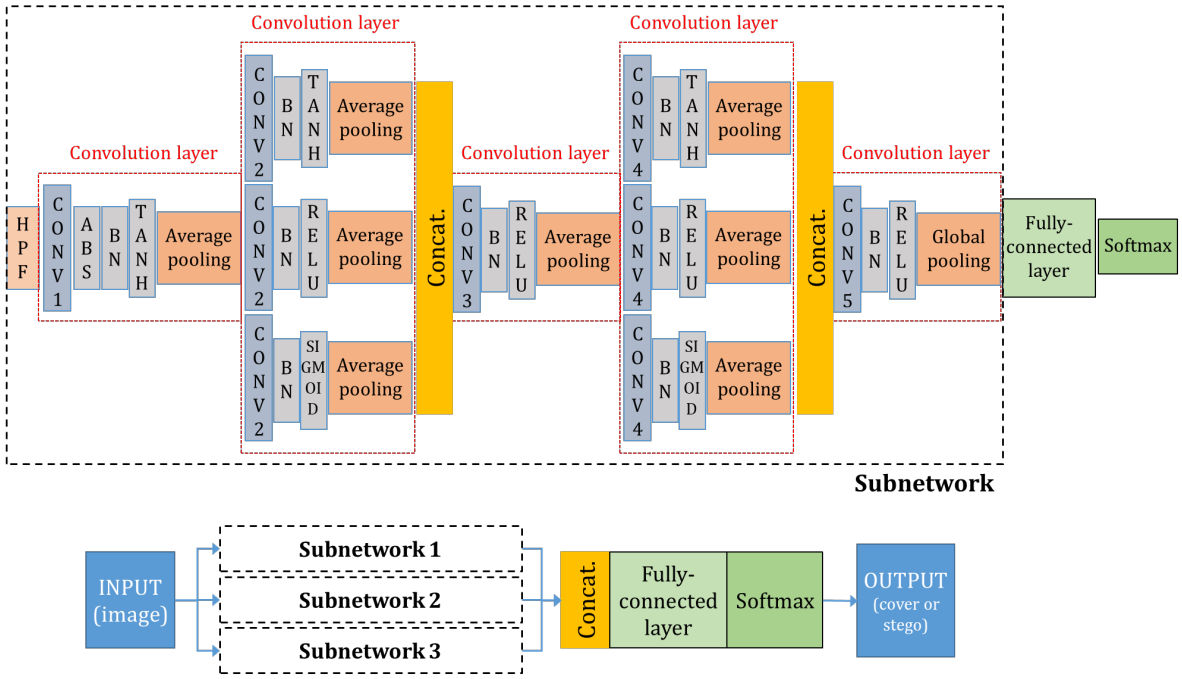


Figure 6. Structure of Rest-Net [19], which comprises three three subnetworks that are a modification of the initial CNN [17] and uses transfer learning.

Yedroudj-Net [20] has a similar structure with the initial CNN [17], but uses linear SRM filters in the preprocessing stage, and has two additional fully connected layers (Figure 7). It removes the average pooling process in the first convolutional layer to prevent loss of information caused by pooling. In the first two convolutional layers, it uses the TLU function instead of the tanh function to remove the strong, but statistically insignificant information. It has an additional scaling process after batch normalization. Yedroudj-Net has achieved approximately 4–5% improvement in accuracy in binary classification in comparison with the initial CNN [17].

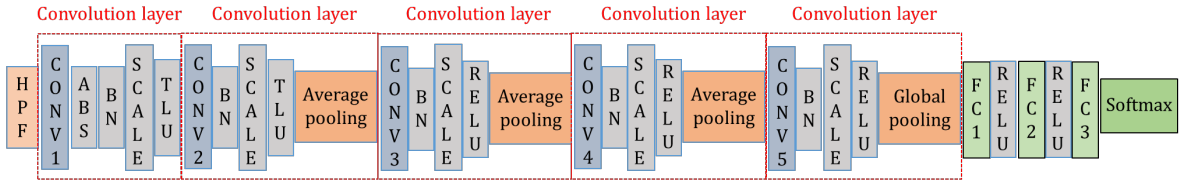


Figure 7. Structure of Yedroudj-Net [20].

Deep residual networks for image steganalysis have also been proposed [22,23]. These networks could be made much deeper by employing residual shortcuts (Figure 8). In [22], without fixing the preprocessing filters or initializing the filter coefficients with the SRM filters, the preprocessing process is significantly expanded using several convolutional and residual layers to realize a completely data-driven steganalysis.

