

# Fooling Vision and Language Models Despite Localization and Attention Mechanism

Xiaojun Xu<sup>1,2</sup>, Xinyun Chen<sup>2</sup>, Chang Liu<sup>2</sup>, Anna Rohrbach<sup>2,3</sup>, Trevor Darrell<sup>2</sup>, Dawn Song<sup>2</sup>

<sup>1</sup>Shanghai Jiao Tong University, <sup>2</sup>EECS, UC Berkeley, <sup>3</sup>MPI for Informatics

## Abstract

*Adversarial attacks are known to succeed on classifiers, but it has been an open question whether more complex vision systems are vulnerable. In this paper, we study adversarial examples for vision and language models, which incorporate natural language understanding and complex structures such as attention, localization, and modular architectures. In particular, we investigate attacks on a dense captioning model and on two visual question answering (VQA) models. Our evaluation shows that we can generate adversarial examples with a high success rate (i.e., > 90%) for these models. Our work sheds new light on understanding adversarial attacks on vision systems which have a language component and shows that attention, bounding box localization, and compositional internal structures are vulnerable to adversarial attacks. These observations will inform future work towards building effective defenses.*

## 1. Introduction

Machine learning, especially deep learning, has achieved great success in various application scenarios, such as image classification, speech recognition, and machine translation. However, recent studies prove the existence of adversarial examples for many vision-based learning models, which may hinder the adoption of deep learning techniques to security-sensitive applications [19, 43, 56, 65]. Most existing works consider image classification and demonstrate that it is almost always possible to fool these models to classify an adversarially generated image as a class specified by the adversary [66]. Albeit numerous defenses have been proposed [19, 60, 52, 67, 51, 48], almost all of them are later shown to be broken [8, 23, 9].

Recently, there has been an increasing interest in whether adversarial examples are practical enough to attack more complex vision systems [44, 45, 6]. In the latest results of this debate, Lu et al. show that previous adversarial examples constructed to fool CNN-based classifiers cannot fool state-of-the-art detectors [45]. We are interested in whether

other forms of localization and/or language context offer effective defense.

In this work, we extend the investigation towards more complex models that not only include a vision component but also a *language component* to deepen our understanding of the practicality of adversarial examples. In particular, we investigate two classes of systems. First, we are interested in dense captioning systems, such as DenseCap [30], which identify regions of interest first and then generate captions for each region. Second, we are interested in visual question answering (VQA) systems, which answer a natural language question based on a given image input. The state-of-the-art VQA systems typically compute attention maps based on the input and then answer the question based on the attended image regions. Therefore, both types of models have a *localization component*, and thus they are good targets for studying whether localization can help prevent adversarial attacks. Further, we explore state-of-the-art VQA models based on Neural Modular Networks [25], and evaluate whether such compositional architectures are also vulnerable to adversarial attacks; in these models, a new network architecture is instantiated for each question type, potentially providing a buffer against attacks.

We evaluate adversarial examples against these vision and language models. We find that in most cases, the attacks can successfully fool the victim models despite their internal localization component via attention heatmaps or region proposals, and/or modular structures. Our study shows that, in an online (non-physical) setting when the attackers have full access to the victim model including its localization component (white-box attack), the generated adversarial examples can fool the entire model regardless of the localization component. Therefore, our evaluation results provide further evidence that employing a localization in combination with a classifier may not be sufficient to defend against adversarial examples, at least in non-physical settings.

We also make the following additional contributions. First, we develop a novel attack approach for VQA models, which significantly outperforms the previous state-of-the-art attacks. Second, we observe and analyze the effect of a

*language prior* in attacking VQA models, and define a principle which explains which adversarial examples are likely to fail. In particular, when the target answer is not compatible with the question, it is difficult to find a successful adversarial attack using existing approaches. To sum up, our work sheds new light on understanding adversarial attacks on vision and language systems and shows that attention, bounding box localization and compositional internal structures are vulnerable to adversarial attacks. These observations will inform future work towards building effective defenses.

## 2. Related Work

In the following, we first review recent work on image captioning and visual question answering. We focus on the models that incorporate some form of localization, e.g. soft attention or bounding box detection. We then review the state-of-the-art methods to generate adversarial examples as well as defense strategies against these methods.

**Image Captioning** Most recent image captioning approaches have an encoder-decoder architecture [11, 12, 32, 33, 50, 69]. A spatial attention mechanism for image captioning was first introduced by [73]. They explored soft attention [7] as well as hard attention. Others have adopted this idea [15, 42, 46, 76] or extended it to perform attention over semantic concepts, or attributes [77, 79]. Recently [61] proposed an end-to-end model which regresses a set of image regions and learns to associate caption words to these regions. Notably, [2, 12, 32] exploited object detection responses as input to the captioning system. As opposed to image captioning of the entire image, [30] have proposed *dense captioning*, which requires localization and description of image regions (typically bounding boxes). Some other dense captioning approaches include [40, 74].

