Universal Adversarial Perturbations Through the Lens of Deep Steganography: Towards a Fourier Perspective

Chaoning Zhang*, Philipp Benz*, Adil Karjauv, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)
chaoningzhang1990@gmail.com, pbenz@kaist.ac.kr, mikolez@gmail.com, iskweon77@kaist.ac.kr

Abstract

The booming interest in adversarial attacks stems from a misalignment between human vision and a deep neural network (DNN), i.e., a human imperceptible perturbation fools the DNN. Moreover, a single perturbation, often called universal adversarial perturbation (UAP), can be generated to fool the DNN for most images. A similar misalignment phenomenon has also been observed in the deep steganography task, where a decoder network can retrieve a secret image from a slightly perturbed cover image. We attempt explaining the success of both in a unified manner from the Fourier perspective. We perform task-specific and joint analysis and reveal that (a) frequency is a key factor that influences their performance based on the proposed entropy metric for quantifying the frequency distribution; (b) their success can be attributed to a DNN being highly sensitive to high-frequency content. We also perform feature layer analysis for providing deep insight on model generalization and robustness. Additionally, we propose two new variants of universal perturbations: (1) high-pass UAP (HP-UAP) being less visible to the human eye; (2) Universal Secret Adversarial Perturbation (USAP) that simultaneously achieves attack and hiding.

Introduction

Deep learning has achieved large success in a wide range of vision applications, such as recognition (Zhang et al. 2019, 2021a), segmentation (Vania, Mureja, and Lee 2019; Kim et al. 2020; Pan et al. 2020) as well as scene understanding (Lee et al. 2019b;a; Zhang et al. 2020d; Argaw et al. 2021b,a). Nonetheless, the vulnerability of deep neural networks (DNNs) to adversarial examples (Szegedy et al. 2013) has attracted significant attention in recent years. In machine learning, there is a surging interest in understanding the reason for the success of the adversarial attack (AA) (Szegedy et al. 2013; Zhang et al. 2020b). The root reason for this booming interest lies in the misalignment between human vision and DNN perception (see Figure 1). A similar misalignment phenomenon has also been observed in deep steganography (DS) (Baluja 2017; Zhang et al. 2020c), where a decoder network retrieves a secret image from a slightly perturbed cover image, often referred to as container

* Equal contribution
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
One work (Geirhos et al. 2019) shows that DNNs are more biased towards HF content. We verify this by presenting extensive analysis. We acknowledge that this does not constitute a major discovery, instead, we apply it to explain the model robustness to UAPs in the context of independent yet conflicting features in the $I + P$. Regarding the Fourier perspective to model robustness, adversarial perturbations are widely known to have the HF property, motivated by which several defense explanations for IDPs do not come to a consensus that can be directly used to explain the existence of UAPs. The image-agnostic nature of UAPs requires a specific explanation. Early works (Moosavi-Dezfooli et al. 2017a,b; Jetley, Lord, and Torr 2018; Moosavi-Dezfooli et al. 2019) focused on why a single UAP can fool most samples across the decision boundary and they attributed the existence of UAPs to the large curvature of the decision boundary. (Zhang et al. 2020b) shows that UAPs have independent semantic features that dominate the image features. Recently, class-wise UAPs (Zhang et al. 2020a) and double targeted UAPs (Benz et al. 2020) have also been investigated for making the universal attack more stealthy. Refer to (Zhang et al. 2021b) for a survey on universal attack.

