# Asset Embeddings

## Introduction

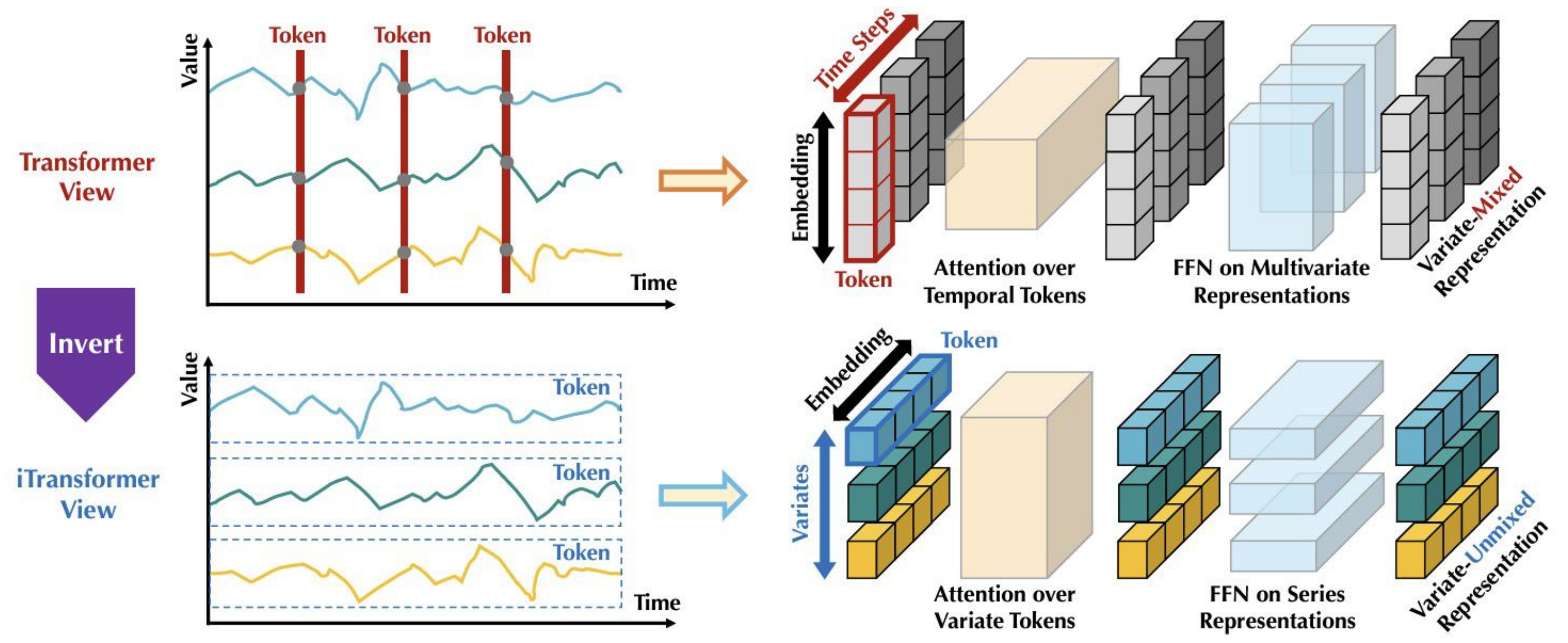
The objective of asset embeddings is to create **unified representations** that encapsulate all relevant information about an asset, including **return time series**, **asset characteristics**, and **text-based information**. Drawing on methodologies from computer vision (CV) and natural language processing (NLP), this approach can be implemented through two primary methods: self-supervised learning and multi-objective learning.

In **self-supervised learning**, the model learns to recover masked portions of data, enhancing its ability to capture underlying structures without relying on labeled data. This method allows for the addition of a task-specific prediction head when applied to downstream tasks. Alternatively, **multi-objective learning** involves training the model on multiple tasks simultaneously, such as portfolio construction, performance metric prediction, and security selection, to produce embeddings that are robust across different financial applications.

