MCS 2018

Adversarial Attacks on Black Box Face Recognition

atmyre, mortido, snakers41, stalkermustang

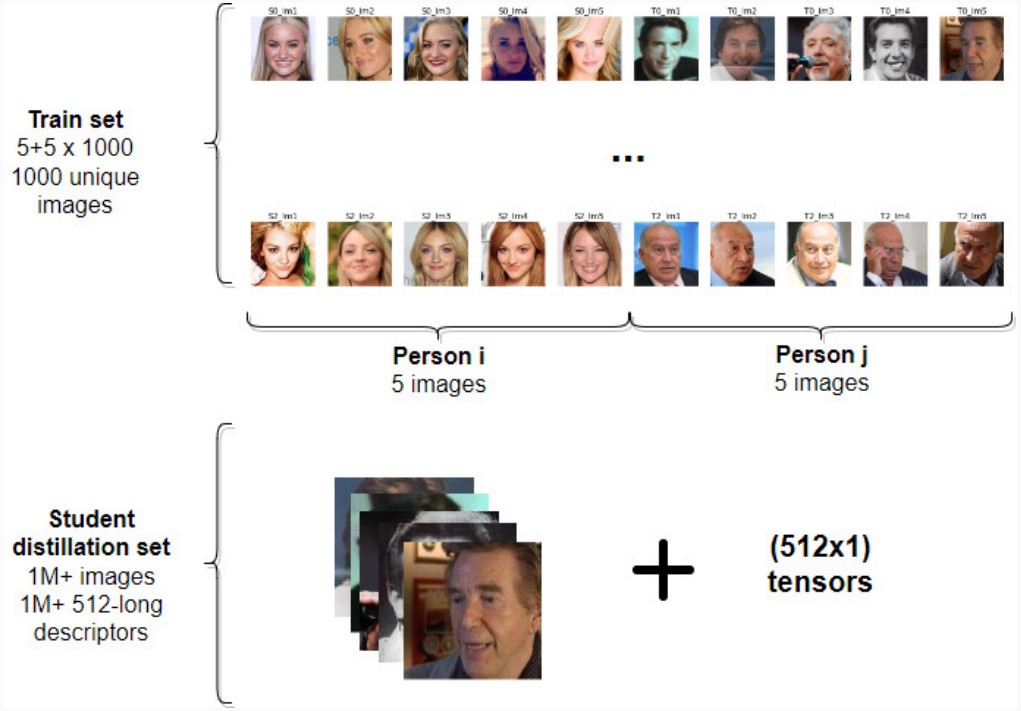
June 7, 2018

Homeless Nonames

Competition overview

How is it this competition different?

- Adequately good baseline solution
- Relatively small attack dataset (1000 pairs, 1000 unique images)
- Small images (250x250, attacks performed on 112x112 images)
- Compiled black box models (which do not always work)



Student Model Results

Table 1: Student Model distillation results

Architecture	Pre-trained (*)	LR regime	Best MSE	Epochs
ResNet18	Yes	(1)	4.51 * 1e-4	25
ResNet34	Yes	(1)	3.36 * 1e-4	25
DenseNet161	Yes	(2)	3.08 * 1e-4	25
XCeption	No	(4)	4.57 * 1e-4	23
ResNet50	No	(4)	4.07 * 1e-4	11

(1) 1e-3 + pre-trained on ImageNet + LR decay + adam
(2) 1e-4 + pre-trained on ImageNet + LR decay + adam
(3) 1e-3 + pre-trained on ImageNet + LR decay + adam
(4) manual adjustments each epoch

Attack results

Attack	Hack	Student CNNs	BB score	LB pub	LB priv
FGSM	-	DenseNet161	1.25	-	-
FGSM	(1)	DenseNet161	1.16	-	-
FGVM	(3)	2 CNNs	0.97	1.05	-
FGVM	(4)	5 CNNs	0.91	-	-
FGVM + 1 pixel	(4)	5 CNNs	0.90	0.99	-
FGVM + 6 pixel	(4)	5 CNNs	0.87	-	-
FGVM + 16 pixel	(5)	5 CNNs	0.87	-	0.96

