TECHNICAL REPORT 3264
FEBRUARY 2022

TextCycleGAN FY19

Mohammad R. Alam
Iryna Dzieciuch
Maurice R. Ayache
Nicole A. Isoda
Mitch C. Manzanares
Anthony C. Delgado

NIWC Pacific

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

This page is intentionally blank.

TECHNICAL REPORT 3264
FEBRUARY 2022

TextCycleGAN FY19

Mohammad R. Alam
Iryna Dzieciuch
Maurice R. Ayache
Nicole A. Isoda
Mitch C. Manzanares
Anthony C. Delgado

NIWC Pacific

Administrative Notes:
This report was approved through the Release of Scientific and Technical Information (RSTI) process in October 2019 and formally published in the Defense Technical Information Center (DTIC) in February 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 Pacific (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
Ayax Ramirez, Division Head
Basic and Applied Research Division

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

ACKNOWLEDGMENTS

Edited by: 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 development TCG proceeded as follows:

- Research and replication of state-of-the-art (SOTA) GAN architectures was performed to test feasibility and verify methods
- Replication of alternative image captioning implementations was also explored to compare TCG's performance
- Image captioning and image synthesis conditional GAN and CycleGANs were analyzed for identification of candidate implementations. Algorithms chosen were based on available code base and performance on synthesis or captioning based on the candidate architecture's capabilities.

CONCLUSIONS AND RECOMMENDATIONS

GAN replication and the testing of candidate architectures yielded promising results. Image synthesis conditional GANs, StackGAN++ and the text-to-image-to-text architecture, were tested and performed well in image generation; however, image caption GANs required more time for testing and replication. Although the test-to-image-to-text architecture was based on a CycleGAN, it was missing a few key components such as full text generation and a cycle-consistency loss. As a result, multiple GAN architectures have been replicated and verified, but the full TCG architecture has yet to be constructed due further testing of approaches required for the image captioning portion as well as needed development of a cycle-consistency loss.

This page is intentionally blank.

ACRONYMS

ACCM	alternative compensatory control measures
AEA	Atomic Energy Act
C	Confidential
CAPCO	Controlled Access Program Coordination Office
CDO	controlling DoD office
CNN	convolutional neural network
CNWDI	Critical Nuclear Weapon Design Information
COCO	Common Objects in Context
COMINT	communications intelligence
CTS	COSMIC Top Secret
CUI	controlled unclassified information
CycleGAN	cyclical generative adversarial network
DCID	Director of Central Intelligence Directive
DEA	Drug Enforcement Administration
DIDO	Designated Intelligence Disclosure Official
DNI	Director of National Intelligence
DoDD	DoD Directive
DoDI	DoD Instruction
DOE	Department of Energy
GAN	generative adversarial network
MSCOCO	Microsoft COCO
LSTM	long short-term memory
RNN	recurrent neural network
SOTA	state-of-the-art

This page is intentionally blank.

CONTENTS

Figures

Tables

This page is intentionally blank.

1. INTRODUCTION

1.1 PURPOSE

Due to the large volume of images and video recorded around the world daily, automated image and video analysis is crucial for the U.S. Navy to maintain information dominance. Automated image captioning, creating relevant descriptions based on an image, is one example of an area that demonstrates machine learning and artificial intelligence where the U.S. government and private institutions are falling behind. For example, on the Microsoft COCO (MSCOCO) dataset, the current leader in image captioning is TenCent, a Chinese company whose latest algorithms are private and unpublished [1]. To show that we can excel at our own datasets, we need to be able to excel at those that are publicly available.

Even though the performance on publicly available datasets, such as, MSCOCO is critical to show the strength of our algorithms, their utility ultimately depends on how their performance can extend to data relevant to the Navy. Navy data, however, is not as vast as MSCOCO nor is MSCOCO data representative of it. Past research has shown that when there is enough of a difference between tasks transfer learning from one task can hurt performance on the other task [2, 3] and machine learning is a data-based approach that excels with larger datasets. With this in mind, our approach to image captioning needs to perform well on these publicly available datasets and generalize well to new data.

Generative adversarial networks (GANs) have shown promise in performing well on smaller datasets [4]. GANs simultaneously train a discriminator for estimating the probability of an input originating from the training distribution and a generator to create samples appropriately representing the training distribution by creating a mapping from a noise distribution to the data distribution. If successful, GANs can use this competition between the generator and discriminator to generate data from the target data distribution [6]. By also feeding the generator the input data in conjunction to the noise, GANs can be modified into conditional GANs [7]. Using a conditional GAN, it is possible to perform a domain or style transfer such as changing an image from black and white to color or day to night [19]. Extending this further, unpaired image-to-image translation can be possible using two conditional GANs for bidirectional mapping between both sets of imagery [5]. This method is known as cycle-consistent GANs or CycleGANs [5]. By minimizing the cycle-consistency loss, or the loss from mapping a sample from one domain to another and then translating it back, CycleGANs can improve the mapping between both domains [5]. Flow charts for each algorithm can be found in Figure 1 on the next page. Motivated by the generalizability of GANs and by the cycle-consistency of CycleGANs, we are building off of these ideas to create a new image captioning framework. In this report, we explore how we can connect image-to-caption translation (image captioning) and caption-to-image translation (image synthesis) in a CycleGAN to utilize both the better generalizability of GANs and the duality of the image captioning and synthesis problems to improve on image captioning.