## Data

At this stage, we use daily time series data as the foundation for constructing asset embeddings. The data includes:

1. **Mutual Fund Return Time Series:** Daily returns from a diverse set of mutual funds, normalized to facilitate consistent embedding generation. This dataset covers various investment strategies and market sectors.
2. **S&P 500 Index:** The daily closing values of the S&P 500 index are incorporated to enable multivariate time series prediction, allowing the model to account for broader market conditions.
3. **Macroeconomic Variables:** Macroeconomic variables are included after dimensionality reduction through Principal Component Analysis (PCA).


Transformer View
Value
Token
Token
Token
Time
Invert
iTransformer View
Value
Token
Token
Token
Time
Time Steps
Embedding
Token
Attention over Temporal Tokens
FFN on Multivariate Representations
Variate-Mixed Representation
Embedding
Token
Variates
Attention over Variate Tokens
FFN on Series Representations
Variate-Unmixed Representation

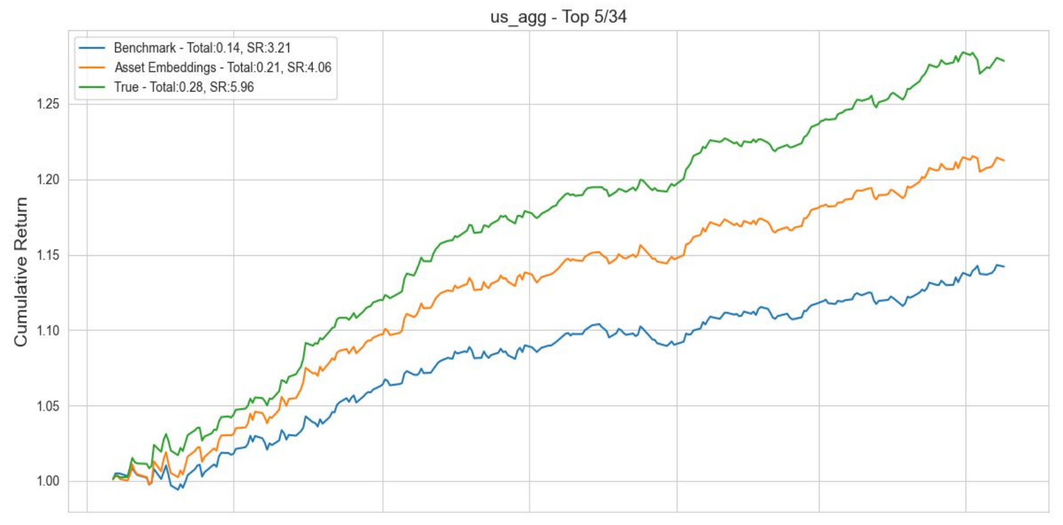
## Model Architectures

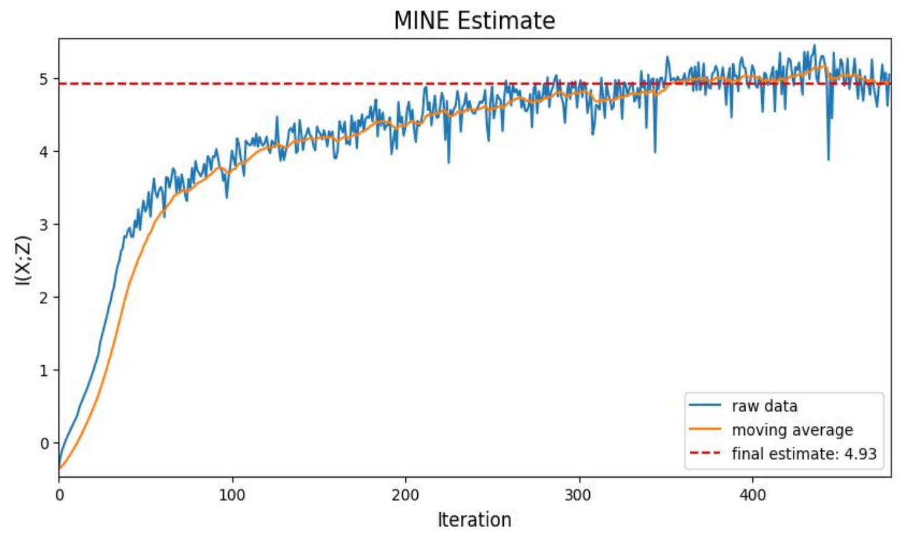
We utilized advanced models from the field of **long sequence time series forecasting** to develop asset embeddings. The models include:

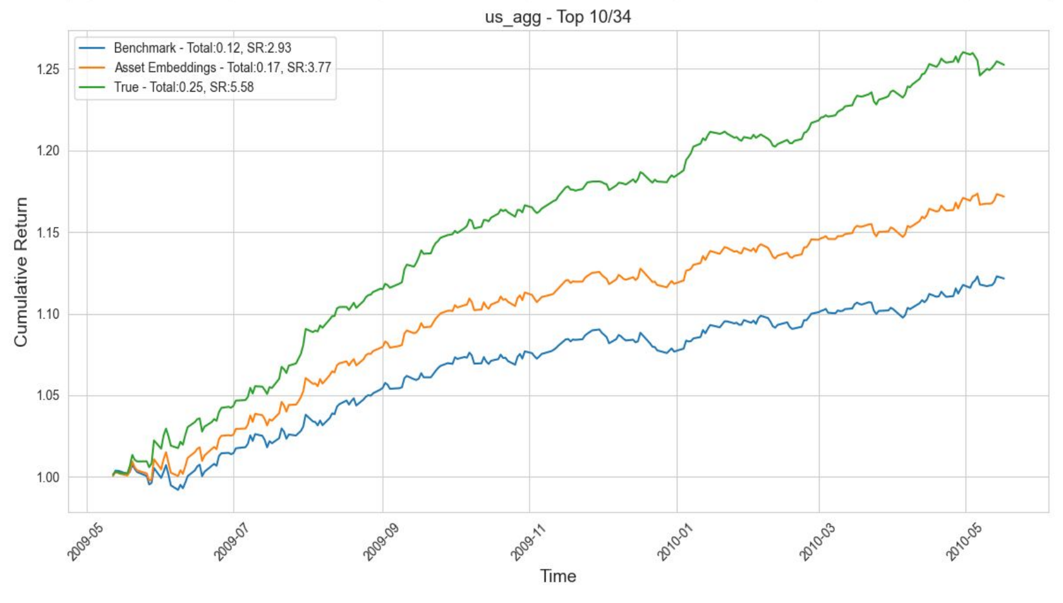
1. **Linear/RLinear/DLinear (2023):** These models map past time steps to future ones through simple MLP. RLinear uses **Reversible Instance Normalization (RevIN)** to address non-stationarity, while DLinear applies **series decomposition** to separate trend and seasonal components.
2. **PatchTST (2023):** This model divides the time series into **smaller patches**, processing each independently to capture local patterns and temporal dynamics.
3. **iTransformer (2023):** iTransformer generates **embeddings for each data channel**, using attention mechanisms to integrate information across different features within the multivariate time series.

