# Evaluating Generative Models of Video

Sahil Bansal
Tokyo Institute of Technology
bansal.s.aa@m.titech.ac.jp

17 March 2019

## Introduction

There has been remarkable progress made in generative models for images. Learning generative models of video is, however, a much harder task. Generating video requires the model to capture the temporal dynamics of a scene, in addition to the visual presentation [9]. There is also a lack of good qualitative Metric for the Generative Models for Video. Two main metrics that are currently used to evaluate quality of video: PSNR, and SSIM do not correlate well with subjective video quality. A good qualitative metric for video should consider visual quality, temporal coherence and diversity of samples. In this proposal, I present some ideas on a new evaluation metric that would consider these criteria.

## Methodology

In the paper [9], authors modified FID [4] score to work on videos. FID score needs the real world data distribution and the distribution defined by generative Model. In other words we will need both the original dataset and the generated images to do the evaluation. This also holds true for FVD score on videos. Inception score can be used without the real distribution. This means we can evaluate the Inception score without the original dataset. The versatility of Inception score is the primary reasons I think we should modify Inception score to work on video as well. IS score necessitates a labeled dataset and is shown to be less robust to noise than FID.

Both Frechet Inception Distance (FID) and Inception Score (IS) [8] have shown promising results. However because they are both one-dimensional scores, they are unable to distinguish between different failure cases. Sajjadi provides a novel definition of Precision, Recall [7]. His approach lets us quantify the degree of mode dropping and mode inventing. This lets us assess whether the model is producing low-quality images or dropping modes. Applying this metric on videos would give us a more functional score and allow us to more carefully analyze our generative model's weakness. Video Precision and Recall can't distinguish between static and temporal noise.

We need a pre-trained network that can suitably suitable represent videos to use these metrics. The network should consider the temporal coherence [9] of the visual content across a sequence of frames along with visual presentation. Inflated 3D Convnet(I3D) is trained to perform action-recognition on the Kinetics dataset. I3D network is commonly used for extracting features for video.

We can test these metrics by adding various types of noise to real videos. I would add static noise to individual frames and temporal noise. For static noise, we can add black rectangle at random location, Gaussian blur and Gaussian Noise [9]. Temporal Noise can be added by locally swapping a number of frames, interleaving the sequence of frames corresponding to multiple different videos [9]. The *Noise Study* will test the sensitivity of our metric to various types of noise. We should also conduct a large-scale human study to confirm that new metric correlates well with qualitative human judgment of generated videos.

We should have 30-50 evaluators from different education backgrounds for the human evaluation. We show all the evaluators the videos and ask them to rank all the videos from 1 to 4 (good to bad) with respect to the two criteria: 1) Visual Quality: how are the visual features are these generated videos? 2) Coherence: judge the temporal connection and readability of the videos. We then show the evaluator 10 videos generated from 5 different model and ask to from 1 to 4 (good to bad) with respect to diversity. Furthermore, we average the ranking on each criterion of all the generated videos by each method and obtain three metrics [6]. I am planning to use BAIR [3], Kinetics-400 [2] and HMDB51 [5] datasets for my study. These datasets are used in [9] for their study.

## Related Works

The Inception Score [8] is a metric for automatically evaluating the quality of image generative models. There are 2 criteria that IS uses:

1. The quality of generated images
2. their diversity

To get a high IS score, generative model should generate sharp images with high diversity.

FID embeds a set of generated samples into a feature space given by a specific layer of Inception Net (or any CNN). Viewing the embedding layer as a continuous multivariate Gaussian, the mean and covariance are estimated for both the generated data and the real data. The Fréchet distance between these two Gaussians (a.k.a Wasserstein-2 distance) is then used to quantify the quality of generated samples. [1]

## References

[1] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *arXiv preprint arXiv:1710.05268*, 2017.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[5] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering '12*, pages 571–582. Springer, 2013.

[6] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798. ACM, 2017.

[7] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pages 5234–5243, 2018.

[8] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[9] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.