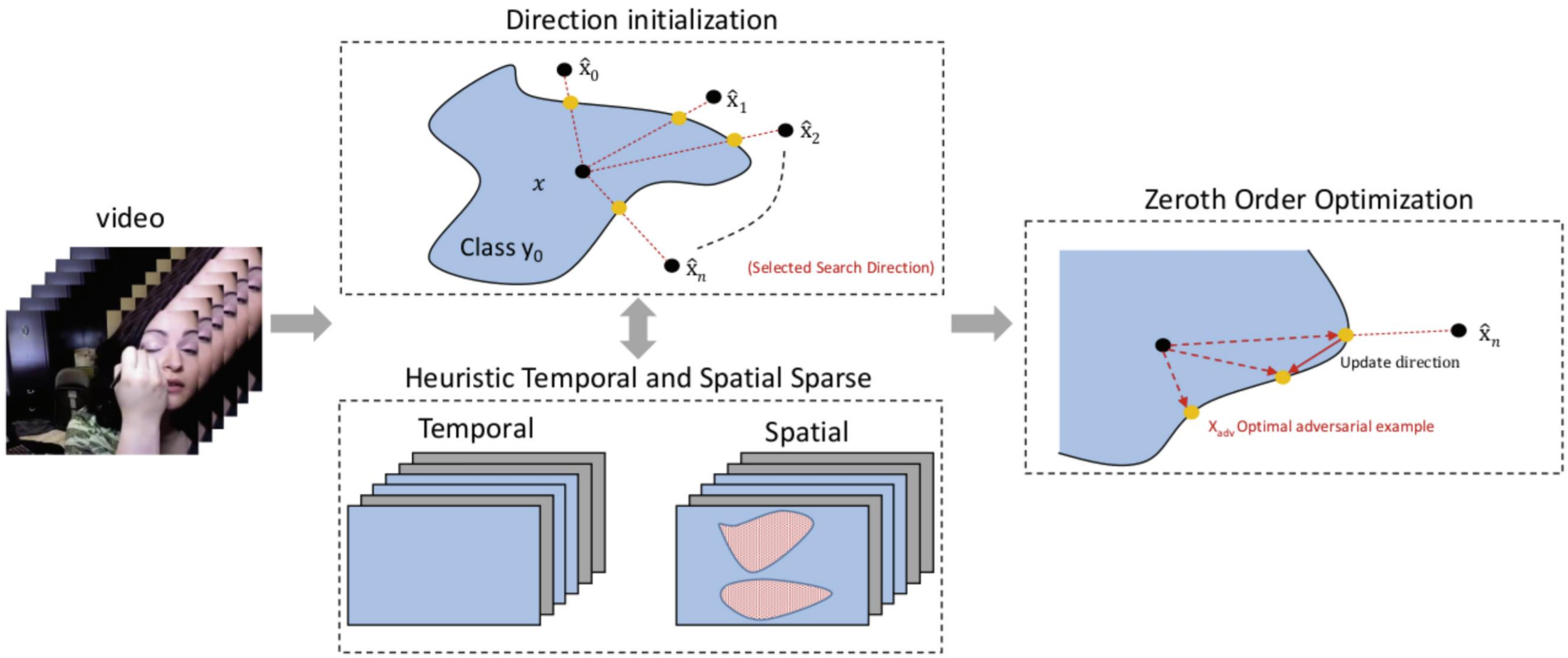
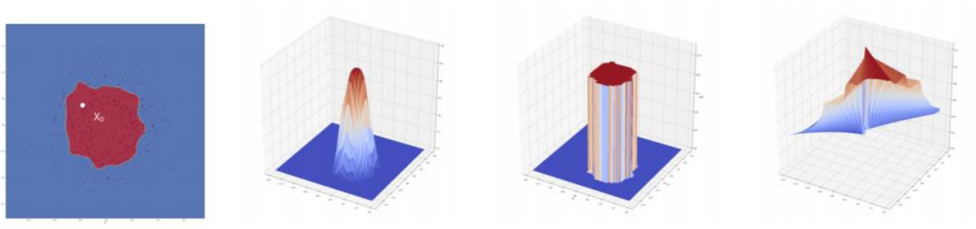
Heuristic Black-Box Adversarial Attacks on Video Recognition Models

Introduction



Methodology

1. Opt-attack



(a) Decision boundary of $f(\boldsymbol{x})$ (b) $\mathscr{L}(Z(\boldsymbol{x}))$ (c) $\mathscr{L}(f(\boldsymbol{x}))$ (d) $g(\boldsymbol{\theta})$

