Báo cáo BTL Cuối kỳ Môn Học máy Nhóm 28

Lê Hoàng Lan
Mã số sinh viên: 22028013
K67-CA-CLC1
22028013@vnu.edu.vn

***Tóm tắt nội dung*—Bài báo cáo này sẽ tổng kết lại những gì em làm được trong bài toán phân loại thể loại phim từ bộ dữ liệu lấy từ bộ MoviesLen1M. Bộ dữ liệu bao gồm: ảnh poster phim, tiêu đề phim và đánh giá của người dùng về các bộ phim. Em đã sử dụng một mô hình CNN và mô hình ResNet50 để phân loại các bộ phim dựa trên poster của chúng. Trong 3 mô hình gồm một mô hình CNN, mô hình baseline có sẵn và mô hình pre-trained ResNet50 thì ResNet50 cho ra kết quả tốt nhất. Kết quả của bài toán có thể sẽ tốt hơn nữa nếu có thể áp dụng thêm các mô hình khác, chẳng hạn như ConvNext, DenseNet hay mô hình phát hiện vật thể YOLO.**

***Index Terms*—phân loại thể loại phim, multilabel classification, học sâu, CNN, ResNet50**

I. Hướng tiếp cận

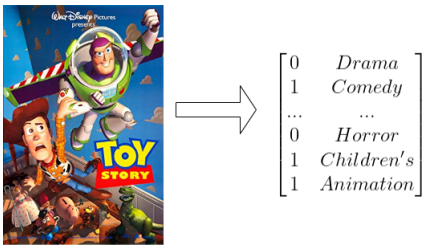
Bài toán được đặt ra là phân loại các nhãn thể loại cho các bộ phim từ tiêu đề, poster và đánh giá của người dùng về chúng. Sau khi phân tích và tìm hiểu, em quyết định lựa chọn poster phim làm đặc điểm (feature) để dự đoán, vì trong ngành làm phim ảnh, các tấm poster đóng vai trò vô cùng quan trọng. Một tấm poster sẽ đem lại cho công chúng ấn tượng đầu về bộ phim và từ đó có thể cảm thấy tò mò, thích thú và quyết định xem phim - nhân tố chính ảnh hưởng đến doanh thu của bộ phim. Chính vì vậy mà các tấm poster phim thường rất được đầu tư để làm sao mang lại nhiều thông tin và ấn tượng tốt nhất có thể về bộ phim.

Từ các tấm poster phim, nhóm sẽ sử dụng các mô hình CNN (mạng tích chập) và một mô hình pretrain ResNet50 để từ đó có thể đưa ra những dự đoán về thể loại phim. ResNet là một mô hình pretrain khá mạnh, nó thậm chí còn từng giành được vị trí thứ nhất cuộc thi ILSVRC với tỉ lệ lỗi chỉ 3.57%. Mô hình này cũng từng được vị trí thứ nhất trong cuộc thi ILSVRC and Coco 2015 với bộ dữ liệu ImageNet Detection. ResNet50 là một biến thể của ResNet, bên cạnh các mô hình khác như ResNet18, ResNet34, ResNet101,...

II. Xử lý dữ liệu

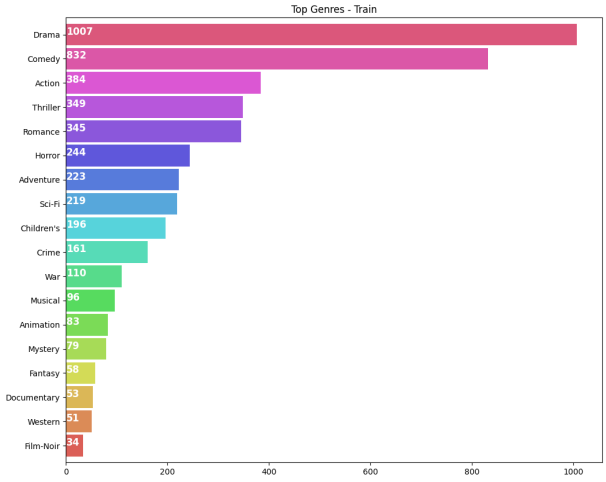
Bộ dữ liệu đã cho được chia thành 2 tập train và test. Tập train gồm 3106 bộ phim và tập test gồm 777 bộ phim. Khi thực hiện train model CNN tự tạo và model ResNet50, những bộ phim bị trùng lặp sẽ bị xóa hết, và chỉ có lần xuất hiện đầu tiên được giữ lại. Trong cả hai tập đều có những bộ phim không tìm được ảnh, vì vậy, với tập train, những bộ phim không có poster sẽ bị xóa, còn với tập test, các bộ phim sẽ được khởi tạo một ma trận RGB ngẫu nhiên. Sau khi tiến hành xóa, tập train còn 2602 bộ phim. Các tấm ảnh poster đều được đưa về kích cỡ 256 x 256 x 3. Đối với mô hình baseline và mô hình CNN, các phần tử trong ma trận ảnh đều được chuẩn hóa về khoảng $[0, 1]$, còn riêng với mô hình ResNet50 thì ma trận ảnh sẽ được đưa vào hàm xử lý của Keras. Các thể loại tương ứng của các bộ phim được chuyển thành vector multihot kích cỡ 18 x 1, với 18 là tổng số các thể loại phim ở trong file genre.dat. Xét một bộ phim bất kỳ, nếu bộ phim này thuộc thể loại x thì giá trị phần tử vector multihot tại chỉ số tương ứng cho x sẽ là 1, còn không thì sẽ là 0.

