Adversarial Perturbations Straight on JPEG Coefficients

Kolja Sielmann[1][0009−0003−2889−5504] and Peer Stelldinger[1][0000−0001−8079−2797]

Hamburg University of Applied Sciences, 20099 Hamburg, Germany
{kolja.sielmann,peer.stelldinger}@haw-hamburg.de

Abstract. Adversarial examples are samples that are close to benign samples with respect to a distance metric, but misclassified by a neural network. While adversarial perturbations of images are usually computed for RGB images, we propose perturbing straight on JPEG coefficients with the ability to individually control the perturbation applied on each color channel and frequency. We find that perturbation as a function of perceptual distance is most efficient for medium frequencies, especially when JPEG compression is used in defense. Overall, we show that attacks on JPEG coefficients are more efficient than state-of-the-art methods that (mainly) apply their perturbation in RGB pixel space. This is partly due to the use of the $\mathrm{YC_bC_r}$ color space, which allows to perturb luma information exclusively, but also due to perturbing the cosine transform coefficients instead of pixels. Moreover, adversarial training using such JPEG attacks with various frequency weighting vectors results in generally strong robustness against RGB and $\mathrm{YC_bC_r}$ attacks as well.

Keywords: Adversarial Attacks · JPEG · Perceptual Distance.

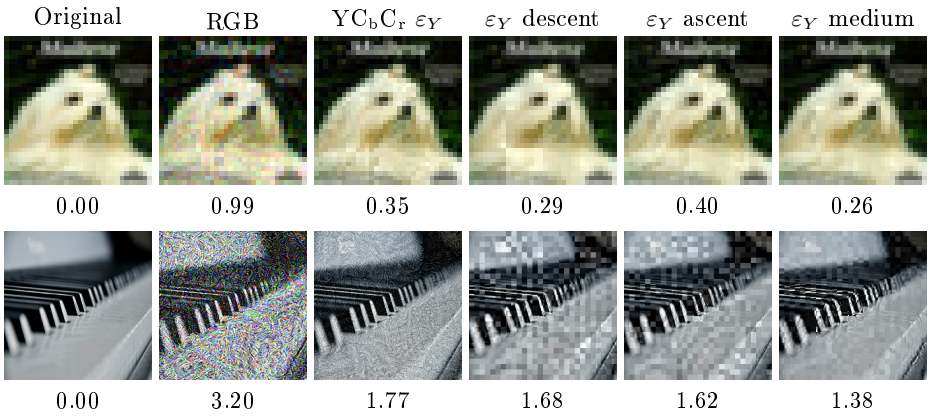


Fig. 1. Adversarial examples with minimum perturbation to force a misclassification on a Densenet^{jq50} for CIFAR10 (top) and IMAGENET (bottom) for attacks on RGB, $\mathrm{YC_bC_r}$ pixels and JPEG coefficients (jq 100) using different frequency weighting vectors. The LPIPS distance is given below each image.

1 Introduction

There has been a large amount of research in recent years on image adversarial attacks and how to defend neural networks against them [3,8,18,29]. Generally, images are represented using RGB pixels and while there has been some research on attacks that try to bypass JPEG compression in defense or when saving the adversarial images [32,33], the main perturbation is mostly still applied in RGB representations. Perturbing straight on JPEG coefficients has only been used as an example of an unforeseen threat model [15,19], but never been analyzed in detail, altough there are several advantages. Adversarial examples exist because neural network classifiers are relying on properties of the data different from those used by humans. In contrast, lossy image compression algorithms like JPEG aim to remove properties of the data which are imperceptible for humans. Thus, a representation of images which separates perceivable from imperceptible parts of the data and enables to control the applied perturbations across frequencies could be a better basis for generating adversarial examples as well as for an adversarial defense.

JPEG compression is known to be a weak defence strategy against adversarial attacks [5,6,10,28]. Attacking straight on JPEG coefficients could increase the success on nets that use JPEG compression in defense, as it could prevent that perturbations are removed during JPEG compression. Moreover, using a $\mathrm{YC_bC_r}$ representation of the image pixels, which is also part of the JPEG compression pipeline, is beneficial for both adversarial attacks and adversarial defense [25]. This leads us to the following questions: Can perturbing straight on JPEG coefficients result in adversarial attacks that are more efficient, i.e. show the same or higher success while being perceptually closer to the original? Can such JPEG attacks indeed bypass JPEG compression with more efficiency than state-of-the-art attacks? Is there a difference between the impact of low-, medium- and high-frequency perturbations on JPEG adversarial attacks? And can these differences be used to achieve more generalising robustness with adversarial training?

Our work is structured as follows: All necessary background on adversarial attacks and defenses, JPEG compression and perceptual metrics is given in section 2. In section 3, we explain our attack method in detail. Section 4 includes the analysis of the efficiency of our approach in comparison with state-of-the-art attacks, followed by a conclusion in section 5. Our implementation is available at https://github.com/KoljaSmn/jpeg-adversarial-attacks.

2 Related Work

2.1 Adversarial Attacks

Following Szegedy et al. [34], given an original input image x with the corresponding ground-truth label y, an image $x' = x + \delta$ is called an adversarial example, if and only if $D(x, x') \leq \varepsilon \wedge f(x') \neq y$, where ε is a hyperparameter limiting the perturbation δ on the original image, D is a distance metric and f is the neural net's output class for a given input. Usually, L_p norms are used

as distance metric. We focus on untargeted attacks that are limited by the L_∞ norm, i.e. attacks that search for an adversarial example x' with a maximum pixel-wise perturbation of ε: $L_\infty(x' - x) = \max_i |x'_i - x_i| \leq \varepsilon$. For generating adversarial examples, we use the Basic Iterative Method (BIM) [18] which iteratively updates the image in the direction of the loss gradient by step size α,

$$x'_t = \text{Clip}_{x,\varepsilon}(x'_{t-1} + \alpha \cdot \text{sign}(\nabla_{x'_{t-1}} \text{J}(x'_{t-1}, y))), \tag{1}$$

where J is the categorical crossentropy loss for some source model.