**Visual Question Answering.** Early neural models for visual question answering (VQA) were largely inspired by image captioning approaches, e.g. relying on a CNN for image encoding and a RNN for question encoding [17, 49, 62]. Inspired by [73], a large number of works have adopted an attention mechanism for VQA [16, 47, 64, 72, 75, 80]. Semantic attention has been explored by [78]. Other directions explored by recent work include Dynamic Memory Networks (DMN) [36, 71], and dynamic parameter layers (DPP) [55]. Recently a new line of work focused on developing more compositional approaches to VQA, namely neural module networks [3, 4, 25, 29]. These approaches have shown an advantage over prior work for visual question answering which involve complex reasoning.

**Adversarial Examples.** Existing works on adversarial example generation mainly focus on image classification models. Several different approaches have been pro-

posed for generating adversarial examples, including fast gradient-based methods [19, 43], optimization-based methods [66, 10], and others [58, 54]. In particular, Carlini et al. [10] proposed the state-of-the-art attacks under constraints on  $L_0$ ,  $L_2$ , and  $L_\infty$  norms. Our work improves [10] on both attack success rate and adversarial probability.

Another line of research studies adversarial examples against deep neural networks for other tasks, such as recurrent neural networks for text processing [59, 28], deep reinforcement learning models for game playing [41, 26, 34], semantic segmentation [14, 70], and object detection [24, 70]. To our best knowledge, our work is the first to study adversarial examples against vision-language models.

While our work assumes that models are known to the attacker, prior works demonstrate that adversarial examples can transfer between different deep neural networks for image classification [66, 19, 43, 56, 58, 53], which can be used for black-box attacks. We briefly analyze the transferability of VQA models in the supplemental material.

**Defense against Adversarial Examples.** On the defense side, numerous strategies have been proposed against adversarial examples [19, 60, 52]. Early attempts to build a defense using distillation [60] were soon identified as vulnerable [8]. Some recent proposals attempt to build a *detector* to distinguish adversarial examples from natural images [52, 21, 18, 13]. Others study ensembles of different models and defense strategies to see whether that helps to increase the robustness of deep neural networks [67, 68, 51]. However, He et al. show that with the knowledge of the detector network and the defense strategies being used, an attacker can generate adversarial examples that can mislead the model, while still bypassing the detector [23].

The most promising line of defense strategies is called *adversarial training* [19, 37, 67, 48]. The idea is to generate adaptive adversarial examples and train the model on them iteratively. The latest results along the line [48] show that such an approach can build a robust MNIST model. But the same approach currently fails on extending to CIFAR-10.

## 3. Generating Targeted Adversarial Examples

In this section, we first present a generic adversarial example generation algorithm, and then our implementations for dense captioning models and VQA models.

### 3.1. Background: targeted adversarial examples for a classification model

Consider a classification model  $f_\theta(x)$ , where  $\theta$  is the parameters and  $x$  is the input. Given a source image  $x$ , a targeted adversarial example is defined as  $x^*$  such that

$$f_\theta(x^*) = y^t \wedge d(x^*, x) \leq B \quad (1)$$

where  $y^t$  is the target label, and  $d(x^*, x) \leq B$  says that the distance between  $x$  and  $x^*$  is bounded by a constant  $B$ .

Without loss of generality,  $f_\theta(x)$  predicts the dimension of the largest softmax output. We denote  $J_\theta(x)$  as the softmax output, then a standard training algorithm typically optimizes the empirical loss  $\sum_i \mathcal{L}(J_\theta(x_i), y_i)$  with respect to  $\theta$  using a gradient decent-based approach. Existing adversarial example generation algorithms leverage the fact that  $J_\theta(x)$  is differentiable, and thus solve (1) by optimizing the following objective:

$$\operatorname{argmin}_{x^*} \mathcal{L}(J_\theta(x^*), y^t) + \lambda d(x^*, x) \quad (2)$$

where  $\lambda > 0$  is a hyper-parameter. In fact, the state-of-the-art attack [10] approximates the solution to (2) using Adam.

### 3.2. Targeted adversarial examples for DenseCap

The DenseCap model [30] predicts  $M = 1000$  regions, ranks them based on confidence, and then generates a caption for each region. It uses a localization network, similar to Fast R-CNN [63], for predicting regions. For each region, the model uses a CNN to compute the embedding and then uses an RNN to generate a sequence of tokens from the embedding to form the caption.

To train the DenseCap model, Johnson et al. include five terms in the loss: four for training the region proposal network, and the last one to train the RNN caption generator. To fool the model to predict the wrong target caption, we can leverage a similar process as discussed above. Note that existing works [24, 70] have demonstrated that an object detection/segmentation model can be fooled by adversarial examples. In this work, we focus on generating adversarial examples to fool the captioning module of the model, while retaining the proposed regions unchanged.

To achieve this goal, assuming the target caption is  $C^t$  and the ground truth regions for a source image are  $\{R_i\}$ , we construct a new set of target region-caption pairs  $\{(R_i, C^t)\}$ . Using these target region-caption pairs as the new ‘‘ground truth’’, we can use the DenseCap loss, with addition of the  $\lambda d(x^*, x)$  term as in (2), as the new objective, and minimize it with respect to  $x^*$ .

