Machine learning là : tìm kiếm dữ liệu hữu ích dựa trên của dữ liệu đầu vào trong khoảng không gian xác định trước, sử dụng hướng dẫn từ tín hiệu phản hồi

Deep in deep learning : các lớp biểu diễn

Tập dữ liệu (dataset): là tập hợp của các quan sát hay các mẫu dữ liệu và các nhãn nếu có được sử dụng để xây dựng mô hình.

Layer: là module xử lý dữ liệu nhận 1 hoặc nhiều tensors và xuất ra 1 hoặc nhiều tensors khác

Model- networks of layer:  Deep-learning model là biểu đồ theo chu kỳ có hướng của các lớp

Function loss- hàm mất mát , objective function-hàm mục tiêu: xác định độ chính xác của kết quả

label: nhãn là giá trị hoặc tập các giá trị được gán cho 1 mẫu dữ liệu để nhận dạng , lấy thông tin từ về mẫu đó

Optimizer- tối ưu hóa

baseline: đường cơ sở - giá trị  được tính toán mục đích là để đánh giá kết quả, hiệu suất của các mô hình

Cách thức hoạt động của deep learning

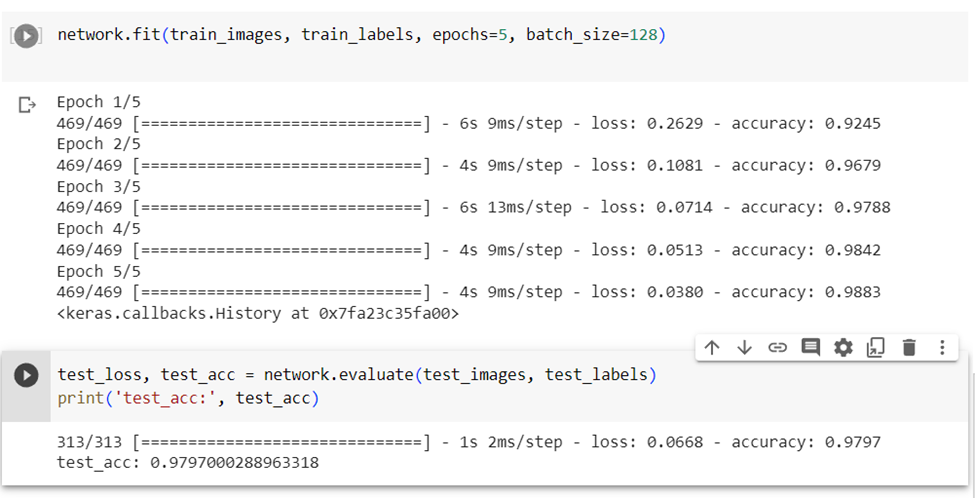
the loss of the network over the training data: độ mất mát trong quá trình đào tạo

the accuracy of the network over the training data: độ chính xác trong quá trình đào tạo

kết quả thực hiện

 2.1  cái nhìn đầu tiên về neural network

ví dụ sử dụng thư viện Keras trong bài toán xử lý chữ số viết tay



 kết quả sau khi training

ở đây có 2 giá trị là loss và accuracy chính là :

the loss of network over the training data: độ mất mát trên dữ liệu huấn luyện

the accuracy of the network over the training data: độ chính xác trên dữ liệu huấn luyện

·        test-set accuracy : độ chính xác trên tập huấn luyện ,

o   training set : là tập dữ liệu sử dụng để huấn luyện mô hình

·        training set accuracy : độ chính xác trên tập kiểm tra

o   Testing Set: tập dữ liệu không được sử dụng trong quá trình huấn luyện

3. bắt đầu với neural network

3.1 cấu tạo neural network

Việc huấn luyện các mạng nơron thông qua các đối tượng

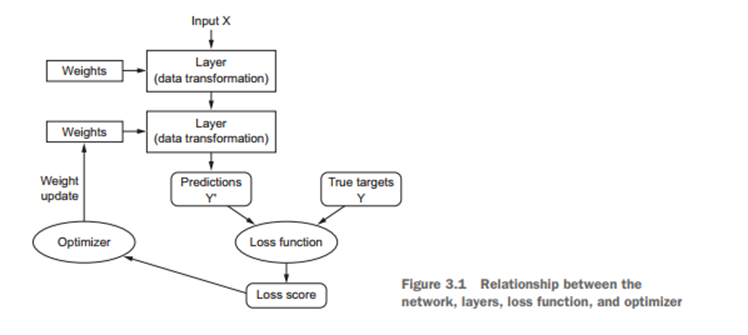
o   Layer: là module xử lý dữ liệu nhận 1 hoặc nhiều tensors và xuất ra 1 hoặc nhiều tensors khác, kết hợp thành network ( model)

o   Input data , target – dữ liệu đầu vào và mục tiêu

o   Loss function: xác định (feedback signal ) cho biết mức độ mất mát , sai số

o   Optimizer – trình tối ưu hóa: xác định cách thức học được tiến hành

Network bao gồm các layer được liên kết với nhau, ánh xạ dữ liệu đầu vào (input data) thành các dự đoán; loss function ( hàm mất mát) dựa vào đó so sánh dự đoán này với target tạo ra loss value ( giá trị mất mát) – đây là thước đo phù hợp của các dự đoán của network với những gì được mong đợi. The Optimizer( trình tối ưu hóa) dựa trên giá trị mất mát này để cập nhật trọng số của network.



