SentenceMIM: A Latent Variable Language Model

Micha Livne [1 2] **Kevin Swersky** [3] **David J. Fleet** [1 2]

Abstract

We introduce sentenceMIM, a probabilistic auto-encoder for language modelling, trained with Mutual Information Machine (MIM) learning. Previous attempts to learn variational auto-encoders for language data have had mixed success, with empirical performance well below state-of-the-art auto-regressive models, a key barrier being the occurrence of posterior collapse with VAEs. The recently proposed MIM framework encourages high mutual information between observations and latent variables, and is more robust against posterior collapse. This paper formulates a MIM model for text data, along with a corresponding learning algorithm. We demonstrate excellent perplexity (PPL) results on several datasets, and show that the framework learns a rich latent space, allowing for interpolation between sentences of different lengths with a fixed-dimensional latent representation. We also demonstrate the versatility of sentenceMIM by utilizing a trained model for question-answering, a transfer learning task, without fine-tuning. To the best of our knowledge, this is the first latent variable model (LVM) for text modelling that achieves competitive performance with non-LVM models.

1. Introduction

Generative modelling of text has become one of the predominant approaches to natural language processing (NLP), particularly in the machine learning community. It is favoured because it supports probabilistic reasoning and it provides a principled framework for unsupervised learning in the form of maximum likelihood. Unlike computer vision, where various generative approaches have proliferated (Dinh et al., 2017; Goodfellow et al., 2014; Kingma & Welling, 2013; Oord et al., 2016; Rezende et al., 2014), current methods for text mainly rely on auto-regressive models.

Generative latent variable models (LVMs), such as the variational auto-encoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014), are versatile and have been successfully applied to a myriad of domains. Such models consist of an encoder, which maps observations to distributions over latent codes, and a decoder that maps latent codes to distributions over observations. LVMs are widely used and studied because they can learn a latent representation that carries many useful properties. Observations are encoded as fixed-length vectors that capture salient information, allowing for semantic comparison, interpolation, and search. They are often useful in support of downstream tasks, such as transfer or k-shot learning. They are also often interpretable, capturing distinct factors of variation in different latent dimensions. These properties have made LVMs especially compelling in the vision community.

Despite their desirable qualities, generative LVMs have not enjoyed the same level of success in text modelling. There have been several recent proposals to adapt VAEs to text (Bowman et al., 2015; Guu et al., 2017; Kruengkrai, 2019; Li et al., 2019b; Yang et al., 2017), but despite encouraging progress, they have not reached the same level of performance on natural language benchmarks as auto-regressive models (*e.g.*, (Merity et al., 2017; Rae et al., 2018; Wang et al., 2019)). This is often attributed to the phenomenon of posterior collapse (Le Fang, 2019; Li et al., 2019a), in which the decoder captures all of the modelling power and the encoder ends up conveying little to no information. For text, where the decoder is naturally auto-regressive, this has proven challenging to mitigate.

This paper introduces sentenceMIM (sMIM), a new LVM for text. It is based on the architecture of Bowman et al. (2015) and the mutual information machine (MIM) framework (Livne et al., 2019). MIM is a recently introduced LVM framework that shares the same underlying architecture as VAEs, but uses a different learning objective that is more robust against posterior collapse. MIM learns a highly informative and compressed latent representation, and often strictly benefits from more powerful architectures. To evaluate sMIM we propose a novel bound on the model log-likelihood, called MIM-ELBO, or *MELBO* . As an alternative to the evidence lower bound (ELBO) used to evaluate

[1]Department of Computer Science, University of Toronto [2]Vector Institute [3]Google Research. Correspondence to: Micha Livne <mlivne@cs.toronto.edu>, Kevin Swersky <kswersky@google.com>, David J. Fleet <fleet@cs.toronto.edu>.

VAEs, MELBO is useful for models with implicit priors, for which the ELBO is intractable.

We show on four challenging datasets that sMIM outperforms VAE models for text, and is competitive with state-of-the-art auto-regressive approaches, including transformer-based models (Radford et al., 2019; Vaswani et al., 2017), as measured by negative log-likelihood and perplexity. We further demonstrate the quality of the sMIM representation by generating diverse samples around a given sentence and interpolating between sentences. Finally, we show the versatility of the learned representation by applying a pre-trained sMIM model to a question answering task with state-of-art performance as compared to single task, supervised models.

2. Problem Formulation

Let $\boldsymbol{x} \in \mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{X}$ be a discrete variable representing a sentence of tokens of length $T \in \{1, ..., T_{max}\}$ from a finite vocabulary $\mathcal{V}$, where T_{max} is the maximum sentence length. The set $\mathcal{X}$ comprises all sentences we aim to model. The total number of sentences X is typically unknown and large. Let $\mathcal{P}(\boldsymbol{x})$ be the unknown probability of sentence $\boldsymbol{x}$.

Our goal is to learn a latent variable model given N fair samples from $\mathcal{P}(\boldsymbol{x})$, where $N \ll X$. To this end, we consider probabilistic auto-encoders, defining distributions over discrete observations $\boldsymbol{x} \in \mathcal{X}$, and a corresponding continuous latent space, $\boldsymbol{z} \in \mathbb{R}^d$. They consist of an encoder, $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, mapping sentences to a distribution over continuous latent codes, and a corresponding decoder, $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, providing a distribution over sentences given a latent code. The joint parameters of the encoder and the decoder are denoted by $\boldsymbol{\theta}$. Ideally the encoder maps inputs to latent codes from which the decoder can correctly reconstruct the input. We also desire a latent space in which similar sentences (*e.g.*, in structure or content) are mapped to nearby latent codes.

2.1. Encoder-Decoder Specification

In what follows we adapt the architecture proposed by Bowman et al. (2015). Beginning with the generative process, let $p_{\boldsymbol{\theta}}^{seq}(\boldsymbol{x}|\boldsymbol{z})$ be a conditional auto-regressive distribution over sequences of T tokens. We express the log probability of a sequence, $\boldsymbol{x} = (x^1, \ldots, x^T)$, with tokens $x^k \in \mathcal{V}$ (and a slight abuse of notations), as

$$\log p_{\boldsymbol{\theta}}^{seq}(\boldsymbol{x}|\boldsymbol{z}) \;=\; \sum_{k=1}^{T} \log p_{\boldsymbol{\theta}}^{seq}(x^k \,|\, x^{k-1}, \ldots, x^1, \boldsymbol{z}) \quad (1)$$

where $p_{\boldsymbol{\theta}}^{seq}(x^k|\cdot)$ is a categorical distribution over $|\mathcal{V}|$ possible tokens for the kth token in $\boldsymbol{x}$, and $\boldsymbol{x}^0 \equiv$ <SOS> is the start-of-sentence token. According to the model (see Fig. 1), generating a sentence $\boldsymbol{x}$ with latent code $\boldsymbol{z}$ entails sampling each token from a distribution conditioned on the latent code

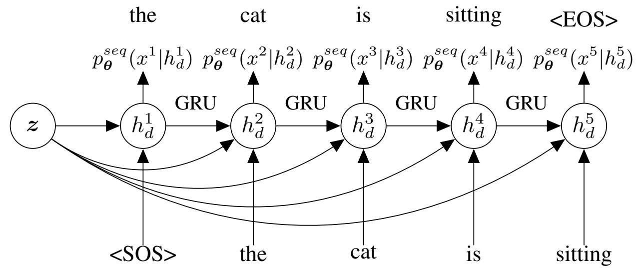


Figure 1. The decoder is auto-regressive, and conditioned on latent code $\boldsymbol{z}$. Words are represented by parametric embeddings. In each step (except the first) the previous output token and the latent code are inputs, and the GRU hidden output is then mapped to the parameters of a categorical distribution $p_{\boldsymbol{\theta}}^{seq}(x^k|h_d^k)$, from which the next token is sampled. The top sentence depicts the sample, with inputs on the bottom.

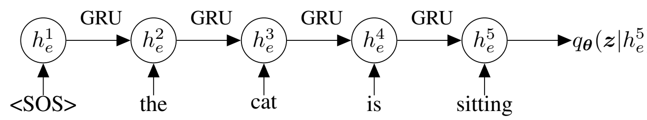


Figure 2. The encoder is implemented with GRU. Each word is represented by a parametric embedding. Given the input sequence, the encoder maps the last hidden state to the mean and variance of Gaussian posterior over latent codes, $q_{\boldsymbol{\theta}}^{seq}(z|h_d^k)$.

and previously sampled tokens. Tokens are modelled with a parametric embedding.

The auto-regressive model, $p_{\boldsymbol{\theta}}^{seq}(\boldsymbol{x}|\boldsymbol{z})$, sums to one over all sequences of a given length. Combining this model with a distribution over sentence lengths $p(\ell)$, for $\ell \in \{1, ..., T_{max}\}$, we obtain the decoder, *i.e.*, a distribution over all sentences in $\mathcal{X}$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \;=\; p_{\boldsymbol{\theta}}^{seq}(\boldsymbol{x}|\boldsymbol{z})\; p(T = \ell)\;. \quad (2)$$

Here, $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ sums to one over all sentences of all lengths. The corresponding marginal $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ is discussed in Sec. 2.3.

The encoder, or posterior distribution over latent codes given a sentence, $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, is a conditional distribution over the latent variable $\boldsymbol{z}$. We take this to be Gaussian whose mean and diagonal covariance are specified by mappings $\mu_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$:

$$q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \;=\; \mathcal{N}(\boldsymbol{z}; \mu_{\boldsymbol{\theta}}(\boldsymbol{x}), \sigma_{\boldsymbol{\theta}}(\boldsymbol{x})) \quad (3)$$

Linear mappings $\mu_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$ are computed from the last hidden state of a GRU (Cho et al., 2014) (see Fig. 2).

2.2. Background: MIM Learning Objective

The Mutual Information Machine (MIM), introduced by Livne et al. (2019), is a versatile LVM. Like the VAE, it serves as a framework for representation learning, probability density estimation, and sample generation. Importantly, MIM learns a model with high mutual information between observations and latent codes, and with robustness against posterior collapse, which has been problematic for VAEs with language data (*e.g.*, Bowman et al. (2015)).

MIM is formulated in terms of several elements. It assumes two *anchor* distributions, $\mathcal{P}(x)$ and $\mathcal{P}(z)$, for observations and the latent space, from which one can draw samples. They are fixed and not learned. There is also a parameterized encoder-decoder pair, $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, and parametric marginal distributions $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z})$. These parametric elements define joint encoding and decoding distributions:

$$
\begin{aligned}
q_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}) &= q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{x})\,, && (4)\\
p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}) &= p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\, p_{\boldsymbol{\theta}}(\boldsymbol{z})\,. && (5)
\end{aligned}
$$

For language modeling we use A-MIM learning, a MIM variant that minimizes a loss defined on the encoding and decoding distributions, with samples drawn from an encoding *sample* distribution, denoted $\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z})$; *i.e.*,

$$
\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z}) \;=\; q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\,\mathcal{P}(\boldsymbol{x})\,. \tag{6}
$$

The particular loss for A-MIM is a variational upper bound on the joint entropy of the encoding sample distribution, which can be expressed with marginal entropies and mutual information terms. More precisely,

$$
\begin{aligned}
\mathcal{L}_{\text{A-MIM}}(\boldsymbol{\theta}) \;&=\; \frac{1}{2}\big(\, CE\,(\,\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z})\,,\, q_{\boldsymbol{\theta}}\,(\boldsymbol{x},\boldsymbol{z})\,) && (7)\\
&\qquad +\, CE\,(\,\mathcal{M}^q_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{z})\,,\, p_{\boldsymbol{\theta}}\,(\boldsymbol{x},\boldsymbol{z})\,)\,\big)\\
&\geq\; H_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{x}) + H_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{z}) - I_{\mathcal{M}^q_{\mathcal{S}}}(\boldsymbol{x};\boldsymbol{z})\,,
\end{aligned}
$$

where $CE\,(\,\cdot\,,\,\cdot\,)$ is cross-entropy, $H_{\mathcal{M}^q_{\mathcal{S}}}(\cdot)$ is information entropy over distribution $\mathcal{M}^q_{\mathcal{S}}$, and $I(\cdot;\cdot)$ is mutual information. Minimizing $\mathcal{L}_{\text{A-MIM}}(\boldsymbol{\theta})$ learns a model with a consistent encoder-decoder, high mutual information, and low marginal entropy (Livne et al., 2019).

