ON THE EFFECTIVENESS OF VISION TRANSFORMERS FOR ZERO-SHOT FACE ANTI-SPOOFING

Anjith George\textsuperscript{a}  
Sébastien Marcel

Idiap-RR-30-2020

NOVEMBER 2020

\textsuperscript{a}Idiap Research Institute
On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing

Anjith George and Sébastien Marcel
Idiap Research Institute
Rue Marconi 19, CH - 1920, Martigny, Switzerland
{anjith.george, sebastien.marcel}@idiap.ch

Abstract

The vulnerability of face recognition systems to presentation attacks has limited their application in security-critical scenarios. Automatic methods of detecting such malicious attempts are essential for the safe use of facial recognition technology. Although various methods have been suggested for detecting such attacks, most of them over-fit the training set and fail in generalizing to unseen attacks and environments. In this work, we use transfer learning from the vision transformer model for the zero-shot anti-spoofing task. The effectiveness of the proposed approach is demonstrated through experiments in publicly available datasets. The proposed approach outperforms the state of the art methods in the zero-shot protocols in the HQ-WMCA and SiW-M datasets by a large margin. Besides, the model achieves a significant boost in cross-database performance as well.

1. Introduction

Face recognition offers a simple yet convenient way for access control. Though face recognition systems have become ubiquitous [23], its vulnerability to presentation attacks (a.k.a. spoofing attacks) [29], [22] limits the application of these systems in safety-critical applications. An unprotected face recognition (FR) system might be fooled by merely presenting artifacts like a photograph or video in front of the camera. The artifact used for such an attack is known as a presentation attack instrument (PAI).

As the name indicates, presentation attack detection (PAD) systems try to protect FR systems against such malicious attempts. Though a wide variety of presentation attacks are possible, the majority of the research efforts have focussed on the detection of 2D attacks such as prints and replays, mainly due to the easiness of producing such attack instruments. Most of the research in PAD focus on the detection of these attacks using the RGB spectrum alone, using either feature-based methods or Convolutional Neural Network (CNN) based approaches. Several feature-based methods using color, texture, motion, liveliness cues, histogram features [7], local binary pattern [28], [10] and motion patterns [2] have seen proposed over the years for performing PAD. However, most of the recent state of the art results are from CNN based methods. Specifically, CNNs using auxiliary information in the form of binary or depth supervision have shown to improve performance greatly [3, 15]. Nevertheless, the majority of these methods perform well only in case of 2D attacks and the performance of such methods degrade when evaluated against sophisticated 3D and partial attacks [27]. Even in the case of 2D attacks, these models often fail to generalize towards unseen attacks and environments.

Recently, several multi-channel approaches have been proposed to address the limitations of PAD systems [18, 31]. Though such methods achieve superior performance as compared to RGB methods, the cost of additional hardware required limits the application of such methods to protect legacy RGB based FR systems. Hence, it is desirable to have a robust RGB based PAD method that is robust against a wide variety of 2D, 3D, and partial attacks. Ideally, a PAD system should be able to generalize well to unseen attacks and environments.
In most cases, the amount of data available to train a PAD model is very limited. This limits the possibility of training deep architectures from scratch. And from literature, transfer learning from a pre-trained network has proven to be an effective strategy to deal with the limited data problem. Moreover, a pre-trained network could also help with addressing the domain shift as the model has seen a wide variety of images in different environments.

In this work, we investigate the effectiveness of the vision transformer model [12] for the zero-shot presentation attack detection problem. We compare the proposed method with both state of the art methods and fine-tuned CNN models. Specifically, we investigate the performance of this method in challenging unseen attack and cross-database scenarios.

To the best knowledge of the authors, this is the first work using vision transformers for the presentation attack detection task. The main contributions of this work are listed below.