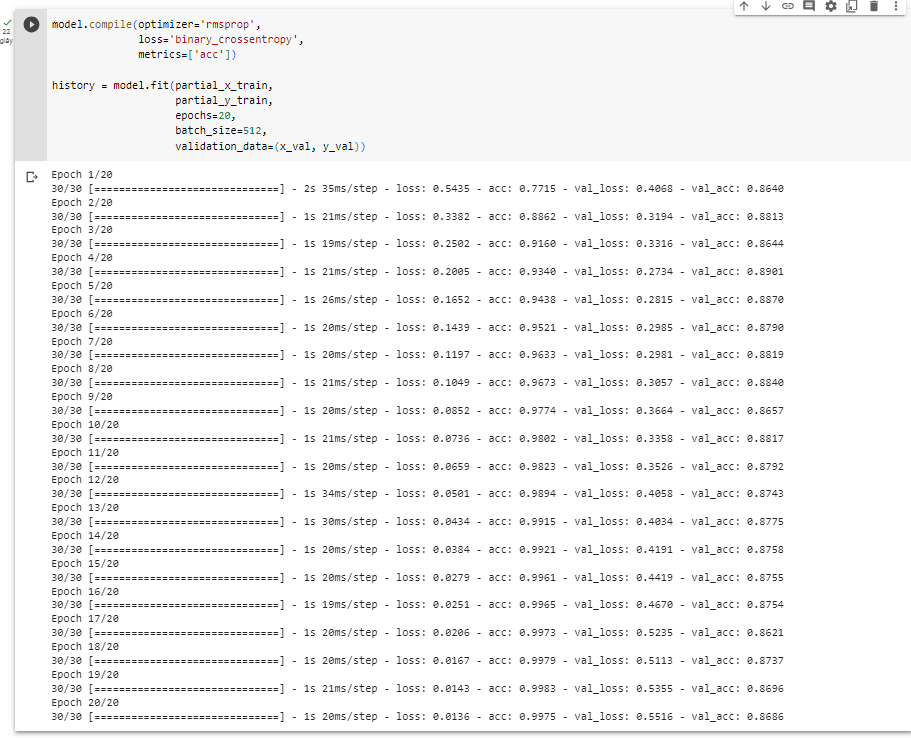
3.4 bài toán phân loại đánh giá phim

3.4.2 chuẩn bị dữ liệu

Không thể đưa trực tiếp danh sách các số nguyên vào mạng neural mà phải  biến đổi dữ liệu thành các tensor

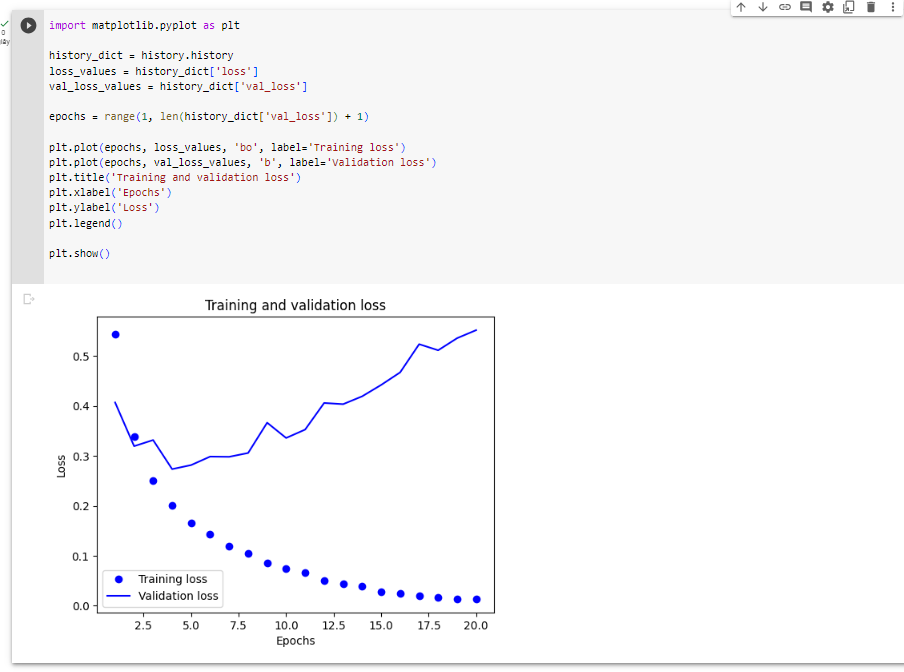
Phân loại nhị phân - bài toán phân loại đánh giá phim

**Figure 3.8. Training and validation accuracy**



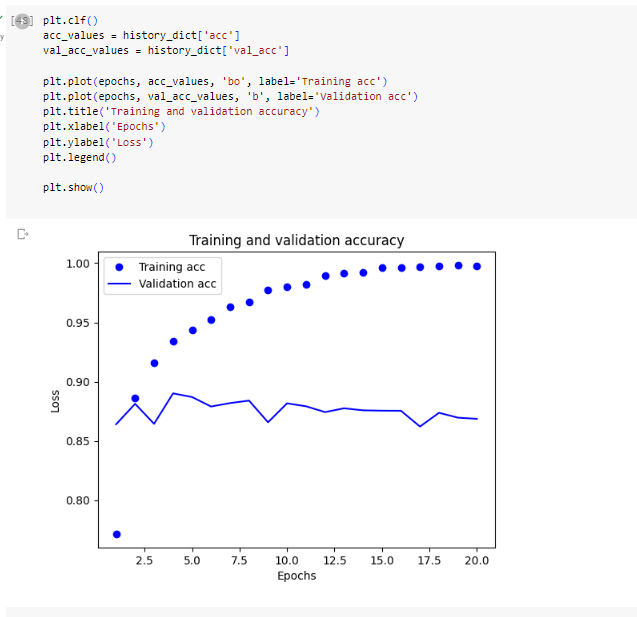
**Listing 3.9. Plotting the training and validation loss**