### 3.3. Targeted adversarial examples for VQA models

We now briefly present our novel targeted adversarial attack against VQA models. More details can be found in Appendix A in the supplemental materials. Our design is inspired by two goals: (1) maximizing the probability of the target answer, which is equivalent to the confidence score of the model’s prediction; and (2) removing the preference of adversarial examples with smaller distance to the source image, as long as this distance is small enough (i.e., below an upper bound). Our evaluation shows that our algorithm performs better than the previous state-of-the-art [10].

---

#### Algorithm 1 Targeted Adversarial Generation Algorithm against a VQA model

---

**Input:**  $\theta, x, Q, y^t, B, \epsilon, \lambda_1, \lambda_2, \eta, \mathbf{maxitr}$   
**Output:**  $x^*$

- 1  $x^1 \leftarrow x + \delta$  for  $\delta$  sampled from a uniform distribution between  $[-B, B]$ ;
- 2 **for**  $i = 1 \rightarrow \mathbf{maxitr}$  **do**
- 3  $y^p \leftarrow f_\theta(x^i, Q)$ ;
- 4 **if**  $y^p = y^t$  and  $i > 50$  **then**
- 5 **return**  $x^i$  as  $x^*$ ;
- 6  $x^{i+1} \leftarrow \mathbf{update}(x^i, \eta, \nabla_x \xi(y^p))$ ;
- 7 **return**  $x^{\mathbf{maxitr}+1}$  as  $x^*$ ;

---

A VQA model takes an additional natural language input  $Q$ , and predicts an answer from a candidate set of  $K$  answers. Similar to (1), a targeted adversarial example  $x^*$  given a question  $Q$  is defined to be a solution to:

$$f_\theta(x^*, Q) = y^t \wedge d(x^*, x) \leq B \quad (3)$$

We employ Algorithm 1 to generate the adversarial example  $x^*$ . The algorithm takes as input: model parameters  $\theta$ , source image  $x$ , question  $Q$ , target answer  $y^t$ , the distance bound  $B$ , and several hyper-parameters:  $\epsilon, \lambda_1, \lambda_2, \eta, \mathbf{maxitr}$ . This algorithm iteratively approximates the optimal solution to the following objective:

$$\begin{aligned} \xi(y^p) = & \mathcal{L}(J_\theta(x^*, Q), y^t) \\ & + \lambda_1 \cdot \mathbf{1}(y^t \neq y^p) \cdot (\tau - \mathcal{L}(J_\theta(x^*, Q), y^p)) \\ & + \lambda_2 \cdot \operatorname{ReLU}(d(x^*, x) - B + \epsilon) \end{aligned} \quad (4)$$

and returns the final result as output. There are two terminating conditions: (1) after at least 50 iterations, if the prediction matches the target, then the algorithm stops and returns the current  $x^i$  as output; or (2) after a maximal number of iterations ( $\mathbf{maxitr}$ ), if the prediction still does not match the target, the algorithm returns  $x^{\mathbf{maxitr}+1}$  as output.

We now take a closer look at (4).  $y^p$  denotes the prediction in each iteration. The objective (4) contains three components. The first is the same as in (2). The second component maximizes the difference between  $J_\theta(x, Q)$  and the prediction  $y^p$  when  $y^p$  is not the target  $y^t$ .  $\tau$  is a constant, e.g.,  $\log(K)$ , set to ensure that the second component is always non-negative. The third component models the constraint  $d(x^*, x) \leq B$  in (3).  $\epsilon$  is a small constant set to  $(\mathcal{L}(f_\theta(x, Q), y^t) + \lambda_1 \tau) / \lambda_2$  ensures that the adversarial example  $x^*$  which optimizes (4) always satisfies  $d(x^*, x) \leq B$ . By using a ReLU function, our attack no longer minimizes the distance  $d(x^*, x)$  if it is smaller than  $B - \epsilon$ . In practice we choose  $d(x, x^*) = \|x - x^*\|_2 / \sqrt{N}$  and set  $B = 20$ . Other hyper-parameters  $\eta, \mathbf{maxitr}$  are the learning rate and the maximal number of iterations. We defer a formal analysis to the supplemental material.

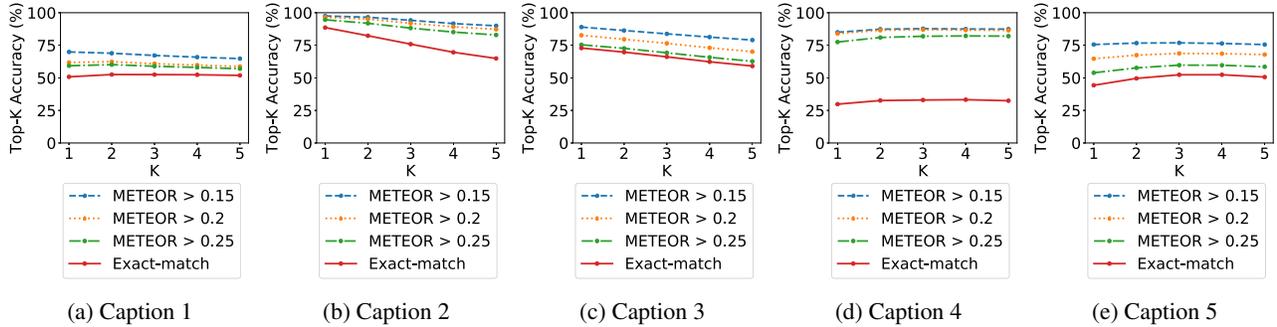


Figure 1: Top- $K$  accuracy on the Caption A dataset averaged across 1000 images generated with each target caption

