1. **Giới thiệu về CLIP (Contrastive Language-Image Pre-Training)**

Contrastive Language-Image Pre-Training làmột kiến trúc học đa phương thức do OpenAI phát triển, học các khái niệm thị giác từ giám sát ngôn ngữ tự nhiên. CLIP thu hẹp khoảng cách giữa dữ liệu văn bản và hình ảnh bằng cách huấn luyện đồng thời mô hình trên tập dữ liệu quy mô lớn gồm hình ảnh và mô tả văn bản tương ứng.

1. **Contrastive learning là gì ?**

Ý tưởng chính của Contrastive learning là tìm ra các cặp đặc trưng của dữ liệu có tính tương đồng - tương phản nhau trong bộ dataset. Từ đó, với những cặp dữ liệu mang tính tương đồng ta có thể "kéo" chúng lại gần để học được những đặc trưng cấp cao hơn của nhau, và ngược lại với những cặp những liệu tương phản sẽ bị "đẩy" ra xa. Để làm được điều này, ta sẽ cần sử dụng các similarity metric để tính toán khoảng cách giữa các embedding vector biểu diễn các điểm dữ liệu với nhau. Ví dụ, ta đã có 1 điểm dữ liệu gốc gọi là anchor, sau đó có thể dùng thêm các kỹ thuật augmentation khác nhau để có thêm 1 biến thể từ anchor gốc gọi là positive sample, và phần còn lại của batch / dataset sẽ được coi là negative sample. Sau đó model sẽ được train để có thể phân biệt được positive sample với negative sample từ 1 cụm dữ liệu.

1. **Kiến trúc CLIP**

**Contrastive Language-Image Pre-Training** sử dụng kiến trúc mã hóa kép (dual-encoder) để ánh xạ hình ảnh và văn bản vào latent space chung. Mô hình hoạt động bằng cách huấn luyện đồng thời hai bộ mã hóa: một cho hình ảnh (Vision Transformer) và một cho văn bản (mô hình ngôn ngữ dựa trên Transformer).

**( Latent space** là gì ? Latent space là không gian trừu tượng nơi các đặc trưng của dữ liệu được mã hóa. Khi dữ liệu như hình ảnh, văn bản hoặc âm thanh được đưa vào một mô hình học máy, nó thường trải qua một quá trình chuyển đổi để tạo ra các biểu diễn trong latent space.

Ví dụ:

**Hình Ảnh**: Giả sử ta có một tập dữ liệu hình ảnh về động vật. Latent space có thể chứa các vector biểu diễn các đặc trưng như "có lông", "có vây", "có đuôi", giúp mô hình phân loại và sinh ra hình ảnh động vật mới.

**Văn Bản**: Trong các mô hình ngôn ngữ, latent space có thể chứa các vector biểu diễn ý nghĩa của từ hoặc câu, cho phép mô hình hiểu và tạo ra ngôn ngữ tự nhiên.)

1. **Bộ mã hóa hình ảnh (Image Encoder)**

**Kiến trúc cơ bản**: Có thể là một mạng CNN như ResNet hoặc Vision Transformer (ViT).

**Xử lý đầu vào**: Bộ mã hóa hình ảnh trích xuất các đặc trưng nổi bật từ đầu vào thị giác. Bộ mã hóa này nhận một "hình ảnh làm đầu vào" và tạo ra một vector đặc trưng toàn cục đại diện cho nội dung của hình ảnh.

1. **Bộ mã hóa văn bản (Text Encoder)**

**Kiến trúc cơ bản:** Thường là một mô hình Transformer dạng mã hóa (encoder-only), tương tự như BERT.

**Xử lý đầu vào:** Bộ mã hóa văn bản mã hóa ý nghĩa ngữ nghĩa của mô tả văn bản tương ứng. Nó nhận một "chú thích văn bản/nhãn làm đầu vào" và tạo ra một một vector đặc trưng toàn cục khác.