- Introduces a simple yet effective Vision Transformer based PAD framework.
- Shows the effectiveness of vision transformer framework in an unrelated downstream task while adapting only a minimal number of parameters in the training stage.
- The proposed approach has been extensively evaluated in challenging unseen and cross-database conditions and it achieves the state of the art performance, outperforming other baselines by a large margin.

Additionally, the source code and protocols to reproduce the results are available publicly* to allow further extension of the work.

2. Related work

2.1. Presentation Attack Detection

The majority of face presentation attack detection methods deal with the detection of 2D attacks. And most of these methods capitalize on capturing the quality degradation of the samples during recapture. Feature-based methods such as as motion patterns [2], Local Binary Patterns (LBP) [7], image quality features [14], and image distortion analysis [36] have been utilized over the years for PAD. Most of the recent PAD methods use CNN based approaches [26, 15, 3]. Since many of these methods depend on the quality degradation during recapture to distinguish attacks from bonafide, they may not be suitable for the detection of sophisticated attacks like 3D masks and partial attacks. Since the quality of attack instruments improve over time, the PAD methods used should be robust against unseen attacks as well.

Multi-channel methods have been proposed in the literature as a solution to handle a wide variety of attacks [33, 34, 33, 13, 11, 6, 5]. The core idea of multi-channel spectral methods is the usage of complementary information from different channels making it extremely hard for attackers to fool the PAD systems. An attacker would have to replicate the properties of a bonafide sample across different sensing domains, making it difficult depending on the channels used in the PAD system. In [18], George et al. presented a multi-channel face presentation attack detection framework using color, depth, infrared and thermal channels. Several recent works have also achieved good PAD performance utilizing multi-channel information [20, 17, 16].

Though multi-channel information could alleviate the issues with the PAD systems, the cost of the hardware increases with additional channels used. Moreover, it limits the widespread adoption of the PAD system, leaving legacy face recognition systems unprotected. An RGB only PAD framework, which performs robustly across unseen attacks and environments is essential for the secure usage of legacy face recognition systems.

2.2. Transformer Models

The Transformer models proposed in [35] introduced a novel approach towards sequence transduction tasks obviating the need for convolution or recurrence mechanisms. Transformers essentially capitalize on the attention mechanism to model dependencies between input and output. Opposed to other recurrent methods, Transformer allows significant parallelization in sequence tasks, while achieving the state of the art results in many tasks. There have been several attempts to use the Transformer framework for vision tasks, some of the methods used initial layers of pre-trained models [25] to get features to be used in the Transformer model. Various forms of position encodings, i.e., fixed or learned are also added to the features before passing to the Transformer layers. The work in [9] used the encoder-decoder structure of the Transformer for the object detection task.

3. Proposed method

In this work, we propose to use transfer learning from a pre-trained Vision Transformer model for zero-shot face anti-spoofing task. The different stages of the framework are detailed in the following sections.

3.1. Preprocessing

The CNN model accepts images of resolution $224 \times 224$. To avoid the contribution from the background and other database biases, we crop the face regions in the preprocessing stage. First, face detection and landmark localization
are performed using the MTCNN [38] algorithm. The detected faces are aligned so that the eye centers are horizontally aligned. After this alignment, the images are cropped to a resolution of $224 \times 224$.

3.2. Network Architecture

The proposed framework uses the recently proposed vision transformer [12] architecture as its back-bone. The details are given in the following sections.

3.2.1 The Vision Transformer model

Transformers were initially proposed by Vaswani et al. [35] for machine translation applications. These models leverage the attention mechanism and this framework has found applications in many natural language, audio, and vision tasks. The attention layers [4] aggregate information from the entire length of the input sequence. Transformers introduced self-attention layers that scan through and update each element in a sequence using the information from the whole sequence. Essentially, they explicitly model all the pairwise interactions between the components in the input sequence. Recently authors in [12] applied the standard transformer with minimal modifications for the image classification task. An image is divided into patches, and embeddings obtained from the patches are used as the sequence input for the transformer. The vision transformers introduce a new way for image classification instead of using convolutional layers. A sequence of image patches is used as the input followed by transformer layers.

