FOOLED BY IMAGINATION: ADVERSARIAL ATTACK TO IMAGE CAPTIONING VIA PERTURBATION IN COMPLEX DOMAIN

Shaofeng Zhang[1*], *Zheng Wang*[12†*], *Xing Xu*[1], *Xiang Guan*[1], *Yang Yang*[1]

[1]University of Electronic Science and Technology of China, Chengdu, 611731, China
[2]Institute of Electronic and Information Engineering of UESTC, Guangdong, 523808, China
ssfzhang@foxmail.com, zh_wang@hotmail.com,xing.xu@uestc.edu.cn, {duochuan.gx, dlyyang}@gmail.com

ABSTRACT

Adversarial attacks are very successful on image classification, but there are few researches on vision-language systems, such as image captioning. In this paper, we study the robustness of a CNN+RNN based image captioning system being subjected to adversarial noises in complex domain. In particular, we propose **Fooled-by-Imagination**, a novel algorithm for crafting adversarial examples with semantic embedding of targeted caption as perturbation in complex domain. The proposed algorithm explores the great merit of complex values in introducing imaginary part for modeling adversarial perturbation, and maintains the similarity of the image in real part. Our approach provides two evaluation approaches, which check whether neural image captioning systems can be fooled to output some randomly chosen captions or keywords. Besides, our method has good transferability under black-box setting. At last, our extensive experiments show that our algorithm can successfully craft visually-similar adversarial examples with randomly targeted captions or keywords at a higher success rate.

Index Terms— Image Captioning, Adversarial Attack, Complex Domain

1. INTRODUCTION

Deep learning has achieved great success in various application scenarios, ranging from the pure vision tasks, such as object detection, to the comprehensive visual-language tasks, for example video understanding [1],cross-media retrieval [2], visual question answer [3], action recognition [4]. The state-of-the-art of those multiple tasks have been remarkably promoted. However, recent studies discover that deep neural networks (DNNs) are vulnerability to adversarial examples, which have undergone small, carefully crafted perturbations, and can easily fool a DNN model into making irrelevant classification. Thus, this may severely hamper the application of deep learning techniques to safety-critical fields, such as auto-driving and face recognition. Albeit numerous algorithms and models have been proposed to defense the adversarial attack [5], almost all of them are later shown to be broken.

Recently, researchers have shown an increasing interest in whether adversarial examples are practical enough to attack more complex systems, for instance image retrieval [6] and video recognition [7]. Although all these adversarial examples display their excellent attack performance on their corresponding tasks, they are all limited to a single modal data. Moreover the security for cross-media intelligent technology is also particularly important. Image captioning task is the simplest and most typical task for vision-language in cross-media processing systems. So in this work, we extend the investigation towards image captioning models, that not only include a vision component but also a language part to deepen our understanding of the practicality of adversarial examples, and at the same time, explore the robustness of the CNN+RNN based architecture for image captioning system. Note that there are a few studies also focusing on image captioning. Xu et al. [8] propose to fool an image captioning system to generate some targeted partial captions for an image polluted by adversarial noises, even the targeted captions are totally irrelevant to the image content.

Note that crafting adversarial examples in image captioning tasks is strictly harder than in well-studied image classification tasks, due to the main reasons [9]: (i) class attack v.s. caption attack and (ii) CNN v.s. CNN+RNN. In this paper, we address the aforementioned challenges by proposing a novel algorithm termed as Fooled-by-Imagination. The proposed algorithm explores the great merit of complex values in introducing imaginary part for modeling adversarial perturbation with the semantic embedding of targeted sentence, but maintains the similarity of the image in real part. The proposed Fooled-by-Imagination algorithm can craft adversarial examples in neural image captioning adaptive to different scenar-

This work was supported in part by the National Natural Science Foundation of China under Project 61976049, the Fundamental Research Funds for the Central Universities under Project ZYGX2019Z015, and Dongguan Songshan Lake Introduction Program of Leading Innovative and Entrepreneurial Talents.

†Corresponding Author.

*Zheng Wang and Shaofeng Zhang contributed equally to this work.