1. **Không Gian Đặc Trưng Chung**

Cả hai bộ mã hóa tạo ra các embeddings trong một không gian vector chung. Không gian này cho phép CLIP so sánh các biểu diễn văn bản và hình ảnh, từ đó học được các mối quan hệ tiềm ẩn giữa chúng.

1. **Quá trình trainning CLIP**

* **Bước 1: Contrastive Pre-training:** CLIP được huấn luyện trước trên tập dữ liệu quy mô lớn. Trong quá trình huấn luyện trước, mô hình được cung cấp các cặp hình ảnh và chú thích văn bản. Một số cặp là khớp chính xác (chú thích mô tả đúng hình ảnh), trong khi các cặp khác không khớp. Mô hình tạo ra các nhúng trong không gian đặc trưng chung.
* **Bước 2:** **Tạo bộ phân loại dữ liệu từ văn bản nhãn**

Với mỗi hình ảnh, nhiều mô tả văn bản được tạo ra, bao gồm một mô tả đúng và một số mô tả sai. Điều này tạo ra tập hợp các cặp mẫu: **Mẫu tích cực** (cặp hình ảnh-văn bản khớp), **Mẫu tiêu cực** (cặp hình ảnh-văn bản không khớp). Các mô tả văn bản được đưa vào bộ mã hóa văn bản, tạo ra các nhúng đặc trưng cho từng lớp**.** Trong giai đoạn này Contrastive Loss Function đóng một vai trò quan trọng: Nếu mô hình gán độ tương đồng cao cho một cặp sai(negative pair) hàm mất mát sẽ tạo ra **loss value** lớn. Giá trị loss value lớn này được sử dụng trong quá trình tối ưu hóa (qua gradient descent và backpropagation) để điều chỉnh trọng số của mô hình, làm cho mô hình "học" rằng cặp sai không nên có độ tương đồng cao.

-> đẩy các vector của cặp sai (negative pairs) ra xa nhau trong không gian đặc trưng làm cho các biểu diễn của hình ảnh và văn bản không liên quan trở nên khác biệt. Nếu các vector biểu diễn của cặp đúng (hình ảnh và văn bản khớp nhau) có **độ tương đồng cao** trong không gian đặc biệt chung, hàm mất mát sẽ tối ưu hóa để giữ các vector của cặp đúng gần nhau, giảm giá trị lost value. -> đảm bảo rằng các biểu diễn của cặp đúng (positive pairs) được giữ gần nhau trong không gian tiềm ẩn, tạo ra một ánh xạ chính xác giữa hình ảnh và văn bản.

**Vd**: Nếu mô hình nhận diện rằng hình ảnh một con mèo và chú thích "một con mèo" có độ tương đồng cao, Contrastive Loss Function sẽ tạo ra một loss value thấp, "thưởng" mô hình bằng cách củng cố hành vi này. Mô hình sẽ tiếp tục tối ưu hóa để giữ các vector của cặp này gần nhau trong không gian đặc biệt chung.

* **Bước 3**: **Zero-shot Prediction**

Bây giờ, bộ mã hóa văn bản đã được huấn luyện được sử dụng như một bộ phân loại không cần huấn luyện trước (zero-shot classifier). Với một hình ảnh mới, CLIP có thể thực hiện dự đoán zero-shot.

Quá trình này được thực hiện bằng cách đưa hình ảnh qua bộ mã hóa hình ảnh và sử dụng bộ phân loại dữ liệu mà không cần tinh chỉnh (fine-tuning).

CLIP tính toán **độ tương đồng cosine** giữa các nhúng (embeddings) của tất cả các cặp hình ảnh và mô tả văn bản.

Mô hình tối ưu hóa các tham số của các bộ mã hóa để:

+Tăng độ tương đồng của các cặp đúng (correct pairs).

+Giảm độ tương đồng của các cặp sai (incorrect pairs).

=> hình ảnh và văn bản có liên quan về mặt ngữ nghĩa được ánh xạ gần nhau.

Lớp được dự đoán là lớp có logit value cao nhất.