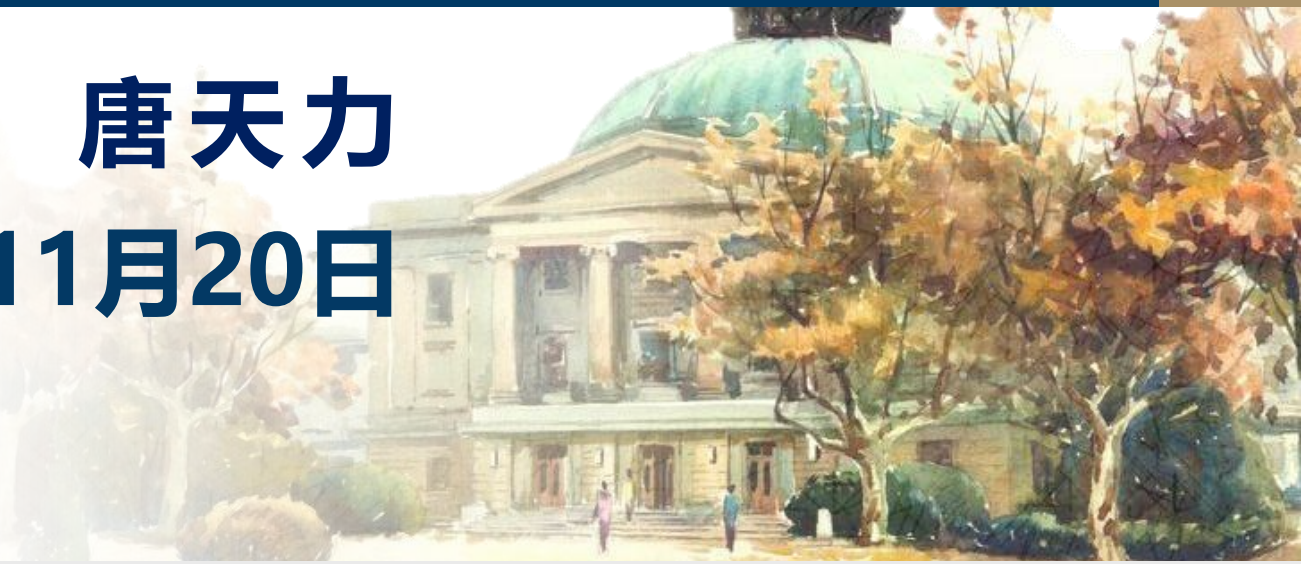
生成对抗网络在交通数据增强中的应用

汇报人：唐天力

2022年11月20日



CONTENTS

生成对抗网络的基础架构

東南大學
SOUTHEAST UNIVERSITY

AI绘画

《Theatre d'Opera Spatial》
By Jason Allen (Midjourney)

生成对抗网络
Generative Adversarial Nets
GAN

"Since 2010, Bengio's papers on generative deep learning, in particular the Generative Adversarial Networks (GANs) developed with Ian Goodfellow, have spawned a revolution in computer vision and computer graphics. In one fascinating application of this work, computers can actually create original images, reminiscent of the creativity that is considered a hallmark of human intelligence."

——ACM A.M. Turing Award 2018

- 对抗学习
- 自监督学习
- 无监督学习

Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie*, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair†, Aaron Courville, Yoshua Bengio‡
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

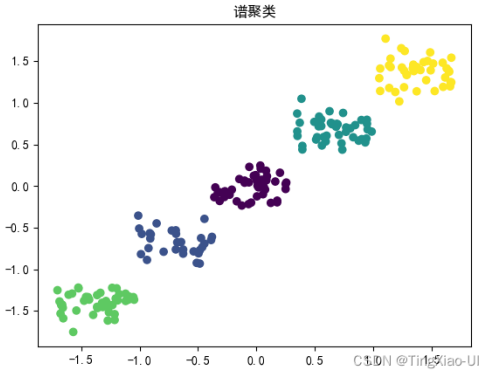
We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

生成对抗网络的基础架构

判别式模型

- 拟合条件概率分布：$P(y|x)$
- 反向传播、Dropout

“猫”



