CSCD94 Introduction and Week 1 Update

Dominik Luszczynski

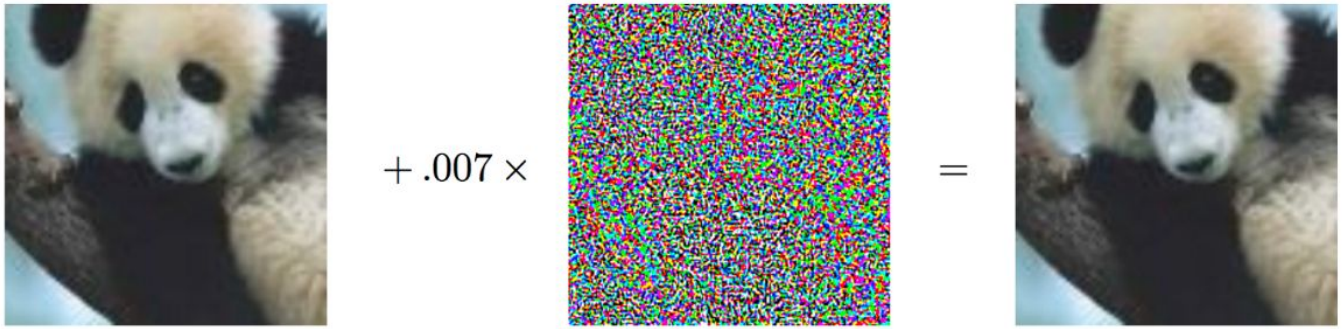
Security in ML

- **Adversarial attacks** - adding small perturbations to the input data to get large differences in model output [1].
- **Data Poisoning** - involves manipulating the training data to make the model behave in an undesirable way (eg. label flipping) [2].
- **Model Inversion** - “reverse engineer the model to extract information from it” [2].

OWASP Top 10: https://owasp.org/www-project-machine-learning-security-top-10/

Adversarial Attacks

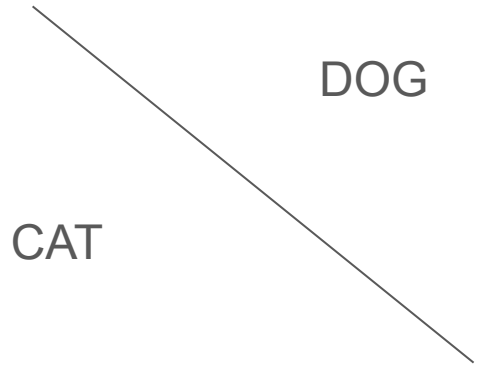
Goal: Manipulate the output of a model by altering/perturbing the input data [1].



https://buaq.net/go-77550.html

Common Adversarial Attack Methods

- White-Box Methods (requires prior knowledge of the model)
  - Fast Gradient Sign Method (requires knowledge of the gradients)