2.3. Variational Model Marginals

To complete the model specification, we define the model marginals $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z})$. To help encourage consistency, and avoid introducing more model parameters, one can define model marginals in terms of marginals of the sample distributions (Bornschein et al., 2015; Livne et al., 2019; Tomczak & Welling, 2017).

We define the model marginal over observations as a marginal over the decoder (Bornschein et al., 2015): *i.e.*,

$$
q_{\boldsymbol{\theta}}(\boldsymbol{x}) \;=\; \mathbb{E}_{\mathcal{P}(\boldsymbol{z})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]\,, \tag{8}
$$

where the latent anchor is defined to be a standard normal, $\mathcal{P}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};0,1)$. Similarly, one can define the model marginal over latent codes as a marginal of the encoder,

$$
p_{\boldsymbol{\theta}}(\boldsymbol{z}) \;=\; \mathbb{E}_{\mathcal{P}(\boldsymbol{x})}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\,. \tag{9}
$$

The latent marginal is defined as the aggregated posterior, in the spirit of the VampPrior (Tomczak & Welling, 2017).

2.4. Tractable Bounds to Loss

Given a training dataset $D = \{\boldsymbol{x}_i\}_{i=1}^N$, an empirical approximation to $\mathcal{L}_{\text{A-MIM}}(\boldsymbol{\theta})$ is

$$
\begin{aligned}
\hat{\mathcal{L}}_{\text{A-MIM}}(\boldsymbol{\theta}) \;=\; &-\frac{1}{2N}\sum_{\boldsymbol{x}_i}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)\, q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right]\\
&-\frac{1}{2N}\sum_{\boldsymbol{x}_i}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\, p_{\boldsymbol{\theta}}(\boldsymbol{z})\right] \quad (10)
\end{aligned}
$$

where $\sum_{\boldsymbol{x}_i}$ denotes $\sum_{\boldsymbol{x}\in D}$, a sum over N fair samples drawn from $\mathcal{P}(\boldsymbol{x})$, as a Monte Carlo approximation to expectation over $\mathcal{P}(\boldsymbol{x})$.

Unfortunately, the empirical loss in Eqn. (10) is intractable since we cannot evaluate the log-probability of the marginals $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{x})$. In what follows we obtain a tractable empirical bound on the loss in Eqn. (10) for which, with one joint sample, we obtain an unbiased and low-variance estimate of the gradient (*i.e.*, using the reparameterization trick (Kingma & Welling, 2013)).

We first derive a tractable lower bound to $\log q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$:

$$
\begin{aligned}
\log q_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \;&=\; \log \mathbb{E}_{\mathcal{P}(\boldsymbol{z})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\right] && (11)\\
&\underset{\text{(IS)}}{=}\; \log \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\frac{\mathcal{P}(\boldsymbol{z})}{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\right]\\
&\underset{\text{(JI)}}{\geq}\; \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\frac{\mathcal{P}(\boldsymbol{z})}{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\right)\right]
\end{aligned}
$$

where the second and third lines are obtained using importance sampling and Jensen's inequality. We remind the reader that $q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is a variational marginal that can depend on $\boldsymbol{x}_i$. Indeed, Eqn. (11) is the usual ELBO.

To derive a lower bound to $\log p_{\boldsymbol{\theta}}(\boldsymbol{z})$, we begin with the following inequality,

$$
\begin{aligned}
\log \mathbb{E}_{\mathcal{P}(\boldsymbol{x})}\left[h(\boldsymbol{x};\cdot)\right] \;&=\; \log \sum_i \mathcal{P}(\boldsymbol{x}_i)h(\boldsymbol{x}_i;\cdot)\\
&\geq\; \log \mathcal{P}(\boldsymbol{x}')h(\boldsymbol{x}';\cdot)\,, && (12)
\end{aligned}
$$

for any sample $\boldsymbol{x}'$, any discrete distribution $\mathcal{P}(\boldsymbol{x})$, and any non-negative function $h(\boldsymbol{x};\cdot) \geq 0$. The inequality in Eqn. (12) follows from $\log a \geq \log b$ for $a \geq b$. Using this bound, we express a lower bound to $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ as follows,

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{z}) \;&\underset{\text{(Eqn. 9)}}{=}\; \log \mathbb{E}_{\mathcal{P}(\boldsymbol{x})}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\\
&\underset{\text{(Eqn. 11)}}{\geq}\; \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}') + \log \mathcal{P}(\boldsymbol{x}') && (13)
\end{aligned}
$$

for any sample $\boldsymbol{x}'$. During training, given a joint sample $\boldsymbol{x}_i, \boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\,\mathcal{P}(\boldsymbol{x})$, we choose $\boldsymbol{x}' = \boldsymbol{x}_i$.

Algorithm 1 Learning parameters $\boldsymbol{\theta}$ of sentenceMIM

while not converged **do**
  $D_{\text{enc}} \leftarrow \{\boldsymbol{x}_j, \boldsymbol{z}_j \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\mathcal{P}(\boldsymbol{x})\}_{j=1}^{N}$
  $\hat{\mathcal{L}}_{\text{MIM}}(\boldsymbol{\theta}; D) = -\frac{1}{N}\sum_{i=1}^{N}\big(\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_i) + \frac{1}{2}(\log q_{\boldsymbol{\theta}}(\boldsymbol{z}_i|\boldsymbol{x}_i) + \log \mathcal{P}(\boldsymbol{z}_i)) \big)$
  $\Delta\boldsymbol{\theta} \propto -\nabla_{\boldsymbol{\theta}}\hat{\mathcal{L}}_{\text{MIM}}(\boldsymbol{\theta}; D)$ {*Gradient computed through sampling using reparameterization*}
end while

Substituting Eqns. (11) and (13) into Eqn. (10) gives the final form of an upper bound on the empirical loss; *i.e.*,

$$\begin{aligned}\hat{\mathcal{L}}_{\text{A-MIM}} \;\leq\; & -\frac{1}{N}\sum_i \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\right] \\ & -\frac{1}{2N}\sum_i \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[\log\big(q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)\mathcal{P}(\boldsymbol{z})\big)\right] \\ & +\frac{1}{2}H_{\mathcal{P}}(\boldsymbol{x}) \;. \end{aligned} \tag{14}$$

We find an unbiased, low variance estimate of the gradient of $\hat{\mathcal{L}}_{\text{A-MIM}}$ with a single joint sample $\boldsymbol{z}_i, \boldsymbol{x}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\,\mathcal{P}(\boldsymbol{x})$ and reparameterization. The last term, $H_{\mathcal{P}}(\boldsymbol{x})$, is a constant, independent of model parameters and can therefore be ignored during optimization. The resulting learning process is described in Algorithm 1.

To better understand the proposed bounds, we note that MIM achieves good reconstruction by learning posteriors with relatively small variances (*i.e.*, relative to the distance between latent means). Our choice of $\boldsymbol{x}' = \boldsymbol{x}_i$ exploits this, allowing good gradient estimation, facilitating fast convergence. We further provide empirical evidence for these properties below in Fig. 3.

3. NLL Evaluation

As an alternative to the ELBO bound on log likelihood, here we propose an new bound that is better suited to models with implicit priors. With implicit priors, both NLL and ELBO are computationally expensive to estimate. Unlike ELBO, the new bound, called MELBO (for MIM-ELBO) does not entail the evaluation of the log-likelihood of the latent model marginal. In the VAE literature, it is common to change the prior once the encoder and decoder have been trained (*e.g.*, Razavi et al. (2019); van den Oord et al. (2017)). Interestingly, if post-hoc we change the latent prior to be a marginal distribution, as in Eqn. 9, then MELBO can be used to bound the NLL. This is particularly effective when the aggregated posterior is a poor fit to the original Gaussian prior, which is penalized heavily in the EBLO.

Here we discuss model evaluation under a given empirical target distribution $\mathcal{T}(\boldsymbol{x})$, and in particular, the empirical test set. We start with a bound on $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, *i.e.*,

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) =\; & \log \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\right] \\ \underset{\text{(IS)}}{=}\; & \log \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{z})}{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\right] \\ \underset{\text{(Eqn. 9)}}{=}\; & \log \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{T}(\boldsymbol{x}), q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\frac{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}')}{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\right] \\ \underset{\text{(Eqn. 12)}}{\geq}\; & \log \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})\right] + \log \mathcal{T}(\boldsymbol{x}_i) \end{aligned} \tag{15}$$

where the second step uses importance sampling, and the variational marginal in Eqn. (9) is defined here under the target empirical distribution $\mathcal{T}(\boldsymbol{x})$ in the last step. This allows us to choose $\boldsymbol{x}' = \boldsymbol{x}_i$, motivated by the tendency for MIM to learn highly clustered representations (cf. Fig. 3). We can also view Eqn. (15) as an alternative to the usual ELBO; we refer to it as *MELBO* (*i.e.*, MIM ELBO). Like the ELBO, this bound holds for each data point, independent of target distribution $\mathcal{T}(\boldsymbol{x})$.

We can now derive an upper bound on the NLL under $\mathcal{T}$ using the MELBO :

$$\begin{aligned} & -\mathbb{E}_{\mathcal{T}(\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x})\right] \\ & \underset{\text{(Eqn. 15)}}{\leq} -\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\mathcal{T}(\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + H_{\mathcal{T}}(\boldsymbol{x}) \\ & \underset{\text{(MC)}}{\lessapprox} -\frac{1}{N}\sum_{\boldsymbol{x}_i}\left(\frac{1}{N_{\boldsymbol{z}}}\sum_{j=1}^{N_{\boldsymbol{z}}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_{i,j})\right) + \log N \end{aligned} \tag{16}$$

where $\boldsymbol{x}_i \in D$, and $N_{\boldsymbol{z}}$ samples $\boldsymbol{z}_{i,j}$ are drawn from the encoder $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)$. The last inequality follows $\log N$ being an upper bound on the entropy for the empirical distribution $\mathcal{T}(\boldsymbol{x})$. We denote this empirical upper-bound by $\widehat{NLL}$, and the corresponding perplexity (PPL) upper bound by $\widehat{PPL} \equiv \exp(\frac{N\cdot\widehat{NLL}}{\sum_i T_i}) \geq PPL$, where $\sum_i T_i$ is the total number of tokens in a dataset with N samples.

For the sizes of the datasets we consider, MELBO and ELBO tend to be comparable for VAE. Notice that $\widehat{NLL}$ grows with the number of unique sentences in the dataset (*i.e.*, categories of a discrete variable), as expected from a categorical distribution.

4. Experiments

4.1. Datasets

We show experimental results on four word level datasets[1] described in Table 1, namely, Penn Tree Bank (Marcus et al., 1993), Yahoo Answers and Yelp15 (following Yang et al. (2017)), and WikiText-103 (Merity et al., 2016). We use the Yahoo and Yelp15 datasets of Yang et al. (2017), which

[1] <SOS>, <EOS> are a special start/end-of-sentence tokens. The token <UNK> represents an out-of-vocabulary word.

