UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

5º Workshop de Medições da RNP - 2023
PMon

Estudo de Técnicas de Aprendizado Profundo para Reprodução de Dados de Monitoramento de Redes com Garantias de Anonimização

Coordenação: Antonio “Guto” Rocha (UFF)

Vinícius F. S. Mota (UFES)

Bolsista PMon: Iran F. Ribeiro (UFES)

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Agenda

- Motivação
- Redes Generativas Adversárias (GANs)
- Evolução das GANs: De geração de imagens a geração de dados de monitoramento
- Resultados preliminares
- Próximos passos

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Motivação

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Necessidade versus disponibilidade de dados para monitoramento de redes

Por um lado...

- Pesquisadores ávidos por dados
- acesso aos dados é requisito para ML
  - Predição de falhas em links
  - Detecção de ataques
  - entre outros...

Do outro lado...

- Esforço crescente para disponibilização
- RNP vem padronizando o acesso:
  - Catálogo de dados
  - GT Borescope
  - ....

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Necessidade versus disponibilidade de dados para monitoramento de redes

Por um lado...

- Pesquisadores ávidos por dados
- acesso aos dados é requisito para ML
  - Predição de falhas em links
  - Detecção de ataques
  - entre outros...

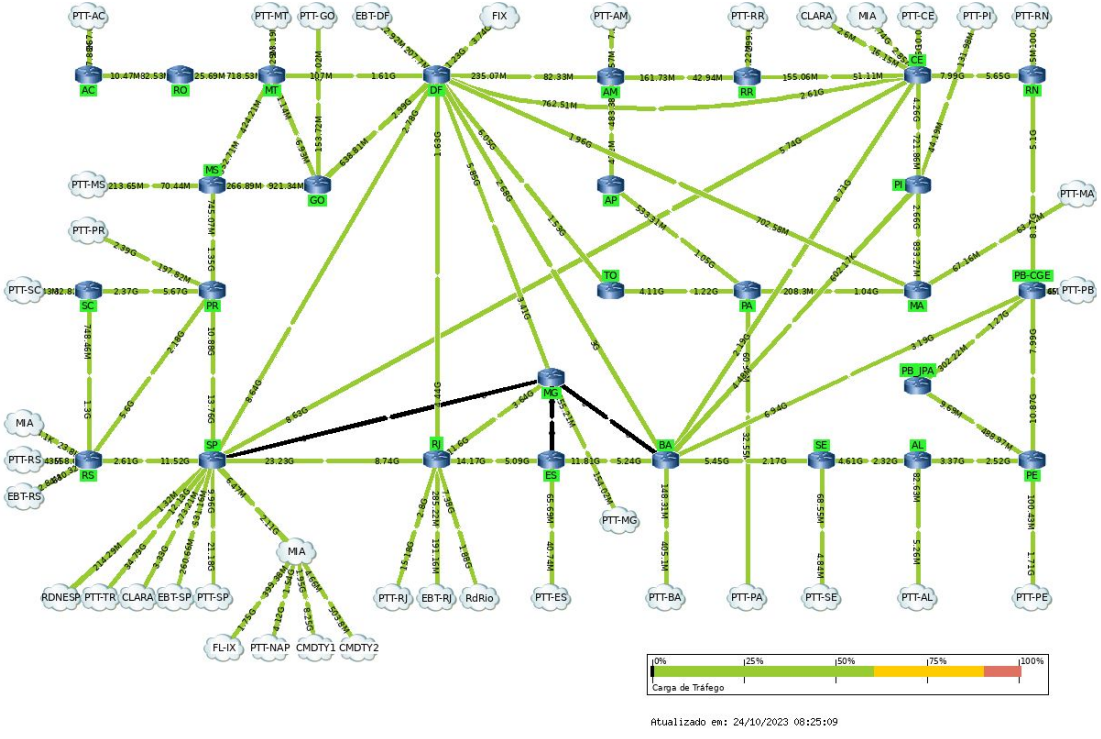
Do outro lado...

- Esforço crescente para disponibilização
- RNP vem padronizando o acesso:
  - Catálogo de dados
  - GT Borescope
  - ....

Dados podem conter informações sensíveis!!!

A disponibilização requer cuidados com a aderência à **LGPD** e questões de segurança!!!!

No contexto de monitoramento: Amostras de fluxos de dados no Backbone



Características:

- Temporalidade importa
- Campos sensíveis
  - IPsrc, IPdest, port, …

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Uma abordagem

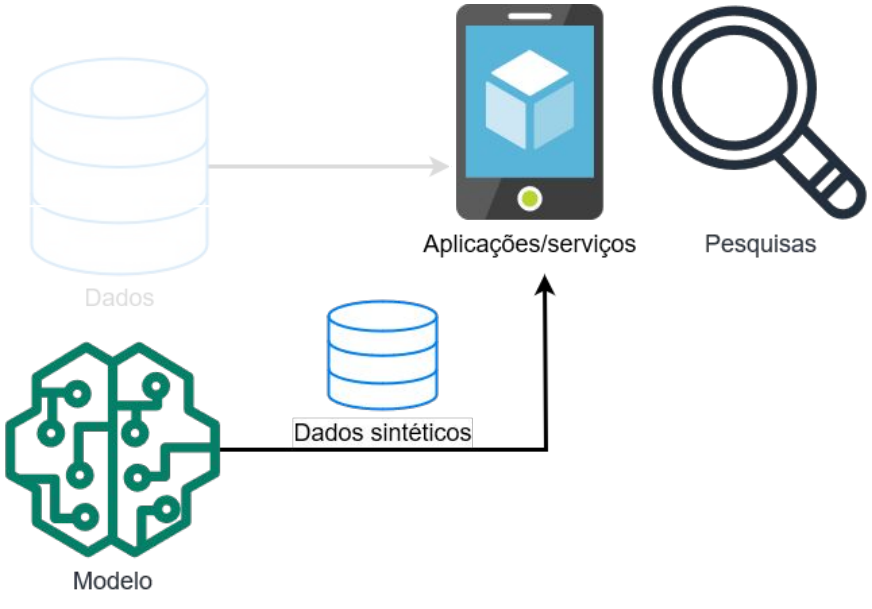
***Dados sintéticos* que mimetizam os dados reais**

- Já amplamente utilizado em simulação de redes para gerar
  - Tráfego de dados
  - Fluxos seguindo alguma distribuição
  - Mobilidade dos nós da rede

Motivação

***Dados sintéticos* por modelos de simulação**

- Desvantagens
  - Parametrização pode ser difícil
  - E, principalmente, não representar um cenário real

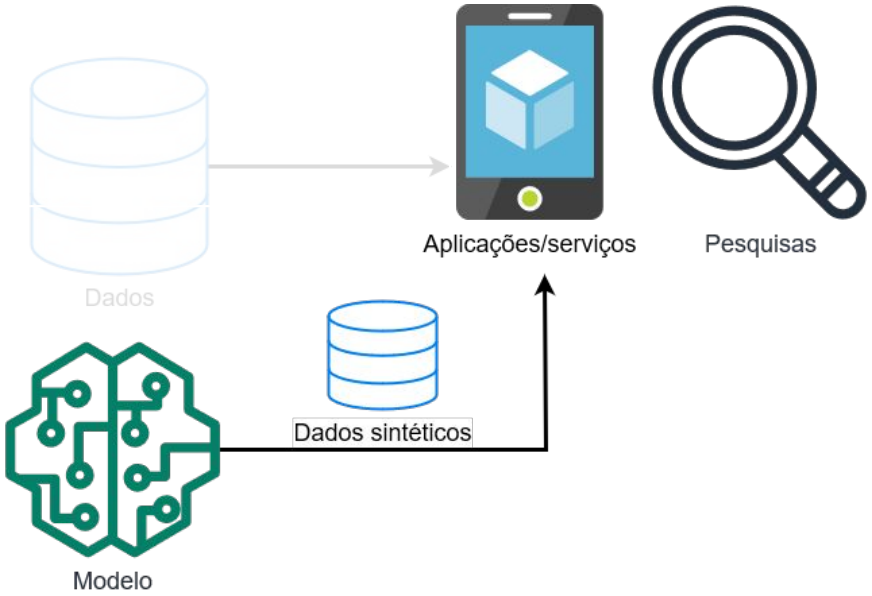


Motivação

***Dados sintéticos* por modelos de simulação**

- Desvantagens
  - Parametrização pode ser difícil
  - E, principalmente, não representar um cenário real