On Universal Adversarial Attack. The reason for the existence of IDP has been analyzed from various perspectives (Qiu et al. 2019), such as local linearity (Goodfellow, Shlens, and Szegedy 2015; Tabacof and Valle 2016), input high-dimension (Shafahi et al. 2019; Fawzi, Fawzi, and Fawzi 2018; Mahloujifar, Dyochnos, and Mahmoody 2019; Gilmer et al. 2018), limited sample (Schmidt et al. 2018; Tanay and Griffin 2016), boundary tilting (Tanay and Griffin 2016), test error in noise (Fawzi, Moosavi-Dezfooli, and Frossard 2016; Gilmer et al. 2019; Cohen, Rosenfeld, and Kolter 2019), non-robust features (Bubeck et al. 2019; Nakkinan 2019; Ilyas et al. 2019), batch normalization (Benz et al. 2021; Benz, Zhang, and Kweon 2020) etc. These explanations for IDPs do not come to a consensus that can be directly used to explain the existence of UAPs. The image-agnostic nature of UAPs requires a specific explanation. Early works (Moosavi-Dezfooli et al. 2017a,b; Jetley, Lord, and Torr 2018; Moosavi-Dezfooli et al. 2019) focused on why a single UAP can fool most samples across the decision boundary and they attributed the existence of UAPs to the large curvature of the decision boundary. (Zhang et al. 2020b) shows that UAPs have independent semantic features that dominate the image features. Recently, class-wise UAPs (Zhang et al. 2020a) and double targeted UAPs (Benz et al. 2020) have also been investigated for making the universal attack more stealthy. Refer to (Zhang et al. 2021b) for a survey on universal attack.

When Adversarial Examples Meet Deep Steganography. Applying deep learning to steganography (Sharda and Budhiraja 2013; Shivaram et al. 2013) has recently become an active research field. Hiding binary messages has been explored in (Hayes and Danezis 2017; Zhu et al. 2018; Wengrowski and Dana 2019) and hiding image (or videos) has been explored in (Baluja 2017; Weng et al. 2018; Mishra et al. 2019). It is crucial to understand how the DNN works in DS. (Baluja 2017, 2019) disproves the possibility of the secret image being hidden in the least significant bit (LSB).

Figure 1: Misalignment between human perception and DNN perception under the universal framework. $D$ stands for Decoder DNN, while $T$ is the Target DNN. (a) USP induced misalignment; (b) UAP induced misalignment. In both (a) and (b): given $||P|| \ll ||I||$, $H(I + P) \approx H(I)$ while $M(I + P) \approx M(P)$, where $I$ indicates a clean image, $P$ is an amplified perturbation, and $I + P$ is a perturbed image. To both sides example images and their Fourier images for the respective task are shown. From top to bottom the images represent: clean image ($I$), amplified perturbation ($P$), and perturbed image ($I + P$). The corresponding Fourier images show that $P$ has HF property contrary to that of $I$. 

(a) USP induced misalignment

(b) UAP induced misalignment
Recent work (Zhang et al. 2020c) shows that the success of DS can be attributed to the frequency discrepancy between cover image and encoded secret image. Refer to (Zhang et al. 2021c) for a survey on deep hiding. Joint investigation of AA and DS has also been investigated by proposing a unified notion of black-box attacks against both tasks (Quiring, Arp, and Rieck 2018), applying the lesson in multimedia forensics to detect adversarial examples (Schöttle et al. 2018). Our work differentiates by focusing on the “universal” property with a Fourier perspective.

**Fourier Transform and Frequency**

Since a large portion of the analysis in this work is dependent on the understanding of image frequency, here we summarize the main points regarding the Fourier transform which is one basic tool to perform image frequency analysis. Sharp contrast edges in the spatial image are considered as HF content, while smooth or constant patches are LF (Lim 1990). Natural images have the Fourier spectrum concentrated in low-medium frequency range that are in the center of the Fourier image. For performing frequency filtering, we define $X_f = \mathcal{F}^{-1}(f(\mathcal{F}(X), bw))$, where $f$ indicates frequency filtering with the bandwidth $bw$. For high-pass (HP) filtering, $f(z(i,j), bw)=z(i,j)$ if $|i-W/2| > bw/2$ or $|j-H/2| > bw/2$, otherwise zero; for low-pass (LP) filtering, $f(z(i,j), bw)=z(i,j)$ if $|i-W/2| <= bw/2$ and $|j-H/2| <= bw/2$, otherwise zero. $W$ and $H$ are image width and height. Fourier images provide a qualitative presentation for the frequency analysis. No metric has been found to quantify the frequency distribution; to facilitate quantitative cosine similarity analysis in this work, we introduce one simple metric: entropy of the Fourier image $z$, i.e. $E(z) = -\sum_i \sum_j p(z(i,j)) \log(p(z(i,j)))$ with $p(z(i,j))$ referring to element probability. Higher entropy indicates more energy being spread to HF regions of $z$, thus indicating the image has more HF content. Note that the entropy is calculated on the transform image $z(i,j)$ instead of the original image.

