Generative Models - DRE7053
Lecture 3

NORA Summer School 2024

Rogelio A Mancisidor
Assistant Professor
Department of Data Science and Analytics
BI Norwegian Business School

June 10-14, 2024

Outline

Outline

Unsupervised Clustering

- The main idea in [12] is to induce a flexible latent space in an unsupervised setting
- This can be achieved in different ways, for example

$$p(z) = \sum_c p(c)p(z|c),$$

where $p(c) \sim \text{Cat}(\pi_0)$, and $p(z|c)$ is a conditional prior distribution.

- The marginal distribution $p(z)$ is a mixture model!

Generative and Inference Models

- The generative model is factorized as $p(\boldsymbol{x}, \boldsymbol{z}, c) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}|c)p(c)$, with distributions

$$p(c) \sim \text{Cat}(\pi_0) \tag{1}$$
$$p(\boldsymbol{z}|c) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2) \tag{2}$$
$$p(\boldsymbol{x}|\boldsymbol{z}) \sim f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{3}$$

- The inference model is factorized as $q(\boldsymbol{z}, c) = q(\boldsymbol{z}|c)q(c|\boldsymbol{x})$, where

$$q(\boldsymbol{z}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \tag{4}$$
$$q(c|\boldsymbol{x}) \sim \text{Cat}(\boldsymbol{\pi}) \tag{5}$$

