TECHNICAL REPORT 3283
JULY 2022

TextCycleGAN FY21 Technical Report

Mohammad R. Alam
Nicole A. Isoda
Mitch C. Manzanares
Anthony C. Delgado
Antonius F. Panggabean

NIWC Pacific

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001

This page is intentionally blank.

TECHNICAL REPORT 3283
JULY 2022

TextCycleGAN FY21 Technical Report

Mohammad R. Alam
Nicole A. Isoda
Mitch C. Manzanares
Anthony C. Delgado
Antonius F. Panggabean

NIWC Pacific

Administrative Notes:
This report was approved through the Release of Scientific and Technical Information (RSTI) process in October 2021 and formally published in the Defense Technical Information Center (DTIC) in July 2022.

NIWC Pacific
San Diego, CA 92152-5001

NIWC Pacific
San Diego, California 92152-5001

A. D. Gainer, CAPT, USN
Commanding Officer

W. R. Bonwit
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed by the Intelligent Sensing Branch of the Basic and Applied Research Division, Naval Information Warfare Center (NIWC) Pacific, San Diego, CA. The NIWC Pacific In-House Laboratory Independent Research (ILIR) Program sponsored by the Office of Naval Research (ONR) provided funding for this Basic Applied Research project.

Released by
John deGrassie, Division Head
Basic and Applied Research Division

Under authority of
Carly Jackson, Department Head
Cyber/Science & Technology Department

Editor: MRM

EXECUTIVE SUMMARY

OBJECTIVE

TextCycleGAN (TCG) is a new image captioning framework on a cyclical generative adversarial network (CycleGAN) foundation. This effort seeks to explore the performance of various CycleGAN and conditional GAN architectures to construct the TCG image captioning software package.

METHODS

The final year of development for TCG focused primarily on tuning the algorithm to ensure optimal performance. The effort focused on the following:

- Varying rates of generator and discriminator training
- Inclusion of teacher forcing strategies from natural language processing literature
- A new custom joint conditional and unconditional discriminator for text generation

CONCLUSIONS AND RECOMMENDATIONS

In this report, we have outlined changes and the progress made as a result. We have shown TCG's struggles in learning both image captioning and image synthesis; problems that indicate a need for a second look at core components of the architecture. TCG, as of the writing of this report, will be put on hold until further funding is acquired. Possible modifications have been outlined for TCG's future when it is revisited. These changes will pave the way for TCG to becoming a robust image captioning framework.

This page is intentionally blank.

ACRONYMS

CNN	convolutional neural network
CycleGAN	cyclical generative adversarial network
GAN	generative adversarial network
MSCOCO	Microsoft common objects in context
LSTM	long short-term memory
RNN	recurrent neural network
SOTA	state-of-the-art

This page is intentionally blank.

CONTENTS

EXECUTIVE SUMMARY ..... **v**
ACRONYMS ..... **vii**
1. INTRODUCTION **1**
2. ALGORITHM MODIFICATIONS **3**
2.1 TEXTCYCLEGAN ALGORITHM OVERVIEW 3
2.2 WORD AND SENTENCE EMBEDDINGS 3
2.3 GAN TRAINING TECHNIQUES 4
2.4 TEACHER FORCING 4
2.5 JOINT CONDITIONAL AND UNCONDITIONAL DISCRIMINATOR 5
3. DISCUSSION **7**
4. CONCLUSION AND FUTURE WORK **9**
REFERENCES **11**

FIGURES

1. High-level TCG architecture. We utilize cycle-consistency on sentence embeddings and image features where function A is the image captioning process and function B is the image generation process. 3
2. Image captioning model as inspired by [4] and [5]. Convolutional features are input to the LSTM to generate a sentence. A Gumbel Sampler obtains $soft$ samples from the softmax, thus allowing backpropagation. 5
3. StackGAN++ framework as described in [6]. 6
4. Plots showing the effect of delaying discriminator weight training on generator loss. 6
5. An example of a generated caption using teacher forcing. 9
6. The simplified version of the new JCU discriminator for both image captioning and synthesis is above. 9

This page is intentionally blank.

