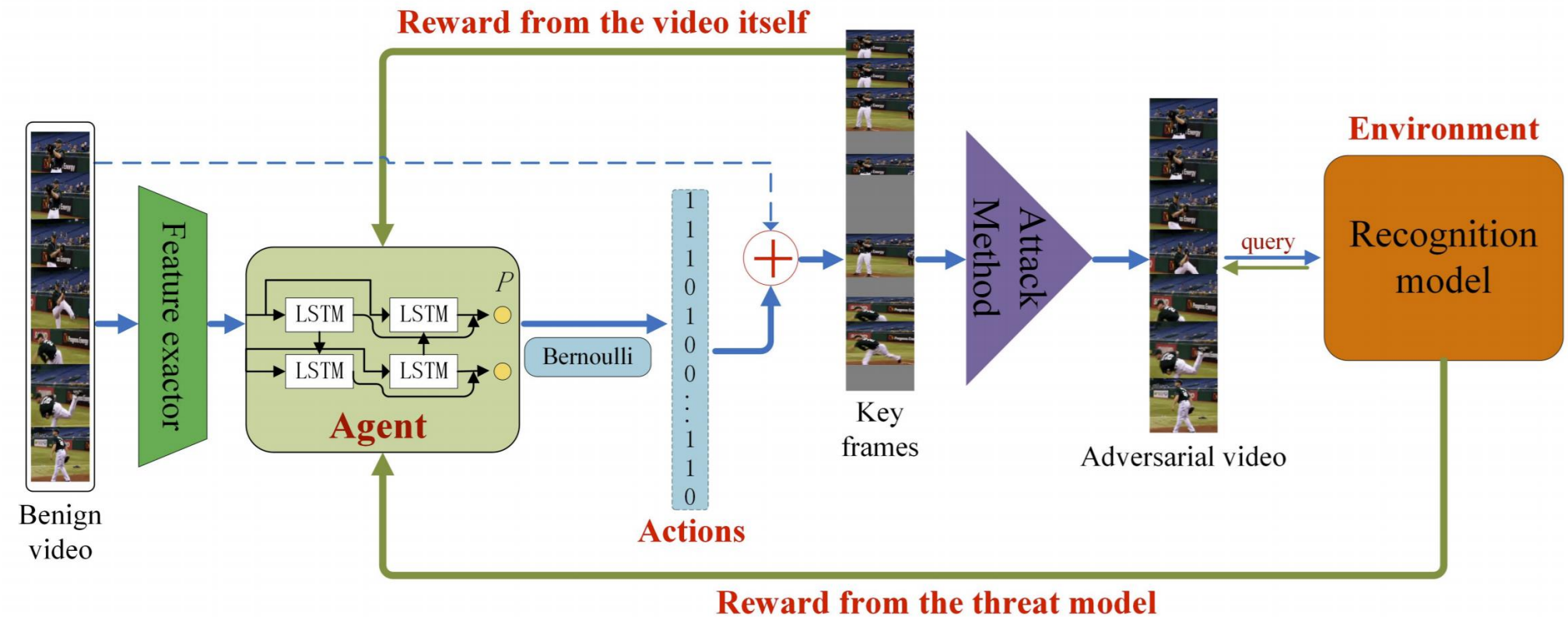
Sparse Black-box Video Attack with Reinforcement Learning

