

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



BÁO CÁO HỌC PHẦN DỰ ÁN CÔNG NGHỆ

Đề tài

ỨNG DỤNG MÔ HÌNH TRANSFORMER TRONG DỰ ĐOÁN GIÁ VÀNG

Sinh viên: Chu Thân Nhất - 22022578

Giảng viên hướng dẫn: PGS.TS Nguyễn Phương Thái

Ý kiến đánh giá:

.....

.....

.....

.....

Điểm số: Điểm chữ:

Hà Nội, ngày tháng năm 2025

Giảng viên đánh giá

(Ký, ghi rõ họ tên)

LỜI CẢM ƠN

Lời đầu tiên, tôi xin cảm ơn chân thành đến PGS.TS Nguyễn Phương Thái, giảng viên hướng dẫn, đã tận tình chỉ bảo, cung cấp những ý kiến quý giá và định hướng rõ ràng trong quá trình thực hiện dự án. Tôi cũng xin cảm ơn các bạn trong nhóm và các thầy cô trong nhóm AI tài chính đã cung cấp các dữ liệu tài chính quý giá giúp tôi xây dựng mô hình Tranformer, cũng như Trường Đại học Công Nghệ, Đại học Quốc gia Hà Nội đã tạo điều kiện thuận lợi để tôi hoàn thành học phần Dự án.

TỔNG QUAN

Giá vàng (XAUUSD là một trong những chỉ số tài chính quan trọng, ảnh hưởng đến các quyết định đầu tư và chính sách kinh tế toàn cầu. Tuy nhiên, giá vàng biến động mạnh do các yếu tố cung cầu, lãi suất, lạm phát và tâm lý thị trường, khiến việc dự đoán xu hướng giá trở thành một bài toán phức tạp và không dễ dàng. Dự án này đề xuất một giải pháp sử dụng mô hình Tranformer kết hợp Time2vec để dự đoán xu hướng giá vàng, phân loại thành ba nhãn: BUY (mua), SELL (bán), HOLD (giữ).

Mô hình được xây dựng trên bộ dữ liệu giá vàng XAUUSD, bao gồm giá mở, cao nhất, thấp nhất, đóng cửa và khối lượng giao dịch. Dữ liệu được bổ sung các chỉ báo kỹ thuật RSI, MACD, Bollinger Bands, và các đặc trưng thời gian như giờ, ngày, trong tuần. Mô hình đạt độ chính xác 66% trên tập kiểm tra (test) với 977044 mẫu. Báo cáo này trình bày chi tiết quá trình xử lý dữ liệu, xây dựng mô hình, đánh giá hiệu suất và đề xuất các cải tiến như sử dụng. Dự án này không chỉ mang lại giá trị thực tiễn trong lĩnh vực tài chính mà còn giúp tôi rèn luyện kỹ năng lập trình, phân tích dữ liệu và nghiên cứu khoa học.

MỤC LỤC

LỜI CẢM ƠN	1
TỔNG QUAN	2
MỤC LỤC	3
DANH SÁCH HÌNH ẢNH	4
I. GIỚI THIỆU	5
1.1. Giá vàng và vai trò trong tài chính	5
1.2. Dự đoán giá vàng	5
1.3. Mục tiêu bài toán	6
II. YÊU CẦU BÀI TOÁN	6
2.1. Tập dữ liệu	6
2.2. Xử lý dữ liệu	7
2.3. Mô tả bài toán	8
III. TÓM TẮT LÝ THUYẾT, GIẢI PHÁP	9
3.1. Các lý thuyết	9
3.2. Cách giải quyết	10
3.3. Liên hệ và so sánh với mô hình khác	13
IV. KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN TIẾP THEO	14
4.1 Kết quả	14
4.2 Kiến thức thu được và hướng phát triển	15
V. KẾT LUẬN	17
THAM KHẢO	18

DANH SÁCH HÌNH ẢNH

Hình 1: Giá vàng trung bình hằng năm từ 2000-2024	5
Hình 2: Phân chia dữ liệu XAUUSD.csv.....	7
Hình 3: Biểu diễn HOLD, BUY, SELL trong 12 tiếng.....	8
Hình 4: Các đặc trưng.....	8
Hình 5: Sơ đồ kiến trúc Transformer.....	9
Hình 6: Confusion Matrix của tập Test.....	14
Hình 7: Biểu đồ Train Loss và Validation Loss khi train mô hình.....	15