While trained with large amounts of data the vision transformer models outperform the state of the art methods in many vision benchmarks. However, retraining such a large model from scratch is very computationally expensive. However, fine-tuning offers a way to utilize these powerful models in limited data scenarios without requiring much computational power.

We use the model trained with $16 \times 16$ patches, meaning the input sequence length will be the number of patches $\frac{HW}{16}$, where $H$ and $W$ are the height and width of the input image in pixels. In addition to patch embeddings, a 1D positional embedding is also added to retain positional information. After the transformer layers, an MLP head consisting of a fully connected layer added for the classification task.

In this work, we investigate the transferability of the pre-trained Vision Transformer model for the PAD task. Specifically, we replace the last layer with a fully connected layer with one output node and we fine-tune the model using binary cross-entropy loss (BCE). We have conducted experiments with adapting a different set of layers to find the effect of fine-tuning when trained with a small dataset. The framework used for PAD is depicted in Fig. 1.

3.2.2 Implementation details

We adapted the Vision Transformer base network described in [12], from the open-source implementation provided in [37]. Pretrained weights provided was used to initialize the network. Specifically, we used the “base” variant of the pre-trained model, made available for an image size of $224 \times 224$, with a patch size of $16 \times 16$. We used the model with the same resolution as other baselines ($224 \times 224$), to enable ready comparison between other ImageNet pre-trained models. Data augmentation was performed during the training phase with random horizontal flips with a probability of 0.5. The network was supervised with binary cross-entropy loss (BCE), with a fixed learning rate of $1 \times 10^{-4}$ and a weight decay parameter of $1 \times 10^{-5}$. A batch size of 16 was used (due to the large size of the model, and memory constraints) during training. We used the standard Adam Optimizer [24], for training the model on a GPU grid for 20 epochs. The best model was selected based on the minimum loss in the validation set. The architecture was implemented using the PyTorch [32] library, and the training and evaluation components were implemented using the Bob* [1] library to make it easy to reproduce the results.

4. Experiments

Details of the databases used and the experimental results with the proposed approach are detailed in this section.

4.1. Databases

A wide variety of attacks are required to evaluate the performance of algorithms against unseen attacks. Most of the publicly available PAD datasets are limited to 2D print and replay attacks. Hence, we selected two publicly available datasets that contain a wide variety of 2D, 3D, and partial attacks, namely HQ-WMCA and SiW-M datasets. .

4.1.1 HQ-WMCA dataset

The High-Quality Wide Multi-Channel Attack (HQ-WMCA) dataset [20, 30] consists of 2904 short multi-modal video recordings of both bonafide and presentation attacks. The database includes both obfuscation and impersonation attacks and the attack categories present are print, replay, rigid mask, paper mask, flexible mask, mannequin, glasses, makeup, tattoo, and wig (Fig. 2). The number of bonafide subjects available is 51 and the dataset contains several data streams captured synchronously such as color, depth, thermal, infrared (spectra), and short-wave infrared (spectra). In this work, we utilize only the RGB data stream from the dataset. The RGB videos are captured at a resolution of $1920 \times 1200$.

*https://www.idiap.ch/software/bob/
4.1.2 Protocols in HQ-WMCA

We have created leave-one-out (LOO) attack protocols in HQ-WMCA by leaving out one attack in the train and development set. The evaluation set consists of bonafide and the attack which was left out. These sub-protocols constitute the zero-shot (or leave-one-out) protocols, which emulate the scenario of encountering an unseen attack in a real-world scenario. We have also used the grandtest protocol which consists of attacks distributed in the train, development, and test sets (with disjoint identities across folds), specifically for cross-database performance evaluation.