Introduction



Methodology

Key frame selection

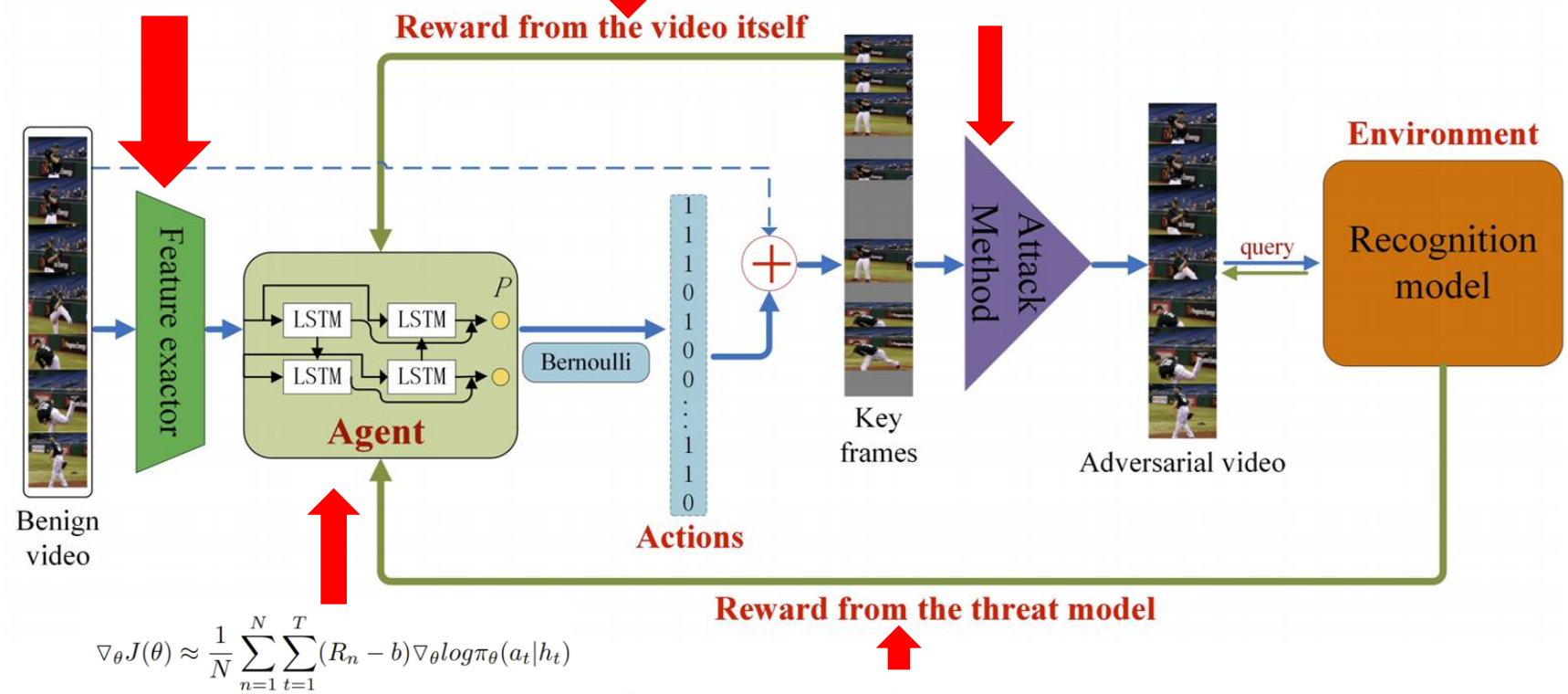
$$R_{rep} = exp(-\frac{1}{T}\sum_{t=1}^{T} min_{t' \in K} \|v_t - v_{t'}\|_2), \quad (5)$$

$$R_{div} = \frac{1}{|K|(|K|-1)} \sum_{t \in K} \sum_{t' \in K, t' \neq t} d(v_t, v_{t'}), \quad (6)$$

FGSM+NES

$$x_{adv} = x + \alpha \cdot sign(\nabla_x l_{adv}(x))$$

$$\nabla E[F(x)] \approx \frac{1}{\sigma n} \sum_{i=1}^{n} \delta_i P(y|x + \sigma \cdot \delta_i)$$



$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} (R_n - b) \nabla_\theta log \pi_\theta(a_t|h_t)$$

$$R_{attack} = \begin{cases} 0.999 \times exp(\frac{-\mathbb{P}}{0.05}) & 30000 > Q > 15000 \\ exp(\frac{-\mathbb{P}}{0.05}) & Q \leqslant 15000 \\ -1 & Q > 30000, \end{cases}$$

Methodology

Algorithm 1: Our SVA targeted attack

Input : The classifier $F(\cdot)$, target class y_{adv} and clean video x.
Output : Adversarial video x_{adv}.
Parameters: Perturbation bound ϵ_{adv}, epsilon decay $\triangle_\epsilon$, *FGSM* step size α.

for $i = 1$ *to epochs* **do**
  $M \leftarrow Agent(x)$, $\epsilon \leftarrow 1$, $x_{adv} \leftarrow$ video of the target y_{adv}.
  $x_{adv} \leftarrow x \times (1 - M) + x_{adv} * M$.
  while $\epsilon > \epsilon_{adv}$ **do**
    $v = 0$, $h = \phi(x_{adv})$.
    $\widehat{v} = v + \nabla_v l_{adv}(x_{adv} + h)$, $\widehat{g} = sign(\widehat{v} \times M)$, $\widehat{\epsilon} \leftarrow \epsilon - \triangle_\epsilon$.
    $\widehat{x}_{adv} \leftarrow CLIP(x_{adv} - \alpha \cdot \widehat{g}, x - \widehat{\epsilon}, x + \widehat{\epsilon})$.
    if $y_{adv} = F(\widehat{x}_{adv})$ **then**
      $x_{adv} \leftarrow \widehat{x}_{adv}$, $\epsilon \leftarrow \widehat{\epsilon}$.
    else
      $\widehat{x}_{adv} \leftarrow CLIP(x_{adv} - \alpha \cdot \widehat{g}, x - \epsilon, x + \epsilon)$.
      if $y_{adv} = F(\widehat{x}_{adv})$ **then**
        $x_{adv} \leftarrow \widehat{x}_{adv}$.
      end
    end
    Adjust $\triangle_\epsilon$ according to the current situation.
  end
  Compute rewards R_{div}, R_{rep} and R_{attack} and update *Agent*.
end
return x_{adv}

