Deep Neural Networks and Tabular Data

Presentation By Jimmy Liang

Introduction

Deep Neural Networks (DNNs) has produced incredible results in the past few years in the fields of computer vision, audio, video, as well as natural language processing.

But, its usage with tabular data, which most business processes rely on, has failed to meet the predictive capability as well as explainability of 'classical' machine learning models such as gradient boosted trees. This presentation goes over the current state of using DNNs with tabular data and future directions

Contents

Challenges

What are some issues with using DNNs with tabular data:

- Data quality
- Spatial dependencies
- Preprocessing
- Model sensitivity

Data Quality

Real world tabular data often contains missing values, outliers, inconsistent, or erroneous data.
Tabular data are often high-dimensional with relatively small sample sizes.
Tabular data is often expensive to obtain and hard to come by, and the dataset are often class imbalanced.

Spatial Correlation

Unlike images, audio, and videos where neighboring pixels and bits provide spatial context, there are no such relationships available in tabular data.
Research has hypothesized that even if there are spatial correlations between variables in tabular data, it is rather complex and irregular and difficult to determine.

Preprocessing

Tabular data is hard to preprocess. One of the main challenges is how to convert categorical features into numerical representations without creating very sparse matrixes. Another issue to watch out for is inadvertently encoding an alignment or ordering based on the numbering system used.

Some implementation attempts to resolve this issue by encoding the categorical features in an embedding space.

Model Sensitivity

Unlike classical machine learning models such as decision trees, DNNs are very sensitive to small changes in the input data.
Tabular data are often highly variable from one sample to the next.

Current State

Now that we discussed the challenges, what are some approaches to overcome them?

We go over:

- Data Transformation
- Hybrid Models
- Transformers
- Regularization

Methods

Data Transformation

Transforming the tabular data with various techniques to overcome issues with categorical variables.

Hybrid Models

Combine deep neural networks with classical machine learning techniques such as decision trees.

Transformers

Building on the success of transformers in natural language processing, using the attention mechanism on tabular data

Regularization

Utilizing the theory that strong regularization will help overcome the model sensitivity due to the extreme flexibility of deep learning models.

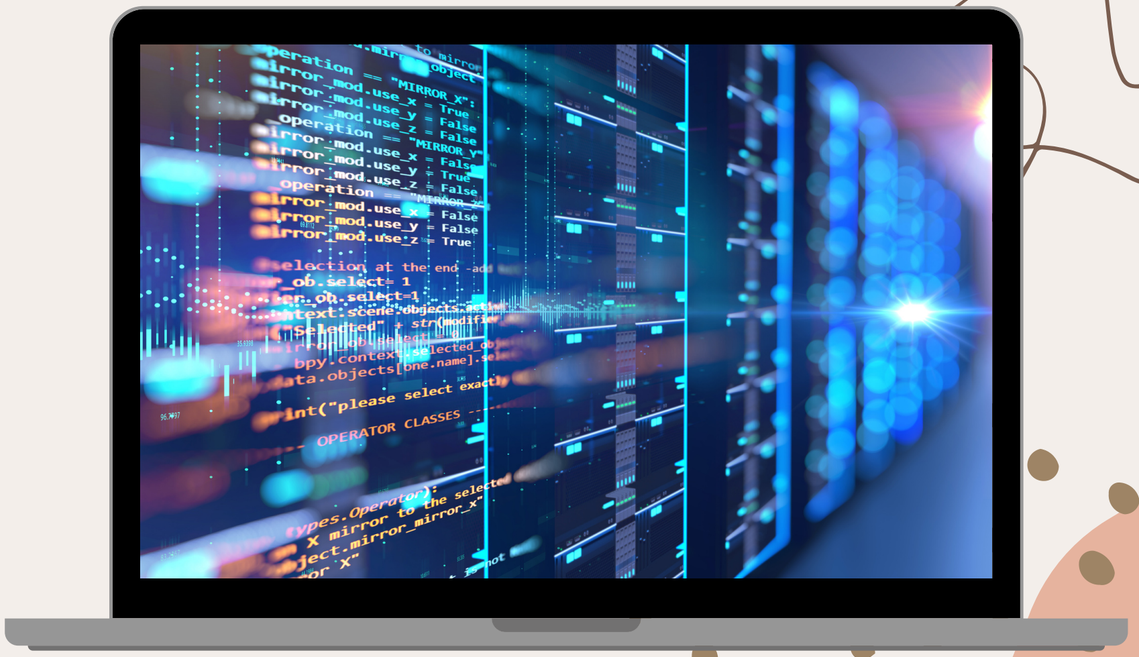
Data Transformation

Single Dimensional

Deterministic, can be used before training. Can be as simple as ordinal encoding, binary encoding, leave-one-out encoding, hash-based encoding.

Multi-dimensional

Using self or semi-supervised techniques to encode the categorical values into a dense embedding space.



Hybrid Models

Fully Differentiable

Permits end to end optimization using gradient descent. Highly efficient on GPU.