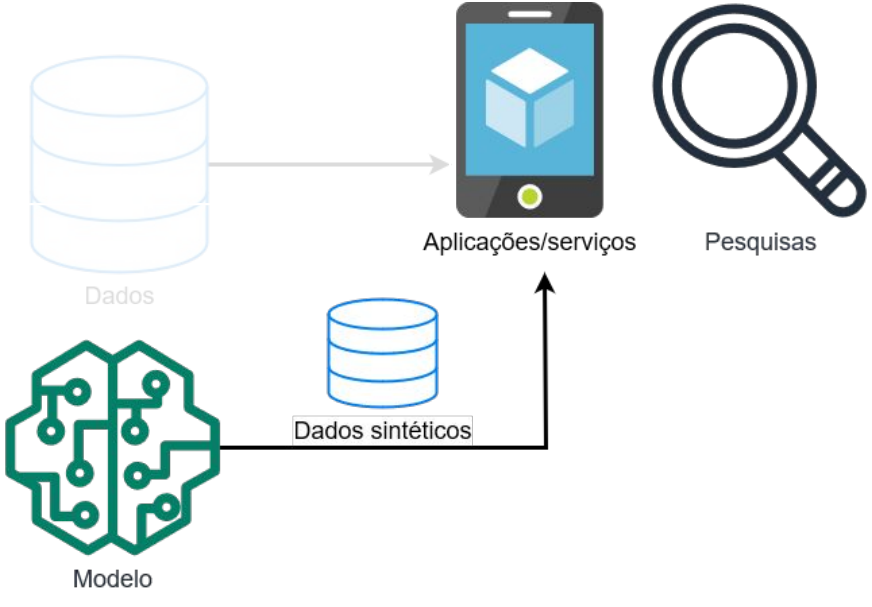
- Vantagens
  - Maior reprodutibilidade



Motivação

***Dados sintéticos* por modelos generativos**

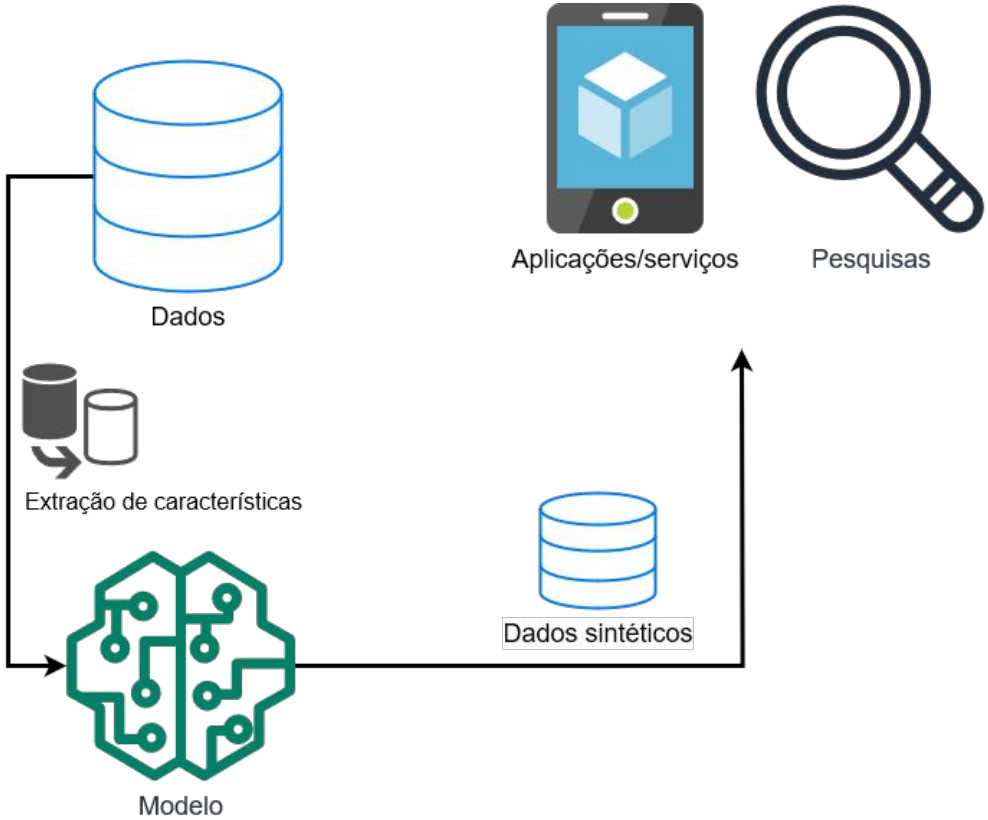
- Visa capturar características de um dado real
- Preservar características estatísticas mantendo a privacidade de dados sensíveis
- Aumentar base de dados (data augmentation)
- Reprodutibilidade



Motivação

Pergunta: Dada as limitações, como obter um modelo gerador que capture principais características dos datasets reais?

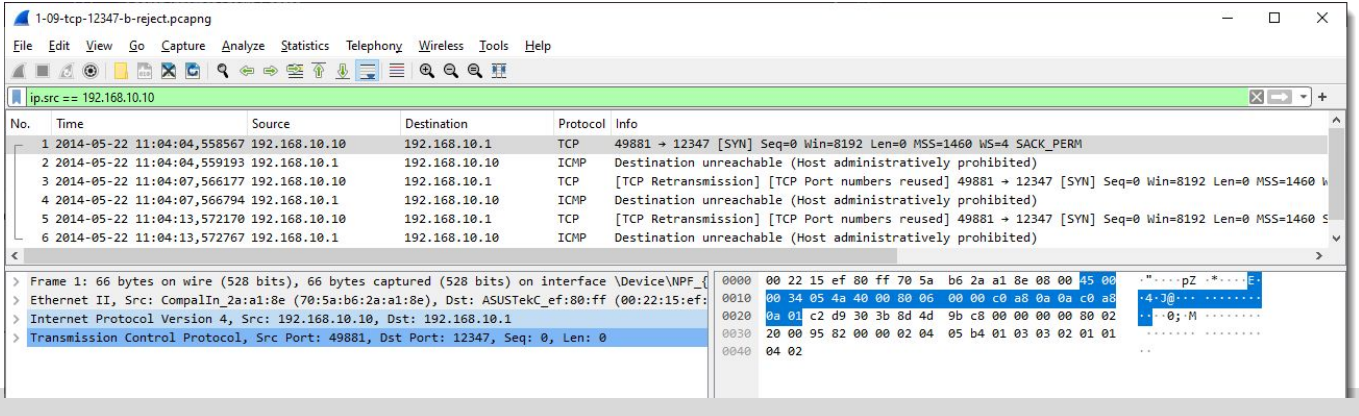
Hipótese: **Redes Generativas Adversárias** podem gerar séries temporais multivariadas.