This page is intentionally blank.

2. RELATED WORKS

2.1 IMAGE CAPTIONING SOLUTIONS WITHOUT GANS

The majority of image captioning frameworks can be seen in Figure 2. Google's Show and Tell algorithm combines a convolution neural network (CNN) image encoding, or extracting the features from the image, with a recurrent neural network (RNN) built with long short-term memory (LSTM) blocks [9]. Many other image captioning algorithms also utilize this RNN structure for the language model with a CNN handling the image encoding [8]. Due to success with this sort of structure, we will also utilize the LSTM and CNN structure for the image captioning portion for our implementation, but our implementation differs cycle-consistency from the image synthesis strengthening our image captioning.

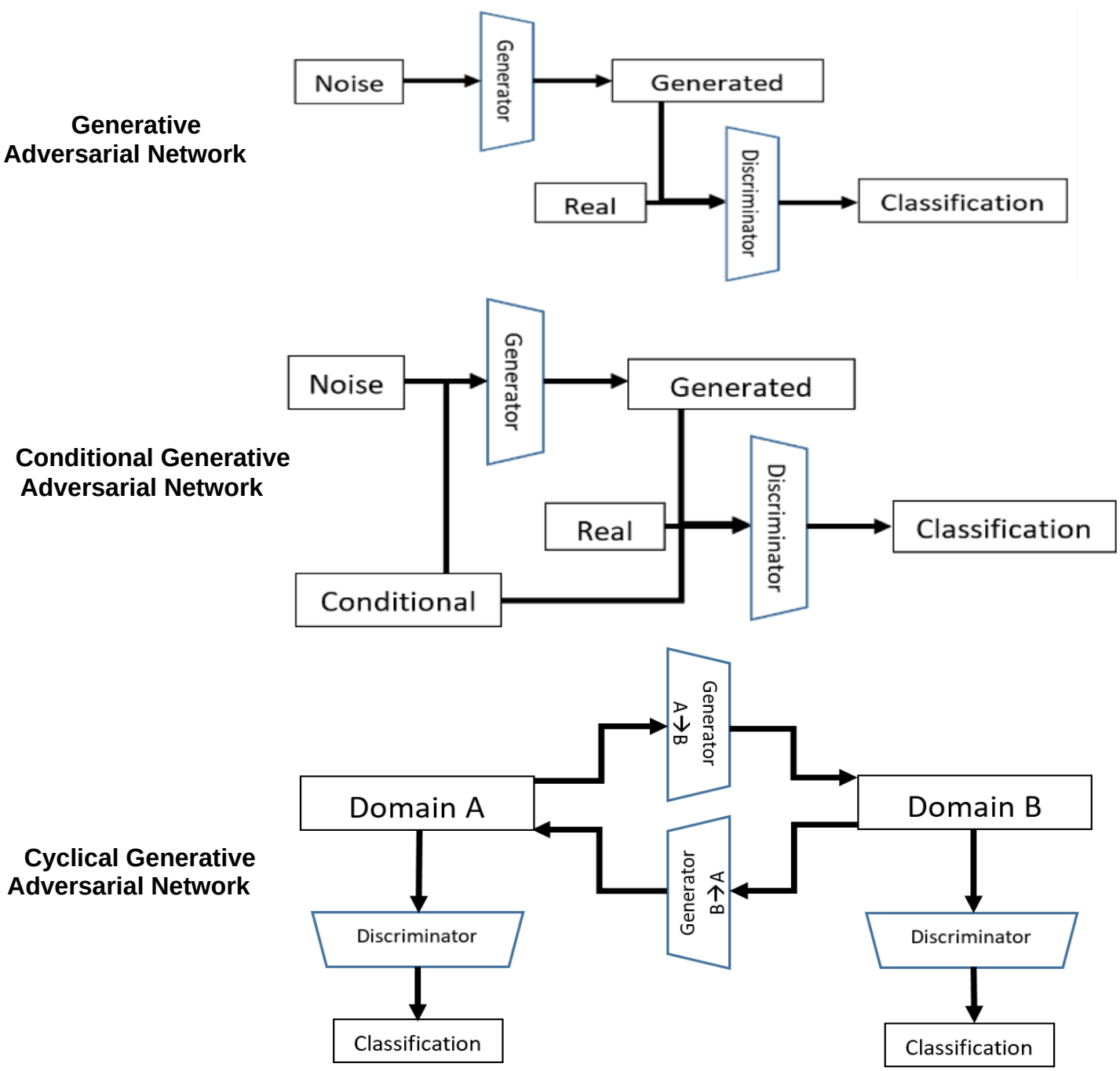

Figure 1. Above are flow charts for each of the types of GANs mentioned in this paper. The generative adversarial network flow chart shows the basic structure of GAN with the generator transforming noise into generated examples and both generated and real examples being assigned labels in the discriminator. Conditional GAN is very similar with the only difference coming from the conditional example or structure applied to the input noise for generation and discriminator for context-based discrimination. CycleGAN can take these a step further by connecting two GAN architectures to allow for domain transfer using the two GANs.

