Project Document: Deep(FAMS)

Mohammad Alyetama
Ali Al-Ramini
Fangyi Li

Contents

Abstract

We present two project ideas that employ deep learning techniques. The first idea involves optimizing generative adversarial networks to generate good results with limited data using a recently proposed mechanism, the adaptive discriminator augmentation. The goal of this idea (Idea 1) is to evaluate the reproducibility and generalizability of the results generated by this proposed mechanism.

The second idea requires using big cycling data to analyze the effect of temporal factors on bicycling. This project idea aims to create a model that correcting map cycling patterns in the State of Nebraska.

Chapter 1

Milestone 1: Project Ideas

1.1 Introduction

1.1.1 Project Idea 1

Generative adversarial networks always seek to increase the dataset size to generate desirable results as an approach to generative modeling. The massive demand for data is challenging to lots of areas, and thus placing constraints on more widely being used of this practical approach. However, the dataset size reduction can cause overfitting issues in checking step by discriminator sub-model. Applying dataset augmentation can generate the augmented distribution, which is highly undesirable to generate samples. An attempt to apply a wide range of augmentation by Kerras et al. is trying to demonstrate GAN can use a smaller dataset size but prevent the augmentation ”leaking” at the same time. In this project, we will follow Karras et al. methods of adaptive discriminator augmentation mechanism [11] to try to reproduce their result.

1.1.2 Project Idea 2

There is a great interest in improving the cycling infrastructure in Nebraska by analyzing cyclic data. Toward this goal, we acquired the Strava Metro data covering Nebraska's state from January 2017 to December 2019. Strava provides data into four categories edges, nodes, origin-destination polygons, and shapefiles that contain spatial attributes to create maps using GIS software. Edges are street segments between nodes. In other words, a set of edges form a street. Nodes are the intersections between edges, and origin-destination polygons divide a space into smaller areas. Every category of Strava data is divided into yearly, monthly, and hourly data. The Strava data, made available by the social fitness network company Strava, includes raw data for the hour-by-hour counts for bicycle trips that have been mostly incorporated into existing maps for better visualization. In this project, we propose building a deep learning regression model that utilizes Strava data in addition to weather data to predict the number of cycling trips in Nebraska.

1.2 Project Idea 1: Evaluating the Reproducibility of Training GAN With Limited Data

Generative adversarial networks (GAN) is a generative modeling approach that uses deep learning techniques to automatically discover and learn the input patterns to generate an output that would plausibly appear as if it was sampled from the original input data [3]. For example, a GAN that uses convolutional neural networks (CNNs) can take pictures of humans as input data and generate new pictures of humans with plausible characteristics that look superficially genuine. This approach gained popularity after it was introduced by Goodfellow and his colleagues in 2014 [3]. In their paper, they describe GAN as a structure with two essential sub-models: (1) a generator model that learns to generate superficially plausible data, which the discriminator takes as negative training examples, and (2) a discriminator model that learns to discriminate between generated and real data, and penalizes the generator if an implausible result is detected [1]. The generator and the discriminator are neural networks that are directly connected. This connection allows the discriminator to send a signal, through backpropagation, that the generator uses to update its weights. Specifically, the generator samples a vector that is randomly drawn from a Gaussian distribution and use it to seed its generative process and match it to a distribution of interest, creating a "latent space." With sufficient training, the generator model can learn the input data's statistical latent space and create output data similar to what is observed in the input data [13]. An example of such data is then processed by the discriminator model and attempt to distinguish it from the real distribution of the data (i.e., a binary class of fake/real). The generator model's goal here is to maximize the error of the discriminator model. In other words, the generator model becomes more effective the more the discriminator process fake data as real data. The GAN approach has been used in fashion, science, and video games with impressive results [5].

However, a significant challenge in this area is the large number of data required to build a good GAN model, which is in some cases not available for researchers interested in applying GAN to their research question. GAN typically requires a large dataset because, with smaller datasets, the discriminator model ends up overfitting to training data examples, and the training eventually diverges [18]. While dataset augmentation is typically applied in such situations to solve overfitting [17], it cannot achieve this in GAN models. The inability to solve this problem with dataset augmentation stems from GAN's ability to employ this technique without learning the augmented distribution and leaking these results to the model [18] causing undesirable outcomes. Therefore, the challenge is to demonstrate that GAN can be used with smaller datasets without the pitfalls mentioned above.