Objetivo

A partir de fluxos de rede reais, gerar fluxos sintéticos que:

- Preserve a temporalidade dos dados
- O modelo possa gerar dados sintéticos com diferenças em relação ao real, mas as distribuições de todos os campos sejam semelhantes aos dados reais



UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

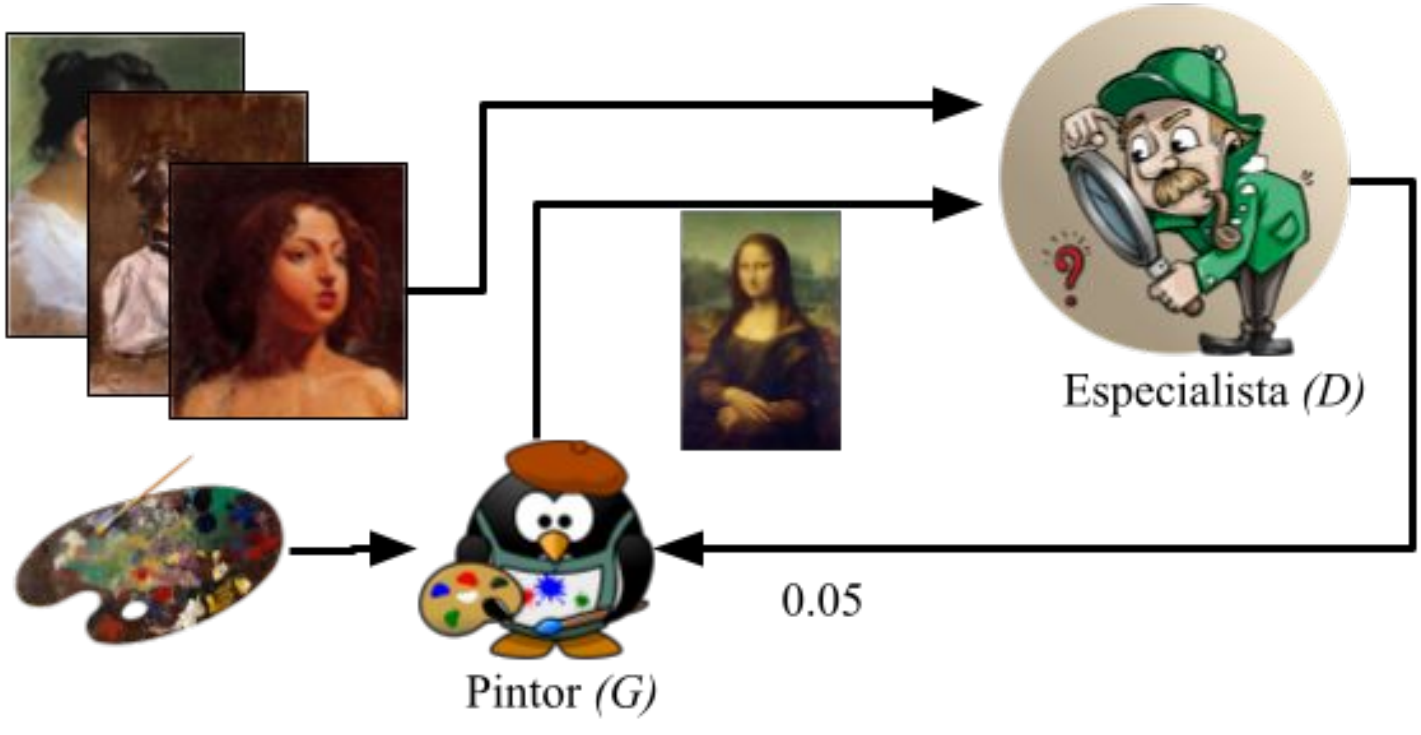
Metodologia

- Revisão da literatura em busca de ferramentas que gerem dados tabulares
- Revisão de datasets públicos de fluxo de rede
- Captura de fluxo de dados “in loco”. Neste caso, foi utilizado uma coleta de monitoramento de fluxos interna, como parte do trabalho de doutorado de integrante do grupo de pesquisa.
- Testes e avaliações de algoritmo para geração de dados de fluxo.
    - Algoritmos e modelos atuais funcionam para dados de fluxo de rede?
    - Caso contrário, é possível adaptá-los ?

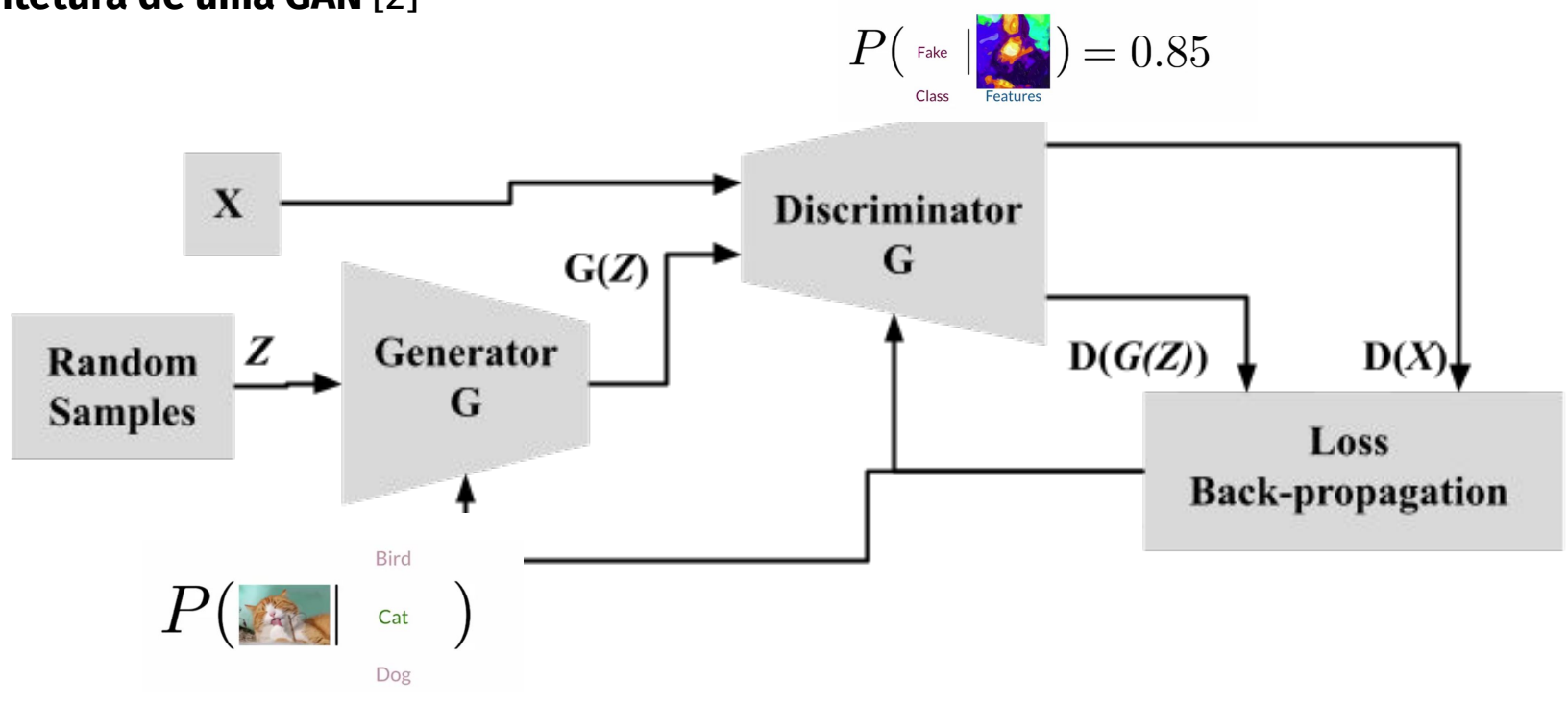
UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Redes Generativas Adversárias (GANs)

Intuição



Arquitetura de uma GAN [2]



UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

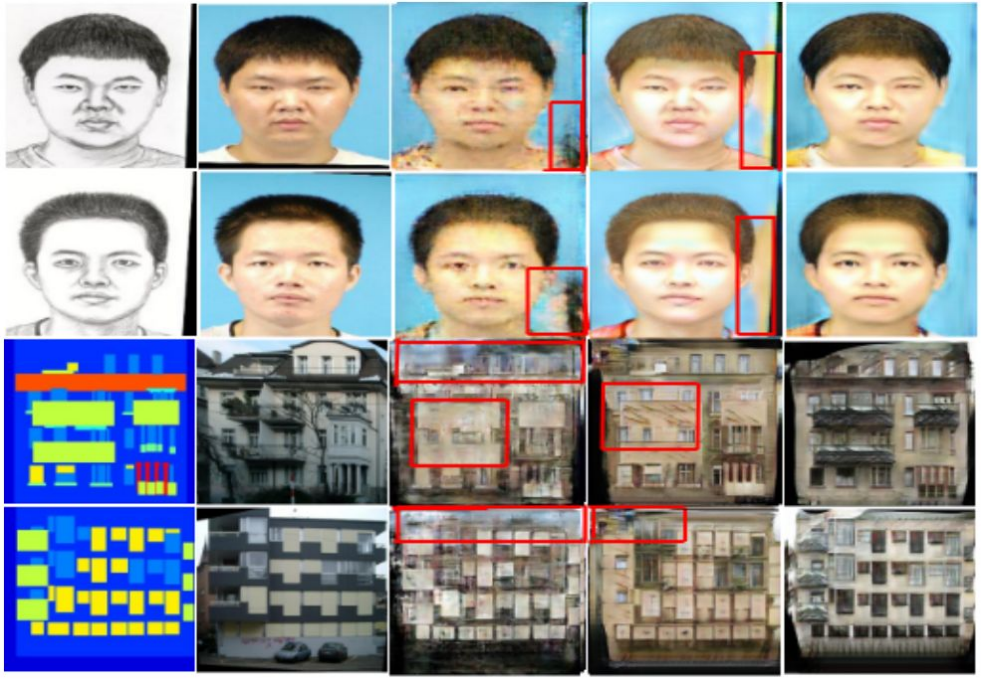
Aplicações

- Geração de imagens realistas [3]