từ kết quả huấn luyện vẽ biểu đồ training và validation loss



**Listing 3.10. Plotting the training and validation accuracy**

từ kết quả training vẽ biểu đồ training and validation acc

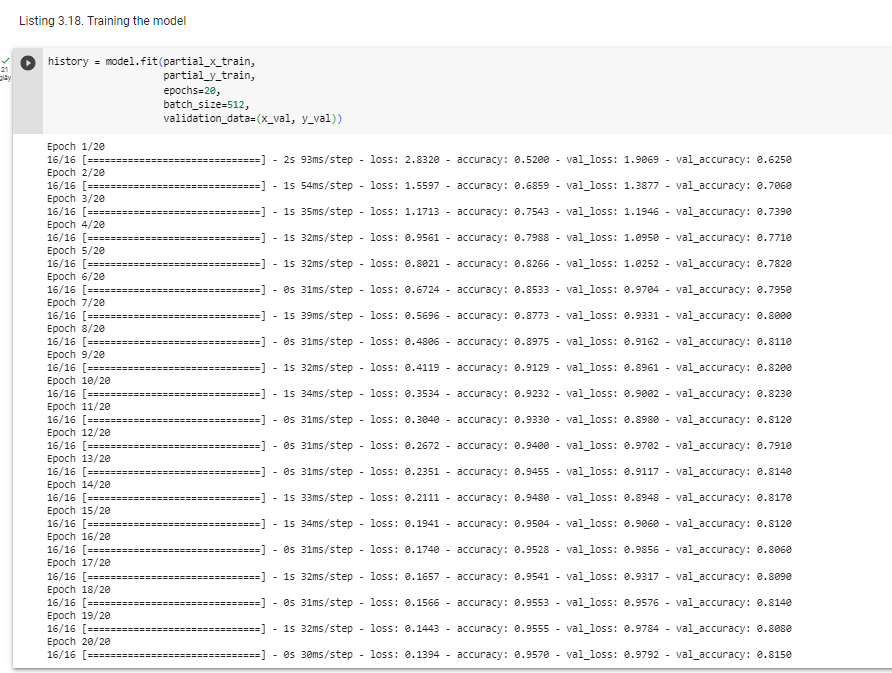


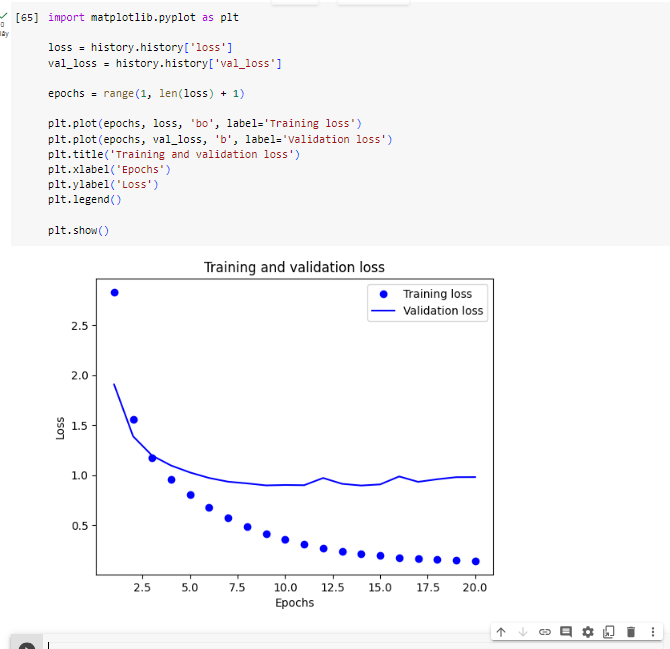
training loss giảm theo từng epoch và training  training accuracy tăng trên từng epoch

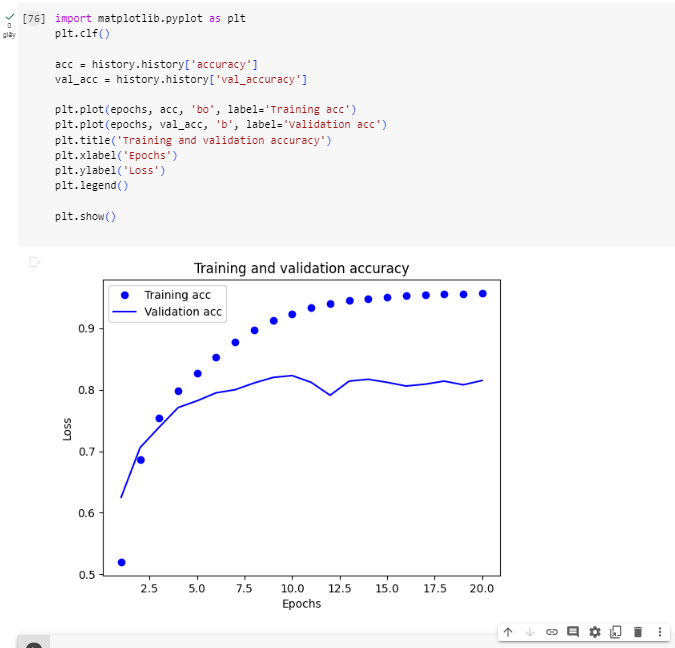
còn validation loss, validation accuracy  đạt đỉnh sau 4 epoch => vấn đề overfitting

overfitting là trường hợp mô hình quá tập trung vào training data mà  không tổng quát hóa cho dữ liệu mới

