Báo cáo Ariel Data Challenge 2025: Xây dựng mô hình học sâu dự đoán phổ quá cảnh

Đỗ Đức Thắng
VNU-UET
FIT
HaNoi, VietNam
23020158@vnu.edu.vn

***Tóm tắt nội dung*—Việc đặc trưng hóa bầu khí quyển của các hành tinh ngoài hệ Mặt Trời thông qua phân tích phổ là một bài toán phức tạp. Cuộc thi Ariel Data Challenge 2025 do NeurIPS tổ chức phối hợp với Cơ quan Vũ trụ châu Âu (ESA) đã tạo điều kiện cho cộng đồng nghiên cứu áp dụng các kỹ thuật học máy nhằm trích xuất thành phần khí quyển từ dữ liệu phổ mô phỏng. Trong báo cáo này, tôi tập trung vào việc xây dựng và tối ưu hóa mô hình dự đoán phổ quá cảnh từ các tín hiệu đo đạc đã qua tiền xử lí nhằm loại bỏ nhiễu kết hợp với các đặc trưng của hành tinh và ngôi sao chủ.**

***Index Terms*—học máy, học sâu**

I. Giới thiệu

Đặc trưng hóa bầu khí quyển các hành tinh ngoài hệ Mặt Trời là một trong những bài toán trọng yếu của thiên văn học hiện đại. Việc trích xuất chính xác các thành phần hóa học từ phổ quá cảnh (transit spectrum) quan sát được đóng vai trò then chốt để hiểu về tiến hóa hành tinh, thành phần khí quyển cũng như khả năng tồn tại sự sống. Cuộc thi Ariel Data Challenge 2025, thuộc chuỗi kì thi được NeurIPS tổ chức để thử nghiệm các phương pháp học máy xử lý tín hiệu phổ mô phỏng với độ nhiễu cao và nhiều hiệu ứng vật lý giống thiết bị thực tế.

Trong cuộc thi, tôi phát triển một mô hình dự đoán phổ dựa trên các trụ cột chính:

- Tiền xử lý dữ liệu (Signal pre-processing): Làm sạch dữ liệu, loại bỏ nhiễu do thiết bị đo đạc và tính toán độ sâu phổ trắng (White light transit depth)
- Cải thiện độ chính xác bằng Residual Network (ResNet): Mô hình sử dụng dữ liệu đầu vào gồm độ sâu phổ trắng và các đặc trưng của hành tinh để dự đoán phổ quá cảnh.
- Ước lượng độ không chắc chắn của dự đoán

II. Xử lí dữ liệu

2.1 Dữ liệu

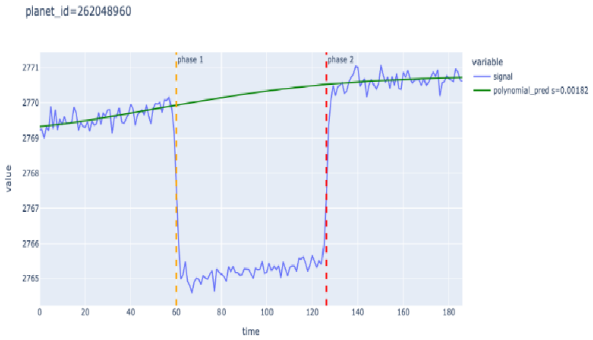
- Đối với mỗi hành tinh, dữ liệu được cho gồm có:
  - Tín hiệu FGS1:(FGS1_signal_0.parquet) với kích thước [135000, 32 × 32] (135000 bước thời gian, mỗi bước tương ứng với 0,1 giây và 32 × 32 là dữ liệu từ cảm biến)
  - Tín hiệu AIRS-CH0: (AIRS-CH0_signal_0.parquet) với kích thước [11250, 32 × 356] (11250 bước thời gian, mỗi bước tương ứng với 1,2 giây và 32 × 356 là dữ liệu từ cảm biến, gồm 32 chiều không gian và 356 bước sóng khác nhau)
- Bộ dữ liệu hiệu chỉnh gồm các thành phần: ảnh tối dùng để loại bỏ nhiễu nhiệt và độ lệch cảm biến; ảnh xác định điểm ảnh chết hoặc nóng để loại bỏ các điểm không hoạt động chính xác; ảnh phẳng dùng để hiệu chỉnh sự không đồng đều về độ nhạy giữa các điểm ảnh; thông tin hiệu chỉnh phi tuyến giúp khôi phục tín hiệu chính xác khi cảm biến gần bão hòa; và ảnh nhiễu đọc ghi lại nhiễu điện tử phát sinh trong quá trình đọc dữ liệu từ cảm biến
- Các đặc trưng của hành tinh gồm mã định danh, bán kính ngôi sao, khối lượng ngôi sao, nhiệt độ ngôi sao, khối lượng hành tinh, độ lệch tâm quỹ đạo, chu kỳ quỹ đạo, bán kính trục lớn, độ nghiêng quỹ đạo.