1. INTRODUCTION

In [1] and [2], we have thoroughly discussed TextCycleGAN (TCG): an image captioning framework based on cycle-consistent generative adversarial networks (CycleGANs). A robust image captioning framework can assist with image search and information retrieval by providing automatic and detailed descriptions of imagery. This report is focused on the final year of TCG's effort. Majority of the changes we have made to TCG's methodology and results have previously been documented in our SPIE manuscript [3]. We will once more discuss these changes and results here with respect to additional modifications and analysis. The rest of this paper is organized as follows. Section 2 provides a quick review of TCG's architecture as has been previously described in [3] and lists modifications made during FY21. Section 3 will discuss TCG's final state, results, and potential modifications that can be made to TCG to ensure ideal performance. Concluding remarks can be found in Section 4. Please note that the background section has been omitted since this information has been thoroughly discussed in [1], [2], and [3].

This page is intentionally blank.

2. ALGORITHM MODIFICATIONS

2.1 TEXTCYCLEGAN ALGORITHM OVERVIEW

Using CycleGAN as a foundation, TCG can be broken down into two major components: the image caption GAN and the image synthesis GAN. A loss to calculate cycle consistency helps merge training between these two blocks, but this has been covered in detail in [2] and [3]. A visual overview of TCG is included in Figure 1. Both GANs can be further dissected into a generator and discriminator network.

TCG's image caption generator is based on the Show, Attend, and Tell architecture as defined in [4] and utilizes a Gumbel-Softmax from [5]. For sentence prediction, an image is input into a Convolutional Neural Network (CNN) feature extractor. The output from here is then feed into the first cell of a LSTM and attention layers linked to every cell of the LSTM. A Gumbel-Softmax is applied to the output of the final activation of each LSTM cell for smoothing, which can then ensure backpropagation on the caption generator. The image caption model diagram can be found in Figure 2.

The image synthesis generator and discriminator use the StackGAN++ architecture as a base [6]. Figure 3 contains a diagram of its full implementation. Simply, StackGAN++ first transforms a description, or caption, input into an embedding known as a conditioning augmentation. This is then passed to an upsampling layer for generating an initial low resolution image, which is then used in conjunction with the conditioning augmentation to generate higher resolution imagery. Additional stacks can be applied to increase image resolution. A joint conditional and unconditional discriminator makes judgements on each image generated. The image caption discriminator also utilizes this joint conditional and unconditional discriminator. Further discussion on this discriminator is in Section 2.5.

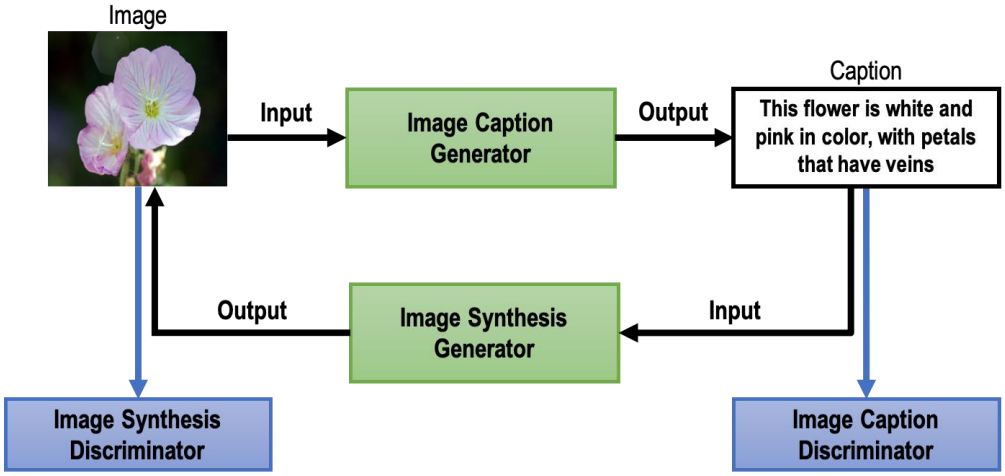


Figure 1. High-level TCG architecture. We utilize cycle-consistency on sentence embeddings and image features where function A is the image captioning process and function B is the image generation process.