3.5 phân loại đa lớp- phân loại các bản tin

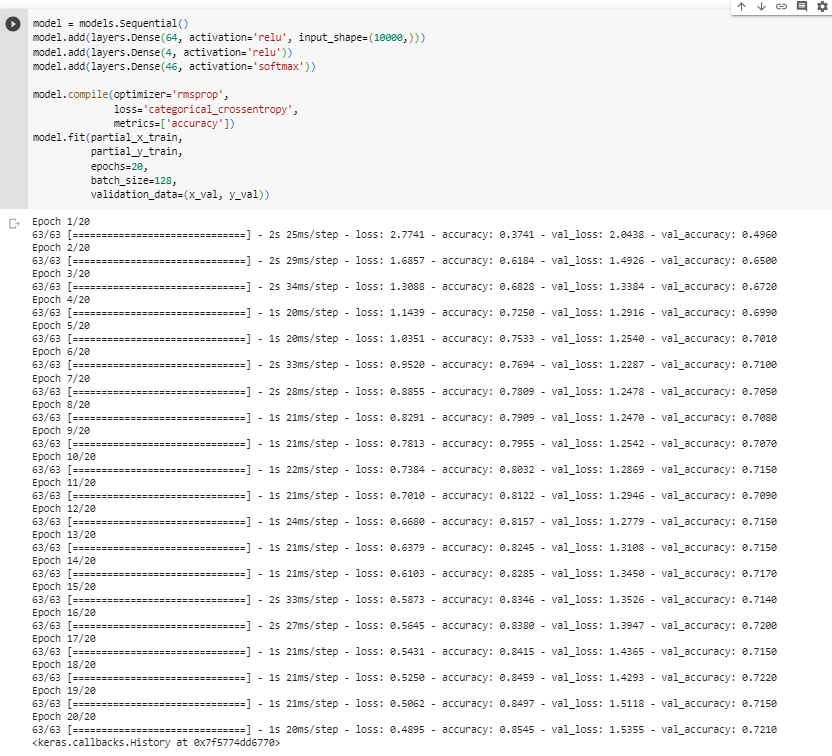






trình trạng overfitting sau 9 -10 enpouch

thử với trường hợp các lớp trung gian ít hơn 46 chiều ( số trường hợp đầu ra)



ta thấy kết quả ~ 0.7 => độ chính xác bị giảm

3.6 dự đoán giá nhà : bài toán hồi quy (regression )

cần đoán giá trị liên tục thay vì rời rạc như 2 ví dụ trên

trong bài toán này các mẫu có 13 đặc điểm, tỷ lệ tội phạm đầu người, số phòng tb mỗi nhà, khả năng tiếp cận cao tốc…. vì có nhiều đặc điểm không tương đồng ( phạm vi giá trị) vì vậy trong bước chuẩn bị dữ liệu cũng rất quan trọng.

 nếu đưa vào mạng neural các giá trị có phạm vi khác nhau thì mạng có thể tự động thích ứng với dữ liệu không đồng nhất như vậy khiến cho quá trình học khó khăn hơn.

trong bài toán này sử dụng phương pháp chuẩn hóa theo tính năng ( feature-wise normalization)

trong trường hợp có quá ít dữ liệu nên để đảm bảo mô hình được tin cậy thì có phương pháp xác thực chéo K-fold (K-fold cross-validation)

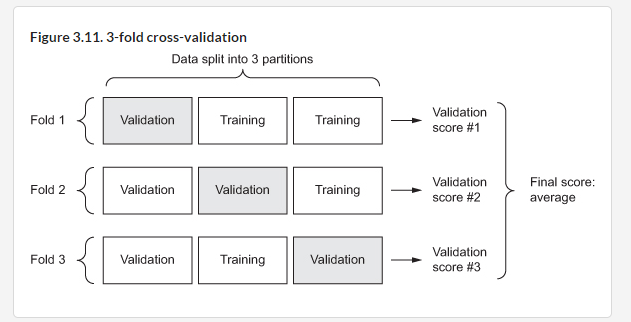
Xáo trộn dataset một cách ngẫu nhiên

  Chia dataset thành k nhóm bằng nhau (k-1 nhóm training và 1 nhóm Validation)

  Tiến hành huấn luyện trên tập training và Validation đã chia

Lặp lại quá trình K lần ( mỗi lần chọn 1 nhóm Validation khác nhau)

Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá

ví dụ về 3-fold cross-validation



các lần chạy cho giá trị validation scores khác nhau từ 2.19 -> 2.71 điểm trung bình là 2.51 là giá trị tin cậy hơn bất kì giá trị nào khác

4. nguyên tắc cơ bản của học máy

4 nhánh của ML

học có giám sát: học cách ánh xạ input data dựa trên targets

học không giám sát: tìm kiếm sự biến đổi dữ liệu input data mà ko có  sự hỗ trợ của targets cốt lõi của data analytics

học tự giám sát: trường hợp cụ thể của học có giám sát, học trong trường ko có targets do con người chú thích mà các nhãn do con người chú thích mà thông quá input data chúng tạo ra các nhãn

học tăng cường: mốt tác nhân nhận thông tin về môi trường của nó và học cách chọn các hành động để tối đa hóa phần thưởng. ví dụ điển hiền là các game chơi cờ

quy trình

1. xác định vấn đề và lập bộ dữ liệu: vấn đề gặp phải là gì, đầu vào là gì, bài toán giải quyết là gì, mục tiêu là gì, thu thập dữ liệu như thế nào
2. lựa chọn độ đo thành công thích hợp để đánh giá hiệu suất mô hình
3. lựa chọn phương pháp đánh giá mô hình :
4. chuẩn bị dữ liệu : là quá trình tiền xử lý và biến đổi dữ liệu phù hợp với mô hình để mang lại kết quả tốt ( gồm các bước như định dạng dữ liệu về tensor, tiền xử lý dữ liệu bị thiếu, loại bỏ nhiễu
5. phát triển mô hình tốt hơn so với một mô hình cơ bản
6. mở rộng quy mô
7. chính quy hóa mô hình của bạn và điều chỉnh siêu tham số của bạn