Algorithm 1 Compute $g(\boldsymbol{\theta})$ locally

Input: Hard-label model f, original image x_0, query direction $\boldsymbol{\theta}$, previous value v, increase/decrease ratio $\alpha = 0.01$, stopping tolerance ϵ (maximum tolerance of computed error)
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$
if $f(\boldsymbol{x}_0 + v\boldsymbol{\theta}) = y_0$ **then**
  $v_{left} \leftarrow v, v_{right} \leftarrow (1+\alpha)v$
  while $f(\boldsymbol{x}_0 + v_{right}\boldsymbol{\theta}) = y_0$ **do**
    $v_{right} \leftarrow (1+\alpha)v_{right}$
else
  $v_{right} \leftarrow v, v_{left} \leftarrow (1-\alpha)v$
  while $f(\boldsymbol{x}_0 + v_{left}\boldsymbol{\theta}) \neq y_0$ **do**
    $v_{left} \leftarrow (1-\alpha)v_{left}$
Binary Search within $[v_{left}, v_{right}]$
while $v_{right} - v_{left} > \epsilon$ **do**
  $v_{mid} \leftarrow (v_{right} + v_{left})/2$
  if $f(\boldsymbol{x}_0 + v_{mid}\boldsymbol{\theta}) = y_0$ **then**
    $v_{left} \leftarrow v_{mid}$
  else
    $v_{right} \leftarrow v_{mid}$
return v_{right}

Algorithm 2 RGF for hard-label black-box attack

Input: Hard-label model f, original image x_0, initial $\boldsymbol{\theta}_0$.
for $t = 0, 1, 2, \ldots, T$ **do**
  Randomly choose $\boldsymbol{u}_t$ from a zero-mean Gaussian distribution
  Evaluate $g(\boldsymbol{\theta}_t)$ and $g(\boldsymbol{\theta}_t + \beta\boldsymbol{u})$ using Algorithm 1
  Compute $\hat{\boldsymbol{g}} = \dfrac{g(\boldsymbol{\theta}_t + \beta\boldsymbol{u}) - g(\boldsymbol{\theta}_t)}{\beta} \cdot \boldsymbol{u}$
  Update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\hat{\boldsymbol{g}}$
return $\boldsymbol{x}_0 + g(\boldsymbol{\theta}_T)\boldsymbol{\theta}_T$

Sparse Black-box Video Attack with Reinforcement Learning

1. Targeted attack: firstly replace some key frames with the corelative frames of the target video.

2. Variant of SVA:SVAL

$$L_{percentage} = \|\frac{1}{T}\sum_{t=1}^{T} p_t + S - 1\|, \qquad (11)$$

Methodology

2. Heuristic Temporal Sparsity and Spatial Sparsity

Algorithm 1: Heuristic temporal selection algorithm for the targeted attack.

Input : DNN F, clean video x, true label y, target class y_{adv}, initial mask $M \in \{1\}^{T\times W\times H\times C}$, an empty array A.
Output : Mask of key frames M.
Parameter: Bound ω.

$\hat{x} \leftarrow$ a video sample of target class y_{adv};
$p, k \leftarrow \hat{x} - x, 0$;
for $t \leftarrow 1$ **to** T **do**
  $M_t \leftarrow \mathrm{DELFRAME}(M, t)$; `// the values of i-th frame are equal to 0.`
  $\widehat{y}, P(\widehat{y}|(p \times M_t + x)) \leftarrow F(p \times M_t + x)$;
  if $\widehat{y} = y_{adv}$ **then**
    $A[k], k \leftarrow (t, P(\widehat{y}|(p \times M_t + x))), k + 1$;
end
$A \leftarrow \mathrm{SORTED}(A)$; `// indexes of frames are sorted in descending order by` $P(\widehat{y}|(p \times M_t + x))$.
$\theta_{init} \leftarrow \frac{p}{\|p\|}$ **for** $i \leftarrow 1$ **to** k **do**
  $\widehat{M} \leftarrow \mathrm{DELFRAME}(M, A[i])$;
  $\widehat{p} \leftarrow p \times \widehat{M}$;
  $\theta \leftarrow \frac{\widehat{p}}{\|\widehat{p}\|}$;
  $\widehat{y}, P(\widehat{y}|(x + \widehat{p})) \leftarrow F(x + \widehat{p})$;
  if $\widehat{y} = y_{adv}$ **then**
    if $\mathrm{MAP}(g(\theta) \times \theta) \leq \omega$ **then**
      if $\mathrm{LENS}(\widehat{M}) < \mathrm{LENS}(M)$ **then** `// the number of key frames.`
        $M, \theta_{init} \leftarrow \widehat{M}, \theta$;
    else
      if $\mathrm{MAP}(g(\theta) \times \theta) < \mathrm{MAP}(g(\theta_{init}) \times \theta_{init})$ **then**
        $M, \theta_{init} \leftarrow \widehat{M}, \theta$;
    end
end
return M

Algorithm 2: Heuristic-based targeted attack algorithm.

Input : DNN F, clean video x, true label y, target class y_{adv}, an empty array A
Output : Adversarial example x_{adv}.
Parameter: ω, φ, the number of update iterations I.

