Supplementary material: *SMEBA* Repository

PaperID 1421

Brief introduction

SMEBA Repository is a software toolbox for boosting adversarial attack and defense research in pixel-to-pixel vision tasks. Notice that, there is another toolbox about adversarial attack and defense, called as *CleverHans* (Papernot et al. 2016). *CleverHans* focuses on image classification task. The proposed *SMEBA* Repository serves as a good complement to *CleverHans*, especially in the pixel-to-pixel tasks.

The content of *SMEBA* Repository

1) "**Attack_methods_library.py**"

In current version, *SMEBA* Repository integrates 14 gradient descent based adversarial attack methods in this script. Please see Table.1 for more details.

2) "**Pretrained_models.py**"

In current version, *SMEBA* Repository integrates 16 state-of-the-art source models, as shown in the Table. 2.

3) "**Loss_functions.py**"

In current version, *SMEBA* Repository supports 5 loss functions for evaluating the fooling ability, *i.e. KL, CC, NSS, BCE* and *MAE* losses, and 5 perceptual constraints for computing the perceptibility loss, *i.e.* $\mathbf{L_1}$ norm, $\mathbf{L_2}$ norm, $\mathbf{L_\infty}$ norm, *SSIM* and *MS-SSIM*.

4) "**config_global.py**"

This script defines some global variables, explained below. In current version, *SMEBA* Repository supports *output-space* attack, *feature-space* attack, *hybrid-space* (*output-space & feature-space*) attack, *targeted* attack, *non-targeted* attack, *clipped box-constraint*, and *automatic box-constraint*. Users can select their desired attack version in this script.

5) "**options.py**"

Users can select the guide image in this script, and define the saving path of produced visualization results.

6) "**Main_Ensemble_Attack.py**"

This script is the main function for ensemble attack that misleads multiple threat models simultaneously. Users can select any attack to mislead any combination of threat models from library. The adversarial example produced by ensemble attack gains the higher transferability under *black-box* attack setting..

7) "**Main_Single_Image_Space_Attack.py**"

This script is the main function for attacking single threat model from *output-space*.

8) "**Main_Single_Feature_Space_Attack.py**"

This script is the main function for attacking single threat model from *feature-space* and *hybrid-space*. *Hybrid-space* attack produces diverse adversarial examples, which could be used to boost the adversarial training.

9) "**data**"

This folder provides dataset alignment and image load functions.

10) "**DCN_lib**"

This folder provides a library for (modulated) 2D deformable convolution with CUDA acceleration.

11) supporting flexible visualization

In current version, *SMEBA* Repository provides a flexible visualization interface for users. Users can define the desired output (the intermediate feature maps across different layers) in "**Pretrained_models.py**" ("**return**" function of each threat model), and save the visualization results at the specific path defined in the main functions.

12) adversarial training defense

In current version, *SMEBA* Repository supports 2 defensive strategies, *i.e. adversarial training* (injecting the adversarial examples produced by *output-space* attasks into the training set), *hybrid adversarial training* (injecting the adversarial examples produced by *output-space*, and *feature-space* attasks into the training set). We will add another defense in the future, *i.e. model distillation* (using a pre-trained teacher model to supervise and boost the training of a student model).

References

Papernot, N.; Goodfollow, I.; Sheatsley, R.; Feinman, R.; and McDaniel, P. 2016. cleverhans v2. 0.0: an adversarial machine learning library. In *arXiv preprint*.

Table 1: The library of Gradient back-propagation based adversarial attack methods.

Number	Method	Gradient Descent Method	Frequency	Perceptual Constraint	Transferability	Goal	Dataset	Box-constraint
$attack_1$	*AdaEq2M*	1_{st}-randn momentum & 2_{nd}-plus momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_2$	Adam-decay-IM	Adam: vanilla 1_{st} momentum & 2_{nd}-decay momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_3$	Adam-vanilla-IM	Adam: vanilla 1_{st} momentum & vanilla 2_{nd} momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_4$	MSGD-MIM	MSGD: vanilla 1_{st} momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_5$	SGD-IFGSM	SGD: no momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_6$	SGD-IFGV	SGD: no momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_7$	AdaGrad-plus-IM	AdaGrad: 2_{nd}-plus momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_8$	AdaGrad-decay-IM	AdaGrad: 2_{nd}-decay momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_9$	*AdaEq2M-V2*	vanilla 1_{st} momentum & 2_{nd}-plus momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{10}$	Hot&Cold	SGD: no momentum	Iterative	*SSIM*	white-box	Targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{11}$	PGD	SGD: no momentum Start from a random state	Iterative	L_1, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{12}$	Adam-decay-C&W	Adam: only using short-term momentums	Iterative	L_0, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	automatic
$attack_{13}$	*AdaEq2M*-C&W	1_{st}-randn momentum & 2_{nd}-plus momentum	Iterative	L_0, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	automatic
$attack_{14}$	**proposed *SMBEA***	short-term momentums long-term momentums	Iterative	L_1, L_2, L_∞	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped

Table 2: The library of source models: state-of-the-art deep saliency models.

Number	Model	Backbone Network	Multi-scale	Convolution Form	Attention Module	Refinement	Adversarial Training
$model_1$	$\mathrm{GazeGAN}_1$	U-Net	Single-stream	Standard Convolution	∅	∅	∅
$model_2$	$\mathrm{SALICON}_2$	VGG16	Single-stream	Standard Convolution	∅	∅	∅
$model_3$	Global-pix2pix	ResNet	Single-stream	Standard Convolution	∅	∅	∅
$model_4$	$\mathrm{GazeGAN}_2$	U-Net	Multi-stream	Standard Convolution	∅	∅	YES
$model_5$	DCN_LSTM_1	VGG16 & Inception-ResNet	Multi-stream	Standard Convolution & Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	ConvLSTM	∅
$model_6$	DCN_2	VGG16 & Inception-ResNet	Multi-stream	Standard Convolution & Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	∅	∅
$model_7$	SAM_VGG_1	VGG16	Single-stream	Standard Convolution & Dilated Convolution	∅	ConvLSTM	∅
$model_8$	SAM_VGG_2	VGG16	Single-stream	Standard Convolution & Dilated Convolution	Softmax Spatial Attention	ConvLSTM	∅
$model_9$	SAM_ResNet	ResNet	Single-stream	Standard Convolution & Dilated Convolution	Softmax Spatial Attention	ConvLSTM	YES
$model_{10}$	CSC_Net	U-Net	Multi-stream	Standard Convolution	Softmax Spatial Attention	∅	∅
$model_{11}$	SalGAN_BCE	VGG16	Single-stream	Standard Convolution	∅	∅	∅
$model_{12}$	DCN_Inception	Inception-ResNet	Single-stream	Standard Convolution	∅	∅	∅
$model_{13}$	DeepGaze_only_VGG	VGG19	Single-stream	Standard Convolution	∅	∅	∅
$model_{14}$	DCN_SAM_VGG	VGG19	Single-stream	Standard Convolution & Modulated Deformable Convolution	Softmax Spatial Attention	ConvLSTM	∅
$model_{15}$	Local-pix2pix	ResNet	Multi-stream	Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	∅	∅
$model_{16}$	DenseSal	DenseNet	Single-stream	Standard Convolution	∅	∅	∅