## 4. Experiments With Dense Captioning

In this section, we evaluate our attacks on DenseCap [30], the state-of-the-art dense captioning model. DenseCap employs a region proposal network to first identify the bounding boxes of objects, and then generates captions for each bounding box. We obtain the pre-trained model from their website<sup>1</sup>.

To evaluate the attack, we use Visual Genome dataset [35], which was originally used to evaluate DenseCap in [30]. For an extensive evaluation, we create the following three attack sets from Visual Genome:

- 1) **Caption A.** We randomly select 5 captions as the target captions and 1000 images as the source images;
- 2) **Caption B.** We randomly select 1000 captions as target captions and 5 images as source images;
- 3) **Gold.** We select 100 images where DenseCap model generates correct captions and manually select target captions irrelevant to the images.

For each caption-image pair, we set the caption as the target, and the image as the source to generate an adversarial example. To evaluate the attack effectiveness, we measure the percentage of top- $K$  predictions from generated adversarial examples that *match* the target captions. We consider two metrics to determine caption matching:

- 1) **Exact-match.** The two captions are identical.
- 2) **METEOR $> \omega$ .** The METEOR score [38] between the two captions is above a threshold  $\omega$ . We consider the threshold  $\omega$  to be 0.15, 0.2, or 0.25, similar to [30].

Formally, we measure  $Acc_{\mu,K}(x^*, C^t) = \sum_{i=1}^K \mu(C^t, C_i) / K$  where  $C^t$  is the target caption,  $x^*$  is the adversarial example,  $C_i$  for  $i = 1, \dots, K$  are the top- $K$  predictions for  $x^*$ , and  $\mu$  is the matching metric (i.e., Exact-match or METEOR $> \omega$ ).

### 4.1. Results and Observations

The evaluation results on Caption A are presented in Figure 1. Each subfigure shows the results for one target caption. For each caption and each  $K \in \{1, 2, 3, 4, 5\}$ , we

compute  $Acc_{\mu,K}$  for each of the 1000 randomly selected images, and report the average value of  $Acc_{\mu,K}$  across 1000 images. Each plot contains such 5 top- $K$  accuracy values for each metric described above (see the legend).

We observe that using the metric derived from METEOR score, the accuracy is higher than using the Exact-match metric. This is intuitive, since Exact-match is an over-conservative metric, which may treat a semantically correct caption as a wrong answer. In contrast, using METEOR score as the metric can mitigate this issue. Even with Exact-match, we observe that all captions have an average top- $K$  accuracy above 30%. Further, for target captions Caption 1-3, the top-1 accuracy is always above 50%. That means, at least 500 generated adversarial examples can successfully fool the DenseCap system to produce the exact target captions with the highest confidence score.

We further investigate the number of attack “failures” among caption-image pairs in Caption A. The attack *fails* if none of the top-5 predictions matches the target based on METEOR $> 0.15$ . We find only 17 such caption-image pairs, i.e., 0.35% of the entire set, which lead to adversarial attack failure. This means that for the rest 99.65% caption-image pairs, the attacks are successful in the sense that there exists at least one prediction for each adversarial example that matches the target caption. The 17 cases can be found in Appendix B in the supplemental material.

The results on Caption B set are similar, and we observe that 97.24% of the caption-image pairs can be successfully fooled in the sense described above. For the Gold set we find that our attack fails only on one image. Due to space limitations, we defer detailed results on Caption B and Gold sets to Appendix B in the supplemental materials.

Note that the attack does not achieve a 100% success rate. We attribute it to two reasons: (1) it is challenging to train an RNN-based caption generation model to generate the exactly matching captions; and (2) the DenseCap network involves randomness, and thus may not produce the same results for all runs. Still, we observe that the attack success rate is over 97%, and thus we conclude that the DenseCap model can be fooled by adversarial examples.