Deriving the ELBO

Qualitative Results

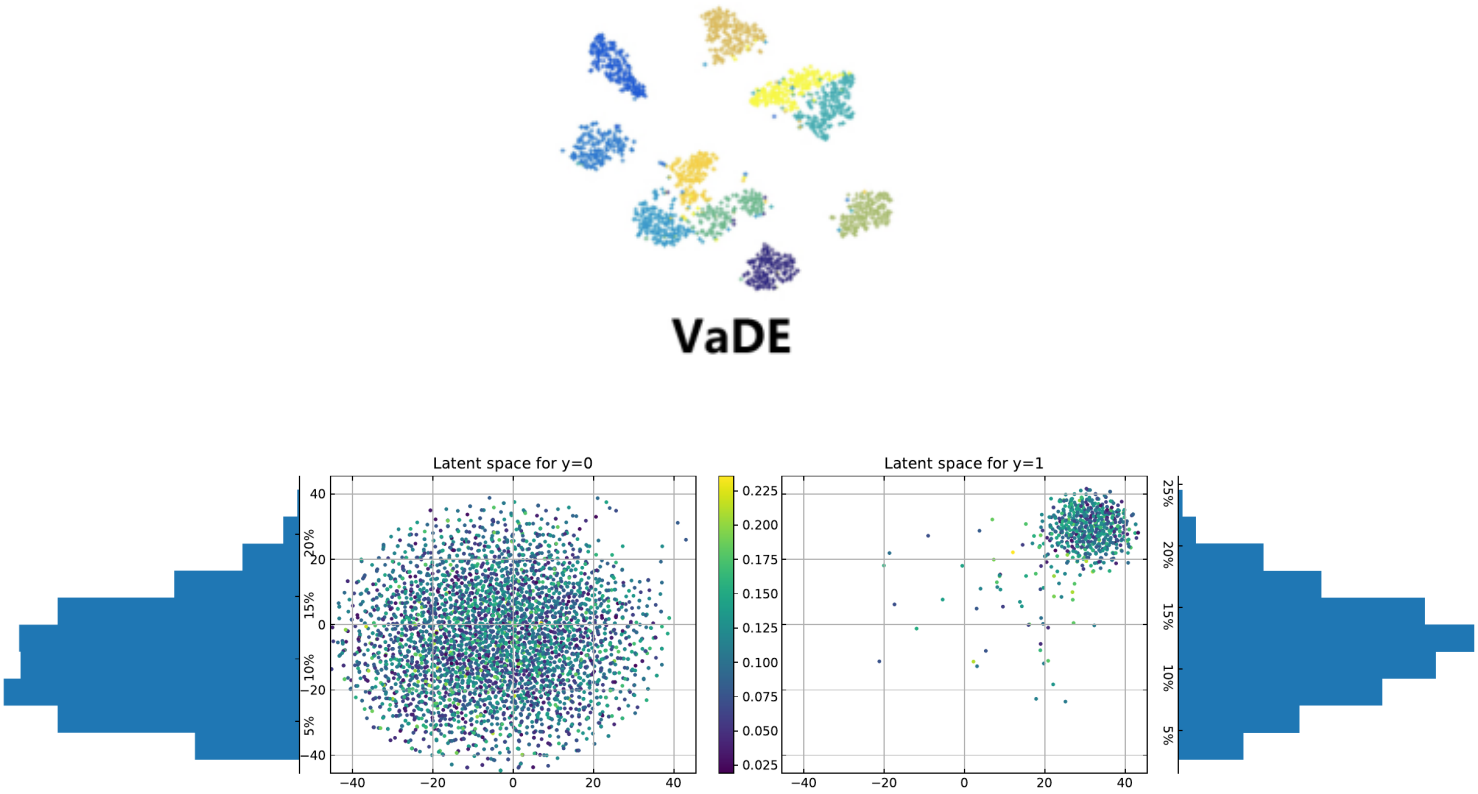


Figure: *Deep generative models for reject inference in credit scoring*

Things to keep in mind

- Why do we model the prior distribution as a GMM and not the posterior?
- Note the effect of the KL divergence term in the ELBO
- Unsupervised generative clustering is difficult to train!
- VaDE initialize the values for $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c^2$ using the expectation-maximization algorithm for GMMs, otherwise the ELBO cannot be optimized.

Outline

What is Multimodal Learning?

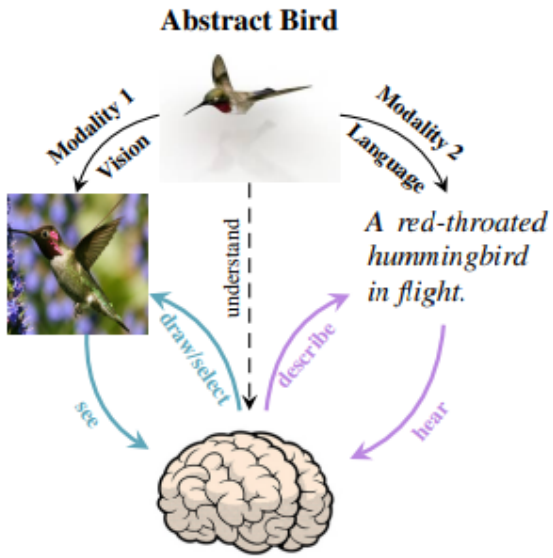


Figure: Borrowed from Shi et al. 2019

Multimodal Learning

- Multimodal learning (MML) uses different measurement modalities of the same object to learn *shared representations*.

$$f : (x_1, x_2) \rightarrow z$$

- Two major approaches:
  1. Joint representation: projects all modalities into a common space.
  2. Coordinated representation: assumes that each representation exists in their own space.

Canonical Correlation Analysis

- First coordinated representation approach introduced by Hotelling (1936):
- Canonical Correlation Analysis seeks the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in

$$\boldsymbol{U} = \boldsymbol{a}^T \boldsymbol{x}_1$$
$$\boldsymbol{V} = \boldsymbol{b}^T \boldsymbol{x}_2$$

such that $corr(\boldsymbol{U}, \boldsymbol{V})$ is maximized. That is,

$$\arg \max_{\boldsymbol{a}, \boldsymbol{b}} corr(\boldsymbol{U}, \boldsymbol{V})$$

