Rethinking Style and Content Disentanglement in Variational Autoencoders

Paper Summary

Daniel Braithwaite

Abstract

This paper considers the problem of separating style and class on digit datasets, specifically SVHN. The authors demonstrate that the VAE objective is not sufficient to enforce this desired separation, yet the CVAE consistently learns a disentangled representation. The authors identify that the shared encoders and decoders play a part in this bias towards disentangled representations, as does the use of stochastic gradient descent.

1 Setup

The context of this paper is learning a latent representation that is independent from a given class label. Specifically, the problem the authors refer to is separating the digit style from the digit class. This is referred to as disentangling the style and class information.

Consider three random variables, X, Y, and Z, that represent the data, class labels, and latent features respectively. Consequently, the dataset consists of, $D = \{x_i, y_i\}$. Then a conditional variational autoencoder can be trained using the following variational lower bound,

$$\begin{aligned}
\log p_\theta(x, y) &= \log \int_z p_\theta(x, y, z) \; dz && (1)\\
&= \log \int_z p_\theta(x, y, z) \frac{q_\phi(z|x, y)}{q_\phi(z|x, y)} \; dz && (2)\\
&= \log \int_z q_\phi(z|x, y) \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)} \; dz && (3)\\
&= \log \mathop{\mathbb{E}}_{z \sim q_\phi(Z|x,y)} \Big[\frac{p_\theta(x, y, z)}{q_\phi(z|x, y)} \Big] && (4)\\
&\geq \mathop{\mathbb{E}}_{z \sim q_\phi(Z|x,y)} \Big[\log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)} \Big] && (5)
\end{aligned}$$

where $q_\phi(Z|X, Y)$ is an additional distribution, also implemented by a neural network. Then the objective function is constructed by taking the expectation of (5),

$$L(q_\phi, p_\theta) = \mathop{\mathbb{E}}_{x,y \sim p_D(X,Y)} \; \mathop{\mathbb{E}}_{z \sim q_\phi(Z|x,y)} \log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)}, \qquad (6)$$

The neural networks that implement q_ϕ and p_θ are given by $\omega_\phi(x,y) = \{\mu_\phi(x,y), \Sigma_\phi(x,y)\}$, and $\omega_\theta(y,z) = \{\mu_\theta(y,z), \Sigma_\theta(y,z)\}$.

2 Contributions

The authors define the following two conditions that must hold for the model to have disentangled the style and class.

- **y is label preserving:** For a fixed $y \in \mathcal{Y}$, $\mu_\theta(y,z)$ outputs images of the same class $\forall z \in \mathcal{Z}$
- **z is style preserving:** For a fixed $z \in \mathcal{Z}$, $\mu_\theta(y,z)$ outputs images with the same style $\forall y \in \mathcal{Y}$.

The authors now show that the VAE objective encourages that y is label preserving, but does not encourage the style preservation.

We can rewrite (6) in the following way,

$$(6) = \underset{x,y\sim p_D(X,Y)}{\mathbb{E}} \; \underset{z\sim q_\phi(Z|x,y)}{\mathbb{E}} \log \frac{p_\theta(z|x,y)p_\theta(x,y)}{q_\phi(z|x,y)} \tag{7}$$

$$= \underset{x\sim p_D(X)}{\mathbb{E}} \; \underset{y\sim p_D(Y|x)}{\mathbb{E}} \log \int_z q_\phi(z|x,y)\frac{p_\theta(z|x,y)p_\theta(x,y)}{q_\phi(z|x,y)} \; dz \tag{8}$$

$$= \underset{x\sim p_D(X)}{\mathbb{E}} \log \int_y \int_z p_D(y|x)q_\phi(z|x,y)\frac{p_\theta(y|x)p_\theta(z|x,y)}{p_D(y|x)q_\phi(z|x,y)} + \log p_D(x|y)p_\theta(x) \; dz \; dy \tag{9}$$

$$= \underset{x\sim p_D(X)}{\mathbb{E}} \big[\log p_\theta(x) - D_{KL}[p_\theta(y|x)p_\theta(z|x,y)||p_D(y|x)q_\phi(z|x,y)]\big] \tag{10}$$

(10) shows that $p_\theta(Y|X)$ is driven towards $p_D(Y|X)$, which the authors claim shows that the VAE objective encourages that p_θ is label preserving.

Y is a known variable, as such, it only contains information about the class label and nothing relating to the style. Consequently, only Z can contain information about the style, the question is, will Z represent the style in a style-preserving way. The authors now show that the VAE objective does not encourage Z to be style preserving.

Proposition 1. *Let p(Z) and p(Y) be fixed and let $p^*(x|y,z)$ be a generator that is style preserving with corresponding true posterior q^*. If p, and q are* ***infinite capacity models****, then there exists $p^{'}$ and $q^{'}$ such that $L(p^*,q^*) = L(p^{'},q^{'})$ but $p^{'}$ is not style preserving.*

Proof. Let Z be the random variable for $p(Z)$, and consider a set of distinct, distrubution preserving transforms, $\{T_i, Z \to Z\}_{i\in Y}$. We have that $T_i(Z) = Z, \forall\ i$ and $\forall\ i,j\ \exists z \in Z$ such that $T_i(z) \neq T_j(z)$.

Let $p^{'}(x|y,z) = p^*(x|y,T_y(z))$, and let $q^{'}$ be the corresponding posterior. It follows that $L(p^*,q^*) = L(p^{'},q^{'})$. Moreover, since the transformations are distinct, $\exists i,j$ and $z \in Z$ such that $p^{'}(x|i,z)$ and $q^{'}(x|j,z)$ produce different styles. □

Proposition 1 demonstrates that it is possible to transform a model with a disentangled representation into a entangled one, without adjusting the loss. One concrete example of this is when the distributions are Gaussian, then label dependent rotations convert a disentangled representation into an entangled one. The authors state that any learning algorithm based on minimising $d(p_D(x, y)||p(x, y))$ according to the divergence d, cannot differentiate p^* and p'. The following question still remains, *why does the model learn a label and style preserving representation if the VAE objective does not strictly enforce it?*

The first hypothesis is that it is a result of a shared encoder, decoder or both. To test this hypothesis, the authors construct an alternate form of the VAE, consisting of 10 different neural networks, one for each class. By removing the shared parameters between the VAEs for the different classes, it is unlikely that a style preserving latent representation will be learned. Consequently, this alternative form can be used to study the disentanglement bias of specific parts of the network. The authors find that it is only necessary to have a shared encoder or decoder to achieve disentangled representations.

Next it is shown that stochastic gradient descent prefers style preserving representations. A CVAE is used to learn a style preserving representation, from which an entangled representation can be constructed. The authors then train an encoder and decoder separately on both the disentangled and entangled representations, finding that while the models converge to the same loss, the disentangled representation converges faster in both cases. This indicates that stochastic gradient descent has a preference for the disentangled model.

3 Observations

Doesn't this make sense? Proposition 1 assumes infinite capacity models, but in the real world, models have finite capacity. Since the model is already given y does it make sense that the model should instead incentivise packing as much style information as possible into z. Still this is a relatively informal argument.