生成式模型

- 拟合联合概率分布：$P(x,y)$
- 极大似然估算

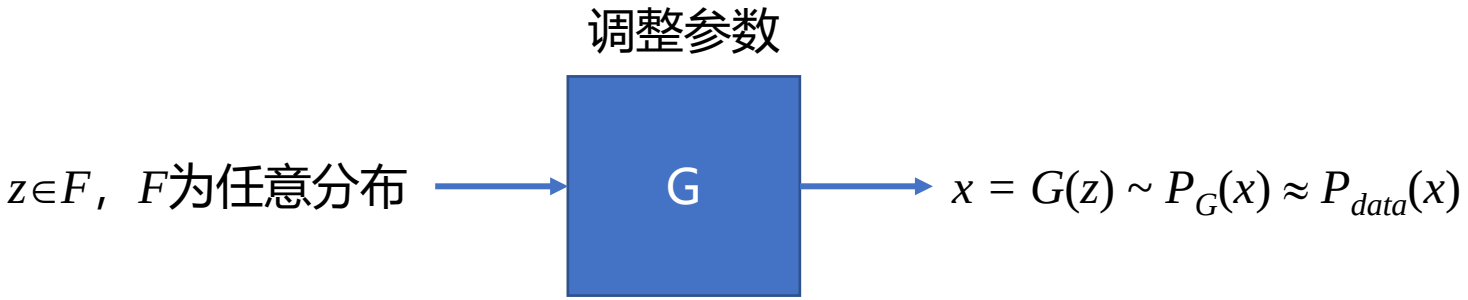
生成对抗网络的基础架构

主要组成

- 生成器G（Generator）
- 判别器D（Discriminator）

网络目的

- 生成器G能学习到样本的真实分布$P_{data}(x)$
- G能生成之前不存在又很真实的数据



生成对抗网络的基础架构

生成器G

蝴蝶：彩色
枯叶：棕色

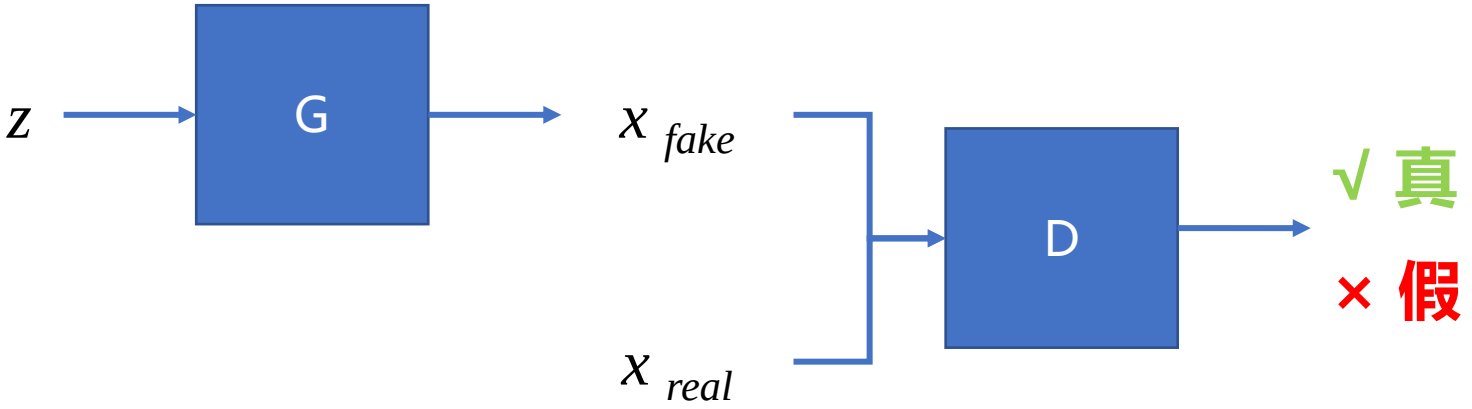
蝴蝶：没有叶脉
枯叶：有叶脉

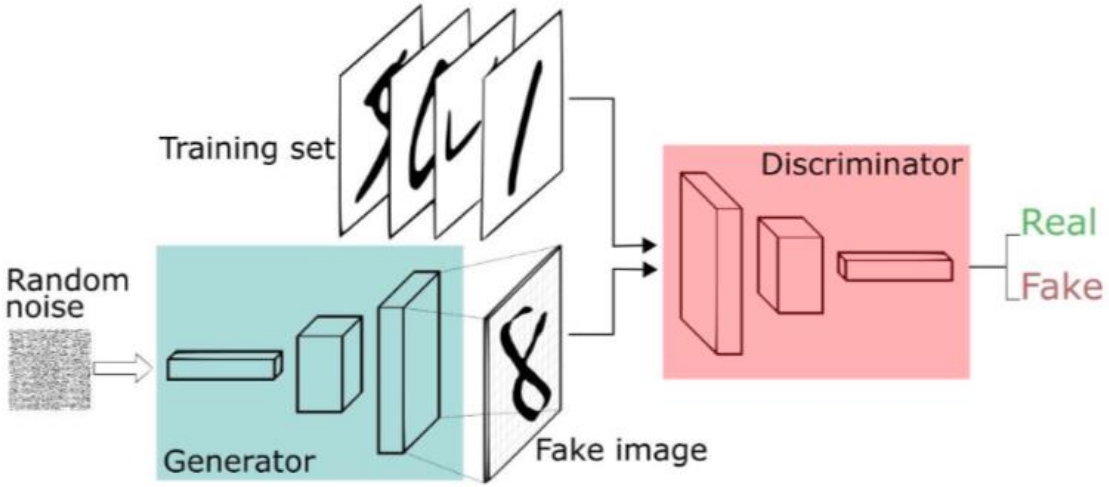
。。。

判别器D

生成对抗网络的基础架构

网络结构





生成对抗网络的基础架构

口 数学表达

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

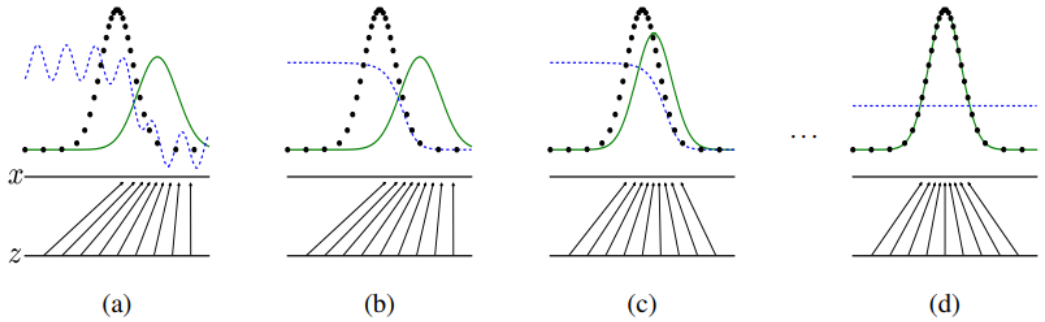
（1）固定G，调整D，最大化V(D,G)，导致D(x)→1，D(G(z))→0

（2）固定D，调整G，最小化max V(D,G)，导致D(G(z))→1
D

极大极小零和博弈

纳什均衡

判别器完全无法区分两个样本



生成对抗网络的基础架构

□ AI绘画

《Beyond the Stars: At Galaxy's Edge》
by Julie Dillon

生成对抗网络的基础架构

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]. \quad \Longrightarrow \quad P_G(x) \approx P_{data}(x)$$

$$\begin{aligned} V(G,D) &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_z p_{\boldsymbol{z}}(\boldsymbol{z}) \log(1-D(g(\boldsymbol{z})))dz \\ &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \end{aligned} \quad \Longrightarrow \quad D(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$$

对D求导

$$\begin{aligned} & V(D,G) \\ &= \int_x p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x)+p_g(x)} + p_g(x) \log \frac{p_g(x)}{p_{data}(x)+p_g(x)} dx \\ &= \int_x p_{data}(x) \log \frac{\frac{1}{2}p_{data}(x)}{\frac{p_{data}(x)+p_g(x)}{2}} + p_g(x) \log \frac{\frac{1}{2}p_g(x)}{\frac{p_{data}(x)+p_g(x)}{2}} dx \end{aligned}$$

生成对抗网络的基础架构

KL散度 $$KL(P||Q) = \sum P(x)\log\frac{P(x)}{P(x)+G(x)}$$

$$-2\log 2 + KL\left(p_{data}(x)\left\|\frac{p_{data}(x)+p_g(x)}{2}\right.\right) + KL\left(p_g(x)\left\|\frac{p_{data}(x)+p_g(x)}{2}\right.\right)$$