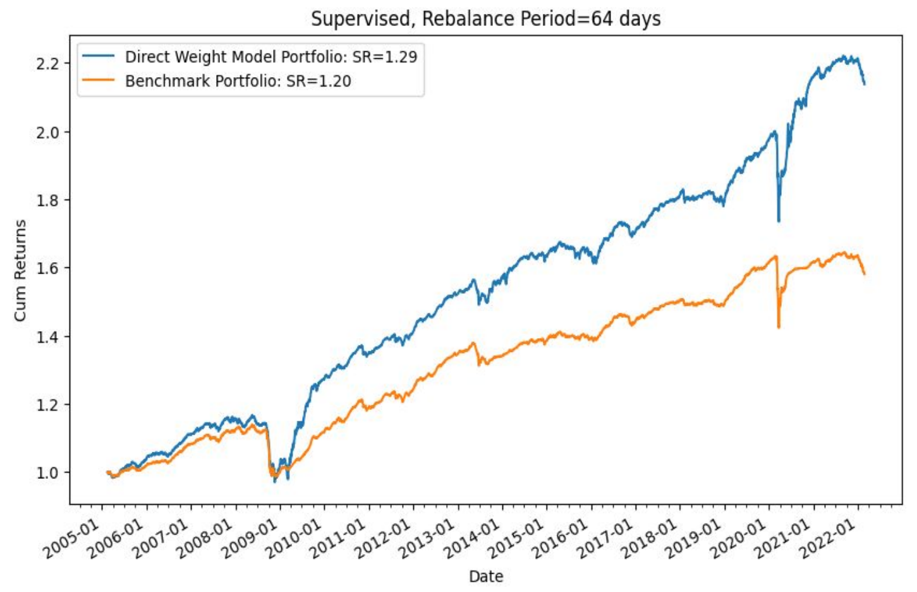
## Results

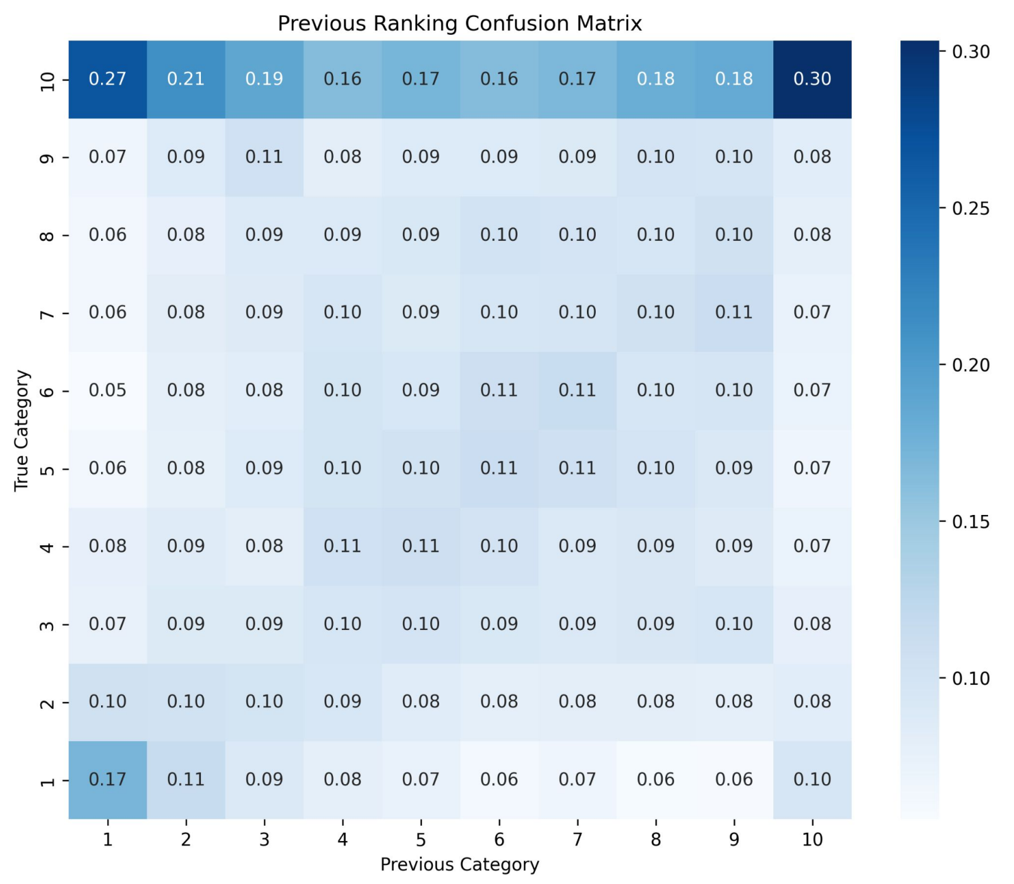
1. **MINE Mutual Information Estimation:** We measure the **information entropy** of the original data and the asset embeddings to assess information retention. Additionally, we calculate the **mutual information** between them to evaluate whether the model captures and adds valuable information. For instance, the input data's entropy is **4.00 nats**, while the embeddings' entropy is **7.59 nats**, with a mutual information of **4.93 nats**. This suggests that the multivariate time series model not only preserves original information but also integrates additional, likely market-related, data into the embeddings.