- Note: $\boldsymbol{V}$ and $\boldsymbol{U}$ are shared *(linear)* representations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Learning Setting in Multimodal VAEs

- Fusion setting: data from all modalities are available at all phases.
- Cross modality learning: data from multiple modalities is available only during feature learning; during the testing phase, a(some) modality(ies) is(are) missing.
- The goal is to approximate the joint posterior

$$p(\boldsymbol{z}|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$$

and to maximize the marginal likelihood of all modalities

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$$

Theoretical Framework

- We assume M modalities $\mathbb{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M)$ that are conditionally independent given the latent variable $\boldsymbol{z}$
- Therefore, the generative model becomes

$$\begin{aligned} p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M, \boldsymbol{z}) &= p(\boldsymbol{z}) p(\boldsymbol{x}_1|\boldsymbol{z}) p(\boldsymbol{x}_2|\boldsymbol{z}) \cdots p(\boldsymbol{x}_M|\boldsymbol{z}) \\ p(\mathbb{X}, \boldsymbol{z}) &\equiv p(\boldsymbol{z}) p(\mathbb{X}|\boldsymbol{z}) \end{aligned} \tag{6}$$

- Given M modalities, there are 2^M subsets in the powerset $\mathcal{P}(\mathbb{X})$, where only $K = 2^M - 1$ includes any of the modalities
- We need to consider all K subsets for cross-modality generation in any direction (the prior takes care of the empty set)

The Multimodal ELBO

- Assume that we observe 2 modalities, i.e $\mathbb{X} = (\boldsymbol{x}1, \boldsymbol{x}2)$
- We need to minimize

$$\sum_k KL[q(\boldsymbol{z}|\mathbb{X}_k)||p(\boldsymbol{z}|\mathbb{X})] \tag{7}$$

for all $\mathbb{X}_k \in \mathcal{P}(\mathbb{X})$ (excluding the empty set)

Joint Posterior Distributions

- Note the joint posterior distribution $q(\boldsymbol{z}|\boldsymbol{x}_1, \boldsymbol{x}_2)$
- We need to estimate $K - M$ joint posterior distributions, which are called *consensus distributions*
- If we observe, say, 5 modalities, we need to estimate $31 - 5 = 26$ consensus distributions
- Scalable multimodal learning deals with efficient methods that
  - scale linearly on M
  - estimate consensus distributions based on unimodal distributions, which are called *expert distributions*
- Mixture of Experts (MoE) and Product of Experts (PoE) are the common choices to approximate the joint posterior distributions

Product of Experts

- Product of Experts (PoE) simply multiply expert distributions

$$q(\boldsymbol{z}|x_1, \boldsymbol{x}_2) = \frac{1}{\mathcal{Z}} q(\boldsymbol{z}|\boldsymbol{x}_1) q(\boldsymbol{z}|\boldsymbol{x}_2) \tag{8}$$

- If experts are Gaussian (as they always are in VAEs):