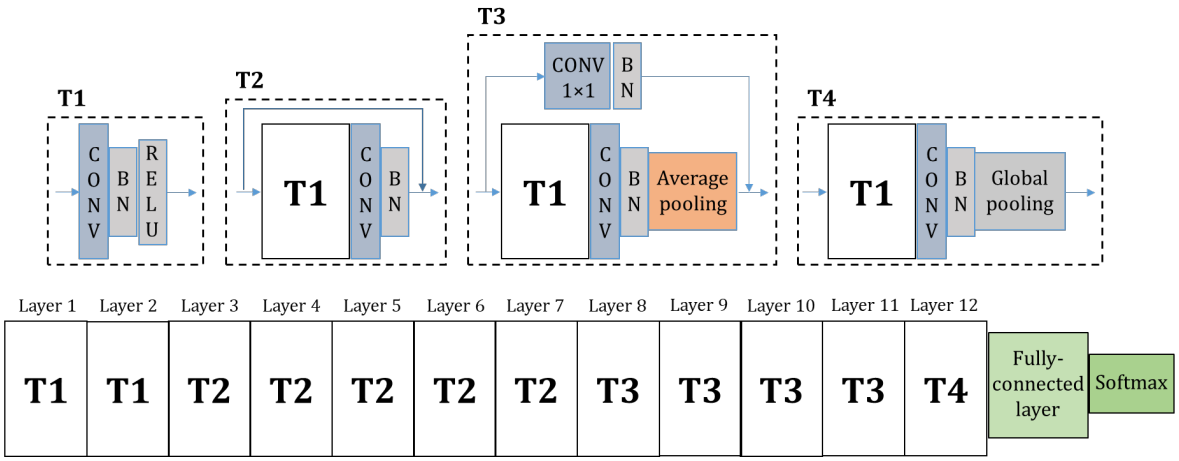


Figure 8. Structure of a deep residual network used in [22].

Ke et al. proposed a multi-column CNN that extracts various features using filters of different sizes in convolutional layers and allows the input image to be of an arbitrary size or resolution [24]. As a multi-task learning approach, Yu et al. extended a CNN by adding fully convolutional networks that take the output of each convolutional layer as the input for a pixel binary classification that estimates whether or not each pixel in an image has been modified because of steganography [26].

Wu et al. proposed a new normalization, called shared normalization, that uses the same mean and standard deviation, instead of the minibatch mean and standard deviation, for all training and test batches to normalize each input batch and address the limitation of batch normalization for image steganalysis [21]. Meanwhile, Ni et al. proposed a selective ensemble method that can choose to join or delete a base classifier by reinforcement learning to reduce the number of base classifiers while ensuring the classification performance [25].

As such, the existing CNN-based staganalytic methods could successfully increase the classification accuracy by deepening or widening the CNNs and using various types of preprocessing filters. However, these methods aimed for the binary classification of cover and stego images, and, thus, may not be available for the N-ary ($N > 2$) classification. Two adaptive steganographic methods, namely WOW and UNIWARD, embed a small amount of messages in a similar manner (Section 2.1); hence, the binary classifiers are very likely to misclassify the WOW and UNIWARD stego images.

3. Proposed Method

3.1. Similarity between WOW and UNIWARD

The adaptive steganographic methods, namely WOW and UNIWARD, use directional filters to analyze how different the differences from the neighboring pixels (i.e., the degree of image distortion) are when a message is embedded into each pixel of an image, and then selectively embed the message into a pixel with a small degree of image distortion. WOW and UNIWARD use different functions to measure the image distortion, but their processes of embedding the message are very similar; thus, the existing CNN-based binary classifiers become confused when discriminating WOW and UNIWARD, and are very likely to make an incorrect classification.

We conducted an experiment in which UNIWARD stego images were input to a binary classifier that had been trained for WOW and vice versa to demonstrate the difficulty of discriminating WOW and UNIWARD using binary classifiers. The CNN used in the literature [17] (Figure 4) was used for the experiment. The other experimental conditions were the same as those given in Section 4.

Table 1 shows that, even when two different steganographic methods (i.e., WOW and UNIWARD) were used in the training and testing phases, respectively, the classification rates for the stego images were still high. For example, the classification rates were 67.13% when the UNIWARD stego images were input into the classifier trained using the WOW stego images. In other words, it is very likely

that they are confused with each other because they are too similar to discriminate. Therefore, using existing binary classifiers to classify WOW and UNIWARD is ineffective.

Table 1. Cross identification between WOW and UNIWARD ($bpp = 0.4$).