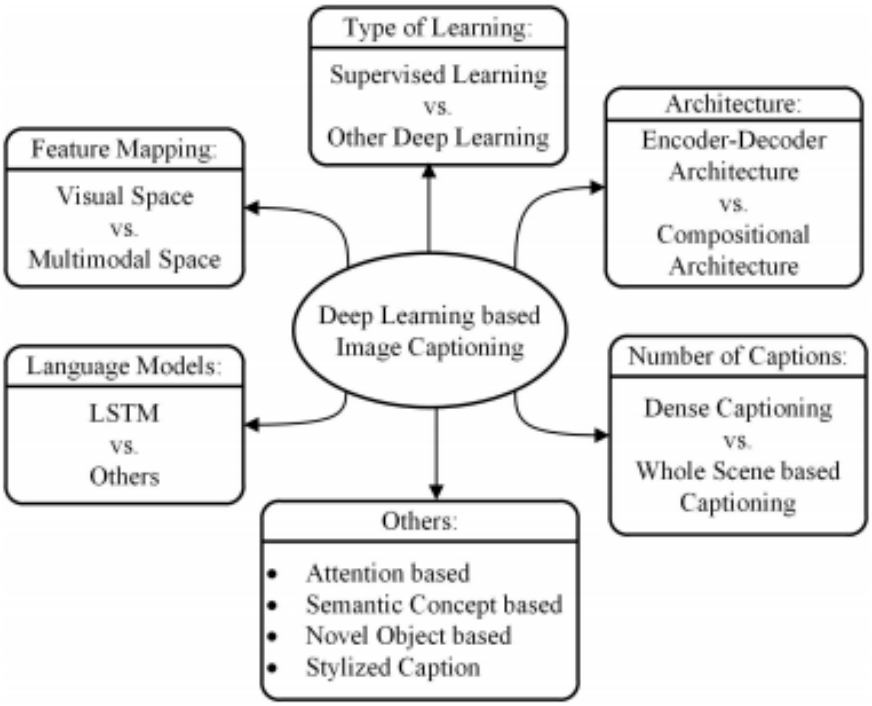

Figure 2. This is an overall taxonomy of deep learning-based image captioning from Hossain's A Comprehensive Survey of Deep Learning for Image Captioning [8]. Our implementation focuses on whole-scene based captioning with an encoder-decoder architecture using supervised learning to optimize CNNs and LSTMs for feature mapping based on a multimodal space.

2.2 IMAGE CAPTIONING SOLUTIONS WITH GANS

We are not the first to use a conditional GAN for image captioning. To avoid confusion and decrease complexity, the rest of this section and the whole methods section will refer to conditional GANs when speaking of GANs. One big issue with using GANs for captioning or any sparse dataset is with backpropagation, since discrete data which makes it difficult to differentiate [8]. Dai and Shetty have already approached the problem with their own implementations [11, 10, 8]. Dai creates his own captioning metric to make the captions more realistic and reinforcement learning to update the gradient without backpropagation [10]. Shetty uses a Gumbel sampler to allow for backpropagation to still work [11]. Neither implementation takes advantage of the inverse problem of image synthesis with a CycleGAN; however, our implementation will take advantage of these implementations.

2.3 IMAGE SYNTHESIS WITH GANS

Similarly, image synthesis is also a problem previously handled using GANs. One prominent example of this is in StackGAN, which is a cascading conditional GAN that is conditioned on an embedding of the text description [12]. Each stage of StackGAN generates a higher resolution image by conditioning on the previous stage's output [12]. An updated version of the model called StackGAN++ by modifying the structure by combining all of the cascades into a tree where all generators can simultaneously be trained and allow for a more end-to-end training scheme rather than separate conditional GANs being trained together [15]. Although this work is not related to image captioning, it is still important for this effort since it is the foundation for the image synthesis portion of our CycleGAN.

2.4 IMAGE CAPTIONING/SYNTHESIS WITH CYCLEGANS

There have been a few CycleGAN implementations of image captioning that came forth while we began working on this effort. One of which is text-to-image-to-text, an image synthesis algorithm that uses some of CycleGAN's structure [13]. Although it borrows from CycleGAN, Gorti's text-to-image-to-text is not a full CycleGAN, since it does not have a cycle-consistency loss for image-to-text-to-image [13]. Another CycleGAN implementation is Feng's Unsupervised Image Captioning method, which uses conditional GANs for image and text reconstruction [14]. This differs from our methodology since it uses outputs from these conditional GANs to feed a reinforcement learning method for improving the sentence captioning.