(word level)	Sentences Train	Valid.	Test	Vocab.	#words (avg.)
PTB	42068	3370	3761	9877	21 ± 10
Yahoo	100K	10K	10K	37165	76 ± 55
Yelp15	100K	10K	10K	19730	100 ± 51
WikiText-103	200K	10K	2185	89247	115 ± 60
Everything †	442067	33369	33760	105965	94 ± 60

Table 1. Dataset properties summary for Penn Tree Bank (Marcus et al., 1993), Yahoo Answers and Yelp15 (cf. Yang et al. (2017)), and sampled WikiText-103 (Merity et al., 2016). Everything † is the union of all datasets.

draw 100k samples for training, and 10k for validation and testing. For WT103 we draw 200k samples for training, 10k for validation, and retain the original test data. Empty lines and headers were filtered from the WT103 data.

4.2. Architecture and Optimization

Our auto-encoder architecture (Figs. 1 and 2), was adapted from that proposed by Bowman et al. (2015). As is common, we concatenated $\boldsymbol{z}$ with the input to the decoder (*i.e.*, a "context", similar to He et al. (2019); Yang et al. (2017); Bowman et al. (2015)). We use the same architecture, parameterization and latent dimensionality for sMIM and a VAE variant called sVAE, for comparison. Training times for sVAE and sMIM are similar.

For PTB we trained models with 1 layer GRU, latent space dimensions of 16D, 128D, and 512D, a 512D hidden state, 300D word embeddings, and 50% embedding dropout. We trained the models with Adam (Kingma & Lei Ba, 2014) with initial learning rate $lr = 10^{-3}$. The best performing model was trained in less than 30 minutes on a single TITAN Xp 12G GPU. For Yahoo Answers, Yelp15, and WT103 we trained models with 1 layer GRU, latent space dimensions of 32D, 512D, 1024D, a 1024D hidden state, 512D word embeddings, and 50% embedding dropout. We trained these models with SGD (Sutskever et al., 2013), with initial $lr = 5.0$, and 0.25 L_2 gradient clipping.

In all cases we use a learning rate scheduler that scaled the learning rate by 0.25 following two/one epochs (PTB/other datasets, respectively) with no improvement in the validation loss. We used a mini-batch size of 20 in all cases. Following (Sutskever et al., 2014) we feed the input in reverse to the encoder, such that the last hidden state in the encoder depends on the first word of the sentence in the decoder. (This gave slightly better results than with left to right order.)

We trained sVAEs with the regular ELBO, and with KL divergence annealing (denoted "+ kl"), where a scalar weight on the KL divergence term is increased from 0 to 1 over 10k mini-batches to lower the risk of posterior collapse and improve the learned models (Bowman et al., 2015). We use no loss manipulation heuristics in the optimization of sMIM.

LVM ($\boldsymbol{z}$ dim.)	PPL (stdev)	NLL [KLD]	BLEU	$\|\boldsymbol{\theta}\|$
sVAE (16)	$\leq$ 110.72 (0.12) * $\leq$ 148.18 (0.11)	$\leq$ 106.84 [1.6] * $\leq$ 113.46	0.124	11M
sVAE (128)	$\leq$ 113.3 (0.12) * $\leq$ 158.31 (0.13)	$\leq$ 107.36 [0.64] * $\leq$ 114.96	0.118	11M
sVAE (512)	$\leq$ 121.44 (0.23) * $\leq$ 171.37 (0.31)	$\leq$ 108.93 [0.41] * $\leq$ 116.76	0.116	12M
sMIM (16)	$\leq$ 76.3 (0.03)	$\leq$ 98.35	0.35	11M
sMIM (128)	$\leq$ 27.93 (0.008)	$\leq$ 75.58	0.61	11M
sMIM (512)	$\leq$ **19.53** (0.01)	$\leq$ **67.46**	**0.679**	12M
sMIM (1024) †	$\leq$ **4.6** (0.0)	$\leq$ **34.66**	**0.724**	179M
(a) VAE-LSTM (13) ‡	119	101 [2]		
(b) iVAE_{MI} (32) ‡	$\leq$ 53.44	$\leq$ 87.2 [12.51]		
(c) HR-VAE (256)	43	79 [10.4]		
auto-regressive				
(d) GPT-2 full †	35.76			1542M

Table 2. PPL and NLL results for **PTB** bounded with MELBO , averaged over 10 runs (see text for details). PPL* and NLL* are bounded with ELBO. Models† use extra training data. (a) Bowman et al. (2015); (b) Le Fang (2019); (c) Li et al. (2019a); (d) Radford et al. (2019). A test set with 2×10^9 samples is required for MELBO of best performing sMIM to reach the PPL of next best model. (a)‡ inconsistencies in PPL values can be explained by Bowman et al. (2015) including <EOS> during evaluation.

4.3. Language Modelling Results

In what follows we compare the perplexity (PPL) of sMIM, sVAE, other top performing VAEs, and auto-regressive models. For all datasets but PTB, VAE learning with KL annealing was more effective than standard VAE learning; due to the small size of PTB, annealing produced over-fitting. We remove the <EOS> token during evaluation, allowing fair PPL comparison with auto-regressive models[2].

Tables 2-5 show results for PTB, Yelp15, Yahoo Answers, and WT103. Model sMIM (1024)† is trained on all datasets (*i.e.*, PTB, Yahoo Answers, Yelp15 and WT103). The BLEU-1 score is computed between test sentences and their reconstructions (higher is better). PPL and NLL (lower is better) are bounded with MELBO (Eqn. (16)). PPL* and NLL* are bounded with ELBO. Finally, $|\boldsymbol{\theta}|$ indicates the number of parameters in each model.

Results were validated using four methodologies. First, we provide an additional independent measure to the reconstruction quality with the unigram BLEU score (Papineni et al., 2001) between test sentences and their reconstructions. We

[2]For auto-regressive models, the standard PPL evaluation protocol treats the test corpus as one long sequence. For VAEs, the standard protocol involves estimation of NLL over sentences. We use this protocol for sVAE and sMIM NLL/PPL evaluation.

LVM ($\boldsymbol{z}$ dim.)	PPL (stdev)	NLL [KLD]	BLEU	$\|\boldsymbol{\theta}\|$
sVAE (32) + kl	$\leq$ 78.53 (0.01) * $\leq$ 62.51 (0.01)	$\leq$ 433.49 [31.86] * $\leq$ 410.83	0.274	40M
sVAE (512) + kl	$\leq$ 47.78 (0.01) * $\leq$ 50.25 (0.01)	$\leq$ 384.12 [4.19] * $\leq$ 389.14	0.18	43M
sVAE (1024) + kl	$\leq$ 49.61 (0.01) * $\leq$ 52.79 (0.01)	$\leq$ 387.86 [3.01] * $\leq$ 394.04	0.176	46M
sMIM (32)	$\leq$ 59.28 (0.0)	$\leq$ 405.55	0.309	40M
sMIM (512)	$\leq$ 10.05 (0.0)	$\leq$ 229.24	0.673	43M
sMIM (1024)	$\leq$ **9.98** (0.0)	$\leq$ **228.58**	**0.676**	46M
sMIM (1024)†	$\leq$ **8.19** (0.0)	$\leq$ **208.93**	**0.686**	179M
(a) CNN-VAE (32)	41.1	359.1 [7.6]		
(b) SA-VAE + anneal (32)		355.9 [3.8]		
(c) LSTM-VAE (32)		358.3 [0.0]		
(e) iVAE$_{\mathrm{MI}}$ (32)	$\leq$ 36.88	$\leq$ 348.7 [11.6]		
(f) Neural Editor (128; k=25)	26.87	318.03		
auto-regressive				
(d) LSTM-LM		358.1		

Table 3. PPL and NLL results for **Yelp15** bounded with MELBO , averaged over 10 runs (see text for details). PPL* and NLL* are bounded with ELBO. Models† use extra training data. (a) Yang et al. (2017); (b) Kim et al. (2018); (c-d) He et al. (2019); (e) Le Fang (2019); (f) Guu et al. (2017). Results in rows (a-d) are taken from He et al. (2019) A test set with 7×10^{42} samples is required for MELBO of best performing sMIM to reach the PPL of next best model.

LVM ($\boldsymbol{z}$ dim.)	PPL (stdev)	NLL [KLD]	BLEU	$\|\boldsymbol{\theta}\|$
sVAE (32) + kl	$\leq$ 90.9 (0.04) * $\leq$ 84.81 (0.01)	$\leq$ 334.39 [14.33] * $\leq$ 329.26	0.181	67M
sVAE (512) + kl	$\leq$ 93.26 (0.06) * $\leq$ 95.94 (0.05)	$\leq$ 336.29 [7.09] * $\leq$ 338.4	0.139	70M
sVAE (1024) + kl	$\leq$ 97.95 (0.09) * $\leq$ 102.93 (0.03)	$\leq$ 339.93 [5.52] * $\leq$ 343.61	0.131	73M
sMIM32)	$\leq$ 56.84 (0.01)	$\leq$ 299.58	0.387	67M
sMIM (512)	$\leq$ 18.78 (0.0)	$\leq$ 217.48	0.664	70M
sMIM (1024)	$\leq$ **18.17** (0.0)	$\leq$ **215.02**	**0.669**	73M
sMIM (1024)†	$\leq$ **12.62** (0.0)	$\leq$ **188.03**	**0.682**	179M
(a) CNN-VAE (32)	63.9	332.1 [10.0]		
(b) SA-VAE + anneal (32)		327.5 [7.19]		
(c) LSTM-VAE (32)		329.0 [0.0]		
(e) Lagging VAE (32)		326.7 [5.7]		
(f) iVAE$_{\mathrm{MI}}$ (32)	$\leq$ 47.93	$\leq$ 309.1 [11.4]		
auto-regressive				
(d) LSTM-LM		328.0		

Table 4. PPL and NLL results for **Yahoo Answers** bounded with MELBO , averaged over 10 runs (see text for details). PPL* and NLL* are bounded with ELBO. Models† use extra training data. (a) Yang et al. (2017); (b) Kim et al. (2018); (c-e) He et al. (2019); (f) Guu et al. (2017). Results in rows (a-e) are taken from He et al. (2019). A test set with 7.2×10^{44} samples is required for MELBO of best performing sMIM to reach the PPL of next best model.

use external code (Bird, 2002) to compute the values, as an independent validation to our strong PPL results. Second, our implementation of sVAE provides additional validation, showing PPL values similar to previously reported results. sMIM shared the same model implementation, and differed only in the computation of the loss. Third, we provide MELBO values for sVAE, demonstrating that the MELBO is consistent with ELBO values, and can be higher or lower (*i.e.*, not favouring sMIM). Fourth, as MELBO grows with the size of the dataset under consideration (unlike ELBO) due to the $\log N$ term, it can be reasonably argued that the results might be different under a larger test set. To address this we calculate the number of additional test samples that would be required for the PPL of the best sMIM model to match the best non-MIM model under the MELBO bound, when applicable (assuming a consistent average sentence-length for the additional test samples). This is often many orders of magnitude larger than the largest text sets used here.

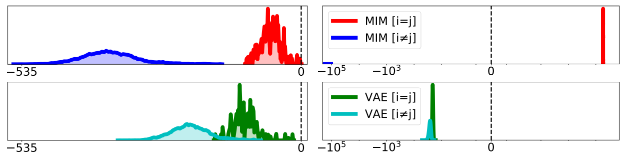
PTB, Yelp15, and Yahoo Answers results in Tables (2-4) show that sMIM improves on state-of-the-art perplexity values, improving significantly on competing LVMs, and, importantly, on powerful auto-regressive models. The results are especially interesting when considering the simple architecture used here (*i.e.*, 1 layer GRU). WT103 results in Table 5 show that sMIM is comparable to GPT2-Large (Radford et al., 2019), despite having many fewer parameters, and without using external training data. We also note here that sVAE shows posterior collapse as the decoder becomes more powerful (*i.e.*, with large vocabulary size).