Training	Testing	Classification Rates (%)			
		For Cover	For UNIWARD Stego	For WOW Stego	Total
UNIWARD	UNIWARD	84.02	73.56	-	78.80
WOW	WOW	77.63	-	78.25	77.94
UNIWARD	WOW	84.02	-	60.48	72.25
WOW	UNIWARD	77.63	67.13	-	72.38

3.2. Combining Pre-Trained Binary Classifiers to Discriminate WOW and UNIWARD

We attempted to train two CNN-based binary classifiers for WOW and UNIWARD and simply combine the two classifiers in parallel to determine the result of the classifier with a higher probability as a final result (Figure 9). This was based on the assumption that the results of the classifier with a greater probability would be right if different classification results are obtained by the two classifiers. Table 2 presents the classification results for the cover, WOW stego, and UNIWARD stego images (the details for the experimental conditions are given in Section 4). The classification rates for the WOW and UNIWARD stego images significantly decreased because of the similarity between WOW and UNIWARD. In other words, the simple combination of two binary classifiers is not useful for discriminating WOW and UNIWARD.

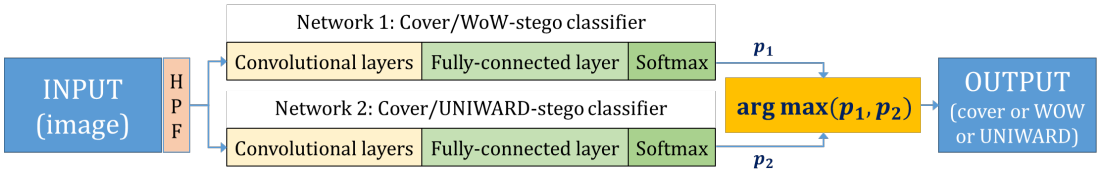


Figure 9. Combining two binary classifiers in parallel for the ternary classification.

Table 2. Ternary classification rates obtained by simply combining two binary classifiers separately trained for WOW and UNIWARD ($bpp = 0.4$).

For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
85.23	50.13	46.24	60.53

We also conducted an experiment for ternary classification through transfer learning. The network parameters of classifiers were fixed after training each binary classifier for WOW and UNIWARD. The fully connected layer was then removed from each binary classifier, and a common fully connected layer was added and trained for the ternary classification (Figure 10). Table 3 shows the classification results for the cover, WOW stego, and UNIWARD stego images (the details for the experimental conditions are given in Section 4). The classification rates for the WOW and UNIWARD stego images were very low, and were lower than those obtained by simply combining two binary classifiers in parallel. This result indicates that the network parameters in the fully connected layer were not correctly trained because of the similarity between WOW and UNIWARD.

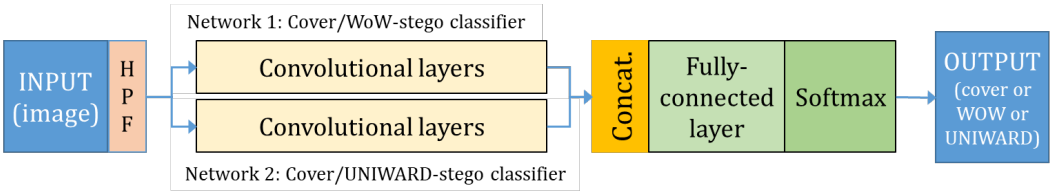


Figure 10. Ternary classification through transfer learning.

Table 3. Ternary classification rates obtained by transfer learning ($bpp = 0.4$).

For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
72.23	45.56	42.90	53.56

Referring to the abovementioned experiments, a new single network should be designed to simultaneously learn the cover, WOW stego, and UNIWARD stego images from the beginning and correctly classify the similar steganographic methods, WOW and UNIWARD.

3.3. Designing a CNN for Ternary Classification

The CNN used in [17] is the most basic CNN for image steganalysis, and most conventional CNNs are its modifications. Therefore, it was used as the base CNN herein. First, the base CNN was tested for ternary classification without modification. The cover, WOW stego, and UNIWARD stego images were simultaneously learned in a single network (Figure 4). Table 4 presents the classification rates (the details for the experimental conditions are given in Section 4), which are better than those obtained by combining pre-trained binary classifiers. However, the cover images were relatively well classified at approximately 84%, but the WOW and UNIWARD stego images were rarely classified as expected. In conclusion, the network structure should be extended, and the preprocessing filter for extracting the steganalytic features should be more carefully designed to make the classifier originally developed for the binary classification between the cover and stego images available for ternary classification.

