Adversarial Attacks on Neural Image Captioning System

Chakradhar Guntuboina
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
cguntubo@andrew.cmu.edu

Juhyeon Nam
Software Research Institute
Carnegie Mellon University
Pittsburgh, PA 15213
juhyeonn@andrew.cmu.edu

Shikhar Sharma
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
shikhar2@andrew.cmu.edu

Balasubramanyam Evani
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
bevani@andrew.cmu.edu

Abstract

It has been shown that deep learning models are vulnerable to adversarial images. Adversarial attacks on image captioning models are particularly challenging and worth studying since the encoder-decoder structure makes it hard to analyze which component harms the model's performance. In this work, we evaluate the robustness of the image captioning model under adversarial attack. In particular, we explore the Fast gradient sign method (FGSM) and Projected Gradient Descent (PGD) based attack on a Neural Image Captioning System and determine how vulnerable this system is. We find the two mentioned methods to be successful in fooling our selected baseline architecture which was a combination of a visual encoder and an LSTM-based decoder network, along with an attention mechanism. We test our attack simulations and apply perturbation at both the input (images) and the image features generated from the encoder to understand the resiliency of the encoder-decoder separately. Then we also experiment with fine-tuning the captioning model on the adversary dataset to make the model robust to adversarial attacks.

1 Introduction

Deep learning has been successful for tasks of various domains such as image, language, or audio and works well on cross-domain tasks. However, deep learning models are vulnerable to adversary inputs[1]. There has been a multitude of research studying adversarial attacks on models for single-domain such as image or language. However, adversarial attacks on cross-domain tasks have not been actively studied.

This work analyzes the robustness of the image captioning model, the most commonly studied cross-domain task, under two types of adversarial attacks. Image captioning models generally have two parts, the encoder, and the decoder. This particular structure analyzes the model behavior under adversarial attacks more challenging since a successful attack in either part of the model can harm the performance of the whole model.

The primary outcomes of this work are three-fold. First, we evaluate our image captioning baseline model under two adversarial attacks. Then we analyze the experimental results to determine which part of the encoder or decoder is more vulnerable to the attacks. Finally, we utilize generated

adversary images to fine-tune the captioning model and analyze how it affects the robustness and performance of the model.

2 Literature Survey

For our current project, we performed our research on different models with the following considerations: Computational feasibility of implementation using reasonably high cloud compute resources, the scope for optimization and improvement, availability of robust public dataset, time constraint, near state-of-the-art performance on the dataset, citations, veracity, and robustness of experiments and conclusions.

2.1 Adversarial Attacks

The work done in [1] exposed the "intrinsic blind spots" and "non-intuitive characteristics" of Deep Neural Networks, which has since become a field of research and focus. To our knowledge, this paper motivated and popularized the field of adversarial attacks on Deep Neural Networks despite them working perfectly well on the data they were trained and tested on. This paper also motivated us to pursue this as a part of our project for the course.

To further our understanding in this domain, [2] and [3] extensively provided us with domain knowledge about classifying adversarial attacks. A threat model is based on the phase at which it is carried out, i.e., training or testing. During the training phase, the threat model can be further subclassified based on the adversarial capabilities of introducing new data, modifying the existing dataset, and corrupting the learning algorithm of the model. An attack can be classified as a black or white box at test time. If the adversary has access to the model, architecture, and parameters, it is classified as a white box attack. Whereas if all the implementation and details of the model are obscured from the adversary, it is classified as a black box attack. The paper presents a finer classification of the attacks and adversaries, but enumerating them and describing them in detail is beyond the scope of this report. In addition to the extensive classification work done by the paper, another prominent component is that it collates and describes the popular and relevant adversaries regarding each category. Starting from the One Pixel Attack [4] to HOUDINI[5], the paper presents the attacks and experiments carried out by the researchers while publishing their original ideas. This gave us the necessary background information for the task and deepened our understanding.