Geração de rostos de pessoas

Aplicações

- Geração de imagens
- Tradução de Imagem para imagem[4]



Linhas 1 e 2: de esboço para imagem reais.
Linhas 3 e 4: imagens rotuladas para fachadas de prédios.

Aplicações

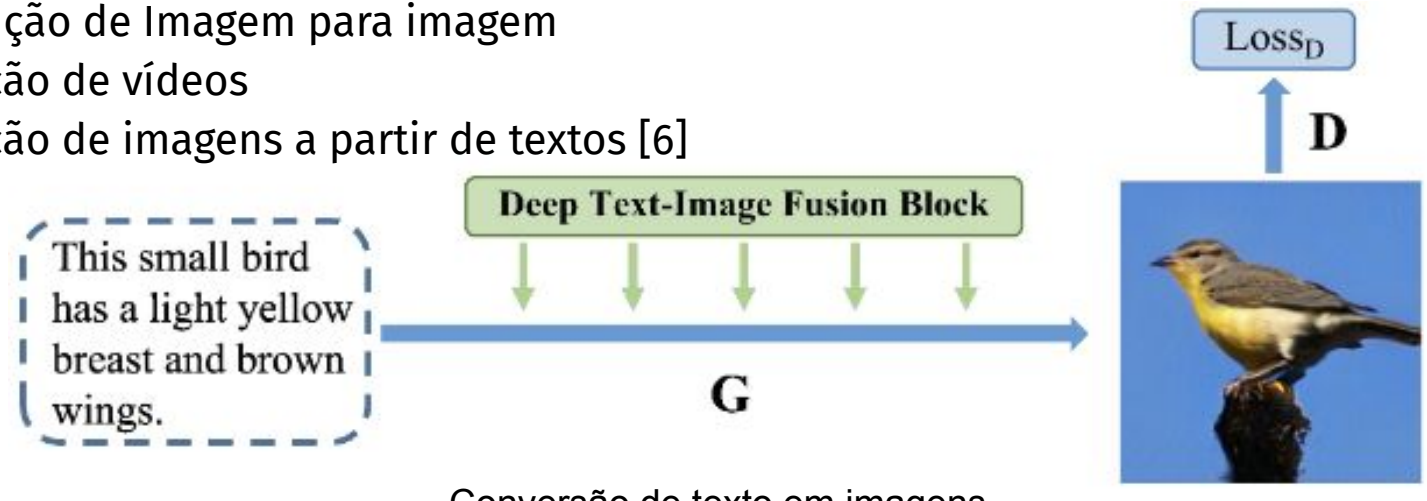
- Geração de imagens
- Tradução de Imagem para imagem
- Geração de vídeos [5]

Input | DeblurGAN-v2 [1] | Ours

Correção de foco em vídeos.

Aplicações

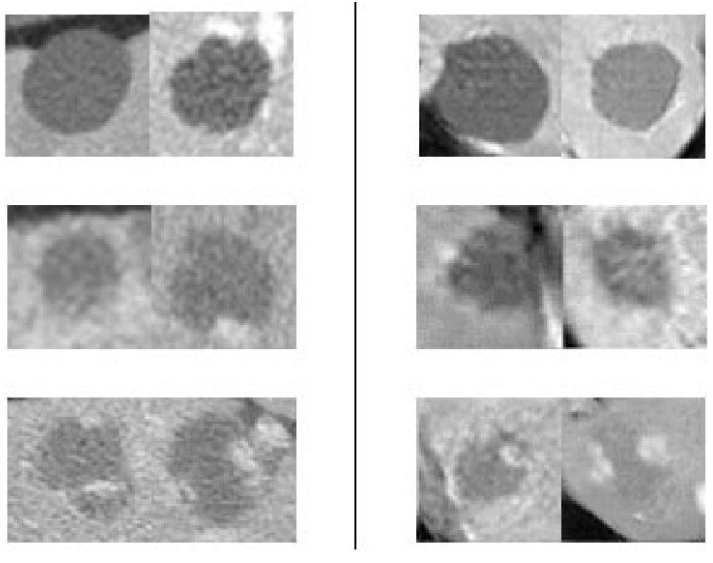
- Geração de imagens
- Tradução de Imagem para imagem
- Geração de vídeos
- Geração de imagens a partir de textos [6]



Conversão de texto em imagens

Aplicações

- Geração de imagens
- Tradução de Imagem para imagem
- Geração de vídeos
- Geração de imagens a partir de textos
- Data augmentation [7]



Geração de imagens de câncer de fígado

Problemas comuns

- **Dimensão do tempo**: Originalmente não consideram a **temporalidade** dos dados ;
- **Instabilidade**: Uma alteração **pequena** no Gerador pode provocar uma **grande** alteração no Discriminador;
- **Métricas de avaliação**: Não há um consenso em termos da melhor métrica para avaliar uma GAN

IGLESIAS, Guillermo; TALAVERA, Edgar; DÍAZ-ÁLVAREZ, Alberto. A survey on GANs for computer vision: Recent research, analysis and taxonomy. **Computer Science Review**, v. 48, p. 100553, 2023.

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Evolução das GANs: De geração de imagens a geração de dados de monitoramento

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

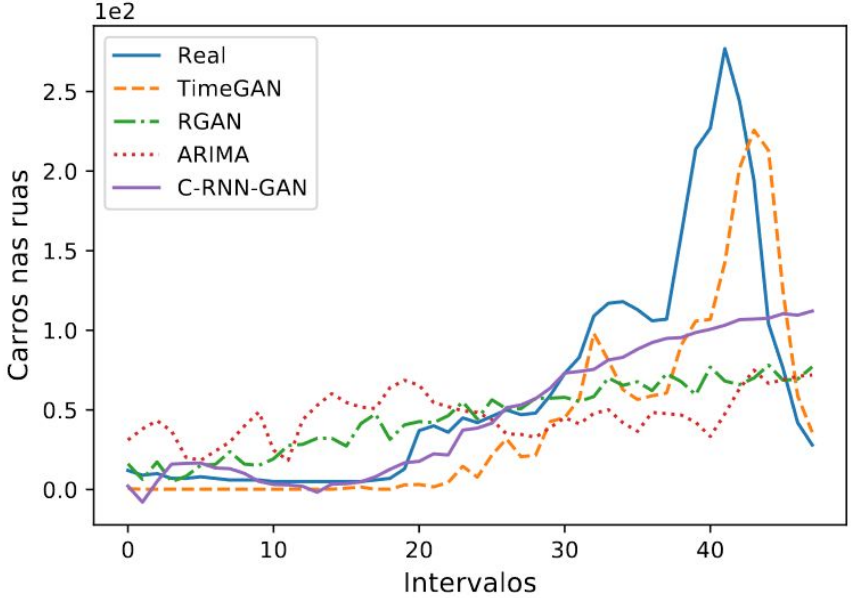
GANs para geração Séries Temporais Sintéticas

- Similaridade: dados sintéticos devem possuir as principais características dos dados reais
- Variabilidade: cada dado sintético gerado deve ser diferente
- Privacidade: não deve ser possível inferir informações dos usuários a partir dos dados sintéticos