Identify applicable funding agency here. If none, delete this.

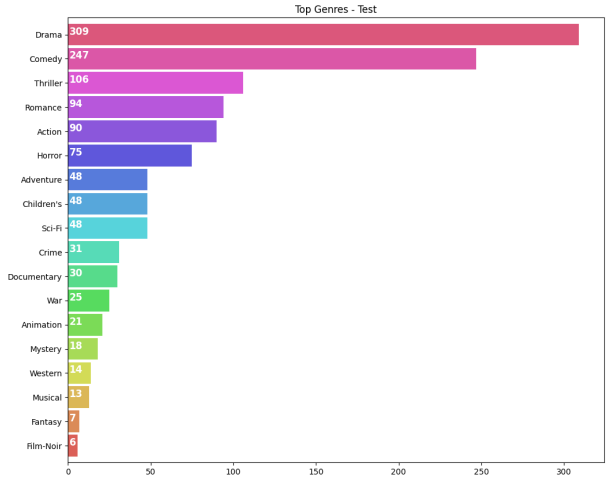


Hình 1. Chuyển các nhãn thể loại phim sang vector multihot

Bộ dữ liệu đã cho là một bộ dữ liệu không cân bằng. Thống kê số bộ phim thuộc từng thể loại được thể hiện như hình dưới đây. Có thể thấy rằng Drama là thể loại xuất hiện nhiều nhất, với lần lượt 1007 và 309 bộ phim trong tập train và test thuộc thể loại này. Trong khi đó, có những thể loại lại có rất ít bộ phim, chẳng hạn như Film-Noir, Fantasy hay Western.



Hình 2. Phân phối thể loại phim tập train.



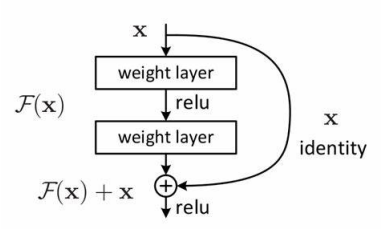
Hình 3. Phân phối thể loại phim tập test.

III. Các mô hình được sử dụng

A. Mô hình ResNet50

Một vấn đề thường gặp phải với các mô hình Deep Learning (học sâu) đó là hiện tượng Vanishing Gradient - hiện tượng đạo hàm giảm về 0 thường xảy ra khi số lượng lớp mạng tăng lên, dẫn đến việc các trọng số của mô hình không được cập nhật làm độ chính xác của mô hình giảm đột ngột.

Chính vì vậy, mạng ResNet ra đời để giải quyết tình trạng này. Mạng ResNet sử dụng các Residual Block có kết nối "tắt"(skipping connection) để xuyên qua một hay nhiều lớp. Các kết nối tắt này bổ sung đầu vào X vào đầu ra của layer, từ đó tránh được hiện tượng đạo hàm bằng 0 và ta có thể train được các lớp sâu hơn của mô hình Deep Learning.



Hình 4. Hình minh họa của một Residual Block. Giá trị được học thời điểm hiện tại là $F(x)$ qua kết nối "tắt"cho ra giá trị output thật là $F(x) + x$

ResNet50 gồm 50 lớp bao gồm lớp chập, lớp residual, pooling, lớp đặc và kết nối tắt.

Tổng quan cấu trúc của ResNet50:

- Input Layer: Lấy một ảnh kích cỡ 224 x 224.
- Lớp chập (Convolutional Layer): Đầu tiên là lớp chập 7x7 convolution với stride = 2, theo sau là một lớp batch normalization và hàm kích hoạt ReLU. Tiếp theo đó là một lớp pooling.
- Residual Blocks: Resnet50 gồm 16 Residual Block chia vào 4 stage. Mỗi stage có nhiều Residual Block với số lượng khác nhau và một số lượng bộ lọc khác nhau.
  - Stage 1: Bắt đầu là lớp tích chập giảm mẫu, sau đó là 3 Residual Block.
  - Stage 2: Có 4 Residual Block.
  - Stage 3: Có 6 Residual Block.
  - Stage 4: Có 3 Residual Block.