**Methods for USP and UAP**

Technically, UAPs are crafted to attack a target DNN while DS learns a pair of DNNs for encoding/decoding. Both tasks share a misalignment phenomenon between the human observer and the involved DNN. Specifically, in both cases, a human observer finds that the perturbed image looks natural, but the DNN gets fooled (for AA) or reveals a hidden image (for DS). Motivated by the observation of shared misalignment phenomenon, we deem it meaningful to study the two tasks in parallel to provide a unified perspective on this phenomenon.

**Adopted USP Generation Method**

Our adopted universal secret perturbation (USP) framework (Zhang et al. 2020c) is shown in Figure 2. Through a decoder DNN, a secret image $S$ is transformed into a secret perturbation $S_p$, i.e. USP. This $S_p$ can be randomly added to any cover $C$, resulting in container $C'$. From $C'$, the decoder retrieves the hidden secret image $S'$. Following (Zhang et al. 2020c) we use the average pixel discrepancy (APD), defined as the $L_1$-norm of the gap between two images, to measure the hiding and revealing performance.

Quantitative results evaluated on the ImageNet validation dataset are shown in Table 1. The two scenarios of IDP and USP are performed with the same procedure as in (Zhang et al. 2020c). The qualitative results are shown in Figure 3, where the difference between $C$ and $C'$ as well as that between $S$ and $S'$ are challenging to identify.

**Adopted UAP Generation Method**

The adopted procedure for generating universal perturbation is illustrated in Algorithm 1, where a differentiable frequency filter $\mathcal{F}$ is adopted to control the frequency of the UAP. We treat the $\mathcal{F}$ as all-frequency pass at this stage, which makes it similar to the UAP algorithm introduced in (Zhang et al. 2020b,a). For $\mathcal{C}$, we adopt the widely used negative cross-entropy loss. Except for the image-agnostic nature, this algorithm can be seen adapted from the widely used PGD attack (Madry et al. 2018; Athalye, Carlini, and Wagner 2018). The vanilla UAP (Moosavi-Dezfooli et al. 2017a) generation process uses DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) to generate a perturbation to push a single sample over the decision boundary and accumulates those perturbations to the final UAP. The
Table 1: Performance comparison for the IDP and USP generation frameworks. We report APD for both cover image (cAPD) and secret image (sAPD). For the secret image, we report the results with the container image (sAPD(C′)) or only perturbation (sAPD(S_p)) as the input to the decoder network. N/A indicates revealing fails thus not available.

<table>
<thead>
<tr>
<th>meta-archs</th>
<th>cAPD</th>
<th>sAPD (C′)</th>
<th>sAPD (S_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDP</td>
<td>2.44</td>
<td>3.42</td>
<td>N/A</td>
</tr>
<tr>
<td>USP</td>
<td>2.37</td>
<td>3.52</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Table 2: Performance for untargeted (top) and targeted (bottom) attack with target class “red panda”. The reported values are untargeted and targeted fooling ratio (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>AlexNet</th>
<th>GoogleNet</th>
<th>VGG16</th>
<th>VGG19</th>
<th>ResNet152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our UAP</td>
<td>94.36</td>
<td>86.03</td>
<td>92.58</td>
<td>94.4</td>
<td>86.67</td>
</tr>
<tr>
<td>Our HP-UAP</td>
<td>91.1</td>
<td>84.4</td>
<td>92.3</td>
<td>90.1</td>
<td>78.4</td>
</tr>
<tr>
<td>Our targ. UAP</td>
<td>73.77</td>
<td>68.87</td>
<td>81.59</td>
<td>78.67</td>
<td>74.0</td>
</tr>
</tbody>
</table>