2.2 Perceptual Metrics

Usually, L_p norms measured in the images' RGB representations are used to evaluate the success of adversarial attacks. These are not really suitable for measuring perceptual differences in real-world scenarios though: First, the RGB color model is based more on physiological properties [26], than on perceptual ones, and it is not perceptually uniform [21]. Second, L_p norms only compute pixel-wise differences and cannot measure structural differences.

Recently, there has been a development towards using perceptually more meaningful distances for measuring and minimizing the distortions created by adversarial attacks: Zhao et al. [39] minimize the CIEDE2000 distance which has been designed to measure perceived color distances [20]. Others have used the Learned Perceptual Image Patch Similarity (LPIPS) to either measure or minimize the perceptual distortion [13,19]. LPIPS is a perceptual loss function that was proposed by Zhang et al. [38] and uses the differences between activations of some convolutional layers in a pretrained network. By relying on differences in feature spaces, LPIPS can also measure structural differences and it has shown to be closer to human perception than pixel-wise distance metrics [38].

The superiority of LPIPS as a perceptual metric can also be illustrated using a simple example, shown in Figure 2, where every pixel in the background was disturbed by adding/subtracting 3 from every RGB channel in two different patterns. On the left, the direction was arranged as a chessboard and on the right it was chosen randomly. The randomly arranged perturbation is not visible without zooming in, while the left image is clearly distinguishable. As the background contains the same pixels on both perturbed images, but the arrangement varies, pixel-wise distances like CIEDE2000 L_2 do not vary, while LPIPS does, corresponding to human perception. In our experiments, we use a VGG-16 net and train the LPIPS model in the same way as in the original paper [38].

2.3 JPEG Compression and JPEG-resistant attacks

JPEG compression builds on the idea that high frequency image content can be altered more before becoming noticeable. In addition to that it takes advantage of human color perception being of lower resolution than human brightness perception. JPEG compression consists of the following steps: First, the pixels are

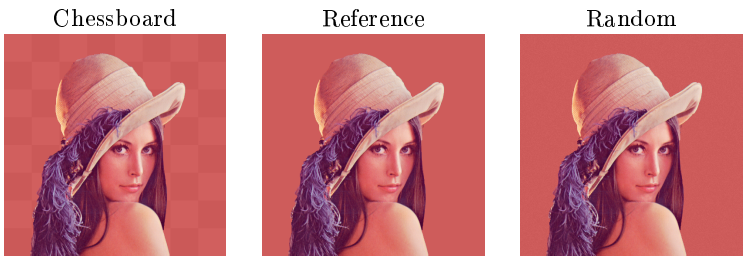


Fig. 2. The reference image's unicoloured background was perturbed with some noise, by adding or subtracting 3 from each RGB channel value. The arrangement of subtracting/adding varies between both patches.

transformed from RGB to $\mathrm{YC_bC_r}$ color space. Second, usually the color channels are subsampled. Third, the three channels are each divided into blocks of size 8x8 pixels which are then replaced by their 64 discrete cosine transform (DCT) [1] coefficients. The main part of the lossy data reduction follows in step four, where the coefficients are divided by some quantization thresholds, which depend on the JPEG quality, and then quantized.

There has been little research on attacks that perturb JPEG coefficients or try to bypass JPEG compression. Kang et al. used JPEG attacks as an example of an unforeseen threat model, but do not give much detail on their attack and do not analyze the attack's success for different parameters [15].

Shin & Song [33] proposed an attack that perturbs the images' RGB pixel representation, but includes an approximation of JPEG compression in the source model. The perturbation of their BIM variant is then given by

$$x'_t = x'_{t-1} + \alpha \cdot \mathrm{sign}(\nabla_{x'_{t-1}} \mathrm{J}(JPEG^{jq}_{\mathrm{approx}}(x'_{t-1}), y)), \tag{2}$$

where $JPEG^{jq}_{\mathrm{approx}}(x)$ is an approximation of JPEG compression, where the rounding during the quantization step is replaced by a differentiable approximation. They also propose an ensemble attack that combines gradients using different JPEG qualities.

Shi et al. [32] proposed an attack that first applies an RGB attack (e.g. BIM) and then compresses the image using a fast adversarial rounding scheme[1] that performs JPEG compression but does not round every coefficient to the nearest integer, but the most important ones in the gradient's direction.

While both approaches can make the perturbation more robust towards JPEG compression, the perturbation is still applied in the RGB pixel representation though. As our attack perturbs straight on JPEG coefficients, it does not have to include an approximation of JPEG compression in the target model or

[1] For targeted attacks, they also propose an iterative rounding scheme. As we only consider untargeted attacks, we will only use the fast adversarial rounding.

make the perturbation robust against JPEG compression using a sophisticated rounding scheme. Thus, our approach is technically straightforward.

2.4 Adversarial Defenses

Goodfellow et al. introduced the concept of defending a neural network against adversarial attacks by adversarial training [8]. The simple idea is to use adversarial examples constructed on the network itself in addition to benign samples for training. In this work, we will use the adversarial training approach by Madry et al., that uses BIM images during training [22].

