敵対的生成ネットワークを利用した疑似トラヒックデータ生成手法

長岡技術科学大学 大学院工学研究科
電気電子情報工学専攻
山際 哲哉
渡部 康平
中川 健治

背景

- 近年，ユーザーが利用する端末が多様化し，ネットワークトラヒックの性質も多様化している
- ネットワークを快適に利用できる最適なシステム提供が必要
- その際トラヒックジェネレータを使って，シミュレーションやテストを行う
- 例：キャパシティプランニング
  - ITシステムの構築の際に使われる
  - あるトラヒックを処理するのに，どのくらい増強すれば良いか

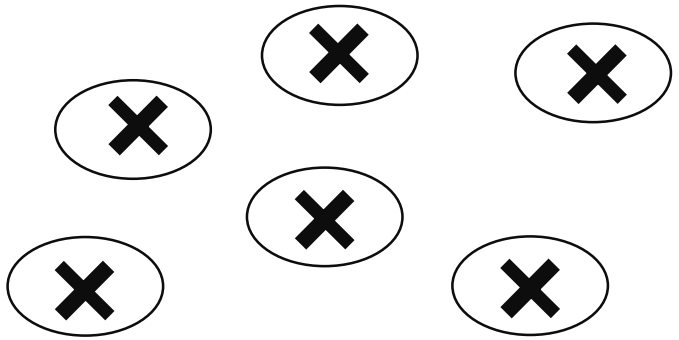
- 公開されているデータセットが少ない
- リアルなトラヒックを作るのは困難
  - 統計学的知識やパラメータ設定

目的

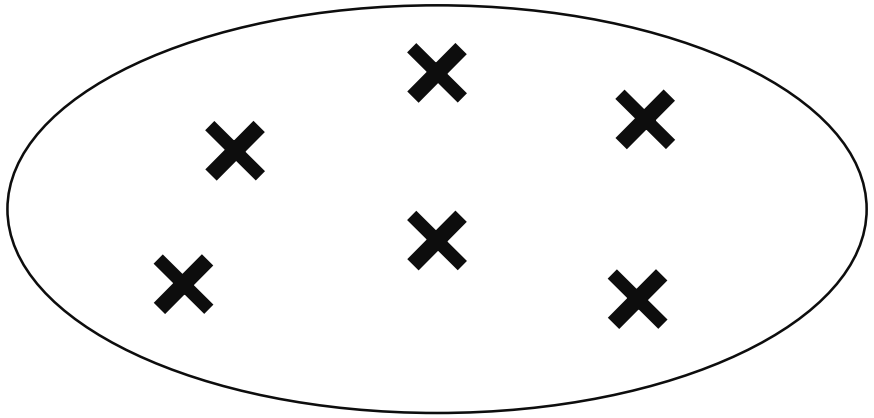
- 機械学習のGANを使って，誰でも簡単に多くのリアルなトラヒックを生成する

- 既存のトラヒックジェネレータとGANで作るデータの違い

既存のトラヒックジェネレータで作る似たデータ



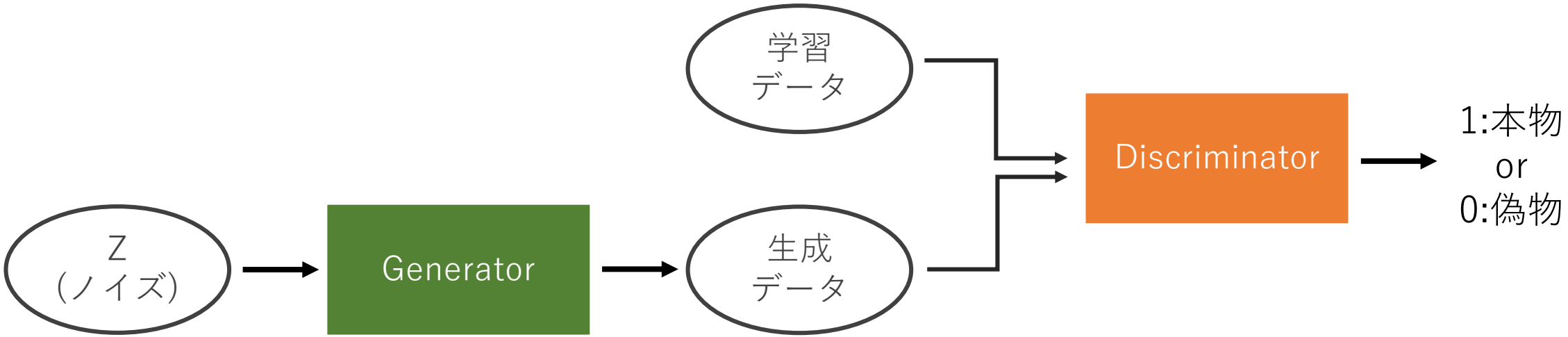
GANで作る似たデータ



×：サンプル
〇：生成しうる範囲

敵対的生成ネットワーク(GAN)[1]