978-1-7281-1331-9/20/$31.00

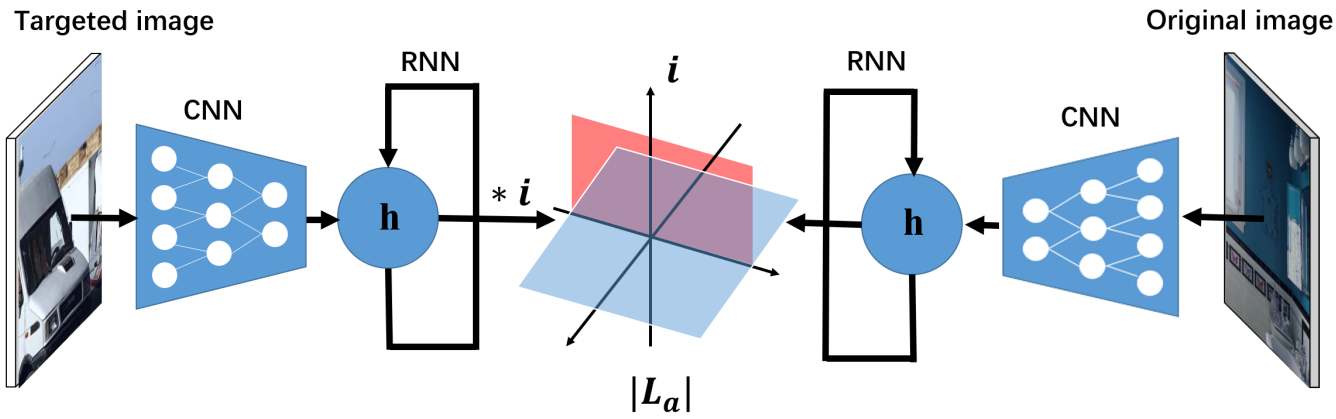

Fig. 1. Overview of Fooled-by-Imagination. The blue plane is the real field, and the red plane for the imaginary field.

ios, that are targeted sentence attack and targeted keywords attack [9]. Regardless of whatever attack, their semantic embedding are acquired from the pre-tained attacked model. The only difference is that keywords attack should choose the location of keywords and then place their semantic embedding to the corresponding places.

The major contributions of this work are three-fold:

- Within our best knowledge, we are the first to craft adversarial examples for image captioning systems in complex domain. Due to the higher dimension, the imperceptibility of perturbation by human eyes maintains well.
- The crafted perturbation is based on semantic embedding of the targeted caption, which could adapt to different attack settings, including caption and keywords attack.
- The experimental results on MSCOCO demonstrate that the success rate of our Fooled-by-Imagination to attack image captioning system based on CNN+RNN, is superior to state-of-the-art methods, whether under white-box or black-box.

2. RELATED WORK

Image Captioning. Most recent image captioning approaches based on deep learning adopt an encoder-decoder architecture that first uses a CNN model as visual feature extractor, followed by a RNN model as the decoder for generating caption of the image [10, 11, 12, 13]. Representative works under this framework are mainly differed by the underlying CNN and RNN architectures, and whether or not the attention mechanisms are considered. Other lines of research generate image captions using semantic information or via a compositional approach.

Adversarial Examples. Existing works on adversarial example generation mainly focus on image classification models. Several different approaches have been proposed for generating adversarial examples, including FGSM(Fast Gradient Sign Method) [14] and IFGSM (Iterative Fast Gradient Sign Method) [15], optimization-based methods [16], and others. In particular, Carlini et al. [16] propose the state-of-the-art attacks under constraints on L_0, L_2, and L norms. Another line of research studies adversarial examples against deep neural networks for vision-language tasks, such as image captioning and visual question answering [9, 17, 8]. Show-and-Fool [9] is the first algorithm for crafting adversarial examples in neural image captioning. The proposed algorithm provides two evaluation approaches, which check whether neural image captioning systems can be mislead to output some randomly chosen captions or keywords.