Since adversarial attacks rely on adding small perturbations to images, other defense methods try to alter or remove such perturbations. I.e. it has been shown, that JPEG compression can reverse the drop in classification accuracy for FGSM [8] attacks as long as the maximal perturbation ε is small enough [6]. It is also known, that training on images of different JPEG compression rates can be used as defense strategy [5,10], although these defenses are only regarded as weak defense strategies [28].

2.5 Adversarial Attacks and Defenses from a Frequency Perspective

By selectively adding noise to different frequencies, Tsuzuku and Sato showed that neural networks are sensitive to the directions of fourier basis functions [36]. Guo et al. and Sharma et al. both masked gradients for high DCT frequencies during RGB pixel perturbation [9,31] and argue that low-frequency perturbations can circumvent certain defenses. Both do not use perceptual distances to evaluate the attacks though, but RGB L_2 distances, or respectively, the input parameter ε, which does not reflect the structural difference of low- and high-frequency perturbations though. Yin et al. argue that adversarial training does increase the robustness on high DFT frequencies, but leads to vulnerability on low frequencies. They also find that neural networks can be successfully trained using high-frequency information that is barely visible to human [37].

Bernhard et al. [2] showed that networks that rely on low frequency information tend to achieve higher adversarial robustness. The intuition behind this is that humans mainly rely on low-frequency information as well. A classifier that uses low-frequency information instead of high frequency components that are barely visible for humans, should thus be more aligned to the human perception; a network relying on exactly the same information as humans, would contradict the existence of adversarial examples. Additionally, Bernhard et al. found that adversarial perturbations are inefficient when limited to high frequencies. Similarly, Maiya et al. [23] state that adversarial perturbations are not necessarily a high-frequency phenomenon, but their distribution across frequencies is dataset-dependent. For CIFAR10 [17], the undefended models are most sensitive on high frequencies, while for IMAGENET [30] they are more sensitive for lower frequencies. For the adversarially trained models the robustness across the whole spectrum increases and in case of CIFAR10 the sensitivity even reverses towards low frequencies.

However, none of these works [2,9,23,31,36,37] are JPEG-related, as they still perturb RGB pixels and perform the DCT/DFT on the whole image and not on 8×8 blocks. Another difference is that they mask some frequencies, while we just weight the perturbations differently.

3 Proposed Method

Let $x^{jq} = (Y, C_b, C_r)$ be a quantized JPEG image of quality jq. For an image of shape $h \times w$, the luma channel Y has shape $(h/8, w/8, 64)$, and the chroma channels also have shape $(h/8, w/8, 64)$, as we do not use chroma subsampling in our attacks in order to focus on the distortion created by the adversarial perturbation. To enable individual control over the perturbation made on each channel, we define three L_∞-balls that limit the relative perturbation made on each channel. For that, we define nine variables, three relative perturbation budgets $\varepsilon_Y^{\text{rel}}, \varepsilon_{C_b}^{\text{rel}}, \varepsilon_{C_r}^{\text{rel}} \in \mathcal{R}_{\geq 0}$, relative step sizes $\alpha_Y^{\text{rel}}, \alpha_{C_b}^{\text{rel}}, \alpha_{C_r}^{\text{rel}} \in \mathcal{R}_{\geq 0}$ and three weighting vectors $\lambda_Y, \lambda_{C_b}, \lambda_{C_r} \in [0,1]^{64}$, where the ε values control the amount of perturbation made on each channel and the λ values determine how much perturbation is permitted for every DCT frequency component. From these relative budgets and the masking vectors, we then compute absolute limits $\varepsilon_Y^{\text{abs}}, \varepsilon_{C_b}^{\text{abs}}, \varepsilon_{C_r}^{\text{abs}} \in \mathcal{R}_{\geq 0}^{(h/8)\times(w/8)\times 64}$ by

$$\varepsilon_Y^{\text{abs}} = \varepsilon_Y^{\text{rel}} \cdot \lambda_Y \cdot |Y|, \quad \varepsilon_{C_b}^{\text{abs}} = \varepsilon_{C_b}^{\text{rel}} \cdot \lambda_{C_b} \cdot |C_b|, \quad \varepsilon_{C_r}^{\text{abs}} = \varepsilon_{C_r}^{\text{rel}} \cdot \lambda_{C_r} \cdot |C_r|. \tag{3}$$

The absolute step sizes $\alpha_Y^{\text{abs}}, \alpha_{C_b}^{\text{abs}}, \alpha_{C_r}^{\text{abs}} \in \mathcal{R}_{\geq 0}^{(h/8)\times(w/8)\times 64}$ are computed correspondingly. For BIM, a single perturbation step is defined by

$$\begin{aligned}
Y'_t &= Y'_{t-1} + \text{sign}(\nabla_{Y'_{t-1}}(\text{J}(\text{rgb}(x'_t), y))) \cdot \alpha_Y^{\text{abs}} \\
C_{b}'_t &= C_{b}'_{t-1} + \text{sign}(\nabla_{C_{b}'_{t-1}}(\text{J}(\text{rgb}(x'_t), y))) \cdot \alpha_{C_b}^{\text{abs}} \\
C_{r}'_t &= C_{r}'_{t-1} + \text{sign}(\nabla_{C_{r}'_{t-1}}(\text{J}(\text{rgb}(x'_t), y))) \cdot \alpha_{C_r}^{\text{abs}},
\end{aligned} \tag{4}$$

where $\text{rgb}(x)$ denotes the transformation from JPEG to unquantized RGB data for JPEG image x. The JPEG to RGB conversion is implemented in a differentiable way using standard convolutional layers with fixed weights. After each update step, the coefficients are clipped to be within each L_∞-ball. After T iterations, the coefficients are rounded to the nearest integer.