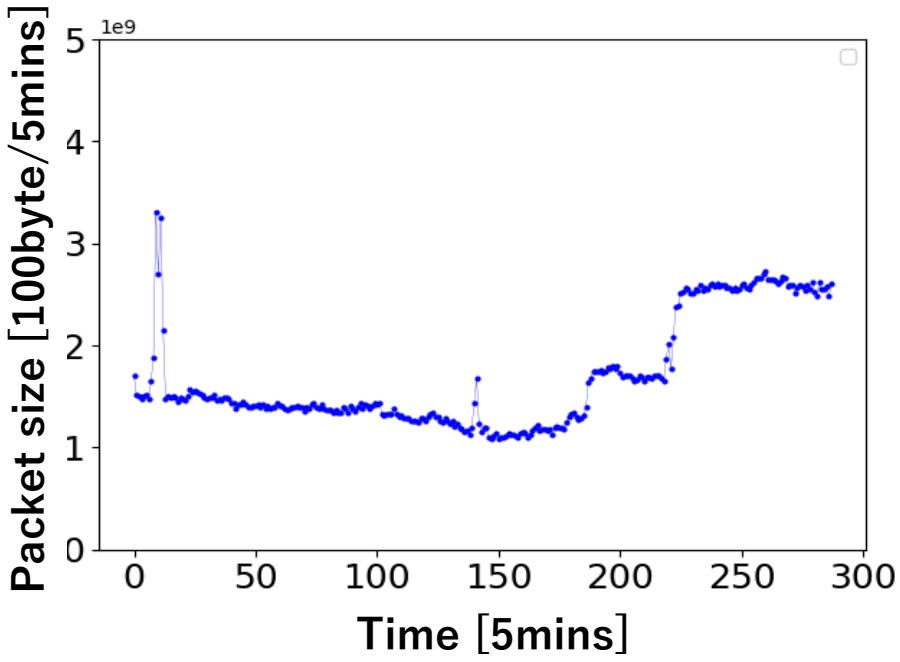
- 本物らしい画像生成の教師なし学習モデル
- 識別器(Discriminator)と生成器(Generator)の2つで構成されている
  - Discriminatorは学習データと生成データを判別する
    - 学習データなら1，生成データなら0となるように学習する
  - GeneratorはDiscriminatorを騙せるほど類似したデータを生成する
    - Discriminatorの出力が1に近づけるように学習する