4.1.3 SiW-M dataset

The Spoof in the Wild database with Multiple Attack Types (SiW-M) [27] again consists of a wide variety of attacks captured using an RGB camera. The number of subjects present is 493, with 660 bonafide and 968 attack samples with a total of 1628 files. There are 13 different subcategories of attacks, collected in different sessions, pose, lighting, and expression (PIE) variations. The attacks consist of various types of masks, makeups, partial attacks, and 2D attacks. The RGB videos are available in 1920 × 1080 resolution.

4.1.4 Protocols in SiW-M

We use the leave-one-out (LOO) testing protocols available with the SiW-M [27] dataset for our experiments. The protocols available with the dataset consists of only train and eval sets. In each of the LOO protocol, the training set consists of 80% percentage of the bonafide data and 12 types of spoof samples. The test set consists of 20% of bonafide data and the attack which was left out in the training set. We created a subset of the train set (5%), as the dev set for model selection. In addition to the protocols available with the dataset, a grandtest protocol was also created (as done in [17]) specifically for cross-database testing with attacks distributed more or less equally across folds.

4.2. Metrics

For the evaluations in HQ-WMCA dataset, we have used the ISO/IEC 30107-3 metrics [22], Attack Presentation Classification Error Rate (APCER), and Bonafide Presentation Classification Error Rate (BPCER) along with the Average Classification Error Rate (ACER) in the eval set. We compute the threshold in the dev set for a BPCER value of 1%. The ACER in the eval set is calculated as the average of APCER and BPCER computed at this threshold.

\[
ACER = \frac{APCER + BPCER}{2}.
\]  

For the SiW-M database, to enable comparison with other state of the art methods, we follow the same method of reporting results as compared to [27]. Specifically, we apply a predefined threshold on the eval set of all the protocols. The ACER, APCER, and BPCER are computed using a fixed threshold of 0.5 on all the sub-protocols. Additionally, the equal error rate (EER) is also reported in the eval set.

For cross-database testing, Half Total Error Rate (HTER) is used following the convention in [17], which computes the average of False Rejection Rate (FRR) and the False Acceptance Rate (FAR):

\[
HTER = \frac{FRR + FAR}{2}.
\]

HTER is computed in the eval set using the threshold computed in the dev set using the EER criterion.

4.3. Baseline methods

We have implemented three CNN based baselines to compare with the proposed method. Since the proposed method is based on transfer learning, we compare the proposed method with transfer learning from two popular architectures for image classification namely ResNet and DenseNet architectures. Besides, we have implemented DeepPixBiS architecture from literature which was specifically designed for presentation attack detection task. In addition to the implemented baselines, we compare the proposed approach with state of the art methods from the literature in the SiW-M dataset. The details of the baseline methods implemented are given below.
• ResNetPAD: Here we take the standard pre-trained ResNet model [19], specifically we used the ResNet101 variant of the architecture. We replace the final layer with a new fully-connected layer making it suitable for binary classification. And while training, only the final fully connected layer is adapted.

• DenseNetPAD: Similarly, here we take the standard DenseNet model [21] for fine-tuning. We used the DenseNet161 variant of the architecture in our experiments. Here again, we replace the final layer with a new fully-connected layer for binary classification and during training, only the last layer is adapted.

• DeepPixBiS: This is a CNN based system [15] which achieved good intra as well as cross-database performance in challenging OULU-NPU [8] dataset. The network was trained using both binary and pixel-wise binary loss function. The usage of pixel-wise loss acts as an auxiliary loss function forcing the network to learn a robust classifier.

• ViTranZFAS: This is our final proposed framework. Essentially, we take the pre-trained vision transformer model [12] and remove the final classification head. A new fully connected layer is added on top of the embedding followed by a sigmoid layer. The network is then trained using binary cross-entropy loss function, adapting only the final fully connected layer during training.

The base architecture and pre-trained weights can be obtained from the following package *.

5. Experiments

We have conducted an extensive set of experiments in both intra and intra-dataset scenarios in both HQ-WMCA and SiW-M datasets. Specifically, we evaluate the baselines and the proposed approach in unseen attack environments (zero-shot) protocols as it indicates the performance of these PAD systems encountering real-world attacks that were not seen during training time.