4 Experiments and Results

We assume a black-box setting in which a ResNet [11] is used as a source model, and several, partially defended, DenseNets [12] are used as transfer models. Densenet^{jqQ} denotes a normally trained DenseNet, where the input is JPEG compressed with quality Q at inference time. Densenet_M^{RGB} denotes a net that is adversarially trained with Madry et al.'s method [22] that uses RGB BIM to create adversarial images during training. In our experiments, we use all 10000

test images for CIFAR10 and a 10000 image subset of the validation dataset for IMAGENET. We incrementally increase the perturbation bound ε or, respectively, $\varepsilon_Y^{\text{rel}}, \varepsilon_{C_b}^{\text{rel}}, \varepsilon_{C_r}^{\text{rel}}$, [2] measure the success rates and perceptual distances for each attack and then, plot the success rate in dependence of the perceptual distance, which we call the efficiency of an attack.[3]

4.1 Varying Luma and Chroma Perturbations

In our first experiment, we compare the success of our JPEG attacks across color channels. Figure 3 illustrates the attack efficiency on each channel. ε_{all} implies that the same $\varepsilon_Y^{\text{rel}} = \varepsilon_{C_b}^{\text{rel}} = \varepsilon_{C_r}^{\text{rel}}$ is used for all three channels. The other attacks are performed on just one channel each, while the other ε's are set to 0.

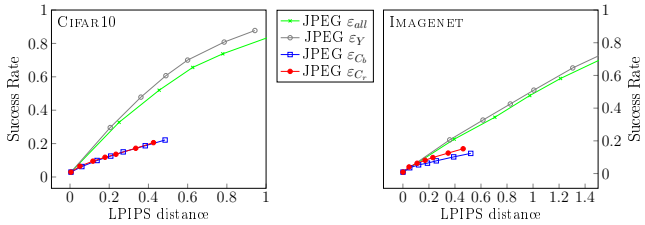


Fig. 3. LPIPS efficiency for unmasked ($\lambda_{all} = 1$) BIM on a Densenet^{jq50}.

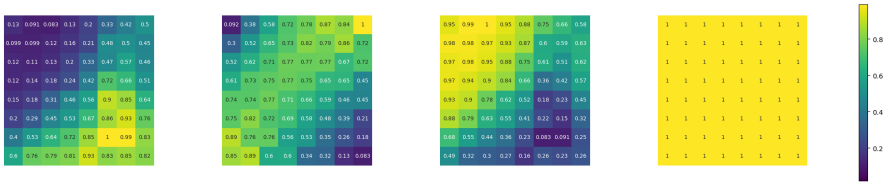
The results confirm Pestana et al. [25], in that adversarial perturbations are much more efficient in the luma than in the chroma channels for both datasets. This is in accordance to neural networks being known to primarily classify based on the image's shapes and textures [7,25]. As luma attacks perform best, our JPEG attacks will only perturb the luma channel in the following.

4.2 Varying Perturbations across Frequencies

Up to now, we applied the same relative perturbation bounds to all JPEG coefficients of one channel although their impact may be very different. One could assume that high frequency perturbations are perceived as noise and are thus less visible than low frequency perturbations. This is in accordance to the behaviour of JPEG compression, which preferably removes high frequency components. On the other hand, low frequency components are perceived as less prominent when high frequency components are visible at the same time, which is the basis of e.g. hybrid image optical illusions [24]. We test this in a second experiment by applying different weighting vectors λ_Y that are illustrated in fig. 4. The

[2] The step sizes are chosen as $\alpha = \frac{\varepsilon}{T}$ for RGB and correspondingly for JPEG attacks. We always use $T = 10$ iterations in our experiments.

[3] For comparison, attacking images with RGB FGSM/BIM and $\varepsilon = 8$ results in an average CIEDE2000 L_2 distance of 203.61/107.17 and a LPIPS distance of 0.85/0.37.



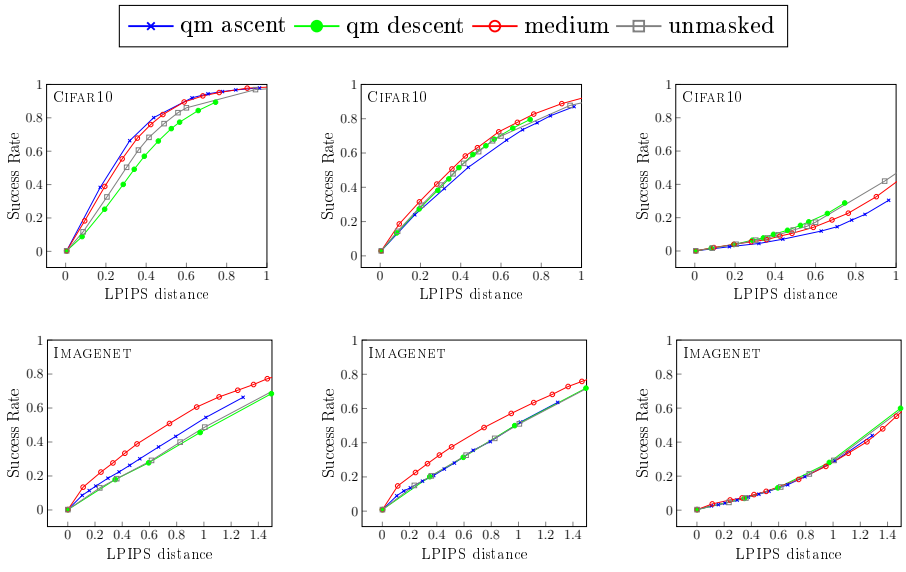
(a) qm ascent (b) medium (c) qm descent (d) unmasked