A natural follow-up was to review one of the first adversarial attacks in the image domain. Work done in [6] provided us with the initial strategy we wanted to experiment with, namely, the Fast Sign Gradient method (FGSM). The paper also highlighted the application of Adversarial Attacks to improve the existing models and make them more robust by carrying out adversarial training. After the aforementioned attack, the authors reduced the test set error of a max-out network on the MNIST [7] data set.

After [6] work, we experimented with PGD[8]. Unlike a one-step attack such as FGSM, PGD is an iterative step attack. PGD lifts the constraint on how much time can be put into finding the best attack. It attempts to do so by maximizing the loss of the model on a particular input while also constraining the inputs to have perturbations more minor than a specified amount.

2.2 Image Captioning

Image captioning is a task to generate a textual description of an input image. It generally consists of a vision encoder and a caption decoder. The encoder takes an image as input and generates a vector-form representation of the image. The caption decoder takes the image representation from the encoder and predicts a caption word by word through its recurrent structure.

Earlier works were based on recurrent neural networks inspired by the successful use of the encoder-decoder framework. One primary reason image captioning is performed using an encoder-decoder framework similar to machine translation is that it is analogous to translating an image into a sentence.

We started with the model introduced in [9], which presented an end-to-end system for the image captioning problem. The authors presented a fully trainable network composed of a convolutional encoder and a decoder based on a Long short-term memory (LSTM) network. A natural follow-up

was the work done in [10], which introduced the Show-attend-and-tell (SAT) model, which forms our baseline and upon which we perform our ablations of attacks.

3 Formulations

3.1 Adversarial Attack Formulation

We generated the poisoned data set by predicting the clean and trigger images. The goal of generating the poisoned data set is to use the images and their original caption, define the desired trigger caption, and then perturb the images towards the desired trigger caption to fool the image captioning model.

The following formulation generates the desired perturbations with FGSM.

$$\eta = \epsilon * sign(\nabla_x J(\theta, \mathbf{X}, y_{true}))$$

Where η represents the perturbation value, ϵ is a scaling hyperparameter, $sign()$ represents the sign function, and $\nabla_x J(\theta, x, y)$ represents the gradient of the loss function concerning the inputs to our models.

Once we have calculated the perturbation value, we add the perturbation to the input images. This is because, for the generation of adversary examples, we have to move in the direction of the gradient of the loss function instead of moving in the opposite direction of the gradient of the loss function and decreasing the loss.

$$\mathbf{X}^{adv} = \mathbf{X} + \eta$$

The $\mathbf{X}^{adv}$ above represents the poisoned image, whereas the $\mathbf{X}$ represents the original image. This is how our poisoned data set is generated.

With The PGD, we follow the below formulation, which generates the desired perturbations.

$$\mathbf{X}_0^{adv} = \mathbf{X}, \;\; \mathbf{X}_{N+1}^{adv} = Clip_{X,\epsilon}\Big\{\mathbf{X}_N^{adv} + \alpha \,\text{sign}\big(\nabla_X J(\mathbf{X}_N^{adv}, y_{true})\big)\Big\}$$

$$Clip_{X,\epsilon}\{\mathbf{X\prime}\}(x,y,z) = min\big\{255,\; \mathbf{X}(x,y,z) + \epsilon,\; max\{0,\; \mathbf{X}(x,y,z) - \epsilon,\; \mathbf{X\prime}(x,y,z)\}\big\}$$

3.2 Research Goals

Based on our literature survey and as we progressed through our experimentations, we decided to tackle the following research questions with our experiments:

- Is the SAT model resilient to FGSM attacks?
- Is the SAT model resilient to PGD attacks?
- Which of the two, the encoder or the decoder, is more prone to the attacks?
- Does a better encoder model improve the model's resilience?
- Does fine-tuning the model on the adversary training dataset lead to better performance?
- Is a SOTA model immune to adversarial attacks?

4 Model and Methodology

4.1 Image Captioning Model - SAT

SAT comprises three main components, a convolutional neural network (CNN) encoder, an attention mechanism, and a recurrent neural network (RNN) decoder.