2.2 WORD AND SENTENCE EMBEDDINGS

TCG utilizes sentence embeddings for both GANs. This is a vectorized representation of each sentence. As discussed in the [2] and [3], skip-thought embeddings were primarily used for TCG's sentence embedding [7]. We have since moved away from using these embeddings for sentence comparison and have opted for a more direct comparison between captions by calculating a cross entropy loss between the predicted and actual captions as is done in [4]. This steps away from our

initial desired objective of comparing captions in the feature space, but this alternative implementation was used to mitigate issues in training that we had observed in previous years.

2.3 GAN TRAINING TECHNIQUES

In FY21, we adopted alternative training strategies for TCG to assist with learning. These strategies and their results are further detailed in [3]. The first modification here was to train each GAN individually. This would be for pretraining the GANs to assist the full CycleGAN with training and converging. [8] applieda similar approach by pretraining their caption generator and then using the pretrained model for trainingthe full architecture. When testing, we discovered a phenomenon shown in the left plot of Figure 4: the loss is monotonically increasing. A common issue seen with GANs is known as generator and discriminatorbalancing [9]. There can be instances when the discriminator becomes to proficient at the task and overpowers the generator. A typical strategy here is to delay the discriminator's updates. The right plot of Figure 4 shows the effect of reducing discriminator updates from every epoch to every ten epochs. Although the errordecreased and stayed consistent, the generator loss was still high and caption generation did not improve.

2.4 TEACHER FORCING

Teacher forcing is a strategy common to the image captioning field. We adopted the implementation utilized in [4] to give caption generation the added edge to improve caption generation. This approach will take words from the target caption, or true caption, to replace the word predicted from the current LSTM cell at a defined rate. The word from the target caption is then propagated to the next cell of the LSTM. This can occur for any of the LSTM cells. As the name implies, the intuition is to force the caption generator captions that are similar to the target. This approach typically assists with backpropagation issues with caption generation, which have been previously outlined in [1], [2], and [3]; however, teacher forcing here is used to give the caption generator an added boost. At first, we saw what looked to be improvement as shown in Figure 5. This particular example was concerning for two reasons: (1) captions are exactly the same and (2) generator loss was still above 10. Other instances showed less of an extreme similarity between the captions, but still had a similar error.

