CCA vs Cross-Attention Transformers

Vladimirov Eduard, group M05-304a

5 октября 2024 г.

Introduction

Finding meaningful relationships between multiple sets of variables is an important problem. Canonical Correlation Analysis (CCA) is a well-established statistical technique that addresses this challenge by identifying linear relationships between two multivariate datasets. On the other hand, cross-attention mechanism, is designed to capture and leverage dependencies between different input sequences. Both CCA and cross-attention are fundamentally concerned with learning projections that reveal underlying dependencies between distinct sets of data. By examining their theoretical connections and differences, we explore how CCA and cross-attention can complement each other in understanding and modeling relationships in high-dimensional spaces.

CCA

Given two sets of vectors $X \in \mathbb{R}^{n_1 \times m}$ and $Y \in \mathbb{R}^{n_2 \times m}$, where m denotes the number of vectors, CCA learns two linear transformations $A \in \mathbb{R}^{n_1 \times r}$ and $B \in \mathbb{R}^{n_2 \times r}$ such that the correlation between $A^T X$ and $B^T Y$ is maximized.

Note the covariances of X and Y as $S_{11} = \frac{1}{m} X X^\intercal \in \mathbb{R}^{n_1 \times n_1}, S_{22} = \frac{1}{m} Y Y^\intercal \in \mathbb{R}^{n_2 \times n_2}$, and the cross-covariance of X, Y as $S_{12} = \frac{1}{m} X Y^\intercal \in \mathbb{R}^{n_1 \times n_2}$. The CCA objective is

$$\begin{aligned} A^*, B^* &= \arg\max_{A,B} \operatorname{corr}(A^T X, B^T Y) \\ &= \arg\max_{A,B} \frac{A^T S_{12} B}{\sqrt{A^T S_{11} A \cdot B^T S_{22} B}} \end{aligned} \tag{1}$$

The solution of the above equation is fixed and can be solved in multiple ways. One method suggested by (Martin and Maes 1979) lets U, S, V^T be the Singular Value Decomposition (SVD) of the matrix $Z = S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}}$. Then A^*, B^* and the total maximum canonical correlation are

$$\begin{aligned} A^* &= S_{11}^{-\frac{1}{2}} U = \left(\frac{1}{m} X X^T\right)^{-\frac{1}{2}} U \\ B^* &= S_{22}^{-\frac{1}{2}} V = \left(\frac{1}{m} Y Y^T\right)^{-\frac{1}{2}} V \\ &\text{corr}(A^{*T} X, B^{*T} Y) = \text{trace}(Z^T Z)^{\frac{1}{2}}. \end{aligned} \tag{2}$$

One limitation of CCA is that it only considers linear transformations, which limits its expressive power. In contrast, deep learning models employ non-linear mappings such as self-attention and cross-attention to learn more complex representations.

Self-attention and cross-attention

Attention mechanisms are used to determine the relevance of different parts of the input data.

The self-attention mechanism is defined as follows:

$$\begin{aligned} &\text{attn} : \mathbb{R}^{m \times d} \times \mathbb{R}^{m \times d} \times \mathbb{R}^{m \times d} \longrightarrow \mathbb{R}^{m \times d} \\ &\text{attn}(Q, K, V) = \varphi\left(\frac{Q K^\top}{\sqrt{d}}\right) V \end{aligned} \tag{3}$$

where $Q, K, V \in \mathbb{R}^{m \times d}$ represent the queries, keys, and values, respectively, and $\varphi : \mathbb{R}^{m \times m} \longrightarrow \mathbb{R}^{m \times m}$ is row-wise applied nonlinear function, usually softmax. The dot product between Q and K determines the attention weights, which are normalized using the softmax function. The result is then applied to the values V to generate the output.