<sup>1</sup><https://github.com/jcjohnson/densecap>

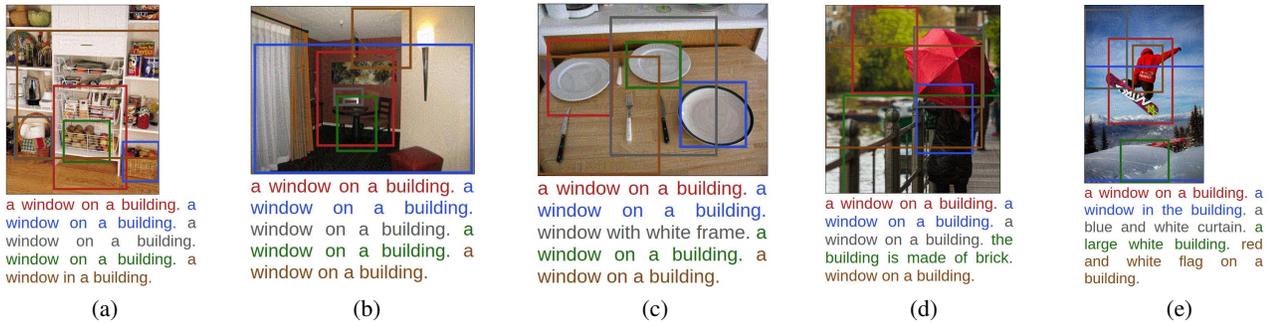


Figure 2: Adversarial examples generated from different images with the target caption to be “a window on a building”.

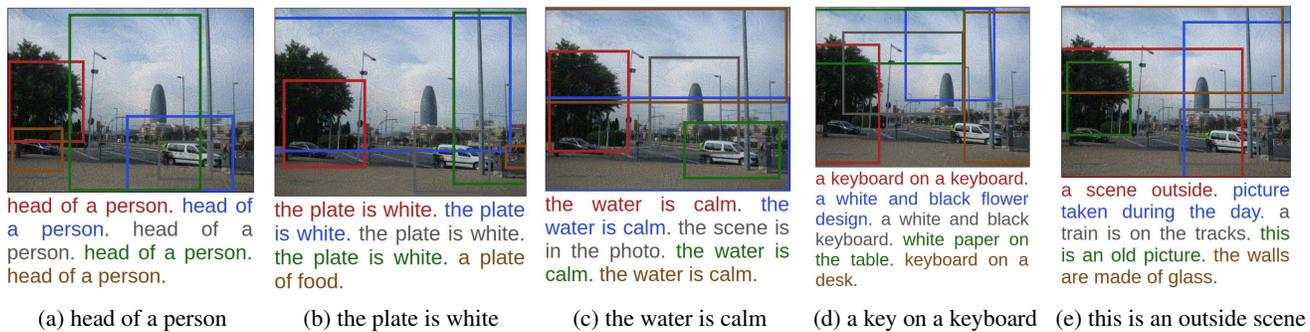


Figure 3: Adversarial examples generated from Image 4 with different target captions (shown as sub-figure captions).

## 4.2. Qualitative Study

We conduct qualitative study to investigate the generated adversarial examples and their predictions. In Figure 2, we present five adversarial examples generated for the same target caption. We see that most of the predicted captions exactly match the target (e.g., all top-5 predictions for Figure 2a and Figure 2b), or be semantically equivalent to the target (e.g., the top-2 prediction for Figure 2e). We further examine the bounding boxes of the regions proposed by the model. We find that the model localizes objects in the adversarial examples, although the caption generation module of the model is completely fooled. For example, in Figure 2c, the model can successfully identify the plates, but label all of them as “a window on a building”.

To further understand this effect, in Figure 3, we show the adversarial examples generated from the same source image but with different target captions. We observe that all adversarial images look identical to each other, and the regions proposed for different images are also similar. For example, we observe that the top proposed regions for the first four images all circumscribe the tree on the left. However, the top captions generated for this region are all different, and match the target captions very well.

## 5. Experiments with VQA

In this section, we evaluate the previous state-of-the-art attack [10] and our novel algorithm on two VQA models.

We also investigate the effect of adversarial attacks on *attention maps* of the VQA models to gain more insights about the way the attacks work. Finally, we analyze the successes and failures of our attacks with respect to *language prior*. More results on qualitative study, transferability, and further investigations to the failure cases can be found in Appendix D and E in the supplemental materials.

### 5.1. Models

We experiment with two state-of-the-art models for open-ended visual question answering, namely the MCB model [16], which is the winner of the VQA challenge in 2016, and the compositional model N2NMN [25]. Both models achieve similar performance on the VQA benchmark [5], while being very different in terms of internal structures. MCB relies on a single monolithic network architecture for all questions, while N2NMN dynamically predicts a network layout for every given question. In our experiments we investigate whether such compositional dynamic architecture is more resilient than the monolithic one.

We retrieve the pre-trained model of MCB from their website<sup>2</sup>, and the pre-trained model of N2NMN by contacting the authors through email directly. The code implementing N2NMN is acquired from the website.<sup>3</sup> Notice that the MCB model is trained not only on the VQA dataset but

<sup>2</sup><https://github.com/akirafukui/vqa-mcb>

<sup>3</sup><https://github.com/ronghanghu/n2nmn>

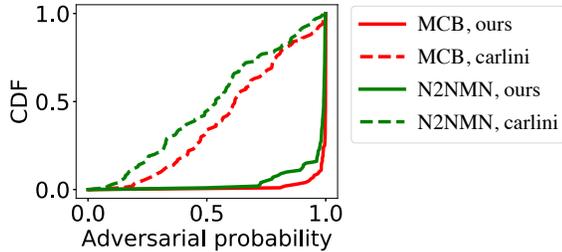


Figure 4: CDF of adversarial probability on the Gold set.

also on the Visual Genome dataset [35], while the N2NMN model only considers the VQA dataset.