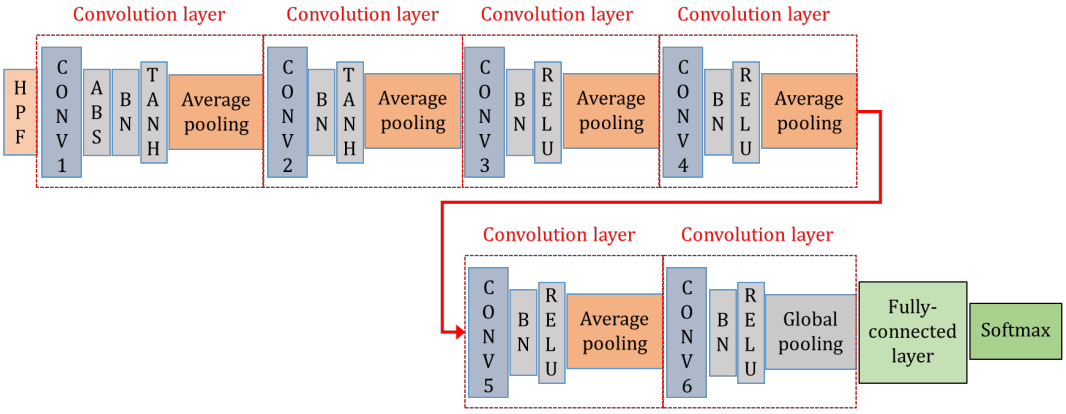
Table 4. Ternary classification rates when simultaneously learning the cover, WOW stego, and UNIWARD stego images using the conventional classifier [17] ($bpp = 0.4$).

For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
84.05	56.45	50.39	63.63

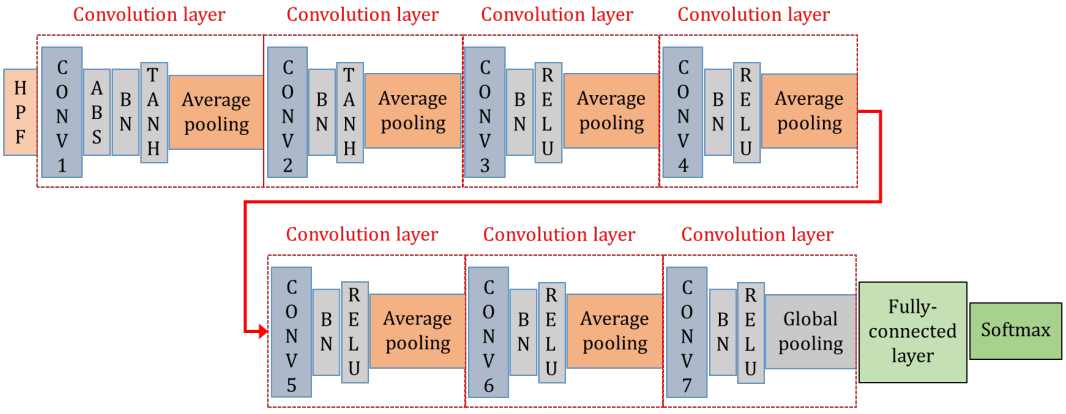
We tried to extend the based CNN by attempting to add more convolutional layers (each comprising convolution, normalization, activation, and pooling operations) because more classification power would be required for the ternary classification compared to the binary classification. Figure 11 shows the structure of the networks extended with additional convolutional layers. Table 5 displays the classification rates of the extended networks (the details for the experimental conditions are given in Section 4). As a result, adding convolutional layers improved the classification rates by 2–4%; however, the classification rates rather became lower with two or more additional convolutional layers, indicating that the network needs to be deeper for ternary classification, but the depth should be properly adjusted.

Table 5. Ternary classification rates of deeper networks in Figure 11 ($bpp = 0.4$).