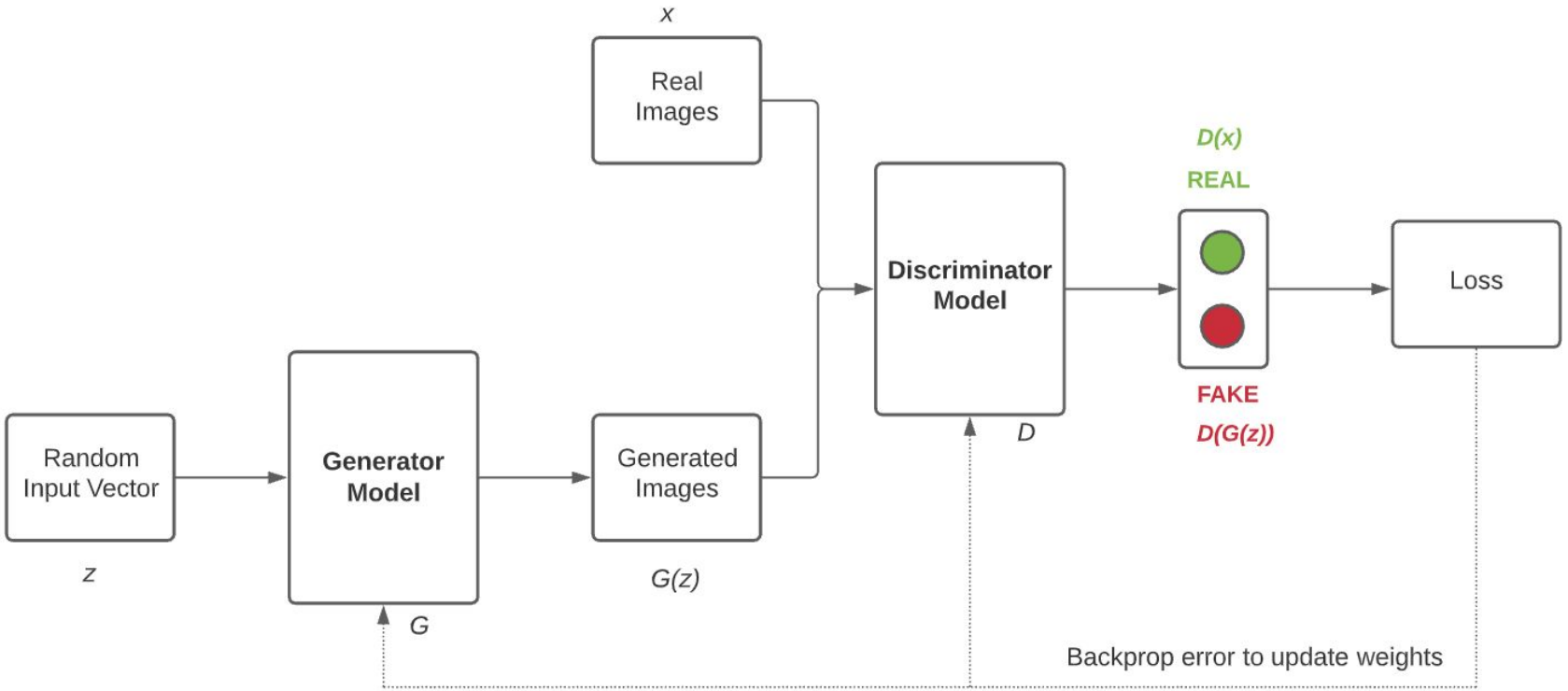
Common Adversarial Attack Methods

- Black-Box Methods (no prior knowledge of model)
  - These methods generate adversarial examples through the outputs of the model [1].
    - One Pixel Attack (good for small images) [3].

Fig. 2. One-pixel attacks on ImageNet dataset where the modified pixels are highlighted with red circles. The original class labels are in black color while the target class labels and their corresponding confidence are given below.

Image taken from [3]