### 5.2. Datasets

To evaluate different adversarial example generation algorithms we derive three datasets from the VQA dataset [5]. In particular, we choose source images and question-answer targets from the VQA validation set as follows:

- 1) VQA-A: We randomly select 6,000 question-answer pairs and 5 source images to constitute 30,000 triples;
- 2) VQA-B: We randomly select 5,000 source images and 10 question-answer pairs to construct 50,000 triples;
- 3) Gold: We manually select 100 triples, such that MCB and N2NMN models can correctly answer the questions based on the images, and the target answers are plausible for the questions but incorrect for the images.

For each triple of question-answer-image, we generate an adversarial example close to the source image using the answer as the target. More details can be found in Appendix C in the supplemental materials.

### 5.3. Evaluation metrics

Given a set of question-answer pairs  $(Q, y^t)$  and the generated adversarial examples  $\{x^*\}$ , we evaluate two metrics: the *attack success rate* and the *adversarial probability*.

**Attack success rate.** The attack is considered successful if  $f_\theta(x^*, Q) = y^t$ . The attack success rate is computed as the percentage of successful attacks over all triples in a dataset.

**Adversarial probability.** The *adversarial probability* is computed as  $J_\theta(x^*, Q)_{y^t}$ , where  $J(\cdot, \cdot)_i$  indicates the  $i$ -th dimension of the softmax output. Adversarial probability indicates the confidence score of the model to predict the target answer  $y^t$ , and thus provides a fine-grained metric.

### 5.4. Results

Here we report the overall success of adversarial attacks on VQA models, and also compare our new algorithm described above with the performance of the previous attack algorithm (CW [10]) applied to this novel VQA setting. We present the quantitative results below, and defer more qualitative results to the supplemental material.

Image #		1	2	3	4	5
MCB	ours	94.67	94.78	94.97	95.02	95.15
	CW [10]	94.10	94.28	94.27	94.52	94.78
N2NMN	ours	94.25	94.53	95.57	95.80	96.15
	CW [10]	93.82	93.78	95.02	95.08	95.37

Table 1: Attack success rate (%) on VQA-A.

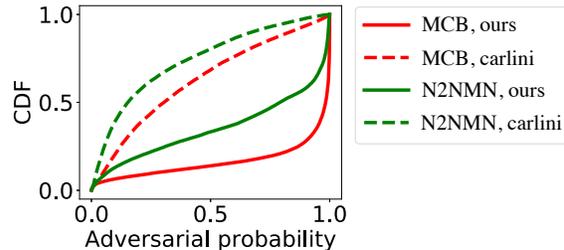


Figure 5: CDF of adversarial probability on VQA-A.

**Gold.** For both MCB and N2NMN models, we achieve 100% attack success rate using either approach. Note that both models can correctly answer all the questions on the original source images. The 100% attack success rate for both VQA models shows that both of them are vulnerable to targeted adversarial examples.

We inspect the adversarial probabilities of the generated adversarial examples, and plot the Cumulative Distribution Function (CDF) in Figure 4. Note that a lower CDF curve indicates a higher probability in general. From the figure we observe that the CDF curve of N2NMN is above MCB’s, indicating that N2NMN is slightly more resilient than MCB. However, we also observe that for both models, almost in all cases the adversarial probability is above 0.7. Thus, we conclude that our attack is very successful at misleading the VQA models to predict the target answers. We also observe that the CDF curve of CW attack is much higher than ours, showing that our approach is more effective at achieving a high adversarial probability. Overall, we show that such attacks can be performed very successfully for target answers that are meaningful to questions.

**VQA-A.** We further investigate VQA adversarial examples across a wide range of target question-answer pairs. We separately compute the attack success rate using each image as the source. The results are presented in Table 1, and the corresponding CDF curves are plotted in Figure 5. We can draw similar conclusions as for the Gold set: (1) the attack success rate is high, i.e.,  $> 90\%$ ; (2) the adversarial probability of our attack is high; and (3) our attack is more effective than CW attack.

We observe that the attack’s performance against N2NMN model is worse than against MCB. In particular, from Figure 5, we see that the adversarial probability of at-

Original image	Benign attention maps		Adversarial attention maps	
	MCB Attention	N2NMN Attention	MCB Attention	N2NMN Attention
<b>What is the man holding?</b> 	Original answer: <b>racket</b>		Target: <b>phone</b>	
<b>What does this sign say?</b> 	Original answer: <b>stop</b>		Target: <b>one way</b>	
<b>What type of vehicle is this?</b> 	Original answer: <b>train</b>		Target: <b>bus</b>	

Table 2: Attention maps of benign and adversarial images on MCB and N2NMN models.

tacks generated on the N2NMN model is significantly lower than the MCB model. This further shows that N2NMN model is somewhat more robust against adversarial attacks. We also observe that the attack success rate with respect to different images does not vary too much. We hypothesize that the attack success is not sensitive to a source image, but more dependent on a target question-answer pair. Our further investigations on VQA-B and language priors below provide more evidence to confirm this hypothesis.