FGSM - Fast Gradient Sign Method

FGVM - Fast Gradient Value Method

Final heuristics

FGVM

- Noise **eps * clamp(grad / grad.std(), -2, 2)**
- Ensemble of several CNNs via weighting their gradients
- Save changes only if it reduces mean loss
- Use target combinations for more robust targeting

Genetic One Pixel

- popsize = 30
- max_iter = 5

Student CNN distillation

What worked

- Transfer learning
- ADAM + clever LR regime to avoid under-fitting
- Best architectures are reasonably heavy **ResNet34** and **DenseNet161**

What did not

- Inception-based architectures (not-suitable due to high down-sampling)
- VGG based architectures (overfitting)
- "Light" architectures (SqueezeNet / MobileNet - underfitting)
- Image augmentations (w/o modifying descriptors)
- Working with 224x224 images

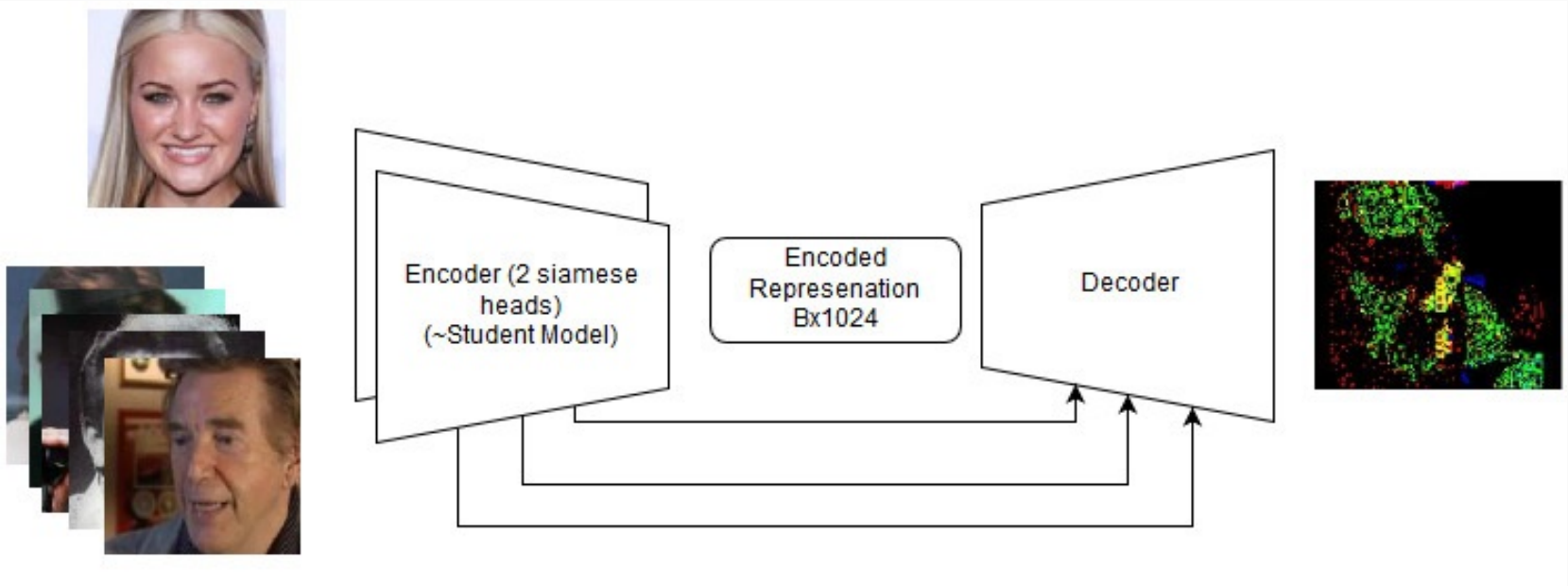
Other Attack Approaches

We also tried:

- **FGVM with momentum**
  https://arxiv.org/abs/1710.06081v3
- **CW** – good for white-box attacks
  https://arxiv.org/abs/1608.04644

End-to-end architectures (1)