Algorithm 1: Universal attack algorithm

```
Input: Dataset X, Loss L, Target Model M,
frequency Filter F, batch size b

v ← 0  ▶ Initialization
for iteration = 1,...,N do
    B ← X: |B| = b  ▶ Randomly sample
    g_v ← E_{x,y~B} [\nabla_v L(M(x + F(v)), y)]
    v ← Adam(g_v)  ▶ Update perturbation
    v ← Clamp(v, -\epsilon, \epsilon)  ▶ Clamping
end
```

adopted algorithm is different from the vanilla UAP algorithm (Moosavi-Dezfooli et al. 2017a) by replacing the relatively cumbersome DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) perturbation optimization with simple batch gradients. ADAM optimizer (Kingma and Ba 2015) is adopted for updating the perturbation values. A similar ADAM based approach has also been adopted for universal adversarial training (Shafahi et al. 2020).

Following (Moosavi-Dezfooli et al. 2017a; Poursaeed et al. 2018; Zhang et al. 2020b), we generate the perturbation with $\epsilon = 10/255$ on the ImageNet training dataset and evaluate it on the ImageNet validation dataset. The results for untargeted and targeted UAPs are shown in Table 2. Our simple algorithm achieves high (targeted) fooling ratio.

Explaining the USP Induced Misalignment

In the whole pipeline from $S$ through $S_p$ to $S′$, in essence, the role of the $C$ is just like noise. It is counter-intuitive that the pipeline still works well under such large disturbance($||I|| \gg ||P||$). Due to the independent property of $S_p$, we can visualize $S_p$ directly, which is very crucial for qualitatively understanding how the secret image $S$ is encoded in $S_r$ (Zhang et al. 2020c). The visualization in Figure 4 clearly shows that $S_p$ has very HF content.

Why does USP Have High Frequency? The decoder network recovers $S′$ from $S_p$ but with the existence of $C$ as a disturbance. Intuitively its role can be decomposed into two parts: distinguishing $S_p$ from $C$ in $C′$ and transforming $S_p$ to $S′$. We conjecture that secret perturbation having high frequency mainly facilitates the role of distinguishing. To verify this, we design a toy task of scale hiding, where we assume/force the encoder to perform a trivial transformation as $S_p = Encoder(S) = S/10$. We then only train the decoder network to perform the inverse up-scaling transformation with the natural $C$ as the disturbance. After the model is trained, we evaluate it in two scenarios: with and without the $C$. The revealing results are present in the supplementary 1. We observe that the secret image can be recovered reasonably well without the $C$ but fails to work with the $C$. This suggests the transformation $S_p$ to $S'$ has been trained well but still is not robust to the disturbance of $C$, which indicates trivial encoding just performing the magnitude change fails. Since natural images $C$ mainly have LF content, it is not surprising that $S_p$ is trained to have HF content, which significantly facilitates the decoder to distinguish $S_p$ from $C$. The decoder network is implicitly trained to ignore LF content in $C$, while transforming the HF $S_p$ back to $S′$. Thus, the revealing performance can be significantly influenced by the image frequency property.

Frequency: A Key Factor for Performance. We perform analysis with three types of images: artificial flat images with constant values in each RGB channel, natural images, and noise sampled from a uniform distribution of 0 to 1. The results are shown in Table 3. Note that flat images are LF while noise images have HF property. The secret APD performance decreases with the increase of frequency for both secret images and cover images. Since the secret perturbation $S_p$ mainly has high frequency, the increase of frequency in the cover images will disrupt more on the $S_p$, resulting in the performance to decrease. The task complexity also increases with the increase in the frequency of secret images. Revealing fails when either $S$ or $C$ is random noise.

Explaining the UAP Induced Misalignment

Inspired by the above explanation of USP induced misalignment from the Fourier perspective, we extend it to understanding the success of UAP by showing that the target DNN is highly sensitive to HF content.

