Establishing an evaluation metric to quantify climate change image realism*

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
1 November 2019

REVISED
14 January 2020

ACCEPTED FOR PUBLICATION
13 February 2020

PUBLISHED
7 April 2020

PAPER

Establishing an evaluation metric to quantify climate change image realism*

Sharon Zhou[1,3], **Alexandra Luccioni**[2,3,4], **Gautier Cosne**[2], **Michael S Bernstein**[1] **and Yoshua Bengio**[2]

[1] Stanford University, United States of America
[2] Mila, Université de Montréal, Canada
[3] Equal contribution.
[4] Author to whom any correspondence should be addressed.

E-mail: sharonz@cs.stanford.edu, luccionis@mila.quebec, cosnegau@mila.quebec, msb@cs.stanford.edu and yoshua.bengio@mila.quebec

Keywords: deep learning, generative adversarial networks, climate change, generative network evaluation, domain transfer

Abstract

With success on controlled tasks, deep generative models are being increasingly applied to humanitarian applications (Nie *et al* 2017 *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 417–25, Yanardag *et al* 2017 *Deep Empathy*). In this paper, we focus on the evaluation of a conditional generative model that illustrates the consequences of climate change-induced flooding to encourage public interest and awareness on the issue. Because metrics for comparing the realism of different modes in a conditional generative model do not exist, we propose several automated and human-based methods for evaluation. To do this, we adapt several existing metrics and assess the automated metrics against gold standard human evaluation. We find that using Fréchet Inception Distance with embeddings from an intermediary Inception-v3 layer that precedes the auxiliary classifier produces results most correlated with human realism. While insufficient alone to establish a human-correlated automatic evaluation metric, we believe this work begins to bridge the gap between human and automated generative evaluation procedures, and to generate more realistic images of the future consequences of climate change.

1. Introduction

Historically, climate change has been an issue around which it is hard to mobilize collective action, notably because public awareness and concern around it do not match the magnitude of its threat to our species and our environment [1, 2]. One reason for this mismatch is that it is difficult for people to mentally simulate the complex and probabilistic effects of climate change, which are often perceived to be distant in terms of time and space [3]. Climate communication literature has asserted that effective communication arises from messages that are both emotionally charged and personally relevant over traditional forms of expert communication such as scientific reports [4], and that images in particular are key to increasing the awareness and concern regarding the issue of climate change [5]. With this in mind, our project leverages the MUNIT architecture [6] to perform cross-domain multimodal mapping between a street-level image without any flooding to multiple versions of this image under diverse flood transformations, to visually represent the impact of climate change-induced flooding on a personal level (for results of our model, see figure 1).

Generally speaking, generative models suffer from a lack of strong evaluation methods for comparing across both different models and different modes of the same model. Undeniably, much of the utility of generative models arises from their ability to produce diverse, realistic outputs, in addition to controlling generation—such as over specific modes, class labels [7], or visual attributes [8]—using conditional constraints. Conditional GANs have two inputs: the conditioning input (in our case, the image of a non-flooded house) and the random noise Z