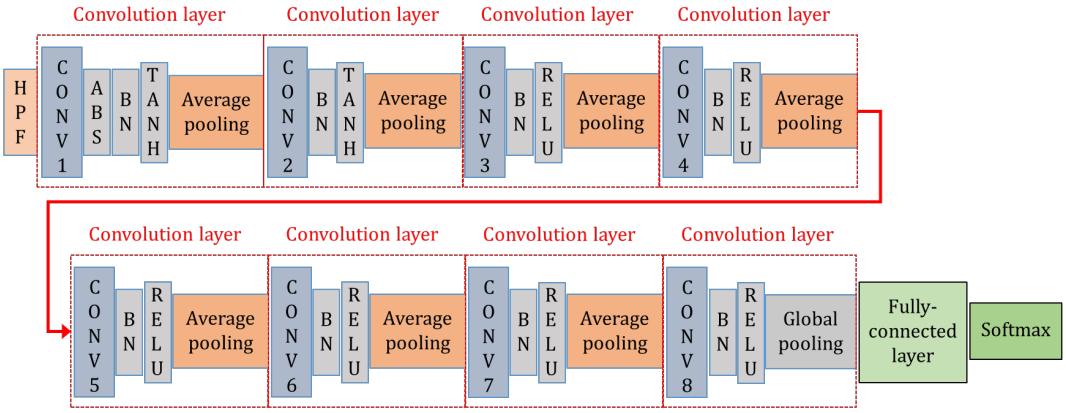
	For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
Figure 11a	75.81	67.48	59.83	67.70
Figure 11b	72.40	58.18	69.80	66.79
Figure 11c	61.24	60.18	75.27	65.56



(**a**) With an additional convolutional layer



(**b**) With two additional convolutional layers



(**c**) With three additional convolutional layers

Figure 11. Extending the conventional network [17] with additional convolutional layers.

We also attempted to use a deep residual network (Figure 12a) or a convolution-stacked network (Figure 12b), where the convolutional blocks were stacked as done in [29] because those residual or convolution-stacked networks demonstrated a significantly improved performance in image recognition. However, as shown in Table 6, the classification rates were not good, indicating that these networks were not suitable for image steganalysis or for ternary classification.

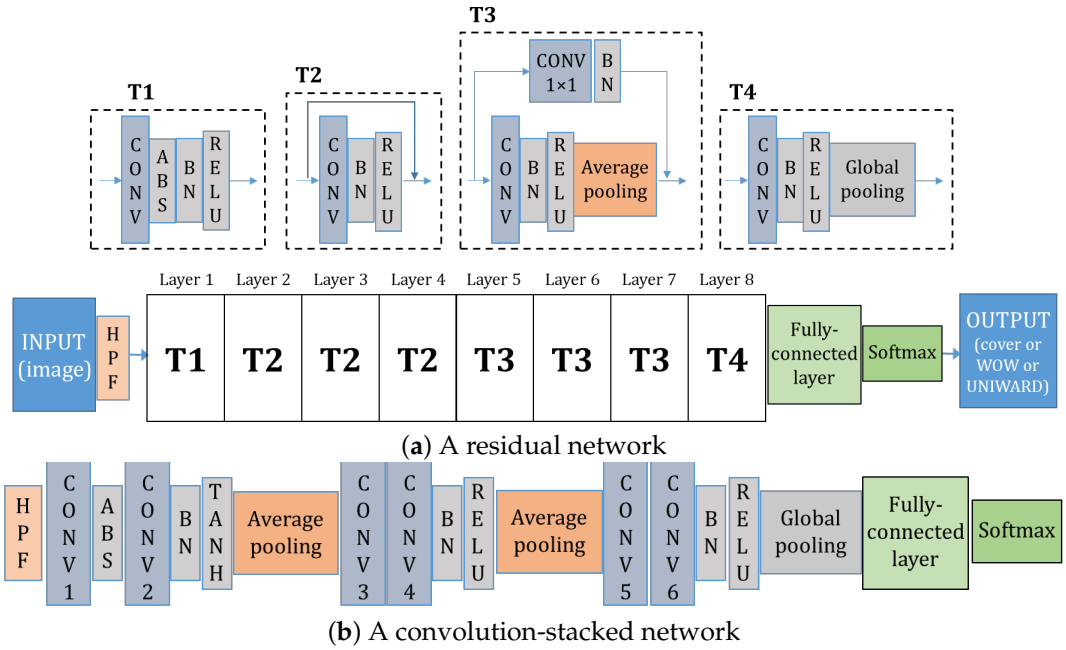


Figure 12. Deep residual network and convolution-stacked network for ternary classification.

Table 6. Ternary classification rates of the residual and convolution-stacked networks of Figure 12 ($bpp = 0.4$).