Complex-value related Study. Recent years, more researchers turn their attention to study complex-valued related work for its richer representational capacity, and apply the algorithm corresponding to complex-value to various areas of deep learning. Reichert et al. [18] indicate that complex-valued neural networks has been investigated long before the earliest deep learning breakthroughs. Trabelsi et al. [19] propose some complex-valued algorithms for batch-normalization, weight initialization strategies, and explore complex convolutional network architectures for image classification and speech spectrum prediction. Wolter et al. [20] firstly introduce complex to gated recurrent network, and achieve the state-of-art performance on the adding problem. To our best knowledge, our work is the first to study adversarial examples against vision-language models with perturbation in complex domain.

3. THE PROPOSED APPROACH

Firstly, we formally review the approach to crafting adversarial examples for image classification. The problem of finding an adversarial example for a given image I can be cast as the following optimization problem:

$$\begin{array}{ll} \min\limits_{\delta} & c \cdot loss(I+\delta) + \|\delta\|_2^2 \\ \text{s.t.} & I+\delta \in [-1,1]^n. \end{array} \tag{1}$$

Here δ denotes the adversarial perturbation to I. $\|\delta\|_2^2 = \|(I+\delta)-I\|_2^2$ is an l_2 distance metric between the benign image and the adversarial image. loss$(\cdot)$ is an attack loss function which takes different forms in different attacking settings. The term $c > 0$ is a pre-specified regularization constant. Intuitively, with larger c, the attack is more likely to succeed but at the price of higher distortion on δ, namely the adversarial example could be perceived easily by human eyes.

Then, we expound our simply but effectively and originally method to attack image captioning system based on complex field. Carlini and Wagner [16] have reported that in CNN-based image classification, using logits in the attack loss function could produce better adversarial examples, which is also adpoted by Show-and-Fool [9]. The objective function of attack on image caption using logits, can express as follow:

$$\mathcal{L} = H(f(I_o), t)), \tag{2}$$

where f is the image caption model, such as Show-and-Tell [10] and so on, I_o is the original image in data set, and t is the target label.

Different from the traditional attack strategy, our main idea is based on semantic embedding layer. Specifically, our goal is to add some invisible perturbation with caption semantic embedding of the targeted image, so that the attacked image captioning model could generate wrong caption and misconceive the visual content. Hence, the loss function in real number field can be expressed as follows:

$$\mathcal{L} = H(E(I_o), E(I_t))), \tag{3}$$

where E is the semantic embedding function, in our algorithm, we adopt the semantic embedding layer in the image caption model, and I_t is the targeted image.

Compared to the real number field $\mathbb{R}$, the complex number field $\mathbb{C}$ is a two-dimensional number field that provides a higher dimensional, more abstract perspective. Inspired by this thought, we extend our loss function to complex set, where the perturbation could be more invisible but better attack effect than the classical adversarial attack, only carrying perturbation in real domain. Therefore, the new novel loss function could be expressed as:

$$\mathcal{L}_a = H(E(I_o), E(I_t) * i)) \tag{4}$$

where i represents imaginary part, and it is the main reason that we term our approach as **Fooled-by-Imagination**.

As aforementioned, we consider l_2 distance to measure the similarity between benign image and adversarial image. Thus the objective function of the optimization is shown in Equation 5. Here, $E^2(I_o)$ and $E^2(I_t)$ in the expansion equation are in real domain for the loss, and $E(I_o) * E(I_t) * i$ are the imaginary part.

$$\begin{aligned} \mathcal{L}_a &= (E(I_t) * i - E(I_o))^2 \\ &= E^2(I_o) - E^2(I_t) - 2 * E(I_o) * E(I_t) * i. \end{aligned} \tag{5}$$

At the same distance metric of l_2, compared to the traditional method, we can conclude that our loss function lowers with two targeted semantic embedding and add the perturbation embedding in complex domain. It assure that our adversarial example are more invisible and better effect.

Next, we compute the module length of our loss function like Equation 6. This function is the key part to generate good targeted adversarial examples.

$$\begin{aligned} |\mathcal{L}_a| &= \sqrt{(E^2(I_o) - E^2(I_t))^2 + 4 * E^2(I_o) * E^2(I_t)} \\ &= E^2(I_o) + E^2(I_t). \end{aligned} \tag{6}$$

To further control the distance of original images and generated adversarial examples, we add a distance loss on L_2 norm:

$$\mathcal{L}_b = \|I_a - I_o\|_2, \tag{7}$$

where the I_a is the generated adversarial examples. In the end, the whole objective function of our simple but effective approach is the weighted sum of the two loss.

$$\mathcal{L} = |\mathcal{L}_a| + \alpha * \mathcal{L}_b, \tag{8}$$

where α is the parameter to balance the two loss. In order to further understand our proposed method, we detail our algorithm process in Algorithm 1.

