Adversarial Video Captioning

Regular Paper

Suman K. Adari
University of Florida

Washington Garcia
University of Florida

Kevin Butler
University of Florida

***Abstract*—In recent years, developments in the field of computer vision have allowed deep learning-based techniques to surpass human-level performance. However, these advances have also culminated in the advent of adversarial machine learning techniques, capable of launching targeted image captioning attacks that easily fool deep learning models. Although attacks in the image domain are well studied, little work has been done in the video domain. In this paper, we show it is possible to extend prior attacks in the image domain to the video captioning task, without heavily affecting the video's playback quality. We demonstrate our attack against a state-of-the-art video captioning model, by extending a prior image captioning attack known as Show and Fool. To the best of our knowledge, this is the first successful method for targeted attacks against a video captioning model, which is able to inject 'subliminal' perturbations into the video stream, and force the model to output a chosen caption with up to 0.981 cosine similarity, achieving near-perfect similarity to chosen target captions.**

I. Introduction

Remarkable progress has been made in the task of fooling machine learning models. Various methods have been employed by researchers in their attempts to force a misclassification by the model, but the core objective is to inject noise into an adversarial sample such that, to the human eye, the sample appears the same as the original, but is interpreted differently by the model [1]–[3]. Over time, methods have been refined [4], and new attack domains considered [5]. However, there has been no work targeting video captioning models, in which the adversarial video would lead to the generation of a chosen caption.

Regarding targeted image captioning, a recently introduced method known as Show-and-fool [6] has been able to produce perturbations that generate a target caption given either the caption itself or a list of targeted keywords with a high success rate. From a probabilistic perspective, the method aims to minimize the perturbation while maximizing the probability of the chosen caption to appear. This is solved by framing the task as an optimization problem.

In this paper, we extend the concept behind Show-and-Fool to the video domain, with the goal of a targeted video captioning attack. We show that it is feasible to launch an attack against complex video captioning networks, by leveraging recent adversarial machine learning attacks.

- To the best of our knowledge, we are the first instance of a successful targeted video captioning attack on a CNN-GRU encoder-decoder network, which manages to achieve a cosine similarity of up to 0.98 among produced captions.
- We show that attacks in the video domain are more challenging than attacks in the image domain due to the addition of the time dimension.

II. Background

A. Video Captioning

Deep Neural Networks (DNNs) are well known for their high performance on visual classification [7], with some architectures beating human performance [8], [9]. Advances in the image classification task translated well to the video captioning task, with current methods leveraging the deep architecture known as *encoder-decoder* to combine convolutional and recurrent neural networks [10], [11]. State-of-the-art video captioning models exploit a combination of temporal information and multi-caption datasets to produce increasingly detailed descriptions of video scenes [12]–[14].

B. Adversarial Machine Learning

DNNs and human vision differ in how they classify objects, as shown when perturbed images that otherwise look indistinguishable from the original to humans are completely misclassified by the model. Much of recent work has focused on methods to generate adversarial images, spanning evolutionary methods [15] to the more recent optimization-based approaches [2], [4], [16]. Carlini et al. [5] re-formulate the adversarial sample crafting problem by optimizing directly over the model's loss, rather than an abstraction as in previous methods. This is made possible by a re-parameterization of the objective function, as we show later.

C. Adversarial Attacks on Image Captioning

Most recent work has been aimed towards generating samples that force misclassification of an image. That is, they only focus on the image classification task, rather than the more complex image captioning task. Chen et al. [6] introduce a targeted image captioning attack known as Show-and-Fool, which applies a method similar to the Carlini et al. [5] attack, and performs optimization directly over the model's loss function:

$$\min_{\delta} c \cdot \text{loss}(I + \delta) + \|\delta\|_2^2 \tag{1}$$

constrained such that

$$I + \delta \in [-1, 1]^n.$$

For efficiency, Chen et al. transform this constrained optimization problem into an unconstrained problem by using new parameters, $y \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$, where

$$
\begin{aligned}
y &= arctanh(I) \\
w &= arctanh(I+\delta) - y.
\end{aligned}
$$

Thus the constrained optimization in (1) is equivalent to

$$
\begin{aligned}
\min_{w \in R^n} \; & c \cdot \text{loss}(tanh(w+y)) \\
& + \|tanh(w+y) - tanh(y)\|_2^2
\end{aligned} \tag{2}
$$