JSD散度 $$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(P||M) \qquad M = \frac{1}{2}(P+Q)$$

$$-2\log 2 + KL\left(p_{data}(x)\left\|\frac{p_{data}(x)+p_g(x)}{2}\right.\right) + KL\left(p_g(x)\left\|\frac{p_{data}(x)+p_g(x)}{2}\right.\right)$$

$$\min_G -2\log 2 + 2JSD(P_{data}(x)||P_G(x))$$

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]. \quad \Longrightarrow \quad P_G(x) \approx P_{data}(x)$$

生成对抗网络的基础架构

缺点

- 没有显示表示$P_G(x)$
- Helvetica scenario

优点

- 不需要使用马尔科夫链
- 只需要用反向传播和梯度下降
- 神经网络中大量的网络结构、训练技巧、损失函数可以直接运用
- 对抗学习的思想避免了过拟合问题

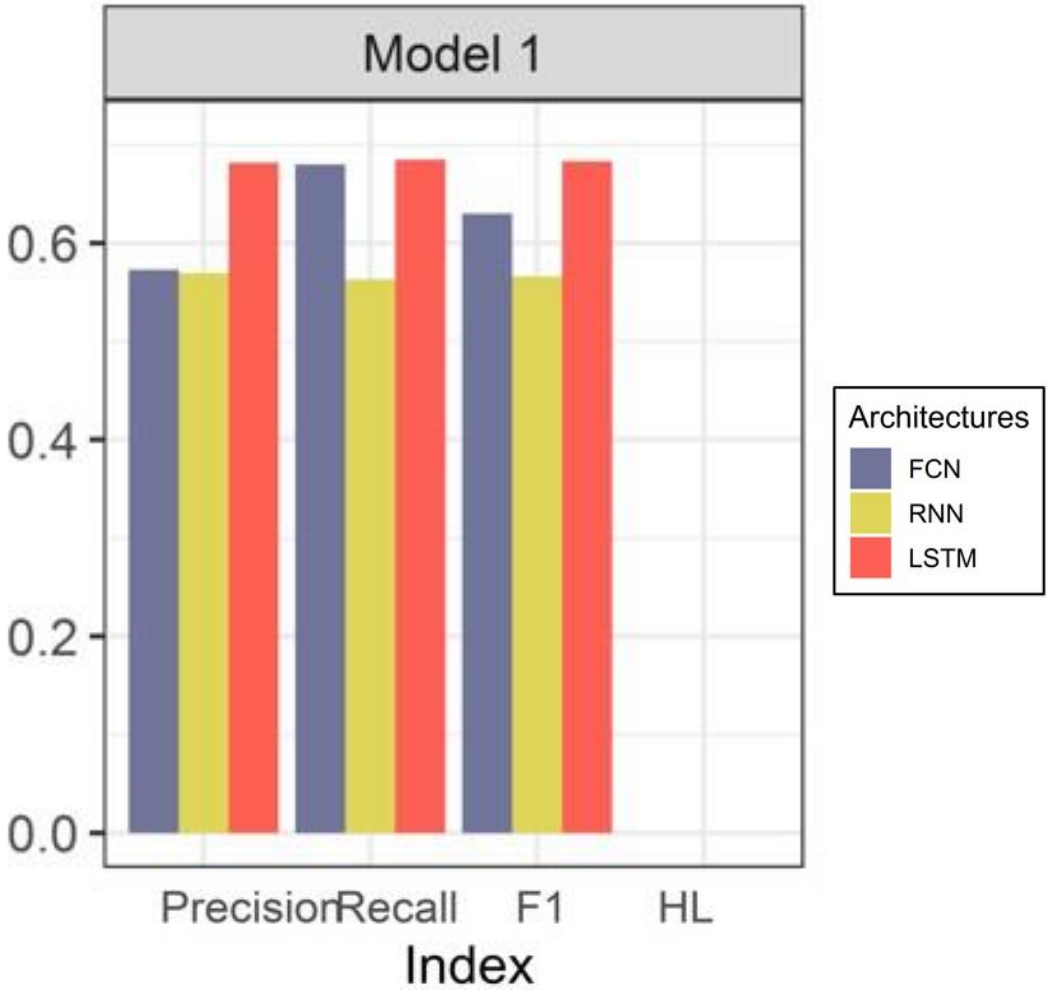
交通数据增强应用

東南大學
SOUTHEAST UNIVERSITY

生成对抗网络在交通数据增强应用

背景

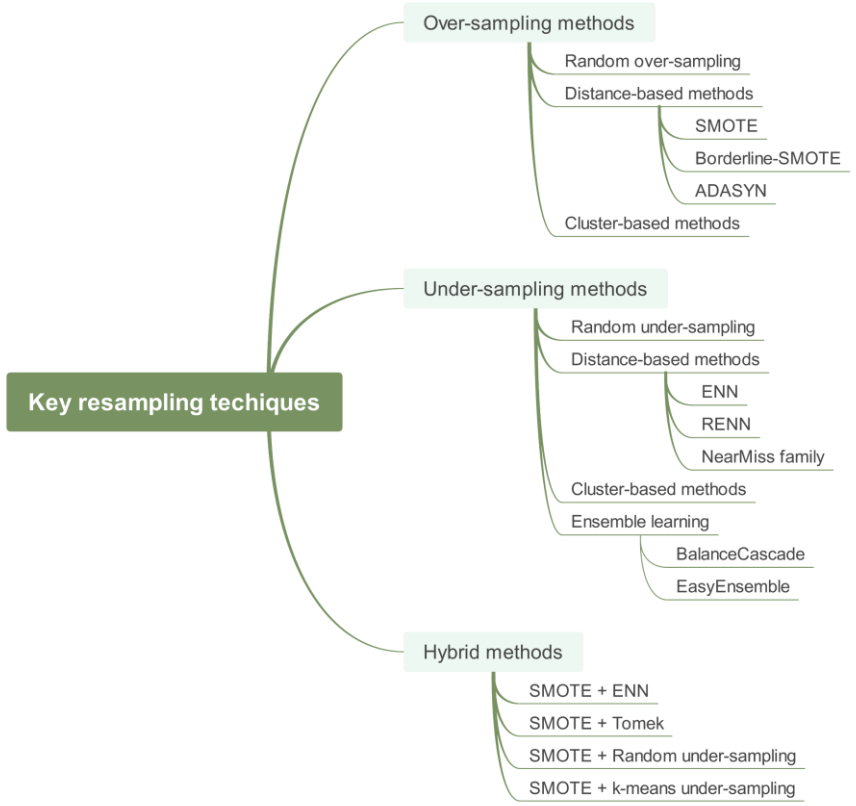
- 个体公交出行预测：某个公交用户在某个时间窗口是否会出行
- 101850用户*19小时=1935150条实例/天
- 出行数据：43003
  不出行数据：1892147
  数据不均衡比：1/44
- 不均衡数据集危害：
  - 小比例实例可能会被视为噪声数据，反之亦然
  - 小比例实例在面对高维度特征时信息密度低



生成对抗网络在交通数据增强应用

数据重采样

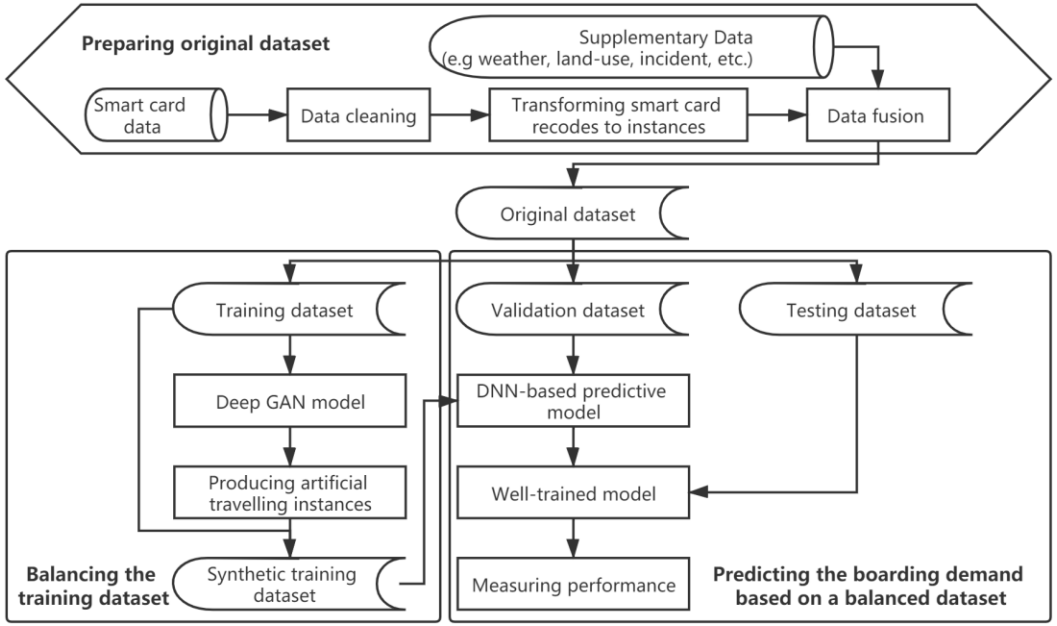
- 过采样：
  - 生成的数据很容易受到离群值的影响
- 欠采样：
  - 失去部分数据的信息的代价
- 混合采样