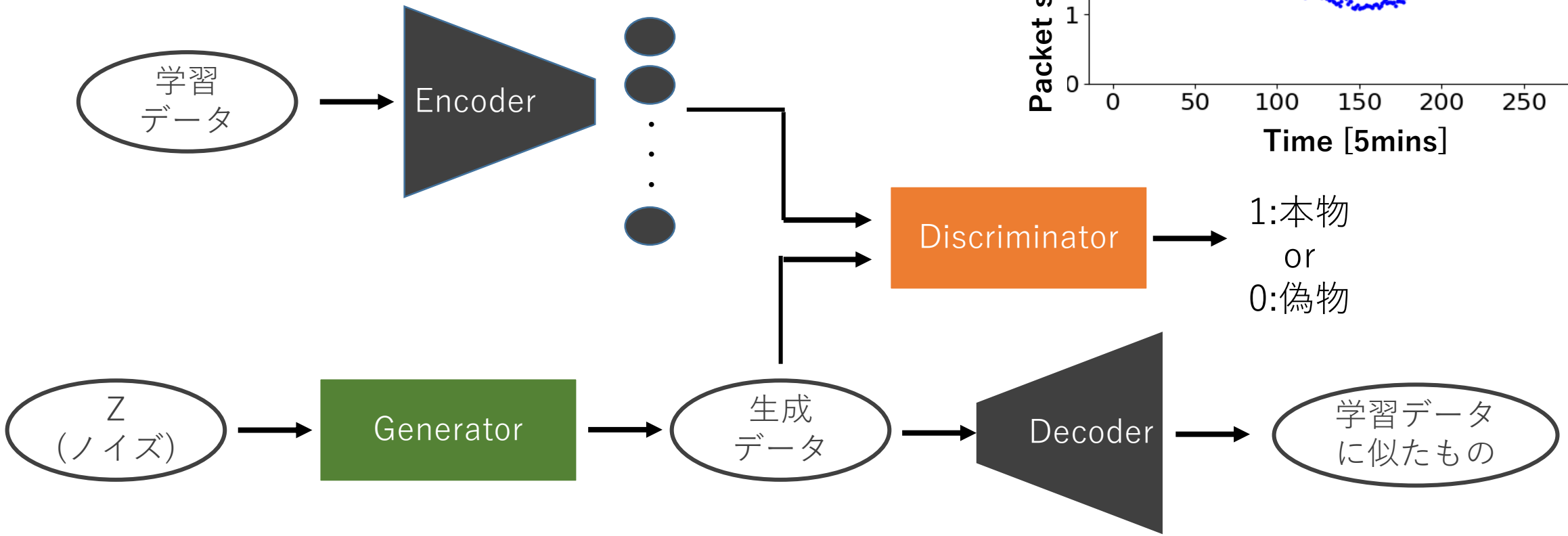


[1] I.J.Goodfellow et al. Generative Adversarial Nets. arXiv:1406.2661,2014.

実験手法

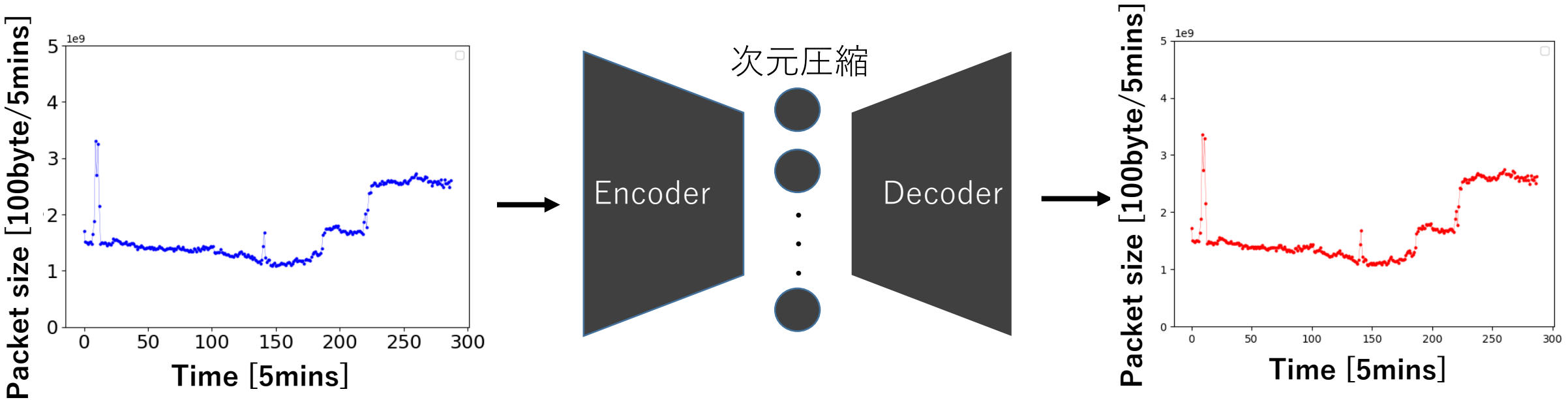
- Encoderを用いてGANに学習
  - 学習データは時系列(各次元はトラヒック量)
  - データにバースト性があり離散値のため
    通常のGANの学習が困難





Encoder

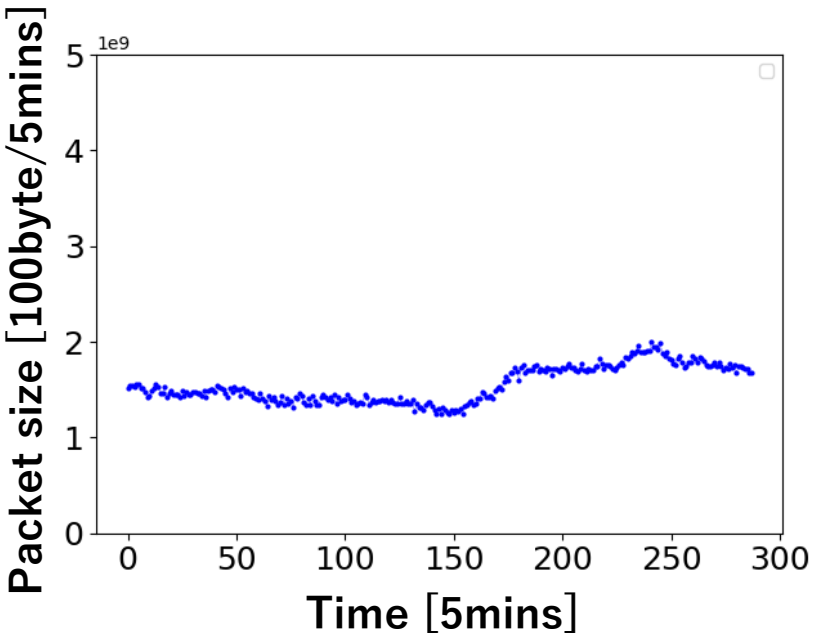
- データの次元圧縮として使われている
- Encoder，Decoder
  - データの次元圧縮，データの復元
- Encode後の次元は各学習データの特徴量を示している