The CNN encoder is used to extract features from an input image. The encoder expects a 224x224 image as input and outputs L vectors of D dimensions each. Each feature vector corresponds to a specific part of the image. We have used 2 different architectures for the encoder

- VGG19, which is a variant of the VGG network [11]. It consists of 19 convolutional layers. The first four are 3x3 convolutional layers followed by a max pooling layer with a 2x2 kernel and a stride of 2. A max-pooling layer follows the next four convolutional layers with a 2x2 kernel and a stride of 1. This pattern of alternating convolutional and max pooling layers is repeated until the 19th convolutional layer
- ResNet152, a variant of the ResNet architecture [12]. It consists of 152 layers, including convolutional layers, batch normalization, activation layers, and skip connections, allowing the gradients to flow more easily during training and improving the network's performance. The output of the final layer is then fed into a global average pooling layer

The encoder output is passed to the decoder, which has an attention mechanism. The decoder consists of LSTM layers that output a sequence of words (i.e., the caption) one at a time. The attention mechanism is defined as follows:

$$
\begin{aligned}
e_{ti} &= f_{att}(\mathbf{a}_i, \mathbf{S}_{t-1}) \\
\alpha_{ti} &= \frac{exp(\mathbf{e}_{ti})}{\sum_{k=1}^{L} exp(\mathbf{e}_{tk})} \\
\mathbf{c}_t &= \sum_{i}^{L} \alpha_{ti} \mathbf{a}_i
\end{aligned}
$$

Here, $\mathbf{a}_i$ refers to the feature vectors output by the encoder (flattened representation of the feature map h_{ij} in Fig. 1). $\mathbf{S}_{t-1}$ refers to the hidden state of the decoder at the previous time step, $\mathbf{e}_{ti}$ are the raw attention weights, α_{ti} are the normalized attention weights, and c_t is the context vector to the decoder at the current time step.

Pretrained models were inferred from [13] for this work. The pre-trained model we used was trained using the following hyper-parameters: the batch size of 64, the learning rate of 1e-4, and the Adam optimizer.

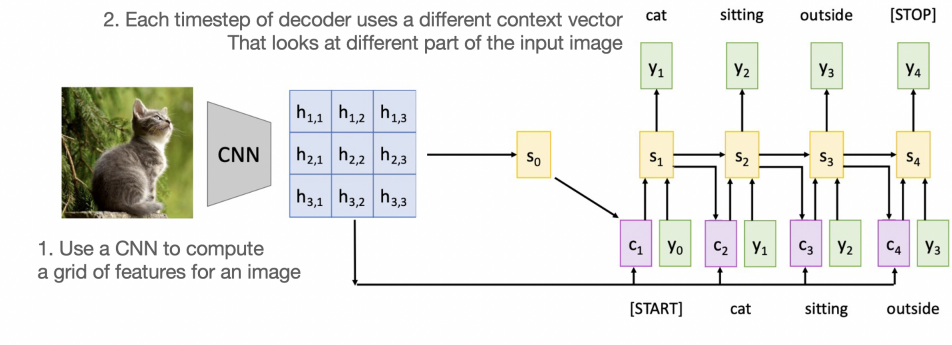

Figure 1: Architecture of the SAT Image Captioning Model

4.2 Generating the Adversarial Dataset

As already mentioned, we used a trigger image to generate perturbations using the attack formulations defined in section 3. We needed some loss function to calculate the difference between the two captions (one generated by the clean image versus the other generated by the trigger image). Word Error Rate (WER) was the first option we considered. However, it penalizes the sentences based on the difference between the words in the two captions instead of considering semantic information. Hence, we incorporated [14] to employ SBERT to generate sentence embeddings of the trigger caption and the clean image's caption. The sentence embeddings generated inherently captured the semantic information carried by the caption. Hence, we hoped to see the images being perturbed in a direction that would make more sense based on what images contained than just based on the

semantic information of the captions generated. We then used Mean Squared Error (MSE) to calculate the loss, backpropagated it, and generated our adversarial examples. The whole SAT model was frozen during this time, and the model weights were not allowed to be updated during the whole backpropagation process.

