Characterizing Labels with CCVAE: Latent Representations and Interventions

Ambre Etienne - Mahdi Ranjbar

Probabilistic Graphical Model

2023

Introduction

The quest to comprehend and manipulate the characteristic factors of perceptual observations has long been central to machine intelligence. Specifically, this involves learning important factors from data that represent characteristics we can grasp.

Variational autoencoders (VAEs) offer a versatile framework for meaningful representation learning, often leveraging labels. However, conventional approaches directly associating latent variables with labels can limit manipulation tasks. In response, we propose the Characteristic Capturing VAE (CCVAE) [2], explicitly capturing label characteristics in the latent space, offering enhanced manipulation capabilities while maintaining predictive accuracy.

VAEs

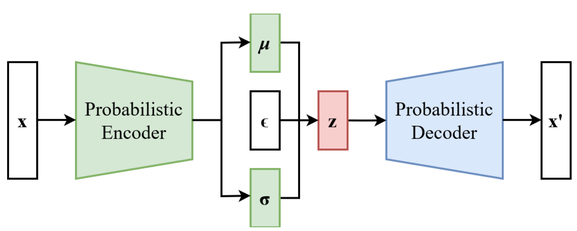

Figure 1:VAE

We examine VAEs as they combine deep autoencoder unsupervised learning with generative latent-variable models 2, portraying data as distributions. VAEs utilize neural networks for encoding and decoding, employing variational inference to learn representations and build an approximate posterior. The goal is to maximize the marginal likelihood using the evidence lower bound (ELBO).

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(z,x)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log(\frac{p_\theta(z,x)}{q_\phi(z|x)})\right] \equiv \mathcal{L}(x;\phi,\theta)$$

SSVAEs

Semi-supervised Variational Autoencoders (SSVAEs) address scenarios where a subset of data has corresponding labels. They decompose the log-marginal likelihood into the sum of labeled and unlabeled data log-likelihoods, each bounded by their respective Evidence Lower Bounds (ELBOs).

SSVAEs commonly split the latent space into z_y (representing labeled information) and $z_{\setminus y}$ (for unlabeled data), and then directly fix $z_y = y$ whenever the label is provided. The focus here is on using label information to structure the latent space, aiming to encapsulate and disentangle characteristics associated with the labels rather than solely improving generation fidelity.

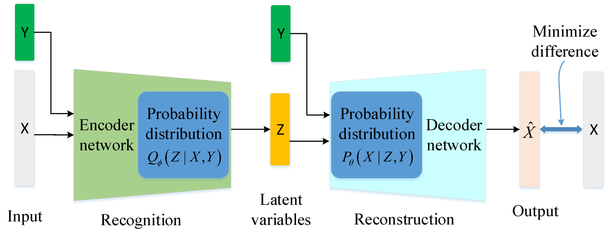

Figure 2:Conditional VAE

CCVAEs

Assuming that labels directly correspond to a part of the latent space (z_y) can cause issues in capturing rich label-related characteristics and manipulating data. For example, difficulty manipulating a characteristic without modifying the labels.

The article from Tom Joy [2] proposes Characteristic-Capturing Variational Autoencoders (CCVAEs) to address issues in treating labels as direct components of the latent space in VAEs. It suggests conditioning latent variables on labels to capture characteristics effectively. CCVAEs split the latent space into characteristic (z_c) and non-characteristic ($z_{\setminus c}$) components.

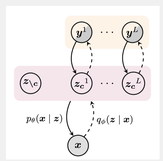

Figure 3:CCVAE graphical model

However, the characteristics of different labels become entangled within z_c. To disentangle it, we partition the latent space, such that the classification of particular labels y_i only has access to particular latents z_c^i and thus $\log q_\varphi(y|z_c) = \sum_i \log q_{\varphi_i}(y^i|z_c^i)$. This forces the characteristic information needed to classify y^i to be stored only in the corresponding z_c^i. Additionally, we can introduce a factorized set of generative models $p(z_c|y) = \prod_i p(z_c^i|y^i)$, enabling easy generation and manipulation of z_c^i individually as shown in Figure 3.

Interventions - multi-label context

To demonstrate the qualitative aspects of representation disentanglement, we often assess it through independent exploration of latent dimensions, commonly known as latent traversals. Our study highlights the efficacy of CCVAE in interventions, showcasing its capacity to selectively isolate label-specific characteristics and provide precise control over interventions.

Our evaluation is conducted in a multi-label context using the CelebA dataset, in which we extract a detailed description of 18 distinct labels. The encoder and decoder architectures are adapted from Higgins et al. (2016), with specific modifications to accommodate label-predictive distribution and conditional prior.

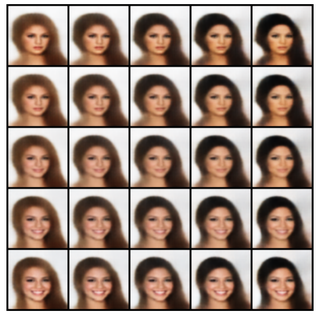

Figure 4:CCVAE latent walk Smiling and Black Hair

In Figure 4, we observe how CCVAE allows the manipulation of one characteristic at a time, transitioning from a neutral expression to a smile and changing hair color from blond to brown.

Interventions - multi-class context

CCVAE demonstrates its versatility not only in addressing multi-label challenges but also in effectively handling multi-class problems. In this context, we present results obtained from the multi-class setting using FashionMNIST dataset, even though this aspect is somewhat tangential to our primary focus.

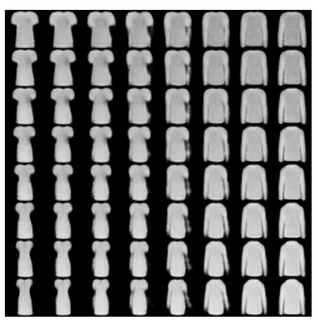

Figure 5:CCVAE latent traversals for FashionMNIST

In 5, we observe CCVAE's ability to transition from a t-shirt to a dress by elongating the length.

Extended work

We extended our work to the the domain of NLP and considered to study a deep generative model for unsupervised text style transfer [1]. This model uses a probabilistic approach, hypothesizing a latent sequence to transform sequences between domains without supervision.

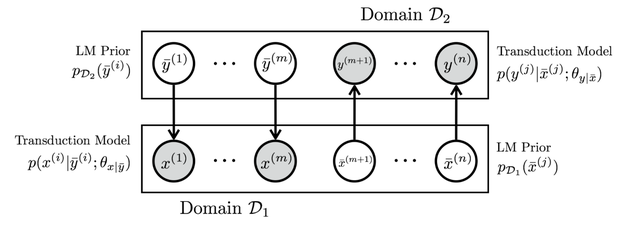

Figure 6:Graphical Model for Style Transfer

Original	Style Transfer Model
The service was incredibly slow, and the staff was unhelpful.	The service was unhurried, and the staff was unreservedly helpful.
This product is a complete waste of money, don't buy it.	This product is worth every penny, definitely buy it.
The customer support is terrible, never responds on time.	The customer support is exceptional, always responsive.
The book was poorly written and lacked any depth.	The book was well-written and full of depth.

Table 1:Inference examples: negative to positive Sentiment Transfer

References

[1] Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*, 2020.

[2] Tom Joy, Sebastian M. Schmon, Philip H. S. Torr, N. Siddharth, and Tom Rainforth. Capturing label characteristics in vaes, 2022.