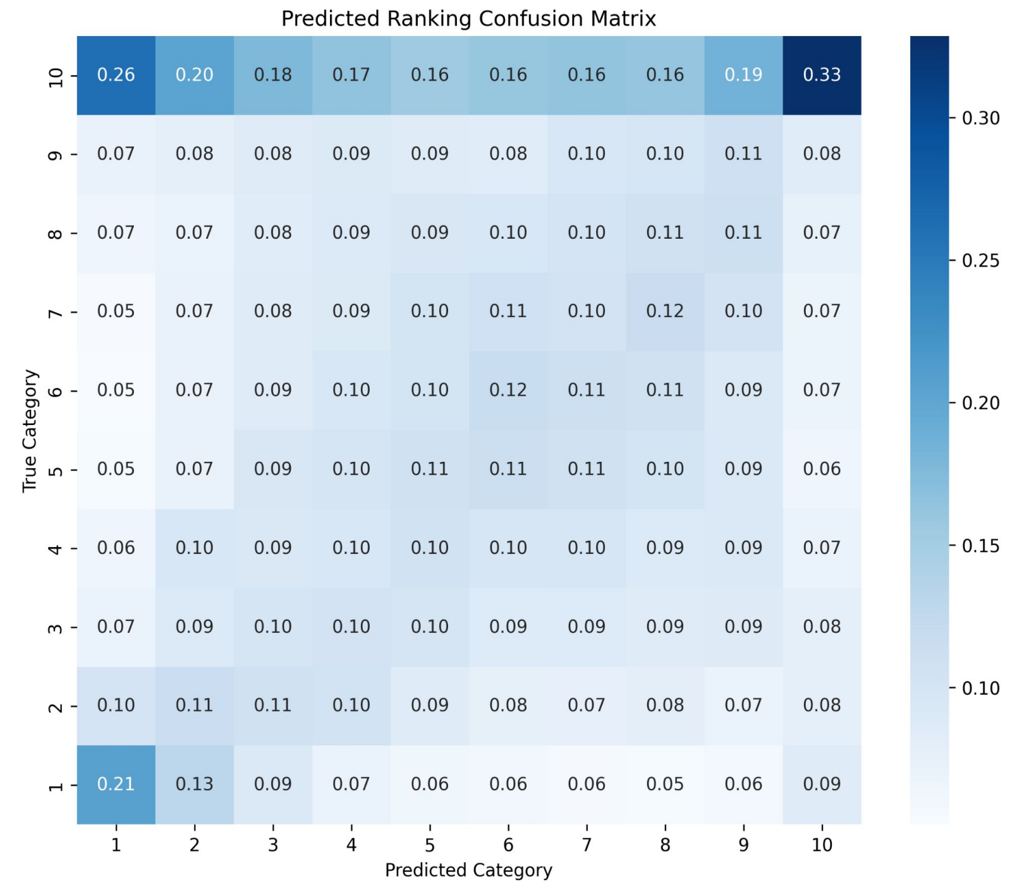












2. **Portfolio Construction:**
   **Return Prediction:** We predict future returns and construct a Markowitz global minimum variance portfolio. This portfolio is compared with one constructed using the returns and covariance matrix from the input data.
   **Direct Portfolio Weight Prediction:** Instead of predicting returns, our model directly predicts portfolio weights, using future weights as labels. In rolling retrain with a 64-day rebalance, our model achieves a **Sharpe Ratio of 1.29**, compared to the **benchmark's 1.20**.
3. **Security Selection and Ranking:** The model ranks assets using listwise methods like ListNet or NeuralNDCG. We then create equal-weighted portfolios from the **top 5** or **10** ranked assets and compare them to portfolios based on known previous rankings. Our model outperformed the benchmark in **21 out of 24** mutual fund portfolios.

## Conclusion & Future Direction

Among the evaluation methods employed, our models consistently demonstrated **superior performance** compared to the benchmark models. At the current stage, we focused on testing the self-supervised learning approach for generating embeddings. However, this method did not exhibit significant advantages over traditional supervised methods in all four evaluation methods.

As the next step, we plan to explore **multi-objective learning** for embedding generation. This approach may offer improved performance by simultaneously optimizing for multiple tasks. Additionally, our current models are based solely on time series data. Future work will involve incorporating tabular data, such as **mutual fund characteristics**, using **multi-modal deep learning** techniques. This will allow us to assess the impact of integrating diverse data types on the effectiveness of asset embeddings.

Supervisors: Professor Ali Hirsa, Sikun Xu
Group Member: Chenkai Li