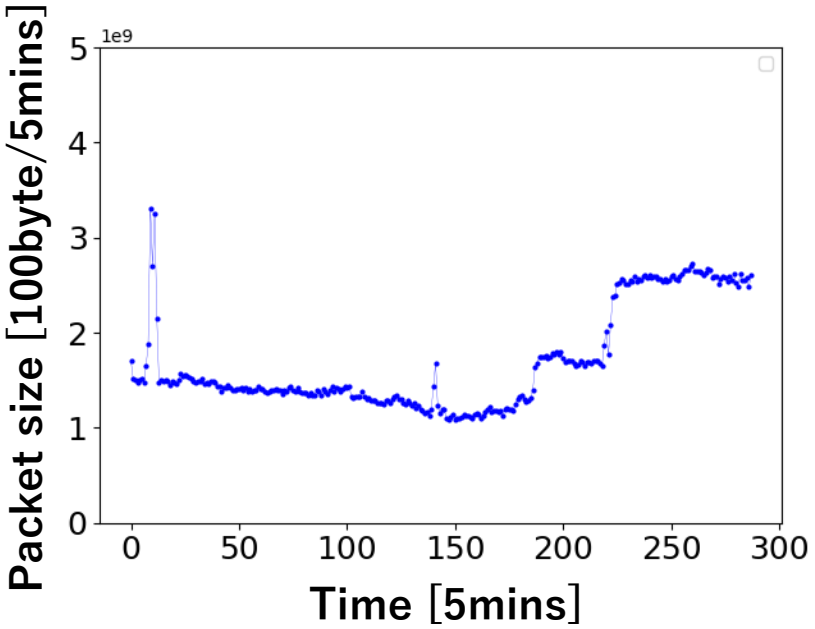


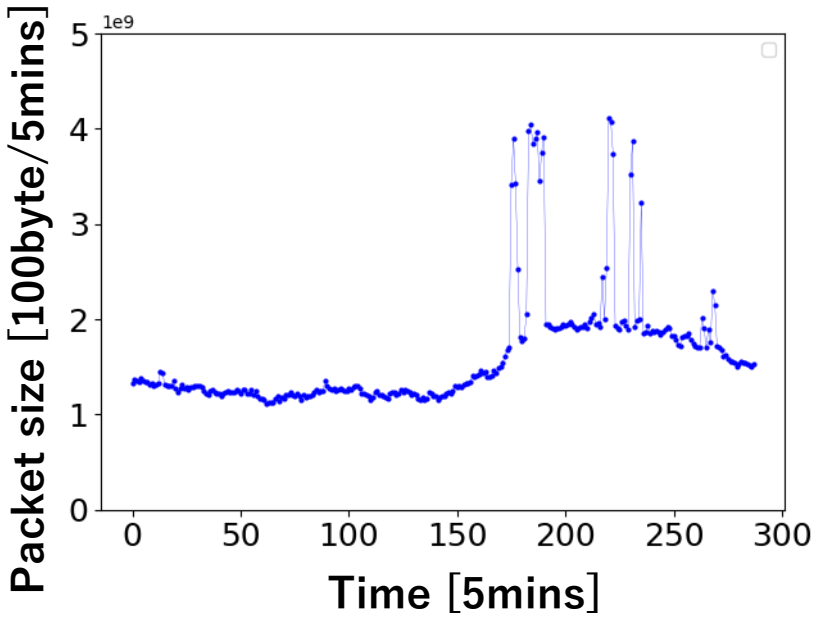
- 次元圧縮することでGANの生成範囲を狭めることができる

トラヒックデータ

- 公開されているトラヒックデータ[2]を学習させる
  - 1日の5分間毎に何バイト通信されたかを表したデータ
  - 1日分を1つの学習データ
  - 168日分