  Mỗi một khối dư trong một stage lại có cấu trúc tương tự nhau:
  - Kết nối trượt định danh: Một kết nối trượt bỏ qua các lớp tích chập chính.
  - Các lớp tích chập: Một chuỗi các tích chập 3x3, chuẩn hóa theo lô và các hàm kích hoạt ReLU.
  - Các lớp cổ chai (BottleNeck): Một lớp tích chập 1x1 được sử dụng để giảm số lượng bộ lọc trước các tích chập 3x3 chính.
  - Kết nối trượt chuyển đổi (nếu cần): Khi kích thước đầu vào và đầu ra của một khối dư khác nhau, sử dụng một lớp tích chập 1x1 để chuyển đổi đầu vào để phù hợp với hình dạng đầu ra.
- Global Average Pooling: Sau Residual Block cuối cùng, áp dụng một lớp Global Average Pooling để giảm kích thước không gian thành một bản đồ đặc trưng 1x1.
- Các lớp kết nối đầy đủ (Fully Connected Layer): Lớp Global Average Pooling được tiếp theo bởi một lớp kết nối đầy đủ với hàm kích hoạt softmax, tạo ra xác suất lớp cuối cùng.

Nhóm sử dụng mô hình pretrain trên tập dữ liệu ImageNet với một số điều chỉnh. Lớp output gồm 1000 nút của mô hình ResNet50 được thay thế bằng một lớp mạng đặc 1024, sau đó Droput với hệ số 0.3 trước khi được đưa đến mạng output cuối cùng gồm 18 nút. Hàm activation softmax của ResNet50 cũng được thay thế thành hàm sigmoid vì đây là bài toán phân loại đa nhãn (Multiple classification), còn hàm mất mát được đổi thành binary cross-entropy loss. Hàm sigmoid và hàm mất mát có công thức như sau:

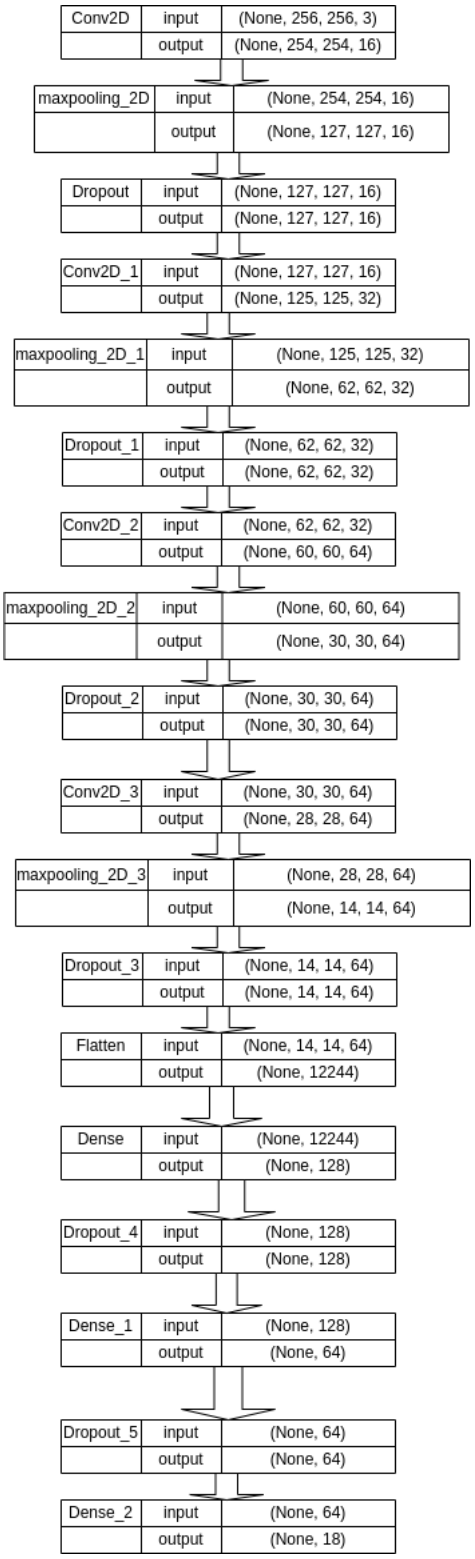
$$\sigma(z) = \frac{1}{1 + e^{(} - z)} \tag{1}$$

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{N} y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i}) \tag{2}$$

Ngoài ra, để tăng hiệu quả của quá trình training, em sử dụng data augmentation, bởi kích thước của tập dữ liệu train khá nhỏ nên cần thêm các dữ liệu khác để tăng hiệu quả.