A recent attempt to solve this problem was proposed by Kerras et al. [11], demonstrating that it is possible to obtain good results using limited data. The critical point in their proposed approach is that we can prevent overfitting and augmentations leak by applying a wide variety of augmentation methods. Their work demonstrates the validity of their approach by describing a set of conditions that allows controlling the augmentations leak problem and then proposes an adaptive discriminator augmentation pipeline that can dynamically control the strength of the augmentation. This is a novel approach they propose contrasting the convention of tuning the augmentation strength manually, a resources-consuming process. The process of building an adaptive discriminator augmentation mechanism, as described by the authors, is achieved by (a) Declare an overfitting heuristic, r, in which a value of zero represents no overfitting and a value of 1 means perfect overfitting. (b) Adjust augmentation probability (p) until the heuristic reaches a target value, which in turn can be processed by: (1) initializing the augmentation strength to zero, then adjusting p every four mini-batches based on an overfitting heuristic (rf). (2) if rf shows too much overfitting, it is countered by a fixed increment of p (or fixed decrement if rf shows too little overfitting). (3) The adjustment size is then set in a way that allows p to quickly rise from 0 to 1 while clamping p after every step from below to zero (adaptive). (4) The results from adaptive versus fixed p are then compared. Using such mechanisms on limited data, the authors demonstrated that their adaptive discriminator augmentation approach improved the quality of the result and stabilized training with a minimal effect on resources consumption, showing that their strategy is both viable and cost-efficient [11].

In this project, we will attempt to reproduce the results reported by Karras et al. using their adaptive discriminator augmentation mechanism. We expect our reproduced work to support Kerras et al.'s claim made in the paper; that is, good results can be obtained in a GAN model with only a few thousand training images. The original article's hypothesis was tested using five small datasets (METFACES, BRECAHAD, AFHQ, and CIFAR-10). Here, we plan to test adaptive discriminator augmentation on additional datasets to find out whether the scope of Kerras et al. paper is generalizable. This would be a pivotal point in our project because the authors of the article claim that their model's strength stems from its ability to work despite variations between datasets in content and size. This is a central argument to their approach. They experimentally demonstrate that a set of fixed augmentation parameters (as opposed to adaptive parameters) will miss the utmost advantage a GAN model can achieve. Overall, our results will allow us to evaluate the reproducibility of the results, the readability of the source code, and experiment with the ability to generalize the approach described in the paper with different datasets.

1.3 Project Idea 2: Using App Data to Model Bicycling Patterns in Nebraska

Across the United States, cycling is becoming increasingly popular as users shift travel modes amid concerns of health, physical activity, air, and environmental quality, and to escape roadway congestion. Unfortunately, the infrastructure in the U.S. traditionally caters to automobile traffic creating impediments for bikers and impacting their safety. To accommodate cycling, a major challenge is the lack of machine learning model representation of the available data to assess the attributes of present assets accurately and to inform additional investments to integrate bicycles into our transportation system. Toward that end, this project uses citywide bicycle travel data (i.e., Strava Metro Data) to provide a comprehensive description of daily cycling in a mid-size American state as a proof-of-concept approach to planning for cycling.