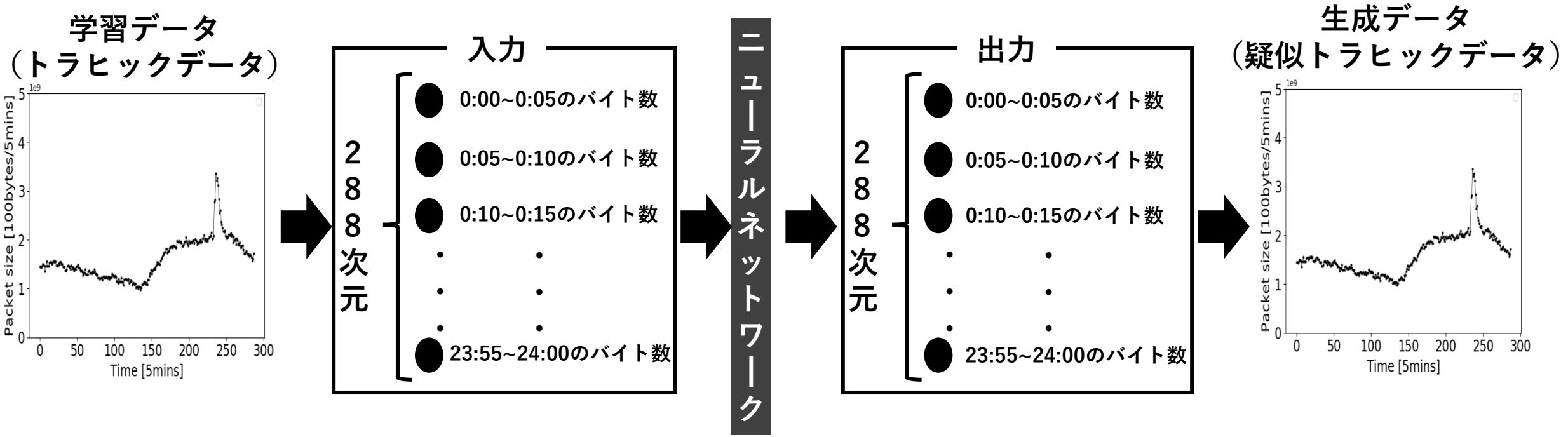


- トラヒックデータをGANによって生成する

[2] http://www.cs.utexas.edu/~yzhang/research/AbileneTM/

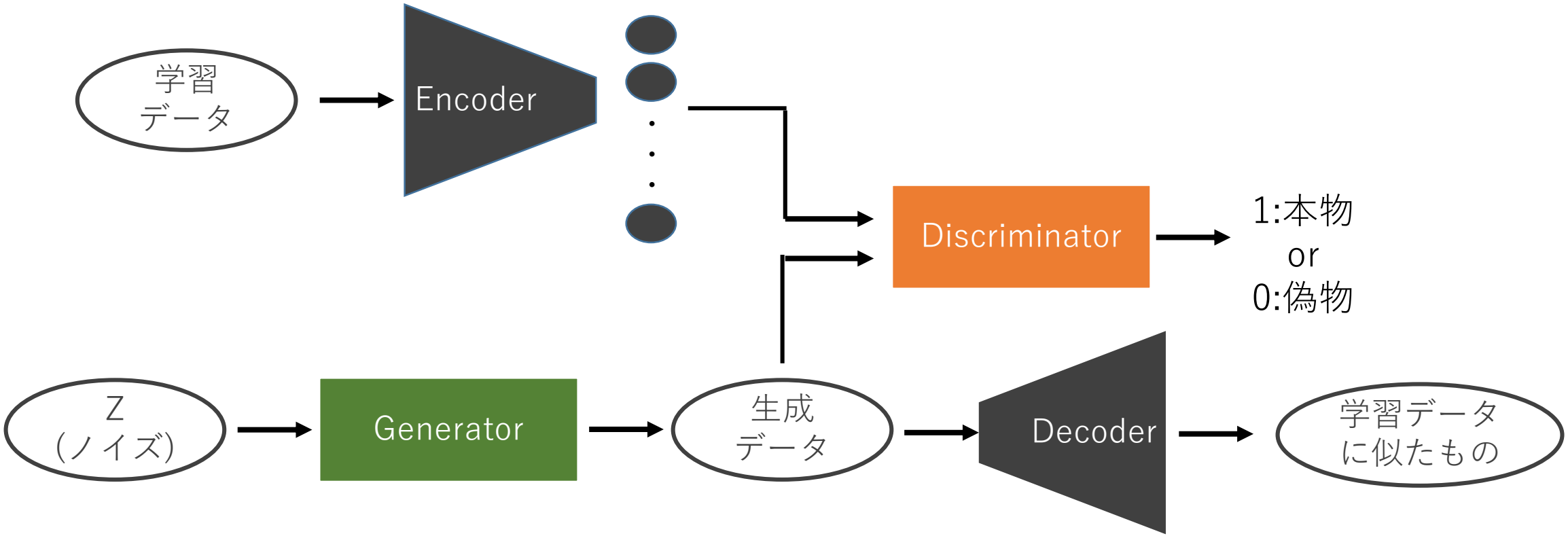
トラヒックデータの入出力

- トラヒックデータをニューラルネットワークに入力し，疑似トラヒックデータを出力する
  - 入力：時間間隔毎のバイト数を成分に持つベクトル
  - 出力：疑似トラヒックデータを表す同様のベクトル