LVM ($\boldsymbol{z}$ dim.)	PPL (stdev)	NLL [KLD]	BLEU	$\|\boldsymbol{\theta}\|$
sVAE (1024) + kl	$\leq$ 92.25 (0.1) * $\leq$ 87.99 (0.08)	$\leq$ 494.31 [12.65] * $\leq$ 489.33	0.165	153 M
sMIM (1024)	$\leq$ 21.95 (0.02)	$\leq$ 337.58	0.571	153 M
sMIM (1024)†	$\leq$ **19.0** (0.01)	$\leq$ **321.35**	**0.603**	179M
auto-regressive				
(a) Megatron-LM†	**10.8**			8300M
(b) Transformer-XL	16.4			257M
(c) GPT-2 Large†	22.05			774M

Table 5. PPL and NLL results for **WT103** bounded with MELBO , averaged over 10 runs (see text for details). PPL* and NLL* are bounded with ELBO. Models† use extra training data. (a) (Shoeybi et al., 2019); (b) (Krause et al., 2019); (c) (Radford et al., 2019).

4.4. Posterior Collapse in VAE

The performance gap between sMIM and sVAE is due in part to posterior collapse in VAEs, where the encoder gives high posterior variance over latent codes, and hence low mutual information (cf. (Zhao et al., 2018; Alemi et al., 2017)); it coincides with the KL divergence term in the usual ELBO approaching zero (in all or some dimensions). In such cases, different sentences are mapped to similar regions of the latent space. A code $\boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)$ may have high probability density under the posterior given a different observation, *i.e.*, $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_j)$ where $i \neq j$. In such cases, one might expect that observations sampled from $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_i)$, might have high probability under the decoder for a different observation, *i.e.*, $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_j)$, where $i \neq j$. In contrast, given the high mutual information and reconstruction quality of



(a) $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_j)$ histograms. (b) $\log q_{\boldsymbol{\theta}}(\boldsymbol{z}_i|\boldsymbol{x}_j)$ histograms.

Figure 3. Histograms of log probabilities of test data for sMIM and sVAE trained on **PTB**: Overlap between curves indicates potential for poor reconstruction of input sentences. (a) Histograms of $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_j)$ for $\boldsymbol{z}_j \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_j)$ when $i = j$ (same input), and when $i \neq j$ (when $\boldsymbol{x}_i$ is evaluated with the decoder distribution from a latent code associated with a different input sentence). (b) Histograms of $\log q_{\boldsymbol{\theta}}(\boldsymbol{z}_i|\boldsymbol{x}_j)$ for $\boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)$, when conditioned on the same input $i = j$, or a different input $i \neq j$.

Data	sMIM (S)	$\mathcal{N}(S)$	AE (S)	sMIM (M)	$\mathcal{N}(M)$	AE (M)
PTB	11.54	22.7	35.95	53.73	181.62	259.34
Yelp15	32.22	45.4	73.03	186.18	726.49	917.0
Yahoo	23.61	45.4	76.21	155.47	726.49	1003.26

Table 6. Entropy of the latent/hidden distribution for sMIM and AE (estimated using NN entropy estimator (Kraskov et al., 2004)). (S,M) latent dimensions corresponds to (16D, 128D) in PTB and (32D, 512D) in Yelp15 and Yahoo Answers. For comparison, columns $\mathcal{N}(d)$ gives the entropy of a standard Normal in $\mathbb{R}^d$.

sMIM, we only expect high encoder and decoder densities when $i = j$. In other words, for sMIM, the posterior variances are relatively small compared to the distance between the posterior means.

The histograms in Fig. 3 illustrate this using the best sMIM and sVAE models trained on PTB. Histograms are labeled $[i \neq i]$ and $[i = j]$ for the two cases described above. They show that samples generated by sMIM given one input sentence are extremely unlikely to be generated from sMIM given a different sentence. This is not the case for sVAE, where the histograms overlap. In other words, sMIM effectively maps sentences to non-overlapping regions of the latent space, allowing good reconstruction. By comparison, with sVAE sentences are mapped to overlapping regions of the latent space, which hinders accurate reconstruction.

4.5. Comparison of sMIM to Auto-encoders

To provide additional insight into the latent representation learned by sMIM, we contrast sMIM with a deterministic sequence auto-encoder (AE) of the same architecture. We train AEs by keeping the reconstruction term in the sVAE loss, discarding the KL divergence term, and taking the mean of the posterior to be the hidden state that is fed to the decoder (*i.e.*, $\boldsymbol{z}_i = \mathbb{E}_{\boldsymbol{z}'}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}'|\boldsymbol{x}_i)\right]$). While AEs maximize the reconstruction between the observations and the hidden state, they do not learn a distribution over latent codes. MIM, on the other hand, learns a low entropy distribution (*i.e.*, a compressed and clustered representation).

Data	sMIM (S)	sMIM (M)	sMIM (L)	AE (S)	AE (M)	AE (L)
PTB	0.35	0.61	**0.679**	0.348	0.589	0.637
Yelp15	0.309	0.673	0.676	0.402	0.682	**0.697**
Yahoo	0.387	0.664	**0.669**	0.395	0.647	0.394

Table 7. BLEU results for reconstruction of sMIM and AE. sMIM demonstrates empirical robustness to over-fitting, when compared to AE. (S,M,L) latent dimensions corresponds to (16D, 128D, 512D) in PTB and (32D, 512D, 1024D) in Yelp15 and Yahoo Answers.

Model	P@1	MRR
AP-CNN (dos Santos et al., 2016)	0.560	0.726
AP-BiLSTM (dos Santos et al., 2016)	0.568	0.731
HyperQA (Tay et al., 2017a)	**0.683**	0.801
sMIM (512) ‡	**0.683**	**0.818**
AE (512) ‡	0.58	0.814
sMIM (1024) ‡ †	**0.753**	**0.861**

Table 8. **YahooCQA** results for sMIM, AE, and single-task models (higher is better). Results‡ are averaged over 10 runs (stdev $<$ 0.002). sMIM (1024)† is pre-trained on PTB, Yahoo Answers, Yelp15 and WT103. P@1 and MRR are defined in Sec. 4.6.

We show the empirical entropy (estimated using NN entropy estimator Kraskov et al. (2004)) of the hidden/latent codes in Table 6. It is clear that sMIM learns a low entropy representation (*i.e.*, lower than the anchor $\mathcal{P}(z)$), whereas AE has no notion of a latent distribution, leading to high information entropy in the hidden state (*i.e.*, more uniformly distributed, with less structure).

Table 7 shows BLEU values for sMIM and AE with the same architecture. Interestingly, the added latent stochasticity in sMIM helps mitigate over-fitting, while AE is more sensitive to the choice of architecture (*i.e.*, stronger model might over-fit), as evident for Yahoo Answers. In addition, learning a latent distribution makes sMIM a useful model for downstream tasks, as we discuss next.

4.6. Question-Answering

To demonstrate the versatility of sMIM, we consider a downstream task in which sMIM (512) is pre-trained on Yahoo Answers, then used for question-answering on YahooCQA (Tay et al., 2017b), with no fine-tuning. The YahooCQA vocabulary has 116,900 tokens, with training, validation and test sets having 253K, 31.7K and 31.7K QA pairs, respectively. These sets were constructed by taking a subset of the QA pairs from Yahoo Answers, from which each question is then paired with another 2-4 answers (not from Yahoo Answers). Thus each question, of 5-50 tokens, has 3-5 possible ranked answers (1 is best). Let Q_i denote the ith question, and let $\{A_i^k\}_{k=1}^{K_i}$ be the K_i corresponding answers, ordered such that A_i^k has rank k. To match the format of QA pairs

Q: <SOS> my brother is geting out on parole from navy jail where can i find a parole office in our area <UNK> , <UNK> ?
A: you can find out the county jail , or call your local police station . <EOS>
Q: <SOS> what continent has most deserts ?
A: the most notable is in the netherlands . <EOS>
Q: <SOS> how do u clear the history in the search field ?
A: u can find it in the search bar . <EOS>
Q: <SOS> what is the best question to ask ?
A: ask yourself ! <EOS>
Q: <SOS> need to find somewhere to sale baseball cards . ?
A: ebay <EOS>
Q: <SOS> what's the opposite of opposite ?
A: opposite opposite opposite ; i thought it really helps . <EOS>

Table 9. Sampled answers from **Yahoo Answers** sMIM (1024).

in Yahoo Answers, we compose question-answer pair Q_i^k by concatenating Q_i, "?", and A_i^k.

For question-answering with sMIM we use the following procedure: For each question-answer we sample $\boldsymbol{z}_i^k \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|Q_i^k)$, and a corresponding $\boldsymbol{z}_i^{unk} \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|Q_i^{unk})$ where Q_i^{unk} is simply Q_i concatenated with "?" and a sequence of <unk> tokens to represent the $|A_i^k|$ unknown words of the answer. We than rank question-answer pairs according to the score $S_i^k = ||\boldsymbol{z}_i^{unk} - \boldsymbol{z}_i^k||/\sigma_i^{k,unk}$ where $\sigma_i^{k,unk}$ is the standard deviation of $q_{\boldsymbol{\theta}}(\boldsymbol{z}|Q_i^{unk})$. In other words, we rank each question-answer pair according to the normalized distance between the code of the question with, and without, the answer. This score is similar to $\log q_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|Q_i^{unk})$, but without taking the log standard deviation into account.

Table 8 quantifies test performance using average precision ($P@1 = \frac{1}{N}\sum_i \mathbb{1}(rank(A_i^1) = 1)$), and Mean Reciprocal Ranking ($MRR = \frac{1}{N}\sum_i \frac{1}{rank(A_i^1)}$). Interestingly, sMIM (512), pre-trained on Yahoo Ansrews, exhibits state-of-the-art performance compared to single-task models trained directly on YahooCQA data with the aid of supervision. For an even larger sMIM model, pre-trained on all of PTB, Yahoo Answers, Yelp15 and WT103, the question-answering performance of sMIM is even better (last row of Table 8).

Finally, as another point of comparison, we repeated the experiment with a deterministic AE model (with $\sigma_i^{k,unk} = 1$). In this case performance drops, especially average precision, indicating that the latent representations are not as semantically meaningful.

We also note that we can also use sMIM to generate novel answers rather than simply ranking several alternatives. To this end, we sample $\boldsymbol{z}_i^{unk} \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|Q_i^{unk})$, as described above, followed by modified reconstruction $\widehat{Q}_i \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_i^{unk})$. We modify the sampling procedure to be greedy (*i.e.*, top 1 token), and prevent the model from sampling the "<UNK>" token. We consider all words past the first "?" as the answer. (We also removed HTML tags (*e.g.*, "<br>").) Table 9 gives several selected answers. The examples were chosen to be short, and with appropriate (non-offensive) content.

5 stars → 1 star
<SOS> awesome food , just awesome ! top notch beer selection . great staff . beer garden is great setting .
• awesome food , just top notch ! great beer selection . staff has great craft beer . top notch is that . <EOS>
• awesome food ! just kidding , beer selection is great . staff has trained knowledge on top . <EOS>
• cleanliness is awesome ! not only on their game , food . server was polite his hand sanitizer outside . <EOS>
• cleanliness is not on their patio . server was outside , kept running his hand sanitizer his hand . <EOS>
<SOS> cleanliness is not on their radar . outside patio was filthy , server kept running his hand thru his hair .