$$\boldsymbol{\mu} = (\sum_{i}^{M} \mu_i \tau_i)(\sum_{i}^{M} \tau_i)^{-1}$$

$$\boldsymbol{\sigma}^2 = (\sum_{i}^{M} \tau_i)^{-1}$$

where $\tau = 1/\sigma^2$ is the precision

Mixture of Experts

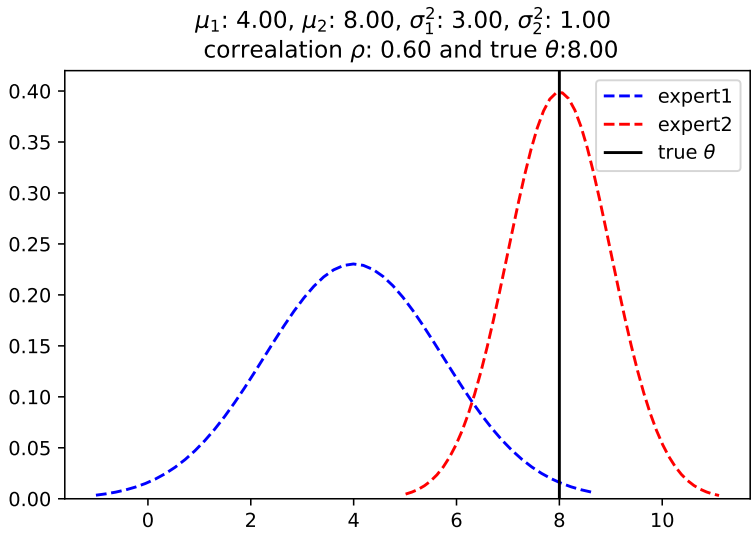
- MoE simply estimates consensus distributions as mixture models

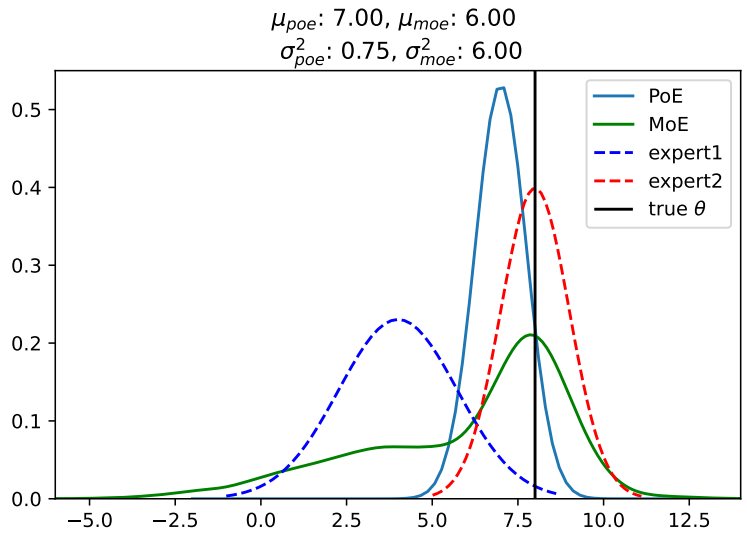
$$q(\boldsymbol{z}|x_1, \boldsymbol{x}_2) = \frac{1}{2}q(\boldsymbol{z}|\boldsymbol{x}_1) + \frac{1}{2}q(\boldsymbol{z}|\boldsymbol{x}_2) \tag{9}$$

- If experts are Gaussian (as they always are in VAEs):

$$\begin{aligned}\boldsymbol{\mu} &= \frac{1}{2}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2 \\ \boldsymbol{\sigma}^2 &= \frac{1}{2}\boldsymbol{\sigma}_1^2 + \frac{1}{2}\boldsymbol{\sigma}_2^2 + \frac{1}{2}\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^2\end{aligned}$$

MoE vs PoE





- PoE has always smaller variance than any of the experts
- MoE is never "sharper" than any of the experts

Consensus of Dependent Experts

- Both PoE and MoE assume independence between expert distributions
- Consensus of Dependent Experts (CoDE) accounts for the dependency between experts!
- For $\rho = 0$, CoDE recovers the PoE parameters!
- Therefore, CoDE is a more general method

CoDE Intuition

- The main idea is to think about estimates from expert distributions as observations from a likelihood function
- We assume a (non-informative) prior distribution and derive the posterior distribution in principled Bayesian manner
- The posterior is the consensus distribution that we are interested in!

CoDE

Definition

All M distributions $q(\boldsymbol{z}|\overline{\overline{\mathbb{X}}}_k = 1)$, where $\overline{\overline{\mathbb{X}}}_k$ denotes the cardinality of $\mathbb{X}_k$, are considered expert distributions, estimating the remaining unknown distributions $q(\boldsymbol{z}|\overline{\overline{\mathbb{X}}}_k > 1)$, which are called consensus distributions.