6.  deep learning for text and sequences

2 mô hình phổ biến trong xử lý chuỗi là

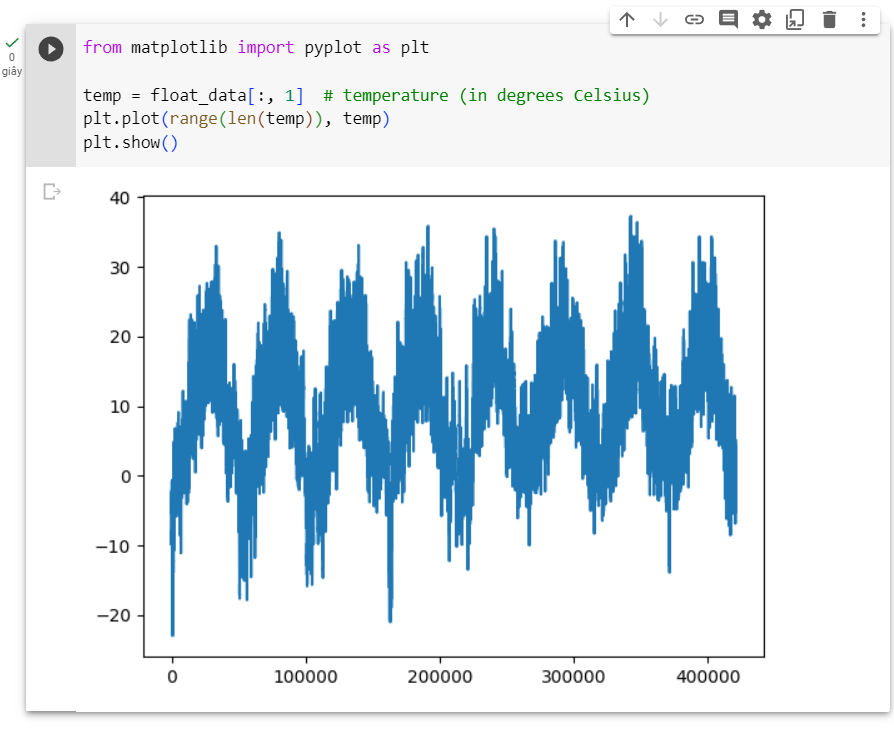
* recurrent neural networks
* 1D convnets

GRU, LSTM

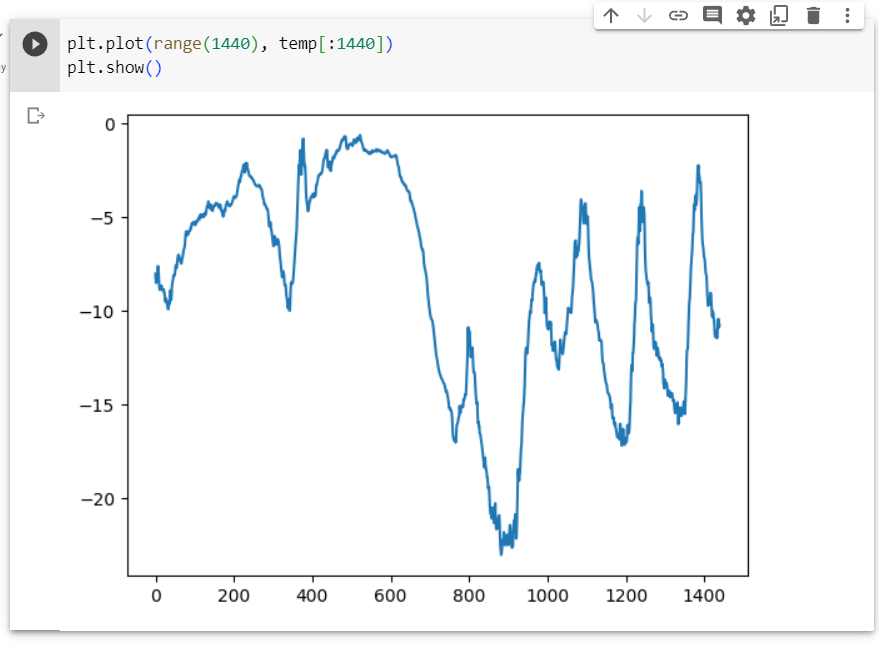
6.2 Recurrent Neural Networks (RNNs)