3. METHODS AND CURRENT STATUS

3.1 OVERALL ARCHITECTURE

Our image captioning framework is based on the CycleGAN architecture for image-to-captions and captions-to-image domain transfers. Initially, we planned on building off of Gorti's Text-to-Image-to-Text implementation to create TextCycleGAN, but their implementation proved too difficult to modify. As a result of these experiences, we decided to split our framework into two separate pieces: an image captioning GAN and an image synthesis GAN. We will be building the framework in Python 3. Using the CycleGAN structure will allow us to optimize both functions simultaneously and use errors from both functions to improve each other. Figure 3 has a flow chart of an overview of our final design. For mapping image captioning, we will currently are using Dai's image captioning GAN, but will be moving over to Shetty's image captioning GAN since the implementation has a method to allow usage of backpropagation and is already written in Python. Figure 4 and Figure 3 show a high-level diagram of our image captioning GAN. For image synthesis, we will utilize StackGAN++'s implementation. An overview of StackGAN++'s architecture can be found in Figure 5. As a part of each function, we need a word embedding to serialize the sentences from the captions for analysis. Currently, word2vec is a likely candidate. Lastly, the CycleGAN is incomplete without any form of cycle-consistency. Development on cycle-consistency is currently in progress. Although unfinished, we believe our framework will be successful due to success found in each of the conditional GANs and using the CycleGAN structure can only improve each function.

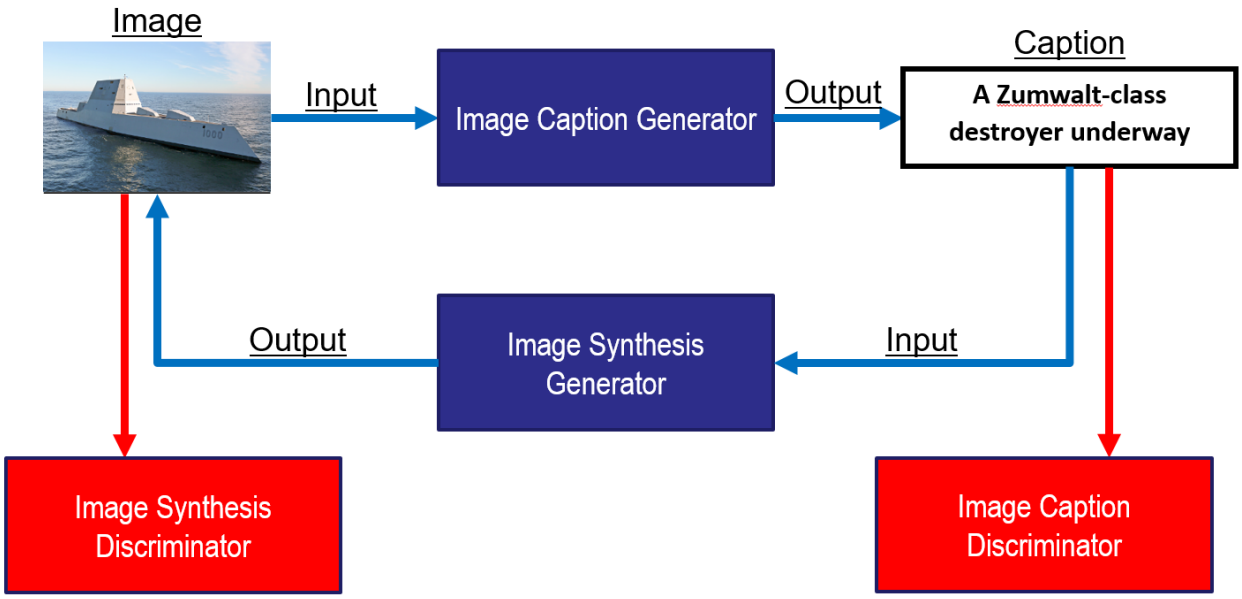

Figure 3. This is our image captioning framework. Similar to the CycleGAN architecture shown before, this diagram shows our architecture which utilizes both an image caption GAN and an image synthesis GAN to create TextCycleGAN.