2.2 Tiền xử lí

- Mô hình áp dụng quy trình tiền xử lý dữ liệu theo các bước:
  - ADC (Analog-to-Digital Conversion): chuyển đổi tín hiệu thô sang giá trị số (đơn vị electrons) bằng hệ số gain và offset
  - Loại bỏ dòng tối (Dark current subtraction): xác định và loại bỏ pixel lỗi (hot/dead) để loại bỏ tín hiệu nền không mong muốn từ cảm biến.
  - Hiệu chỉnh phi tuyến (Linearity correction): hiệu chỉnh phi tuyến sử dụng polynomial model từ dữ liệu
  - Loại bỏ dark-current (Dark-Frame Subtraction): Trừ ảnh tối đã được hiệu chỉnh (loại bỏ pixel chết) khỏi ảnh gốc theo thời gian phơi sáng để loại bỏ tín hiệu nền không mong muốn từ cảm biến.
  - Lấy ảnh hiệu chỉnh đôi (Correlated Double Sampling): Tính hiệu giữa ảnh cuối và ảnh đầu mỗi chu kỳ phơi sáng để loại bỏ nhiễu và ổn định tín hiệu.
  - Gộp theo thời gian (Time Binning): Gộp nhiều khung hình theo thời gian để giảm dung lượng và làm mượt chuỗi thời gian quan sát.
  - Hiệu chỉnh trường phẳng (Flat Field Correction) Chia mỗi pixel cho giá trị tương ứng trong ảnh trường phẳng để hiệu chỉnh sự không đồng đều trong độ nhạy của từng pixel

– Loại bỏ các bước sóng ở ngoài rìa và tại các thời điểm đo đầu và cuối để tránh nhiễu, chỉ giữ lại các bước sóng ở trung tâm

2.3 Tính white light transit depth



Hình 1. Ví dụ về hành tinh id: 28204960

–

– Phương pháp Tính toán Độ sâu Quá cảnh bằng Tối ưu hóa

∗ Nguyên lý cốt lõi là mô hình hóa đường cong ánh sáng quan sát được như một hợp thành của hai thành phần chính:
(1) Sự biến đổi nhỏ của độ sáng ngôi sao theo thời gian
(2) Sự giảm đột ngột độ sáng gây ra bởi sự kiện quá cảnh.

∗ Quy trình bao gồm ba bước chính: xác định các pha quá cảnh, xây dựng hàm mục tiêu dựa trên mô hình khử xu hướng, và tối ưu hóa số để tìm tham số độ sâu.