	For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
Figure 12a	80.03	23.52	4.30	35.95
Figure 12b	83.87	62.43	32.87	59.72

As explained in Section 2.3, the CNN-based classifiers for image steganalysis have preprocessing filters to facilitate the extraction of steganalytic features from images. Many conventional methods tried to use various preprocessing filters for performance improvement. For the ternary classification, we decided to use the SRM filters mostly used in conventional methods and conducted an experiment to determine their performance. The base CNN was used with three different preprocessing filter sets: 30 SRM filters (Figure 3), three groups of 10 SRM filters, and 10 selected SRM filters (Figure 13). The second filter set was obtained by dividing 30 SRM filters into three groups of 10 (using different numbers of groups was worse [28]). The filters of each group were applied to the input image. Ten filtered results were generated by performing the element-wise sum between the filtered results of each group [28]. The third filter set is a new one proposed herein. More effective filters were selected from 30 SRM filters. Each of the 30 SRM filters was applied to the arbitrary cover and stego images. The differences between the filtered cover and stego images were then computed (Figure 14). Subsequently, 10 filters with higher differences were selected, assuming that those filters would extract steganalytic features from the images well. For all of the filter sets, eight feature maps were generated in the first convolutional layer and doubled in the subsequent convolutional layers. Tables 4 and 7 (the details for the experimental conditions are given in Section 4) show that the classification rates of the base CNN did not increase as the number of filters increased, unlike expected. The results of the three groups of 10 SRM filters were better than those of the others, indicating that simply increasing the number of filters does not guarantee performance improvements, and finding the appropriate filters for a given CNN is necessary.

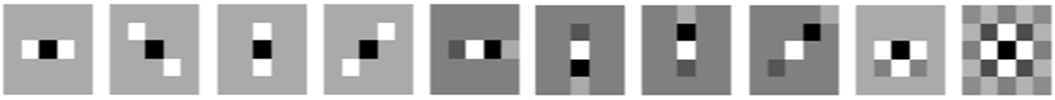


Figure 13. Ten selected SRM filters. They can better detect tiny variation on images, among 30 SRM filters of Figure 3.

Figure 14. Selection of more effective SRM filters: a large difference (e.g., 1.229 and 7.234) depending on the filter between the filtered cover and stego images is found after each SRM filter is applied.

Table 7. Ternary classification rates of the base CNN with different preprocessing filters ($bpp = 0.4$).

Preprocessing Filters	For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
30 SRM	50.43	19.98	90.90	53.70
Three groups of 10 SRM	65.44	51.01	75.26	63.90
10 selected SRM	68.55	73.15	42.41	61.37

Together with increasing the number of filters, we also attempted to increase the feature maps in the first convolutional layers from 8 to 60. Table 8 shows that the classification rates of the base CNN became significantly lower, except for the 10 selected SRM filters, when the number of feature maps increased. Unlike most conventional CNNs that achieve performance improvement by using more filters or feature maps, the base CNN had a better performance with a small number of filters maybe because the base CNN failed to learn a large amount of information extracted by many filters or feature maps. From these results, we conclude that the base CNN should be deeper such that more filters or feature maps can be used.

Table 8. Ternary classification rates of the base CNN with different preprocessing filters when increasing the feature maps in the first convolutional layers to 60 ($bpp = 0.4$).

Preprocessing Filters	For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
5×5 HPF	90.44	40.15	44.89	58.49
30 SRM	4.54	87.39	26.04	39.32
Three groups of 10 SRM	79.17	32.52	67.39	59.69
10 selected SRM	46.08	45.27	88.99	60.11

3.4. Proposed Classifier for Ternary Classification

We proposed a CNN-based classifier for the ternary classification. The base CNN [17] was extended with an additional convolutional layer. The feature maps were increased to 60 in the first

convolutional layer and doubled in the subsequent convolutional layers: thus, 1920 feature maps were fed into the fully connected layer. Ten selected SRM filters were used as the preprocessing filters.

4. Experimental Results and Discussion

All the experiments presented in the previous sections and in this section were conducted with the following conditions: 10,000 gray scale images of 512 × 512 in BOSSBase 1.01 [30] were quartered, and the resulting 40,000 images were divided into the training and testing sets, each comprising 30,000 and 10,000 images, respectively. The stego images for both sets were generated with a random payload of $bpp = 0.4$ (In most steganalytic studies, 0.1, 0.2, and 0.4 bpp have been used for testing steganalytic methods. However, when using adaptive steganographic methods, 0.1 and 0.2 bpp are too small to identify the stego images, even in binary classification [31]. The average PSNRs of the WOW and UNIWARD stego images of 0.4 bpp are 58.76 and 59.36 dB, respectively; thus, the image quality of the stego images of 0.4 bpp is still very high.) using WOW and UNIWARD. As a result, 90,000 (30,000 for cover, WOW stego, and UNIWARD stego images each) training images of 256 × 256 and 30,000 (10,000 for cover, WOW stego, and UNIWARD stego images each) testing images were used. For training, a momentum optimizer [32] with a momentum value of 0.9 was used. The learning rate started at 0.001 and decreased to 90% in every 5000 iterations. The minibatch size was 64 (32 pairs of cover and stego images). The other hyperparameters were set the same as in the conventional method [17]. All CNNs were implemented using the TensorFlow library [33].