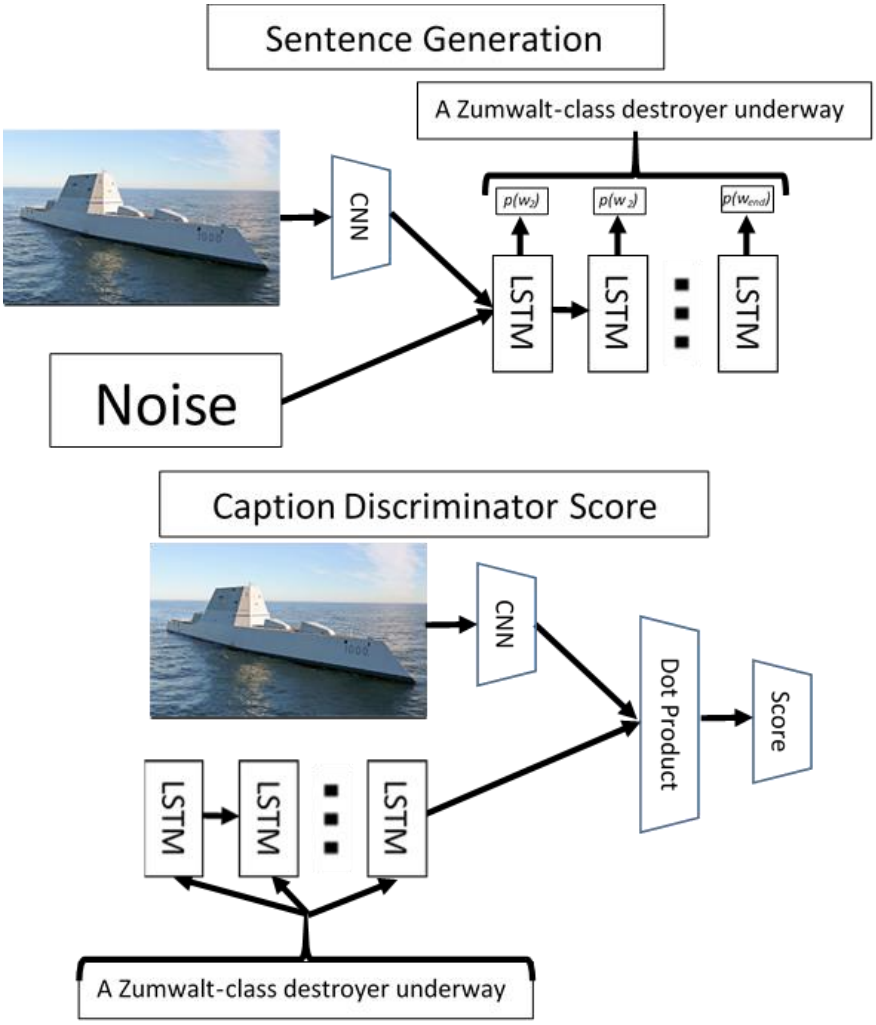

Figure 4. This is the structure for our image captioning GAN. This is currently built off of Dai's image captioning GAN implementation [10], but we hope to move to Shetty's implementation [11].

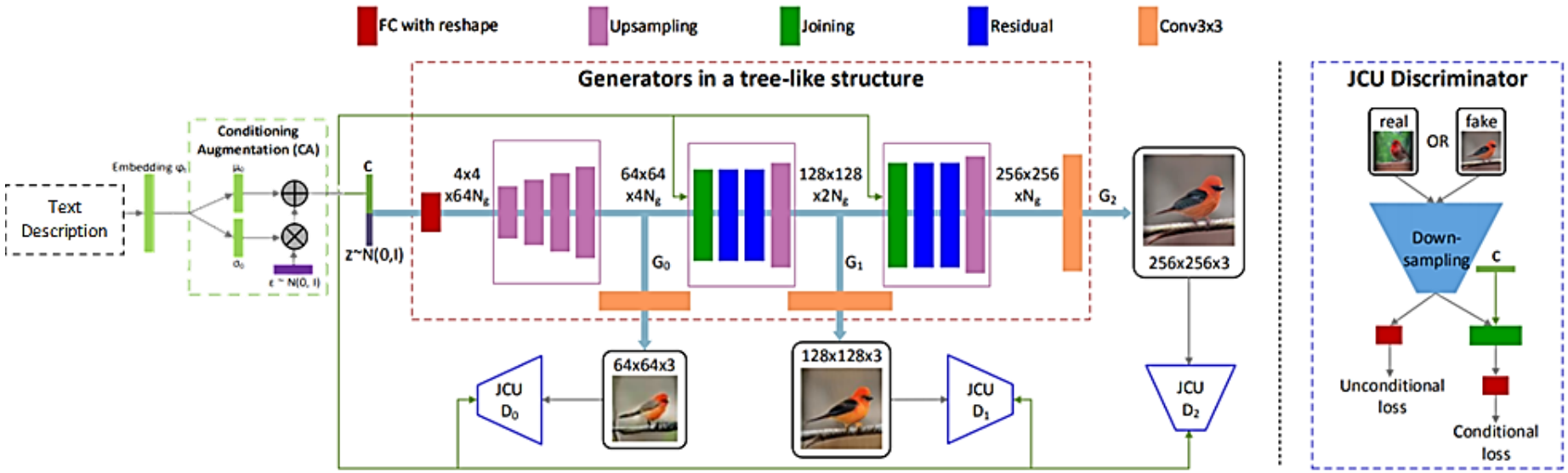

Figure 5. This is the structure for our image synthesis GAN. This is currently completely built off StackGAN++ [15], which has shown much promise.