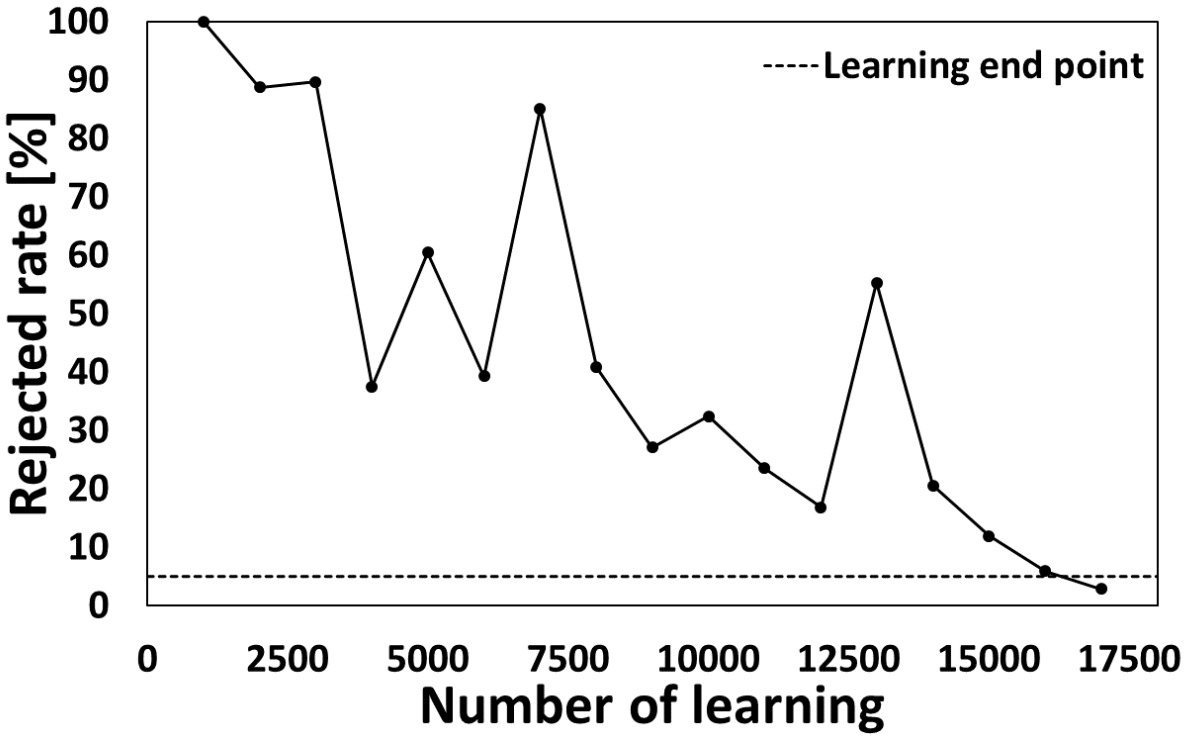
実験

- 学習方法
  1. トラヒックデータをEncoder，Decoderに学習させる
  2. EncoderとGANを組み合わせて学習させる（GANのみを学習）
  3. Generatorからの生成データをDecoderに入力し，疑似トラヒックデータを生成する



評価方法

- KS(コルモゴロフ-スミルノフ)検定
  - 仮説「標本XとYが同一の母集団の確率分布から発生」
  - 棄却された場合：確率分布が一致していない
- 5分間毎のバイト数における学習データと疑似トラヒックデータをそれぞれ168サンプルで検定する
  - 100セットを行う
- KS検定で棄却される割合が5%を下回ったとき，GANの学習を終了させる
  - 有意水準を5%に設定しているため