I. GIỚI THIỆU

1.1. Giá vàng và vai trò trong tài chính

Giá vàng (XAUUSD), biểu thị giá trị của một ounce vàng so với đồng USD, là một chỉ số tài chính quan trọng trên thị trường ngoại hối. Vàng được xem là tài sản trú ẩn an toàn, đặc biệt trong các giai đoạn kinh tế vật ộn, lạm phát cao, hoặc biến động chính trị. Giá vàng chịu ảnh hưởng từ nhiều yếu tố như chính sách tiền tệ của Mỹ (FED), lãi suất, cung cầu trên thị trường, và tâm lý nhà đầu tư. Tại Việt Nam, tính từ năm 2000 đến năm 2024 giá vàng đã tăng gấp 8 lần, vàng không chỉ là công cụ đầu tư mà còn gắn liền với văn hóa, như mua vàng tích trữ hoặc sử dụng trong các dịp lễ tết. Do đó, việc dự đoán chính xác xu hướng giá vàng có ý nghĩa lớn đối với nhà đầu tư cá nhân, tổ chức tài chính, và các nhà hoạch định chính sách.



Hình 1: Giá vàng trung bình hàng năm từ 2000–2024

1.2. Dự đoán giá vàng

Dự đoán giá vàng là một bài toán phức tạp do tính chất phi tuyến và biến động mạnh của dữ liệu. Các phương pháp truyền thống như hồi quy tuyến tính, ARIMA, hoặc phân tích kỹ thuật đơn giản thường không đạt hiệu quả cao vì không thể nắm bắt được các mẫu (trend) phức tạp trong dữ liệu chuỗi thời gian. Gần đây, các mô hình học sâu như LSTM (Long Short-Term Memory), GRU và Transformer đã cho thấy tiềm năng vượt trội nhờ khả năng học các phụ thuộc dài hạn và xử lý dữ liệu

đã biến. Trong đó Tranformer, với cơ chế attention, đặc biệt phù hợp để dự đoán xu hướng giá vàng dựa trên dữ liệu lịch sử và các chỉ báo kỹ thuật.

1.3. Mục tiêu bài toán

Dự án này nhằm xây dựng một mô hình học sâu sử dụng kiến trúc Tranformer kết hợp Time2vec để dự đoán xu hướng giá vàng, phân loại thành ba nhãn: BUY, SELL, HOLD. Mô hình được huấn luyện trên bộ dữ liệu giá vàng XAUUSD, bao gồm giá mở, cao nhất, thấp nhất, giá đóng, khối lượng giao dịch (OHLCV) và các chỉ báo kỹ thuật như RSI, MACD, Bollinger Bands. Mục tiêu cụ thể bao gồm:

- Đạt độ chính xác cao trong việc phân loại xu hướng giá vàng
- Đánh giá hiệu quả mô hình Tranformer
- Đề xuất các cải tiến để tăng tính thực tiễn tài chính

II. YÊU CẦU BÀI TOÁN

2.1. Tập dữ liệu

Bộ dữ liệu được sử dụng trong dự án là XAUUSD.csv, từ 14/6/2004 đến 11/3/2025 chứa dữ liệu giá vàng theo khung thời gian 1 phút, bao gồm các cột:

- Date, Time: Thời gian giao dịch
- Open, High, Low, Close: Giá mở, cao nhất, thấp nhất, đóng cửa (USD/ounce)
- Volume: Khối lượng giao dịch

Tổng cộng có hơn 6,5 triệu bản ghi giao dịch, được chia thành 3 tập:

- Train (70%): Dữ liệu huấn luyện mô hình (4560131 mẫu)
- Validation (15%): Dữ liệu xác thực để điều chỉnh siêu tham số (977170 mẫu)
- Test(15%): Dữ liệu kiểm tra (977172 mẫu)

Tập dữ liệu	Bắt đầu	Kết thúc	Khoảng thời gian
Train	2004-06-14 04:17:00	2019-08-22 16:59:00	~15 năm
Validation	2019-08-22 17:00:00	2022-06-07 00:34:00	~2 năm 10 tháng
Test	2022-06-07 00:35:00	2025-03-11 07:15:00	~2 năm 9 tháng

Hình 2: Phân chia dữ liệu XAUUSD.csv

2.2 Xử lý dữ liệu

Dữ liệu được sắp xếp theo thời gian để đảm bảo tính liên tục của chuỗi thời gian, các bước xử lý bao gồm:

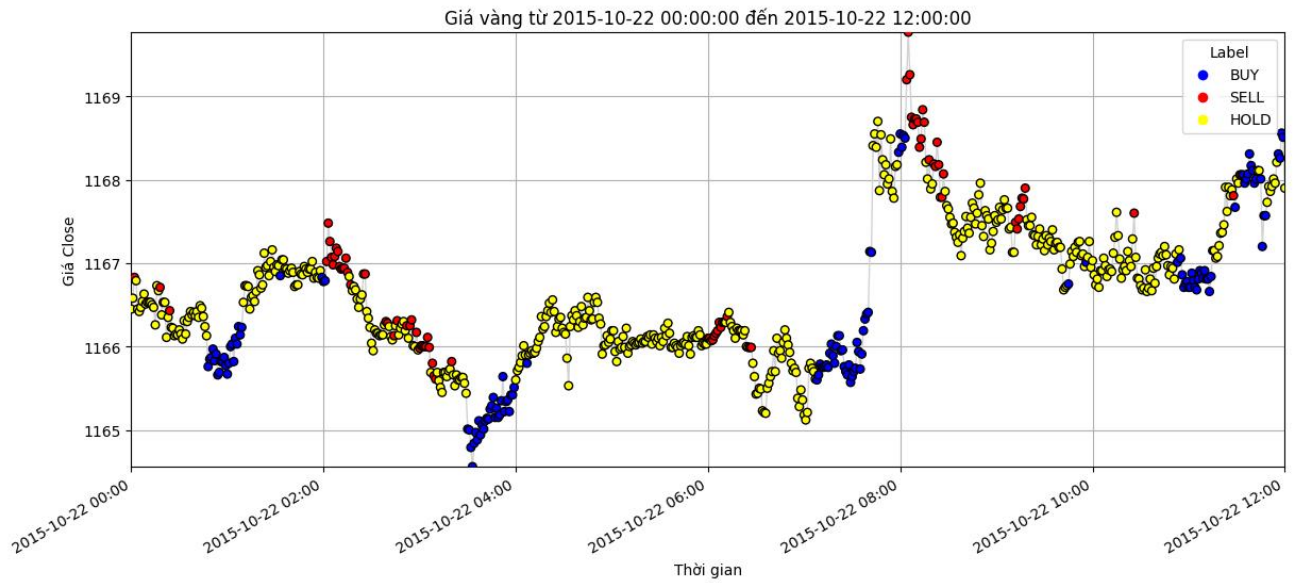
- Tạo cột timestamp bằng cách kết hợp Data và Time, đồng thời kiểm tra timestamp có hợp lệ hay không.
- Thêm các chỉ báo kỹ thuật:
 - 4 chỉ báo động lượng: RSI, Momentum, CMO, Williams_%R
 - 9 chỉ báo biến động: ATR, BB_Mid, BB_Upper, BB_Lower, BB_Bandwidth, KC_High, KC_Low, DC_High, DC_Low
 - 6 chỉ báo xu hướng: SMA_20, EMA_20, DPO, MACD, MACD_Hist, Mass_Index
 - 5 chỉ báo khối lượng: AD, CMF, Force_Index, MFI, OBV.
- Nhãn HOLD, BUY, SELL được gán dựa trên độ biến động giá sau 30 phút.

Gán nhãn đơn giản

+ $BUY \leftrightarrow Pricet + \Delta > Pricet * (1 + threshold)$.

+ $SELL \leftrightarrow Pricet + \Delta < Pricet * (1 - threshold)$.

+ $HOLD \leftrightarrow otherwise$.



Hình 3: Biểu diễn HOLD, BUY, SELL trong 12 tiếng

- Từ timestamp ta cần thêm các đặc trưng theo giờ, ngày trong tuần, phút trong ngày và chỉ số index. Tất cả được chuẩn hóa về [0,1]
- Xử lý NaN/Inf: Thay thế NaN bằng phương pháp bfill và ffill, loại bỏ các hàng còn NaN nếu có.

Kết hợp với Open, High, Low, Close, Volume, tổng cộng có 33 đặc trưng.

```
train_processed.head()
```

	Open	High	Low	Close	Volume	Label	RSI	Momentum	CMO	Williams_%R	...	Mass_Index	AD	CMF	Force_Index	MFI	OBV	hour	day_of_week	minute_of_day	index
0	0.000000	0.000000	0.000000	0.000000	2	0	51.296441	0.0	382.5	-50.0	...	20.836749	2.0	0.0	-0.015863	42.871896	2	0.173913	0.0	0.178596	0.000000e+00
1	0.000000	-0.052247	0.000000	-0.052247	1	2	51.296441	0.0	382.5	-50.0	...	20.836749	2.0	0.0	-0.015863	42.871896	1	0.173913	0.0	0.180681	2.192920e-07
2	0.052274	0.052274	0.000000	0.000000	2	2	51.296441	0.0	382.5	-50.0	...	20.836749	0.0	0.0	-0.015863	42.871896	3	0.173913	0.0	0.182766	4.385840e-07
3	0.078370	0.078370	0.130685	0.130685	1	1	51.296441	0.0	382.5	-50.0	...	20.836749	0.0	0.0	-0.015863	42.871896	4	0.173913	0.0	0.184851	6.578760e-07
4	-0.130514	-0.130514	-0.130514	-0.130514	1	2	51.296441	0.0	382.5	-50.0	...	20.836749	0.0	0.0	-0.015863	42.871896	3	0.173913	0.0	0.185546	8.771680e-07

5 rows × 34 columns

Hình 4: Các đặc trưng

2.3 Mô tả bài toán

(*) - Đầu vào:

- Dữ liệu giá vàng lịch sử (OHLC, Volume) trong khoảng thời gian 128 phút trước đó (sequence length = 128).
- Các chỉ báo kỹ thuật: RSI, Momentum, MACD, Bollinger Bands, ATR, v.v. (33 đặc trưng).
- Đặc trưng thời gian: giờ, ngày trong tuần, phút trong ngày (được mã hóa bằng Time2Vec).