CoDE

Definition

Let $\boldsymbol{\theta}^k = (\theta_1^k, \theta_2^k, \cdots, \theta_D^k)^T$ denote the latent variable $\boldsymbol{z} \in \mathbb{R}^D$ for the k-th consensus distribution $q(\boldsymbol{z}|\mathbb{X}_k)$. Each j-th expert distribution provides a point estimate μ_j^d on the d-th dimension of θ_d^k, and the uncertainty about the estimation is expressed in the parameter σ_j^d of each expert distribution.

Definition

The error of estimation of the j-th expert in the d-th dimension is $e_j^d = \mu_j^d - \theta_d^k$ and the error of estimation of all experts in the d-th dimension is $\boldsymbol{e}_d = (e_1^d, e_2^d, \cdots, e_{M'}^d)^T$, where M' is the number of experts who evaluate the subset $\mathbb{X}_k$. The overall error of estimation on $\boldsymbol{\theta}^k$ is $\boldsymbol{e}^k = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_D)^T$ with distribution $\boldsymbol{e}^k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^k)$ where $\boldsymbol{\Sigma}^k$

$$\boldsymbol{\Sigma}^k = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_D \end{bmatrix} \text{ where } \boldsymbol{\Sigma}_d = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,M'} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,M'} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M',1} & \sigma_{M',2} & \cdots & \sigma_{M'}^2 \end{bmatrix}$$

and $\mathbf{0}$ is a $[M' \times M']$ matrix.

Lemma 2

Assuming a flat prior distribution on the parameter $\boldsymbol{\theta}^k$ of the k-th consensus distribution and known covariance matrix $\boldsymbol{\Sigma}$, the posterior distribution is

$$h(\boldsymbol{\theta}^k|\boldsymbol{\mu}^k) \sim \mathcal{N}(\boldsymbol{\mathcal{A}}_k^{-1}\boldsymbol{\mathcal{B}}_k, \boldsymbol{\mathcal{A}}_k^{-1}), \tag{10}$$

where $\boldsymbol{\mathcal{A}}_k = \boldsymbol{u}^t\boldsymbol{\Sigma}_k^{-1}\boldsymbol{u}$, $\boldsymbol{\mathcal{B}}_k = \boldsymbol{u}^t\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}^k$, $\boldsymbol{\Sigma}_k^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}^k$, and $\boldsymbol{u}$ is a $[M^{'} \cdot D \times D]$ design matrix with size $M^{'}$ vectors of 1s along the diagonal and 0s elsewhere, and $\boldsymbol{\mu}^k = (\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \cdots, \boldsymbol{\mu}^D)^T$ be the vector containing estimates of all expert distributions about all dimensions of $\boldsymbol{\theta}^k$, where $\boldsymbol{\mu}^d = (\mu_1^d, \mu_2^d, \cdots, \mu_{M'}^d)^T$.

CoDE - A Simple Example

Outline

Generative modeling with the posterior distribution

Algorithm 1 Generative Modeling with Multimodal VAEs

$\theta, \phi \leftarrow$ Optimized trainable parameters
repeat for $i = 1, \cdots, N$
  $\mathbb{X}^i|\text{available} \leftarrow$ Random multimodal sample from available modalities
  for $\mathbb{X}_k^i \in \mathcal{P}(\mathbb{X}^i)$ **do**
    $\boldsymbol{\epsilon}^i \leftarrow$ Random samples from $\mathcal{N}(\mathbf{0}, \mathbf{1})$
    $\boldsymbol{\mu}^k, \boldsymbol{\sigma}^k = q_k(\boldsymbol{z}|\mathbb{X}_k^i)$
    $\boldsymbol{z}_k^i = \boldsymbol{\mu}^k + \boldsymbol{\sigma}^k \boldsymbol{\epsilon}^i \leftarrow$ latent from posterior
    for $m \in M$ **do**
      $\boldsymbol{x}_{m|k} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{z}_k^i) \leftarrow$ Generate from the m-th likelihood
    end for
  end for
until
return $\boldsymbol{x}_{m|k} \cdots \boldsymbol{x}_{M|K}$