3.2 WORD EMBEDDING

A critical part of working with textual data is in embedding it for use in machine learning. TextCycleGAN utilizes text as both an output and an input. An alternative approach would be to use a one-hot-encoding approach for every word in the dictionary, but this becomes unfeasible and very sparse. Word embeddings can also be trained with the rest of the model [9]. In practice, the image captioning GANs would not necessarily output a sequence of words, but rather a sequence of optimal word embeddings [9, 13, 10, 11]. Likewise, image synthesis GANs also use a text embedding instead of the entire description as an input [13, 12, 15]. With this in mind, we will also be utilizing word embeddings in our implementation for both parts of CycleGAN.

An initial milestone for TextCycleGAN was to recreate and understand the work of Gorti, et al. They used the skip-thought method of word embedding to encode descriptions prior to generating related images. After some research into the optimality of using skip-thoughts versus other word embedding algorithms such as word2vec or GloVe (global vectors for word representation), we have come to understand this naive difference between skip-thoughts and word2vec: context. Skip-thoughts focus on gaining context from surrounding sentences (e.g. good for understanding paragraphs) [18], whereas word2vec considers a more sentence-level context [16, 17]. Our goal is to create sentence descriptions of each input image and use sentence descriptions for the image synthesis portion. With this in mind, we have chosen word2vec as our word embedding for both output captions and input descriptions for image captioning and synthesis respectively.

3.3 IMAGE CAPTIONING CONDITIONAL GAN

To develop the image captioning GAN, we initially considered using Dai's implementation. His model showed considerable promise and we believed we could improve on it with the CycleGAN structure. He discusses how text generation involves creating distinct tokens that correspond to particular words, which becomes difficult to apply backpropagation directly [10]. Restricted by this, he utilized reinforcement learning approaches with policy gradients and utilized a score for each caption based on its input image to assist in training the network [10]. Although successful in captioning, this method soon proved difficult for us to implement, since his implementation of the algorithm was written in Lua scripting and the policy gradient method would need us to either drastically modify CycleGAN to use it or use a different structure. Our current implementation is still based off of Dai's work.

Recently, we discovered Shetty's Image Captioning GAN, which uses a Gumbel softmax sampler that is able to produce continuous outputs, which allow for backpropagation and, therefore, end-to-end training [11]. Where Dai used an evaluator to score each caption based on the input image [10], Shetty compares multiple output captions with each other in embedded space to test for diversity and compares the same multiple output captions with the image in embedded space to test for relevancy [11]. The results from Shetty's work are good and his implementation should fit well with ours. In the near future, we plan to move to Shetty's image captioning GAN as the foundation for ours. Additionally, Shetty uses more traditional metrics rather than defining a new one [11].

3.4 IMAGE SYNTHESIS CONDITIONAL GAN

For image synthesis, Text-to-Image-to-Text was considered to be the foundation for the image synthesis portion. Since they built off of the CycleGAN architecture [13], it would be a natural progression to build off of this implementation and build our image captioning CycleGAN; however, their CycleGAN did not include an image-to-image cycle-consistency loss and their implementation utilized a pre-trained image captioning GAN [13]. Essentially, it was building off of StackGAN and utilizing a different word embedding to accomplish the same task. As a result, StackGAN and StackGAN++ produced higher resolution and more diverse imagery than Text-to-Image-to-Text. As a result, our best option was to move away from Text-to-Image-to-Text and utilize what it was built on: StackGAN.

StackGAN takes an input text description encodes it into an embedding, combines the embedding with Gaussian noise using their Conditioning Augmentation, and then inputs the result into multiple stage of GANs, where the first stage generates a low resolution image and additional stages take the previous stage's generated image as an additional input to create a higher resolution image [12]. Results from this work seemed very promising alone, but the authors of the paper improved on the implementation by transforming StackGAN from multiple cascading GANs to multiple GANs connected in a tree-like structure [15]. This updated implementation creates tensors at each GAN and each of these tensors can be used to create an image of resolution similar to the size of the tensor, which allows for a more end-to-end framework [15]. The structure, implementation, and robustness of StackGAN++ is very favorable and will make for a great image synthesis GAN for our CycleGAN.

3.5 CURRENT STATUS AND TASKS REMAINING

Summarily, TextCycleGAN is built on two major pieces: an image captioning GAN and an image synthesis GAN. We will be using Shetty's image captioning GAN and StackGAN++ to build TextCycleGAN. We are currently still replicating Shetty's work, but we have replicated the original GAN by Goodfellow, the original CycleGAN, Text-to-Image-to-Text, and StackGAN++. You can find example outputs from the two in the results section. Additionally, to combine StackGAN++ and Shetty's image captioning GAN into one CycleGAN, we still need to establish a cycle-consistency loss for image-to-text-to-image and text-to-image-to-text.

This page is intentionally blank.

4. RESULTS

Replicating the original GAN and CycleGAN papers were crucial to see if the rest of the work could be successful. Figure 6 shows the results from generating MNIST images using GAN. Figure 7 shows results from changing winter to summer and vice-versa with CycleGAN respectively. By Epoch 40, GAN is able to generate some good images of handwritten digits but with few artifacts. By Epoch 80, those artifacts become minimal and the digits are clearer. CycleGAN successfully transforms the summer scenery into winter scenes and vice versa. The examples shown show instances of trees withering or greening or snow being added or removed. The only issue that can be seen from the transformation are from VGG artifacts that are mentioned in the original paper. Additionally, the transformations are successfully removed when mapped back to the originating domain.