(*) - Đầu ra:

- Nhãn phân loại: BUY (0), SELL (1), HOLD (2), dựa trên biến động giá Close sau 30 phút với ngưỡng $\pm 0.05\%$.

Mục tiêu:

Xây dựng mô hình Transformer dự đoán chính xác nhãn BUY, SELL, HOLD.

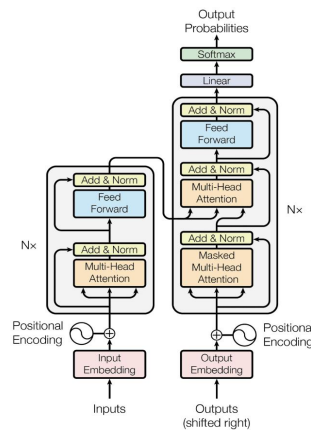
Đánh giá mô hình trên tập test và kiểm tra hiệu quả giao dịch thực tế thông qua backtesting.

III. TÓM TẮT LÝ THUYẾT, GIẢI PHÁP

3.1. Các lý thuyết

3.1.1. Transformer

Transformer, được giới thiệu trong bài báo “Attention is All You Need” là một kiến trúc học sâu dựa trên cơ chế attention để xử lý dữ liệu chuỗi. Không giống LSTM hay GRU, Transformer không sử dụng cấu trúc tuần tự, mà xử lý toàn bộ chuỗi đầu vào song song, giúp tăng tốc độ huấn luyện và học các phụ thuộc dài hạn hiệu quả. Trong dự đoán chuỗi thời gian, Transformer được áp dụng để nắm bắt các mẫu phức tạp trong dữ liệu giá vàng, đặc biệt khi kết hợp với các đặc trưng như chỉ báo kỹ thuật.



Hình 5: Sơ đồ kiến trúc Transformer

3.1.2. Time2Vec

Time2Vec là một phương pháp mã hóa đặc trưng thời gian thành vector giúp mô hình học các mẫu mang tính chu kỳ (ví dụ: biến động giá theo giờ hoặc ngày). Time2Vec chuyển các giá trị thời gian (như phút trong ngày) thành một vector biểu diễn, kết hợp các hàm tuyến tính và phi tuyến tính (sin, cos) để mô tả tính tuần hoàn.

Trong dự án, Time2Vec được sử dụng để mã hóa đặc trưng `minute_of_day` (phút trong ngày), làm tăng khả năng dự đoán của mô hình Transformer. Nếu không có nó, cơ chế attention không hiểu được tính thứ tự và chỉ ra dấu hiệu về các sự kiện. Tóm lại

Time2Vec mã hóa một giá trị thời gian t , thành một vector bằng cách kết hợp

- Thành phần tuyến tính: biểu diễn xu hướng tuyến tính của thời gian
- Thành phần tuần hoàn: biểu diễn mẫu lặp lại có chu kỳ bằng hàm sin/cos

Công thức:

$$\text{Time2Vec}(t)[i] = \begin{cases} \omega_i t + \phi_i & \text{if } i = 0 \\ \sin(\omega_i t + \phi_i) & \text{if } 1 \leq i < k \end{cases}$$

3.2. Cách giải quyết

Dữ liệu ban đầu sẽ được tiền xử lý và thêm các chỉ báo kỹ thuật và gán nhãn [Mục 2.2].

3.2.1. Mã hóa đặc trưng thời gian với Time2Vec

Luồng dữ liệu qua Time2Vec

Sử dụng Time2Vec để mã hóa đặc trưng thời gian, cụ thể là cột `minute_of_day` (phút trong ngày, từ 0 đến 1439), thành một vector biểu diễn có kích thước 32 chiều.

Đầu vào: Giá trị `minute_of_day` được chuẩn hóa về khoảng $[0,1]$ bằng cách chia cho 1440 (tổng số phút trong ngày)

Đầu ra: Một vector 32 chiều cho mỗi mẫu dữ liệu, biểu diễn tính tuần hoàn của thời gian. Vector này được nối với các đặc trưng khác (giá OHLC, chỉ báo) để tạo đầu vào cho Transformer.