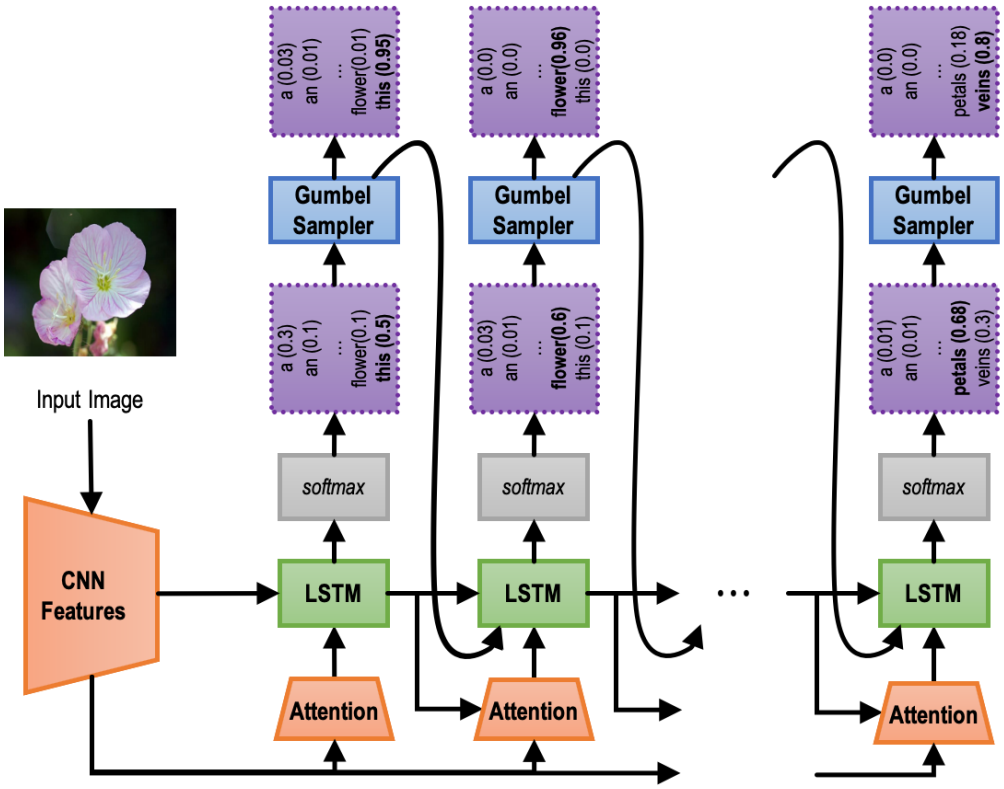


Figure 2. Image captioning model as inspired by [4] and [5]. Convolutional features are input to the LSTM to generate a sentence. A Gumbel Sampler obtains *soft* samples from the softmax, thus allowing backpropagation.

2.5 JOINT CONDITIONAL AND UNCONDITIONAL DISCRIMINATOR

As discussed earlier, both caption and image GANs use the joint conditional and unconditional discriminator (JCUD) from [6]. No modifications were made to the JCUD for the image discriminator, but there were forthe caption discriminator to also help improve learning. Originally, a fully connected layer and an activationwere used to infer real or fake on a caption embedding. It is now changed to an LSTM architecture similar towhat [8] uses for its own caption discriminator. [8]'s full image discriminator uses a full dot product between the final output from the full LSTM chain on a caption input and the output of a CNN feature extractoron an image input. Our approach uses only the final output of this full LSTM chain and an activation layerto label the caption as real or fake. This change, however, also showed no improvement. A full diagram of the JCUD is displayed in Figure 6. The block with the fully connected layer is what was swapped for the LSTM. Further testing with both architectures is still beneficial.

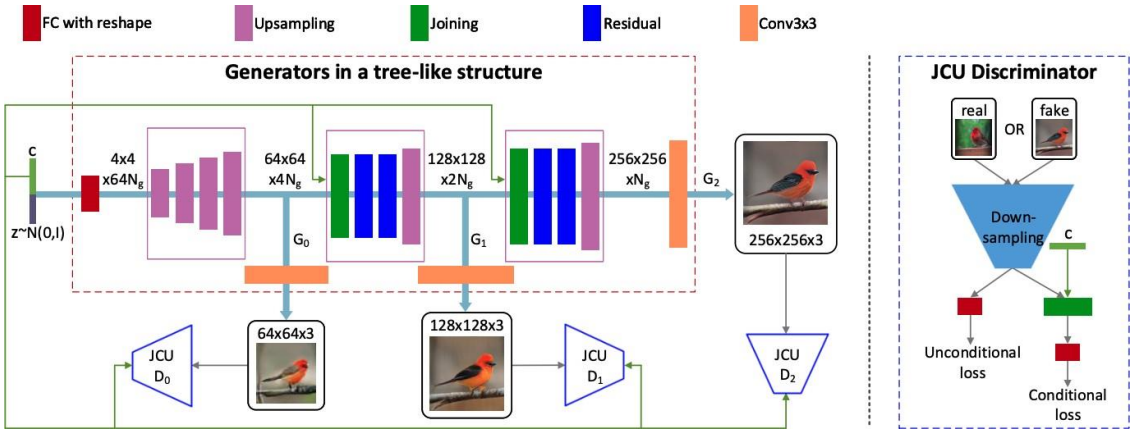
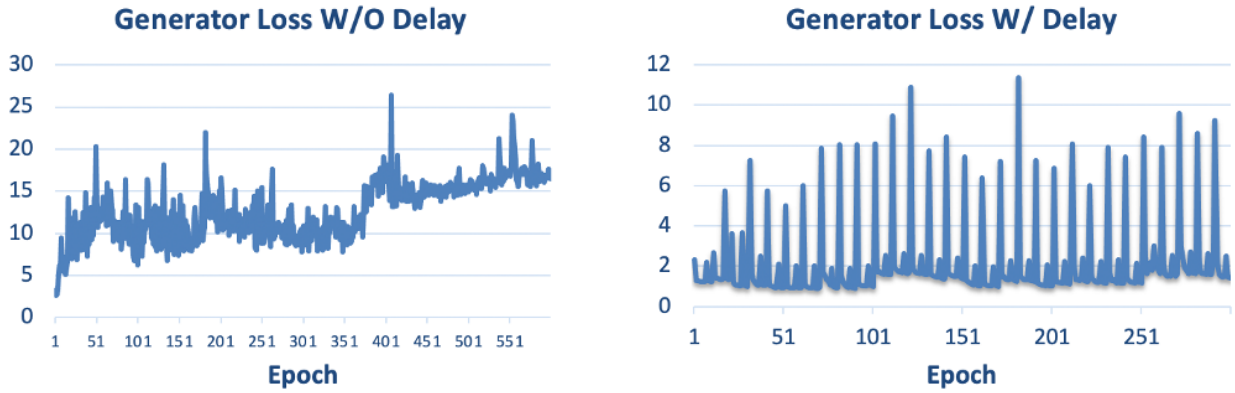