Fig. 4. DCT weighting vectors (λ). Frequencies increase from top-left to bottom-right.

qm descent/ascent weighting vectors are based on the luma quantization matrix for JPEG quality 50, the medium vector was determined by applying absolute perturbations using the qm descent vector and then extracting the resulting relative perturbations. Figure 5 summarizes the results of the experiment for both datasets. For CIFAR10, we observe that the ascent weighting vector that concentrates perturbations on high frequencies is the most efficient on the undefended DenseNet, closely followed by the medium masking vector. When Densenet^{jq50} is considered, the ascent vector becomes least successful because the perturbation made on high frequencies is removed during JPEG compression. The JPEG quality used for computing the qm descent/ascent weighting vectors does only slightly influence the resulting weighting vectors and thus, the attack's efficiency. I.e., weighting vectors that use JPEG qualities that are different from the one used in defense (50 in this case) do not yield a significant reduction in efficiency. On the adversarially trained net, the order is reversed compared to the undefended net. As already stated by Yin et al. [37], adversarial training does lead to more robustness on high frequencies but vulnerability on low ones, at least for CIFAR10.

For IMAGENET, the difference between the efficiency of low-frequency and high-frequency perturbations is smaller which indicates that the undefended net is indeed more sensitive towards low-frequency perturbations than nets trained on CIFAR10 as already found by Maiya et al. [23], but the medium vector shows even more success. Again, we observe that using JPEG compression in defense decreases the success of high-frequency perturbations as the ascent vector's efficiency is decreased, while the efficiency of medium and low frequency perturbation is less affected by JPEG compression. While a lower JPEG quality could reduce their efficiency too, the relative results should remain the same. Here, we can not observe any outstanding vulnerabilities resulting from the adversarial training as all vectors are similarly successful.

Contrary to the general assumption that adversarial perturbations are mainly a high-frequency phenomenon, the experiments show that medium frequency perturbations are the most efficient (IMAGENET), or at least approximately on par with the best other perturbations (CIFAR10). Note that these observation should apply for perturbations in the medium frequencies in general and we do not state that our selection of the medium vector is optimal. The results also



(a) Densenet (b) Densenet^{jq50} (c) Densenet$^{RGB}_{M}$

Fig. 5. LPIPS efficiency of JPEG luma BIM attacks.

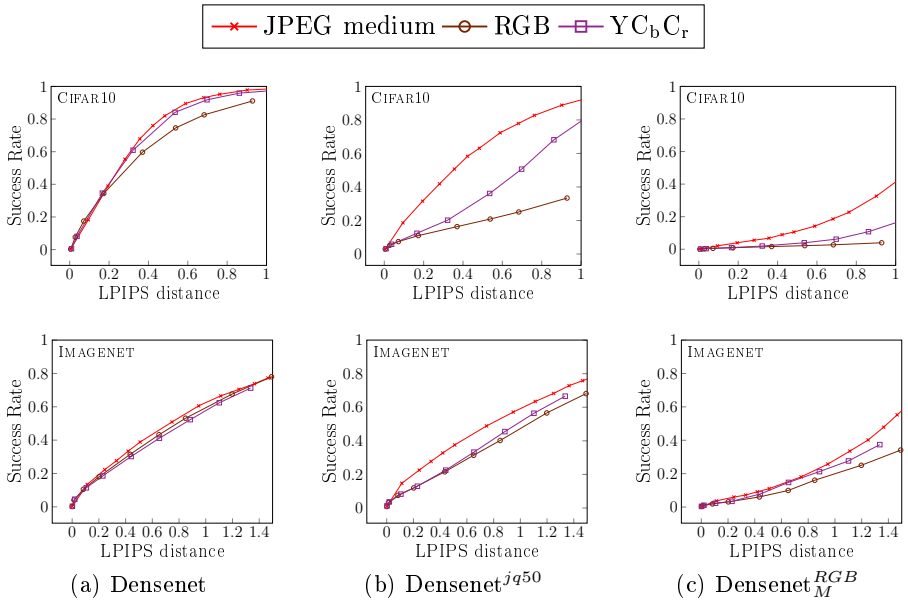
correspond to findings in steganography and watermarking [4,14,16,27], where information is preferably added in medium frequency components since these have less contribution towards energy and perceived image details. Following these results, we will use the medium weighting vector in the following experiments for comparison with state-of-the-art attacks.

4.3 Comparison of Adversarial Attacks on JPEG Coefficients to YC_bC_r and RGB Pixel Representations

In this section, we will analyze whether our attacks straight on JPEG coefficients are advantageous over attacks on pixel representations, such as the usual RGB attacks. However, to determine whether the advantages of JPEG attacks are due to using the YC_bC_r color model or using DCT coefficients, our experiments also include pure YC_bC_r pixel attacks, where we only perturb the luma channel but perform absolute perturbations on pixel values similar to standard RGB attacks.