Experiment

Table 2. The results of SVAL on C3D with UCF-101 under different sparsity (S).

		S(%)						
	Metrics	10	20	30	40	50	60	70
Un-targeted Attack	MAP	5.5395	5.3805	5.3550	-	**3.2895**	-	-
	FR(%)	100.0	100.00	100.00	80.0	**100.00**	80.0	60.0
Targeted Attack	MAP	8.7538	**6.6218**	-	-	-	-	-
	FR(%)	100.0	**100.0**	60.0	60.0	40.0	20.0	0.0

Table 3. The video attack results of four attack algorithms in the un-targeted mode.

Dataset	Target Model	Attack Model	Metrics & Un-targeted Attack			
			MAP	S(%)	Q	FR(%)
UCF-101	C3D	Opt-attack	4.2540	0.00	15076.23	74.0
		Heuristic-attack	3.2980	22.08	13609.91	79.0
		SVAL(ours)	3.1765	50.00	**8367.78**	83.0
		SVA(ours)	**2.4450**	**63.14**	9402.28	**86.0**
	LRCN	Opt-attack	2.8320	0.00	9032.68	57.0
		Heuristic-attack	2.6940	17.19	9460.38	49.0
		SVAL(ours)	2.4976	60.00	**4131.57**	**68.0**
		SVA(ours)	**2.396**	**62.14**	6132.38	**63.0**
HMDB-51	C3D	Opt-attack	2.8930	0.00	13274.14	76.0
		Heuristic-attack	2.4960	25.68	11870.69	78.0
		SVAL(ours)	2.4482	60.00	**10727.93**	94.0
		SVA(ours)	**2.3940**	**51.37**	24948.67	**98.0**
	LRCN	Opt-attack	2.7586	0.00	18207.11	62.0
		Heuristic-attack	2.6110	27.32	15663.41	66.0
		SVAL(ours)	**1.9479**	**70.00**	**10891.67**	**68.0**
		SVA(ours)	3.1570	62.50	18868.09	64.0

Experiment

Table 4. The video attack results of four attack algorithms in the targeted mode.

Dataset	Target Model	Attack Model	Metrics & Targeted Attack			
			MAP	S(%)	Q	FR(%)
UCF-101	C3D	Opt-attack	-	-	> 60000	-
		Heuristic-attack	-	-	> 60000	-
		SVAL(ours)	6.7672	20.00	43797.0	**38.0**
		SVA(ours)	**3.6450**	**57.24**	**36497.5**	32.0
	LRCN	Opt-attack	-	-	> 60000	-
		Heuristic-attack	-	-	> 60000	-
		SVAL(ours)	5.8834	20.00	**49065.3**	39.0
		SVA(ours)	**3.270**	**56.64**	57850.4	**41.0**
HMDB-51	C3D	Opt-attack	-	-	> 60000	-
		Heuristic-attack	-	-	> 60000	-
		SVAL(ours)	6.9279	30.00	47190.3	**40.0**
		SVA(ours)	**3.8960**	**62.15**	**42900.3**	38.0
	LRCN	Opt-attack	-	-	> 60000	-
		Heuristic-attack	-	-	> 60000	-
		SVAL(ours)	6.2861	20.00	**43880.5**	32.0
		SVA(ours)	**3.5170**	**66.77**	47681.9	**36.0**

Table 5. The ablation study of the proposed method SVA in untargeted setting.

Metrics	Modules			
	No RL	$\text{SVA}_{R_{attack}}$	$\text{SVA}_{R_{attack+rep}}$	SVA
MAP	6.5037	2.3723	2.0321	1.8624
S(%)	0.00	62.35	68.75	74.65

Conclusion

Advantages:

Reduced the query times
Lower Mean Absolute Perturbation(MAP)
Fewer frames perturbated

Disadvantages:

Weak transferability
Several frames are directly replaced for targeted attack, which is easy to percept.