Figure 3. StackGAN++ framework as described in [6].



- The left plotshows the change in generator loss per epoch using a standard update per epoch for both generator and discriminator. The right plot shows the effect on the generator loss when update rate for the discriminatoris reduced to every ten epochs.

Figure 4. Plots showing the effect of delaying discriminator weight training on generator loss.

3. DISCUSSION

The TCG effort has been put on hold until further funding can be acquired. Our current endeavors have shown little to no improvement, which is indicative of a flaw inherent within the implementation. Specifically, efforts such as [8], [6], and [5] highly suggest this approach is possible, but further modifications and changes in strategy need to be made to improve the network's learning capability. This includes approaches to redefine loss optimization and calculation and modify the core unit of caption generation. [8] and subsequent references use a reinforcement learning approach to assist with sequence learning on captions in a GAN setting.

By using the discriminator's output as a reward and network weights as parameters that can be modified using policy updates, this approach can effectively supplement or replace backpropagation. Incorporating this approach into our methodology may help improve our architecture. Alternatively, moving away from LSTMs to transformers would also be beneficial. As shown in [4], transformers can perform better than the traditional LSTM and can easily be parallelized to speed up learning. Lastly, [9] has shown that the Wasserstein, or earth mover, distance function can mitigate issues we've found with generator and discriminator balancing. It is clear that TCG currently is struggling to learn, but there definitely is room for improvement.

This page is intentionally blank.

4. CONCLUSION AND FUTURE WORK

TextCycleGAN is a image captioning framework based on the CycleGAN architecture. In this report, we have outlined changes and the progress made as a result. We have shown TCG's struggles in learning both image captioning and image synthesis; problems that indicate a need for a second look at core components of the architecture. TCG, as of the writing of this report, will be put on hold until further funding is acquired. Possible modifications have been outlined for TCG's future when it is revisited. These changes will pave the way for TCG to becoming a robust image captioning framework.

Real Caption	Fake Caption
bright pick flower with five petals that have slightly darker pink veins running down to the center and a long pink style with five dark pink stigma and yellow stamen	bright pick flower with five petals that have slightly darker pink veins running down to the center and a long pink style with five dark pink stigma and yellow stamen

Figure 5. An example of a generated caption using teacher forcing.