The results (fig. 6) show, that our JPEG attack is more successful than the YC_bC_r and RGB attack on both CIFAR10 and IMAGENET, especially on the net defended with JPEG compression. There are two main reasons for the superiority of the JPEG approach: First, as shown in section 4.1 and already stated by Pestana et al. [25], adversarial attacks are much more effective when only luma information is perturbed, as the shape and textures that neural networks rely on for classification are mainly located here. Attacking RGB pixels always implies that color information is also changed, resulting in more perceptual difference

than needed and thus, less efficiency. Second, perturbing JPEG coefficients allows to control how perturbations are distributed across frequencies and fixing 0-coefficients (as being done implicitly by our approach) avoids perturbing high-frequency information that would be removed during the JPEG compression in defense anyway. This proofs that the advantage of attacking on JPEG coefficients is not exclusively reasoned by the use of the YC_bC_r color model, but also because of the DCT representation.

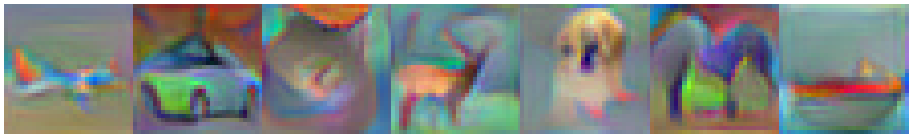


(a) Densenet (b) Densenet^{jq50} (c) Densenet$^{RGB}_{M}$

Fig. 6. LPIPS efficiency of JPEG (jq 100), RGB and YC_bC_r BIM.

The Densenet$^{RGB}_{M}$s are basically robust against the RGB attack, while they still are vulnerable towards our JPEG attack. This alone is not surprising as this RGB attack is what is used during the adversarial training and adversarially trained nets are known to be vulnerable towards unseen threat models [15,19], but the difference between the JPEG and the YC_bC_r attack is worth mentioning as both are attacking on luma exclusively. For CIFAR10, we also adversarially trained a net with JPEG ε_{all} and ε_Y attacks using our four frequency vectors with probability ratios medium:ascent:descent:unmasked 8:5:4:1. The ε_{all}-attacks are weighted twice as much as the ε_Y-attacks. The JPEG adversarial training leads to high overall robustness regarding not only JPEG but also YC_bC_r and RGB attacks: the success rates for ~ 0.9 LPIPS distance are 8.60% (JPEG medium), 2.14% (YC_bC_r) and 6.01% (RGB), compared to 32.53%, 10.72% and 3.97% for an RGB defense. However, the clean accuracy drops from 82.09% (Densenet$^{RGB}_{M}$) to 74.74%, as expected, since "robustness may be at odds with accuracy" [35]. The adversarial training using JPEG attacks leads to the

(a) Densenet$_M^{RGB}$



(b) Densenet$_M^{JPEG}$

Fig. 7. Starting from unicolored images (R=G=B=128), each model's loss was minimized in 100 steps of gradient descent for some of CIFAR10's classes.

net relying on low-frequency information even more than the Densenet$_M^{RGB}$, as fig. 7 illustrates. The reliance on the general composition of images and abstract, coarse structures rather than superficial high-frequency information results in better robustness against small perturbations and better generalization.

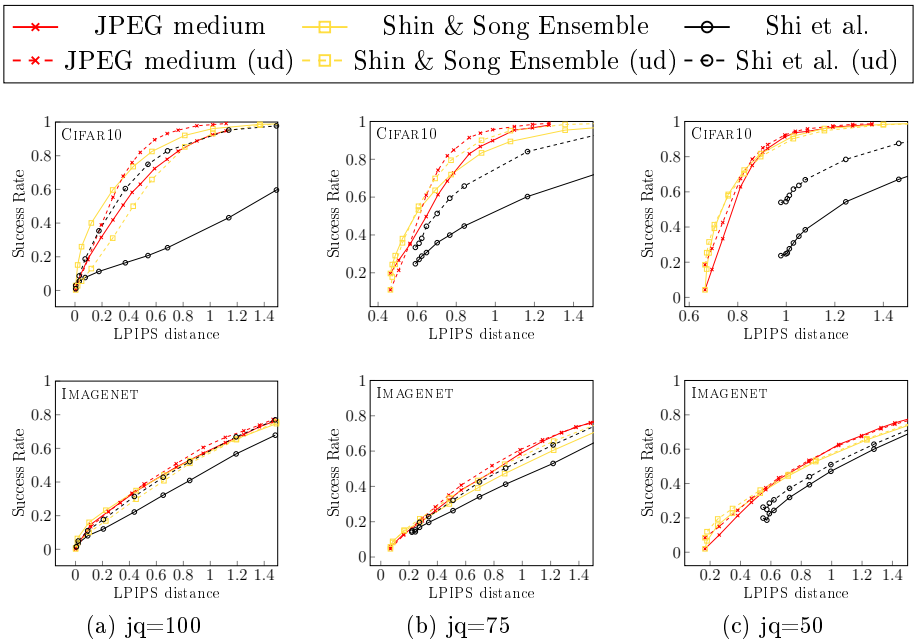
As the examples in Figure 1 show, RGB and $\mathrm{YC_bC_r}$ pixel attacks often result in clearly visible colour or high-frequency noise, while the perturbations from the JPEG medium attack are generally less visible. One problem of our JPEG attacks is that some 8×8 JPEG blocks are clearly visible when there are too strong perturbations on low frequencies.

4.4 Comparison with JPEG-resistant attacks

One of our main motivations for proposing attacks straight on JPEG coefficients was bypassing JPEG compression in defense, or when saving the adversarial images. Thus, we compare our approach to Shin & Song's [33] and Shi et al.'s [32] attacks, for three JPEG qualities used in attack. The experiment is conducted for both CIFAR10 and IMAGENET (fig. 8). As Shin & Song return uncompressed images, we compress them to the same JPEG qualities for comparison.

