- 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:

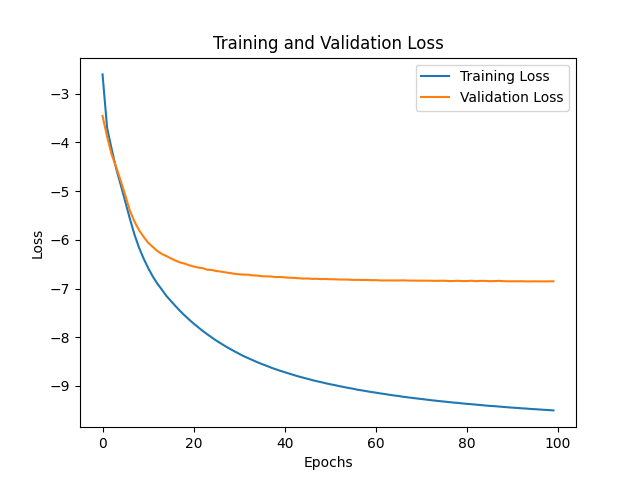
+ Đầu vào: Hai điểm dữ liệu X1 và X2.

+ 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.

Source code output:

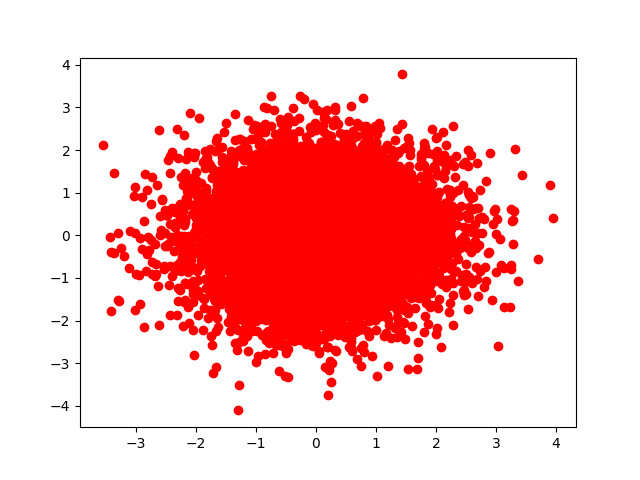


**Training Loss Trend**: The training loss starts at approximately **-3** and shows a sharp decline until around **20 epochs**, indicating rapid learning. After this point, the loss continues to decrease but at a slower rate, suggesting diminishing returns on learning as the model begins to converge.

**Validation Loss Plateau**: The validation loss begins just below **-4** and decreases steadily until it plateaus around **40 epochs**. This plateau suggests that the model is not generalizing well beyond the training data, which could be a sign of overfitting.

**Overfitting Concern**: The divergence between the training and validation loss after around **40 epochs** implies that the model’s performance on the training data is improving, while its performance on unseen data is not. This is a classic indication of overfitting, where the model learns the training data too well, including noise and outliers, at the expense of its ability to generalize.

**Recommendations**: To address potential overfitting, consider implementing techniques such as early stopping, regularization, or using a more complex model if the current one is too simple. Additionally, expanding the dataset or performing data augmentation could help improve the model’s generalization capabilities.



**Data Distribution**: The red dots form an elliptical shape, suggesting a **bivariate normal distribution**. This indicates that the two variables represented on the x and y-axes have a **linear relationship** and are likely **correlated**.

**Data Range**: The data points are spread across a numerical range from **-4 to 4** on both axes. This range is symmetric around the origin, which is typical for standardized data in statistical analysis.

**Data Density**: The dense concentration of dots at the center of the plot implies that the **mean values** for both variables are close to zero, and most data points fall within one standard deviation of the mean.

**Outliers**: The presence of scattered points away from the dense center suggests **outliers** or **extreme values** that do not follow the general pattern of the data.