$M \leftarrow \mathrm{SPATIAL}(x, \varphi)$;
$M \leftarrow \text{Algorithm } 1(F, x, y, y_{adv}, M, A, \omega)$;
$\theta = \frac{\hat{x} - x}{\|\hat{x} - x\|}$;
$\theta = \frac{\theta \times M}{\|\theta \times M\|}$;
for $t \leftarrow 1$ **to** I **do**
  $\hat{g} = \frac{g(\theta + \beta\mathbf{u}) - g(\theta)}{\beta} \cdot \mathbf{u}$;
  $\theta = \theta - \eta\hat{g}$;
end
$x_{adv} = x + g(\theta) \times \theta$;
return x_{adv}

Experiment

Table 2: Results of our algorithm with various ω in the untargeted attack.

ω	$FR(\%)$	MQ	MAP	MAP^*	$S(\%)$
0	100	16085.0	3.7033	3.8449	17.69
3	100	16085.0	3.6858	3.9667	25.19
6	100	15996.0	3.7471	4.0328	23.94
9	100	17527.0	3.7757	4.2862	34.19
12	100	15912.5	3.8169	4.3646	36.44
15	100	16795.0	3.7274	4.3429	36.69
∞	100	14382.0	3.6039	7.9585	83.75

Table 4: Results of our algorithm with various ω in the targeted attack.

ω	$FR(\%)$	MQ	MAP	MAP^*	$S(\%)$
0	100	302230.50	9.7547	10.5442	8.54
15	100	302230.50	9.7178	10.6463	11.67
30	100	323615.50	8.5328	11.1309	26.88
45	100	307470.00	10.6991	14.8790	35.00
∞	100	209826.00	5.0075	16.6886	71.98

Table 3: Results of our algorithm with various φ in the untargeted attack.

φ	$FR(\%)$	MQ	MAP	MAP^*	$S(\%)$
0.2	90	8770.0	1.5890	8.7153	85.00
0.4	100	12336.0	2.6273	7.0203	68.84
0.6	100	14125.0	3.2194	5.7604	54.25
0.8	100	13845.0	3.4507	4.6347	40.33
1.0	100	16085.0	3.6858	3.9667	25.19

Table 5: Results of our algorithm with various φ in the targeted attack.

φ	$FR(\%)$	MQ	MAP	MAP^*	$S(\%)$
0.2	100	142253.5	11.7693	17.6957	44.17
0.4	100	146720.0	13.4624	18.9002	36.54
0.6	100	175194.5	11.4973	16.2451	34.58
0.8	100	191216.0	10.7961	13.4766	22.58
1.0	100	323615.0	8.5328	11.1309	26.88

Experiment

Table 6: Untargeted and targeted attacks against C3D/LRCN Models. For all attack models, the Fooling Rate (FR) is 100%.

Dataset	Target Model	Attack Model	Untargeted attacks				Targeted attacks			
			MQ	MAP	MAP^*	$S(\%)$	MQ	MAP	MAP^*	$S(\%)$
UCF-101	C3D	Opt-attack (Cheng et al. 2018)	17997.5	4.2540	**4.2540**	0.00	207944.5	9.0906	**9.0906**	0.00
		Our (Temp.)	16292.0	4.0895	4.3642	21.19	313229.0	**7.8069**	10.4700	28.00
		Our (Temp. + Spat.)	**12940.0**	**3.0346**	5.5189	**54.33**	**167217.0**	10.8588	15.4904	**34.28**
	LRCN	Opt-attack (Cheng et al. 2018)	12359.5	**1.8320**	**1.8320**	0.00	445279.0	13.4795	**13.4795**	0.00
		Our (Temp.)	14713.5	1.8754	1.8794	17.19	566719.0	11.7858	14.7894	23.33
		Our (Temp. + Spat.)	**8421.5**	1.8383	3.0848	**47.50**	**399655.0**	**11.2066**	19.8620	**46.92**
HMDB-51	C3D	Opt-attack (Cheng et al. 2018)	14509.5	2.8930	**2.8930**	0.00	205286.5	**6.5704**	**6.5704**	0.00
		Our (Temp.)	13536.5	2.9214	3.2010	26.94	196371.5	8.3599	10.6761	21.88
		Our (Temp. + Spat.)	**10616.0**	**2.3765**	4.4574	**57.04**	**144917.5**	9.6109	12.2993	**28.70**
	LRCN	Opt-attack (Cheng et al. 2018)	18655.0	2.7586	**2.7586**	0.00	224414.0	**3.8598**	**3.8598**	0.00
		Our (Temp.)	15369.5	2.8011	2.8923	24.22	339367.0	4.0618	5.5601	28.75
		Our (Temp. + Spat.)	**13311.5**	**1.5390**	2.8302	**62.03**	**206120.0**	12.7966	18.1835	**42.87**

Conclusion

Advantages:

Fewer query numbers

Disadvantages:

Soft-label black-box attack
The parameter ω has little effect on the result