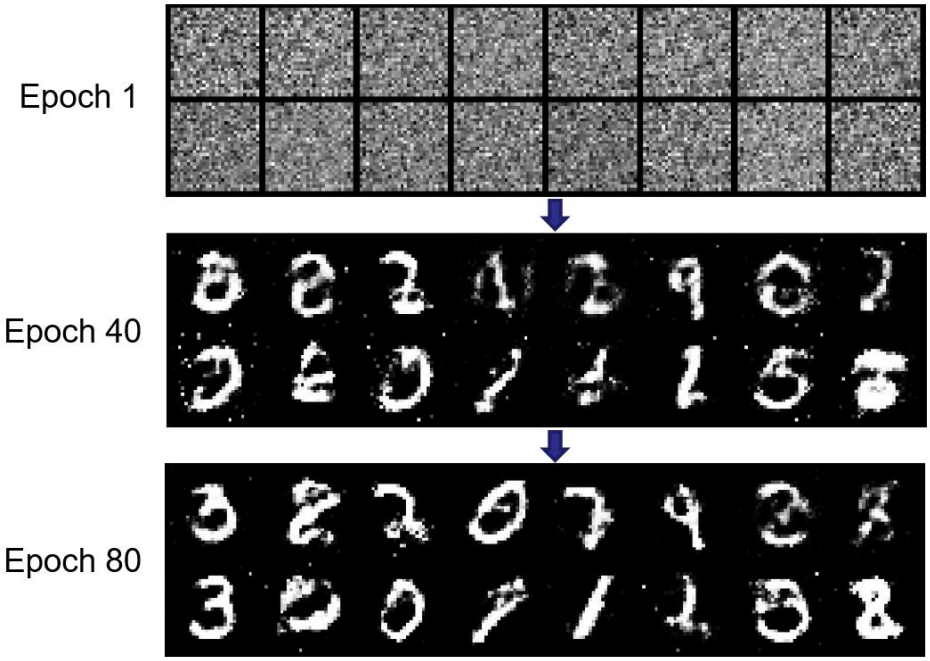

Figure 6. Results from replicating GAN and training the dataset on MNIST. We were successfully able to replicate a few handwritten digits by Epoch 80.

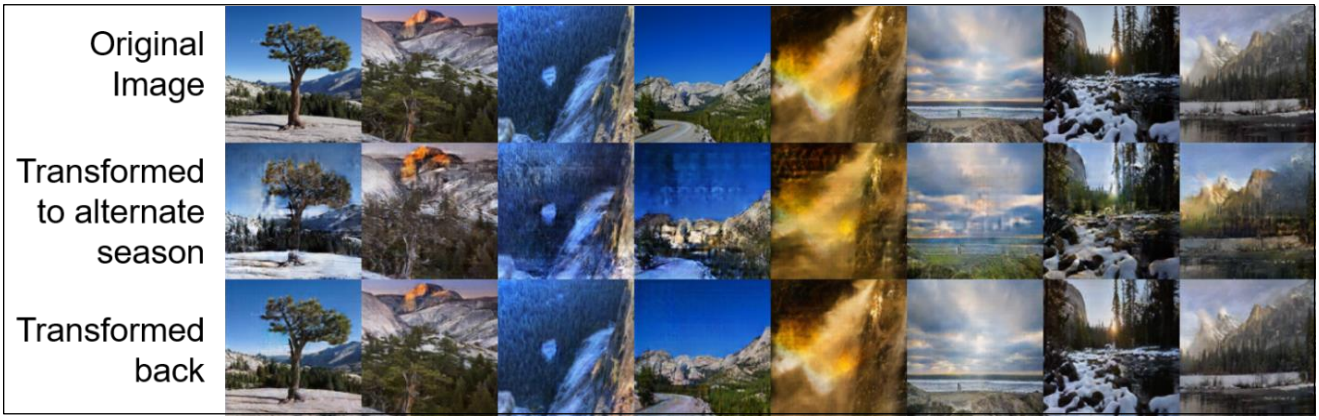

Figure 7. Results from replicating CycleGAN with the winter2summer dataset. We were successfully able to transform scenarios from winter to summer and vice-versa; however, most images displayed JPEG artifacts from the VGG network used in training.