Algorithm 1 Fooled by Imagination: Adversarial Attack to Image Captioning via Perturbation in Complex Domain.

Input: f_{target}: trained image caption model; I_o, y: original data; α: control parameter
Output: I_a: adversarial examples
Init: $E \leftarrow f^E_{target}$; $I_a \leftarrow I_o$
  while $epoch < maxiter$ **do**
    $|L_a| \leftarrow E^2(I_o) + E^2(I_t)$
    $L_b \leftarrow \|I_a - I_o\|_2$
    $L \leftarrow |L_a| + \alpha * L_b$
    $\nabla I_a \leftarrow \partial L(I_a, I_o, f_{target})/\partial I_a$
    $I_a \leftarrow I_a - \nabla I_a$
  end while
  return I_a

Different from Show-and-Fool [9], we don't need to design our loss functions for different attack settings, such as targeted caption and targeted keyword attack. Based on the weight of the semantic embedding layer and the linear layer of vocab, the semantic embedding vector of a single word can

be calculated after transpose. The maximum length of the caption is set to 20. Therefore, the image caption model outputs a vector with dimension of $(20, 512)$ for each image, and 512 is the semantic embedding dimension. Take 1-keyword attack as an example, suppose we want the target keyword to be in the first three places of the caption sentence, thus we replace the first three dimensions of the $(20, 512)$ vector with the targeted keyword semantic embedding, finally we perform the similar iteration, presented in Algorithm 1.

4. EXPERIMENT

We perform extensive experiments to test the effectiveness of our proposed method Fooled-by-Imagination, which crafts the adversarial example in complex domain and aims to attack CNN+RNN based image captioning systems. In our work, we use the pre-trained Tensorflow implementation of Show-and-Tell [10] with VGG-16 as the CNN for visual feature extraction. Although some recent image captioning systems have improved much better performance, Show-and-Tell is the vanilla version of CNN+RNN based framework, where CNN is employed to extract visual feature and RNN for caption generation. We verify our performance on MSCOCO [21], a widely known and large data set. In recent two years, some studies [9, 17, 8] focus to adversarial attack on vision-language system, however to the best of our knowledge, there is no other method based on complex domain to crafting adversarial examples, either for image classification or vision-language system. Besides, we verify the validity of the generated adversarial examples when transferring to another model.

We use ADAM to optimize each step and set the learning rate to 0.001. All the experiments are performed on a server with Nvidia GTX 1080 Ti GPU.

4.1. White-box Attack

For white-box attack, we can use the information of the attacked model, such as the details of architecture and output. In our experiments, we adopt Show-and-Tell as the attacked model, first we extract semantic embedding layer of (CNN+RNN) as E, then we utilize Equation 8 as objective function and iterate to generate adversarial examples.

Sentence Attack. Unlike the image recognition task where all possible labels are predefined, the possibilities of captions for image captioning are almost limitless. However, the captioning model can only generate correlative captions learned from the training set. Therefore, we adopt the same way to Show-and-Fool [9] to ensure that the targeted sentence lies in the sapce where the captioning system can possibly generate.

By adding a little perturbations to the original image, the generated adversarial examples can generate corresponding sentence on the attacked model. After extract semantic embedding layer as E, we then adopt the caption of an random chosen image, which is the targeted caption, to get $E(I_t)$. Finally, according to Algorithm 1, we optimize our model to generate better adversarial examples. For sentence attack, our adversarial examples are visually identical to the original images, as displayed in Figure 2.

