- Deep Canonical Correlation Analysis (DCCA) là một phương pháp học các biến đổi phi tuyến tính phức tạp của hai quan điểm dữ liệu sao cho các biểu diễn kết quả có mối tương quan tuyến tính cao. Các tham số của cả hai biến đổi được học chung để tối đa hóa tổng tương quan (được điều chỉnh). DCCA có thể được xem như một mở rộng phi tuyến tính của phương pháp tuyến tính phân tích tương quan cổ điển (CCA). Đây là một phương pháp thay thế cho phương pháp phi tham số phân tích tương quan cổ điển (KCCA) để học các biến đổi phi tuyến tính có mối tương quan. Khác với KCCA, DCCA không yêu cầu một tích vô hướng, và có những ưu điểm của một phương pháp tham số: thời gian huấn luyện tăng tốt với kích thước dữ liệu và dữ liệu huấn luyện không cần được tham chiếu khi tính toán các biểu diễn của các trường hợp chưa nhìn thấy. Trong các thí nghiệm trên hai bộ dữ liệu thực tế, chúng tôi thấy rằng DCCA học các biểu diễn có mối tương quan đáng kể cao hơn so với những biểu diễn được học bởi CCA và KCCA. Chúng tôi cũng giới thiệu một hàm sigmoid không bão hòa mới dựa trên căn bậc ba có thể hữu ích nhiều hơn trong các mạng nơ-ron feedforward.

- Để giải quyết bài toán DCCA, chúng ta cần tối ưu hóa hàm mất mát tương quan giữa các biểu diễn từ hai mạng neural network độc lập. Hàm mất mát này thường bao gồm việc tối thiểu hóa độ lệch tương quan giữa các biểu diễn của hai tập dữ liệu, đồng thời cũng đảm bảo rằng các biểu diễn tối ưu hóa sẽ giữ lại các thông tin quan trọng và có tác động lớn đến cả hai tập dữ liệu để tìm ra các tham số tối ưu của mô hình. Quá trình này bao gồm việc huấn luyện các mạng sâu để học các biến đổi phi tuyến tính của hai quan điểm dữ liệu sao cho các biểu diễn cuối cùng có mối tương quan cao nhất. Để đạt được điều này, chúng ta sử dụng các phương pháp như lan truyền ngược (backpropagation) để tính gradient của hàm mục tiêu và cập nhật các tham số của mô hình. Việc tinh chỉnh siêu tham số (hyperparameters) như số lớp, kích thước đầu ra, độ rộng của các lớp ẩn, và các tham số điều chuẩn là quan trọng để đạt được hiệu suất tốt nhất của mô hình DCCA trên dữ liệu huấn luyện và dữ liệu kiểm tra.

- Bài toán DCCA có thể được mô tả như sau:

+ Đầu vào: Hai điểm dữ liệu X1 và X2.

+ Đầu ra: Các biểu diễn phi tuyến tính của X1 và X2 sao cho chúng có mối tương quan tuyến tính cao nhất.

+ Mục tiêu: Tối đa hóa tổng tương quan giữa các biểu diễn của hai quan điểm dữ liệu thông qua việc học các biến đổi phi tuyến tính.

+ Phương pháp: Sử dụng mạng sâu để học các biến đổi phi tuyến tính của hai quan điểm dữ liệu và tối ưu hóa các tham số của mô hình để đạt được mối tương quan cao nhất trên dữ liệu kiểm tra.

\*\*Công thức\*\*

Cho hai tập hợp biến số, X ∈Rn\*dx và Y ∈Rn\*dy, trong đó n là số mẫu và dX và dY là chiều của hai tập hợp biến số, tương ứng. Mục tiêu của Deep CCA là học hai mạng neural, fx(.) và fy(.), để ánh xạ X và Y đến một không gian ẩn chia sẻ, nơi mối tương quan giữa hai tập hợp biến số được tối đa hóa.

Hàm mục tiêu của Deep CCA có thể được viết dưới dạng:

Maxfx,fy corr(fx(X), fy(Y))

trong đó corr(a, b) là một metric tương quan, chẳng hạn như hệ số tương quan Pearson hoặc thông tin chung.

\*\*Hàm mất mát\*\*

Hàm mất mát của Deep CCA thường được định nghĩa là:

L = -corr(fx(X), fy(Y)) + lambda\_1 \* ||fx||2 + lambda\_2 \* ||fy||2

trong đó lambda\_1 và lambda\_2 là hệ số điều chỉnh độ phức tạp của mạng neural.

\*\*Tối ưu hóa\*\*

Vấn đề tối ưu hóa trong Deep CCA có thể được giải quyết bằng phương pháp gradient descent ngẫu nhiên (SGD) hoặc các biến thể của nó. Gradient của hàm mất mát đối với tham số mô hình có thể được tính toán bằng phương pháp backpropagation.

∂L/∂θ\_f = - ∂corr/∂F \\* ∂F/∂θ\_f + λ₁ \\* ∂reg\_F/∂θ\_f

∂L/∂θ\_g = - ∂corr/∂G \\* ∂G/∂θ\_g + λ₂ \\* ∂reg\_G/∂θ\_g

Trong đó, θ\_f và θ\_g lần lượt là các tham số mô hình cho f và g.

\*\*Kiến trúc\*\*

Kiến trúc của Deep CCA thường bao gồm hai mạng neural, fx() và fy(), mỗi mạng bao gồm nhiều lớp. Đầu ra của mỗi mạng là một vector trong không gian ẩn chia sẻ. Mối tương quan giữa hai đầu ra được tính toán bằng metric tương quan, chẳng hạn như hệ số tương quan Pearson hoặc thông tin chung.

