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

Đỗ Đức Thắng
Khoa Công nghệ Thông tin, ĐHQGHN - UET
Hà Nội, Việt Nam
23020158@vnu.edu.vn

***Tóm tắt nội dung*—Đặc trưng hóa bầu khí quyển của các ngoại hành tinh từ phổ quá cảnh là một bài toán phức tạp do dữ liệu đo đạc thường chứa nhiều nhiễu và hiệu ứng thiết bị. Cuộc thi Ariel Data Challenge 2025 do NeurIPS và Cơ quan Vũ trụ châu Âu (ESA) tổ chức cung cấp bộ dữ liệu mô phỏng thực tế nhằm thúc đẩy việc áp dụng các kỹ thuật học máy và học sâu trong lĩnh vực này. Báo cáo này trình bày mô hình dự đoán phổ quá cảnh dựa trên *Residual Network* (ResNet), sử dụng đầu vào gồm *white light transit depth* và các đặc trưng vật lý của hành tinh – sao chủ. Quy trình gồm tiền xử lý tín hiệu, tính toán white light transit depth bằng tối ưu hóa và tinh chỉnh giá trị tại bước sóng $0.7\,\mu\mathrm{m}$ bằng ResNet. Kết quả cho thấy ResNet với 80 khối residual giảm sai số MAE tại $0.7\,\mu\mathrm{m}$ khoảng 53.9% so với phương pháp baseline, đồng thời duy trì độ ổn định cao. Mô hình đề xuất minh chứng khả năng kết hợp thông tin vật lý và kiến trúc học sâu để cải thiện độ chính xác trong dự đoán phổ quá cảnh.**

***Index Terms*—Residual Network, học sâu, ước lượng bất định, PyTorch**

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 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 phương pháp dự đoán phổ quá cảnh dựa trên các bước chính:

- Tiền xử lý dữ liệu
- Tính white light transit depth
- Cải thiện độ chính xác bằng Residual Network (ResNet)
- Ước lượng độ không chắc chắn của dự đoán

II. 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 nhằm hiệu chỉnh số liệu giúp giảm thiểu nhiễu phát sinh trong quá trình đo đạc.
- 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.

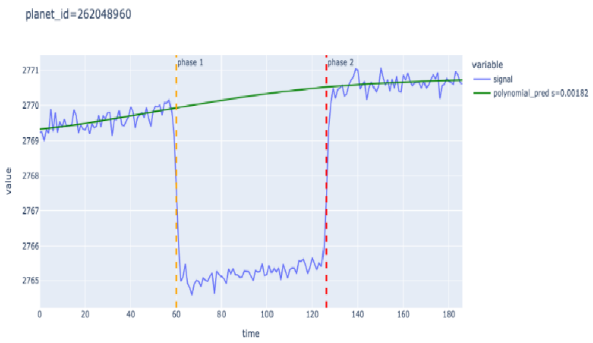
III. Phương pháp

3.1 Tiền xử lý dữ liệu

- 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

3.2 Tính white light transit depth

Phương pháp tính white light transit depth được tham khảo từ notebook Kaggle [1].



Hình 1. Ví dụ về hành tinh id: 28204960

– Phương pháp Tính toán white light transit depth 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 $F_{\text{out-of-transit}}$. White light transit depth, 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 3 $P(t)$. 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 s bằng tối ưu hóa
Mục tiêu của phương pháp là tìm giá trị tối ưu cho white light transit depth 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 white light transit depth được xác định bằng:

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

3.3 Áp dụng mô hình Residual Network

• Residual Network (ResNet) là một kiến trúc mạng nơ-ron sâu được phát triển nhằm giải quyết hiện tượng suy giảm hiệu suất khi tăng số lớp, thông qua cơ chế *residual learning*. Thay vì trực tiếp xấp xỉ ánh xạ mục tiêu $H(x)$ từ đầu vào x, ResNet mô hình hóa một hàm dư:

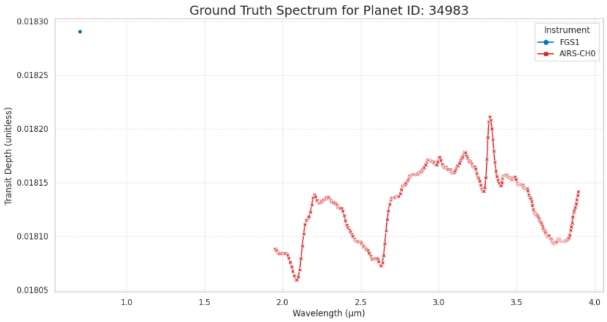
$$F(x) = H(x) - x \tag{1}$$

và tái cấu trúc đầu ra bằng:

$$H(x) = F(x) + x. \tag{2}$$

Cơ chế này cho phép truyền thẳng thông tin và gradient qua nhiều lớp, giảm thiểu *vanishing gradient* và cải thiện khả năng huấn luyện mạng sâu.

- Bối cảnh bài toán: Nhiệm vụ yêu cầu dự đoán 283 giá trị của phổ quá cảnh. Phân tích dữ liệu cho thấy *white light transit depth* là một ước lượng ban đầu khá chính xác cho toàn phổ. Tuy nhiên, tại bước sóng $0.7\,\mu\mathrm{m}$, sai số lớn hơn đáng kể và đồng thời số điểm nhận được ở bước sóng này chiếm tới 40% trong điểm số cuối cùng. Vì vậy, tôi sẽ tinh chỉnh giá trị ở bước sóng này bằng mô hình học sâu và sử dụng giá trị white light transit depth làm kết quả cho 282 bước sóng còn lại.



Hình 2. Phổ của hành tinh 34983 trong tập huấn luyện. Tại bước sóng $0.7\,\mu\mathrm{m}$, giá trị chênh lệch rõ rệt so với các bước sóng khác, gây sai lệch đáng kể khi ước lượng từ white light transit depth.

- Ý tưởng mô hình: Mục tiêu là học phần dư Δ tại bước sóng $0.7\,\mu\mathrm{m}$, được biểu diễn bởi:

$$\text{Spectrum}(\lambda) = \text{WhiteTransitDepth} + \Delta(\lambda, StarInfo) \tag{3}$$

với Δ phụ thuộc vào bước sóng λ và đặc trưng hành tinh, sao chủ $StarInfo$. Với đặc điểm này, ResNet là lựa chọn phù hợp vì có thể giữ nguyên thông tin gốc khi không cần điều chỉnh, đồng thời tập trung học chính xác phần hiệu chỉnh Δ.

3.4 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 các 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.

Đầu vào x của khối 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.

Bảng I
CẤU TRÚC CHI TIẾT CỦA MỘT KHỐI RESIDUAL

Thứ tự	Lớp	Chi tiết / Kích hoạt
1	Linear 1 (fc1)	32 đơn vị
2	Batch Normalization 1 (bn1)	-
3	ReLU	Hàm kích hoạt
4	Dropout	Tỷ lệ: 0.3
5	Linear 2 (fc2)	32 đơn vị
6	Batch Normalization 2 (bn2)	-
7	(Kết nối tắt) + ReLU	Cộng với đầu vào và qua ReLU

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

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:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4}$$

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. THỰC NGHIỆM

4.1 Thiết lập

Mô hình được huấn luyện và đánh giá trên tập dữ liệu của Ariel Data Challenge 2025 bằng phương pháp *K-Fold cross-validation* để tận dụng toàn bộ dữ liệu và giảm thiểu sai lệch do chia tập, với dữ liệu được chia thành 80% cho huấn luyện và 20% kiểm tra (validation). Kết quả cuối cùng được tính bằng trung bình của các chỉ số đánh giá trên tất cả các lần chia.