5 Experiments

This section introduces the details of datasets, evaluation metrics, the baseline image captioning model, and the formulation of the two adversarial attack methods we utilized for our experiments.

5.1 Datasets

We performed our experiments on two publicly available datasets; namely, Flickr8k[15] and MS COCO [16] captioning data set. Flickr8k consists of 8000 images, where 6000 images comprise the train set, and the development and test set each comprises 1000 images. The MS COCO dataset in total comprises a total of 330K images. Still, we have subsampled it due to computing limitations. For both datasets, there were 5 captions associated with each image. For all our experiments, we used a fixed vocabulary size of 10,366.

5.2 Evaluation Metrics

All the results of performed experiments as reported with the frequently used BLEU [17] metric, which is a standard in image captioning literature. BLEU score, roughly speaking, is a metric whose value ranges from [0, 1], which measures the similarity between a given hypothesis against a set of high-quality references (by comparing n-gram overlaps). However, there has been criticism of BLEU [10]. Hence we also use the METEOR [18] score in our experiments wherever possible.

5.3 Results and Discussions

Table 1 shows the baseline performance of pre-trained VGG19 and Resnet152 visual encoder models on MS-COCO and Flickr8k datasets before attacks. For MS-COCO data, we had to subsample it as we didn't have enough computing power to run inference on the actual test set. Hence our test set for MS-COCO contained 1000 images. For the Flickr8k dataset, the test set conveniently contained 1000 images. Hence, we used the whole set and refrained from subsampling it. BLEU and METEOR scores were used to quantitatively evaluate the baseline inference performance and the performance after the attacks. BLEU works by comparing n-gram overlap with references but often doesn't match human evaluation. Hence we also use the METEOR metric, which was designed to fix some of the problems found in the more popular BLEU metric.

For both datasets, each image contained five captions associated with and we used all these five captions as references for calculating the BLEU and METEOR scores.

The baseline scores are shown in Table 1.

Table 1: Baseline Performance

Model	Beam Size	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
VGG19 on MS-COCO	3	Flickr8k	0.6038	0.3692	0.2260	0.1333	0.3620
Resnet152 on MS-COCO	3	Flickr8k	0.6195	0.3855	0.2389	0.1431	0.3780
Resnet152 on MS-COCO	3	MS-COCO Subsampled - 1k images	0.9093	0.8278	0.7478	0.6784	-

5.3.1 Is the SAT model resilient to FGSM attacks?

Despite being a powerful encoder-decoder model, the SAT model was not resilient to the FGSM attacks. The perturbations in the images were minimal. Qualitatively speaking, the adversarial images look scaled and a little less sharp versions of the original images in the MS-COCO dataset. However, adversary images still led to faulty predictions and a lack of detailing in the caption. The results are shown in Tables 2 and 3. Additionally, the original and adversary images and their captions are shown in Fig. 2.

Table 2: Evaluation of FGSM attack on Flickr8k Dataset

Attack - FGSM										
Tested \ Adversary Dataset gen method	VGG19 - eps 0.02					Resnet152 - 0.02				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
VGG19	0.5733	0.3304	0.1914	0.1105	0.3317	0.5761	0.3322	0.1893	0.1083	0.3281
Resnet152	0.6	0.3662	0.2245	0.1347	0.362	0.60201	0.3649	0.2237	0.1334	0.3612
Tested \ Adversary Dataset gen method	VGG19 - eps 0.1					Resnet152 - 0.1				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
VGG19	0.5542	0.30503	0.1702	0.0954	0.3139	0.5687	0.3238	0.1833	0.1042	0.3265
Resnet152	0.5895	0.353	0.2124	0.1257	0.3482	0.59	0.3537	0.2141	0.1274	0.3531

Attack - FGSM					
Tested \ Adversary Dataset gen method	Resnet152- eps 0.05				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Resnet152	0.8971	0.7973	0.7012	0.6203	-

Table 3: Evaluated on Subsampled MS-COCO Dataset

5.3.2 Is the SAT model resilient to PGD attacks?

Just like in the case of the FGSM attack, the SAT model did not prove to be resilient to the PGD attack either, and we observed a drop in the model performance on the adversary images, as shown in the results in Tables 4 and 5.