The proposed classifier was evaluated with different preprocessing filters. As a new preprocessing filter set, 16 Gabor filters were used together with the 10 selected SRM filters, as has been done in [19]. The results in Table 9 are the classification rates for the cover, WOW stego, and UNIWARD stego images obtained using different preprocessing filters.

Table 9. Ternary classification rates of the network of Figure 11a with different preprocessing filters ($bpp = 0.4$).

Preprocessing Filters	For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
5 × 5 HPF	68.52	46.07	61.01	58.53
30 SRM	75.49	51.85	71.46	66.26
Three groups of 10 SRM	76.22	59.66	77.45	70.78
10 selected SRM	75.65	69.71	71.32	72.22
10 selected SRM + 16 Gabor	76.23	56.26	62.10	64.86

Unlike the base CNN, using more filters and feature maps increased the classification rates; however, utilizing too many and different types of filters was not good. The results of the 10 selected SRM filters (i.e., the proposed one) were the best. The experimental results demonstrated that the cover, WOW stego, and UNIWARD stego could be classified with an accuracy of approximately 72% through the single CNN-based ternary classifier proposed herein.

We also attempted to change the tanh functions of the first two convolutional layers to TLU functions, as has been done in [20], and the ReLU functions of the subsequent convolutional layers to leaky ReLU functions, but the classification rates were not good (Table 10).

Table 10. Ternary classification rates when changing the activation functions of the proposed CNN ($bpp = 0.4$).

For Cover (%)	For UNIWARD Stego (%)	For WOW Stego (%)	Total (%)
39.60	72.29	91.87	67.93

5. Conclusions and Future Works

This study proposed a CNN-based ternary classifier to identify cover, WOW stego, and UNIWARD stego images. The existing binary classifiers were designed to learn and detect a specific steganographic

method; hence, they were not suitable for discriminating different steganographic methods. Adaptive steganographic methods, such as WOW and UNIWARD, embed a small amount of the secret message in a similar manner; therefore, discriminating their stego images using the existing binary classifiers or combining them was very difficult. However, the proposed ternary classifier could effectively learn the difference between both steganographic methods and discriminate them with high accuracy. The classification between different steganographic methods using the proposed ternary classifier was the first step in restoring the embedded message instead of simply determining whether or not a message has been embedded.

It was experimentally confirmed that, in designing a CNN-based ternary classifier for image steganalysis, simply expanding the width or depth of the CNN does not guarantee performance improvements. In other words, the CNN width and depth need experimental optimization. This study demonstrated the results of such an experimental optimization.

The proposed method had an accuracy of approximately 72%, which is not very high. Therefore, ways to improve the accuracy by further highlighting the differences between WOW and UNIWARD must be explored in the future. Ways to design a CNN-based classifier suitable for classifying a larger number ($\geq$3) of steganographic methods, including those with other embedding domains (e.g., DCT and wavelet domains), must also be explored.

Author Contributions: Conceptualization, S.K. and H.P.; Funding acquisition, H.P. and J.-I.P.; Methodology, S.K. and H.P.; Software, S.K.; Supervision, H.P. and J.-I.P.; Validation, S.K. and H.P.; Writing—original draft, S.K.; Writing—review and editing, H.P. and J.-I.P.