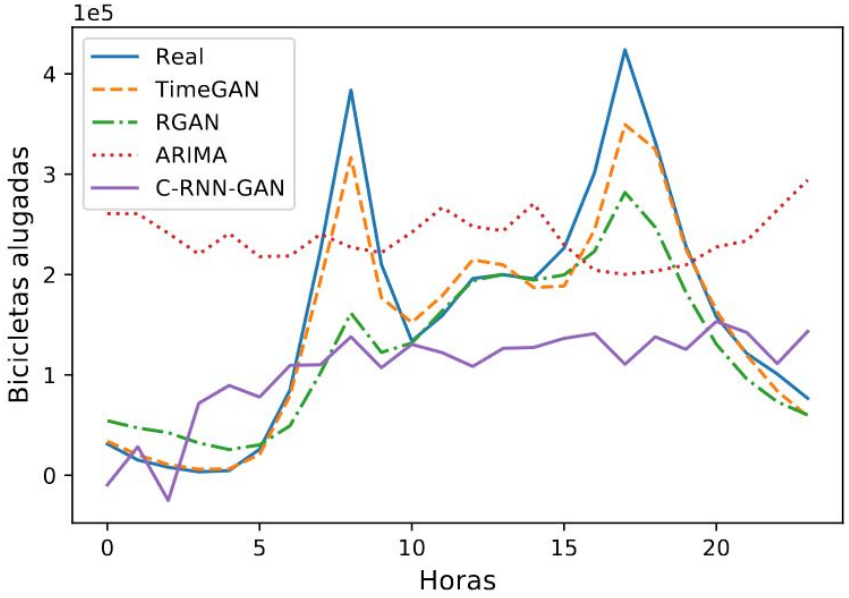
Aplicações

- Dados medicinais: [8]
- Música: [9]
- Mobilidade: [1, 10]
- Tráfego de rede: [11]

Estudos preliminares do grupo de pesquisa - Séries temporais

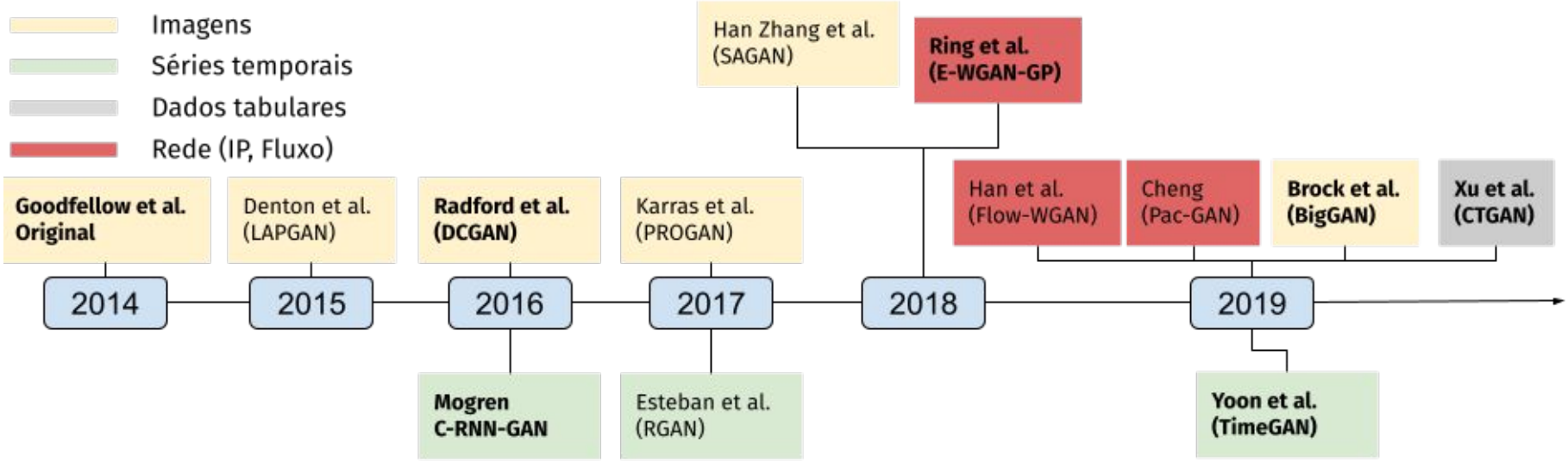


Número de carros em uma rua numa sexta-feira, geradas por diferentes modelos [1]

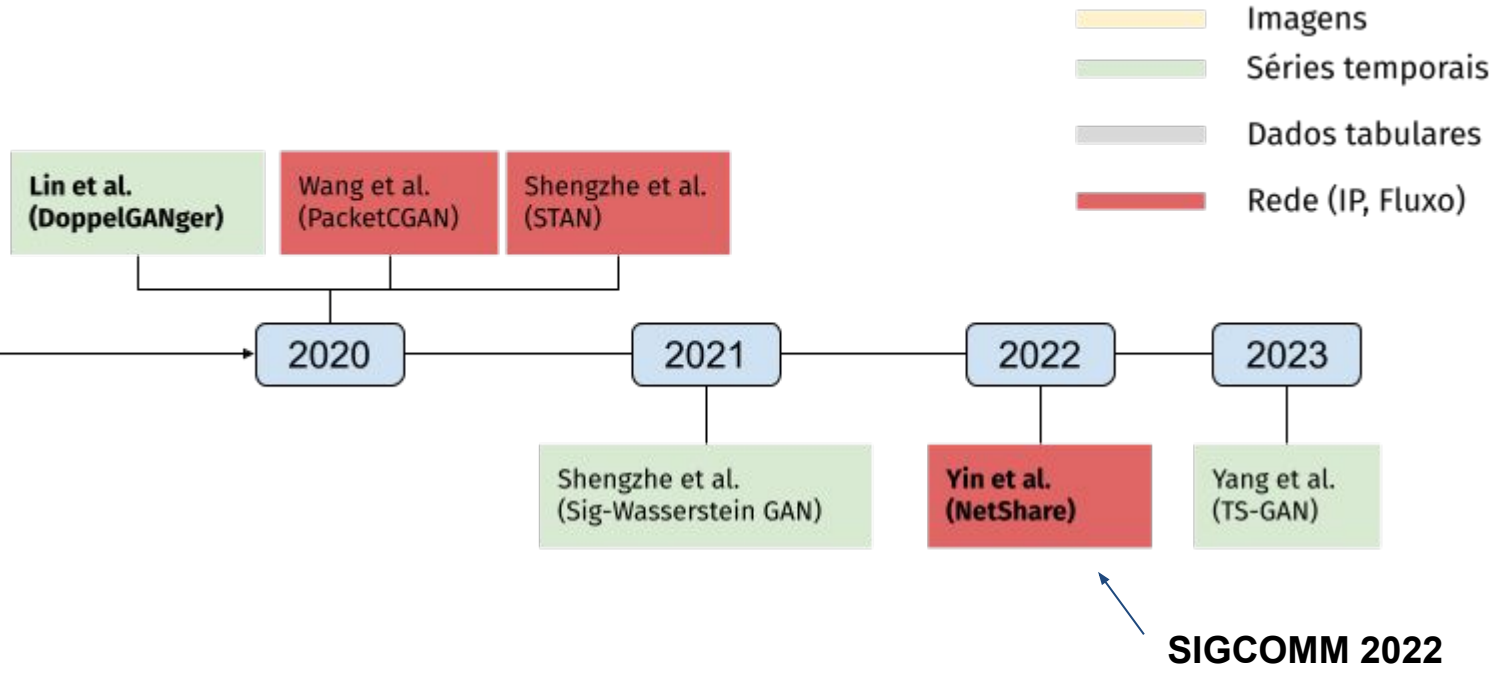


Total de bicicletas alugadas na sexta-feira, geradas por diferentes modelos [1]

Visão Geral GANs e suas aplicações (Etapa 1 deste P-Mon)