3.2.2. Xử lý chuỗi dữ liệu với Transformer Encoder

Transformer Encoder là thành phần chính của mô hình, chịu trách nhiệm học các mẫu phức tạp trong chuỗi dữ liệu giá vàng và các đặc trưng bổ sung. Dữ liệu vào là các chuỗi có độ dài 128 (phút trước đó), mỗi mẫu bao gồm 33 đặc trưng và vector Time2Vec.

Luồng dữ liệu qua Transformer Encoder

Đầu vào:

Một batch dữ liệu có kích thước (batch_size, sequence_length=128, feature_dim=33)

Dữ liệu được chuẩn hóa theo khối (block normalization) để đảm bảo giá trị nằm trong khoảng trung bình 0, độ lệch chuẩn 1.

Xử lý:

1) Lớp Linear: Một lớp fully-connected ánh xạ 69 đặc trưng thành không gian $d_{\text{model}}=32$, tạo tensor (batch_size, 128, 32).

2) Positional Encoding: Thêm thông tin vị trí cho mỗi bước thời gian trong chuỗi, sử dụng hàm sin và cos để mã hóa thứ tự thời gian:

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_{\text{model}}}), \quad PE(pos, 2i + 1) = \cos(pos/10000^{2i/d_{\text{model}}})$$

Điều này giúp Transformer phân biệt các bước thời gian trong chuỗi.

3) Transformer Encoder:

- Số tầng: 1.
- Số đầu attention: 4.
- Kích thước feed-forward: 128.
- Dropout: 0.3.
- Mỗi tầng bao gồm:
 - Multi-Head Self-Attention: Học các mối quan hệ giữa các bước thời gian, xác định các mẫu quan trọng (ví dụ: giá tăng sau khi RSI vượt ngưỡng 70).
 - Feed-Forward Network: Áp dụng biến đổi phi tuyến cho từng bước thời gian.
 - Layer Normalization: Chuẩn hóa đầu ra để ổn định huấn luyện.

4) Global Pooling: Lấy giá trị trung bình của đầu ra Transformer (batch_size, 128, 32) theo chiều chuỗi, tạo tensor (batch_size, 32).

Đầu ra: Một vector 32 chiều cho mỗi batch, chứa các đặc trưng tiềm ẩn đã học từ chuỗi dữ liệu.

Transformer Encoder nắm bắt các phụ thuộc dài hạn và các mẫu phức tạp trong dữ liệu giá vàng, chẳng hạn như mối quan hệ giữa MACD và giá đóng cửa, hoặc tác động của biến động giá trong giờ giao dịch cao điểm.

3.2.3. Dự đoán xu hướng giá vàng

Sau khi Transformer Encoder tạo ra vector đặc trưng tiềm ẩn, một lớp dự đoán được sử dụng để phân loại xu hướng giá vàng thành BUY, SELL, hoặc HOLD.

Luồng dữ liệu qua lớp dự đoán:

- **Đầu vào:** Vector 32 chiều từ Transformer Encoder (batch_size, 32).
- **Xử lý:**
 - Một mạng MLP (Multi-Layer Perceptron) với cấu trúc:
 - Lớp fully-connected: 32 -> 64, kích hoạt ReLU.
 - Dropout: 0.3.
 - Lớp fully-connected: 64 -> 3, xuất ra logits cho 3 nhãn.
 - Hàm kích hoạt Softmax chuyển logits thành xác suất cho BUY, SELL, HOLD.
- **Đầu ra:** Xác suất cho mỗi nhãn (batch_size, 3). Nhãn có xác suất cao nhất được chọn làm dự đoán cuối cùng

Lớp MLP ánh xạ các đặc trưng tiềm ẩn thành dự đoán xu hướng giá, đảm bảo mô hình có thể đưa ra quyết định giao dịch dựa trên các mẫu đã học.

3.2.4. Huấn luyện mô hình

Mô hình được huấn luyện trên tập train (70% dữ liệu) với các tham số:

- ✧ Hàm mất mát: CrossEntropyLoss, phù hợp với bài toán phân loại 3 nhãn.
- ✧ Tối ưu hóa: Adam optimizer với learning rate 0.001, weight decay 1e-5.
- ✧ Scheduler: ReduceLROnPlateau, giảm learning rate khi loss validation không cải thiện.
- ✧ Early stopping: Dừng huấn luyện nếu loss validation không giảm sau 10 epoch.
- ✧ Batch size: 512.
- ✧ Số epoch: Tối đa 40

Quá trình huấn luyện được thực hiện trên GPU T100 trên kaggle, với dữ liệu được chia thành các batch và xử lý qua DataLoader của PyTorch.