Partly Differentiable

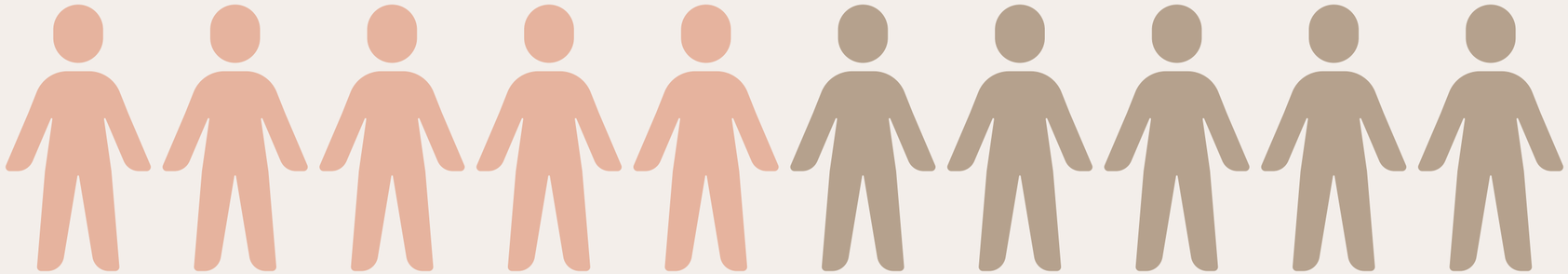
Combining non-differentiable models such as decision trees with deep neural networks. Utilizing different models to handle numerical and categorical features.

Transformers

Lots of research

Lots of recent and active research in this area. TabNet, TabTransformer, ARM-net, SAINT, etc. Utilizes multiple subnetworks and self-attention to handle categorical features and incorporate varying techniques such as decision trees, k-nearest-neighbor, and feature crosses.

Regularization



Regularization Learning Network

Applying a trainable regularization coefficient to lower the overall model sensitivity.

Regularization Cocktail

Applying multiple regularization techniques together. A paper in 2021 used 13 regularization techniques together that outperformed tree-based models.

Data Generation

Why

How

Quality

Why

- Tabular data is difficult and expensive to come by. Training data is usually limited.
- Data augmentation and imputation (filling in missing values)
- Rebalancing imbalanced classes
- Ensure privacy

How

- Generative Adversarial Networks (GANs)
- MedGAN for domain specific generations
- Variational Autoencoders
- Various VAEs, can outperform GANs, but both are considered state of the art.

Quality

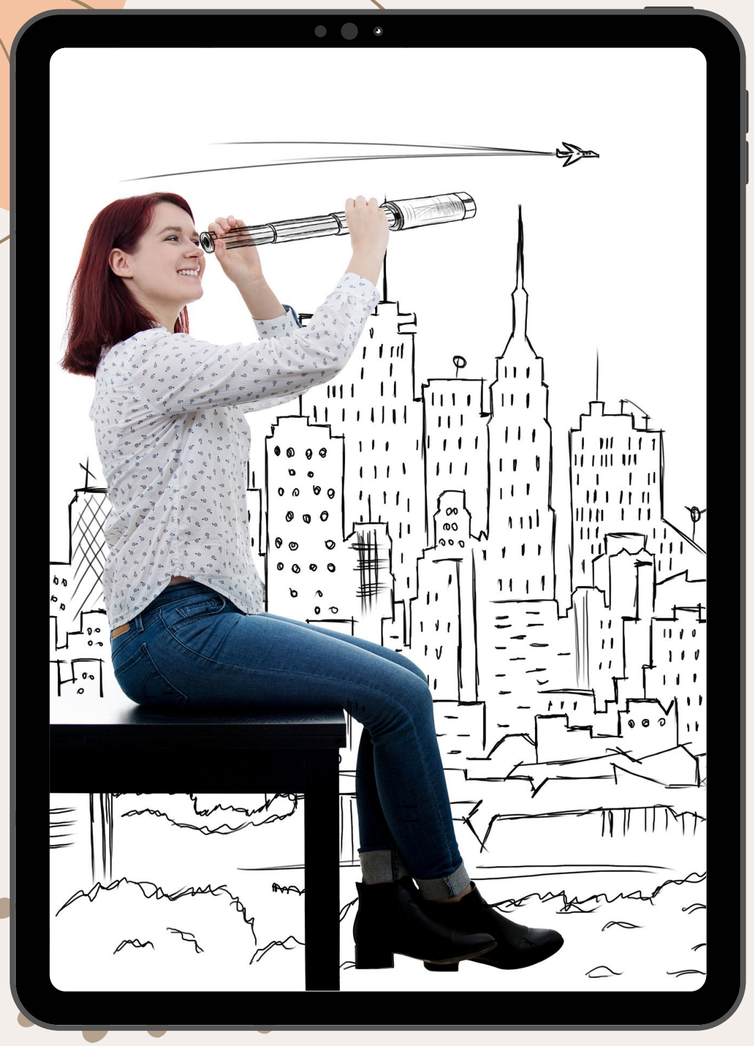
- How to assess quality?
- Typically using a proxy classification task that is trained using generated tabular data.
- The prediction is done using real data to assess the quality of the generated data.
- Another approach is using statistical methods to generate data based on original data's distribution.

Explainability

The ability to understand what the prediction is based on is hugely important in the real world.

Feature Highlighting

Constructing the model that are explainable by design. In cases where the model parameter is not available, use surrogate models or a benchmarking library.



Future Directions

Deep Learning with Tabular Data is an actively researched topic. Tabular data is the most used type of data in businesses and as such, can have the potential to produce the most impact.

Some of the trends and future directions include ->

Trends

Data Preprocessing	Continue to transform into homogeneous representations such as an embedding
Architecture	Transformers have taken the lead, offering multiple advantages such as attention over both categorical and numerical features.
Regularization	It's been shown that combining regularization techniques can help even a vanilla feed forward network.
Data Generation	Generation task is difficult for tabular data as the possible space is infinite. More research needs to happen in this area.
Explainability	Explainable AI is the foundation to ensure equity. DNNs need to do more in this area to match the classical techniques such as decision trees.

Thank You !