đặc điểm : xử lý thông tin tuần tự và phụ thuộc vào ngữ cảnh trước đó

6.3.1 bài toán dự báo nhiệt độ

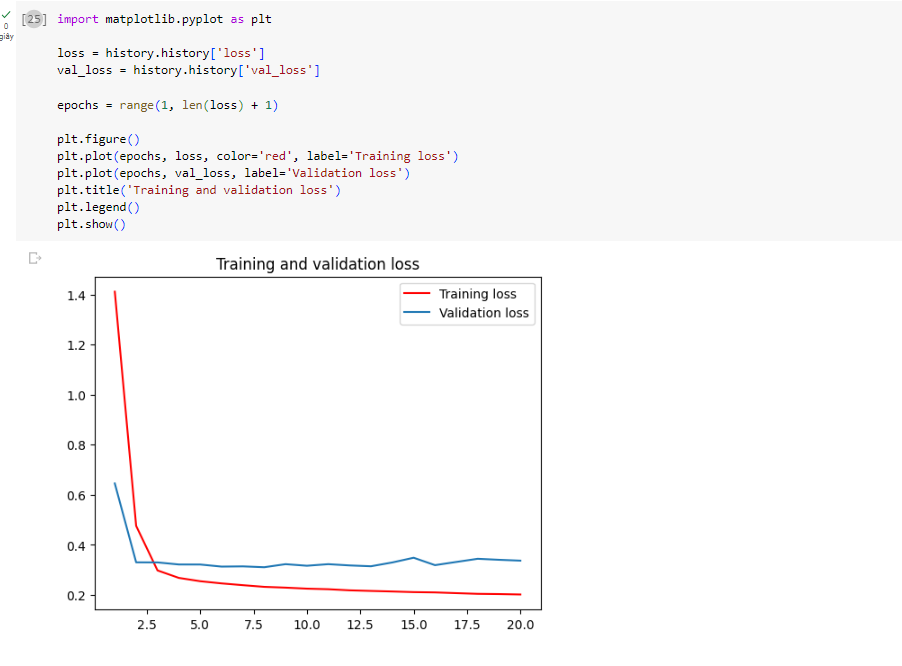


biểu đồ chuỗi thời gian nhiệt độ



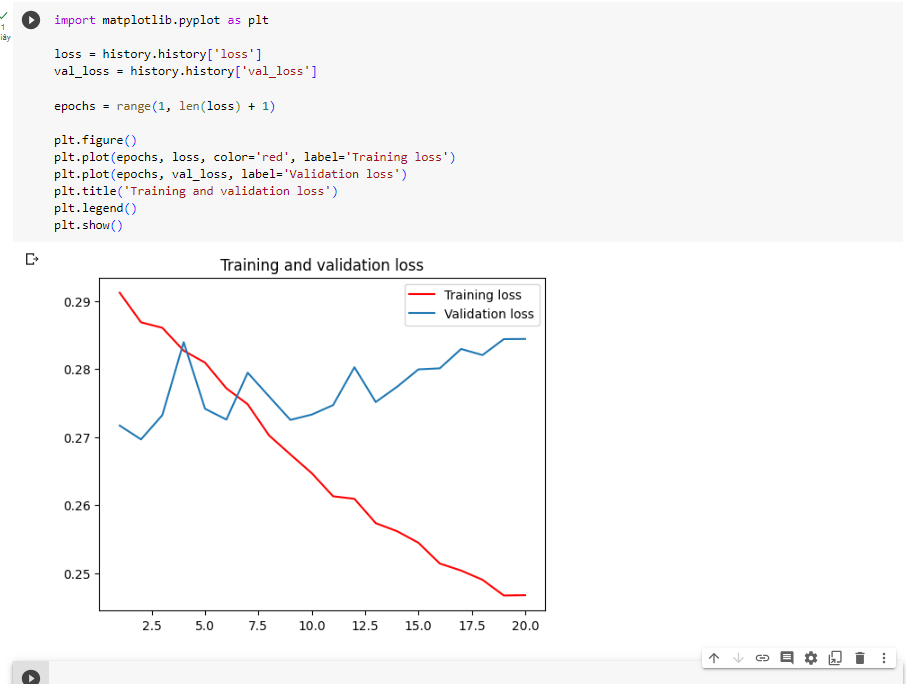
biểu đồ 10 ngày đầu tiên của chuỗi thời gian nhiệt độ

**A Basic Machine-Learning Approach**



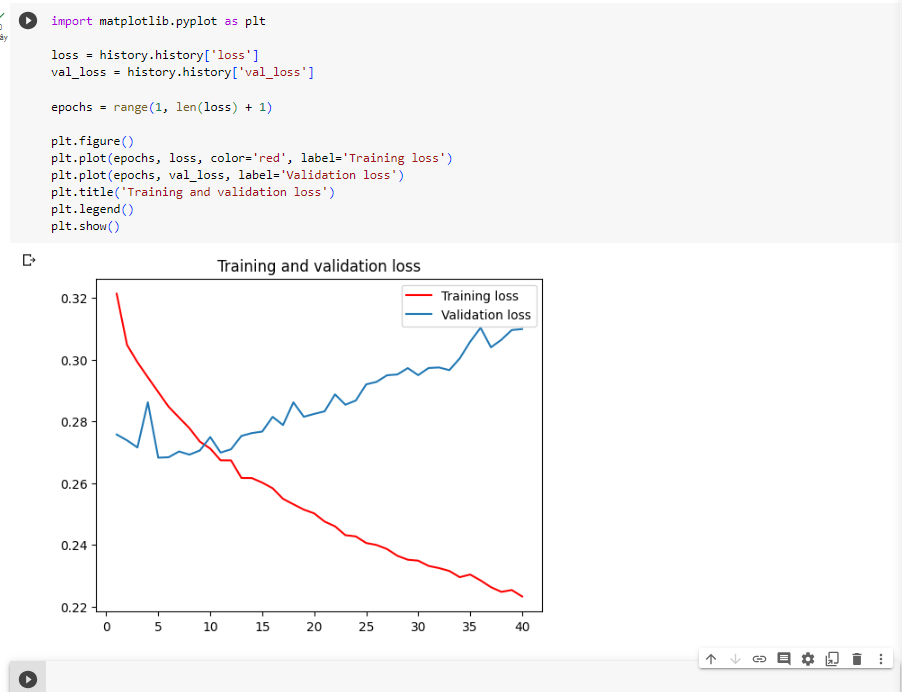
biểu đồ training loss và validation loss

