Evaluating Adversarial Robustness on Document Image Classification

Timothée Fronteau*, Arnaud Paran*, Aymen Shabou*

* DataLab Groupe of Crédit Agricole S.A, France

Abstract

Adversarial attacks and defenses have gained increasing interest on computer vision systems in recent years, but as of today, most investigations are limited to images. However, many artificial intelligence models actually handle documentary data, which is very different from real world images. Hence, in this work, we try to apply the adversarial attack philosophy on real-world documentary data and to protect models against such attacks. To the best of our knowledge, no such work has been conducted by the community in order to study the impact of these attacks on the document image classification task.

1. Methodology

We generate perturbed images from the RVL-CDIP [1] document images dataset under L_∞ and L_2 perturbation constraints, such that the two evaluated models ResNet50 and EfficientNetB0 misclassify the perturbed images (see Figure 1). We focus our work on untargeted **gradient-based**, **transfer-based** and **score-based attacks**, and evaluate the impact of **grey-scale input transformation**, **JPEG input compression** and **adversarial training** on the robustness of these two state-of-the-art visual models.

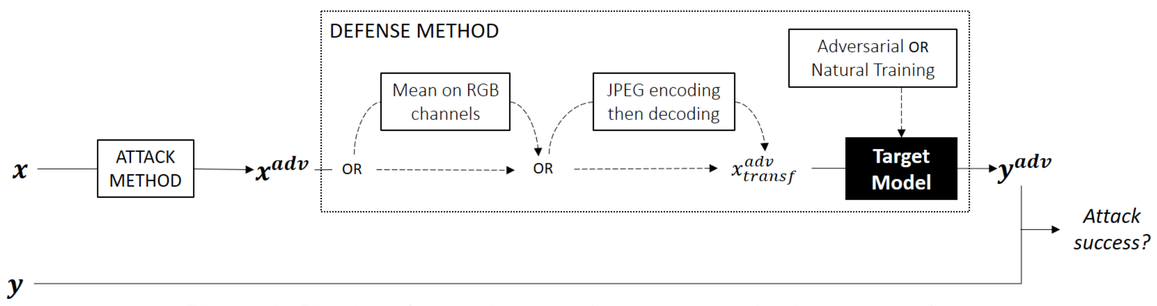


Figure 1: Pipeline for evaluating the adversarial robustness of a defense model under attack

2. Attack Methods

1. **Gradient-based attacks:** We generated adversarial images with the Fast Gradient Sign method [2] (**FGS**) which computes a perturbation in one step using a loss function knowing the target model parameters, as well as the Basic Iterative Method [2] (**BIM**) and the Momentum Iterative Method [2] (**MIM**), which are both sophisticated iterative versions of FGS.

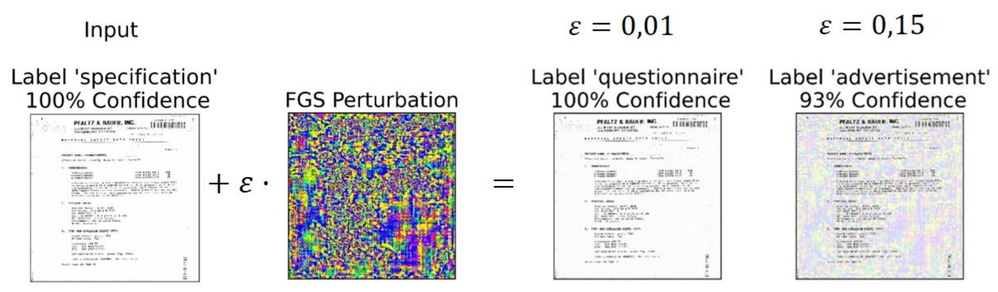


Figure 2: Generation of an adversarial document image with FGM

2. **Transfer-based attacks:** Adversarial images generated with gradient-based attacks on a substitute model have a chance of also leading the target model to misclassification [2]. We used the two most robust of our models to attack other models in a transfer-based setting.
3. **Score-based attack:** We generated adversarial images with the Simultaneous Perturbation Stochastic Approximation method [2] (**SPSA**), which uses the logits output of the target model to *estimate* a gradient and use it to generate a perturbation in a similar way to BIM.