* 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

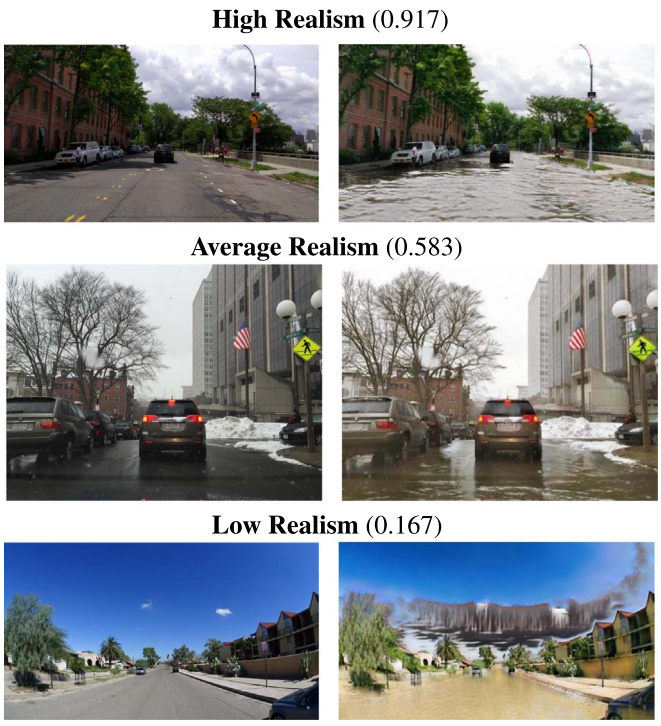

Figure 1. Generated images of flooded scenes from our model on the right (input on the left), spanning a range of scenes including an urban street, a city view, and a suburban scene. These images also span a range of performance on human perceptual realism, from highly realistic on top to highly non-realistic on the bottom. Exact HYPE-Style scores are indicated in parentheses.

which selects a *style*, defined as a mode of the conditional distribution learned in an unsupervised manner. Existing methods for evaluating the quality and diversity of the generated outputs have strong limitations, and are particularly scarce for conditional models. Widely used metrics include using heuristic approximations [9–11] that do not necessarily correlate with human judgment [12], rendering quantitative measurement of progress difficult. We encountered this issue during the course of the development of our model and in this paper we propose generally applicable methods for quantifying the realism of modes learned by a generative model. We start with a human evaluation of the images and styles produced by a multimodal generative model, followed by a comparison of human and automated approaches for evaluating the output of multimodal generative models, illustrated in the context of our image generation task.

2. Related work

To date, there have been two main approaches proposed for generative model evaluation: automated metrics such as Kernel Inception Distance (KID) [13], Inception Score (IS) [10] and Fréchet Inception Distance (FID) [11], which all aim to evaluate both the visual quality and sample diversity of generated samples at the distribution level, and, more recently, human-centered metrics such as HYPE (Human eYe Perceptual Evaluation) [12], which use human evaluators to assess image realism. Both approaches have their advantages and drawbacks: while automated metrics are cheap and easy to compute, they need large sets of both generated and real samples in order to produce reliable scores, which even then are not comparable between different tasks. Human metrics, on the other hand, may be more representative of human perception, but are more costly to compute and can vary depending on task design [14, 15].

Recent work has proposed ways of extending existing automated metrics, for instance by using a modified version of FID for conditional models [16] and sampling heuristics such as the truncation trick [17]. However, these modifications do not evaluate the visual fidelity between different modes, only within them in the case of Fréchet Joint Distance [16], which limits their application in multimodal settings such as ours. Methods for detecting artifacts [18] and artificial fingerprints [19] in generated samples also touch on perceptual fidelity, but either, in the case of artifacts, are a subset of image realism or, in the case of artificial fingerprints, encompass non-perceptual qualities that are imperceptible to a human viewer. Therefore, within the scope of our research, we found no satisfactory automated metric that would allow us to evaluate the realism of the images that we generated, and we endeavoured to find new ways of doing so, which we describe below.

3. Evaluating image realism

The research questions that we aim to answer are as follows: (1) What is the most effective way to evaluate the realism of different styles generated by our model? and (2) Can we propose an automated method that is correlated with human perceptual realism for automatically selecting the best mode on the flood generation task? We frame this task at the style level: for each given style vector, which represents a mode of the conditional distribution, we aggregate across multiple samples conditioned on the same mode. This style-level aggregation avoids evaluating on individual samples, which would produce noisier comparisons. We accomplish this by adapting the HYPE metric for style-level assessment using crowdsourced human evaluation, and call our new metric *HYPE-Style* (see section 3.1). We compare HYPE-Style against various automated metrics, which adapt FID and KID to the style comparison task. For each metric, we also experiment with different Inception layers.