**Train and Evaluate a GRU-based Model**

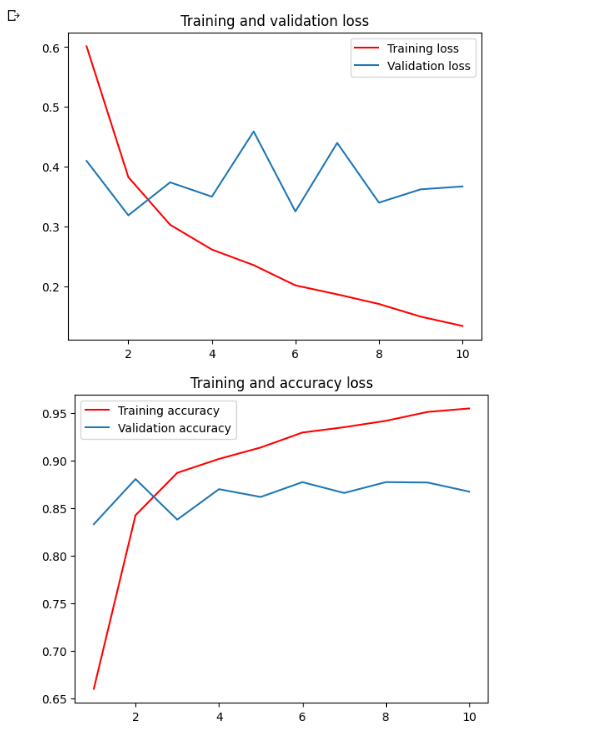


từ biểu đồ training loss cà validation loss có thể thấy tình trạng overfitting. tiếp theo ta sẽ sử dụng recurrent dropout để chống lại tình trạng overfitting

**Train and Evaluate a Dropout-regularized GRU-based Model**



**Train And Evaluate An LSTM Using Reversed Sequences**



**tổng kết**

   Khi tiếp cận 1 bài toán mới cần thiết lập các đường cơ sở, nếu ko có đường cơ sở thì không thể xác định được mô hình, thuật toán mình sử dụng tiến bộ như thế nào

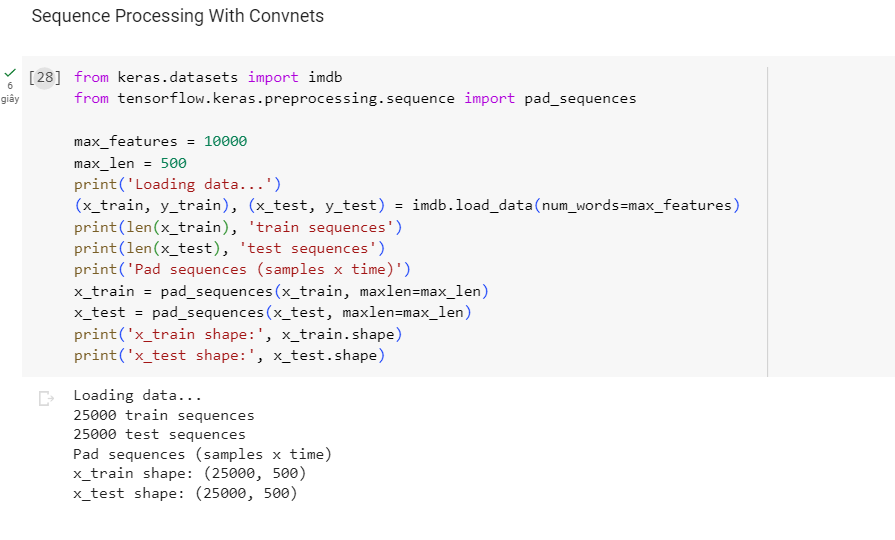
   Hãy thử các mô hình đơn giản trước những mô hình phức tạp, tốn kém đôi khi mô mình đơn giản lại mang tới sự lựa chọn tốt nhất

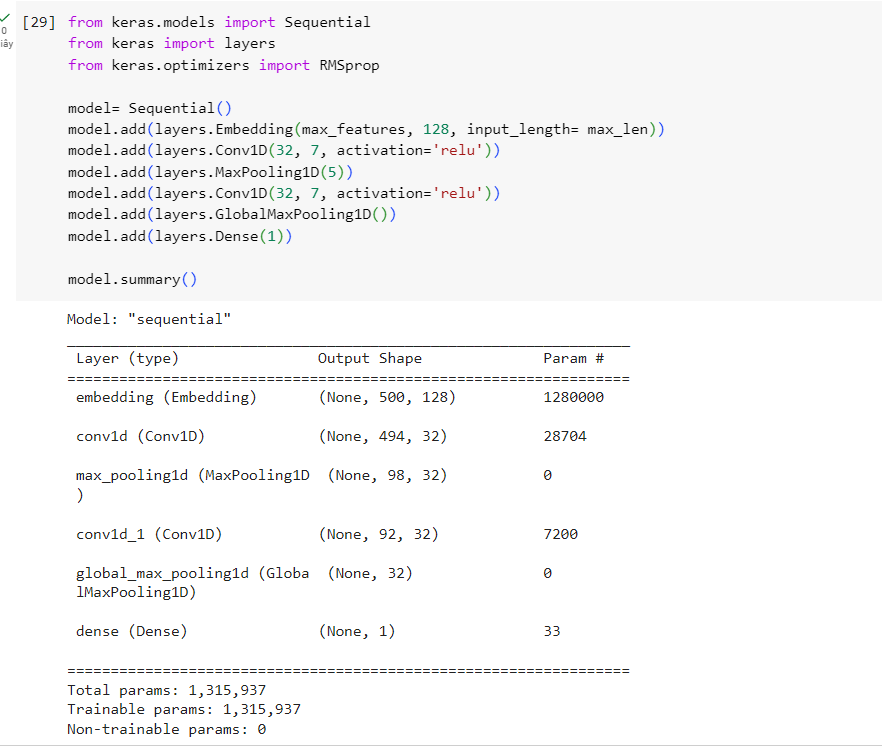
  Khi sử dụng dữ liệu trong đó thứ tự thời gian quan trọng thì recurrent networks phù hợp và hiệu quả .