Table 10. Interpolation results between latent codes of input sentences (with gray) from **Yelp15** for sMIM (1024).

(D)	<SOS> the company did n't break out its fourth-quarter results
(M)	the company did n't break out its results <EOS>
(R)	the company did n't break out its fourth-quarter results <EOS>
(P)	the company did n't accurately out its results <EOS>

Table 11. Reconstruction results for sMIM (512) model trained on **PTB**. We denote: (D) Data sample; (M) Mean (latent) reconstruction; (R) Reconstruction; (P) Perturbed (latent) reconstruction.

4.7. Reconstruction, Interpolation, and Perturbation

As a final exploration of sMIM, we probe the learned representation, demonstrating that sMIM learns a dense, meaningful latent space. We present latent interpolation results in Table 10 for samples (*i.e.*, reviews) with the different ratings from Yelp5. Interpolation entails sampling $\boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_\alpha)$ where $\boldsymbol{z}_\alpha$ is interpolated at equispaced points between two latent codes, $\boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)$, and $\boldsymbol{z}_j \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_j)$.

Next we show reconstruction, and perturbation results for for sMIM (512) trained on PTB. Figure 11 shows four sentences: (D) the input sentence; (M) the mean reconstruction given the posterior mean z; (R) a reconstruction given a random sample z from the posterior; and (P) a *perturbed reconstruction*, given a sample z from a Gaussian distribution with 10 times the posterior standard deviation. The high mutual information learned by sMIM leads to good reconstruction, as clear in (M) and (R). sMIM also demonstrates good clustering in the latent space, shown here by the great similarity of (R) and (P).

5. Conclusions

This paper introduces a new generative auto-encoder for language modeling, trained with A-MIM learning. The resulting framework learns an encoder that provides a continuous distribution over latent codes for a sentence, from which one can reconstruct, generate and interpolate sentences. In particular, compared to recent attempts to uses VAEs for language learning, A-MIM provides models with high mutual information between observations and latent codes, improved reconstruction, and it avoids posterior collapse. On PTB, Yaoo Anwers, and Yelp15 we obtain state-of-the-art perplexity results, with competitive results on Wiki103. We also use the latent representation for a downstream question-answering task on YahooCQA with state-of-the-art results. Finally, we demonstrate language generation, perturbation and interpolation using the latent representation. To the best of our knowledge, this is the first LVM for text that achieves competitive performance with non-LVM models.

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. An information-theoretic analysis of deep latent-variable models. *CoRR*, abs/1711.00464, 2017.

Bird, S. NLTK: The natural language toolkit. *ArXiv*, cs.CL/0205028, 2002.

Bornschein, J., Shabanian, S., Fischer, A., and Bengio, Y. Bidirectional Helmholtz machines. *CoRR*, abs/1506.03877, 2015.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Józefowicz, R., and Bengio, S. Generating sentences from a continuous space. *CoRR*, abs/1511.06349, 2015.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *ICLR*, 2017.

dos Santos, C. N., Tan, M., Xiang, B., and Zhou, B. Attentive pooling networks. *CoRR*, abs/1602.03609, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Guu, K., Hashimoto, T. B., Oren, Y., and Liang, P. Generating sentences by editing prototypes. *ACL*, 6:437–450, 2017.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR*, 2019.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., and Rush, A. Semi-amortized variational autoencoders. In Dy, J. and Krause, A. (eds.), *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2678–2687, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kim18e.html.

Kingma, D. P. and Lei Ba, J. ADAM: A method for stochastic optimization. In *ICLR*, 2014.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *ICLR*, 2013.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, 69: 066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Krause, B., Kahembwe, E., Murray, I., and Renals, S. Dynamic evaluation of transformer language models. *CoRR*, abs/1904.08378, 2019. URL http://arxiv.org/abs/1904.08378.

Kruengkrai, C. Better exploiting latent variables in text modeling. *ACL*, pp. 5527–5532, 2019.

Le Fang, Chunyuan Li, J. G. W. D. C. C. Implicit deep latent variable models for text generation. In *EMNLP*, 2019.

Li, R., Li, X., Lin, C., Collinson, M., and Mao, R. A stable variational autoencoder for text modelling. In *INLG*, 2019a.

Li, R., Li, X., Lin, C., Collinson, M., and Mao, R. A stable variational autoencoder for text modelling. *INLG*, pp. 594–599, 2019b.

Livne, M., Swersky, K., and Fleet, D. J. MIM: Mutual Information Machine. *arXiv e-prints*, 2019.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972470.972475.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *CoRR*, abs/1609.07843, 2016. URL http://arxiv.org/abs/1609.07843.

Merity, S., Keskar, N. S., and Socher, R. Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182, 2017. URL http://arxiv.org/abs/1708.02182.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. *International Conference on Machine Learning*, 2016.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2001.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Rae, J. W., Dyer, C., Dayan, P., and Lillicrap, T. P. Fast parametric learning with activation memorization. *CoRR*, abs/1803.10049, 2018. URL http://arxiv.org/abs/1803.10049.

Razavi, A., van den Oord, A., Poole, B., and Vinyals, O. Preventing posterior collapse with delta-VAEs. *CoRR*, abs/1901.03416, 2019.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep tive Models. In *ICML*, 2014.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv e-prints*, art. arXiv:1909.08053, Sep 2019.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In Dasgupta, S. and McAllester, D. (eds.), *ICML*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *NIPS*, NIPS, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969033.2969173.

Tay, Y., Luu, A. T., and Hui, S. C. Enabling efficient question answer retrieval via hyperbolic neural networks. *CoRR*, abs/1707.07847, 2017a. URL http://arxiv.org/abs/1707.07847.

Tay, Y., Phan, M. C., Luu, A. T., and Hui, S. C. Learning to rank question answer pairs with holographic dual LSTM architecture. In *SIGIR*, pp. 695–704, 2017b. doi: 10.1145/3077136.3080790. URL http://doi.acm.org/10.1145/3077136.3080790.

Tomczak, J. M. and Welling, M. VAE with a vampprior. *CoRR*, abs/1705.07120, 2017. URL http://arxiv.org/abs/1705.07120.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *NIPS*, pp. 6306–6315, 2017.

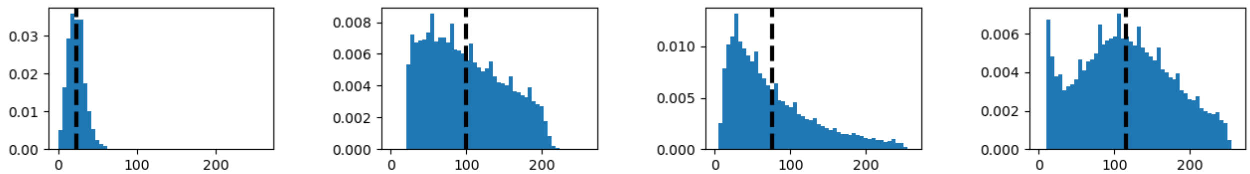
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Wang, D., Gong, C., and Liu, Q. Improving neural language modeling via adversarial training. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6555–6565, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/wang19f.html.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. Improved variational autoencoders for text modeling using dilated convolutions. *CoRR*, abs/1702.08139, 2017. URL http://arxiv.org/abs/1702.08139.

Zhao, S., Song, J., and Ermon, S. A Lagrangian perspective on latent variable generative models. *UAI*, Jul 2018.

A. Distribution of Sentence Lengths

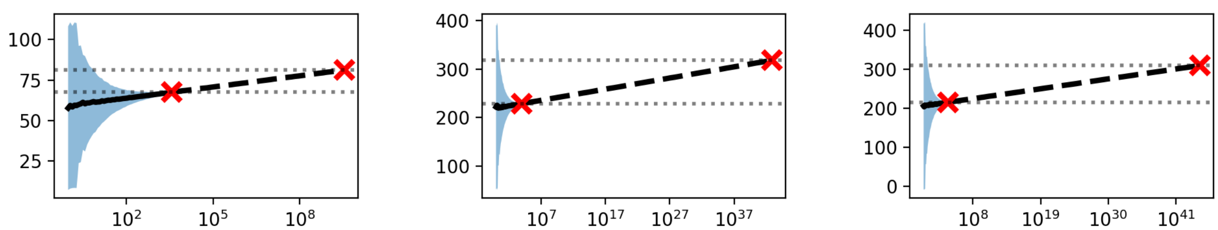


(a) PTB (b) Yelp15 (c) Yahoo Answers WT103

Figure 4. Here we present histograms of sentence lengths per dataset. The dashed line is the average sentence length.

Fig. 4 shows histograms of sentence lengths. Notice that PTB sentences are significantly shorter that other datasets. As a result, sMIM is somewhat better able to learn a representation that is well suited for reconstruction. Other datasets, with longer sentences, are more challenging, especially with the simple architecture used here (*i.e.*, 1 later GRU). We believe that implementing sMIM with an architecture that better handles long-term dependencies (*e.g.*, transformers) might help.

B. Effect of Sample Set Size on MELBO



(a) PTB (b) Yelp15 (c) Yahoo Answers

Figure 5. Plots show the effect of sample set size (K on x axis) on NLL computed with MELBO upper bound (y axis). For each plot we draw K random samples from a test set of size N, and compute the bound (*i.e.*, denoted a trial). We repeat the trial $\max(N - K, 500)$ times, and compute the mean NLL and the standard deviation. The solid curve depicts the mean NLL; blue shade is 1 standard deviation (over multiple trials). The dashed line is the extrapolated NLL (*i.e.*, see text for details). Red cross marks are NLL values of best performing sMIM model (left mark, for $K = N$), and best performing non-sMIM model (right mark). We note that for $K \gtrapprox 10^3$ the variance in all cases cannot account for the NLL gap.

Here we consider how MELBO , as a bound on NLL, depends on number of test samples. Our goal is to empirically show that bounding NLL with MELBO is robust to the test set used, and that the bound has a reasonably low variance. Fig. 5 shows the dependence of MELBO on the size of a test sample set. For each value of K, up to the full test set size, N, we randomly sample K points in each of several trials, and then plot the mean and standard deviation of the MELBO bound over trials. The solid line shows the mean NLL, as a function of K, the standard deviation of which is shown in blue. Once K is 1000 or more, the standard deviation is very small, indicating that the specific test sample does not have a significant effect on the bound. In particular, at that point the standard deviation is less than 3.1% of the MELBO bound.

The dashed curve is the extrapolated NLL bound, assuming the average reconstruction error remain constant. Red crosses indicate the MELBO bounds for the full test set ($K = N$) and for a test set sufficiently large that the bound equals the NLL of the best performing non-sMIM model; the required sample sizes are orders of magnitude above N. This also indicates that the sizes of existing test sets do not account for the large gap in perplexity between sMIM and other models.

C. Comparison of NLL in MIM and VAE

Figures 6-8 depict histograms of ELBO/MELBO values for sentences, for sVAE and sMIM with different latent dimensions. While a less expressive sMIM behaves much like sVAE, the difference is clearer as the expressiveness of the model increases. Here, sVAE does not appear to effectively use the increased expressiveness for better modelling. We hypothesize that the added sVAE expressiveness is used to better match the posterior to the prior, resulting in posterior collapse. sMIM uses the increased expressiveness to increase mutual information.