Quy trình tiền xử lý dữ liệu áp dụng các bước đã trình bày. Các siêu tham số huấn luyện được chọn như sau:

- Số epoch: 200
- Batch size: 32
- Learning rate: 1×10^{-3}, điều chỉnh tự động bằng ReduceLROnPlateau với hệ số giảm 0.5, patience = 10 epoch
- Optimizer: AdamW với weight decay = 1×10^{-4}
- Dropout: 0.3 trong các khối residual
- Seed: 42

Môi trường thực nghiệm:

- Framework: PyTorch

4.2 Kết quả và phân tích

Bảng II trình bày kết quả so sánh giữa mô hình đề xuất và các biến thể ablation. Kết quả cho thấy:

- Sử dụng White Transit Depth đơn thuần đã cho kết quả tốt trên hầu hết các bước sóng, nhưng sai số vẫn cao tại $0.7\,\mu\mathrm{m}$.

- Mạng ResNet giúp giảm đáng kể MAE tại $0.7\,\mu\mathrm{m}$ nhờ khả năng học phần dư Δ.
- Tăng số khối residual từ 20 lên 80 mang lại cải thiện rõ rệt, nhưng đồng thời làm chi phí tính toán tăng.
- Khi tăng lên 120 khối residual, MAE trên tập validation lại tăng, cho thấy mô hình bắt đầu bị overfitting do độ phức tạp vượt quá nhu cầu của dữ liệu.
- So với baseline, ResNet-80 giúp giảm MAE trên tập validation đi 53.9%.

Bảng II
KẾT QUẢ TRÊN TẬP VALIDATION (MAE) TẠI $0.7\,\mu\mathrm{m}$

Mô hình	**MAE**$_{0.7\mu m}$
Baseline (White Transit Depth)	0.001125
ResNet (20 blocks)	0.000950
ResNet (80 blocks)	0.000513
ResNet (120 blocks)	0.000712

V. ƯỚC LƯỢNG ĐỘ KHÔNG CHẮC CHẮN

Độ không chắc chắn trong bài toán này được tính bằng độ lệch chuẩn của sai số dự đoán trên tập validation, phản ánh mức độ phân tán của các sai số quanh giá trị trung bình.

Gọi y_i là giá trị thực và $\hat{y}_i$ là giá trị dự đoán tại mẫu thứ i, sai số được xác định bởi:

$$e_i = y_i - \hat{y}_i. \tag{5}$$

Độ bất định U được tính như sau:

$$U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}, \tag{6}$$

trong đó $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$ là sai số trung bình và n là số lượng mẫu trong tập validation. Công thức này sử dụng mẫu số n do tập validation được coi là toàn bộ tập dữ liệu quan tâm. Từ công thức trên, độ không chắc chắn của mô hình ResNet được sử dụng để dự đoán bước sóng $0.7\,\mu\mathrm{m}$ là 0.00078, độ không chắc chắn của cách ước lượng bằng white light transit depth giá trị của 282 bước sóng còn lại là 0.0009.

VI. KẾT QUẢ

Mô hình đề xuất đạt điểm số 0.327 và xếp hạng 54 trên bảng xếp hạng của cuộc thi *Ariel Data Challenge 2025*.

Đóng góp: Toàn bộ quá trình xây dựng mô hình, thực nghiệm và viết báo cáo được thực hiện bởi tác giả Đỗ Đức Thắng (100%).

VII. KẾT LUẬN

Bài báo cáo đã trình bày quy trình xây dựng mô hình dự đoán phổ quá cảnh từ dữ liệu của Ariel Data Challenge 2025. Phương pháp được sử dụng là sự kết hợp của việc tính toán giá trị white light transit depth với kiến trúc Residual Network cho phép mô hình học phần dư Δ để tinh chỉnh dự đoán tại các bước sóng $0.7\,\mu\mathrm{m}$.

Kết quả thực nghiệm cho thấy mô hình ResNet-80 đạt MAE thấp hơn 53.9% so với phương pháp baseline tại các bước sóng $0.7\,\mu\mathrm{m}$. Phân tích cũng cho thấy việc tăng số khối residual trên mức tối ưu có thể dẫn đến hiện tượng overfitting.

TÀI LIỆU

[1] V. Kudelya, "Neurips: Non-ml transit curve fitting," Kaggle, 2025, accessed: Aug. 10, 2025. [Online]. Available: https://www.kaggle.com/code/vitalykudelya/neurips-non-ml-transit-curve-fitting