On both datasets, we observe that our attack is superior compared to Shi et al.'s approach for all three JPEG qualities. Although the fast adversarial rounding in their attack makes the perturbations more robust towards JPEG compression for a given ε, it also induces a significant perceptual distortion even if $\varepsilon = 0$. Additionally, their attack still perturbs mainly on RGB pixels which results in color perturbations that are partially removed during the chroma subsampling in the JPEG compression in defense anyway.

In comparison with Shin & Song's approach the results are less clear. As it always shows better performance, we only include the ensemble attack from



(a) jq=100 (b) jq=75 (c) jq=50

Fig. 8. LPIPS efficiency on the Densenet^{jq50} and the undefended (ud) DenseNet for BIM attacks, for different JPEG qualities used in attack.

Shin & Song [33] in our experiments.[4] For very small perturbations, the attacks from Shin & Song show more efficiency, while for larger perturbations our attack is often more efficient, especially on IMAGENET. Moreover, the ensemble attack's efficiency is significantly worse than ours when the target model is undefended. We explain this by the chroma subsampling that is used in the attack but not in the undefended net. Thus, the attack could induce color perturbations that are ideal to fool nets that use chroma subsampling in defense. In a black-box setting though, it would be unknown whether and how the target model is defended. Therefore, generalizing for a number of target models is an important measure of an attack's success in the black-box setting. In total, our JPEG attack though seems to generalize very well, as it performs well on both undefended and defended nets, and efficiency barely differs between the attack qualities used in attack. Another advantage is our attack's smaller computation time: On a NVIDIA P6000, attacking the whole CIFAR10 test dataset with 10 iterations took only $72s$ compared to $218s$ for Shin & Song's ensemble attack.

As the sample images in Figure 9 show, both Shin & Song's and Shi et al.'s attacks show significant perturbations (colour noise, high-frequency noise and/or block artifacts), while the medium frequency attack shows less perceptual distortions and a smaller LPIPS distance.

[4] For Shin & Song's ensemble attack, jpeg qualities 90, 75 and 50 were used for all three subfigures.

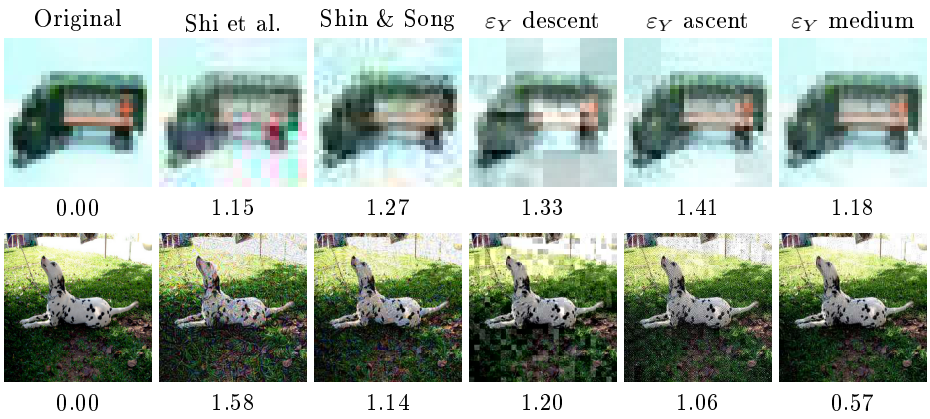


Fig. 9. Adversarial examples with minimum perturbation to force a misclassification on a Densenet^{jq50} for CIFAR10 (top) and IMAGENET. Images are created on a ResNet. The LPIPS distance is given below each image.

5 Conclusion

We introduced a JPEG version of the BASIC ITERATIVE METHOD that allows individual control over the allowed perturbation on each channel and frequency. We found that applying perturbations straight on JPEG coefficients has several advantages that lead to superiority over perturbing straight on RGB or $\mathrm{YC_bC_r}$ pixels:

First, JPEG uses the $\mathrm{YC_bC_r}$ color model which is well-suited for adversarial attacks as it separates chroma and luma information which is often more important for neural networks and thus more efficient to perturb than color channels.

Second, the ability to control the perturbation applied on each frequency allowed us to find that adversarial perturbations on medium frequencies are often more efficient than when they are concentrated on the highest frequencies, especially when JPEG compression is used in defense.

Third, perturbing straight on JPEG coefficients and fixing 0-coefficients allows to apply only perturbations that are not removed during JPEG compression such that it can often bypass JPEG compression more efficiently than state-of-the-art attacks.

Fourth, our approach is much simpler than other methods which try to bypass JPEG compression [32,33], but still generally outperforms them regarding success rate.

Finally, we observed that RGB adversarial training can indeed lead to vulnerability on low frequencies, while adversarial training using multiple, weighted JPEG attacks results in strong overall robustness – not only against JPEG attacks, but also against RGB and $\mathrm{YC_bC_r}$ pixel attacks.

Thus, adversarial perturbations straight on JPEG coefficients leads to more successful attacks and can be used for a generally robust defence strategy.