We analyze Pearson's correlation coefficient r between each proposed automated style ranking method and HYPE-Style to identify the method that is most correlated with human perceptual realism. The measure r has support $[-1, 1]$, where values of 1 and -1 indicate strong positive and negative correlation, respectively, while values around zero indicate low correlation. An r of 1 is the maximum performance achievable on this metric. We also compute the 95% bootstrapped confidence intervals (CIs) on r using 25 replicates in order to determine the separability of the scores. For each replicate i, we compute HYPE-Style and an automated score using images sampled with replacement, from which we calculate r_i. We report the median r values, with 95% bootstrap CIs.

3.1. HYPE-style: human evaluation

In order to establish a human gold standard, we evaluated 500 image-style combinations drawn from our model, based on 25 input images of diverse locations and building types (houses, farms, streets, cities), each with 20 styles generated by our model. To establish the human baseline, we presented 50 images to each of our human evaluators: 25 real flooded images and 25 generated images. Following prior work, evaluators were calibrated and filtered by this tutorial of half real and half generated images, and were given unlimited time to label an image real or fake [12]. For each image, we compute the average error rate, which corresponds to the proportion of human evaluators who judged the image as real. Higher values indicate more realistic images.

We make several modifications to prior work in order to enable intra-style comparisons in conditional generation. Instead of randomly sampling across all generated images, we constrain the procedure in two ways: (1) we require that each style and image combination is evaluated multiple times, so we have comparisons between styles yet still within a given image, and (2) we ensure that evaluators do not see multiple styles generated from a given input image, as this visual redundancy would reveal that they were generated. These two adaptations increase the number of evaluators needed for this task, as evaluators are restrained to a limited set of images sans input redundancy, while still needed to evaluate across different styles for given input images.

We also diverge from the original HYPE metric when calculating scores, aggregating images by style into groups and computing the micro-average of all human evaluator labels within each group. Specifically, for each style s and image x, we have multiple human labels l_x^s marked either 'real' (1) and 'generated' (0) based on human judgments of its realism and we compute HYPE-Style $= \sum_i l_i^s$ for each style s, summing across images of that particular style. Thus, higher scores on generated images indicate higher fool rates and seem more realistic to humans on average. We use these style-level scores as the human baseline, where higher scores indicate more realistic styles, which we call *HYPE-Style*. This human evaluation, while more precise and reliable, is expensive and time-consuming to perform per style: we thus set out to find automated methods that are most correlated with human judgment to assess a much larger set of styles than is cost-efficient for HYPE-Style.

3.2. Automated style ranking methods

We adapted FID and KID to compute distances between real and generated distributions within a single style, and use these as the style scores. We also experimented with different layers of the Inception-v3 architecture trained on ImageNet [10] that span low-level (pool1) to high-level (pool3) features. For our evaluation, we included features from all three pooling layers, as well as the feature map before the auxiliary classifier (pre-aux). In total, we evaluate eight automated methods $\{\text{FID, KID}\} \times \{\text{pool 1, pool 2, pre-aux, pool 3}\}$.

As shown in table 1, both FID and KID using pre-aux embeddings exceed the other metrics in correlating with human HYPE-Style scores, with a moderate correlation ($r = 0.433$ and $r = 0.432$, respectively). Following these metrics, the observed order is: FID using pool 3 embeddings ($r = 0.407$), or the original FID score, then KID using pool 3 embeddings ($r = 0.367$). Finally, FID and KID using pool 2 and pool 1 layers exhibit extremely weak correlation with $r < 0.2$. When comparing performance between layers, KID and FID track each other, with pre-aux embeddings coming first, followed by pool 3, pool 2, and lastly pool 1.