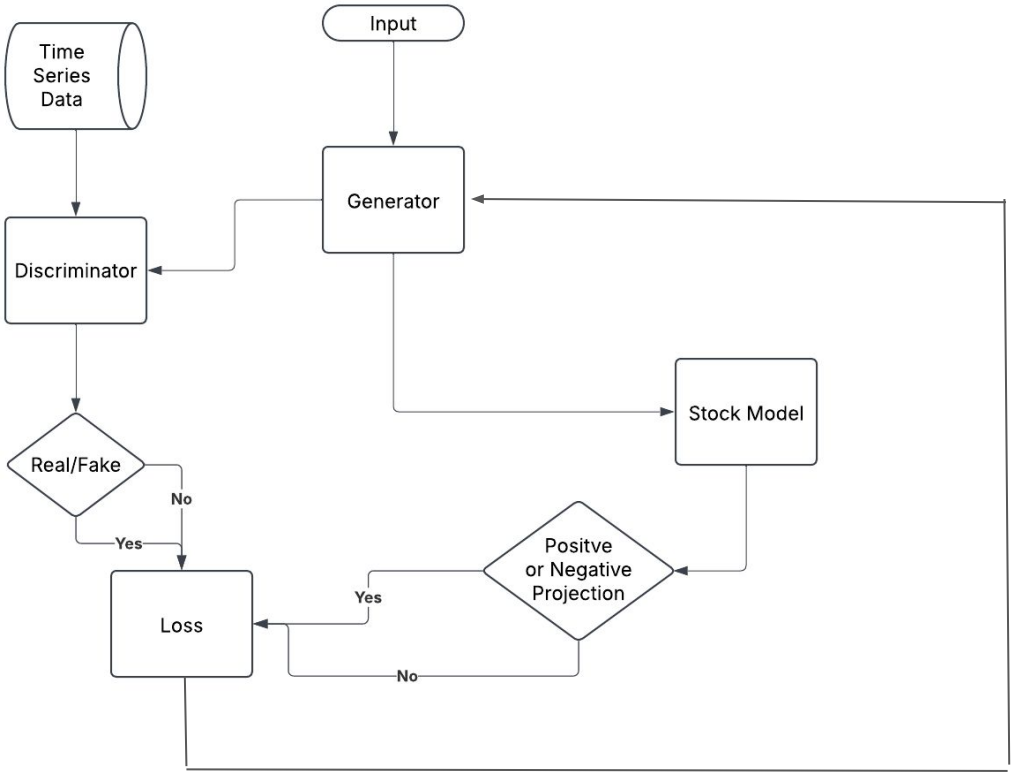
GANs - Generative Adversarial Networks



https://editor.analyticsvidhya.com/uploads/68909architecture.PNG

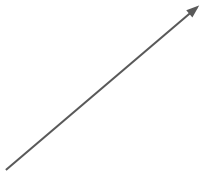
D94 Project: Targeted Adversarial GAN

Model Architecture



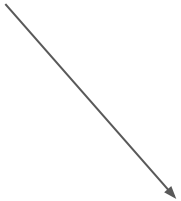
Targeted Manipulation

- With time-series data, there is no clear measure of success regarding whether the projection is incorrect.



Positive Projection

Neutral Projection



Negative Projection

Main Steps

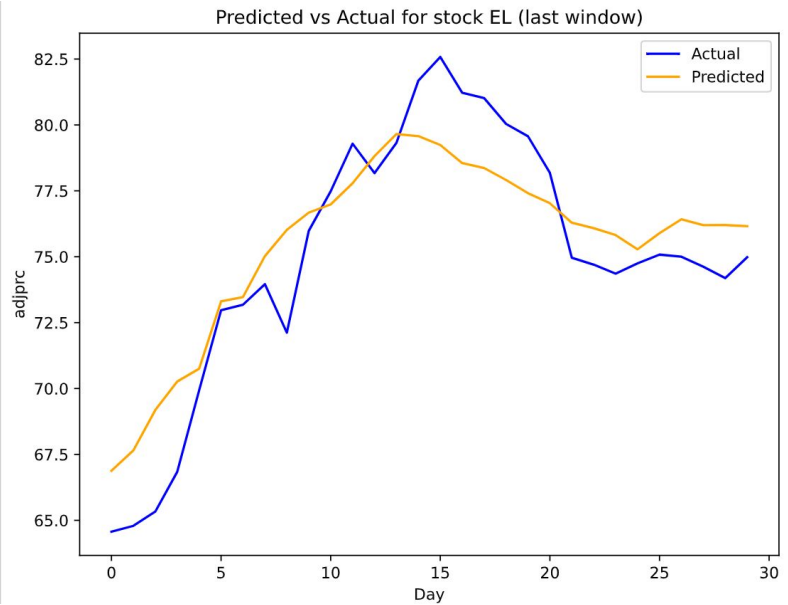
1) Build a stock projection model
2) Implement existing adversarial attacks
3) Create a standard GAN for financial time-series data
4) Augment the GAN with the projection model as the second discriminator (positive projection only)
5) Add additional manipulation modes to the model (neutral and negative projection)

Week 1 Progress - Data Collection and Preprocessing

- Collected daily adjprc (adjusted price) data from the S&P 500 index.
- Created a stratified split based on current price.
- Generated rolling means and standard deviation features based on the adjprc.
- Applied StandardScalar for the derived features fitted to each individual stock.

Week 1 Progress - Model Training

- Used the N-HiTS projection model framework from pytorch-forecasting.



Over the Next Few Weeks…

- Optimize the N-HiTS model (hyperparameters, features, feature scaling)
- Implement various gradient-based adversarial attacks for time series data.
- Start making the GAN.

Citations

[1] J. Chen, H. Zheng, H. Xiong, S. Shen, and M. Su, “MAG-GAN: Massive attack generator via GAN,” *Information sciences*, vol. 536, pp. 67–90, 2020, doi: 10.1016/j.ins.2020.04.019.

[2] Singh, Shain, and Sagar Bhure . “Owasp Machine Learning Security Top Ten.” *OWASP Machine Learning Security Top Ten | OWASP Foundation*, OWASP, 2023, owasp.org/www-project-machine-learning-security-top-10/.

[3] J. Su, D. V. Vargas, and K. Sakurai, “One Pixel Attack for Fooling Deep Neural Networks,” *IEEE transactions on evolutionary computation*, vol. 23, no. 5, pp. 828–841, 2019, doi: 10.1109/TEVC.2019.2890858.