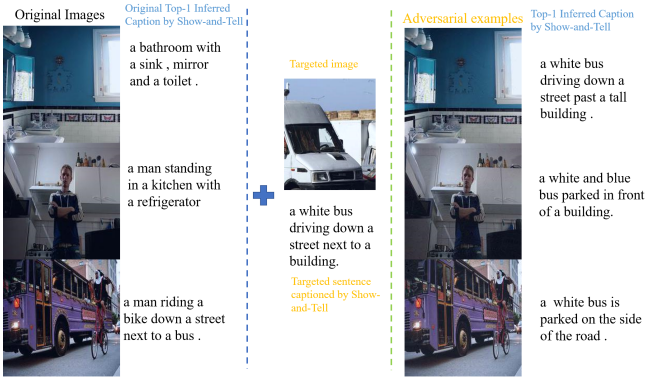

Fig. 2. Adversarial examples crafted by Fooled-by-Imagination using the targeted sentence method.

Table 1. Different sentence attack model's BLEU score

	FGSM	I+FGSM	ShowandFool	Ours
BLEU-1	0.671	0.843	0.723	**0.913**
BLEU-2	0.421	0.633	0.616	**0.856**
BLEU-3	0.399	0.461	0.514	**0.816**
BLEU-4	0.342	0.411	0.417	**0.801**

Like Show-and-Fool [9], we also employ BLEU-1, BLEU-2, BLEU-3, BLEU-4 [22] scores to evaluate the correlations between the inferred captions and the targeted ones. These scores are widely used in NLP community and are adopted by image captioning systems for quality assessment. From the results of Table 1, we conclude that our Fooled-by-Imagination strategy can generate better adversarial examples, especially the correlations of the captions. The main reason may be that our perturbations are based semantic embedding of targeted sentence in complex domain, which is obtained by removing the final layer when input the targeted image to the Show-and-Tell model.

Keyword Attack. In this task, we choose the number of keywords $M = 1, 2, 3$. In the same way to sentence attack, after training the attacked model (Show-and-Tell), we extract the semantic embedding layer as E. For each image the targeted keywords are randomly selected, and then we get the semantic embedding of word through the last linear layer in Show-and-Tell, and finally take the semantic embedding of word as $E(I_t)$. Note that to exclude common words like "a", "the", "and", we look up each word in the targeted sentence and only select nouns, verbs, adjectives or adverbs.

An adversarial image is thought to be successful when its

Table 2. The success rate of different keywords attack model.

	FGSM	I+FGSM	ShowandFool	Ours
1-keyword	0.846	0.912	0.971	**0.979**
2-keywords	0.613	0.926	0.975	**0.981**
3-keywords	0.599	0.871	0.960	**0.963**

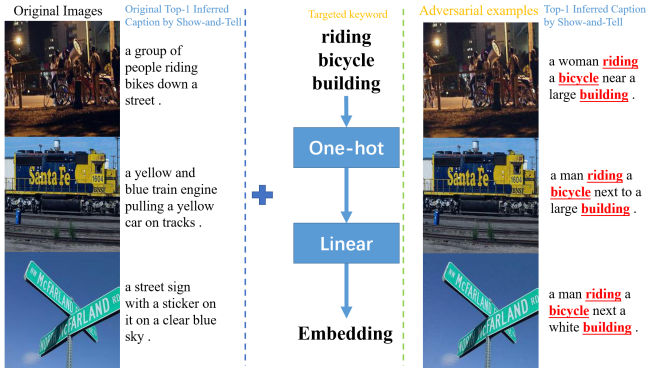

Fig. 3. Adversarial examples crafted by Fooled-by-Imagination using 3-keywords attack.

caption contains all specified keywords, and we use the attack success rate as the evaluation indicator. The overall success rate are shown in Table 2. We can draw a conclusion that our proposed Fooled-by-Imagination method achieves the best success rate (at least 96.3% for 3-keywords case, at least 97.9% for 1-keyword and 98.1% for 2-keywords cases). When compared to Show-and-Fool, our approach boost an even higher performance, but more simple and interesting.

Figure 3 shows some adversarial examples crafted from our targeted keyword method with three keywords - "riding", "bicycle" and "building". Using Fooled-by-Imagination, the top-1 caption of a street sign image becomes "a man riding a bicycle next a white building ", while the adversarial image remains visually indistinguishable to the benign one.

