A VAE for Transformers with Nonparametric Variational Information Bottleneck

James Henderson 1 Fabio Fehr 1 2

1idiap Research Institute 2École polytechnique fédérale de Lausanne

Motivation and Contributions

Summary
We propose a Nonparametric Variational Information Bottleneck (NVIB) layer to regularise a Variational AutoEncoder (VAE) for Transformers. We derive a Bayesian nonparametric formalisation of attention-based latent representations as mixture distributions and of the attention mechanism as Denoising attention. This is then used to regularise and access the posterior distribution returned from the Transformer encoder.

Motivation: This work is motivated by the empirical success of Transformers and inspired by connections between attention-based representations and Bayesian nonparametrics.

Impact: The domain includes: model regularisation and sparsity; deep probabilistic generative models; and learning distributional latent representations for attention-based models.

Contributions:
1. Denoising attention: A Bayesian nonparametric interpretation of the attention mechanism.
2. Nonparametric VIB: A variational Bayesian framework for regularising and generating from Transformer embeddings.

Intuition
To build intuition, the properties of attention and nonparametric distributions are compared:

<table>
<thead>
<tr>
<th>Attention Distributions</th>
<th>Nonparametric Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable number of vectors</td>
<td>Variable number of mixture components</td>
</tr>
<tr>
<td>Permutation invariant</td>
<td>Exchangeable</td>
</tr>
</tbody>
</table>

Model
We define prior and posterior distributions over our attention-based latent representations:

Prior
\[ F = \sum_{n=1}^{K} q_{n} \delta_{x_{n}} \]

Posterior
\[ F = \sum_{n=1}^{K} q_{n} \delta_{x_{n}} \]

- **Bounded Dirichlet Process** prior \( p(F) \)
- \( F \sim BDP(\alpha_{0}, \alpha) \)

- **Bounded Dirichlet Process** posterior \( q(F | x) \)
- \( F \sim BDP(q_{n}) \)

- **Dirichlet Process** prior \( \pi \)

- **Dirichlet Process** posterior \( \pi_{n} \)

Variational Information Bottleneck Loss

The VIB loss minimises the log-likelihood of the observation \( x \) where \( x \) is the input text:

\[ \log p(x) \geq \sum_{i} \log p(z_{i} | x) - D_{KL}(p(z_{i} | x) \| p(z_{i})) \]

- \( L_{B} \) - Reconstruction loss
- \( L_{D} \) - KL-divergence for Dirichlet weights \( \pi \)
- \( L_{G} \) - KL-divergence for Gaussian vectors \( Z \)

Denoising Attention

Regress the scaled dot product attention function such that all operations are in \( Z \) space:

\[ \text{Attention}(u, Z) = \text{softmax} \left( \frac{u^{T} W^{v} Z W^{k}}{\sqrt{d}} \right) Z W^{v} \]

**Figure 1. Standard attention.**

where \( u = u^{T} W^{v} (W^{k} Z)^{T} \) is projected into the space of \( Z \). We interpret \( Z \) as specifying a probability distribution over a vector space, and define a function over such probability distributions which, when given any vector \( u \), always returns the result vector \( \text{Attention}(u, Z) \).

\[ \text{Attention}(u, Z) = \text{softmax} \left( \frac{u^{T} Z}{\sqrt{d}} \right) Z \]

\[ F_{Z} = \sum_{u} \text{exp} \left( \frac{u^{T} z}{\sqrt{d}} \right) I_{u} \]

\[ D_{\text{Att}}(u | F_{Z}) = \int f_{g}(w) p(v | w) dv \]

**Figure 2. Denoising attention.**

where \( I_{u} \) is an impulse distribution at \( u \), \( f_{g} \) is the probability density function for distribution \( g \), and \( v = W^{k} Z \) is a multivariate Gaussian function with diagonal variance of \( \sqrt{2} \). Figure 2 shows that \( D_{\text{Att}}(u | F_{Z}) \) can be interpreted as query denoising.

Nonparametric Variational Information Bottleneck

A VAE for Transformers using Nonparametric Variational Information Bottleneck (NVIB), which regularises the attention-based representations between the Transformer encoder and decoder.

**Figure 3. Nonparametric VAE model, with its NVIB layer.**

- The encoder estimates the posterior pseudo-observation parameters given an input text \( u \).
- The Dirichlet Process is sampled by sampling component weights and vectors separately.
- The decoder reconstructs \( x \) using denoising attention over a sample \( F \) from the posterior.
- The latent distribution is regularised by a KL-divergence for Gaussian vectors.

Smooth Interpolations

The NVIB framework provides a smooth, intuitive interpolation between latent sets of vectors.

**Figure 6. Interpolation BLEU between sentences \( S_{1} \) versus \( S_{2} \).**

Future Work
- Evaluating NVIB on large scale, pretrained models for downstream tasks.
- Interpretation and implementation of NVIB for self attention.

Experiment Setup

Baselines: Nonparametric Variational AutoEncoder (NVAE) differs only from the VAE baselines in the latent representation between the Transformer encoder and decoder.

- VT Variational Transformer (all vectors)
- VTP Variational Transformer Pooled
- VTS Variational Transformer Stride

Data: All models are trained to reconstruct text using the WikiText encyclopedia dataset.

Sparsity Regularisation

The NVAE models dynamically regularise the number of vectors based on the text information.

**Figure 4. Number of latent vectors vs input tokens.**
**Figure 5. BLEU vs proportion of retained vectors.**