The attack success rate is not 100%, which shows that there exist a few question-answer pairs where neither ours nor the CW attack can succeed. In fact, for these question-answer pairs, we have also tried other attack methods and none of them can succeed in fooling the victim VQA model. We find that these question-answer pairs tend to appear infrequently in the training set, and this observation leads to our hypothesis regarding *language prior*. We present more analysis of the language prior in the following section.

**VQA-B.** We test the hypothesis that *the attack success rate is not strongly dependent on the choice of source images* using the VQA-B dataset. In our evaluation, we observe that for 9 out of 10 question-answer pairs, the adversarial examples generated from any of the 5,000 source images fool the victim model with 100% attack success rate. For the one remaining question-answer pair, however, we cannot generate successful adversarial examples from *any* of the source

images. This result further confirms our hypothesis. Interestingly, we observe that the “hard” question-answer pairs for the two VQA models are different. For the MCB model, the question is “Why is the girl standing in the middle of the room with an object in each hand?” with the target answer “playing wii”; for the N2NMN model, the question and answer are “Who manufactured this plane?” and “japan”, respectively. This suggests that the hard question-answer pairs are model-specific, which further motivates us to investigate language prior in VQA models.

### 5.5. Adversarial examples fool attention mechanism

We conduct a qualitative study to gain more insights as to how the attack succeeds. In the following we use the Gold dataset. In particular, both models in our experiments have attention mechanism. That is, to answer a question, a model first computes an *attention map*, which is a weight distribution over local features extracted from a CNN based on the image and the question. Intuitively, a well-performing model should put more weight, i.e. attend to, the image region that is most informative to answer the question.

We demonstrate the attention heatmaps for three source images and their adversarial counterparts in Table 2. We observe that the adversarial examples mislead the VQA models to ignore the regions that support the correct answer to the question. For example, in the second source image both

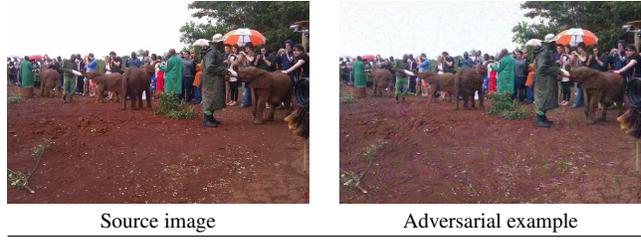
MCB and N2NMN focus on the stop sign when answering the question. The adversarial examples fool MCB and N2NMN to pay attention to the street sign instead, which leads to predicting a one-way traffic sign, likely because both signs are long rectangular metal plates. In the last example, the attention is misled to ignore the rail tracks but focusing on the windows which look similar to those on a bus. Therefore, we observe that adversarial examples can fool both the attention and the classification component of the VQA models to achieve the malicious goal.

### 5.6. Language Prior

We illustrate the *language prior* phenomenon in Figure 6. It provides an example which cannot be successfully attacked by any algorithm in our evaluation. We show an adversarial example generated by our attack algorithm, and the top-5 predictions from the MCB model. Clearly, the model is *confused* about the image and the question. The answer with the highest probability only has a probability of less than 5%. Although the model is confused, it cannot be *fooled* to predict the target answer “partly” to the question “what animal is next to the man?”. This observation is different than those reported in the literature [66], i.e., that targeted adversarial examples can always be successfully generated against an *image classifier* regardless of the image and the target label. We believe that the observed phenomenon is due to the internal mechanism of a VQA model which learns to process natural language questions and predict semantically relevant answers.

In all previous experiments we choose question-answer pairs from the VQA validation set, and thus the answers are likely meaningful to the questions. To evaluate the effect of language prior we construct the *Non-Sense* dataset. Specifically, we choose question-answer pairs, such that answers do not match the questions semantically, as they belong to questions of a different type (e.g. “what color” vs. “how many”). We find that the attack success rates using our approach against MCB and N2NMN are only 7.8% and 4.6% respectively; the corresponding numbers for CW attack are even lower, 6.8% and 3.8%. This experiment further confirms the significance of the language prior.

Prior work has noted the effect of language prior, i.e. that the VQA models capture the training data biases and tend to predict the most frequent answers [1, 20, 27, 31]. We find that N2NMN is more influenced by language prior than MCB. Specifically, N2NMN produces a smaller number of distinct answers, predicting question-relevant answers independent of image content. This may explain why it is more difficult to achieve a high probability on some targets with N2NMN than with MCB. We include more results and analysis in Appendix E in the supplemental materials.



Rank	Answer	Probability
1	yes	0.042
2	middle	0.041
3	on wall	0.040
4	left	0.031
5	background	0.025

Figure 6: The effect of language prior. The target question / answer are “What animal is next to the man?” / “partly”. We show the top-5 predictions from MCB after the attack.