3.3. Liên hệ và so sánh với mô hình khác

Mô hình Transformer kết hợp Time2Vec được so sánh với các mô hình phổ biến trong dự đoán giá vàng, bao gồm ARIMA và LSTM, dựa trên hiệu suất phân loại xu hướng (BUY, SELL, HOLD):

- **ARIMA (Autoregressive Integrated Moving Average):** ARIMA là phương pháp thống kê truyền thống, phù hợp với dữ liệu tuyến tính và ổn định. ARIMA không tận dụng được các chỉ báo kỹ thuật hoặc đặc trưng thời gian, nên dễ dẫn đến hiệu suất thấp.
- **LSTM (Long Short-Term Memory):** LSTM là mô hình học sâu xử lý chuỗi thời gian tuần tự, có khả năng học phụ thuộc dài hạn. LSTM có nhược điểm là thời gian huấn luyện lâu (gấp 2-3 lần Transformer trên cùng phần cứng) do xử lý tuần tự. Trong các bài toán tài chính khác LSTM luôn có lợi nhuận thấp hơn Transformer.
- **Transformer kết hợp Time2Vec (đề xuất):** Mô hình đạt độ chính xác 66% và F1-score trung bình 0.66 trên tập test (977044 mẫu). Transformer vượt trội nhờ cơ chế attention, cho phép học các mối quan hệ phức tạp giữa các bước thời gian và đặc trưng (ví dụ: mối liên hệ giữa MACD và giá đóng cửa). Time2Vec bổ sung thông tin tuần hoàn, cải thiện khả năng dự đoán trong các khung giờ giao dịch đặc thù.

Kết luận: Transformer đạt hiệu suất dự đoán tốt nhất vì những lý do sau:

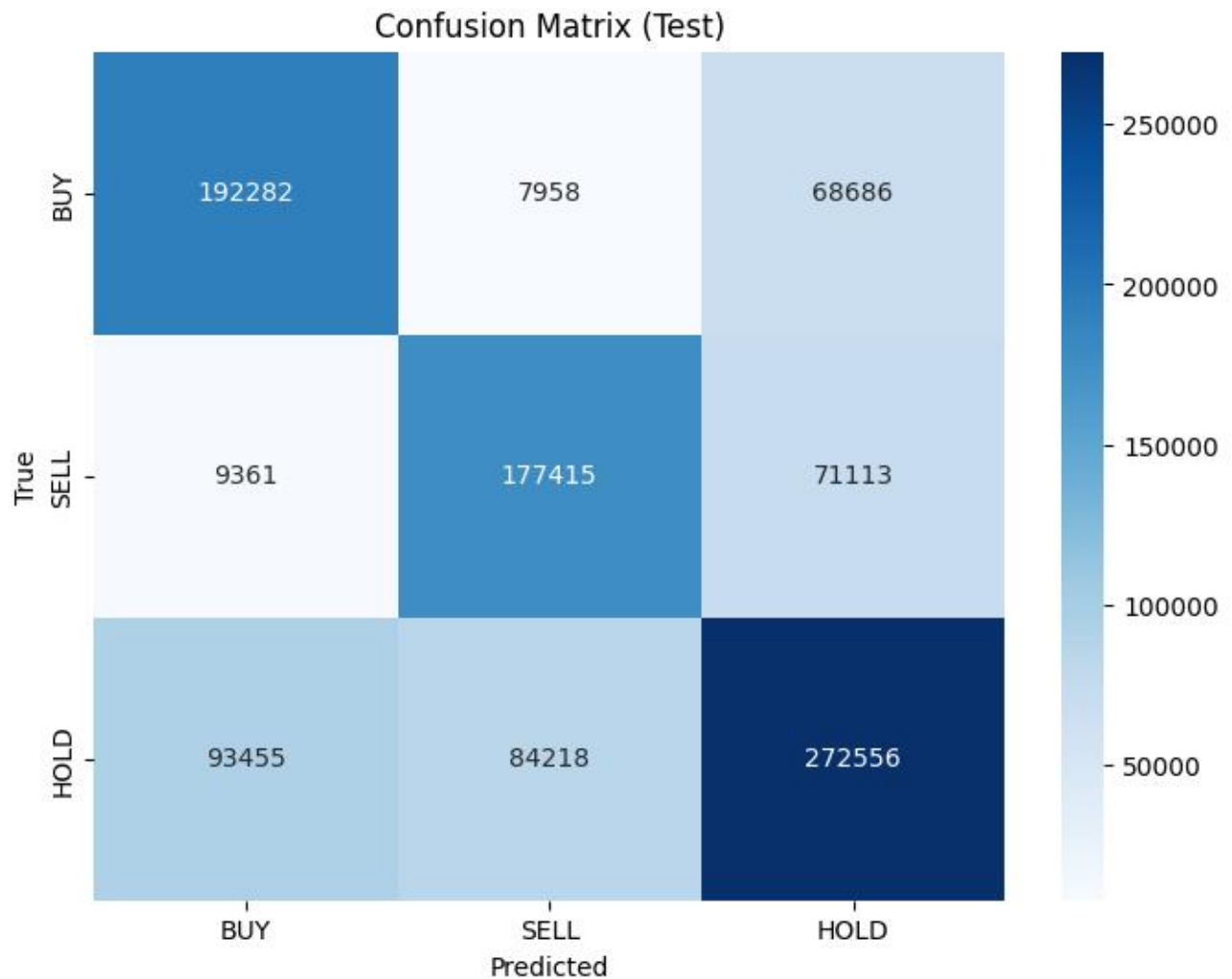
- ✓ Khả năng suy luận lý: Transformer có khả năng xử lý phụ thuộc ngữ cảnh tầm xa vượt trội do có attention và nhúng thời gian.
- ✓ Dữ liệu đầu vào là đa biến sẽ cải thiện hiệu suất của mô hình transformer.