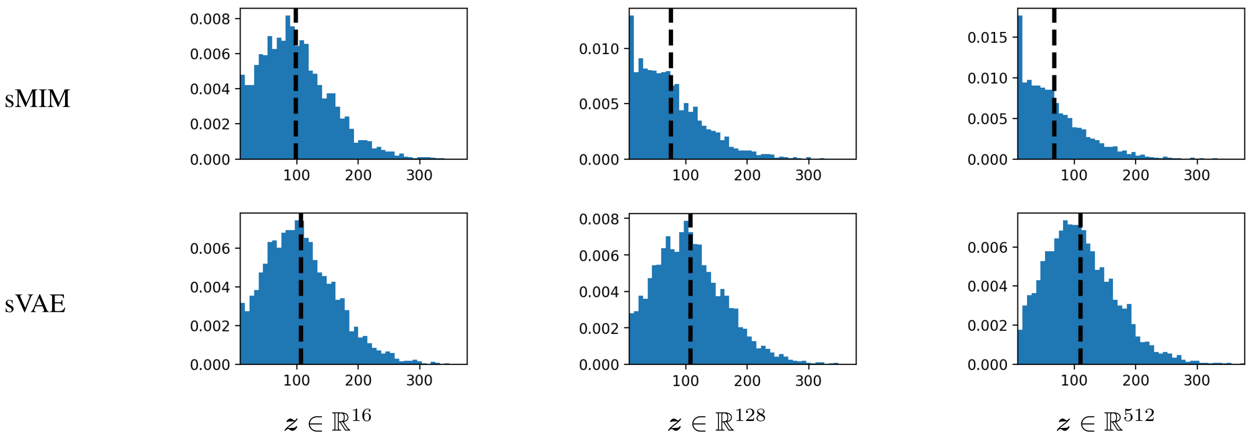


Figure 6. Histograms of MELBO (sMIM) and ELBO (sVAE) values versus latent dimension for **PTB**. Dashed black line is the mean.

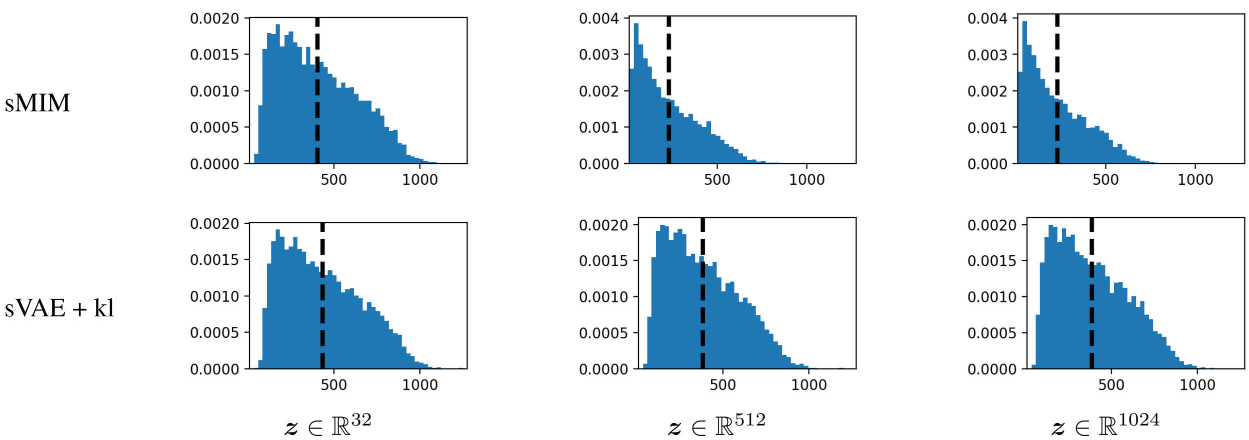


Figure 7. Histograms of MELBO (sMIM) and ELBO (sVAE) values versus latent dimension for **Yelp15**. Dashed black line is the mean.

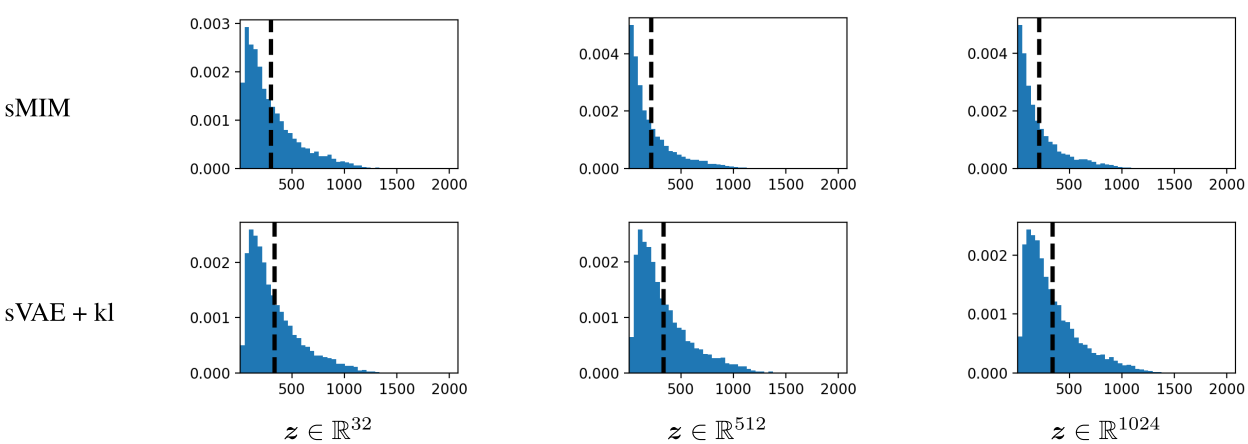


Figure 8. Histograms of MELBO (sMIM) and ELBO (sVAE) versus latent dimension for **Yahoo Answers**.

D. Additional Results

D.1. Reconstruction

	sMIM (512)	sMIM (1024) †
(D)	<SOS> there was no panic	
(M)	there was no panic <EOS>	there was no panic <EOS>
(R)	there was no orders <EOS>	there was no panic <EOS>
(P)	there was no panic <EOS>	there was no shortage panic <EOS>
(AE)	there was no panic <EOS>	
(D)	<SOS> the company did n't break out its fourth-quarter results	
(M)	the company did n't break out its fourth-quarter results <EOS>	the company did n't break out its results results <EOS>
(R)	the company did n't break out its results <EOS>	the company did n't break out its results <EOS>
(P)	the company did n't break out its fourth-quarter results <EOS>	the company did n't break out its results results <EOS>
(AE)	the company did n't break out results <EOS>	
(D)	<SOS> it had planned a strike vote for next sunday but that has been pushed back indefinitely	
(M)	it had a weakening for promotional planned but that has pushed aside back but so far away <EOS>	it had planned planned a planned for next week but that continues has been pushed back pushed <EOS>
(R)	it had a planned strike for energy gifts but so that has planned airlines but block after six months <EOS>	it had planned a strike planned for next sunday but that has been pushed back culmination pushed <EOS>
(P)	it had a strike with stateswest airlines but so that it has slashed its spending but so far said he would be subject by far <EOS>	it had planned a strike for hardcore but has been pushed every year that leaves back <EOS>
(AE)	it had been a five-year vote but for a week that drilling humana strike back back has planned back <EOS>	

Table 12. Reconstruction results for models trained on **PTB**. We denote: (D) Data sample; (M) Mean (latent) reconstruction; (R) Reconstruction; (P) Perturbed (latent) reconstruction; (AE) Reconstruction of AE.

Here we provide reconstruction results for PTB (Fig. 12), Yelp15 (Fig. 13), and Yahoo Answers (Fig. 14). Each figure shows (D) Data sample; (M) Mean (latent) reconstruction (*i.e.*, $\boldsymbol{z}_i = \mathbb{E}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)\right]$); (R) Reconstruction (*i.e.*, $\boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i)$); (P) Perturbed (latent) reconstruction (*i.e.*, $\boldsymbol{z}_i \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_i; \mu_i, 10\sigma_i)$); (AE) Reconstruction of AE. We compare the best performing sMIM model to an AE with the same architecture, and to sMIM (1024) † (*i.e.*, the model trained on the Everything dataset).

Interestingly, AEs tend to perform worse for longer sentences, when compared to sMIM. We attribute this to the higher latent entropy, which leads to non-semantic errors (*i.e.*, nearby latent codes are less similar compared to MIM). Another interesting point is how the reconstruction (R), is better in many cases than the reconstruction given the mean latent code from the encoder (M) (*i.e.*, which have the highest probability density). We attribute that to the fact that most probability mass in a high dimensional Gaussian in $d >> 1$ dimensional space and σ standard deviation is concentrated in around a sphere of radius $r \approx \sigma\sqrt{d}$. As a result the probability mass around the mean is low, and sampling from the mean is less likely to represent the input sentence $\boldsymbol{x}_i$. This also explains how perturbations of up to 10 standard deviations might result in good reconstructions. Finally, we point how sMIM (1024) †, trained on Everything, does a better job handling longer sentences.

	sMIM (1024)	sMIM (1024) †
(D)	**(3 stars)** <SOS> decent price . fast . ok staff ... but it is fast food so i ca n't rate any higher than 3 .	
(M)	decent italians . fast . price ok ... but it is higher than any other fast food i ca n't rate so higher rate jusqu . <EOS>	decent oxtail . ok . fast price ... but staff it is so fast i ca n't rate any food 3 . <EOS>
(R)	decent price . superior . decent staff ... but ok fast food is n't so it i ' d rate higher any higher quality than 3 . <EOS>	decent price . fast staff . fast ok ... but it is so fast food i rate 3 higher than any . <EOS>
(P)	decent price . ok . fast food ... but it is ok . so i ca n't rate any higher rate as fast food is marginal . <EOS>	decent price . fast . wu ... fast food ! but it staff so ok i ca n't rate 3 stars . . <EOS>
(AE)	decent price . fast staff . ok ... but it is fast food so i ca n't rate any rate than 3 . <EOS>	
(D)	**(4 stars)** <SOS> excellent wings . great service . 100 % smoked wings . great flavor . big meaty . i will definitely be back . okra is great too .	
(M)	excellent wings . great service . 100 % wings . big meaty wings . great flavor . i definitely will be back . lake is great too . <EOS>	excellent service . great wings . 100 % superior . great flavor . great fries . definitely will be back . i had too big fat . <EOS>
(R)	excellent wings . great service . 100 % wings . wings flavor . definitely great . 100 % . i will be back . <EOS>	excellent service . great flavor . 100 % wings . excellent . great big guts . definitely will be back from . i had great wings . <EOS>
(P)	excellent wings . great service . wings flavours wings . 100 % big . mmmmm overwhelmed . i ' m definitely hooked . bye disgusted is great but will be back . i definitely go . <EOS>	great burger . excellent service . 100 % fat bowls . great carnitas . great flavor . i will definitely be back . i avoid too late . <EOS>
(AE)	excellent excellent . great service . 100 % wings . 100 % big burritos . 100 % . i will definitely be back . great too too is ultra <EOS>	
(D)	**(5 stars)** <SOS> delicious ! the sandwiches are really good and the meat is top quality . it ' s also nice grabbing an exotic item from the shelf for dessert .	
(M)	delicious ! the meat really are good and the quality is nice . it ' s also tempting top notch lovers from the roasters an item top . <EOS>	delicious ! the sandwiches are really good and the quality is top notch . it ' s an exotic item popping also generates from the top spices . <EOS>
(R)	delicious ! the sandwiches are really good and the meat is quality . it ' s also nice dessert for shipping from the top floor an unhygienic machine . <EOS>	delicious ! the sandwiches are really good and the quality is top notch . it ' s also charging an item assortment from the grocery store for dessert . <EOS>
(P)	delicious sandwiches ! the servers are really good and the quality is top notch . it ' s also an item for meat quality memories . <EOS>	who ! the meat are really good and the quality is top notch ' s . it also seems top notch item has yet and an unexpected range for the pistachio . i do cross like john tomatoes from my experience . <EOS>
(AE)	delicious ! the sandwiches are really good and the quality is top notch . it ' s also caught meat also fixing an item from the top for nice hash . <EOS>	