Generative modeling with the prior distribution

Algorithm 1 Generative Modeling with Multimodal VAEs

$\theta, \phi \leftarrow$ Optimized trainable parameters
repeat for $i = 1, \cdots, N$
  $z^i \sim p(z) \leftarrow$ latent from prior
  for $m \in M$ **do**
    $x_m \sim p_\theta(x_m | z^i) \leftarrow$ Generate from the m-th likelihood
  end for
until
return $x_m \cdots x_M$

- Point-estimate: μ_θ. The mode or median is also possible.
- Random samples: $x^i = \mu_\theta^i + \sigma_\theta^i \epsilon^i$

Image to Text

Tags describing images are generated with the multi-modal learning deep Boltzmann machine (DBM) (Srivastava & Salakhutdinov, 2012) and with CMMD. DBM fails to generate coherent tags in the first 3 images. CMMD is, however, able to generate meaningful tags. In the last image, both models generate coherent tags.

				
Generated tags DBM	water, glass, wine, drink, beer, bubbles, splash, drops, drop	portrait, women, soldier, postcard soldiers, army	nikon, d200, tamron, d300, f28, sb600, d60 nikkor, d50, d90	foliage, autumn, trees, leaves, fall, forest, woods, path
Generated tags MVAE	–	statue	car, performance	a700
Generated tags MMVAE IWAE	canon, night, 2007	nikon, green, lion	flower	trees, autumn
Generated tags CMMD	sign, fisheye	animal, lion, outdoors, zoo, k10d, challengeyouwinner, boston, wildlife	apple, food	nature, light autumn, leaves wood, path, forest

Image to Text and Text to Image

Text to Text (Translation)

Type	Sentence
$x_{\text{en}} \sim p_{\text{data}}$	this was one of the highest points in my life.
$x_{\text{vi}} \sim p(x_{\text{vi}}\|z(x_{\text{en}}))$	Đó là một gian tôi vời của cuộc đời tôi.
$\text{GOOGLE}(x_{\text{vi}})$	It was a great time of my life.
$x_{\text{en}} \sim p_{\text{data}}$	the project's also made a big difference in the lives of the people .
$x_{\text{vi}} \sim p(x_{\text{vi}}\|z(x_{\text{en}}))$	tôi án này được ra một Điều lớn lao cuộc sống của chúng người sống chữa hưởng .
$\text{GOOGLE}(x_{\text{vi}})$	this project is a great thing for the lives of people who live and thrive .
$x_{\text{vi}} \sim p_{\text{data}}$	trước tiên , tại sao chúng lại có ấn tượng xấu như vậy ?
$x_{\text{en}} \sim p(x_{\text{en}}\|z(x_{\text{vi}}))$	first of all, you do not a good job ?
$\text{GOOGLE}(x_{\text{vi}})$	First, why are they so bad?
$x_{\text{vi}} \sim p_{\text{data}}$	Ông ngoại của tôi là một người thật đáng <unk> phục vào thời ấy .
$x_{\text{en}} \sim p(x_{\text{en}}\|z(x_{\text{vi}}))$	grandfather is the best experience of me family .
$\text{GOOGLE}(x_{\text{vi}})$	My grandfather was a worthy person at the time .