実験の構成

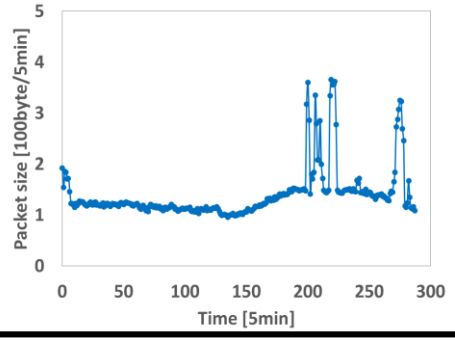
- 各ニューラルネットワークの構成

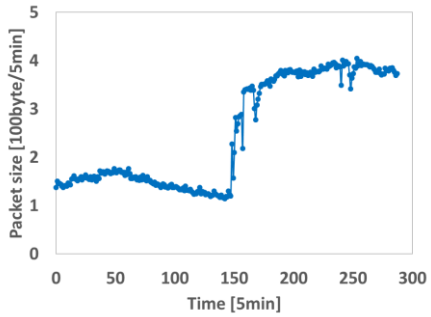
	Generator		Discriminator		Encoder		Decoder	
Layer	Units	Act.	Units	Act.	Units	Act.	Units	Act.
Input	2	-	5	-	288		5	-
Hidden	10	LReLU	50	LReLU	400	LReLU	50	LReLU
Hidden	30	LReLU	40	LReLU	200	LReLU	100	LReLU
Hidden	40	LReLU	30	LReLU	100	LReLU	200	LReLU
Hidden	50	LReLU	10	LReLU	50	LReLU	400	LReLU
Output	5	-	1	Sigmoid	5	Tanh	288	-