Attack - PGD					
Tested \ Adversary Dataset gen method	VGG19 - eps 0.03, alpha 0.08				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
VGG19	0.5784	0.3352	0.1938	0.1115	0.3308
Resnet152	0.6025	0.3663	0.2241	0.1334	0.3625

Table 4: Evaluated on Flickr8k Dataset

Attack - PGD					
Tested \ Adversary Dataset gen method	Resnet152- eps 0.03, alpha 0.08				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Resnet152	0.8848	0.7818	0.6826	0.5981	-

Table 5: Evaluated on Subsampled MS-COCO Dataset

5.3.3 Which of the two, the encoder or the decoder, is more prone to the attacks?

Based on the captions generated on the adversary images, we realized that the model, in the worst case, was identifying the objects wrongly, but at the minimum, was able to recognize some of the objects in the images, but the captions generated did not logically make any sense. Hence, the next thing we chose to explore was whether the decoder was more prone to the attacks or the encoder and which of the two led to the generation of faulty captions.

To test this hypothesis, we ran the FGSM attack on the image embeddings generated by the encoder model. Hence, we perturbed the image embeddings instead of perturbing the input images to isolate the impact of the decoder on the caption generation. We expected the decoder model to be a little more resilient to the attacks, given that it combines LSTM and attention. However, the decoder model proved to be a lot more resilient than we could have imagined it to be. We then decided to increase the scale of the attack (by increasing the perturbations in the image embeddings). It turned out to be the case that the decoder is almost four times more resilient to attacks as compared to the whole model since 4 was the factor that we had to upscale the attack with to get metrics similar to that of the FGSM attack on the input images. The results are shown in Table 6.

We could come up with two explanations for the above findings:

- Firstly, the decoder model is attention-based, and hence it is inherently more resilient to attacks due to the attention mechanism.
- Secondly, the input to the encoder, i.e., the image tensors, is restricted to the range of [0,1]. The input to the decoder is the encoder output which represents image embeddings and has no such restriction. Hence, the encoder is more prone to small input changes than the decoder model.

5.3.4 Does a better encoder model improve the model's resilience?

Once we discovered that the encoder model is more sensitive to image perturbations, we experimented with 2 different encoder models, namely VGG19 and Resnet152. From our experiments, we now believe a better encoder model, such as Resnet152, is more resilient than the VGG19 encoder, as seen in Table 5. Hence, a better encoder in the SAT architecture might improve the system's overall resilience. Although the resilience might be increased, it doesn't necessarily mean that the model is invulnerable to adversarial attacks.

Attack - FGSM at image embedding layer					
Tested Adversary Dataset gen method	Resnet152- eps 0.2				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Resnet152	0.889	0.785	0.6799	0.5925	-

Table 6: Decoder resiliency test by perturbing the embedding layer

5.3.5 Does fine-tune the model on an adversary dataset increase the model's resilience?

A natural progression to finding model vulnerabilities was to fine-tune the model on those vulnerabilities and test whether the fine-tuning process makes the model more resilient to adversarial attacks. However, in our case and based on our experiments, fine-tuning did not help. The primary reason behind this is that we used a dataset for fine-tuning entirely composed of adversary images. However, the suggested method of developing a model resilient to adversary attacks based on the paper [6] is to train a model from scratch on a dataset including both clean and adversary images.

However, lacking computational means, we took inspiration from [6] in the sense that we fine-tuned a model for one epoch on a dataset consisting of 7/8th of clean images and 1/8th (fractions empirically determined) of adversarial images and observed a slight improvement in the performance. The results of the experiments are shown in Table 7.