We follow the notation of Chen et al. and define the target caption as $S = (S_1, S_2, ..., S_t, ..., S_N)$, where S_t is the t-th word in vocabulary $\mathcal{V}$. In their attack, the adversary's goal is to craft an image $I+\delta$ which maximizes the log probability of target caption S, given the set of all possible captions Ω:

$$
\begin{aligned}
logP(S|I+\delta) &= \max_{S' \in \Omega} logP(S^{'}|I+\delta) \\
&= \sum_{t=2}^{N} logP(S_t^{'}|I+\delta, S_1^{'}, \cdots, S_{t-1}^{'})
\end{aligned} \tag{3}
$$

For image captioning, the value of $P(S_t^{'}|I+\delta, S_1^{'}, \cdots, S_{t-1}^{'})$ is computed using a recurrent cell f with hidden state h_{t-1} and input $S_{t-1}^{'}$:

$$
z_t = f(h_{t-1}, S_{t-1}^{'}) \quad and \quad p_t = softmax(z_t) \tag{4}
$$

with z_t, the vector of logits for every word in vocabulary $\mathcal{V}$, taking the form

$$
z_t \equiv [z_t^{(1)}, \cdots, z_t^{(|\mathcal{V}|)}]
$$

The log probability of the target caption is then used directly as a loss function, where the input is the first $N-1$ words of S:

$$
\begin{aligned}
\text{loss}_S(I+\delta) &= -logP(S|I+\delta) \\
&= -\sum_{t=2}^{N} logP(S_t|I+\delta, S_1, ..., S_{t-1})
\end{aligned}
$$

Since the output of the RNN $p_t = \text{softmax}(z_t)$ is a probability distribution on $\mathcal{V}$, the definition of the softmax is used by Chen et al. to simplify the maximum log probability above to

$$
logP(S|I+\delta) \propto \sum_{t=2}^{N} z_t^{(S_t)} = \max_{S' \in \Omega} \sum_{t=2}^{N} z^{(S_t^{'})}. \tag{5}
$$

Applied directly to (1), given some target caption S, this becomes:

$$
\begin{aligned}
\min_{w \in R^n} \; & c \cdot \text{loss}_S(tanh(w+y)) \\
& + \|tanh(w+y) - tanh(y)\|_2^2
\end{aligned} \tag{6}
$$

Hence it suffices to maximize the probability of each word in the target caption. We later show that the result in (6) can be generalized to the temporal domain, and used to attack video captioning models.

D. *Adversarial Attacks on Video*

In the video domain, Hosseini et al. present the only work so far in adversarial video, which fools video classification services by inserting adversarial frames into the video, and return desired keywords [17]. Our approach builds on top of the Show and Fool method to create a targeted attack on the more difficult video captioning task. We successfully attack video captioning models by extending the optimization over L_2 to the temporal domain of videos.

III. Approach

A. *Threat Model*

Our threat model mirrors the white-box threat model described in Chen et al. [6]. That is, the adversary has full knowledge of the model weights, architecture, training algorithm, and training data. The adversary's goal is to force the video captioning model to misclassify an input video sequence V with a target caption S. Transferability of the white-box attack to a black-box scenario is left for future work.

B. *Targeted Caption Optimization*

Similar to the image captioning attack proposed by Chen et al. [6], we rely on the optimization problem of (2). However, we extend the definition of I to a sequence of video frames $\mathbf{V} = [I_0, \cdots, I_n]$. This allows a natural extension of the adversarial image attack to the sequence-to-sequence task of video captioning, where the adversary's goal is to now produce a sequence of perturbations $\Delta = \{\delta_0, \cdots, \delta_n\}$. Keeping the notation of Chen et al., the sequence-to-sequence model estimates the conditional probability of an output sequence $S = (S_1, S_2, \cdots, S_t, \cdots, S_N)$ given the input sequence $\mathbf{V}$. The adversary's goal is to obtain an adversarial tensor $\mathbf{V} + \Delta$ which maximizes the probability