References

1. Ahmed, N., Natarajan, T., Rao, K.: Discrete cosine transform. IEEE Transactions on Computers **C-23**(1), 90–93 (1974). https://doi.org/10.1109/T-C.1974.223784
2. Bernhard, R., Moëllic, P.A., Mermillod, M., Bourrier, Y., Cohendet, R., Solinas, M., Reyboz, M.: Impact of spatial frequency based constraints on adversarial robustness. In: International Joint Conference on Neural Networks, IJCNN (2021). https://doi.org/10.1109/IJCNN52387.2021.9534307
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE (2017). https://doi.org/10.1109/SP.2017.49
4. Chang, C.C., Lin, C.C., Tseng, C.S., Tai, W.L.: Reversible hiding in dct-based compressed images. Information Sciences **177**(13), 2768–2786 (2007), https://www.sciencedirect.com/science/article/pii/S0020025507001016
5. Das, N., Shanbhogue, M., Chen, S., Hohman, F., Chen, L., Kounavis, M.E., Chau, D.H.: Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. CoRR (2017), http://arxiv.org/abs/1705.02900
6. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of JPG compression on adversarial images. CoRR (2016), http://arxiv.org/abs/1608.00853
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations, ICLR (2019), https://openreview.net/forum?id=Bygh9j09KX
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations, ICLR (2015)
9. Guo, C., Frank, J.S., Weinberger, K.Q.: Low frequency adversarial perturbation. In: Uncertainty in Artificial Intelligence Conference. PMLR 115 (2020), https://proceedings.mlr.press/v115/guo20a.html
10. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: International Conference on Learning Representations, ICLR (2018), https://openreview.net/forum?id=SyJ7ClWCb
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016). https://doi.org/10.1109/CVPR.2016.90
12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017). https://doi.org/10.1109/CVPR.2017.243
13. Jordan, M., Manoj, N., Goel, S., Dimakis, A.G.: Quantifying perceptual distortion of adversarial examples. CoRR (2019), https://arxiv.org/abs/1902.08265
14. Kahlessenane, F., Khaldi, A., Kafi, R., Euschi, S.: A robust blind medical image watermarking approach for telemedicine applications. Cluster computing **24**(3), 2069–2082 (2021)
15. Kang, D., Sun, Y., Hendrycks, D., Brown, T., Steinhardt, J.: Testing robustness against unforeseen adversaries. CoRR **abs/1908.08016** (2019), http://arxiv.org/abs/1908.08016
16. Khan, S., Irfan, M., Arif, A., Rizvi, S.T.H., Gul, A., Naeem, M., Ahmad, N.: On hiding secret information in medium frequency dct components using least significant bits steganography. CMES-Computer Modeling in Engineering & Sciences **118**(3) (2019)

17. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (2012), https://www.cs.toronto.edu/~kriz/cifar.html
18. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: International Conference on Learning Representations, ICLR (2017), https://openreview.net/forum?id=HJGU3Rodl
19. Laidlaw, C., Singla, S., Feizi, S.: Perceptual adversarial robustness: Defense against unseen threat models. In: International Conference on Learning Representations, ICLR (2021), https://openreview.net/forum?id=dFwBosAcJkN
20. Luo, M.R., Cui, G., Rigg, B.: The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Research & Application **26**(5), 340–350 (2001). https://doi.org/10.1002/col.1049
21. MacDonald, L.: Using color effectively in computer graphics. IEEE Computer Graphics and Applications **19**(4), 20–35 (1999). https://doi.org/10.1109/38.773961
22. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations, ICLR (2018), https://openreview.net/forum?id=rJzIBfZAb
23. Maiya, S.R., Ehrlich, M., Agarwal, V., Lim, S., Goldstein, T., Shrivastava, A.: A frequency perspective of adversarial robustness. CoRR (2021), https://arxiv.org/abs/2111.00861
24. Oliva, A., Torralba, A., Schyns, P.G.: Hybrid images. ACM Trans. Graph. **25**(3), 527–532 (2006). https://doi.org/10.1145/1141911.1141919
25. Pestana, C., Akhtar, N., Liu, W., Glance, D., Mian, A.: Adversarial attacks and defense on deep learning classification models using YCbCr color images. In: International Joint Conference on Neural Networks, IJCNN (2021). https://doi.org/10.1109/IJCNN52387.2021.9533495
26. Plataniotis, K., Venetsanopoulos, A.N.: Color image processing and applications. Springer (2000)
27. Pradhan, C., Saxena, V., Bisoi, A.K.: Non blind digital watermarking technique using dct and cross chaos map. In: International Conference on Communications, Devices and Intelligent Systems, CODIS (2012). https://doi.org/10.1109/CODIS.2012.6422191
28. Raff, E., Sylvester, J., Forsyth, S., McLean, M.: Barrage of random transforms for adversarially robust defense. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition,CVPR (2019)
29. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial attacks and defenses in deep learning. Engineering **6**(3), 346–360 (2020). https://doi.org/10.1016/j.eng.2019.12.012
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (2014), https://www.image-net.org/
31. Sharma, Y., Ding, G.W., Brubaker, M.A.: On the effectiveness of low frequency perturbations. In: International Joint Conference on Artificial Intelligence, IJCAI (2019)
32. Shi, M., Li, S., Yin, Z., Zhang, X., Qian, Z.: On generating JPEG adversarial images. In: IEEE International Conference on Multimedia and Expo, ICME (2021). https://doi.org/10.1109/ICME51207.2021.9428243
33. Shin, R., Song, D.: JPEG-resistant adversarial images. In: NIPS 2017 Workshop on Machine Learning and Computer Security, NeuRIPS (2017)
34. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations, ICLR (2014)

35. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: International Conference on Learning Representations, ICLR (2019), https://openreview.net/forum?id=SyxAb30cY7
36. Tsuzuku, Y., Sato, I.: On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR (2019)
37. Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. In: Advances in Neural Information Processing Systems, NeuRIPS (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2018)
39. Zhao, Z., Liu, Z., Larson, M.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR (2020)