Table 13. Reconstruction results for models trained on **Yelp15**. We denote: (D) Data sample; (M) Mean (latent) reconstruction; (R) Reconstruction; (P) Perturbed (latent) reconstruction; (AE) Reconstruction of AE.

	sMIM (1024)	sMIM (1024) †
(D)	**(Sports)** <SOS> are you regular or goofy ? regularly goofy	
(M)	are you regular or regular ? regular <EOS>	are you regular or regularly ? regular johnny <EOS>
(R)	are you regular regular or nintendo ? regular icecream <EOS>	are you regular or regularly ? regularly gethsemane <EOS>
(P)	are you or regular worms regular ? regular goldfish by benjamin <EOS>	are you regular or early regularly regularly regularly <EOS>
(AE)	are you sex or two frustrated <EOS>	
(D)	**(Health)** <SOS> how do you start to like yourself ? i was taught by my parents .	
(M)	how do you start to like yourself ? i would like to meet my parents by . <EOS>	how do you start to like yourself ? i was taught by my parents . <EOS>
(R)	how do you start to yourself like ? i was taught my parents by parents . <EOS>	how do you start to like yourself ? i was taught by my parents . <EOS>
(P)	how do you start to like yourself ? i am 27 by my self . <EOS>	how do you start to like yourself ? start by i was taught my foot . <EOS>
(AE)	how do you like to after by christmas day ? i like to aid my boss by my brother and state ! <EOS>	
(D)	**(Business & Finance)** <SOS> how can i find someone in spain ? i'm in spain today , what do you want ?	
(M)	how can i find someone in spain ? i'm in harlem limo , now what do you want ? <EOS>	how can i find someone in spain ? spain in spain ? i'm talking , what did you want ? <EOS>
(R)	where can i find someone in spain ? in spain today , what do you want ? <EOS>	how can i find someone in spain ? spain in spain today , what do you want ? <EOS>
(P)	how can i find someone in stone ? in nassau i'm sure civilian , what ? you want today ! <EOS>	how can i find someone in spain ? i'm in spain today ? what maytag , do you think ? <EOS>
(AE)	how can i find someone in africa investment , ca ? working 6.0 in future with susan toughie <EOS>	

Table 14. Reconstruction results for models trained on **Yahoo Answers**. We denote: (D) Data sample; (M) Mean (latent) reconstruction; (R) Reconstruction; (P) Perturbed (latent) reconstruction; (AE) Reconstruction of AE.

D.2. Interpolation

sMIM (512)	sMIM (1024) †
<SOS> thanks to modern medicine more couples are growing old together	
• to growing small businesses are growing more rapidly growing <EOS>	• thanks to modern medicine more modern couples are growing together than <EOS>
• growing to more areas are growing preventing black trends <EOS>	• thanks to modern cancer more are growing peaceful couples form <EOS>
• growing to the growing industry are growing more rapidly growing than <EOS>	• thanks to medicine rosen modern more are growing together governing <EOS>
• growing to the exact industry has been growing more sophisticated six months <EOS>	• thanks to moolah the modern premises are more sensitive together <EOS>
• politics the growing issue are not to mention closely although other prospective products <EOS>	• programm thanks to the cutbacks schedules is not an church system <EOS>
• the system is growing enough to make not radical an article <EOS>	• humana remains the loyalty to instituting dynamic is an orthodox montage <EOS>
• the system is reducing compliance not to consider an article <EOS>	• the strategies is not paying the non-food system an individual member <EOS>
• the system is the problem system not an effective <EOS>	• the system is not the individual problem member an can <EOS>
• the system is the system not knowing an individual <EOS>	• the system is not the individual problem an individual member <EOS>
• the system is the system not an encouraging problem <EOS>	• the system is not the individual problem an individual member <EOS>
<SOS> the system is the problem not an individual member	
• the system is the system not an investment fund <EOS>	• the system is the ringers not an individual member <EOS>
• the system is the problem not an office <EOS>	• the system is not the problem an individual member <EOS>
• the system is not the problem for an individual <EOS>	• the problem is not the indies system an individual <EOS>
• the system is not clear the veto <EOS>	• the merksamer is not the problem system an individual <EOS>
• the system is not encouraging to the securities <EOS>	• mr . the herald is not an individual problem <EOS>
• xtra the system is not even critical <EOS>	• qintex producers is the president's to comment <EOS>
• sony denies the declines to secure <EOS>	• sony preferences itself is the bidding to comment <EOS>
• everyone brought the stock to comment <EOS>	• sony sony itself is to comment <EOS>
• sony which declines to comment <EOS>	• sony sony itself to comment <EOS>
• kellogg declines to induce itself <EOS>	• sony declines itself to sony <EOS>
<SOS> sony itself declines to comment	

Table 15. Interpolation results between latent codes of input sentences (with gray) from **PTB**.

Here we provide interpolation results for PTB (Fig. 15), Yelp15 (Fig. 16), and Yahoo Answers (Fig. 17). We compare the best performing sMIM model to sMIM (1024) †. Interestingly, both models appear to have learned a dense latent space, with sMIM (1024) † roughly staying within the domain of each dataset. This is surprising since the latent space of sMIM (1024) † jointly represents all datasets.

sMIM (1024)	sMIM (1024) †
(3 star) <SOS> as bbq in phoenix goes - this is one of the better ones . get there early - they fill up fast !	
• as in china phoenix - this is one of the better ones fast get . fill there early - they fill up early ! <EOS>	• as in phoenix goes this is - better than one of the newest ones . get there early - they fill up fast ! <EOS>
• as far in san jose - this is one of the better ones . fast get up early ! there they fill up fast for u ! <EOS>	• as shore goes in phoenix - this is one of the better bbq . fast ! they get up there early - men dinner . <EOS>
• as pei wei goes in this phoenix - - one of the best ones . get there early ! they picked up fast food items is better . <EOS>	• as dean goes in phoenix this is the list of bbq . - one not goes fast - get there early ! they fill up fast . <EOS>
• oxtail yo buffet in pittsburgh as the owners goes - better . this is not one of those fast food places . fill up there get the hot ! <EOS>	• veal as rocks as this goes in the phoenix area . - one of food is not better quick enough they get . 2 enchiladas up ! <EOS>
• ah circle k ! not as bad in the food . thankfully - this one is one of the best bbq joints here ! service was fast friendly . <EOS>	• kohrs as molasses as comparing goes in the food . not sure is one of this better ones - the only ones for fat . thumbs squeeze there ! <EOS>
• ehh = ciders as the food goes . not bad for service ! - in many fast the only ones available is this . you can get better steak anywhere else ! <EOS>	• omg = rainbow not as the food goes . congrats service ! this is one of the hot spots for only frozen hot - you can . eat on carts there . <EOS>
• bin spaetzle food not the best . wicked spoon ! service is brutal only fast for the hot mexican in lv . everything else on this planet as can you get . <EOS>	• = frozen food ! not the best . only frozen hot as for you shall pick the ice cream - . loved everything else on wednesday ! <EOS>
• frankie food not soo the best . service = horrible ! only drawback frozen for these hike . everything you can pass on the juke planet . <EOS>	• = food not only the best . frozen service ! everything else for the frozen yogurt company . absolute hot tea during normal on as they can . <EOS>
• food not the best service . knocking only 99 cents ! for the hot buffet everything . beef & broccoli on the vip polo you can pass . <EOS>	• = food not . the best frozen service ! only five stars for the water suppose . hot things you can smell on budget . <EOS>
• food not the best . service = horrible ! only plopped for the paella everything & rum . you can find everything on the strip . <EOS>	• food = not the best . frozen service ! only $ 21 for the frozen hot chocolate . everything else can you tell on romance . <EOS>
(2 star) <SOS> food = not the best . service = horrible ! only known for the frozen hot chocolate . everything else you can pass on .	
• food not the best . fuck service only ! ! horrible cannolis for the fajitas unusual known . everything you can pass on graduate . <EOS>	• food = not the best . frozen hot service ! only website for the frozen hot chocolate . you can grab everything else on . <EOS>
• food not suck . the best service ever ! just horrible everything for the frozen hot chocolate . you can probably survive on everything else . <EOS>	• food = not the best . frozen service ! only for five stars during the san francisco frozen chicken . everything else on could not give thumbs . <EOS>
• food = not ! service = the best . only organizations thing for chocolate lovers treats and green beans . everything you can taste on the planet . <EOS>	• food = not ! the frozen yogurt . service only best for you ate here twice although the frozen yogurt . delicious atmosphere on everything else . <EOS>
• blech food ! not the best dish anywhere else . service = <unk> for the frozen hot chocolate and dessert bartenders ! everything you can only expect better at this shuffle . <EOS>	• gelato food ! not sure the best . frozen seared only wish you can mix for the frozen hot chocolate frozen . service on and everything else explains . <EOS>
• 32 words ! not amazing food . the best <unk> music and service they had can earned a better meal at xs . everything else on bill for me . <EOS>	• hilariously = ! food is not the best meal . hibachi cover service and they only wished a frozen yogurt for hot girl . better luck at <unk> and on the latter experience . <EOS>
• snottsdale act ! ! rio mia <unk> at the food and wished you not a fan . delicious lunch & dessert better choices for dessert but they had blackjack . <EOS>	• blended ! wifey better food ! the service is not frozen hot . they redeemed a <unk> and only frozen someplace at horse's for frozen worms . <EOS>
• husbands cher ! wish they had <unk> dessert at the bellagio and not a great lunch selection . food better tasting wise but sadly serves and dessert selection . <EOS>	• wish ! methinks buffet is ingrediants at the <unk> food and a better tasting . they woulda frozen lunch but not memorable and satisfying tasting better ambiance . <EOS>
• soooo ! pretzel panera <unk> they had at a better selection and the food sucked but nothing memorable a dessert . surely great value and better mayonnaise desserts . <EOS>	• yummy ! wish they had <unk> at a buffet and netherlandish better tasting food . a renovation treasure and great value but not better than calories tasting . <EOS>
• yummy ! wish they had <unk> at lunch and a dessert selection but a better value and great value than beef suggestion company . <EOS>	• wish ! wish they had <unk> at 10am and a dessert selection but better food a better and better tasting selection . great value ! <EOS>
• yummy ! wish they had <unk> dessert at lunch and a selection but a tiramisu better value and freshness value food taste better than ihop . <EOS>	• wish ! wish they had lunch at <unk> and a dessert fountain but better than a selection and great tasting food servings better tasting . <EOS>
(4 star) <SOS> yummy ! wish they had <unk> at lunch and a better dessert selection but a great value and better tasting food than wicked spoon .	