3. Defense Methods

1. **Grey-scale transformation, JPEG compression:** In the testing phase, adding constraints like averaging the RGB channels of a document image (**Grey**) or encoding and decoding it using JPEG protocol (**JPEG**) can improve adversarial robustness [2]. Unlike JPEG compression, the grey-scale transformation does not affect the accuracy of models on the RVL-CDIP dataset, since it is it is a set of grey-scale images (see Table 1).
2. **Adversarial Training:** A training focusing only on maximizing accuracy on a target dataset by fine-tuning a model on legitimate documents may result in poor adversarial robustness of the model. We compared this "**Natural**" training method with an "**Adversarial**" training where the model is trained against adversarial batches generated with BIM along the whole training phase [2] (see test accuracies in Table 1).

Table 1: Test accuracy of each defense model

Defense Method\Model Backbone	EfficientNet	ResNet
No Defense (Natural)	**0.908**	**0.890**
Grey	**0.908**	**0.890**
JPEG	0.868	0.861
Adversarial	0.890	0.873
Grey + JPEG + Adversarial	0.893	0.872

4. Results

L_∞ and L_2 perturbation budgets: We use L_∞ and L_2 norms to constrain the attacker's capability for generating adversarial document images. We draw curves of model accuracy vs. perturbation budget ε for each attack, where $\|x - x^{adv}\| < \varepsilon$ with x an original image and x^{adv} an adversarial image generated with the former (see Figures 3, 4 and 5).