  để sử dụng dropout với recurrent networks nên sử dụng time- constant dropout mask và recurrent dropout mask

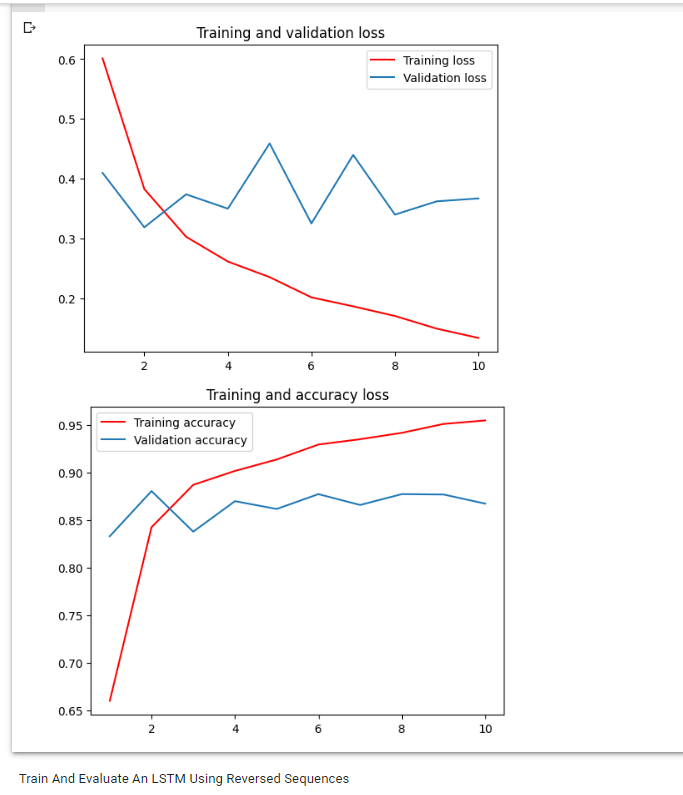
  Stacked RNNs cung cấp sự biểu diễn nhiều hơn so với RNN đơn lẻ nhưng không phải lúc nào cũng hiệu quả

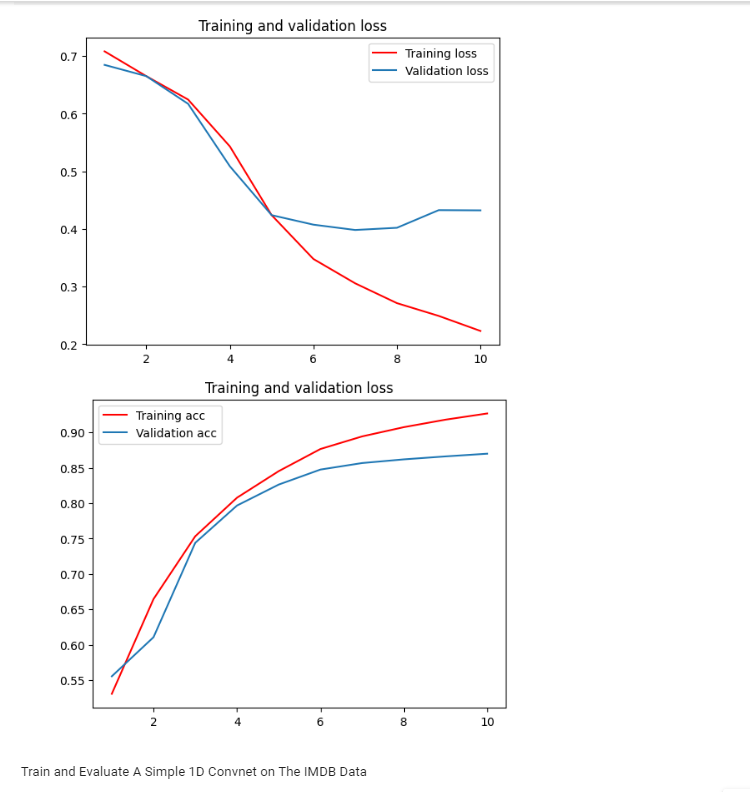
**Sequence Processing With Convnets**





**Train and Evaluate A Simple 1D Convnet on The IMDB Data**





validation accuracy thấp hơn 1 ít nhưng tốc độ lại nhanh hơn đáng kể 1D convert có thể mang lại giải pháp thay thế nhanh , re cho recurrent network

**Combine CNNs and RNNs to Process long Sequences**

**Train and Evaluate a Simple 1D Convnet on The Jena Data**