IV. KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN TIẾP THEO

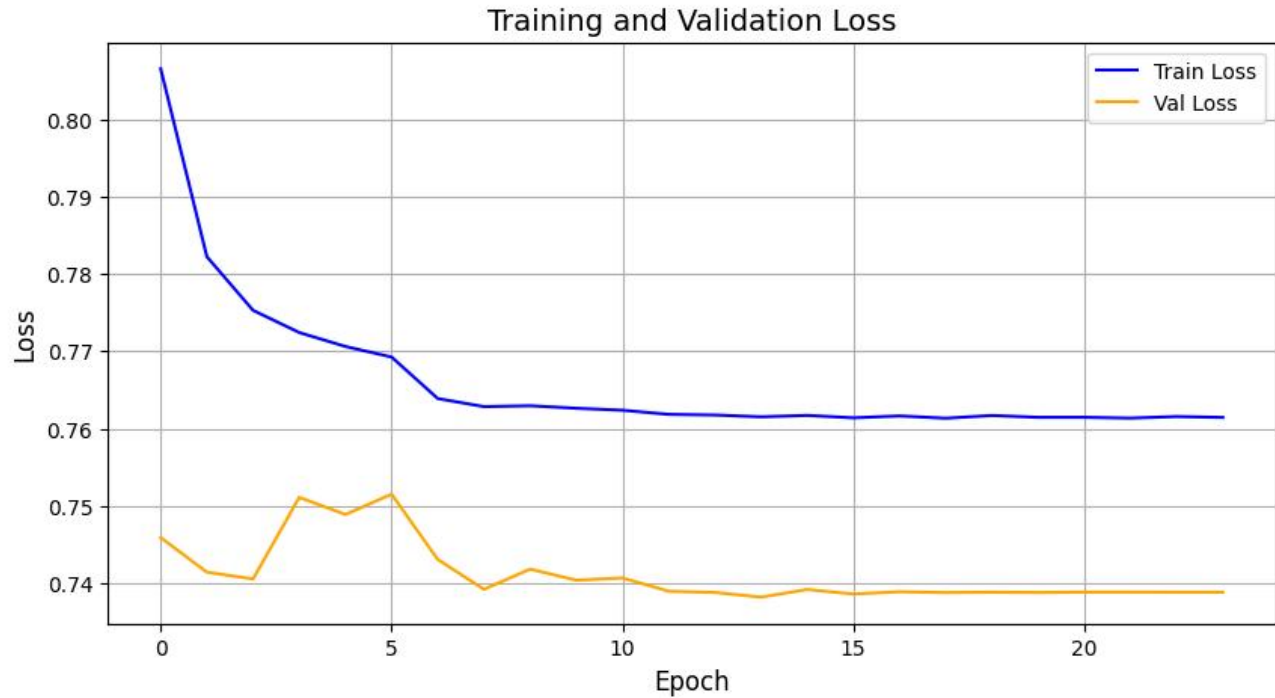
4.1 Kết quả

Mô hình Transformer kết hợp Time2Vec đạt được các kết quả sau trên tập kiểm tra (977,044 mẫu):

- **Độ chính xác tổng thể:** 66%.
- **Classification Report:**
 - **BUY:** Precision 0.65, Recall 0.71, F1-score 0.68, Support 268,926.
 - **SELL:** Precision 0.66, Recall 0.69, F1-score 0.67, Support 257,889.
 - **HOLD:** Precision 0.66, Recall 0.61, F1-score 0.63, Support 450,229.
- **Ma trận nhầm lẫn:** Mô hình dự đoán chính xác phần lớn nhãn BUY và SELL, nhưng nhầm lẫn nhiều giữa HOLD và các nhãn còn lại (do ngưỡng $\pm 0.05\%$ tạo ra nhiều mẫu HOLD gần ranh giới).