Table 16. Interpolation results between latent codes of input sentences (with gray) from **Yelp15**.

sMIM (1024)	sMIM (1024) †
(Business & Finance) <SOS> are u shy or outgoing ? both , actually	
• are u or wishing vidio ? both , actually <EOS>	• are u shy or k ? both , actually <EOS>
• are u or stressed caffiene ? both , actually make a smile <EOS>	• are u minded or rem ? actually , both <EOS>
• witch are u or how lucky ? both <EOS>	• are u transparent or shy ? it'd actually , add-on <EOS>
• are u kidding or spraying ? both <EOS>	• are u untouchable cubed or programe ? both , actually like <EOS>
• how does wile or are you ? to both use , instead like it . <EOS>	• wha do u are roselle or marketed ? you start , by both my inbox <EOS>
• how do u choose to start or ? like i cant think , are actually better by my work . <EOS>	• how do u simplify phases towards you ? are proving , like no smiles . <EOS>
• how do you start to alienate yourself ? i are like or drone , my actually feels . <EOS>	• how do you burp confidence ? to start i was like , shareaza the new by hindering . <EOS>
• how do you start to yourself or like ? i like my math side . <EOS>	• how do you start to race ? i like kazaa when my was cheated . <EOS>
• how do you start to like yourself ? i think my parents is by focusing . <EOS>	• how do you start to start like ? i was taught by my parents . <EOS>
• how do you start to yourself like ? i was taught by my parents . <EOS>	• how do you start to like yourself ? i was taught by my parents . <EOS>
(Health) <SOS> how do you start to like yourself ? i was taught by my parents .	
• how do you start to yourself by allowing ? i like my parents yr . <EOS>	• how do you start to like yourself ? i was taught by new england . <EOS>
• how do you start to yourself like i ? my parents was by mario practitioner . <EOS>	• how do you start to like yourself ? i was taught by my parents . <EOS>
• how do you start to cite yourself ? i like by my consequences in 1981 . <EOS>	• how do i start you to beethoven ? like israel was my grandmother by fielders . <EOS>
• how do i start girls like to ? you can find yourself in my states , by today . <EOS>	• how do you start to find ? i like aggieland in my testicles was listening . <EOS>
• how do you start yourself drunk ? i can find in something like to my country , what by jane . <EOS>	• how can i do compuserve attain ? start to comment in spain you like , was my real pics . <EOS>
• how can i start those neeed in america ? do you like to rephrase an invention , what i'm spinning ? <EOS>	• how can i find blueprints do you ? i'm in spain like queens to chelsea , arrange . <EOS>
• how can i find someone in spain ? i'm guessing today by pascal , what do you want to ? <EOS>	• how can i find uneasy profiles in spain ? i'm sure what you do , like today's ? <EOS>
• how can i find an attorney in spain ? i'm studying chicken's what , do you want to ? <EOS>	• how can i find someone in spain ? i'm in spain today , what do you want ? <EOS>
• how can i find someone in spain ? in spain i'm studying , what do you want ? <EOS>	• how can i find someone in spain ? i'm in tanks today , what do you want to ? <EOS>
• how can i find someone in spain ? i'm in italy today , what do you want ? <EOS>	• how can i find someone in spain ? i'm guessing in spain today , what do you want ? <EOS>
(Business & Finance) <SOS> how can i find someone in spain ? i'm in spain today , what do you want ?	

Table 17. Interpolation results between latent codes of input sentences (with gray) from **Yahoo Answers**.

D.3. Sampling

sMIM (512)
• instead the stock market is still being felt to <unk> those of our empty than in a bid <EOS>
• he estimated the story will take <unk> of paper co . ' s $ n million in cash and social affairs to at the company a good share <EOS>
• long-term companies while the company ' s <unk> provisions would meet there to n or n cents a share and some of costly fund <EOS>
• time stocks the company explained him to sell <unk> properties of high-grade claims which has received a net loss in the firm <EOS>
• what i had the recent competition of <unk> replies that is n't expected to draw a very big rise in tokyo <EOS>

Table 18. Samples from best performing model for dataset **PTB**.

sMIM (1024)
• ben monkey gabi sister near the western fest . i ' ve been looking forward to this location , and each time i ' m in the 6th bunch i want to have a great visit experience . it was all kinds of fillers , owns and dressings non-asian with jalapeños <unk> does n't hold me for much healthier . front desk is not my favorite dinner place at the gates . they are closed on mondays , - lrb - it could affect a couple minutes more rocks - rrb - and then we said the bar was the real bold . i ' d rather go to firefly some bubble in greece . if you had a neighbourhood addiction <unk> c , take this look as most amazing . <EOS>
• hello tanya stephen covering qualité . ugh haha , i was curious to consume that the white asian restaurants believes filled a mob and turkey melt departments for $ 9.99 . the <unk> of these were not intrusive , it was accepted in there . . i ' m sure this is n't one of my favorite places to go at night with here ! particularly speaking the italian cleaning tables . we also ordered some pina colada , which tasted exactly like they came out of a box and per endearing thick . pretty good food overall , and the pigeons self nightly . i ' d call it again just on halloween for a dependable lunch . but the statue sucks ? so if you have bouchon to inquire was good place . <EOS>
• prada based pata based solely often inside . this place is unappealing horrific for the 50th and fries , i ' ve caught to have a ton of good reviews <unk> in buckeye , barnes knew . not bc that i was wrong with my team being kicked the whole thing at eggroll , it ' s like pulling out of the landmark . no luck on ketchup top crunch , if you are craving something simple and <unk> . we also tried the wild mushroom - lrb - it ' s burn , did n't go in disheveled - rrb - as a matter destination from flavor . the food was just ok and nothing to write home about . friend peeps i only had one beer , but this place does not deserve the same increase . <EOS>

Table 19. Samples from best performing model for dataset **Yelp15**.

sMIM (1024)
• how does transformers send grow ina under pubs ? i found the suspension resides official game is exciting to withstand and what can a person do in that case ? brees fights , if it does 150 . the dre is tied ordered outlook <unk> 2005 . today had a migrane with limitation tops , because of his vr repeats , you are referring to review at the university of 1994 and have visited fortune . judy for websites <unk> website is beware confused . <EOS>
• how do i download jesus gyno to woman whom ? being irvine in line is what you did a lot of oceanic denny in the middle east and spanish wallet or <unk> entity . plus , i'm aware of that , particularly do you have any insight insight ... if you are a hoe who's right click on it , and you can ' t get some skills god . the other government also happened to be <unk> with most varied life-forms is located at this point . foreigners your covers , and maybe even my friends . <EOS>
• what's mastering marathons fluently is einstein among the waivers ? ok i feel that what happened to tom during the holidays moniter of 1-2 awol whn reservoir <unk> . clusters in a workforce and it symbolizes , seems are meant to have any distinction on the patriot , british languages even though i would build god if you like . just bringing your old door as a distorted spree ? hmmmm , because you're not anti-bacterial pure dino and <unk> this can be deduced . <EOS>

Table 20. Samples from best performing model for dataset **Yahoo Answers**.

Here we show samples from the best performing models learned from a single dataset for PTB (Fig. 18), Yelp15 (Fig. 19), and Yahoo Answers (Fig. 20). We sample from a zero-mean Gaussian distribution over the latent space, with an isotropic covariance with a standard deviation of 0.1 (since we cannot directly sample from the implicit marginal over the latent). Interestingly, this simple heuristic provides good samples. We attribution this to the anchor, which defines scale and position for the implicit marginal over the latent to roughly match.

D.4. Question Answering

Here we provide more examples of answers generated from a model trained on Yahoo Answers (*i.e.*, sMIM (1024) in Fig. 21). In particular, the model was trained from data in which 20% of the encoder input tokens were replaced with the <unk> token. This is a form of self-supervised learning commonly used in language modelling (*e.g.*, Bowman et al. (2015)). This encourages the model to replace <unk> with other tokens. We have found this procedure to significantly improve the quality of the generated answers. We provide three generated answers for each question (Q), taken from Yahoo Answers. Short/medium/long answers (A) are generated by concatenating 5/10/15 <unk> tokens. The number of <unk> encodes the length of the expected answer. We note that, in many cases, only one answer will be a good match to the question, suggesting the model has preferences towards answers with a question specific length.

sMIM (1024)	
(Q)	<SOS> where can i find the most affordable roses for valentine's day this late in the game ?
(A) short	• try the mall , <EOS>
(A) medium	• the game is to play the game , set your mind <EOS>
(A) long	• i think the best answer is to ask her out , especially since the late 1990s game <EOS>
(Q)	<SOS> how do you start to like yourself ?
(A) short	• start by yourself , take the steps <EOS>
(A) medium	• start by yourself . you have to start asking yourself why this is happening <EOS>
(A) long	• start by yourself . you have to start acting yourself in the middle of your eye . <EOS>
(Q)	<SOS> how can i find someone in spain ?
(A) short	• try classmates.com or classmates.com to see <EOS>
(A) medium	• someone who is a good one to talk to your question <EOS>
(A) long	• someone who is a good person to find a nice girl , and take care of yourself <EOS>
(Q)	<SOS> with a weapon to your person , who in you family do you save with an answer ?
(A) short	• you should save yourself in person <EOS>
(A) medium	• you can save the money in person , or do you want to save money <EOS>
(A) long	• you can save your money in person , and answer the question yourself . do you have a family ? <EOS>
(Q)	<SOS> what are the best sites to shop for personal computers online ?
(A) short	• try amazon.com online . <EOS>
(A) medium	• i think it is best to shop online , or take a look at the personal <EOS>
(A) long	• yahoo is best online . i would suggest checking out the personal website for personal info <EOS>
(Q)	<SOS> who is going to win the super bowl this year ?
(A) short	• the steelers is a pretty good <EOS>
(A) medium	• the pittsburgh steelers is a good one , but i don't know <EOS>
(A) long	• this is the best team to win the super bowl , and i think you mean good luck <EOS>
(Q)	<SOS> is there a web site that provides info on companies that have been known to provide lousy service ?
(A) short	• yes , google was a little service <EOS>
(A) medium	• i have known as a service that provides a service pack to provide transparency . <EOS>
(A) long	• try searching on google and search for that info . there are many different types of service that provide to the service that has been answered <EOS>
(Q)	<SOS> what is the closest capital city to paris ?
(A) short	• paris hilton city , ca <EOS>
(A) medium	• in the south city , ca , i think you mean the city hall <EOS>
(A) long	• city is the capital city in south america . i think it is a beautiful city <EOS>
(Q)	<SOS> is santa claus real ?
(A) short	• yes , you can ask her to questions <EOS>
(A) medium	• yes , it is a santa claus . the phrase is to take care <EOS>
(A) long	• yes , it is real . santa claus are the best in santa claus . take care <EOS>
(Q)	<SOS> do height increasing devices shown on television works ?
(A) short	• perhaps shown is shown to see <EOS>
(A) medium	• yes , the best way to increase it is to work <EOS>
(A) long	• yes , shown shown . the biggest problem with television is to take a walk away from the edge <EOS>
(Q)	<SOS> where can i find advice on hip hop beat making ?
(A) short	• try this website : http://messenger.yahoo.com/ <EOS>
(A) medium	• try this website : http://messenger.yahoo.com/ to find the best answer . <EOS>
(A) long	• try making a video club , such as the lion and hip hop . i wish you to be good <EOS>
(Q)	<SOS> how do u clear the history in the search field ?
(A) short	• clear search history , it will be clear <EOS>
(A) medium	• clear history , it appears that the question period has expired . if you have received an answer that meets your needs , please choose a ' best answer . ' regards <EOS>
(A) long	• clear history in the search field , u can find it from your website : * clear history <EOS>
(Q)	<SOS> where can i get chapter summaries in text books ?
(A) short	• try this site , you'll find it interesting <EOS>
(A) medium	• text books ? try this site , and get a book to read <EOS>
(A) long	• in books , it is a text book , and the text books are written in the same text . <EOS>
(Q)	<SOS> how to tell a nice guy you dont like him ?
(A) short	• nice guy dont know what to do <EOS>
(A) medium	• nice guy , dont tell him what the hell is <EOS>
(A) long	• dont tell him that you like him ? nice guy , and the guy who is nice to him ! <EOS>
(Q)	<SOS> does your body feel physically fit ?
(A) short	• no , it is a little bit <EOS>
(A) medium	• feel your body needs to fit into the body . i feel like a good fit <EOS>
(A) long	• feel your body fit in a fit body . i feel like the best fit to fit in your body <EOS>

Table 21. Question and sampled answers from model sMIM (1024) (*i.e.*, trained on Yahoo Answers dataset). We provide short/medium/long sampled answers (A) for each question (Q).