· Xác định các Pha Quá cảnh.
Để tách riêng tín hiệu quá cảnh, cần xác định hai thời điểm đặc trưng: bắt đầu (T_1) và kết thúc (T_2) của sự kiện. Các mốc này được xác định bằng cách phân tích đạo hàm bậc nhất của chuỗi thời gian quang thông $F(t)$.
Thời điểm bắt đầu T_1 là nơi đạo hàm đạt cực tiểu:

$$\frac{dF}{dt} \to \min$$

Thời điểm kết thúc T_2 là nơi đạo hàm đạt cực đại:

$$\frac{dF}{dt} \to \max$$

· Mô hình hóa
Mô hình giả định rằng tín hiệu quang thông trong vùng quá cảnh, ký hiệu là $F_{\text{in-transit}}$, bị suy giảm một cách có hệ thống so với tín hiệu ngoài vùng quá cảnh $F_{\text{out-of-transit}}$. Độ sâu quá cảnh, ký hiệu là s, được định nghĩa là hệ số tỷ lệ mô tả mức độ suy giảm này.

Để tìm giá trị s, tôi xây dựng một tín hiệu đã hiệu chỉnh $F'(t, s)$, bằng cách áp dụng hệ số nhân $(1 + s)$ cho các điểm dữ liệu nằm trong vùng quá cảnh. Biểu thức được định nghĩa như sau:

$$F'(t,s) = \begin{cases} F(t) \cdot (1+s), & \text{nếu } T_1 \le t \le T_2 \\ F(t), & \text{nếu } t < T_1 \text{ hoặc } t > T_2 \end{cases}$$

Sự biến thiên nội tại của ngôi sao được giả định là một quá trình chậm, có thể mô hình hóa hiệu quả bằng một đa thức bậc thấp $P(t)$. Quá trình khớp và loại bỏ thành phần biến thiên này khỏi tín hiệu được gọi là *khử xu hướng* . Nếu tham số s được lựa chọn chính xác, tín hiệu đã hiệu chỉnh $F'(t, s)$ sẽ chỉ còn lại nhiễu và thành phần dao động chậm của sao, do đó sẽ khớp tốt với mô hình $P(t)$.

· Tính độ sâu quá cảnh s bằng tối ưu hóa
Mục tiêu của phương pháp là tìm giá trị tối ưu cho độ sâu quá cảnh s_{opt}, sao cho tín hiệu sau khi hiệu chỉnh và khử xu hướng khớp tốt nhất với mô hình biến thiên chậm của sao. Việc này được thực hiện bằng cách tối thiểu hóa một hàm mục tiêu $L(s)$, đo lường sai số còn lại giữa tín hiệu đã hiệu chỉnh và mô hình khử xu hướng.
Cụ thể, hàm mục tiêu được định nghĩa là sai số tuyệt đối trung bình (Mean Absolute Error - MAE):

$$L(s) = \frac{1}{N} \sum_{i=1}^{N} |F'(t_i, s) - P(t_i)|$$

Trong đó:

· $F'(t_i, s)$ là tín hiệu quang thông đã hiệu chỉnh tại thời điểm t_i,

· $P(t_i)$ là giá trị từ mô hình đa thức khử xu hướng tại thời điểm đó,

· N là tổng số điểm dữ liệu.

Giá trị tối ưu của độ sâu quá cảnh được xác định bằng:

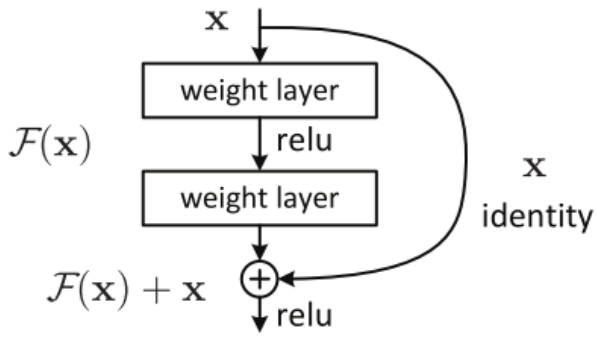
$$s_{\text{opt}} = \arg\min_{s} L(s)$$

III. Cải thiện độ chính xác bằng Residual Network

3.1 Residual Network

Residual Network (ResNet) là một kiến trúc mạng nơ-ron sâu được giới thiệu vào 2015, nổi bật với khả năng huấn luyện hiệu quả các mạng rất sâu nhờ cơ chế *residual learning*. Thay vì học trực tiếp một ánh xạ đầu vào $x \to H(x)$, ResNet học một hàm dư $F(x) = H(x) - x$, và kết hợp lại theo công thức:

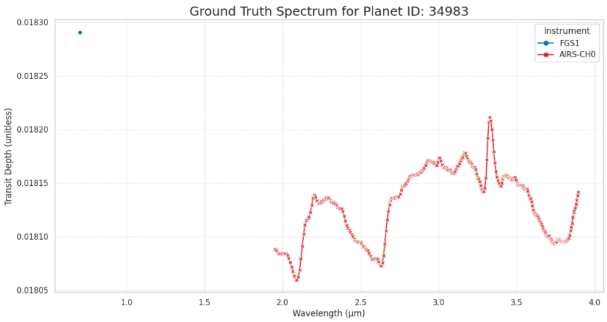
$$H(x) = F(x) + x$$



Hình 2. Cấu trúc của một khối residual

3.2 Áp dụng vào bài toán

Bài toán yêu cầu dự đoán 283 giá trị của phổ truyền qua (transit spectrum), một đại lượng vật lý quan trọng để mô tả khí quyển của ngoại hành tinh. Trong quá trình tiền xử lý, tôi nhận thấy giá trị *white light transit depth* (tức độ sâu quá cảnh trung bình trên toàn bộ dải sóng) đã là một ước lượng rất tốt cho toàn bộ phổ. Tuy nhiên, tại bước sóng $0.7\,\mu\text{m}$, giá trị này có sai số rất lớn so với tại các bước sóng khác.Đồng thời, sai số ở bước sóng này đóng góp rất lớn đến với số điểm cuối cùng. Đây là điểm cần cải thiện bằng mô hình học sâu.



Hình 3. Giá trị phổ của hành tinh 34983

Nhiệm vụ chính của mô hình học cách tinh chỉnh giá trị *white transit depth*. Nói cách khác, mô hình cần học phần dư (sự biến thiên của phổ tại bước sóng $0.7\,\mu\text{m}$) so với giá trị trung bình. Ý tưởng này có thể được biểu diễn bằng công thức:

$$\text{Spectrum}(\lambda) = \text{WhiteTransitDepth} + \Delta(\lambda, StarInfo) \quad (1)$$

Trong đó, nhiệm vụ của mạng nơ-ron là học hàm hiệu chỉnh Δ, một hàm phức tạp phụ thuộc vào bước sóng λ, các đặc tính của hành tinh và ngôi sao chủ $StarInfo$.

Với bản chất này của bài toán, kiến trúc Mạng Residual (ResNet) là một lựa chọn tự nhiên và tối ưu. Các kết nối tắt (skip connections) trong ResNet cho phép mô hình dễ dàng truyền thẳng tín hiệu đầu vào nếu không cần thay đổi, và chỉ tập trung học Δ.

3.3 Kiến trúc mạng

1) Lớp Đầu vào (Input Layer): Dữ liệu đầu vào của mô hình bao gồm các thông tin:

- White Transit Depth: Một giá trị vô hướng.
- Đặc trưng hành tinh và sao chủ: Chỉ sử dụng 2 đặc trưng có tương tác mạnh nhất đến kết quả đầu ra, bao gồm: Ms (khối lượng sao được đo theo đơn vị khối lượng Mặt Trời, $M_\odot$) và i (độ nghiêng quỹ đạo được đo theo đơn vị độ).

Các Khối Residual (Residual Blocks)

Phần lõi của mô hình là một chuỗi gồm 80 khối residual xếp chồng lên nhau, được định nghĩa trong lớp ResidualBlock. Mỗi khối xử lý một vector có chiều là 32. Cấu trúc của một khối được mô tả trong Bảng I.