B. Mô hình CNN

Trong mô hình có sử dụng MaxPooling để giữa các lớp chập nhằm giảm chiều ảnh mà vẫn có thể giữ lại được những thông tin quan trọng mà vẫn tránh được overfitting. Lớp Flatten được dùng để duỗi ma trận 3 chiều thành vector để từ đó có thể sử dụng các lớp liên kết đặc (Dense). Hàm kích hoạt ReLU được sử dụng sau các lớp Flatten và các lớp Convolution, còn hàm sigmoid được dùng để tính đầu ra (như ở ví dụ trên). Tương tự như với mô hình ResNet50, hàm mất mát được dùng là binary cross-entropy loss. Optimizer được dùng là Adam.



Hình 5. Mô hình CNN

IV. Thực nghiệm và kết quả

Threshold và batch size của mô hình baseline là 0.5 và 8 và được train trong 10 epoch. Mô hình ResNet50 có threshold và batch size lần lượt là 0.3 và 32, được train trong 20 epoch. Cuối cùng, mô hình CNN có threshold 0.3 và batch size 32, được train trong 30 epoch.

Do hầu hết các giá trị trong vector nhãn bằng 0, nên accuracy sẽ không phải là một tiêu chí tốt để đánh giá hiệu quả của 3 model, bởi một model mà cho kết quả tất cả các nhãn đều bằng 0 cũng có thể cho accuracy cao. Vì vậy, 2 chỉ số được dùng để đánh giá hiệu quả mô hình là F1-score và mAP@K.

F1-score được định nghĩa như sau:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{3}$$

$$Precision = \frac{Truepostive}{Truepositive + Falsepositive} \tag{4}$$

$$Recall = \frac{Truepostive}{Truepositive + Falsenegative} \tag{5}$$

Về map@K: Trước tiên, ta xét k nhãn được dự đoán có xác suất cao nhất. Precison@K được định nghĩa là Precision tính đến k, tức là bằng số nhãn thật sự được gán cho dữ liệu chia cho k. Nói cách khác:

$$precision@k = \frac{truepositive}{k} \tag{6}$$

AP@K có công thức như sau:

$$AP@k = \frac{1}{k}\sum_{i=1}^{k} precision@k * rel(k) \tag{7}$$

với $rel(k) = 1$ nếu nhãn k là true positive, 0 nếu ngược lại.

Và cuối cùng, mAP@K là lấy trung bình AP@K trên tất cả n dữ liệu.

Model	F1	map@1	map@3	map@5
Baseline	0.0741	0.040	0.045	0.066
CNN	0.1007	0.3745	0.317	0.317
ResNet50	**0.1629**	**0.4337**	**0.3893**	**0.3874**

Bảng I
Kết quả thực nghiệm.

Từ bảng kết quả trên, có thể thấy được rằng mô hình CNN nhỉnh hơn mô hình Baseline một chút còn mô hình ResNet50 có kết quả thể hiện tốt nhất. Điều này cũng chứng tỏ rằng mô hình có độ sâu càng lớn thì sẽ phân tích được kết quả tốt hơn.

Kết luận và các cải tiến có thể trong tương lai

Phân loại thể loại phim là một bài toán phân loại đa nhãn thú vị nhưng cũng nhiều thách thức. Trong bài này, mô hình ResNet50 và một mô hình CNN có độ sâu lớn hơn mô hình baseline đã cho thấy kết quả tốt hơn mô hình baseline. Mặc dù F1-score không cao, nhưng điều này có thể được cải thiện, chẳng hạn bằng cách thêm nhiều dữ liệu phim. Ngoài ra, ta cũng có thể thử một số mô hình khác, như MobileNet, ConvextNet,... hay thử cả mô hình phát hiện vật thể YOLO để phát hiện ra

người, sự vật,... liên quan đến một số thể loại phim nhất định. Bên cạnh sử dụng ảnh, kết hợp cả đánh giá của người xem để phân loại thể loại phim cũng là một hướng đi mà nhóm muốn thử trong tương lai.

V. TÀI LIỆU

TÀI LIỆU

[1] Mô hình ResNet50: ResNet50 paper
[2] Link video: Video thuyết trình
[3] Link Code với mô hình CNN: Link CNN
[4] Link Code với mô hình Resnet50: Link Resnet50
[5] Link Code Baseline: Baseline