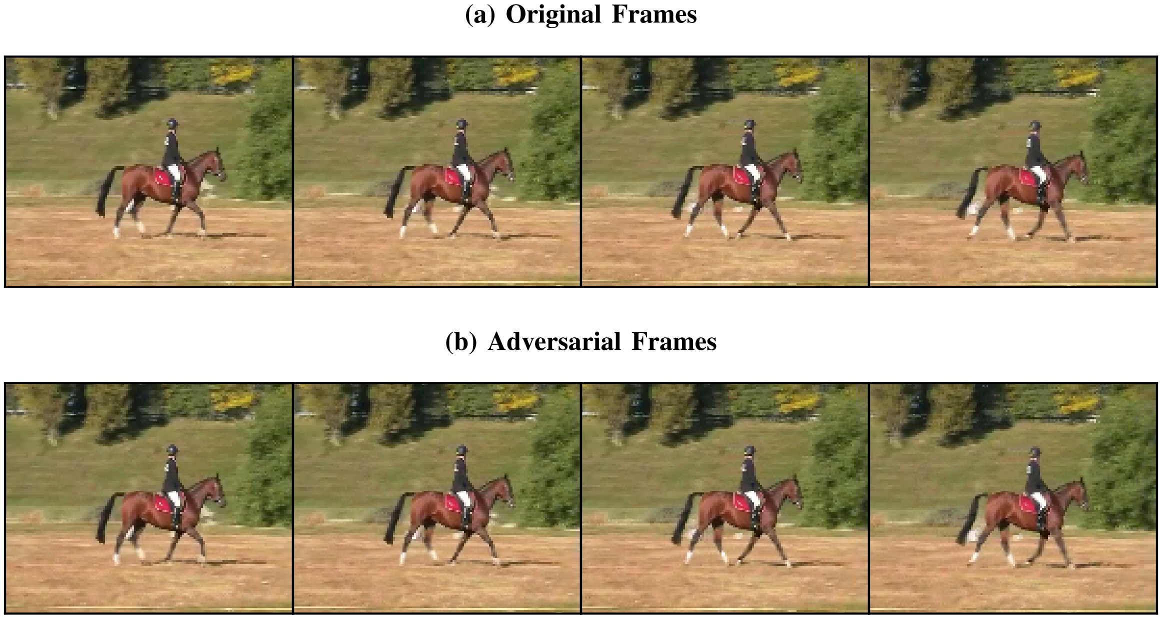
$$
\begin{aligned}
\text{log}P(S|\mathbf{V}+\Delta) &= \max_{S' \in \Omega} \text{log}P(S|\mathbf{V}+\Delta) \\
&= \max_{S' \in \Omega} \text{log}P(S|I_0+\delta_0, \cdots, I_n+\delta_n)
\end{aligned}
$$

As before, the conditional probability is computed by an RNN cell f with hidden state h_{t-1} and input $\mathbf{V}_{t-1}^{'}$. For our particular video captioning network, the output in fact comes from a GRU cell [18], with gating functions omitted for brevity. As before, we use the negative log probability directly as the loss function,

$$
\begin{aligned}
\text{loss}_S(\mathbf{V}+\Delta) &= \text{loss}_S(I_0+\delta_0, \cdots, I_n+\delta_n) \\
&= -logP(S|\mathbf{V}+\Delta) \\
&= -\sum_{t=2}^{N} logP(S_t|\mathbf{V}+\Delta, S_1, ..., S_{t-1})
\end{aligned} \tag{7}
$$

$p(t)$ remains a probability distribution over $\mathcal{V}$, so we have

$$
logP(S|\mathbf{V}+\Delta) \propto \sum_{t=2}^{N} z_t^{(S_t)} = \max_{S' \in \Omega} \sum_{t=2}^{N} z^{(S_t^{'})}. \tag{8}
$$



(a) Original Frames

(b) Adversarial Frames

Fig. 1: Example of an adversarial 4-frame video with $c = 0.01$. The original video (a) is captioned as "man riding on the horse". The chosen target caption is "A women is talking". Our attack manages to successfully fool the video captioning model into captioning the adversarial video (b) as "A women is talking" with nearly indistinguishable perturbations.

Since loss_S now operates over a sequence of frames, we simply adjust (2)

$$\min_{w \in R^n} c \cdot \text{loss}_S(tanh(w_0 + y_0), \cdots, tanh(w_{||\mathbf{V}||} + y_{||\mathbf{V}||}) + \| \frac{1}{||\mathbf{V}||} \sum_{i=1}^{||\mathbf{V}||} tanh(w_i + y_i) - tanh(y_i) \|_2^2 \tag{9}$$

such that

$$\begin{aligned} y_i &= arctanh(I_i) \\ w_i &= arctanh(I_i + \delta_i) - y_i. \end{aligned}$$

Thus we are able to optimize over sequences of frames as a natural extension of (6).

IV. Evaluation

We performed several experiments on the MSVD data set to judge the effectiveness of our attack algorithm. The model used in our experiments is an S2VT model, which in this case is a CNN that feeds into two GRU models, following the encoder-decoder architecture [11]. The network leverages the Attention [12] mechanism to weight frames temporally, and is trained for 1000 epochs on the MSVD dataset, achieving a reasonable Bleu-4 [19] score of 0.38. For the target caption, a caption is chosen randomly from the validation set to be used as part of the targeted attack. A byproduct of this is selection of words that may be outside the model's vocabulary, which appear as $<$ UNK $>$ in final captions.