CelebA Faces generated from the prior

Figure: Left: common latent variable. Right: modality-specific and common latent variables

CelebA Faces generated from the posterior

Figure: Left: common latent variable. Right: modality-specific and common latent variables

Generating Descriptions

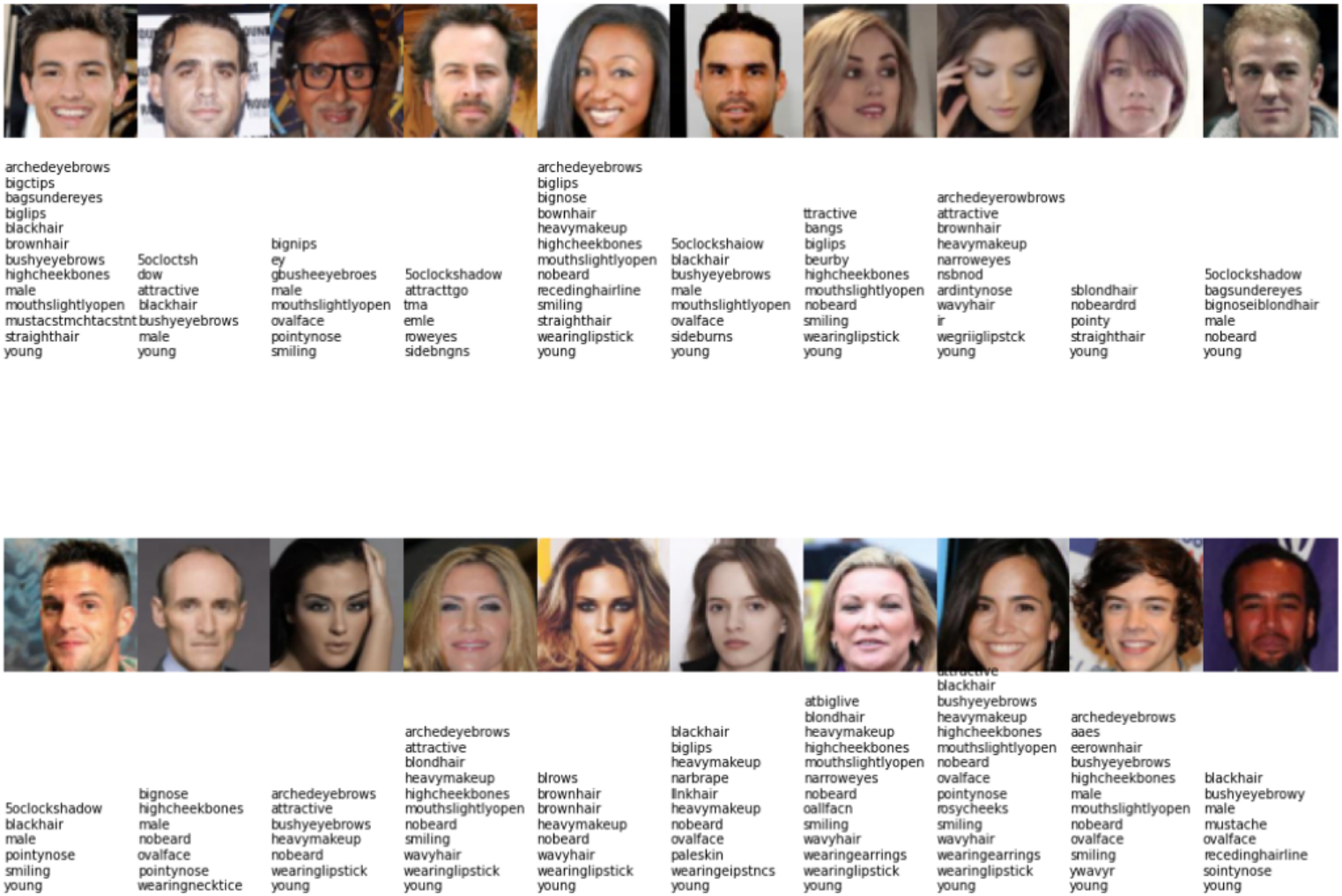


Figure: Top: text generated from image. Right: text generated from posterior

Varying Style and Common Latent Variables

Figure: Left: fixed common latent variables. Right: fixed modality-specific latent variables.