While the original FID paper proposed to use features from the third and last 2048-dimensional pooling layer (pool 3) of an ImageNet-pretrained Inception-v3 network [11], we find empirically that the

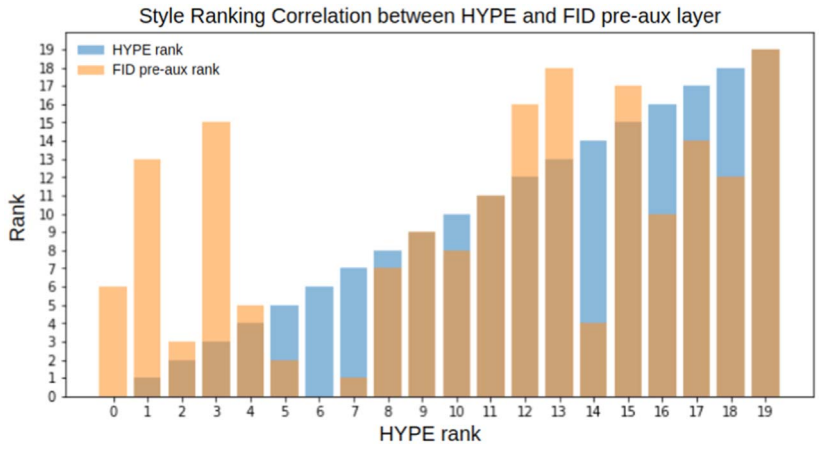

Figure 2. Rank order correlation between HYPE-Style and the best performing automated metric using FID on pre-auxiliary layer embeddings. Pearson's *r* correlation coefficient exhibits moderate correlation (0.433).

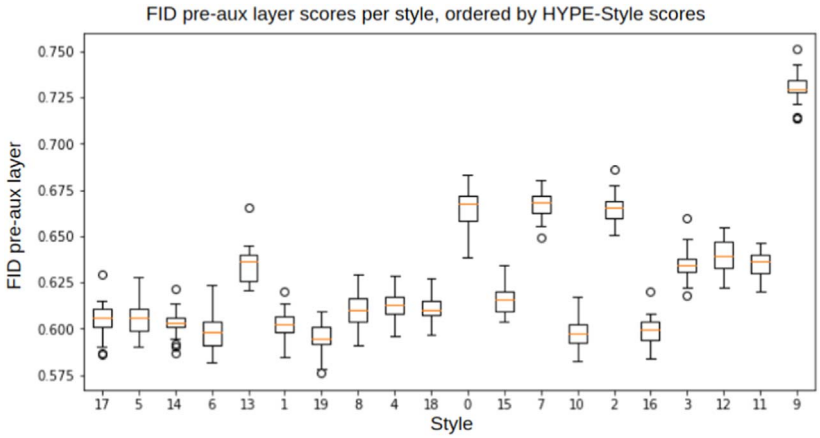

Figure 3. Box plot of FID performance using pre-auxiliary layer embeddings on different styles. The different styles are ordered by HYPE-Style scores to observe moderate correlation.

Table 1. Pearson's *r* correlation coefficient. Results of Pearson's *r* and bootstrap 95% confidence intervals between human HYPE-Style scores and all automated methods across different layers of an ImageNet-pretrained Inception-v3 model, including the three pooling layers (*pool* 1, *pool* 2, *pool* 3) and the layer preceding the auxiliary classifier (*pre-aux*). Higher values indicate greater correlation.

	Pool 1	Pool 2	Pre-aux	Pool 3
FID	0.103 (0.53, 0.153)	0.146 (0.099, 0.193)	**0.433 (0.390, 0.476)**	0.407 (0.366, 0.448)
KID	0.010 (−0.041, 0.061)	0.034 (−0.015, 0.083)	0.432 (0.389, 0.475)	0.367 (0.322, 0.412)

768-dimensional Inception-v3 layer just preceding the auxiliary classifier head (pre-aux) outperforms the pool 3 layer and other earlier pooling layers {pool 1, pool 2}. Intuitively, this is explained by the fact that the pre-aux layer is the most feature-rich layer that is still regularized by the gradients from the auxiliary classifier. This regularization would encourage the layer to encode more general features that are less overfit to ImageNet, which is more useful on this task, whose domain differs significantly from ImageNet. ImageNet itself has, in fact, also been criticized for generalizing poorly to test sets within its own domain [20]. We found that the choice of the pre-aux layer over pool 3 and others is consistent across FID and KID, with scores of 0.433 and 0.432 on the pre-aux layer against 0.407 and 0.367 on pool 3 for FID and KID, respectively. As a note, the difference between the FID layers' *r* values are not fully separable based on their 95% bootstrapped CIs. We show the rank order correlation between HYPE-Style with FID on the pre-aux layer embeddings in figure 2, with exact numbers in figure 3.