4.2. Black-box Attack and Transferability.

For black-box adversarial attack, we have no idea about the attacked model. To achieve this, we trained four Show-and-Tell model A, B, C and D with Resnet-152, Resnet-101, VGG-16 and Inception-V3 respectively as the CNN for visual feature extraction, but all the RNN structures are the same with unidirection LSTM. To verify the transferbility and the effectiveness on black-box, we employ E extracted from A to train our Fooled-by-Imagination method to generate adversarial examples, while we test the success rate of generated adversarial examples to attack model B, C and D. The same is true for other configurations. Here we take 1-keyword attack as an example to comparison, the results are displayed in Table 3. Apparently, our novel strategy has marvelous performance against black-box attack, and can transfer effectively to other models. In particular, among the 12 combinations, the lowest success rate still reaches 90.8%, and the highest rate of black-box attack could even achieve at 97.1%, that is an amazing result.

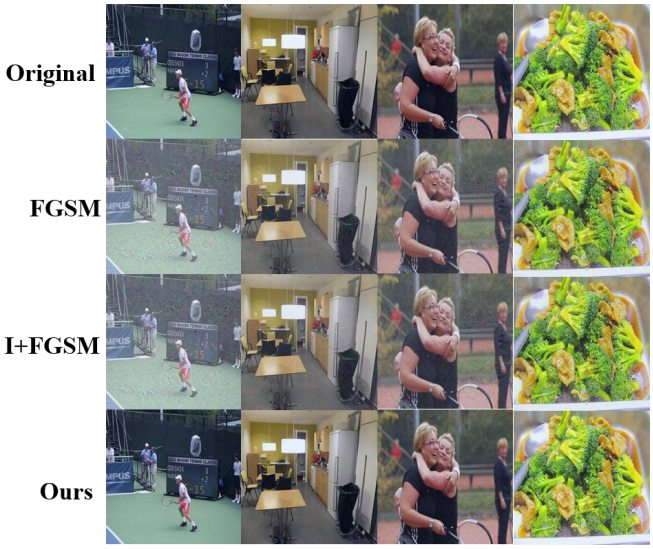

Fig. 4. Adversarial examples generated by different models.

Table 3. Different 1-keyword black-box attack model's success rate.

Algorithm		A	B	C	D
FGSM	A	0.841	0.620	0.614	0.644
	B	0.544	0.833	0.593	0.613
	C	0.571	0.546	0.824	0.513
	D	0.387	0.399	0.411	0.839
I+FGSM	A	0.908	0.703	0.716	0.752
	B	0.681	0.913	0.640	0.691
	C	0.634	0.631	0.896	0.563
	D	0.577	0.596	0.583	0.906
Ours	A	0.977	0.944	0.951	0.962
	B	0.913	0.971	0.908	0.918
	C	0.934	0.938	0.969	0.955
	D	0.917	0.919	0.926	0.981

4.3. Ablation Study

Quality of Adversarial Examples. To compare the quality of adversarial examples generated by different models, we random choose some instances presented in Figure 4, and find that adversarial examples generated by our approach are the most nearest to benign one and remain visually indistinguishable.

To further demonstrate the superiority of the generated adversarial samples, we make comparisons on l_1 and l_2 distance

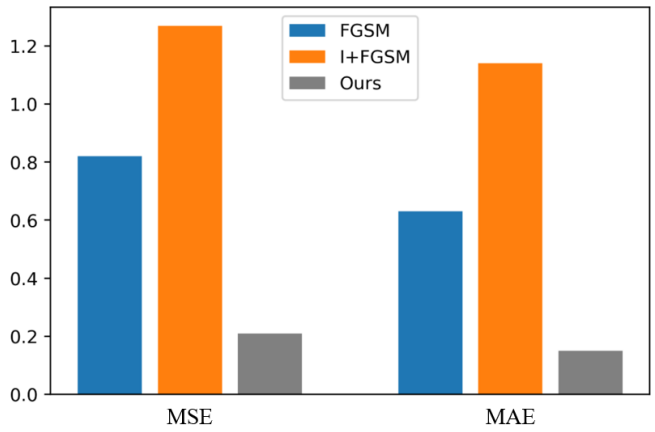

Fig. 5. l_1 and l_2 distance between original image and the generated adversarial examples by different models.

Table 4. Sentence attack's BLEU score with different α.