Disentangling Frequency and Magnitude. We explore

1Supplementary: https://arxiv.org/abs/2102.06479
frequencies of another. The qualitative results with a hybrid image (Oliva, Torralba, and Schyns 2006), which natural images as well as explore which side dominates in show the usefulness of features with LF or HF content in the model to filtered images. For a normally trained DNN, we setup, we evaluate the performance of a normally trained ally forcing the model to specifically learn either LF or HF is to force the model to utilize high or low frequency. Manu-

learning algorithm does not involve any manual intervention images containing both HF content and LF content and the expected, a higher magnitude supple-

mentary, here we summarize the main findings: As


cos


\[ \cos(\cos(I, M(I + P))) \]

\[ \cos(M(I), M(I + P)) \]

\[ \text{Figure 5: The columns for each image triplet indicate HF image, LF image and hybrid image, respectively.} \]

the target DNN’s sensitivity to features of different frequencies. Specifically, we analyze the dominance of two independent inputs on the combined output with the cosine similarity \( \cos \) metric (Zhang et al. 2020b). \( I \) represents a natural image, while \( P \) is an image that extracts the content of a certain frequency range \( \omega \) which is one control variable. We normalize \( P \) to have uniform magnitude and then multiply it by a new magnitude \( m \) which is another control variable. We then calculate \( \cos(M(I), M(I + P)) \) and \( \cos(M(I), M(I + P)) \). For a detailed result, refer to the supplementary, here we summarize the main findings: As expected, a higher magnitude \( m \) leads to higher dominance. On the other hand, we find that \( \omega \) has an (even more) significant influence on the model prediction. Specifically, higher frequency leads to higher dominance.

Hybrid Images: HF vs. LF. The target DNN achieves high accuracy and we are interested in finding out whether HF content or LF content dominantly contributes to the success. Note that the targeted DNN has been trained on natural images containing both HF content and LF content and the learning algorithm does not involve any manual intervention to force the model to utilize high or low frequency. Manu-

ally forcing the model to specifically learn either LF or HF is possible as performed in (Yin et al. 2019). In contrast to their setup, we evaluate the performance of a normally trained model to filtered images. For a normally trained DNN, we show the usefulness of features with LF or HF content in the natural images as well as explore which side dominates in a hybrid image (Oliva, Torralba, and Schyns 2006), which combines the low frequencies of one image with the high frequencies of another. The qualitative results with \( \text{bw} \) of 20 are available in Figure 5. We observe that a hybrid image visually looks more similar to the LF image. The quantitative results of hybrid images are shown in Table 4. In a hybrid setup, the LF image feature is dominated by the HF one.

The hybrid setup is similar to the universal attack setup because the LF content image is not targeted for any specific HF content image and they are randomly combined. Overall, we observe that the LF image content dominates the human vision, while the HF image content dominates the DNN perception, i.e. prediction. Given the dominance of HF content on the DNN, it is not surprising that the optimization-based UAP with HF property can dominate most natural images.

Class-Wise Robustness Imbalance. We randomly choose a targeted class “red panda” for performing a universal attack on VGG19. We find that robust classes have a targeted attack success rate of around 40%, while that for non-robust classes is 100%. One interesting observation from the qualitative results with Fourier analysis in Figure 7 is that all the classes with high robustness have repetitive semantic feature patterns, i.e. HF features, such as the patterns on the feathers of a peacock. The classes with low robustness have LF feature patterns, such as the monotone color of a white washbasin. A Fourier analysis of samples from these classes confirms that robust classes have more HF features, making them more robust against UAP. This analysis shows that there are significant class-wise robustness disparity and the key factor that influences its robustness is their frequency. This also provides extra evidence that the DNN is biased towards HF features. Our work is the first to report and analyze this class-wise imbalanced robustness against UAP.

Feature Layer Analysis. In contrast to prior works performing dominance analysis on the DNN output (Zhang et al. 2020b), we provide fine-grained feature layer analysis with \( \cos \) to provide deep insight on generalization and robustness of a target DNN, e.g. VGG19. In Figure 8, a high \( \cos \) indicates a dominant contribution to the DNN response. First, with the introduced entropy metric, we explore the influence of the image frequency property on its robustness against UAP. We reveal that images of high entropy (HE) (indicating more HF content) are much more robust to UAP on all feature layers, especially on latter layers (see Figure 8 left). For example, at layer of \( \text{conv} 6 \), \( \cos(M(I), M(I + P)) \) is around 0.9 and 0 for images of HE and LE, respectively. The results clearly show that images with more HF content are more robust, which aligns well with our finding on class-wise robustness imbalance. Second, comparing \( \cos(M(I), M(LP(I))) \) and \( \cos(M(I), M(LP(I))) \) shows \( \cos(M(I), M(LP(I))) \) is higher only in the first two layers and then significantly lower in latter layers (see Figure 8 right). In other words, except for the very first few layers, all layers of DNN are much more sensitive to HF instead of LF content. When \( P \) is noise, \( \cos(M(I), M(LP(I))) \) first decreases and then increases again, with the \( \text{conv} 3 \) being the most vulnerable to noise. The influence of random