4. Discussion and future work

In this paper, we contribute a human evaluation metric for evaluating different styles on a generative model. We also evaluate eight different automated methods, finding that using Inception embeddings preceding the auxiliary classifier correlates more with human perception on this task than widely used methods using the last pooling layer. Our work is motivated largely by the dearth of available, reliable evaluation metrics for quantifying the progress of this task.

While none of the automated approaches evaluated comes sufficiently close to HYPE-Style for standalone use, our work still constitutes an initial foray into evaluating style-level attributes of multimodal cross-domain mapping, an area where it remains difficult to use mainstream automated evaluation metrics out of the box. Specifically, FID is a biased estimator and does not perform well on data with few samples. While KID remedies some of this problem, it is still possible that the order of magnitude of data was still insufficient for KID to be consistent and reliable without large number of runs. Nevertheless, both metrics have been shown to correlate imperfectly with human judgment. Their deficiencies lie in their relative insensitivity to visual features over semantic distortions and certain artifacts; these are a result of relying on embeddings from a pretrained ImageNet Inception-v3 model. Specifically, we find that automated metrics fail to detect artifacts in regions outside of the water flooding zones, e.g. the sky, that humans could immediately discern, as well as noise that appears to look like vertical motion blur in only parts of the image. An additional limitation of FID and KID is that they are distribution-level metrics and thus cannot compare individual images.

As future work, we plan to both improve the realism of our generative model and explore improved methods for evaluation, which persists as an open research problem in generative models. For instance, the performance of the pre-auxiliary classifier embeddings suggest that we are operating outside the domain of ImageNet, and from this insight, we are inclined to leverage other embedding spaces, e.g. the Mapillary or Cityscapes datasets [21, 22], which could provide more suitable street-level scenery features that is similar to ours. Using a method that is pretrained on ImageNet, then fine-tuned on a relevant dataset could provide improvements to automated evaluation. We could also explore different methods of measuring precision on generated images [23]. The variance of flooding severity is another area that would require conditional evaluation; that is, provided a certain flood condition, e.g. 2 m sea level rise, what would this look like on a given image? As flooding models grow more precise, we plan to juxtapose generated images of varying severity levels, using automated depth and height estimation techniques to project levels of flooding on streets and buildings.

The ultimate vision of this work is to create an interactive, ML-based website which, given an image from Google StreetView [24] based on a user-chosen location, is able to generate the most realistic image of climate change-induced extreme weather phenomena given the contextual characteristics of that given image and the future climate projections at that given location [25]. While representing flooding realistically is the first step to achieving this goal, particularly given the high population density of coastal regions worldwide, we later aim to represent other catastrophic events that are being aggravated by climate change (e.g. tropical cyclones or wildfires) using a similar approach, in the hopes that these will help raise awareness of the far-reaching future impacts of climate change.

Data availability statement

The data that support the findings of this study are available from the corresponding author, AL, upon reasonable request.

ORCID iDs

Alexandra Luccioni https://orcid.org/0000-0001-6238-7050

References

[1] Pidgeon N 2012 Climate change risk perception and communication: addressing a critical moment? *Risk Anal.: Int. J.* **32** 951–6
[2] Weber E U and Stern P C 2011 Public understanding of climate change in the United States *Am. Psychol.* **66** 315
[3] O'Neill S J and Hulme M 2009 An iconic approach for representing climate change *Glob. Environ. Change* **19** 402–10
[4] Lujala P, Lein H and Rød J K 2015 Climate change, natural hazards, and risk perception: the role of proximity and personal experience *Local Environ.* **20** 489–509
[5] O'Neill S J, Boykoff M, Niemeyer S and Day S A 2013 On the use of imagery for climate change engagement *Glob. Environ. Change* **23** 413–21
[6] Huang X, Liu M-Y, Belongie S and Kautz J 2018 Multimodal unsupervised image-to-image translation *Proc. European Conf. on Computer Vision (ECCV)* pp 172–89
[7] Mirza M and Osindero S 2014 *Conditional generative adversarial nets* arXiv:1411.1784