Hình 6: Confusion Matrix của tập Test



Hình 7: Biểu đồ Train Loss và Validation Loss khi train mô hình
 Khi train mô hình hiện tại dừng tại epochs thứ 24 / 40, với GPU T100

4.2 Kiến thức thu được và hướng phát triển

Dự án mang lại nhiều kỹ năng và kiến thức quý giá, hỗ trợ tôi trong việc nghiên cứu và ứng dụng trí tuệ nhân tạo trong tài chính:

- **Kỹ năng:**
 - Lập trình với Python, sử dụng các thư viện như PyTorch, pandas, numpy để xây dựng và huấn luyện mô hình học sâu.
 - Phân tích dữ liệu chuỗi thời gian, từ xử lý dữ liệu thô đến đánh giá mô hình.
 - Thực hiện backtesting và phân tích các chỉ số tài chính như Sharpe Ratio, Drawdown.
 - Viết báo cáo khoa học, trình bày kết quả nghiên cứu một cách rõ ràng và logic.
 - Xử lý lỗi trong lập trình và nghiên cứu, đặc biệt là lỗi backtesting do tiền xử lý sai.
- **Kiến thức:**
 - Hiểu sâu về kiến trúc Transformer, cơ chế attention, và Time2Vec trong xử lý chuỗi thời gian.
 - Ứng dụng chỉ báo kỹ thuật (RSI, MACD, Bollinger Bands) trong phân tích tài chính.

- Các phương pháp chuẩn hóa dữ liệu và mã hóa thời gian trong bài toán chuỗi thời gian.
- Nguyên lý backtesting và cách đánh giá hiệu quả giao dịch thực tế.

Hướng phát triển tiếp theo:

- **Thêm backtesting:** Chạy lại backtesting trên toàn bộ tập test để đánh giá hiệu quả thực tế.
- **Cải thiện mô hình:**
 - Tăng số tầng Transformer (num_layers=2) và kích thước d_model (64 hoặc 128) để học các mẫu phức tạp hơn.
 - Thử nghiệm các kỹ thuật regularization (như dropout 0.5) để giảm overfitting.
 - Điều chỉnh ngưỡng gán nhãn ($\pm 0.03\%$ thay vì $\pm 0.05\%$) để cân bằng phân bố nhãn.
- **Bổ sung dữ liệu ngoại sinh:** Thêm các yếu tố như lãi suất FED, chỉ số USD (DXY), hoặc tin tức thị trường để tăng độ chính xác dự đoán..
- **Ứng dụng thực tế:** Triển khai mô hình trên nền tảng giao dịch demo (như MetaTrader) để kiểm tra hiệu quả trong điều kiện thị trường thực.

V. KẾT LUẬN

Dự án đã thành công trong việc xây dựng một mô hình Transformer kết hợp Time2Vec để dự đoán xu hướng giá vàng (BUY, SELL, HOLD), đạt độ chính xác 66% trên tập kiểm tra với 977,044 mẫu. Mô hình tận dụng dữ liệu giá vàng XAUUSD, các chỉ báo kỹ thuật, và đặc trưng thời gian để học các mẫu phức tạp trong chuỗi thời gian, chứng minh tiềm năng của học sâu trong phân tích tài chính. Tuy nhiên, do chưa có backtesting nên chưa thể hiện lợi nhuận trên thị trường diễn biến.

Dự án không chỉ mang lại giá trị khoa học mà còn có ý nghĩa thực tiễn, đặc biệt tại Việt Nam, nơi vàng là tài sản đầu tư và tích trữ phổ biến. Các kỹ năng lập trình, phân tích dữ liệu, và tư duy nghiên cứu thu được từ dự án là nền tảng để tôi tiếp tục khám phá các ứng dụng AI trong tài chính. Hướng phát triển tương lai bao gồm phát triển backtesting, cải thiện mô hình với dữ liệu ngoại sinh, và triển khai giải pháp trên các nền tảng giao dịch thực tế. Những bài học từ dự án sẽ hỗ trợ tôi trong các nghiên cứu và ứng dụng sau này.

THAM KHẢO

[1] Vaswani, A., et al., "Attention is All You Need," Advances in Neural Information Processing Systems, 2017. Available: <https://arxiv.org/abs/1706.03762>.

Kazemi, S. M., et al., "Time2Vec: Learning a Vector Representation of Time," 2019.
[2] Available: <https://arxiv.org/abs/1907.05321>.

[3] <https://github.com/zhangmordred/QuantFormer/blob/main/2404.00424v1.pdf>

<https://www.sciencedirect.com/science/article/pii/S0957417424004032>
[4]

<https://arxiv.org/pdf/2211.07400>
[5]