5.1. Results in HQ-WMCA dataset

We have performed experiments using all the LOO protocols in the HQ-WMCA dataset and the results are tabulated in Table 1. The values reported are the ACER in the eval set corresponding to a threshold found from the dev set (using BPCER 1% criterion). It can be seen that the proposed approach achieves much better performance than the baseline methods, achieving an average ACER of 9.020 ± 7.99 %. This result is very promising since only the last fully connected layer was retrained for classification.

5.2. Results in SiW-M dataset

The SiW-M dataset contains a wide variety of attacks. We have performed experiments with the zero-shot protocols and the results are tabulated in Table 2. In this database, the proposed approach achieves a large performance improvement, nearly half of the error rate as compared to the state of the art methods. The proposed approach performs well on most of the sub-protocols and achieves a mean EER of 6.72 ± 5.66 %.

5.3. Ablation studies

It was observed that fine-tuning the last layer alone was achieving the state of the art performance. Here we examine the effectiveness of retraining different sets of layers in the HQ-WMCA dataset. We have considered three different settings for this study they are,

• FC only: Here, only the last fully connected layer is retrained, all the other layers are frozen, this corresponds to the fine-tuning scenario from a pre-trained model.

• All layers: All the layers from the architecture are adapted during training.

• Embedding + FC: Here both the first embedding layers as well as the final fully connected layer are adapted.

The results for this set of experiments are shown in Table 3. Clearly, adapting only the FC layer achieves the best results. Given the large amount of data used for training the pre-trained model, adapting other layers appears to be prohibitive with a limited amount of training data.

Figure 3 shows the t-SNE plots from the Vision Transformer embeddings on both HQ-WMCA and SiW-M datasets (for the eval set in the corresponding grandtest protocols). It can be seen that there is already a good amount of separability between bonafide and the spoof samples in the feature space. This could explain the good performance of the proposed method just by adapting the final fully connected layer.

5.4. Cross-database evaluations

From the previous section, it was clear that the proposed approach achieves much better performance as compared to state of the art methods in unseen attack scenarios. One main issue of PAD methods has been poor cross-database generalization, which is essential to ensure reliable performance in real-world deployment scenarios. To evaluate the generalization, we performed cross-database evaluations between HQ-WMCA and SiW-M dataset. For each

*https://github.com/rwightman/pytorch-image-models
Table 1: Performance of the baseline systems and the proposed method in unseen protocols of HQ-WMCA dataset. The values reported are ACER’s obtained in the eval set with a threshold computed for BPCR 1% in dev set.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Flexiblemask</th>
<th>Glasses</th>
<th>Makeup</th>
<th>Mannequin</th>
<th>Papermask</th>
<th>Rigidmask</th>
<th>Tattoo</th>
<th>Replay</th>
<th>Mean±Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>DenseNetPAD</td>
<td>28.20</td>
<td>45.50</td>
<td>36.60</td>
<td>0.40</td>
<td>6.90</td>
<td>12.70</td>
<td>4.60</td>
<td>32.40</td>
<td>20.91±15.76</td>
</tr>
<tr>
<td>ResNetPAD</td>
<td>39.10</td>
<td>42.00</td>
<td>41.20</td>
<td>2.80</td>
<td>0.50</td>
<td>21.30</td>
<td>28.60</td>
<td>21.50</td>
<td>24.62±15.33</td>
</tr>
<tr>
<td>DeepPixBiS [15]</td>
<td>5.80</td>
<td>49.30</td>
<td>23.80</td>
<td>0.00</td>
<td>25.90</td>
<td>13.60</td>
<td>6.00</td>
<td>6.00</td>
<td>15.55±15.77</td>
</tr>
<tr>
<td>ViTranZFAS</td>
<td>2.60</td>
<td>15.90</td>
<td>25.80</td>
<td>2.70</td>
<td>2.30</td>
<td>9.50</td>
<td>2.40</td>
<td>12.40</td>
<td>9.20±7.99</td>
</tr>
</tbody>
</table>

Table 2: Performance of the proposed framework in the leave one out protocols in SiW-M dataset.