Một điểm quan trọng là kết nối tắt (skip connection): đầu vào x của khối (identity) được cộng trực tiếp với kết quả sau lớp BatchNorm1d thứ hai. Toàn bộ tổng này sau đó được đưa qua một hàm kích hoạt ReLU. Công thức toán học có thể được biểu diễn như sau:

$$\text{output} = \text{ReLU}(\text{BN}_2(\text{FC}_2(\text{Dropout}(\text{ReLU}(\text{BN}_1(\text{FC}_1(x))))))+x)$$

Bảng I
CẤU TRÚC CHI TIẾT CỦA MỘT KHỐI RESIDUAL DỰA TRÊN MÃ NGUỒN.

Thứ tự	Lớp	Chi tiết / Kích hoạt
1	Linear 1 (fc1)	32 đơn vị (units)
2	Batch Normalization 1 (bn1)	-
3	ReLU	Hàm kích hoạt
4	Dropout	Tỷ lệ: 0.2
5	Linear 2 (fc2)	32 đơn vị (units)
6	Batch Normalization 2 (bn2)	-
7	(Kết nối tắt) + ReLU	Cộng với đầu vào và qua ReLU

2) Lớp Đầu ra (Output Layer): Lớp cuối cùng của mạng chứa 1 giá trị, tương ứng với giá trị tại bước sóng $0.7\,\mu\text{m}$.

3) Hàm mất mát:

- Hàm mất mát (Loss Function): Sử dụng hàm mất mát Mean Absolute Error (MAE) làm hàm mất mát trong quá trình huấn luyện. Công thức được định nghĩa như sau:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (2)$$

Trong đó:

- N là tổng số mẫu trong tập dữ liệu.
- y_i là giá trị thực tế của mẫu thứ i.
- $\hat{y}_i$ là giá trị do mô hình dự đoán cho mẫu thứ i.

IV. Ước lượng Độ không chắc chắn (Uncertainty Estimation)

4.1 Phương pháp ước lượng Homoscedastic (Đồng phương sai)

Phương pháp được sử dụng trên một giả định cơ bản gọi là tính đồng phương sai (homoscedasticity). Giả định này cho rằng phương sai của sai số dự đoán là không đổi trên toàn bộ không gian đầu vào.

Nói cách khác, chúng ta giả định rằng mô hình có cùng một mức độ sai sót cơ bản cho dù nó đang dự đoán ở bất kỳ bước sóng nào. Do đó, chúng ta có thể biểu diễn độ không chắc chắn của toàn bộ mô hình bằng một giá trị duy nhất, không đổi.

4.2 Quy trình Tính toán

Giá trị độ không chắc chắn được xác định bằng độ lệch chuẩn của phần dư (residuals) trên tập dữ liệu huấn luyện. Phần dư chính là sai số giữa giá trị thực tế và giá trị mà mô hình dự đoán. Quy trình tính toán như sau:

1) Dự đoán trên tập huấn luyện: Sau khi huấn luyện xong, mô hình được sử dụng để dự đoán lại toàn bộ các mẫu x_i trong tập huấn luyện, thu được các giá trị dự đoán $\hat{y}_i$.
2) Tính toán phần dư: Với mỗi cặp giá trị thực tế y_i và giá trị dự đoán $\hat{y}_i$, phần dư e_i được tính:

$$e_i = y_i - \hat{y}_i \tag{3}$$

3) Tính độ lệch chuẩn: Độ không chắc chắn của mô hình (σ) được định nghĩa là độ lệch chuẩn của tập hợp tất cả các phần dư này:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(e_i - \bar{e})^2} \tag{4}$$

Trong đó N là số lượng mẫu trong tập huấn luyện và $\bar{e}$ là giá trị trung bình của các phần dư (thường gần bằng 0 đối với một mô hình được huấn luyện tốt).

Giá trị 0.0009 chính là kết quả của phép tính độ lệch chuẩn (4) trên tập dữ liệu cụ thể được sử dụng.

V. KẾT QUẢ

TÀI LIỆU