Supplementary Material for: Exact Adversarial Attack to Image Captioning via Structured Output Learning with Latent Variables

Yan Xu[‡†*], Baoyuan Wu[†♯*], Fumin Shen[‡], Yanbo Fan[†], Yong Zhang[†], Heng Tao Shen[‡], Wei Liu[†♯]

[†]Tencent AI Lab, [‡]University of Electronic Science and Technology of China

{xuyan5533,wubaoyuan1987,fumin.shen,fanyanbo0124,zhangyong201303}@gmail.com,

shenhengtao@hotmail.com, wl2223@columbia.edu

The main contents in this manuscript are outlined as follows:

- The attack results on the Show-and-Tell model with different CNN architectures are presented in Section 1.
- The results of transfer attacks among three image captioning models are shown in Section 2.
- We evaluate the influence of hyper-parameters to latent SSVMs in Section 3.
- We present some qualitative results of our methods, and analysis about the relationship between the noise map and the attention map, in Section 4.
- We analyze some failed attacks in Section 5.

1. Attack Results on the Show-and-Tell Model with Different CNN Architectures

In this section, we present the attack results on the Show-and-Tell [4] model, of which the CNN part is specified as Inception-v3 [3] and ResNet-101 [1], respectively. **(1)** As shown in Table 1, the attack performance on Show-and-Tell with Inception-v3 is better than that on Show-and-Tell with ResNet-101 at most cases, with smaller $\|\boldsymbol{\epsilon}\|_2$ and higher SR, Precision and Recall. It demonstrates that the CNN part could significantly influence the attack performance of the CNN+RNN based image captioning system. **(2)** The attack performance on Show-and-Tell with ResNet-101 is much worse than the attack performance on Show-Attend-and-Tell [5] and SCST [2] (see Section 5.2 in the main manuscript), where the CNN parts are also ResNet-101. It demonstrates that the RNN architecture also significantly influences the attack performance. As analyzed from Line 633 to 643 in the main manuscript, in Show-and-Tell, the visual features extracted by CNN are only fed into the starting step of RNNs, while they are fed into RNNs at every step in Show-Attend-and-Tell and SCST. Consequently, the gradients of observed words in targeted partial captions can be directly back-propagated to the input image in Show-Attend-and-Tell and SCST. In contrast, the gradients of both observed words and latent words are firstly multiplied, and then are back-propagated to the input image. Obviously, it is more difficult to enforce the Show-and-Tell model to produce the targeted words/captions, than Show-Attend-and-Tell and SCST.

Show-and-Tell model with Inception-v3								
method	metric	0 latent	1 latent	2 latent	3 latent	1 obser	2 obser	3 obser
GEM	$\|\boldsymbol{\epsilon}\|_2 \downarrow$	4.5959	3.4488	3.3999	3.3783	2.2588	2.5779	2.7472
	SR ↑	0.4404	0.5034	0.4094	0.3408	0.4606	0.4248	0.4962
	Prec ↑	0.6758	0.7475	0.691	0.6455	0.4606	0.5468	0.6403
	Rec ↑	0.6635	0.7344	0.6763	0.626	0.4606	0.5468	0.6403
Latent SSVMs	$\|\boldsymbol{\epsilon}\|_2 \downarrow$	1.7635	4.5913	4.6584	4.7369	4.5513	4.8617	4.933
	SR ↑	0.4924	0.5808	0.4634	0.3978	0.287	0.2118	0.227
	Prec ↑	0.7438	0.7982	0.7257	0.6697	0.287	0.3609	0.4065
	Rec ↑	0.7318	0.7862	0.7122	0.6545	0.287	0.3459	0.3898
Show-and-Tell model with ResNet-101								
method	metric	0 latent	1 latent	2 latent	3 latent	1 obser	2 obser	3 obser
GEM	$\|\boldsymbol{\epsilon}\|_2 \downarrow$	4.9734	4.8364	4.6889	4.525	2.725	3.2965	3.6852
	SR ↑	0.3428	0.2974	0.2471	0.1962	0.3818	0.3092	0.3532
	Prec ↑	0.6072	0.5762	0.5507	0.5209	0.3818	0.4305	0.5142
	Rec ↑	0.594	0.561	0.5339	0.5014	0.3818	0.4305	0.5142
Latent SSVMs	$\|\boldsymbol{\epsilon}\|_2 \downarrow$	4.9449	4.9662	5.0024	4.997	4.0661	4.4297	4.5829
	SR ↑	0.3658	0.2956	0.247	0.1914	0.3084	0.248	0.2842
	Prec ↑	0.6458	0.5867	0.55	0.5074	0.3084	0.3742	0.4545
	Rec ↑	0.6322	0.5721	0.5348	0.4909	0.3084	0.3742	0.4545