<table>
<thead>
<tr>
<th></th>
<th>( C_{\text{Flat}} )</th>
<th>( C_{\text{Natural}} )</th>
<th>( C_{\text{Noise}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{\text{Flat}} )</td>
<td>0.34</td>
<td>1.85</td>
<td>48.06</td>
</tr>
<tr>
<td>( S_{\text{Natural}} )</td>
<td>1.77</td>
<td>3.52</td>
<td>49.48</td>
</tr>
<tr>
<td>( S_{\text{Noise}} )</td>
<td>87.45</td>
<td>98.33</td>
<td>100.47</td>
</tr>
</tbody>
</table>

Table 3: Secret APD performance with three types of images. The rows and columns indicate cover images and secret images, respectively.

<table>
<thead>
<tr>
<th></th>
<th>24</th>
<th>20</th>
<th>16</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF</td>
<td>23.13</td>
<td>31.07</td>
<td>41.79</td>
<td>53.31</td>
</tr>
<tr>
<td>Hybrid HF</td>
<td>15.95</td>
<td>20.39</td>
<td>26.54</td>
<td>34.31</td>
</tr>
<tr>
<td>Hybrid LF</td>
<td>0.87</td>
<td>0.52</td>
<td>0.32</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 4: Top1 accuracy (%) for LF, HF, and hybrid images on the ImageNet val dataset evaluated on the VGG19 network. Hybrid HF indicates the accuracy when the HF images labels are chosen as the ground-truth for the Hybrid images. Parallel reasoning applies to Hybrid LF. The columns indicate the bandwidth.
noise on latter layers is very limited, which provides insight on why DNN is robust against noise but not UAP.

**Joint Analysis for Two Tasks**

**Cross-Task Cosine Similarity Analysis for Class Ranking**

We perform a \( \cos \) analysis between two seemingly unrelated tasks, DS and AA. Specifically, the 1000 ImageNet classes were ranked along the attack success rate metric \( (R_1) \), secret APD metric \( (R_2) \) and the Fourier image entropy metric \( (R_3) \). The ranking plots of \( R_1 \) over \( R_2 \), \( R_5 \) over \( R_1 \), and \( R_3 \) over \( R_2 \) are shown in Figure 6. We find that \( \cos(R_1, R_2) \) is 0.74, indicating high linear correlation for two seemingly unrelated tasks. The fact that class robustness is an indicator of the revealing performance in DS task clearly shows that a certain factor exists to link them and we identify this factor to be frequency. Note that \( R_3 \) is our proposed metric for quantifying the energy distribution (corresponding to each frequency) of Fourier images. \( \cos(R_1, R_3) \) and \( \cos(R_2, R_3) \) are 0.68 and 0.77, respectively, attributing the high correlation between \( R_1 \) ranking and \( R_2 \) ranking to the frequency.

**Investigation of LF Perturbation.** To investigate the behavior of perturbations containing LF features we explore two methods: loss regularization and low-pass filtering. Similar to (Mahendran and Vedaldi 2015) we add a regularization term to the loss function during universal perturbation generation to force the perturbation to be smooth for both tasks. The results are shown in Figure 9. The results show that regularizing the perturbation to enforce smoothness results in a significant performance drop. An LF perturbation can also be enforced by performing an LP filtering to the perturbation before adding the perturbation to the image, for which \( F \) is a differentiable LPF (LP filter) in Algorithm 1. Smoothing the perturbations with an LPF and lead to lower attack success rates, see Figure 10 (top). Overall, regarding model robustness, we find that **UAP that attacks most images is a strictly high-frequency (HF) phenomenon.** Inspired by the above results, we propose a novel high-pass (HP) universal attack, *i.e.* **HP-UAP**, by setting \( F \) to be a differentiable
Figure 10: Examples for LP UAPs (top) and HP UAPs (bottom). The first row shows the perturbations for different bandwidths. The bandwidth (BW) and the achieved fooling ratio (FR) are written above the corresponding perturbation. The second row shows the adversarial example with the predicted class of VGG19 written below. The originally predicted and ground truth class is “fountain pen”.