## 6. Conclusion

In this work, we study adversarial attacks against vision and language models, specifically, dense captioning and visual question answering models. The models in our study are more complex than previously studied image classification models, in the sense that they contain language generation component, localization, attention mechanism, and/or compositional internal structures. Our investigation shows that (1) we can generate targeted adversarial examples against all victim models in our study with a high success rate (i.e., > 90%); and (2) the attacks can either retain the localization output or also fool the attention heatmaps to fool the victim model. While studying attacks on VQA models, as additional contributions, we propose a better attack method than the previous state-of-the-art approach. Also, we observe and evaluate the effect of language prior that may explain which question-answer pairs represent harder targets. Our work sheds new light on understanding adversarial attacks on complex vision and language systems, and these observations will inform future directions towards building effective defenses.

## Acknowledgement

This work was supported in part by the National Science Foundation under Grant No. TWC-1409915, Berkeley DeepDrive, DARPA STAC under Grant No. FA8750-15-2-0104, and Center for Long-Term Cybersecurity. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

- [1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8
- [2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017. 2
- [3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In *Proc. of NAACL*, 2016. 2
- [4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proc. of CVPR*, 2016. 2
- [5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proc. of ICCV*, 2015. 5, 6
- [6] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017. 1
- [7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*, 2015. 2
- [8] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016. 1, 2
- [9] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISeC*, 2017. 1
- [10] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 2, 3, 5, 6, 12, 13
- [11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. of CVPR*, 2015. 2
- [12] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [13] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017. 2
- [14] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101*, 2017. 2
- [15] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. 2016. 2
- [16] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2, 5
- [17] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proc. of NIPS*, 2015. 2
- [18] Z. Gong, W. Wang, and W.-S. Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017. 2
- [19] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proc. of ICLR*, 2015. 1, 2
- [20] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8
- [21] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 2
- [22] 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 (CVPR)*, 2016. 19
- [23] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017. 1, 2
- [24] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3
- [25] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proc. of ICCV*, 2017. 1, 2, 5
- [26] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017. 2
- [27] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision (ECCV)*, 2016. 8
- [28] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 2
- [29] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proc. of ICCV*, 2017. 2
- [30] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016. 1, 2, 3, 4
- [31] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 8
- [32] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*, 2015. 2

- [33] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9:595–603, 2015. 2
- [34] J. Kos and D. Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017. 2
- [35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4, 6
- [36] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2
- [37] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *Proc. of ICLR*, 2017. 2
- [38] M. D. A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 376, 2014. 4
- [39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 12
- [40] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and caption regions. In *Proc. of ICCV*, 2017. 2
- [41] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *Proc. of IJCAI*, 2017. 2
- [42] C. Liu, J. Mao, F. Sha, and A. Yuille. Attention correctness in neural image captioning. In *Proc. of AAAI*, 2017. 2
- [43] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. of ICLR*, 2017. 1, 2, 13, 17, 19
- [44] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017. 1
- [45] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. Standard detectors aren’t (currently) fooled by physical adversarial stop signs. *arXiv preprint arXiv:1710.03337*, 2017. 1
- [46] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. of CVPR*, 2017. 2
- [47] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical co-attention for visual question answering. *Proc. of NIPS*, 2, 2016. 2
- [48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2
- [49] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proc. of CVPR*, pages 1–9, 2015. 2
- [50] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. of ICLR*, 2015. 2
- [51] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In *CCS*, 2017. 1, 2
- [52] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. In *2017*, 2017. 1, 2
- [53] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proc. of CVPR*, 2017. 2, 17
- [54] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. of CVPR*, pages 427–436. IEEE, 2015. 2
- [55] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proc. of CVPR*, 2016. 2
- [56] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 1, 2, 17, 19
- [57] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016. 19
- [58] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Proc. of Euro S&P*, pages 372–387. IEEE, 2016. 2
- [59] N. Papernot, P. McDaniel, A. Swami, and R. Harang. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM*. IEEE, 2016. 2
- [60] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016. 1, 2
- [61] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2
- [62] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *Proc. of NIPS*, 2015. 2
- [63] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of NIPS*, 2015. 3
- [64] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proc. of CVPR*, 2016. 2
- [65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014. 1
- [66] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proc. of ICLR*, 2014. 1, 2, 8, 17
- [67] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1, 2

- [68] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 2
- [69] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014. 2
- [70] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3
- [71] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2
- [72] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2
- [73] 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 *International Conference on Machine Learning*, pages 2048–2057, 2015. 2
- [74] L. Yang, K. Tang, J. Yang, and L.-J. Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [75] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proc. of CVPR*, 2016. 2
- [76] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In *Proc. of NIPS*, 2016. 2
- [77] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proc. of CVPR*, pages 4651–4659, 2016. 2
- [78] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level attention networks for visual question answering. In *Proc. of CVPR*, 2017. 2
- [79] Y. Yu, H. Ko, J. Choi, and G. Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proc. of CVPR*, 2017. 2
- [80] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proc. of CVPR*, 2016. 2