\*\*Ưu điểm\*\*

Deep CCA có một số ưu điểm so với CCA truyền thống, bao gồm:

1. \*\*Mối tương quan phi tuyến\*\*: Deep CCA có thể bắt được mối tương quan phi tuyến giữa các biến số, điều này không thể được bắt bởi CCA truyền thống.

2. \*\*Độ linh hoạt\*\*: Deep CCA có thể học các mối quan hệ phức tạp giữa các biến số bằng cách sử dụng mạng neural.

3. \*\*Khả năng mở rộng\*\*: Deep CCA có thể xử lý dữ liệu cao chiều và các bộ dữ liệu lớn.

\*\*Ứng dụng\*\*

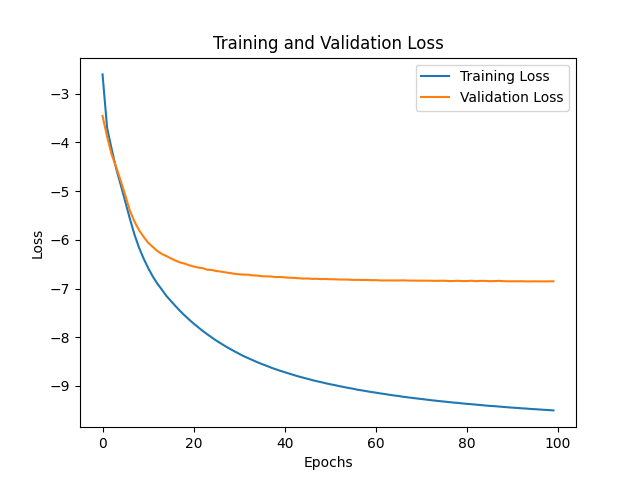
Deep CCA đã được áp dụng vào nhiều lĩnh vực, bao gồm:

1. \*\*Học đa modal\*\*: Deep CCA có thể được sử dụng để học mối tương quan giữa các modal khác nhau, chẳng hạn như hình ảnh và văn bản.

2. \*\*Chuyển giao kiến thức\*\*: Deep CCA có thể được sử dụng để chuyển giao kiến thức từ một miền sang miền khác.

3. \*\*Học đặc trưng\*\*: Deep CCA có thể được sử dụng để học các đặc trưng được tương quan trên các bộ dữ liệu khác nhau.

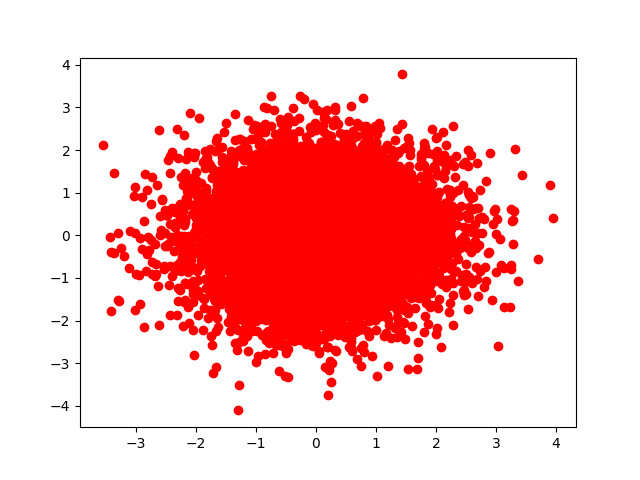
Source code output:



**Training Loss Trend**: Training loss bắt đầu từ khoảng -3 và cho thấy sự suy giảm đột ngột cho đến khoảng 20 epochs, chỉ ra tốc độ học tập nhanh chóng. Sau điểm này, mất mát tiếp tục giảm nhưng với tốc độ chậm hơn, cho thấy hiệu quả học giảm dần khi mô hình bắt đầu hội tụ.

**Validation Loss Plateau**: Validation loss bắt đầu từ khoảng hơn -4 và giảm đều đặn cho đến khi đạt đến plateau quanh 40 epochs. Plateau này cho thấy mô hình không tổng quát hóa tốt ngoài dữ liệu huấn luyện, điều này có thể là dấu hiệu của hiện tượng overfitting.

**Overfitting Concern**: Sự phân kỳ giữa mất mát huấn luyện và mất mát xác thực sau khoảng 40 epochs cho thấy hiệu suất của mô hình trên dữ liệu huấn luyện đang được cải thiện, trong khi hiệu suất trên dữ liệu chưa thấy không được cải thiện. Đây là một dấu hiệu điển hình của overfitting, nơi mô hình học dữ liệu huấn luyện quá tốt, bao gồm cả nhiễu và ngoại lệ, với chi phí khả năng tổng quát hóa của nó.



**Data Distribution**: Các chấm đỏ hình thành một hình elip, cho thấy phân phối chuẩn hai biến. Điều này chỉ ra rằng hai biến thể hiện trên trục x và y có mối quan hệ tuyến tính và có khả năng tương quan.

**Data Range**: Các điểm dữ liệu được trải rộng trên phạm vi từ -4 đến 4 trên cả hai trục. Phạm vi này đối xứng quanh gốc tọa độ, điều này là điển hình cho dữ liệu được chuẩn hóa trong phân tích thống kê.

**Data Density**: Sự tập trung dày đặc của các chấm tại trung tâm đồ thị ngụ ý rằng giá trị trung bình của cả hai biến gần bằng không, và hầu hết các điểm dữ liệu nằm trong một độ lệch chuẩn của giá trị trung bình.

**Outliers**: Sự hiện diện của các điểm rải rác cách xa trung tâm cho thấy các điểm ngoại lệ hoặc giá trị cực đoan không theo khuôn mẫu chung của dữ liệu.