Table 1. Results of adversarial attacks to the Show-and-Tell model, with different CNN architectures.

2. Results of Transfer Attacks among Three Image Captioning Systems

Here we present transfer attacks among different image captioning systems (*i.e.*, SAT, SCST and ST), at the case of

*indicates equal contributions. ♯indicates corresponding authors. This work was done when Yan Xu was an intern at Tencent AI Lab.

λ	0.01				0.1				1			
ζ	0.1	0.5	5	10	0.1	0.5	5	10	0.1	0.5	5	10
$\|\epsilon\|_2 \downarrow$	6.2199	7.4208	10.134	10.1338	4.1528	4.5344	5.1702	5.174	2.116	2.2647	3.2191	3.2207
SR $\uparrow$	0.9728	0.987	0.8932	0.8964	0.9348	0.9822	0.9656	0.9558	0.7708	0.9156	0.8078	0.8026
Prec $\uparrow$	0.9855	0.9921	0.9476	0.9489	0.9632	0.9893	0.9817	0.9808	0.86	0.9508	0.8941	0.8909
Rec $\uparrow$	0.9851	0.9919	0.9458	0.9473	0.9625	0.9891	0.9811	0.98	0.8572	0.9497	0.8909	0.8877

Table 2. Adversarial attacks of targeted complete captions to the Show-Attend-and-Tell model, using latent SSVMs with different hyper-parameters λ and ζ.

Model A $\rightarrow$ Model B	GEM				SSVMs			
	$\|\epsilon\|_2$	SR	Prec	Rec	$\|\epsilon\|_2$	SR	Prec	Rec
SAT $\rightarrow$ ST	4.28	0.002	0.1818	0.1711	5.14	0.002	0.1829	0.1721
SAT $\rightarrow$ SCST	4.28	0.0234	0.2952	0.2883	5.17	0.0103	0.261	0.2514
SCST $\rightarrow$ ST	5.03	0.002	0.1824	0.1716	4.71	0.002	0.1834	0.1724
SCST $\rightarrow$ SAT	5.20	0.0013	0.2105	0.1985	4.70	0.0023	0.2048	0.1939
ST $\rightarrow$ SAT	3.61	0.0003	0.1776	0.1662	3.79	0	0.1778	0.1663
ST $\rightarrow$ SCST	3.6	0.0007	0.1851	0.1742	3.8	0.001	0.1856	0.1747

Table 3. Results of transfer attacks among SAT, SCST and ST models.

adversarial attacks of targeted complete captions. Specifically, we firstly generate one perturbed image to produce a targeted complete caption based on one captioning system. Then, we feed this perturbed image into another captioning system, to check whether the same targeted caption can be predicted. The results are summarized in Table 3. The low values of SR, Prec and Rec demonstrate the poor transferability of targeted adversarial noises among different image captioning systems. Actually, even for one benign image, different image captioning systems are likely to produce different captions. It reveals that the distributions of predicted captions of different captioning systems are significantly different in the structured-output space. Thus, it is not surprising to produce different captions by different image captioning systems for one perturbed image. In future, we plan to explore more details about the commonalities and differences of the caption distributions between different image captioning systems, using the proposed exact adversarial attack methods. It is expected to provide more insights to understand the inner mechanisms of image captioning systems.