<table>
<thead>
<tr>
<th>Adapted Layers</th>
<th>Flexiblemask</th>
<th>Glasses</th>
<th>Makeup</th>
<th>Mannequin</th>
<th>Papermask</th>
<th>Rigidmask</th>
<th>Tattoo</th>
<th>Replay</th>
<th>Mean±Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL Layers</td>
<td>2.40</td>
<td>44.90</td>
<td>21.10</td>
<td>0.00</td>
<td>0.20</td>
<td>21.60</td>
<td>0.60</td>
<td>25.90</td>
<td>14.59±15.42</td>
</tr>
<tr>
<td>Embbedding + FC</td>
<td>5.50</td>
<td>47.00</td>
<td>23.40</td>
<td>1.90</td>
<td>19.40</td>
<td>11.60</td>
<td>2.10</td>
<td>11.50</td>
<td>15.30±13.99</td>
</tr>
<tr>
<td>FC Only</td>
<td>2.60</td>
<td>15.90</td>
<td>25.80</td>
<td>2.70</td>
<td>2.30</td>
<td>9.50</td>
<td>2.40</td>
<td>12.40</td>
<td>9.20±7.99</td>
</tr>
</tbody>
</table>

Table 3: Performance of the Vision Transformer network when fine tuning different set of layers in unseen protocols of HQ-WMCA dataset. The values reported are ACER’s obtained in eval set with a threshold computed for BPCR 1% in dev set.

<table>
<thead>
<tr>
<th>Methods</th>
<th>trained on HQ-WMCA</th>
<th>trained on SiW-M</th>
<th>trained on SiW-M</th>
<th>trained on HQ-WMCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DenseNetPAD</td>
<td>11.40</td>
<td>27.5</td>
<td>10.4</td>
<td>29.3</td>
</tr>
<tr>
<td>ResNetPAD</td>
<td>13.50</td>
<td>25.7</td>
<td>10.4</td>
<td>29.4</td>
</tr>
<tr>
<td>DeepPixBiS [15]</td>
<td>4.60</td>
<td>25.6</td>
<td>14.7</td>
<td>38.1</td>
</tr>
<tr>
<td>ViTranZFAS</td>
<td>5.60</td>
<td>14.7</td>
<td>6.0</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Table 4: The results from the cross-database testing between SiW-M and HQ-WMCA databases. HTER (%) values computed in eval set for threshold computed in dev set using EER criteria are reported in the table.

5.5. Discussions

From the experimental results in both datasets, it can be clearly seen that the proposed method achieves the state of the art performance in challenging unseen attack scenar-
Figure 3. t-SNE plots corresponding to the 768-dimensional feature from the Vision Transformer, the plot already shows separability between bonafide and attacks in the features extracted from the pretrained model.

Figure 4. DET curves for the eval sets in cross-database test, a) trained in HQ-WMCA and evaluated on SiW-M and b) trained in SiW-M evaluated on HQ-WMCA.

6. Conclusions

In this work, we have shown the effectiveness of the vision transformer network for the zero-shot face anti-spoofing task. Essentially, just fine-tuning a pre-trained vision transformer model for the PAD task was sufficient to achieve the state of the art performance. The proposed approach achieves the state of the art performance in unseen attack protocols of two publicly available datasets. In addition to excellent performance in unseen attacks, the proposed approach outperforms the state of the art methods in cross-datasets evaluations by a large margin, indicating the efficacy of the proposed approach in generalizing to both unseen attacks and domains. The vision transformers could prove to be very beneficial in dealing with the...
current limitations of presentation attack detection systems. The datasets, source code and the protocols used are made available publicly to enable further extension of the work.

Acknowledgment

Part of this research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References