Funding: This work was supported by the research fund of the Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and Agency for the Defense Development of Korea.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chan, C.K.; Cheng, L.M. Hiding data in images by simple LSB substitution. *Pattern Recognit.* **2004**, *37*, 469–474. [CrossRef]
2. Wu, D.C.; Tsai, W.H. A steganographic method for images by pixel-value differencing. *Pattern Recognit.* **2003**, *24*, 1613–1626. [CrossRef]
3. Chang, K.C.; Chang, C.P.; Huang, P.S.; Tu, T.M. A novel image steganographic method using tri-way pixel-value differencing. *J. Multimed.* **2008**, *3*, 37–44. [CrossRef]
4. Darabkh, K.A.; Al-Dhamari, A.K.; Jafar, I.F. A new steganographic algorithm based on multi directional PVD and modified LSB. *J. Inf. Technol. Control* **2017**, *46*, 16–36. [CrossRef]
5. Kordov, K.; Stoyanov, B. Least significant bit steganography using Hitzl-Zele chaotic map. *Int. J. Electron. Telecommun.* **2017**, *63*, 417–422. [CrossRef]
6. Stoyanov, B.P.; Zhelezov, S.K.; Kordov, K.M. Least significant bit image steganography algorithm based on chaotic rotation equations. *C. R. Acad. Bulg. Sci.* **2016**, *69*, 845–850.
7. Fridrich, J.; Goljan, M.; Du, R. Detecting LSB steganography in color and gray-scale images. *IEEE Multimed. Mag.* **2001**, *8*, 22–28. [CrossRef]
8. Pevny, T.; Filler, T.; Bas, P. Using high-dimensional image models to perform highly undetectable steganography. In Proceedings of the 12th International Conference on Information Hiding, Calgary, AB, Canada, 28–30 June 2010; pp. 161–177.
9. Holub, V.; Fridrich, J. Designing steganographic distortion using directional filters. In Proceedings of the IEEE Workshop on Information Forensic and Security, Tenerife, Spain, 2 December 2012.
10. Holub, V.; Fridrich, J.; Denemark, T. Universal distortion function for steganography in an arbitrary domain. *EURASIP J. Inf. Secur.* **2014**, *2014*, 1. [CrossRef]
11. Penvy, T.; Bas, P.; Fridrich, J. Steganalysis by subtractive pixel adjacency matrix. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 215–224.
12. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
13. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 275–282.

14. Fridrich, J.; Kodovský, J. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [CrossRef]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Xu, G.; Wu, H. Structure design of convolution neural networks for steganalysis. *IEEE Signal Process. Lett.* **2016**, *23*, 708–712. [CrossRef]
18. Yuan, Y.; Lu, W.; Feng, B.; Weng, J. Steganalysis with CNN using multi-channels filtered residuals. *LNCS* **2017**, *10602*, 110–120.
19. Li, B.; Wei, W.; Ferreira, A.; Tan, S. ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis. *IEEE Signal Process. Lett.* **2018**, *25*, 650–654. [CrossRef]
20. Yedroudj, M.; Comby, F.; Chaumont, M. Yedroudj-Net: An efficient CNN for spatial steganalysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 15–20.
21. Wu, S.; Zhong, S.; Liu, Y. A novel convolutional neural network for image steganalysis with shared normalization. *IEEE Trans. Multimed.* **2019**. [CrossRef]
22. Boroumand, M.; Chen, M.; Fridrich, J. Deep residual network for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1181–1193. [CrossRef]
23. Wu, S.; Zhong, S.; Liu, Y. Deep residual learning for image steganalysis. *Multimed. Tools Appl.* **2018**, *77*, 10437–10453. [CrossRef]
24. Ke, Q.; Ming, L.D.; Daxing, Z. Image steganalysis via multi-column convolutional neural network. In Proceedings of the 14th IEEE International Conference on Signal Processing, Beijing, China, 12–16 August 2018; pp. 550–553.
25. Ni, D.; Feng, G.; Shen, L.; Zhang, X. Selective ensemble classification of image steganalysis via deep Q network. *IEEE Signal Process. Lett.* **2019**, *26*, 1065–1069. [CrossRef]
26. Yu, X.; Tan, H.; Liang, H.; Li, C.T.; Liao, G. A multi-task learning CNN for image steganalysis. In Proceedings of the 10th IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 11–13 December 2018.
27. Zhang, T.; Zhang, H.; Wang, R.; Wu, Y. A new JPEG image steganalysis technique combining rich model features and convolutional neural networks. *Math. Biosci. Eng.* **2019**, *16*, 4069–4081. [CrossRef]
28. Kang, S.; Park, H.; Park, J.-I. Toward ternary classification in CNN-based image steganalysis. In Proceedings of the 15th International Conference on Multimedia Information Technology and Applications, Ho Chi Minh City, Vietnam, 27 June–1 July 2019.
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Available online: https://arxiv.org/pdf/1409.1556.pdf (accessed on 4 September 2019).
30. Bas, P.; Filler, T.; Pevny, T. Break our steganographic system—The ins and outs of organizing BOSS. In Proceedings of the International Workshop on Information Hiding, Prague, Czech Republic, 18–20 May 2011; pp. 59–70.
31. Kim, J.; Park, H.; Park, J.-I. CNN-based image steganalysis using additional data embedding. *Multimed. Tools Appl.* **2019**, in press.
32. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151. [CrossRef]
33. TensorFlow. Available online: https://www.tensorflow.org/ (accessed on 4 September 2019).