For Text-to-Image-to-Text, we tried generating images like that of the Oxford 102 flowers dataset as mentioned in the paper. Shown in Table 1 are a few examples of images created using Text-to-Image-to-Text. The first three images do a wonderful job of creating flowers based on the descriptions; however, the fourth image not only displays a flower with a few errors it is also the same as the second image. This is an example of mode collapse, which is discussed in detail in the CycleGAN paper [5]. The outputs from Text-to-Image-to-Text have many examples like these. On the other hand, in Figure 8 shows a few examples from StackGAN++. These are images trained on and displayed using the CUB Birds dataset mentioned in the StackGAN++ paper [15]. There weren't any instances of mode collapse that we found, but there are some issues when generating pictures of waterfowls and some uncommon birds. These birds would either become some amalgamation of the a blotch on top of colors to mimic a bird-like image or have features of a finch or sparrow on top of the correct bird type. These issues could either stem from insufficient data or weak descriptions found in the captions. Although imperfect, StackGAN++ did a wonderful job at synthesizing these images with high resolution.

Table 1. Results from replicating Text-to-Image-to-Text and tested on the Oxford 102 flower dataset.

Text Description	Image
A beautiful flower with light pink petals and dark pink tips.	
A beautiful pink rose has green sepals and the corolla has 3-4 whorls of broad petals.	
A beautiful yellow and white flower with a large cluster of yellow stamen surrounded by multiple layers of white petals.	
A distinctive flower with a large circular red center, and petals that start red and turn to yellow at the tips.	

The bird is a mixture of red and white with short side wings.	This bird is black with white and has a long, pointy beak.	This greyish-green bird has a white belly, a sharp, down-curved beak and a red eyering.

Figure 8. Results from replicating StackGAN++ and tested on the CUB birds dataset.

This page is intentionally blank.

5. CONCLUSION AND FUTURE WORK

As mentioned prior, we still need to complete TextCycleGAN by combining Shetty's image captioning GAN and StackGAN++ and adding in a cycle consistency loss. Once finished, we plan to train TextCycleGAN on the Microsoft Common Objects in Context dataset and compare our results with Google's Show and Tell. From there, we intend to improve the algorithm by incorporating ideas from Dual Channel-wise Alignment Networks to better capture context from segments of imagery [20] and AttnGAN which is an updated StackGAN++ with an attention framework to better capture context from text. Additionally, we plan on applying TextCycleGAN to naval applications and a more naval captioning setting to see how it performs there.

This page is intentionally blank.

REFERENCES

1. “Captioning Leaderboards.” COCO: Common Objects in Context, http://cocodataset.org/#captions-leaderboard. Accessed 2 Sept. 2019.
2. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
3. M.T. Rosenstein, Z. Marx, L.P. Kaelbling, "To Transfer or Not to Transfer", Proc. Conf. Neural Information Processing Systems (NIPS '05) Workshop Inductive Transfer: 10 Years Later, 2005-Dec.
4. Augustus Odena. “Semi-Supervised Learning with Generative Adversarial Networks", Data Efficient Machine Learning Workshop, ICML 2016
5. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", in IEEE International Conference on Computer Vision (ICCV), 2017
6. Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua. “Generative adversarial nets.” NIPS, 2014
7. M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.
8. MD Hossain, F Sohel, MF Shiratuddin, and H Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys, 51(6):1–36, 2019.
9. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. In Proc. International Conference on Machine Learning http://arxiv.org/abs/1502.03044 (2014).
10. Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. Towards diverse and natural image descriptions via a conditional gan. arXiv preprint arXiv:1703.06029, 2017.
11. Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. arXiv preprint arXiv:1703.10476.
12. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.
13. S. K. Gorti and J. Ma. Text-to-image-to-text translation using cycle consistent adversarial networks. arXiv preprint arXiv:1808.04538, 2018.
14. Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.
15. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. TPAMI.
16. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Accepted to NIPS 2013.
17. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in ICLR, 2013.

18. Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In NIPS, 2015.

19. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-toimage translation with conditional adversarial networks. In CVPR, 2017.

20. Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis. DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation. In ECCV, 2018.

INITIAL DISTRIBUTION

84310	Technical Library/Archives	(1)
71740	M. Alam	(1)
71750	I. Dzieciuch	(1)
55360	M. Ayache	(1)
71740	N. Isoda	(1)
71740	M. Manzanares	(1)
71740	A. Delgado	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (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)*
February 2022	Final	

4. TITLE AND SUBTITLE

TextCycleGAN FY19 Technical Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Mohammad R. Alam
Iryna Dzieciuch
Maurice R. Ayache
Nicole A. Isoda
Mitch C. Manzanares
Anthony C. Delgado

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-3264

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.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

There has been much success with image captioning learned on large datasets, but the problem becomes more complex with smaller datasets. Generative adversarial networks (GANs) have shown promise in learning to generalize mappings from smaller datasets. GANs utilize the competition between a discriminator and a generator to strengthen generation. The discriminator identifies whether an input originated from the training set. The generator tries to create an output that the discriminator will falsely detect as being from the training set. With TextCycleGAN, we utilize cycle-consistent GANs (CycleGANs). CycleGANs utilize multiple GANs to learn a mapping between two domains. With CycleGAN, we can improve image captioning performance on smaller datasets by using both the GAN architecture for better generalization and by learning both translations from image to text and text to image. This will be a low-cost software package of a trained CycleGAN model for image captioning to be applied to naval applications.

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	38	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