3. Attack Performance of Latent SSVMs with Different Hyper-Parameters

As demonstrated in Line 620 to 622 in the main manuscript, the performance of latent SSVMs may be influenced by two hyper-parameters, *i.e.*, λ and ζ (see Eqs. (11) and (12) in the main manuscript). In this section, we present brief experimental analysis based on attacks of targeted complete captions to the Show-Attend-and-Tell model, using latent SSVMs. As shown in Table 2, smaller λ generally leads to higher $\|\epsilon\|_2$, and higher SR, Precision and Recall, and vice versa. λ controls the balance between the noises and attack performance. Given a fixed λ, the best attack performance is obtained at $\zeta = 0.5$, while there is no significant difference between $\zeta = 5$ and $\zeta = 10$. It demonstrates that ζ in a suitable range also has a significant influence to the performance of latent SSVMs. This relationship is not linear, we can see that the performance at $\zeta = 0.5$ surpasses other situations. However, given a fixed ζ (*e.g.*, $\zeta = 5$), the best performance is obtained at $\lambda = 0.1$, rather than $\lambda = 0.01$ or $\lambda = 1$. It tells that these two hyper-parameters are strongly coupled with each other.

4. Qualitative Results of Adversarial Attacks

In this section, we present some qualitative results of adversarial attacks on three state-of-the-art CNN+RNN image captioning models, using the proposed two methods, as shown in Fig. 1. All targeted partial/complete captions are successfully attacked, while the adversarial noises are invisible to human perception.

An interesting observation is that the adversarial noises of Show-Attend-and-Tell [5] (see row 1 to 4) and SCST [2] (see row 5 to 8) distribute at some particular areas of the image, while the noises of Show-and-Tell [4] with Inception-v3 (see row 9 to 12) cover the whole image. There are two important differences between (Show-Attend-and-Tell, SCST) and Show-and-Tell. First, the visual features extracted by the CNN part are fed into the RNN part at each step in Show-Attend-and-Tell and SCST, while those in Show-and-Tell are fed into the RNN part only at the starting step. Second, there is an attention module in Show-Attend-and-Tell and SCST to control the input visual features at each step. As analyzed in Section 1, the first difference is the main reason that the attack performance on Show-and-Tell is much worse than that of the other two models. The second difference may be the main reason to control the noisy areas. To verify this point, we analyze two cases in Fig. 1, where the noisy areas are clearly distinguished, including the cases of the second column and the 4th, 5th rows. As shown in Fig. 2, we present the adversarial image, adversarial noises, and the attention maps at each step. It can be found that most attended points at each step only occur at noisy areas. It is easy to understand this observation, as the noises are positively proportional to the back-propagated gradients, and the gradient back-propagated to the image at each step is positively proportional to the corresponding attention map. Thus, the

