

Cross-modal Variational Alignment of Latent Spaces

Thomas Theodoridis Theocharis Chatzis Vassilios Solachidis Kosmas Dimitropoulos
Petros Daras

Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

Ftomastheod, hatzis, vsol, dimi trop, darasG@i ti . gr

Abstract

*In this paper, we propose a novel cross-modal variational alignment method in order to process and relate information across different modalities. The proposed approach consists of two variational autoencoder (VAE) networks which generate and model the latent space of each modality. The first network is a **multi-modal variational autoencoder** that maps directly one modality to the other, while the second one is a **single-modal variational autoencoder**. In order to associate the two spaces, we apply **variational alignment**, which acts as a **translation mechanism** that projects the latent space of the first VAE onto the one of the single-modal VAE through an intermediate distribution. Experimental results on four well-known datasets, covering two different application domains (food image analysis and 3D hand pose estimation), show the generality of the proposed method and its superiority against a number of state-of-the-art approaches.*

1. Introduction

Cross-modal learning has attracted increasing attention recently due to the rapid growth of multi-modal data (image, video, text, audio, depth, IR etc) and the need for enhanced learning either by leveraging information from one data modality to accomplish a given task in another, or through the synergistic synthesis of information from multiple modalities. Because of their general nature, they have been extensively used in the literature for various problems, such as audio retrieval from text [20], text-to-image and image-to-text retrieval [24], sentiment analysis from video, audio and text sources [22], synchronization among different representations of music, like sheet music and audio recordings [15], recipe (ingredients and instructions) retrieval from images and vice versa [27] and 3D hand pose estimation from images [30]. Recent cross-modal frameworks involve neural networks as encoder and decoder mechanisms in order to transition from one modality to another. Based on the way these frameworks model

the cross-modal objective, they are categorized as discriminative and generative. Approaches that fall into the first category model the probability of an outcome conditioned on the given observation. Generative approaches, on the other hand, model the underlying distribution of the observed variables, thus obtaining valuable information regarding their origin.

Most recent approaches have adopted deep generative models, such as VAEs, GANs or a combination of them, to encode cross-modal data into a shared latent space [30, 34]. However, the main problem in these approaches is the fact that each modality has completely different characteristics from the others and, as a result, it is difficult to efficiently model the heterogeneous modalities (like image, speech or text) into a shared latent space. To address the problem of **learning meaningful mappings among embedding spaces**, we propose a novel variational alignment framework of latent spaces, which performs the mapping of the latent space of one modality onto the one of another modality. More specifically, in this paper we present a cross-modal learning approach consisting of a number of variational autoencoder networks that aim to generate and model the latent space corresponding to each modality and, at the same time, align the different spaces through the modeling of an intermediate latent space, generated by an additional variational autoencoder network. The main contributions of this paper are summarized as follows:

- We introduce a generic cross modal deep learning approach using variational autoencoder networks in order to model the latent spaces of different modalities as probability distributions. More specifically, we propose the use of a pair of multimodal (M_1 -to- M_2) and single-modal (M_2 -to- M_2) variational autoencoders with aligned latent spaces, where the aligned latent space of the first modality can be directly used by the decoder of the single-modal VAE network, outperforming the state-of-the-art in different application domains.
- We propose a novel cross-modal variational alignment

1

2

3