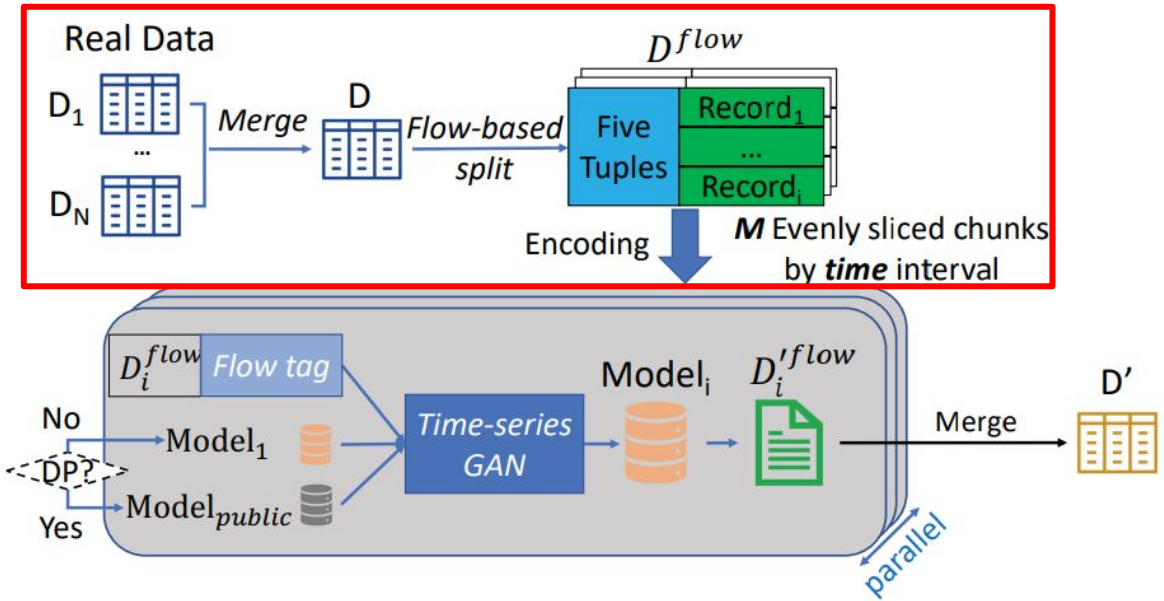


Visão Geral GANs e suas aplicações (Etapa 1 deste P-Mon)



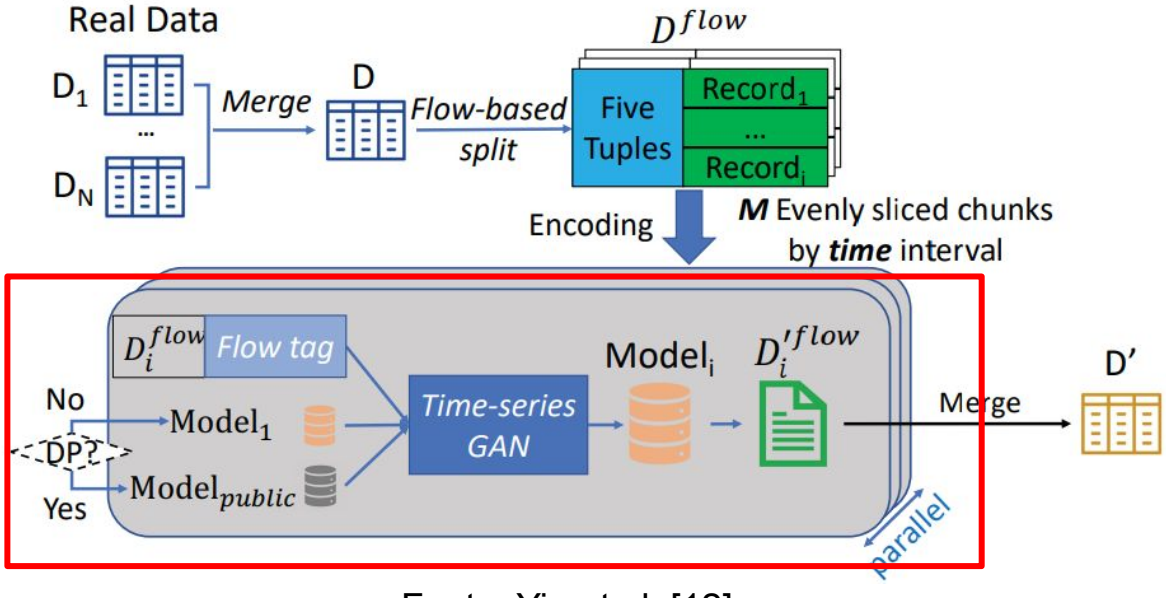
Aprendizado profundo para geração de dados de monitoramento

Netshare



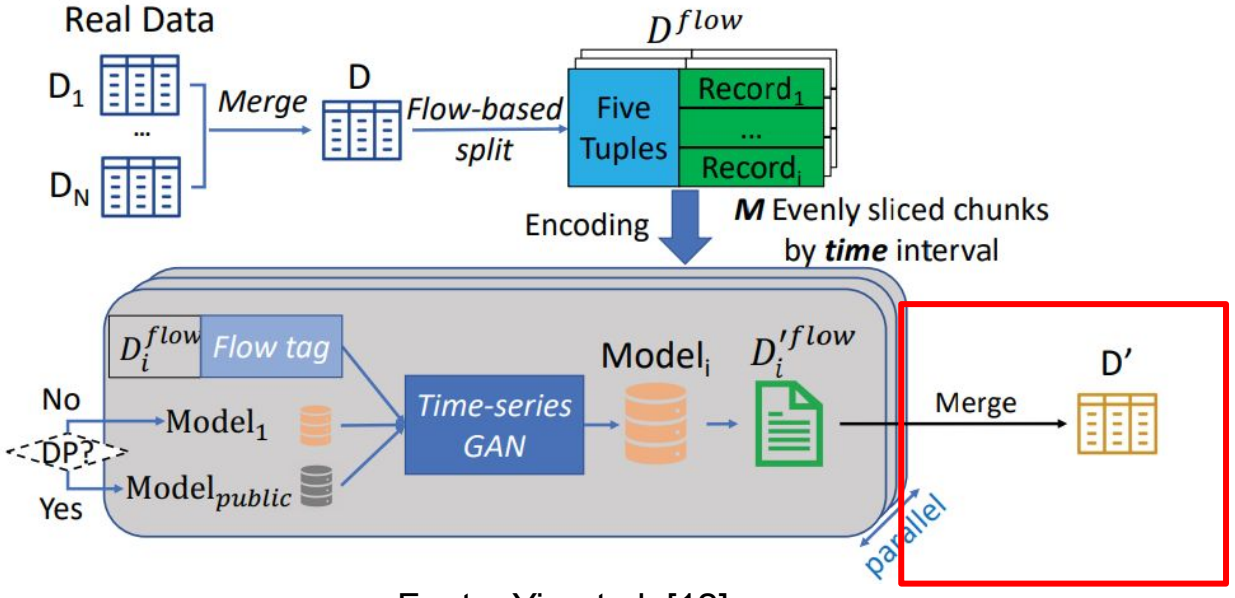
Fonte: Yin et al. [12]

Aprendizado profundo para geração de dados de monitoramento



Fonte: Yin et al. [12]

Aprendizado profundo para geração de dados de monitoramento



Fonte: Yin et al. [12]

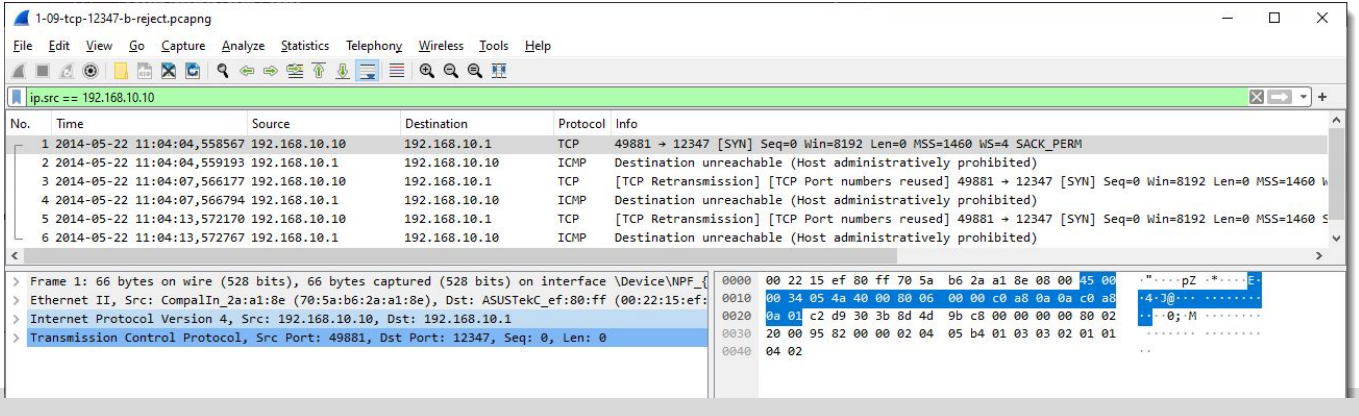
UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