結果

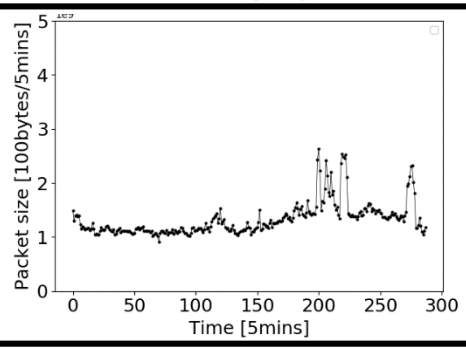
- 生成結果

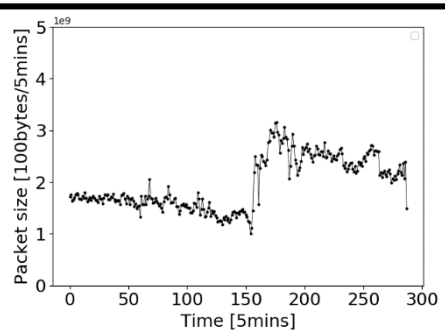
学習データ



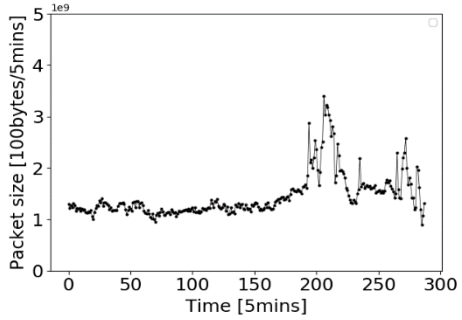


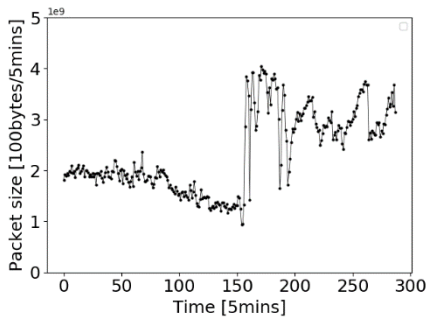
生成データ（類似）





生成データ（一部類似）

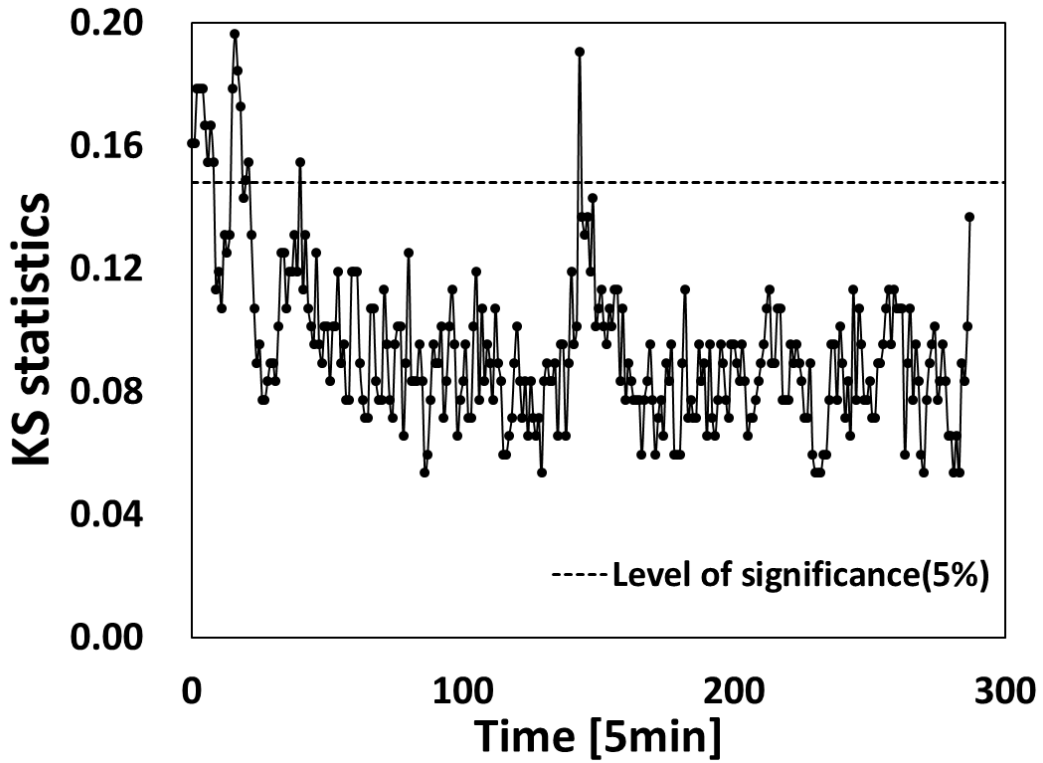




- 学習データに類似しているデータを生成していた

結果

- KS検定
  - 仮説「標本XとYが同一の母集団の確率分布から発生」
  - 棄却された場合：確率分布が一致していない



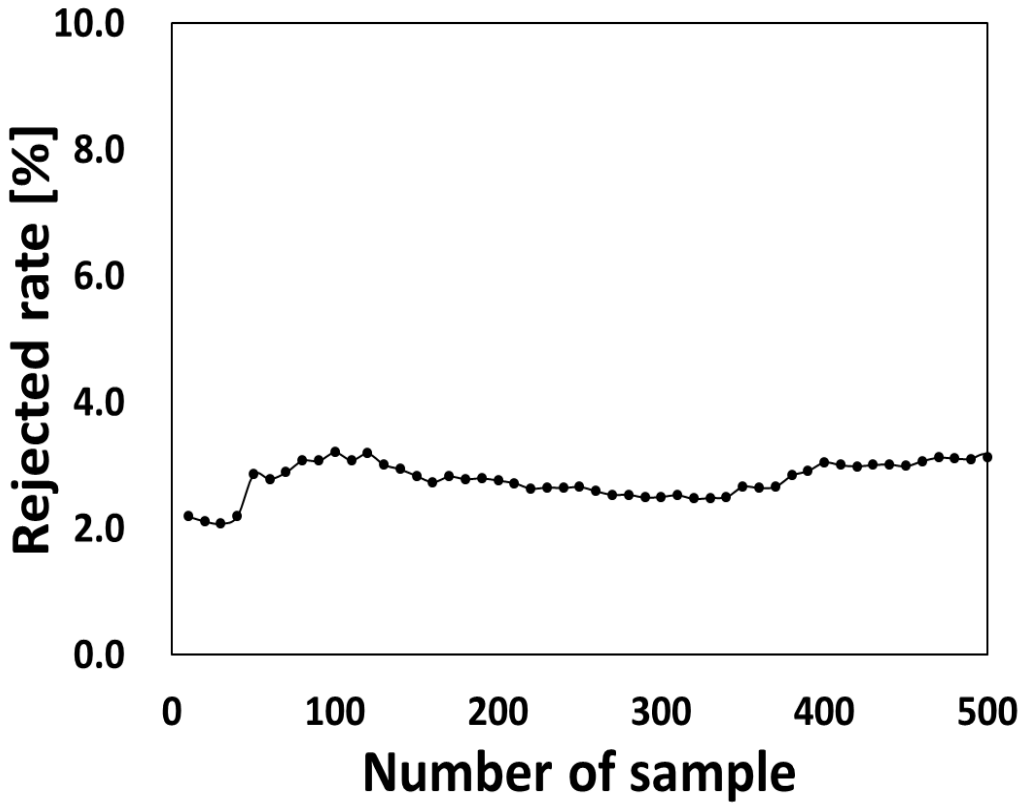
- 生成データは本物のデータの分布とかけ離れていない

結果

- GANはランダムな生成のため，必ずしもKS検定で棄却されない訳ではない
- GANで500セット生成し，棄却される割合を確認したところ，3.1%に収束した
- 有意水準5%に設定しており，実データで検定を行った場合，95%棄却されないが，GANの場合96.9%棄却されていない

- GANの生成は本物のデータとほぼ同等

→本物のデータの再現がほぼできている



まとめ・今後

- まとめ
  - Encoderを用いることでGANを使って，トラヒックデータ（学習データ）に類似したデータを作り出せることを確認した
- 今後
  - より精度の高いGANの構築
  - 多様性のあるトラヒックデータの生成
  - 時系列での評価