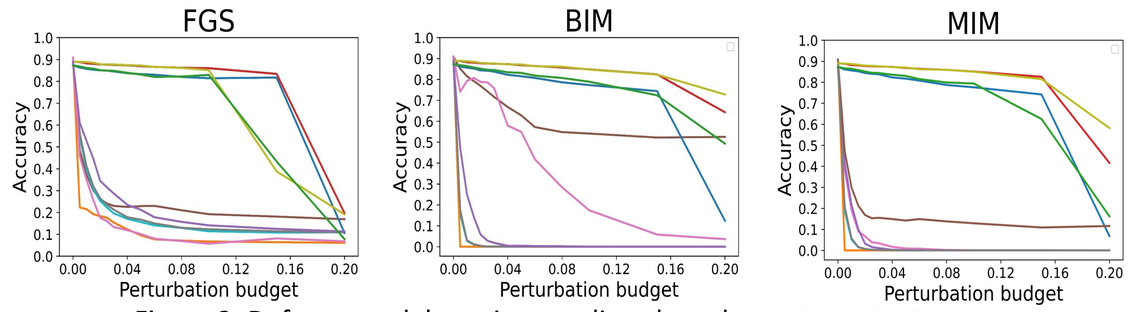


Figure 3: Defense models against gradient-based attacks under L_∞ norm

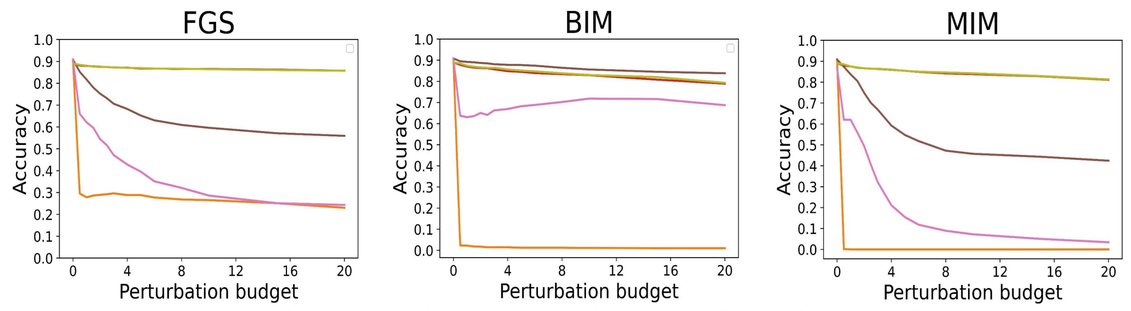


Figure 4: Defense models against gradient-based attacks under L_2 norm

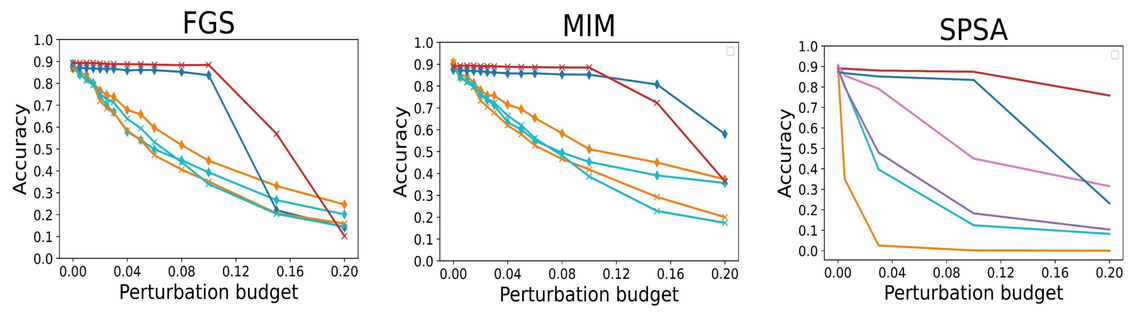
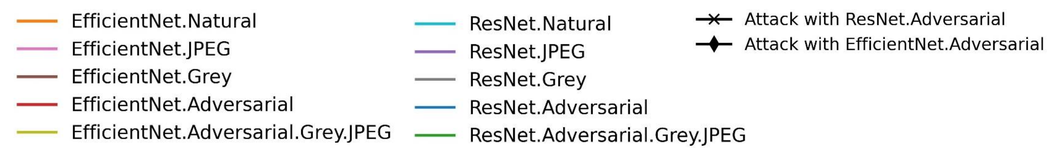


Figure 5: Defense models against transfer-based attacks (FGS, MIM) and against a score-based attack (SPSA) under L_∞ norm

Legend for Figures 3, 4 and 5:



Robustness against gradient-based attacks (Figures 3 and 4):

- Accuracies of undefended models (EfficientNet.Natural, ResNet.Natural) drop drastically from more than 88% to less than 1% under BIM and MIM attack ($\varepsilon = 0{,}01$ and $\varepsilon = 1$ for the L_∞ and L_2 norm respectively).
- The grey-scale transformation and JPEG compression improve inconsistently the adversarial robustness for EfficientNet and ResNet models.
- Accuracies of Adversarially trained models drop by two points compared to undefended models (Table 1), but are very robust to adversarial attacks.

Robustness against transfer-based attacks under L_∞ norm (Figure 5):

- Undefended models are vulnerable to adversarial images generated with more robust models (up to 36 points of accuracy decrease), while adversarially trained models stay robust (less than 4 points of accuracy decrease) for a perturbation budget $\varepsilon = 0{,}10$.

Robustness against a score-based attack under L_∞ norm (Figure 5):

- Undefended models are vulnerable to the SPSA attack, while adversarially trained models are very robust.
- The JPEG compression improves inconsistently the robustness of ResNet and EfficientNet (3-point increase for ResNet vs. 76-point increase for EfficientNet for $\varepsilon = 0{,}03$).

Conclusion and Future Work

Naturally trained, unprotected models are as vulnerable to adversarial attacks on the RVL-CDIP classification task as on the CIFAR-10 or ImageNet classification task, as shown in [2]. However, training document image classification models on adversarial batches highly improves their adversarial robustness, and with an apparent consistency. The efficiency of reactive defenses like JPEG compression or grey-scale transformation to improve adversarial robustness is inconsistent on the two models.

In future works, we explore whether models that combine the visual modality with text and layout information are vulnerable to adversarial document images generated with visual models. We will also study whether attacks on OCR based models can affect their performances.

References

[1] A. W. Harley, A. Ufkes, K. G. Derpanis. *Evaluation of deep convolutional nets for document image classification and retrieval*. In ICDAR, 2015.

[2] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, J. Zhu. *Benchmarking adversarial robustness on image classification*. In Proceedings CVPR, 2020.