

# On variational learning of controllable representations for text without supervision

Xavier Garcia

July 2019

## 1 Comments

### 1.1 High-level

**Latent space manipulations for text** Common generative models usually involve some type of continuous latent variable, which allows for various degrees of control on the attributes of generation. For a given attribute, one can find an *attribute vector*  $z_{atr}$  so that if  $z(x)$  is the latent representation of  $x$ , then  $z(x) + z_{atr}$  will decode to an output which has the desired attribute. While such techniques have worked for images, it doesn't work for text.

**The Vacant Space Hypothesis** One theoretical explanation goes as follows. Notice that the decoder  $p_\theta$  only ever sees samples from  $q_\phi(z|x)$ . Therefore, the decoder has learned that its input distribution is  $q_\phi(z) = \mathbb{E}_x q_\phi(z|x)$ . If  $z(x) + z_{atr}$  has landed in a region of low density for  $q_\phi$ , then this sample could be considered out of distribution and hence we have no guarantees that it should decode to anything. Such “vacant regions” could exist because it is impossible to sample enough samples to cover an unbounded space. Moreover, even if we constrained the latent space, there is no guarantee that the generative model would learn to use all of it anyways.

### 1.2 Medium-level

**Example: Sentiment transfer** Suppose you train a  $\beta$ -VAE on the Yelp restaurant reviews dataset, with a standard setup involving a standard LSTM encoder-decoder setup. A neat observation is that there exists a single dimension (say,  $j$ ) such that if you only used this value to predict sentiment, you would have 90% accuracy, while the other codes have about 50%. This suggests a simple sentiment transfer strategy: Encode your input  $x$  into the latent space to get a representation  $z(x)$ , modify said representation by adding  $z(x) + \lambda e_j$  and then decode this representation. Here  $e_j$  is the vector consisting of all zeros except for a value of 1 at the  $j$ th coordinate. For large  $\lambda$ , the samples have the correct style but are largely irrelevant to the input. For low  $\lambda$ , you don't change

the style. Empirically, they plotted the negative log-likelihood value for 1000 different values of  $\lambda$  and observed the the value rises sharply after some point, empirically verifying that there exists these “gaps” in the space.

**Proposed prior and posterior distributions** One way to get less gaps is to focus on a particular part of space. For example, if you constrained the posterior distribution to be a Gaussian with its mean constrained to some finite region of space, then you would expect at least that points near that region would be seen during training. To do so, suppose you have an orthonormal set of vectors  $\{v_i\}_{i=1}^K$  (which consist of learnable parameters) for some hyperparameter  $K$ . We shall restrict the means of the posterior  $q$  to be within the simplex spanned by these vectors i.e.  $q(\cdot|x) \sim \mathcal{N}(\mu, \sigma^2)$  where  $\mu = \sum_{i=1}^K p_i v_i$  for non-negative scalars  $p_i := p_i(x)$  which sum up to 1. Notice that the explicit formula for the KL terms still remains true, since the posteriors are still Gaussians. However, in this setup, it even simplifies further: Suppose that  $v_i \cdot v_j = \alpha \delta_{ij}$  where  $\alpha > 0$  and  $\delta$  is the Kronecker delta which is 0 unless  $i = j$  in which case its one. Then,

$$\begin{aligned} \text{KL}(q(z|x)||p(z)) &= \frac{1}{2}(\mu^T \mu + \sigma^T \sigma - \sigma^T \log \sigma - K) + \text{Constant} \\ &= \frac{1}{2}(\sum_{i=1}^K \alpha p_i^2 + \sigma^T \sigma - \sigma^T \log \sigma - K) + \text{Constant} \end{aligned}$$

By construction, the KL term cannot collapse. In practice, we enforce the orthogonality by introducing an additional term  $\mathcal{L}_{REG} = ||vv^T - \alpha I||^2$ .

**Using up the space** Even though we have constrained the means, its possible for the model to still force everything to try to minimize the KL as much as possible i.e. make  $p$  be uniform. To do this, we introduce a term to encourage different inputs to have  $p$ . Namely, we’ll encourage the means  $\mu$  to be close to the logits  $h$  of  $p$  and far away from other negative samples. In particular, for a given input  $\mu$  with logit  $h$ , suppose that  $\mu_i$  are different inputs (sampled randomly). Then we define the loss term  $\mathcal{L}_{S-REG}$  as follows:

$$\mathcal{L}_{S-REG} = \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \frac{1}{m} \sum_{i=1}^m \max(0, 1 - h \cdot \mu + h \cdot \mu_i) \right]$$