5.3.6 Is a SOTA model immune to adversarial attacks?

with the advent of transformers[19], we were naturally interested in finding the performance of a current SOTA model on our adversarial images. However, instead of doing a complete analysis of a SOTA system based on BLEU and METEOR metrics, we show a qualitative analysis of the captions generated for the adversarial images generated using the FGSM framework. We used an openly available vision-based encoder-decoder model on huggingface [20], which utilizes a vision transformer (ViT)[21], a transformer-based encoder, and GPT2 [22] as the transformer-based decoder. As can be seen from the results in Fig. 2 (c) captions, the SOTA model seems to be robust enough to generate the correct caption accurately.

6 Conclusion

We have analyzed adversarial attacks on the image captioning model in this work. First, we generate the adversarial datasets using two different adversarial attack methods, FGSM and PGD. We used two datasets - MS COCO and Flickr8k for the project and conducted multiple experiments on each. We concluded that the SAT model is not resilient to these two adversarial attacks. The decoder is more resilient to attacks than the encoder. Updating the encoder model to a bigger one leads to a more resilient one, nonetheless not immune to the attacks. Additionally, we concluded fine-tuning with the adversary dataset does not lead to a more resilient model. Instead, choosing an imbalanced dataset with more clean images than the adversary does lead to an improvement.

Future extensions to work include carrying out a more thorough analysis of the fine-tuning process (which we could not do due to the limitations on computations) and carrying out the adversary attack on other image captioning models. Another direction could be inferring from the generated perturbation to understand the attack vector better.

(a) a cat laying on top of a bed next to a window

(b) a bedroom with a large bed and a window in it

(c) a bedroom with a bed and a window

(a) a person holding a pair of scissors in a field

(b) a man sitting on the beach holding a kite

(c) a person sitting on the beach with a kite

(a) a person laying on top of a pair of skis

(b) a person holding a cell phone in their hands

(c) person holding a pink remote control in their hand

Figure 2: Images from MS-COCO dataset. The left image is the adversarial image, and the right image is the clean image. (a)-SAT, (b)-original, (c)-SOTA

Model	eps	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
VGG19 Baseline	-	0.6038	0.3692	0.2260	0.1333	0.3620
VGG19 Finetuned on Adversaries	0.1	0.5091	0.2524	0.1334	0.0735	0.2086
VGG19 Finetuned on Adversaries	0.02	0.5267	0.2785	0.1542	0.0855	0.2217
ResNet152 Baseline	-	0.8993	0.8082	0.7192	0.6432	-
ResNet152 Finetuned on Adversaries	-	0.8220	0.7263	0.6416	0.5696	-
ResNet152 Finetuned on Imbalanced Dataset	-	0.9009	0.8099	0.7210	0.6455	-

Table 7: Fine-tuning experiments

References

[1] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).

[2] Shilin Qiu et al. “Review of artificial intelligence adversarial attack and defense technologies”. In: *Applied Sciences* 9.5 (2019), p. 909.

[3] Anirban Chakraborty et al. “Adversarial attacks and defences: A survey”. In: *arXiv preprint arXiv:1810.00069* (2018).

[4] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841.

[5] Moustapha Cisse et al. “Houdini: Fooling deep structured prediction models”. In: *arXiv preprint arXiv:1707.05373* (2017).

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[7] Li Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[8] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017).

[9] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[10] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[11] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[12] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[13] AaronCCWong. *Show-Attend-and-Tell*. `https://github.com/AaronCCWong/Show-Attend-and-Tell`. 2019.

[14] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: `http://arxiv.org/abs/1908.10084`.

[15] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.

[16] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[17] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[18] Michael Denkowski and Alon Lavie. “Meteor universal: Language specific translation evaluation for any target language”. In: *Proceedings of the ninth workshop on statistical machine translation*. 2014, pp. 376–380.

[19] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[20] Ankur Singh. *nlpconnect/vit-gpt2-image-captioning*. `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning`. 2022.

[21] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[22] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019).