Triển khai DistilBERT trên tập dữ liệu IMDb với bài toán Reviews Analysis

Liem Thanh

liemkg1234@gmail.com

1. Mô hình

DistilBERT là một mô hình ngôn ngữ được thiết kế để tối ưu hóa sự cân bằng giữa hiệu suất và hiệu quả khi xử lý ngôn ngữ tự nhiên. Nó được phát triển bởi Hugging Face và được giới thiệu trong bài báo khoa học "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter" vào năm 2019. Mục tiêu của DistilBERT là giảm kích thước của mô hình BERT ban đầu mà vẫn giữ được một phần lớn hiệu suất của mô hình gốc.

Đặc điểm của DistilBERT:

- **Kích Thước Nhỏ Gọn**: DistilBERT chỉ có khoảng 40% số lượng tham số so với BERT, điều này làm cho mô hình nhẹ hơn và dễ dàng được triển khai trên các thiết bị có hạn chế về tài nguyên tính toán.
- **Tốc Độ Xử Lý Nhanh**: Nhờ kích thước nhỏ gọn, DistilBERT có thể thực hiện dự đoán nhanh hơn so với BERT, làm tăng hiệu quả trong các ứng dụng thời gian thực và giảm chi phí tính toán.
- **Hiệu Suất Tốt**: Mặc dù nhỏ gọn, nhưng DistilBERT vẫn duy trì được một mức độ hiệu suất cao, đạt khoảng 95% hiệu suất của BERT trên một số bài toán ngôn ngữ tự nhiên.
- **Phương Pháp "Distillation"**: DistilBERT được huấn luyện sử dụng kỹ thuật gọi là "knowledge distillation", nơi kiến thức từ một mô hình lớn (như BERT) được "chưng cất" vào một mô hình nhỏ hơn. Quá trình này bao gồm việc huấn luyện DistilBERT để mô phỏng hành vi của mô hình BERT gốc, thông qua việc tối ưu hóa mục tiêu giữa việc dự đoán của mô hình lớn và mô hình nhỏ.

Do sự cân bằng giữa kích thước, tốc độ và hiệu suất, DistilBERT được sử dụng rộng rãi trong các ứng dụng như xử lý ngôn ngữ tự nhiên, phân tích cảm xúc, phân loại văn bản, và hơn thế nữa. Nó đặc biệt hữu ích trong các môi trường có hạn chế về tài nguyên tính toán, như thiết bị di động hoặc ứng dụng web thời gian thực.

2. Tập dữ liệu:

Internet Movie Database (IMDb) là một tập dữ liệu phổ biến được sử dụng trong nghiên cứu và phát triển các hệ thống xử lý ngôn ngữ tự nhiên (NLP), đặc biệt là cho các ứng dụng phân tích cảm xúc. IMDb là một cơ sở dữ liệu trực tuyến lớn chứa thông tin về phim, chương trình truyền hình, diễn viên, đạo diễn, và nhiều hơn nữa. Tuy nhiên, khi nói đến "tập dữ liệu IMDb" trong ngữ cảnh của NLP và học máy, người ta thường ám chỉ một tập con cụ thể của dữ liệu này được dùng để đánh giá các mô hình về khả năng phân tích cảm xúc.

Đặc điểm của IMDb:

- **Phân loại cảm xúc**: Tập dữ liệu thường được sử dụng để huấn luyện và kiểm thử các mô hình phân loại cảm xúc, với mục tiêu phân loại các đánh giá phim là tích cực hay tiêu cực.
- **Kích thước**: Bao gồm khoảng 50.000 đánh giá, được chia đều giữa tập huấn luyện và tập kiểm thử, với mỗi tập chứa số lượng đánh giá tích cực và tiêu cực tương đương nhau.
- **Đánh giá văn bản**: Mỗi mẫu trong tập dữ liệu là một đoạn văn bản đánh giá phim, thường dài và chi tiết, cung cấp nội dung đầu vào phong phú cho các mô hình học máy và NLP.
- **Đa dạng**: Đánh giá đến từ một loạt các phim và người dùng, đảm bảo sự đa dạng về ý kiến và phong cách viết, điều này giúp tạo ra các mô hình có khả năng tổng quát hóa tốt.

3. Phương pháp huấn luyện

Phương pháp: Học có giám sát.

Mô hình: distilbert-base-uncased, là một phiên bản có 66 triệu parameters, được huấn luyện trên tập dữ liệu không phân biệt chữ hoa, tối ưu hóa cho hiệu quả tính toán mà vẫn duy trì được một mức độ hiệu suất cao.

Tập dữ liệu: IMDb, được chia thành 3 phần: train/validation/test, mô hình được đánh giá độ chính xác trên tập test.

Siêu tham số:

- Epochs: 10
- Batch size: 16
- Max_length_input: 512
- learning_rate=0.00002
- Metrics: Accuracy, Precision, Recall, F1

4. Kết quả

- **Best epoch**: 3
- **Table result**:

	Accuracy	Precision	Recall	F1
test	0.92656	0.9265	0.9265	0.926

5. Vấn đề gặp phải

- Ở phần phân chia tập dữ liệu, tập validation được lấy từ tập train với 5000 samples, tuy nhiên labels trong tập validation chỉ toàn Positive, do đó khi training các chỉ số đánh giá rất thấp (~50%), dẫn đến không thể đánh giá mô hình đã hội tụ.