Various governments and organizations utilize big data to evaluate their cycling infrastructure [6]. Big cycling data are usually collected using live point data, journey data, Bike-Share Programs (BSP), and GPS. Live point data are collected on intersections using cameras on traffic lights, counting stations, or even sensors. While the journey data provide information about the origin and the destination of the trip, it does not provide the trip details. This set of data could be collected from BSP or by other sources like online questionnaires. BSP data are complete and in real-time. However, these data only give information within the area of its location [14, 15]. Strava is considered a GPS program that is made available by the social fitness network company Strava. Strava utilizes the Open Street Map (OSM) to deliver its data. These GPS data are very detailed and historical but represent a small sample of the cyclists' total population. Strava app data contains a vast amount of spatial and temporal details to predict cycling activity patterns. It provides a good approximation of the most-used routes and the peak months and hours. To protect privacy, the Strava data set is combined into population datasets. While a small portion of cyclic may use Strava to log their trips or the app might track trips for users using other transportation methods [2, 6, 15], several studies showed that there is a strong correlation between the Strava data and the ground-truth data obtained from counting stations [9, 10]. Cycling is affected by several factors such as weather, time of the day, infrastructure, congestion, environment, income, public transportation, health, population density, the slope of the street, and cultural view towards cycling [8,12,16]. But traditionally, access to high-quality data has limited our understanding of cycling behavior and route choice in the face of these myriad factors. In this project, we explore the weather effects and weather parameters sensitivity to cycling in Nebraska. This work aims to specify cycling behavior further as it relates to specific factors, but more importantly, to determine the quality of the specified factors in determining the number of cycling trips over a vast area like the State of Nebraska. Using data visualization techniques and Deep learning regression (e.g., ANN, RNN, LSTM, and GRU) [4, 7] we study the most influential time-related factors affecting the cycling patterns in Nebraska. Moreover, cycling is usually categorized into two classes, commuting and recreational. The proposed study will take advantage of the data shared by Strava to predict the number of commutes and recreational activities across all streets.

1.4 Conclusions

1.4.1 Project Idea 1

The core of this project is to test if the methods proposed by Karras et al. can solve the problem of augmentation leak and then ease the burden of huge datasets required by GAN. Given that the proposed mechanism is relatively new and paradigm-shifting, we believe that the idea of reproducing the paper's results would be greatly valuable. Studying such an approach to solve overfitting and subsequent problems while using limited datasets in GAN allows any researcher access to cost-efficient models.

1.4.2 Project Idea 2

This project provides an exciting idea: to create a deep learning model that predicts Nebraska's cycling patterns. With the help of high-quality data sources (e.g., Strava Data and Weather Data), this project's outcomes could be essential towards understanding how cyclists react to temporal variations. However, the spatial factors affecting cycling should also be studied, adding more complexity to the problem. The spatial representation of the data could be tough to map during this short period. Additionally, this idea is essential considering the current situation with COVID-19, adding more complexity and difficulty to reproducing the analysis to agree with the latest status.

Bibliography

[1] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[2] Elliot Fishman. Cycling as transport, 2016.

[3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[4] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[5] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

[6] Wendy Hall, Nigel Shadbolt, Thanassis Tiropanis, Kieron O'Hara, and Tim Davies. Open data and charities. 2012.

[7] Mohamad H Hassoun et al. *Fundamentals of artificial neural networks*. MIT press, 1995.

[8] Hartwig H Hochmair, Eric Bardin, and Ahmed Ahmouda. Estimating bicycle trip volume for miami-dade county from strava tracking data. *Journal of transport geography*, 75:58–69, 2019.

[9] Jinhyun Hong, David Philip McArthur, and Mark Livingston. The evaluation of large cycling infrastructure investments in glasgow using crowdsourced cycle data. *Transportation*, 47(6):2859–2872, 2020.

[10] Ben Jestico, Trisalyn Nelson, and Meghan Winters. Mapping ridership using crowdsourced cycling data. *Journal of transport geography*, 52:90–97, 2016.

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

[12] Walter Musakwa and Kadibetso M Selala. Mapping cycling patterns and trends using strava metro data in the city of johannesburg, south africa. *Data in brief*, 9:898–905, 2016.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[14] Scott Rogers and Nikolaos P Papanikolopoulos. Counting bicycles using computer vision. In *ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 00TH8493)*, pages 33–38. IEEE, 2000.

[15] Gustavo Romanillos, Martin Zaltz Austwick, Dick Ettema, and Joost De Kruijf. Big data and cycling. *Transport Reviews*, 36(1):114–133, 2016.

[16] Avipsa Roy, Trisalyn A Nelson, A Stewart Fotheringham, and Meghan Winters. Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Science*, 3(2):62, 2019.

[17] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[18] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv preprint arXiv:2002.04724*, 2020.