[8] Yan X, Yang J, Sohn K and Lee H 2016 Attribute2image: conditional image generation from visual attributes *European Conf. on Computer Vision (Berlin)* (Springer) pp 776–91
[9] Karras T, Laine S and Aila T 2019 A style-based generator architecture for generative adversarial networks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 4401–10
[10] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X 2016 Improved techniques for training GANs *Advances in Neural Information Processing Systems* pp 2234–42
[11] Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S 2017 GANs trained by a two time-scale update rule converge to a local Nash equilibrium *Advances in Neural Information Processing Systems* pp 6626–37
[12] Zhou S, Gordon M, Krishna R, Narcomey A, Morina D and Bernstein M S 2019 Hype: human eye perceptual evaluation of generative models arXiv:1904.01121
[13] Binkowski M, Sutherland D J, Arbel M and Gretton A 2018 Demystifying MMD GANs arXiv:1801.01401
[14] Le J, Edmonds A, Hester V and Biewald L 2010 Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation* vol 2126, pp 22–32
[15] Mitra T, Hutto C J and Gilbert E 2015 Comparing person-and process-centric strategies for obtaining quality data on Amazon mechanical turk *Proc. 33rd Annual ACM Conf. on Human Factors in Computing Systems* (New York: ACM) pp 1345–54
[16] DeVries T, Romero A, Pineda L, Taylor G W and Drozdzal M 2019 On the evaluation of conditional GANs *CoRR* arXiv:1907.08175
[17] Brock A, Donahue J and Simonyan K 2018 Large scale GAN training for high fidelity natural image synthesis arXiv:1809.11096
[18] Bau D, Zhu J-Y, Strobelt H, Zhou B, Tenenbaum J B, Freeman W T and Torralba A 2018 GAN dissection: visualizing and understanding generative adversarial networks arXiv:1811.10597
[19] Marra F, Gragnaniello D, Verdoliva L and Poggi G 2019 Do GANs leave artificial fingerprints? *2019 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)* (Piscataway, NJ: IEEE) pp 506–11
[20] Recht B, Roelofs R, Schmidt L and Shankar V 2019 Do imagenet classifiers generalize to imagenet? arXiv:1902.10811
[21] Neuhold G, Ollmann T, Rota Bulo S and Kontschieder P 2017 The mapillary vistas dataset for semantic understanding of street scenes *Proc. IEEE Int. Conf. on Computer Vision* pp 4990–9
[22] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B 2016 The cityscapes dataset for semantic urban scene understanding *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 3213–23
[23] Sajjadi M S, Bachem O, Lucic M, Bousquet O and Gelly S 2018 Assessing generative models via precision and recall *Advances in Neural Information Processing Systems* pp 5228–37
[24] Anguelov D, Dulong C, Filip D, Frueh C, Lafon S, Lyon R, Ogale A, Vincent L and Weaver J 2010 Google street view: capturing the world at street level *Computer* **43** 32–8
[25] Fitzpatrick M C and Dunn R R 2019 Contemporary climatic analogs for 540 North American urban areas in the late XXI century *Nat. Commun.* **10** 614
[26] Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q and Shen D 2017 Medical image synthesis with context-aware generative adversarial networks *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 417–25
[27] Yanardag P, Rahwan I, Garcia Herranz M, Fabian C, Rahwan Z, Obradovich N, Dubey A and Cebrian M 2017 Deep Empathy - Can Artificial Intelligence induce empathy? https://deepempathy.mit.edu/