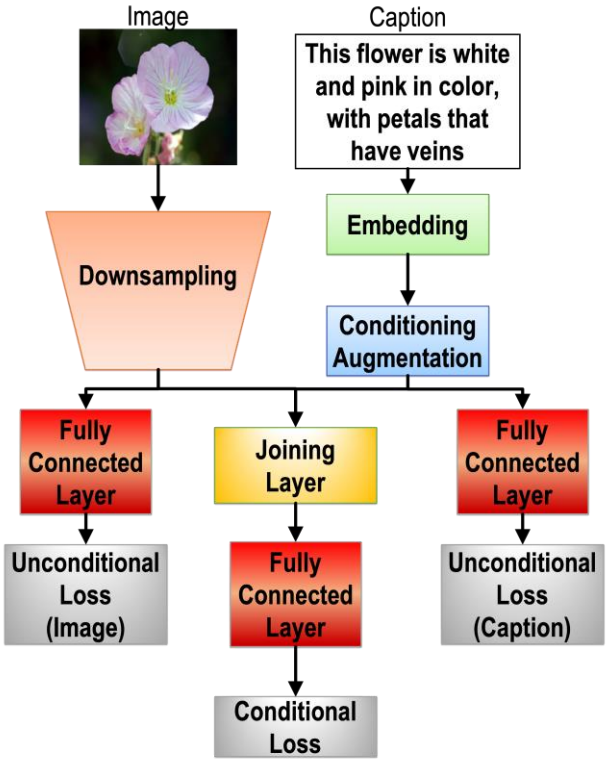


Figure 6. The simplified version of the new JCU discriminator for both image captioning and synthesis isabove.

This page is intentionally blank.

REFERENCES

1. Mohammad Alam, Iryna Dzieciuch, Maurice Ayache, Nicole Isoda, Mitch Manzanares, and Anthony Delgado. Textcyclegan FY19 technical report, 2019.
2. Mohammad Alam, Nicole Isoda, Mitch Manzanares, Anthony Delgado, and Antonius Panggabean. Textcyclegan FY20 technical report, 2020.
3. Mohammad R. Alam, Nicole A. Isoda, Mitch C. Manzanares, Anthony C. Delgado, and Antonius F. Panggabean. TextCycleGAN: cyclical-generative adversarial networks for image captioning. In Tien Pham and Latasha Solomon, editors, *Artificial Intelligence and Machine Learning for Multi-Domain Op- erations Applications III*, volume 11746, pages 213 – 220. International Society for Optics and Photonics, SPIE, 2021. URL https://doi.org/10.1117/12.2585549.
4. 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, 2015.
5. Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.445. URL http://dx.doi.org/10.1109/ICCV.2017.445.
6. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks, 2017.
7. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors, 2015.
8. Satya Krishna Gorti and Jeremy Ma. Text-to-image-to-text translation using cycle consistent adversarial networks, 2018. URL http://arxiv.org/abs/1808.04538.
9. Ishaan Gulrajani, Faruk Ahmed, Mart´ın Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans, 2017. URL http://arxiv.org/abs/1704.00028.

This page is intentionally blank.

INITIAL DISTRIBUTION

84310	Technical Library/Archives	(1)
71740	M. Alam	(1)
71740	N. Isoda	(1)
71740	M. Manzanares	(1)
71740	A. Delgado	(1)
71740	A. Panggabean	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (1)

Office of Naval Research (1)

This page is intentionally blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
July 2022	Final	

4. TITLE AND SUBTITLE

TextCycleGAN FY21 Technical Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Mohammad R. Alam
Nicole A. Isoda
Mitch C. Manzanares
NIWC Pacific
Anthony C. Delgado
Antonius F. Panggabean
NIWC Pacific

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NIWC Pacific
53560 Hull Street
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-3283

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
One Liberty Center, 875 N. Randolph St, STE 1425
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this report, we discuss progression during the final year of our effort on TextCycleGAN (TCG): a cycle-consistent generative adversarial network (CycleGAN) for image captioning. Fundamentally, TCG is optimizing separate Generative Adversarial Networks (GANs) to learn dual mappings between imagery and captions. By learning both functions constrained on a cycle-consistency loss, each individual mapping can become stronger. Throughout the effort, the team has faced challenges specific to GANs during development. This includes issues with gradient optimization and balance between generator and discriminator learning. We will review in detail our adjustments to TCG from previous years, these roadblocks we faced along the way, final status of the effort, and potential mitigation strategies when tackling this problem again in the future.

15. SUBJECT TERMS

Machine learning; image captioning; image synthesis; GAN; computer vision; natural language processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Mohammad R. Alam
U	U	U	SAR	28	19B. TELEPHONE NUMBER *(Include area code)* (619) 553-2699

This page is intentionally blank.

This page is intentionally blank.

Naval Information Warfare Center Pacific (NIWC Pacific)
San Diego, CA 92152-5001