	0.3	0.8	1.0	1.5	2.0
BLEU-1	0.845	0.912	0.911	0.899	0.871
BLEU-2	0.753	0.834	0.847	0.851	0.812
BLEU-3	0.711	0.799	0.801	0.784	0.762
BLEU-4	0.703	0.781	0.799	0.744	0.736
mean	0.753	0.832	**0.840**	0.820	0.795

between original image and the generated adversarial examples by different models. As shown in Figure 5, the distance of l_1 (Mean Absolute Error) and l_2 (Mean Square Error) is far lower than other models, including FGSM and Iterative FGSM.

The Effectiveness of Parameter α. We do several experiments to analyze the importance of the balance parameter α. As Table 4 shows, when $\alpha = 0.8$, the BLEU-1 score is the highest, but the highest mean score is obtained at $\alpha = 1.0$. Table 5 shows keyword attack success rate on different α, the attack success rate achieves the highest at $\alpha = 0.8$, while $\alpha = 1.0$, we acquire the hishest mean success rate. In consequence, the α is set to 1.0 to strive for best performance in our work.

5. CONCLUSION

In this paper, we proposed a novel algorithm, Fooled-by-Imagination, for crafting adversarial examples first with perturbation in complex domain, and providing robustness evaluation of neural image captioning. Our extensive experiments show that our approach yields high attack success rates while the adversarial perturbations are still imperceptible to human eyes. We further demonstrate that Fooled-by-Imagination can generate highly transferable adversarial examples and also achieve black-box attack with high-quality. To the best of our knowledge, this is the very first work on crafting adversarial examples with perturbation in complex domain, especially for vision-language systems, and suggest a possible direction to study weakness of neural image captioning systems.

Table 5. Success rate of 1-keyword attack with different α.

	0.3	0.8	1.0	1.5	2.0
1-keyword	0.963	0.979	0.974	0.971	0.955
2-keyword	0.972	0.973	0.981	0.965	0.952
3-keyword	0.943	0.951	0.962	0.962	0.942
mean	0.960	0.968	**0.972**	0.966	0.950

6. REFERENCES

[1] Zheng Wang, Jie Zhou, Jing Ma, Jingjing Li, Jiangbo Ai, and Yang Yang, "Discovering attractive segments in the user-generated video streams," *Information Processing & Management*, vol. 57, no. 1, pp. 102130, 2020.

[2] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, 2019.

[3] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang, "Cra-net: Composed relation attention network for visual question answering," in *ACM Multimedia*, 2019.

[4] Zheng Wang, Kai Chen, Mingxing Zhang, Peilin He, Yajie Wang, Ping Zhu, and Yang Yang, "Multi-scale aggregation network for temporal action proposals," *Pattern Recognition Letters*, vol. 122, pp. 60 – 65, 2019.

[5] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain, "Adversarial attacks and defenses in images, graphs and text: A review," in *CoRR*, 2019.

[6] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian, "Universal perturbation attack against image retrieval," in *ICCV*, October 2019.

[7] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang, "Black-box adversarial attacks on video recognition models," in *ACM MM*, 2019.

[8] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu, "Exact adversarial attack to image captioning via structured output learning with latent variables," in *CVPR*, June 2019.

[9] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *ACL*, July 2018, pp. 2587–2597.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, June 2015, pp. 3156–3164.

[11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[12] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018, pp. 684–699.

[13] Yi Bin, Yang Yang, Jie Zhou, Zi Huang, and Heng Tao Shen, "Adaptively attending to visual attributes and linguistic knowledge for captioning," in *ACM MM*, 2017.

[14] Goodfellow Ian J, Shlens Jonathon, and Christian Szegedy., "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *CVPR*, June 2016.

[16] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 39–57.

[17] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song, "Can you fool AI with adversarial examples on a visual turing test?," in *CVPR*, 2018.

[18] David P. Reichert and Thomas Serre, "Neuronal synchrony in complex-valued deep networks," *Computer Science*, 2013.

[19] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal, "Deep complex networks," in *ICLR*, 2018.

[20] Moritz Wolter and Angela Yao, "Complex gated recurrent neural networks," in *NeurIPS*, 2018, pp. 10557–10567.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, July 2002, pp. 311–318.