<table>
<thead>
<tr>
<th>Metric</th>
<th>AlexNet</th>
<th>GoogleNet</th>
<th>VGG16</th>
<th>VGG19</th>
<th>ResNet152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fooling Ratio</td>
<td>93.8</td>
<td>85.0</td>
<td>92.7</td>
<td>95.8</td>
<td>90.3</td>
</tr>
<tr>
<td>sAPD</td>
<td>13.6</td>
<td>8.9</td>
<td>14.2</td>
<td>11.1</td>
<td>11.9</td>
</tr>
</tbody>
</table>

Table 5: Performance evaluation of the proposed USAP.

HPF (HP filter) in Algorithm 1. Overall we observe a performance drop in fooling ratio with increasing bw. Results for the HP UAP generated for VGG19 are shown in Figure 10 (bottom). With bw 60, the perturbation is less visible to the human vision and still achieves a fooling ratio of 90.1%, with only a moderate performance drop compared with the 94.4% for bw 0 without filtering.

Universal Secret Adversarial Perturbation

We explore whether a single perturbation can fool the DNN for most images while simultaneously containing the secret information, termed universal secret adversarial perturbation (USAP). As shown in Figure 11, the secret image S is fed to the encoder DNN to produce the USAP $S_p$. It is then added to a random cover image C to produce the container image $C'$, which fools the target DNN and reveals the secret $S'$ if it is fed to the decoder DNN. Note, that during training, only encoder and decoder DNNs are trained, while the target DNN is a fixed pre-trained classifier. Therefore, after training, the encoder DNN can turn any image into USAP, so the approach is different from the original UAP generation where perturbation’s pixels are updated. We adopt the same USP generation network, while adding an additional loss term $NCE(M(C'), y)$ resulting in the loss function: $L(S_p, S, S', C') = ||S_p|| + \beta ||S - S'|| + \gamma NCE(M(C'), y)$, where NCE indicates the negative cross-entropy loss and y indicates the ground-truth label. We set $\beta$ and $\gamma$ to 0.75 and 0.001, respectively. The USAP is constrained to be in $L_{\infty} = 10/255$. The results in Table 5 and Figure 12 demonstrate a high fooling ratio while containing secret information that can successfully be revealed by the decoder. We are the first to show the existence of such perturbation.

Conclusion

This work has jointly analyzed AA and DS for the observed misalignment phenomenon and explained their success from the Fourier perspective. With the proposed metric for quantifying frequency distribution, extensive task-specific and cross-task analysis suggest that frequency is a key factor that influences their performance and their success can be attributed to the DNN being highly sensitive to HF content. Our feature layer analysis sheds new light on model generalization and robustness: (a) Images with more high-frequency content are more robust against UAP; (b) the influence of LF features on the DNN diminishes in the later layers. We also proposed two new variants of universal attacks: HP-UAP that is less visible to the human and USAP that simultaneously achieves attack and hiding.

Ethics Statement

Due to security concerns, adversarial attack and deep steganography have become popular topics in recent years.
We hope that our work will raise awareness of this security concern to the public.

References


Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In ICLR.


Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In ICLR.

Gilmer, J.; Ford, N.; Carlini, N.; and Cubuk, E. 2019. Adversarial examples are a natural consequence of test error in noise. In ICML.


Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In ICLR.


Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In NeurIPS.

Jetley, S.; Lord, N.; and Torr, P. 2018. With friends like these, who needs adversaries? In NeurIPS.


Liu, C.; and JaJa, J. 2019. Feature Prioritization and Regularization Improve Standard Accuracy and Adversarial Robustness. In IJCAI.


Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In CVPR.


3304