Báo cáo kết quả huấn luyện mô hình phân loại cảm xúc

1. Đề bài.

Mục tiêu: Xây dựng mô hình phân loại cảm xúc cho các bài đánh giá phim từ dữ liệu IMDB, phân loại đánh giá là positive (tích cực) hoặc negative (tiêu cực).

Dữ liệu: Sử dụng bộ dữ liệu IMDB, chứa các bài đánh giá cùng nhãn cảm xúc (positive/negative). Dữ liệu được chia thành 2 phần: tập huấn luyện (5,000 mẫu) và tập kiểm tra (5,000 mẫu).

2. Tiền xử lý dữ liệu.

Các bước tiền xử lý em đã thực hiện:

- Loại bỏ thẻ HTML: Sử dụng biểu thức chính quy để loại bỏ các thẻ HTML trong văn bản.
- Loại bỏ ký tự không phải chữ cái: Các ký tự không phải chữ cái được thay thế bằng khoảng trắng.
- Chuyển văn bản thành chữ thường: Tất cả các từ trong văn bản đều được chuyển thành chữ thường để đồng nhất.
- Tách từ (Tokenization): Văn bản được phân tách thành các từ riêng biệt.
- Tạo từ điển (Vocabulary): Các từ trong tập huấn luyện được mã hóa thành các chỉ số số học thông qua một từ điển, với 10,000 từ phổ biến nhất.
- Mã hóa nhãn cảm xúc (positive/negative) bằng cách sử dụng LabelEncoder từ thư viện sklearn.

3. Mô hình.

Mô hình phân loại cảm xúc được xây dựng bằng một mạng nơ-ron có các thành phần chính sau:

- Embedding Layer: Chuyển các chỉ số từ (tokens) thành vector nhúng (embedding vectors).
- Hidden Layers: Các lớp ẩn được xây dựng bằng các lớp Linear, với các kích thước ẩn khác nhau tùy thuộc vào siêu tham số.
- Output Layer: Lớp đầu ra sử dụng hàm kích hoạt softmax để phân loại thành hai nhãn: tích cực và tiêu cực.
- Activation Functions: Các hàm kích hoạt như ReLU và Tanh được sử dụng tùy theo cấu hình siêu tham số.

4. Siêu tham số.

Để tối ưu mô hình, các siêu tham số được thử nghiệm với 5 cấu hình khác nhau. Các siêu tham số bao gồm:

- Batch size: Kích thước của mỗi batch trong quá trình huấn luyện.
- Learning rate: Tốc độ học (learning rate) của bộ tối ưu hóa Adam.
- Hidden sizes: Số lượng và kích thước của các lớp ẩn trong mô hình.
- Activation function: Hàm kích hoạt sử dụng trong các lớp ẩn (ReLU hoặc Tanh).

Các cấu hình siêu tham số em đã thử nghiệm:

- Cấu hình 1: batch_size = 32, learning_rate = 0.001, hidden_sizes = [128, 64], activation = 'relu'
- Cấu hình 2: batch_size = 32, learning_rate = 0.0005, hidden_sizes = [256, 128], activation = 'tanh'
- Cấu hình 3: batch_size = 64, learning_rate = 0.001, hidden_sizes = [256, 128, 64], activation = 'relu'
- Cấu hình 4: batch_size = 16, learning_rate = 0.0005, hidden_sizes = [512, 256], activation = 'tanh'
- Cấu hình 5: batch_size = 64, learning_rate = 0.01, hidden_sizes = [128, 64], activation = 'relu'

5. Kết quả.

Mô hình được huấn luyện và đánh giá trên mỗi cấu hình siêu tham số. Dưới đây là các kết quả trung bình về độ chính xác (accuracy) từ ba lần chạy mô hình trên mỗi cấu hình:

Cấu hình	Accuracy trung bình	Độ lệch chuẩn
Cấu hình 1	0.8638	0.0034
Cấu hình 2	0.8641	0.0032
Cấu hình 3	0.8617	0.0041
Cấu hình 4	0.8592	0.0050
Cấu hình 5	0.8598	0.0046

6. Nhận xét.

Các mô hình với cấu hình có kích thước ẩn nhỏ (như Cấu hình 1 và 2) vẫn cho kết quả khá tốt với độ chính xác xấp xỉ 86%.

Cấu hình với kích thước batch lớn và học nhanh (Cấu hình 5) không mang lại sự cải thiện đáng kể về độ chính xác, có thể do việc học quá nhanh không tối ưu hóa tốt các tham số.

Mô hình với activation = 'relu' dường như hoạt động tốt hơn so với activation = 'tanh'.

Kết quả mô hình cho thấy khả năng phân loại cảm xúc tương đối tốt, với độ chính xác ổn định và ít biến động qua các lần chạy.

7. Kết luận.

Mô hình phân loại cảm xúc trên dữ liệu IMDB đã đạt được độ chính xác cao với các cấu hình khác nhau. Mặc dù có sự khác biệt nhỏ trong kết quả giữa các cấu hình, tất cả đều đạt độ chính xác trên 85%, cho thấy mô hình có khả năng phân loại cảm xúc hiệu quả trên dữ liệu thử nghiệm.