A. Implementation

Due to memory constraints that result from storing the gradients of a CNN and two GRU networks, we chose to perform the attack on the first four frames of the video. When tested on an NVIDIA GTX 1060 GPU with 6 GB of memory, attacking on more than three frames at once would lead to a CUDA out of memory error. Thus our experiments were performed on an NVIDIA GTX 1080-Ti GPU with 12GB of memory, with the maximum number of iterations set to 1,000. As shown in Figure 1, the transcription of the original four-frame video was still meaningful.

Similar to Show-and-Fool, we used the ADAM optimizer with betas set to 0.9 and 0.999 and a learning rate of 0.001 to minimize our loss functions. We selected a random set of eight videos from the MSVD data set to serve as the target videos for each experiment, and used several different values of c in (9) to measure its effect on the performance of the attack. We plan to release the code for our experiments publicly.

B. Results

Visually, we see that the frames in Figure 1(a) and Figure 1(b) look the same, but the model interprets them differently. The attack achieves similar results to the Show-and-Fool attack, except in the temporal domain.

The values in Table I reference average delta, average cosine similarity, and the average number of iterations. Average $||\delta||_2$ is the mean L_2-norm of perturbations across every frame in the video. We use this as a gauge for how much distortion is added across levels of c. Average cosine similarity is the average of each frame's cosine similarity between the final and target

TABLE I: Results of running the adversarial video captioning attack across the subset of 4-frame videos for differing values of c.

c	Avg. $\|\|\delta_i\|\|_2$	Avg. Cosine Similarity	Average Iterations
0.01	3.737	0.934	544.125
0.1	7.227	0.906	474.625
1	8.971	0.908	485.75
10	6.926	0.951	280.75
100	7.602	0.959	329

captions, with identical captions achieving a cosine similarity of 1.0. The cosine similarity was found by subtracting from one the cosine angle formula of the respective vectorized captions. Average number of iterations is the mean of each video's number of iterations that the attack ran for. If the target caption was not reached within 1,000 iterations, the attack would stop.

$c = 100$ and $c = 10$ produce the most optimal results, followed by $c = 0.01$. Since $c = 10$ had the lowest number of average iterations and highest average cosine similarity, it performed the best due to a high proportion of its attacks converging to produce the target caption. The second most optimal performance was delivered by $c = 100$. In every case, the attack manages to achieve nearly identical captions with scores above 0.908 for cosine similarity.

V. Discussion

The decline in average delta from $c = 10$ onward defies the expected trend for average distortion to increase as c increases. A possible explanation is that, having optimized over the norm of the average of the frames, frames with higher attention weights influence the average delta to be higher as they require greater perturbations. Since c scales the contribution of the cost function, the attack takes less iterations to converge as the value of c increases.

While we chose to use four frames as a result of the heavy memory requirements of the captioning model, we plan to scale the attack up to the entire video to investigate the effects of different values of c on the full video sequence in our future work as well as incorporate BLEU scores to evaluate attack performance. One option is to use a window to loop through the video frames and generate adversarial counterparts that would ideally produce the target caption for the whole video when concatenated. Another path is to perturb select frames with the highest attention weights due to their increased influence in determining the produced caption.

VI. Conclusion

In this paper, we extended the targeted image captioning attack technique known as Show-and-Fool to the video domain. The experiments show that for cases with logical target captions, the algorithm produces adversarial perturbations that successfully produce the target caption while remaining nearly imperceptible to human eyes. Although our tested video captioning models are deeper due to the temporal aspect of videos, they are still susceptible to the concepts behind targeted image captioning attacks.

References

[1] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial Classification," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy (Euro S&P)*, pp. 372–387, March 2016.

[4] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIACCS)*, 2017.

[5] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57, 2017.

[6] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2587–2597, Association for Computational Linguistics, 2018.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[10] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *CoRR*, vol. abs/1412.4729, 2014.

[11] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," in *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[13] R. Pasunuru and M. Bansal, "Multi-task video captioning with video and entailment generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1273–1283, 2017.

[14] O. Nina, W. Garcia, S. Clouse, and A. Yilmaz, "MTLE: A multitask learning encoder of visual feature representations for video and movie description," *CoRR*, vol. abs/1809.07257, 2018.

[15] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, June 2015.

[16] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

[17] H. Hosseini, B. Xiao, and R. Poovendran, "Deceiving Google's Cloud Video Intelligence API Built for Summarizing Videos," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2017-July, pp. 1305–1309, 2017.

[18] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2001.