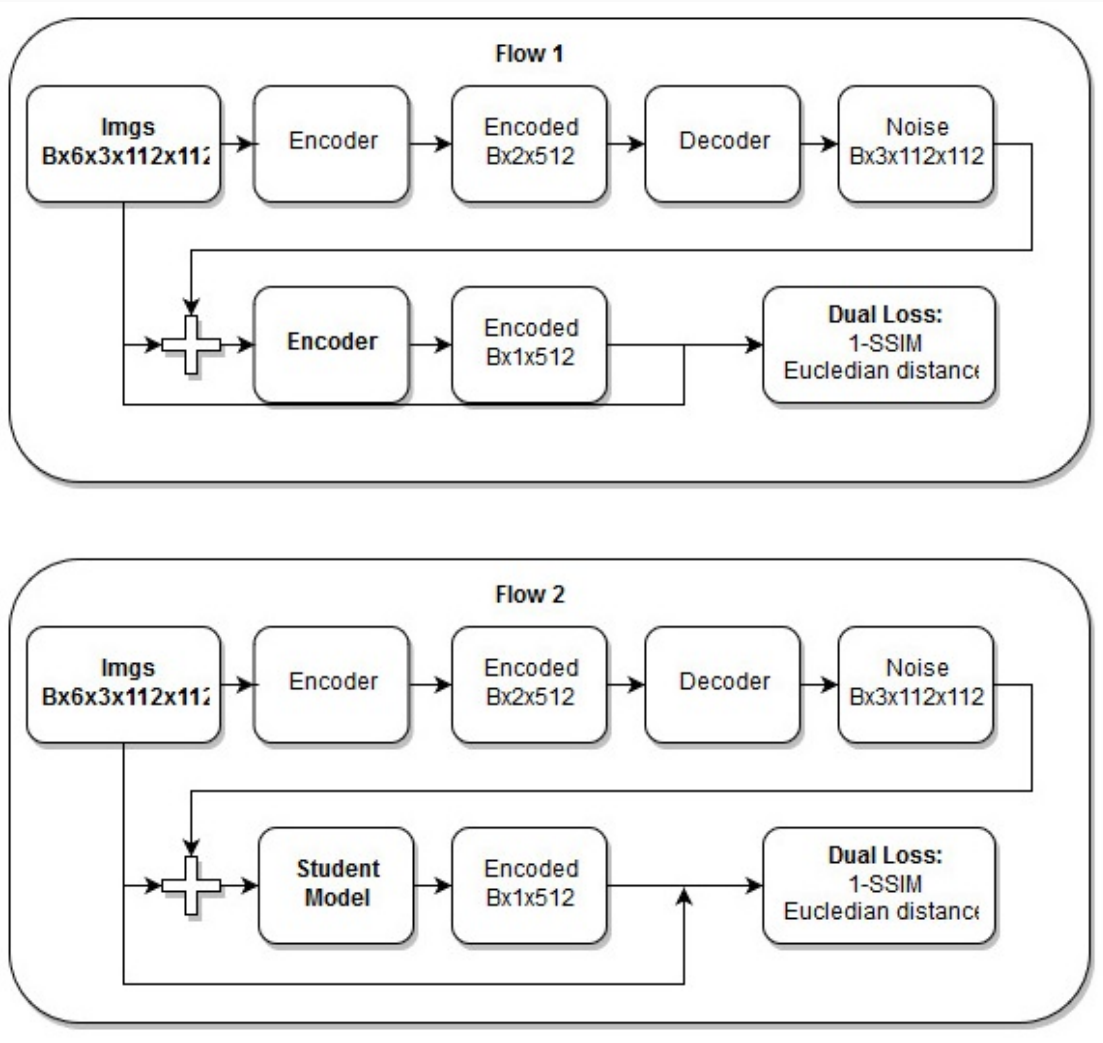
- Key ideas - use a mixture of VAE / Siamese LinkNet
- 2 part loss - PyTorch SSIM + Eucledian distance



End-to-end architectures (2)

Key take-aways

- Performs well on **WB** and poorly on **BB**
- Difficult to balance Loss - use running mean scaling
- Problems with scaling images back - use some eps
- Model parametrization - open question
- Pass image as skip connection



Some fun illustrations