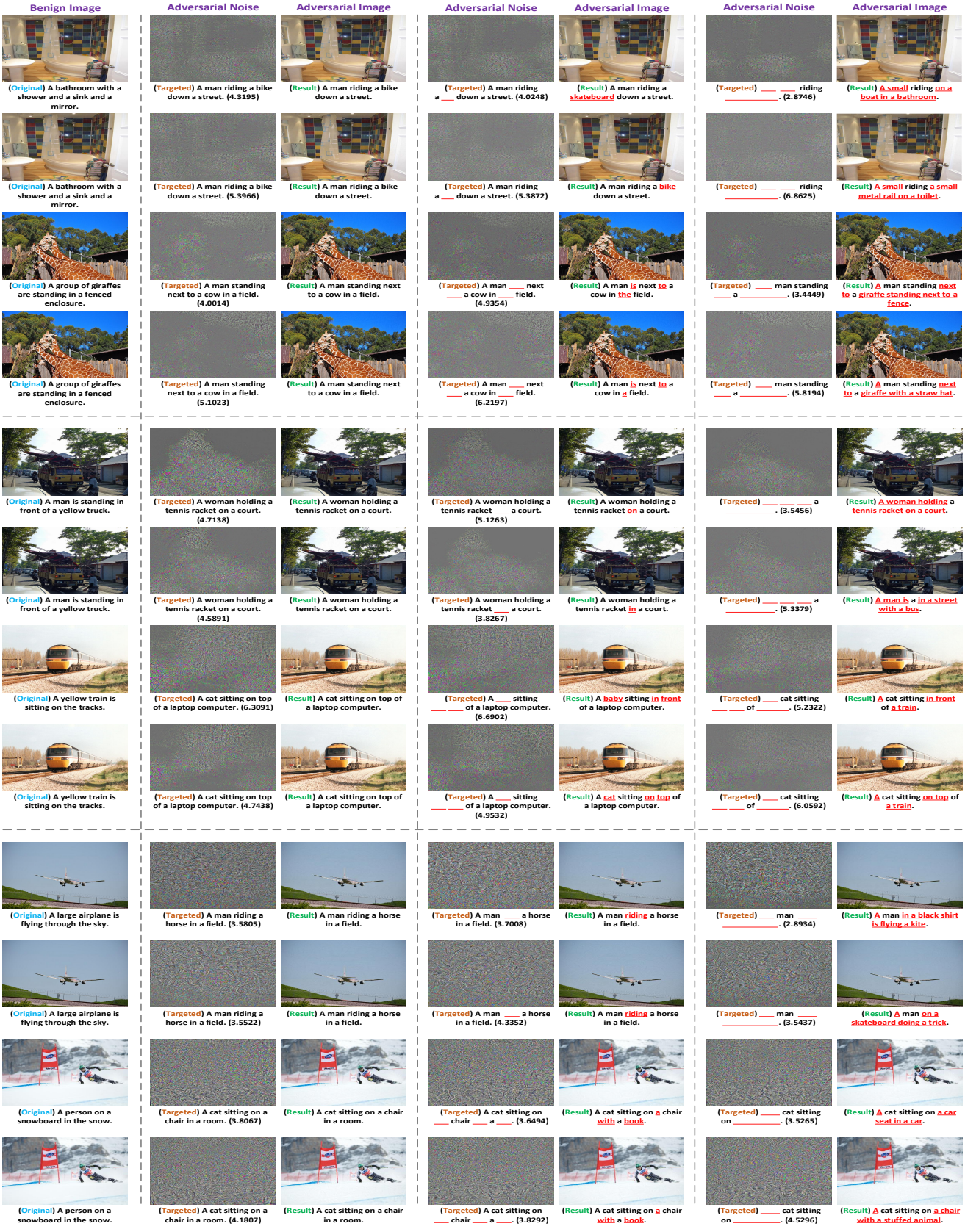

Figure 1. Some qualitative examples of adversarial attacks on Show-Attend-and-Tell [5] (**row 1 to 4**), SCST [2] (**row 5 to 8**), and Show-and-Tell [4] with Inception-v3 (**row 9 to 12**), respectively. **Odd** rows are attack results using the proposed GEM method, while **even** rows are attack results using the proposed latent SSVM method. **(a)** Attacks of targeted complete captions; **(b)** attacks of targeted partial captions with some latent words; **(c)** attacks of targeted partial captions with some observed words.