Self-attention applied to the input $X \in \mathbb{R}^{m \times n_1}$ is computed as:

$$\begin{aligned} &\text{self-attn} : \mathbb{R}^{m \times n_1} \longrightarrow \mathbb{R}^{m \times d} \\ &\text{self-attn}(X) = \text{attn}(X W_q, X W_k, X W_v) \end{aligned} \tag{4}$$

where $W_q, W_k, W_v \in \mathbb{R}^{n_1 \times d}$ — parameter matrices

In multihead attention, several attention heads are used in parallel, where each head computes its own attention weights and outputs. The

outputs are then concatenated and linearly transformed by a weight matrix $W^Q \in \mathbb{R}^{p \cdot d \times d}$:

$$\text{multihead-attn}(Q, K, V) = [\text{head}_1, \ldots, \text{head}_p]W^Q, \tag{5}$$

where $\text{head}_i = \text{self-attn}(X)$

Cross-attention, in contrast, involves attention between two different sets of inputs. It computes attention by using one set of inputs for queries $X_1 \in \mathbb{R}^{m \times d_1}$ and another set for keys and values $X_2 \in \mathbb{R}^{m \times d_2}$:

$$\text{cross-attn}(X_1, X_2) = \text{attn}(X_1 W_q, X_2 W_k, X_2 W_v) \tag{6}$$

CCA and attention

Both CCA and attention mechanisms aim to find relationships between two sets of data. However, they differ significantly in their approach and applications:

Aspect	**Attention**	**Canonical Correlation Analysis (CCA)**
Goal	Identify relevant parts of input sequences	Receive embeddings in the same hidden space + dimensionality reduction
Similarity Measure	$A = \frac{1}{\sqrt{d}} QK^\mathsf{T}$ – attention matrix	$\text{tr}(A^\mathsf{T} S_{12} B)$, s.t. $A^\mathsf{T} S_{11} A = B^\mathsf{T} S_{22} B = I$
Optimization Goal	Minimize task-specific loss	$\max_{A,B} \text{corr}(A^T X, B^T Y)$

Таблица 1: Comparison of Attention Mechanisms and CCA

Note that $A^\mathsf{T} S_{12} B = \frac{1}{m} A^\mathsf{T} X Y^\mathsf{T} B = \frac{1}{m} A^\mathsf{T} X \left(B^\mathsf{T} Y\right)^\mathsf{T} = \frac{1}{m} \widehat{Q} \widehat{K}^\mathsf{T}$. And it's quite similar to attention matrix formula $A = \dfrac{1}{\sqrt{d}} QK^\mathsf{T}$. Especially, in cross attention case, where Q is a linear transformation of X_1 and K is a linear transformation of X_2:

Attn	**Self-attn**	**Cross-attn**	**CCA**	**CCA-X**
Q	$W_Q^\mathsf{T} X$	$W_Q^\mathsf{T} X$	$A^\mathsf{T} X$	$S_{11}^{-\frac{1}{2}} X$
K	$W_K^\mathsf{T} X$	$W_K^\mathsf{T} Y$	$B^\mathsf{T} Y$	$S_{22}^{-\frac{1}{2}} Y$
V	$W_V^\mathsf{T} X$	$W_V^\mathsf{T} Y$	I	$S_{11}^{-\frac{1}{2}} X$
φ	softmax	softmax	Id	SVD_U

Таблица 2: United notation of CCA and attention

Let's view in detail the CCA projection of X to latent space:

$$\begin{aligned} \mathrm{CCA}_{XY}(X) &= U^\mathsf{T} S_{11}^{-\frac{1}{2}} X = U^\mathsf{T} X_1 \\ \mathrm{CCA}_{XY}(Y) &= V^\mathsf{T} S_{22}^{-\frac{1}{2}} Y = V^\mathsf{T} Y_1 \\ Z &= S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}} = \frac{1}{m} X_1 Y_1^\mathsf{T} \end{aligned} \tag{7}$$

Conclusion

CCA and attention mechanism are mechanisms for finding relationships between two datasets. They both exploit cross-correlation matrix $X^\mathsf{T} Y$. However, while CCA focuses on linear transformations for maximizing correlations, attention mechanism uses non-linear mappings and is optimized using task-specific loss functions.