Cross-Modal Retrieval with Correspondence Autoencoders

Abstract:

Cross-modal retrieval is a challenging task that involves retrieving data from one modality based on a query from another modality. This paper introduces a novel approach, namely Correspondence Autoencoders (CAE), for addressing cross-modal retrieval tasks. The proposed method aims to learn effective joint representations that capture the inherent relationships between different modalities, thereby facilitating accurate retrieval.

Introduction:

Cross-modal retrieval is of great significance in various applications such as multimedia information retrieval, image-text matching, and more. Traditional methods often suffer from the modality gap and the difficulty of finding direct correspondences between different types of data. In this paper, we present the Correspondence Autoencoders (CAE) framework, which leverages the power of autoencoders to learn latent representations that encode both modalities' information while emphasizing their shared semantics.

Correspondence Autoencoders (CAE):

The CAE framework is designed to learn joint representations for cross-modal retrieval. It consists of paired autoencoders, each designed to map data from one modality to a common latent space. During training, the autoencoders are constrained to reconstruct their respective input data while also aiming to reconstruct the data from the opposite modality. This encourages the autoencoders to capture the correspondence between modalities.

Learning Objectives:

The CAE framework employs a combination of reconstruction loss and correspondence loss. The reconstruction loss ensures that the autoencoders effectively reconstruct their own modality's data. The correspondence loss

enforces the autoencoders to reconstruct the data from the other modality, thereby aligning the representations.

Experiments and Results:

To evaluate the effectiveness of the proposed CAE framework, extensive experiments were conducted on benchmark datasets. Cross-modal retrieval performance metrics such as precision, recall, and mean average precision were used to assess the framework's efficacy. Comparative analyses were performed against state-of-the-art cross-modal retrieval methods, demonstrating the superiority of CAE in capturing cross-modal relationships.

Conclusion:

In this paper, we introduced Correspondence Autoencoders (CAE) for addressing the cross-modal retrieval challenge. The CAE framework aims to learn joint representations by exploiting correspondence relationships between different modalities. Experimental results validate the effectiveness of CAE in achieving accurate cross-modal retrieval, indicating its potential for various applications in multimedia analysis and beyond.