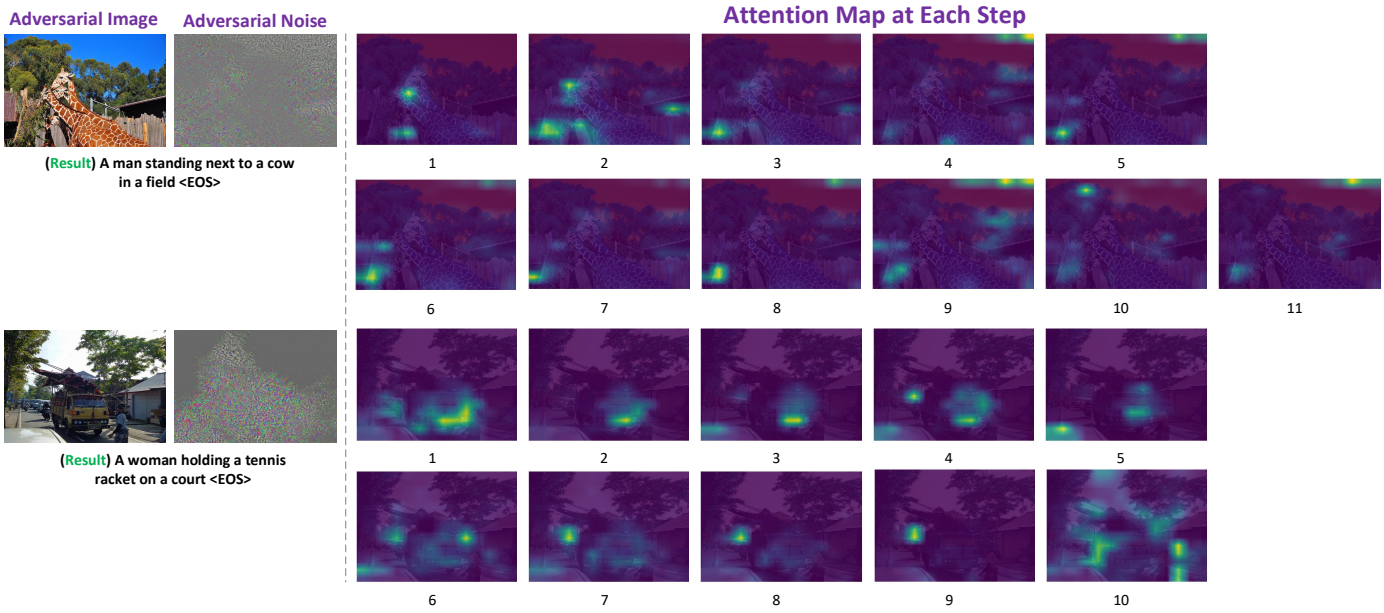

Figure 2. Two examples of adversarial noises and the corresponding attention map at each step.

noises are generally positively proportional to the attention maps. However, as the gradient at one step will also be influenced by the attention maps of other steps, through the cell state in LSTM, it is nontrivial to formulate the relationship between the noise map and the attention maps using a simple equation (*e.g.*, linear). This relationship will be rigorously studied in our future research.

5. Examples of Failed Attacks

In this section, we present some examples of failed attacks, based on the Show-Attend-and-Tell model and the latent SSVM method. Four typical types of failed attacks are shown in Fig. 3. In the **first** row, the adverbial *in the background* is missed in the result, compared to the targeted caption. In the **second** row, there is an additional adverbial *and a cord* in the result, compared to the targeted caption. These two examples could be explained by the fact that the collocation between one adverbial and other components is very flexible in natural languages. In the **third** row, the end words *game of soccer* in the targeted caption are not successfully attacked. We can see that the second latent word is predicted as *soccer*, as *playing soccer* is a frequent collocation. Consequently, as the collocation playing soccer game of a soccer rarely occurs in the training captions, the attacks of the final three words are failed. In the **last** row, the second observed word *in* is not successfully attacked. The reason is that the first observed word *is* is very flexible to collocate with other words in natural languages, so its influence on the predicted word at the next latent location is very weak. However, the predicted word *holding* has a strong influence to its next word, *i.e.*, *in*. Moreover, after the

Figure 3. Some examples of failed adversarial attacks on the Show-Attend-and-Tell model with the latent SSVM method. Note that the missed predictions are highlighted in **dark blue**; the extra predictions are emphasized in **brown**; the incorrect predictions are highlighted in **pink**.

observed *in*, there is no more observed word to provide constraints to *in*. Consequently, the observed word *in* is failed due to the strong influence from its previous prediction *hold-*

ing. **In summary**, the above failed examples reveal that the adversarial attacks of targeted captions could be influenced by many factors, such as frequent collocation, the number of observed words, and the locations of observed words. These observations will be helpful to designing better attack methods to image captioning in our future research.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[2] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 1

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 1, 2, 3

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015. 1, 2, 3