GANs:
Gerando dados de fluxo de Rede
Primeiros resultados

Revisando o Objetivo

A partir de fluxos de rede reais, gerar fluxos sintéticos que:

- Preserve a temporalidade dos dados
- O modelo possa gerar dados sintéticos com diferenças em relação ao real, mas as distribuições de todos os campos sejam semelhantes aos dados reais



UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Primeiros testes

Estudo da Ferramenta NetShare

- Aplicação em dois datasets:
  - Dataset público conhecido como UGR'16
    - 100Kb → Apenas para entender a ferramenta
  - Dataset coletado internamente
    - ~100kb, 5mb, 10mb, 20mb
    - ~2 s, 7 min, 14 min, 27 min
- Intel i9 @3,70Ghz x 20, 128 Gig mem, NVidia GeForce RTX 4090

Métricas:

- Comparação visual das distribuições dos dados
- Score NetShare[1]

[1] https://github.com/netsharecmu/SDMetrics_timeseries

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Fluxo de rede (UGR'16)

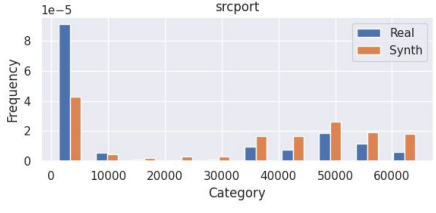
NetFlow v9 collectors in a Spanish ISP

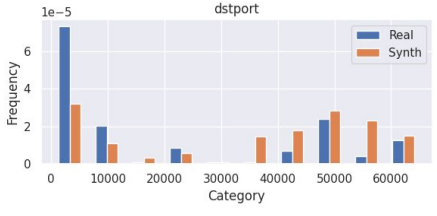
	srcip	dstip	srcport	dstport	proto	ts	td	pkt	byt	type
7	714257883	719036341	62058	54331	UDP	1.458298e+15	305.760	9008	4362440	background
8	2968793991	719036242	46556	21	TCP	1.458298e+15	305.664	204	12060	background
9	3564553392	719035635	38123	22002	TCP	1.458298e+15	303.640	120	6528	background
13	618839282	719035635	36921	40004	TCP	1.458298e+15	300.880	300	16664	background
14	621076693	719036249	161	63061	UDP	1.458298e+15	304.636	521	78848	background

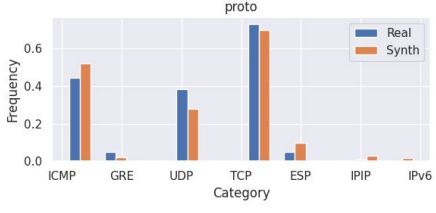
Dataset disponível em: https://nesg.ugr.es/nesg-ugr16/

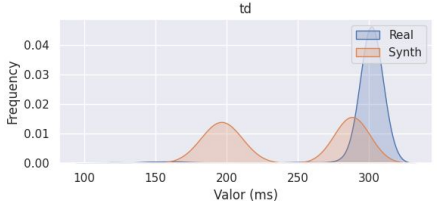
Fluxo de rede (UGR'16)

Utilizando apenas dados de testes: ~100Kb - 3 minutos de fluxo - ~1 min para gerar o modelo









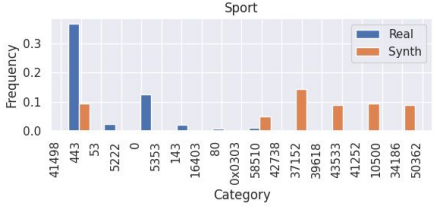
UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

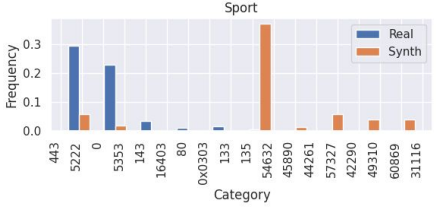
Fluxo de rede

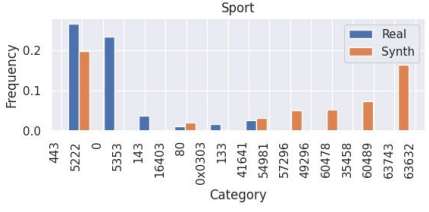
Dataset privado (coletado internamente - NDA)

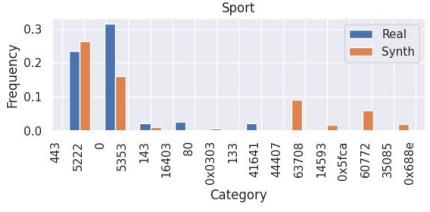
	SrcAddr	DstAddr	Sport	Dport	Proto	StartTime	Dur	TotPkts	TotBytes	State
0	3232236050	3232239482	41498	80	tcp	1.694437e+09	0.900379	3	208	RST
1	3232236050	3232249800	48971	443	tcp	1.694437e+09	0.901932	3	208	RST
2	3232236740	3232257577	35502	443	tcp	1.694437e+09	3.764490	3536	3337189	CON
3	3232251593	3232237444	443	60418	tcp	1.694437e+09	0.111999	553	714050	CON
4	3232258754	3232243501	53	55818	udp	1.694437e+09	0.000000	1	150	INT

Dados categóricos (Source port)





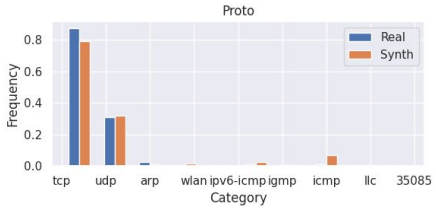




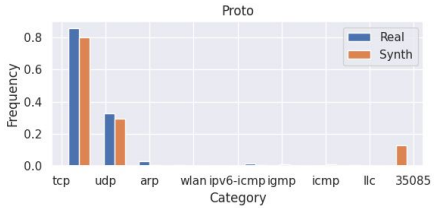
Tamanho	100 kb	5 mb	10 mb	20 mb
Tempo Fluxo	2s	7 min	14 min	27 min
Tempo modelo	1”30	1h	2h	3h40

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

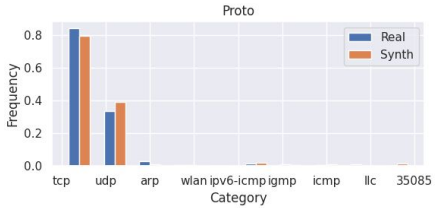
Dados categóricos (Protocolo)



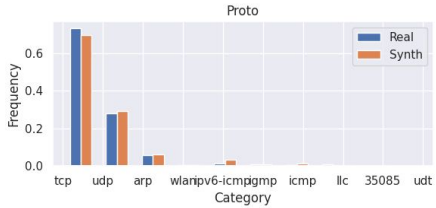
100 kb
2s



5 mb
7 min

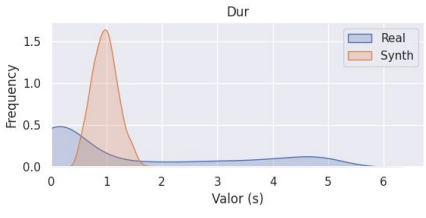


10 mb
14 min

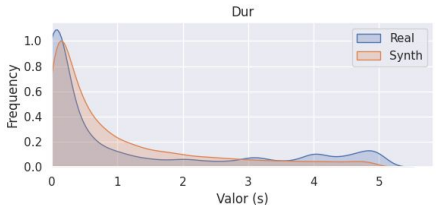


20 mb
27 min

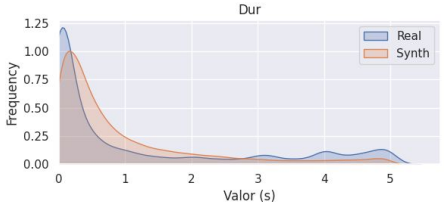
Dados contínuos (Duração)