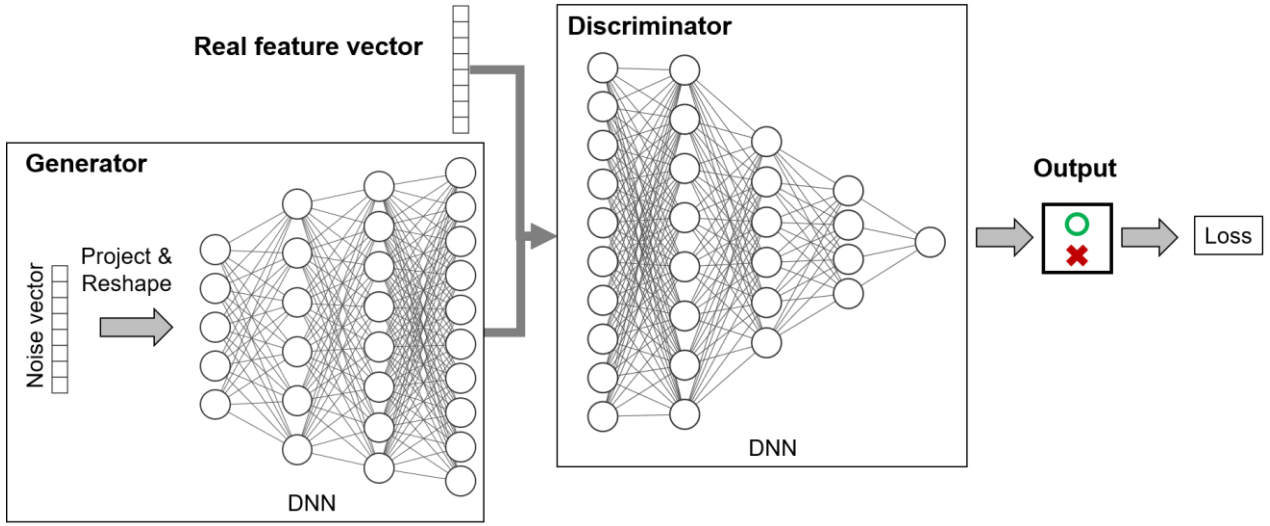


生成对抗网络在交通数据增强应用

模型框架

- Deep-GAN对小比例的出行样本重采样
- 增强后的数据训练DNN预测模型



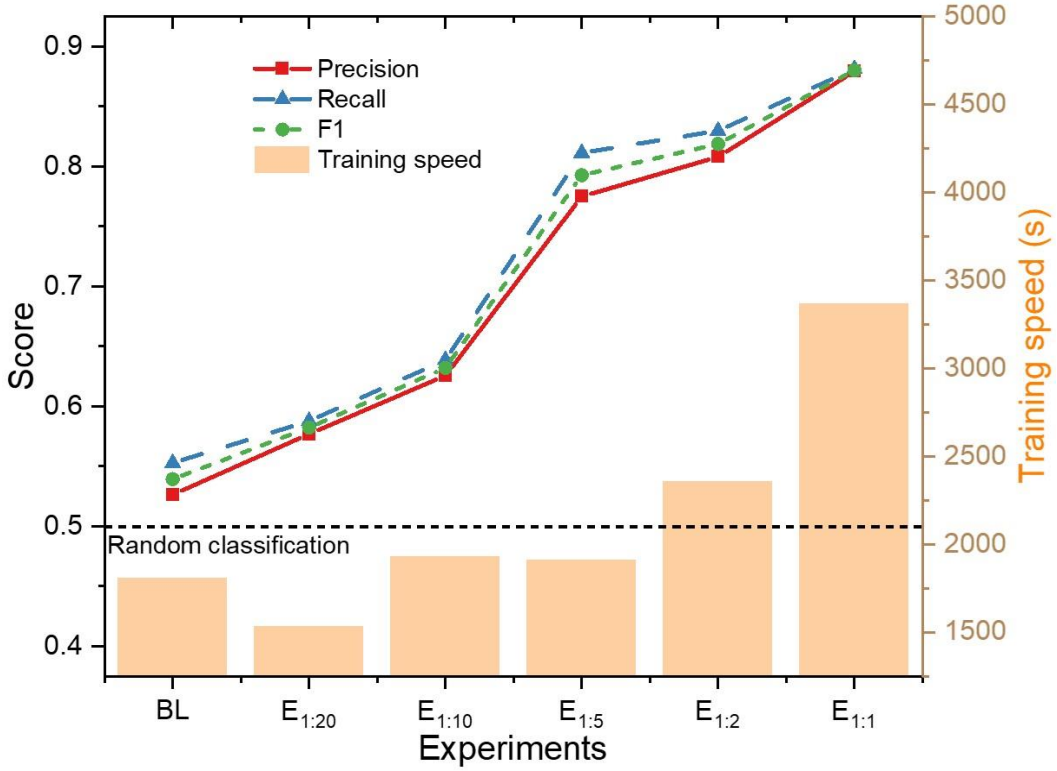


生成对抗网络在交通数据增强应用

口 不均衡比例对预测模型的影响

- 数据均衡，训练出的预测模型性能越好
- 1:5不均衡比可以达到较强的预测性能与较快的学习效率

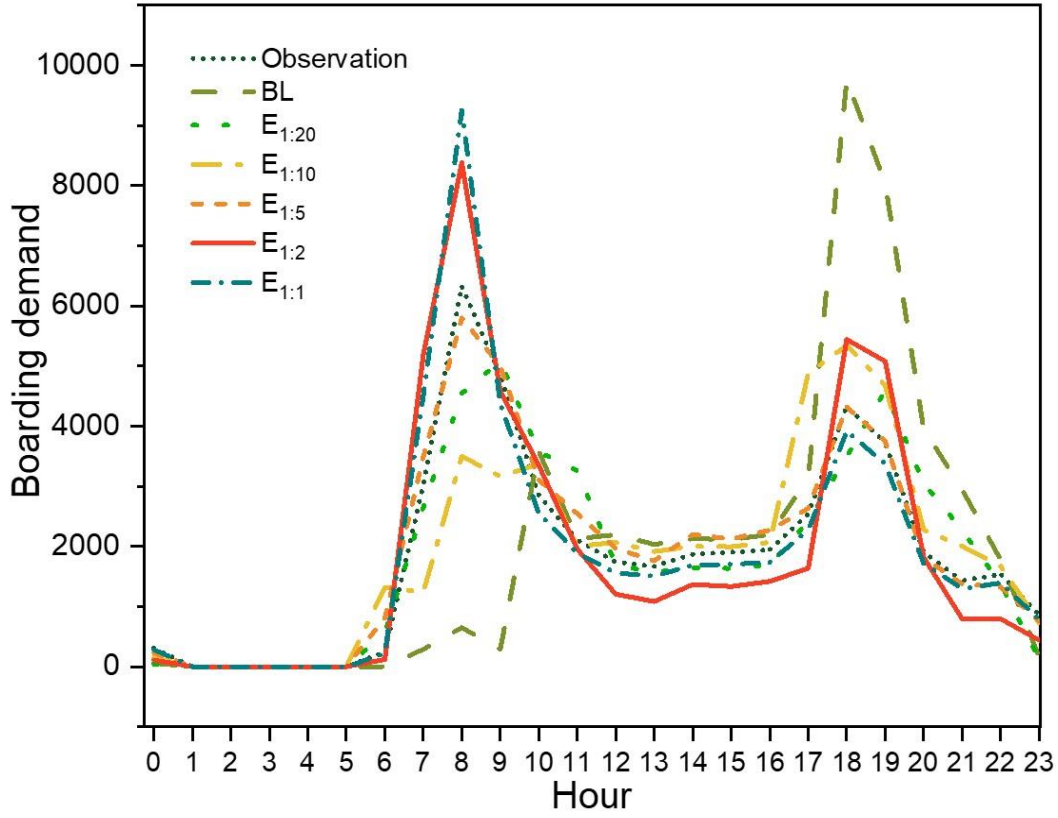
Experiments	Real travelling instances	Synthetic data	Non-travelling instances	Imbalanced rate	Total
BL (original)	819,278	0	36,335,602	1:45	37,154,880
$E_{1:1}$		35,516,324		1:1	72,671,204
$E_{1:2}$		17,348,523		1:2	54,503,403
$E_{1:5}$		6,447,842		1:5	43,602,722
$E_{1:10}$		2,814,282		1:10	39,969,162
$E_{1:20}$		997,502		1:20	38,152,382



生成对抗网络在交通数据增强应用

口 不均衡比例对预测模型的影响

- BL, 1:20 & 1:10：不平衡的数据会导致机器学习模型的不准确
- 1:1 & 1:2：由信息冗余和重复造成的



Experiments	BL	$E_{1:20}$	$E_{1:10}$	$E_{1:5}$	$E_{1:2}$	$E_{1:1}$
RMSPE	0.74	0.49	1.09	0.56	0.38	0.18
RMSE	2483.44	712.35	1114.84	281.26	912.15	785.62

生成对抗网络在交通数据增强应用

重采样方法产生的训练数据对预测模型的影响

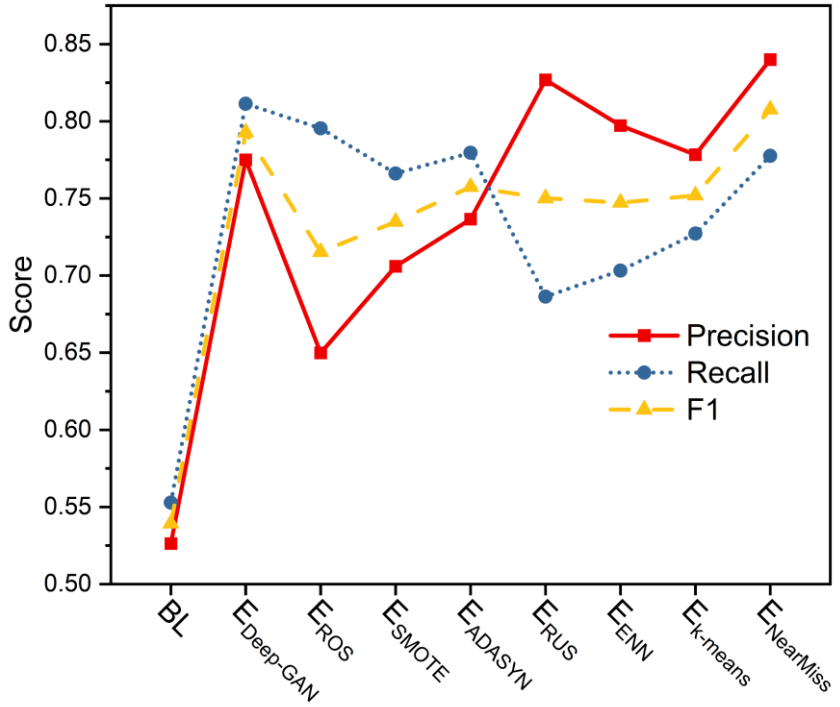
- 出行与非出行数据中存在明显差异
- Deep-Gan能够产生与真实数据相似的合成数据，并且与非出行数据有明显不同
- Deep-GAN确保了合成数据的多样性，后续的预测模型不会过度适应某些数据特征。

Synthetic travelling instances by method	The values of FD	
	Real travelling instances	Non-travelling instances
Real travelling instances	-	455.24
$E_{Deep-GAN}$	87.17	315.32
E_{ROS}	$<10^{-4}$	455.23
E_{SMOTE}	0.53	452.15
E_{ADASYN}	16.96	341.00

生成对抗网络在交通数据增强应用

重采样方法产生的训练数据对预测模型的影响

- 只要通过任何重采样方法降低不平衡比，预测模型的准确性都将得到提高
- 合成训练数据的多样性将有助于预测模型的精度
- 欠采样方法产生的训练数据集比大多数过采样方法更可靠



撸起袖子，加油‘GAN’！