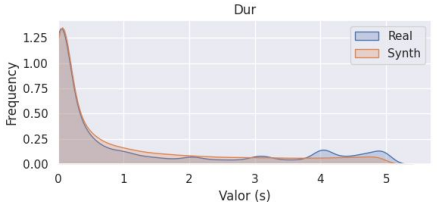
100 kb
2s



5 mb
7 min



10 mb
14 min



20 mb
27 min

Tamanho do dataset x “Score de fidelidade”

(SrcAddr, Sport, Proto)

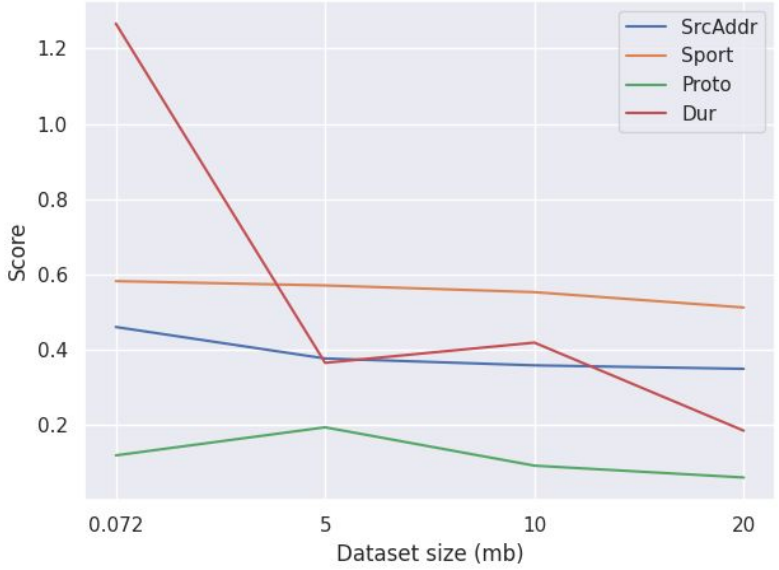
Melhor: 0

Pior: 1

Dur:

Melhor: 0

Pior: Infinito.



UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Próximos passos

- Acesso a dados fornecidos pela RNP
- Aplicações e cenários que utilizem os dados sintéticos
- Explorar novos modelos para geração de dados de monitoramento
- Produzir um relatório/artigo comparativo das técnicas de geração de dados baseadas em aprendizado profundo.

Indo além do P-Mon....

Projeto de Iniciação científica, com bolsa projeto MCTI/FAPESP Porvir 5G, visa avaliar métricas existentes para mensurar similaridade e variabilidade dos dados.

A. As métricas refletem o que acontece nos cenários reais?
B. Métricas de distância de distribuição são boas para observar a similaridade, mas não a temporalidade dos dados.
C. Propor técnicas mais eficientes para avaliação de datasets sintéticos de séries temporais.
D. Deve se também medir a anonimização gerada pelos dados sintéticos, garantindo que informações sensíveis não vaze pelo modelo.

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Dúvidas?

Iran F. Ribeiro - iran.ribeiro@edu.ufes.br
Vinícius FS Mota - vinicius.mota@inf.ufes.br

Núcleo de Estudos em Redes Definidas por Software

UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Referências

[1] RIBEIRO, Iran F. et al. Uma abordagem para geração de séries temporais de mobilidade urbana baseada em aprendizado profundo. In: **Anais do V Workshop de Computação Urbana**. SBC, 2021. p. 251-264.

[2] GOODFELLOW, Ian et al. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

[3] KARRAS, Tero; LAINE, Samuli; AILA, Timo. A style-based generator architecture for generative adversarial networks. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. 2019. p. 4401-4410.

[4] BABU, Kancharagunta Kishan; DUBEY, Shiv Ram. CSGAN: Cyclic-synthesized generative adversarial networks for image-to-image transformation. **Expert Systems with Applications**, v. 169, p. 114431, 2021.

[5] REN, Xuanchi; QIAN, Zian; CHEN, Qifeng. Video deblurring by fitting to test data. **arXiv preprint arXiv:2012.05228**, 2020.

[6] TAO, Ming et al. Df-gan: A simple and effective baseline for text-to-image synthesis. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2022. p. 16515-16525.

Referências

[7] FRID-ADAR, Maayan et al. Synthetic data augmentation using GAN for improved liver lesion classification. In: **2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)**. IEEE, 2018. p. 289-293.

[8] DASH, Saloni et al. Medical time-series data generation using generative adversarial networks. In: **Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18**. Springer International Publishing, 2020. p. 382-391.

[9] DONG, Hao-Wen et al. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2018.

[10] SHIN, Seungjae et al. User mobility synthesis based on generative adversarial networks: A survey. In: **2020 22nd International Conference on Advanced Communication Technology (ICACT)**. IEEE, 2020. p. 94-103.

[11] CHENG, Adriel. PAC-GAN: Packet generation of network traffic using generative adversarial networks. In: **2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)**. IEEE, 2019. p. 0728-0734.

Referências

[12] YIN, Yucheng et al. Practical gan-based synthetic ip header trace generation using netshare. In: **Proceedings of the ACM SIGCOMM 2022 Conference**. 2022. p. 458-472.

[13] DENTON, Emily L. et al. Deep generative image models using a laplacian pyramid of adversarial networks. **Advances in neural information processing systems**, v. 28, 2015.

[14] RADFORD, Alec; METZ, Luke; CHINTALA, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. **arXiv preprint arXiv:1511.06434**, 2015.

[15] MOGREN, Olof. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. **arXiv preprint arXiv:1611.09904**, 2016.

[16] KARRAS, Tero et al. Progressive growing of gans for improved quality, stability, and variation. **arXiv preprint arXiv:1710.10196**, 2017.

Referências

[17] ESTEBAN, Cristóbal; HYLAND, Stephanie L.; RÄTSCH, Gunnar. Real-valued (medical) time series generation with recurrent conditional gans. **arXiv preprint arXiv:1706.02633**, 2017.

[18] ZHANG, Han et al. Self-attention generative adversarial networks. In: **International conference on machine learning**. PMLR, 2019. p. 7354-7363.

[19] RING, Markus et al. Flow-based network traffic generation using generative adversarial networks. **Computers & Security**, v. 82, p. 156-172, 2019.

[20] HAN, Luchao; SHENG, Yiqiang; ZENG, Xuewen. A packet-length-adjustable attention model based on bytes embedding using flow-wgan for smart cybersecurity. **IEEE Access**, v. 7, p. 82913-82926, 2019.

[21] CHENG, Adriel. PAC-GAN: Packet generation of network traffic using generative adversarial networks. In: **2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)**. IEEE, 2019. p. 0728-0734

Referências

[22] BROCK, Andrew; DONAHUE, Jeff; SIMONYAN, Karen. Large scale GAN training for high fidelity natural image synthesis. **arXiv preprint arXiv:1809.11096**, 2018.

[23] XU, Lei et al. Modeling tabular data using conditional gan. **Advances in neural information processing systems**, v. 32, 2019.

[24] WANG, Pan et al. PacketCGAN: Exploratory study of class imbalance for encrypted traffic classification using CGAN. In: **ICC 2020-2020 IEEE International Conference on Communications (ICC)**. IEEE, 2020. p. 1-7.

[25] XU, Shengzhe et al. Stan: Synthetic network traffic generation with generative neural models. In: **Deployable Machine Learning for Security Defense: Second International Workshop, MLHat 2021, Virtual Event, August 15, 2021, Proceedings 2**. Springer International Publishing, 2021. p. 3-29.