A VAE for Transformers with Nonparametric Variational Information Bottleneck

James Henderson [1]   Fabio Fehr [1 2]

[1]Idiap Research Institute   [2]École polytechnique fédérale de Lausanne

Motivation and Contributions

Summary

We propose a *Nonparameteric Variational Information Bottleneck* (NVIB) layer to regularise a Variational AutoEncoder (NVAE) 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 domains this work influences include: 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:

Attention Distributions	Nonparametric Distributions
▪ Variable number of vectors	▪ Variable number of mixture components
▪ Permutation invariant	▪ Exchangeable

Model

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

Prior

$$F = \sum_{i=1}^{\kappa_0} \pi_i \delta_{z_i}$$
$$\boldsymbol{\pi} \sim Dir(\frac{\alpha_0^p}{\kappa_0}, \overset{\kappa_0}{\ldots}, \frac{\alpha_0^p}{\kappa_0})$$
$$\boldsymbol{z}_i \sim G_0^p$$

- **Bounded Dirichlet Process** prior $p(F)$
- $F \sim BDP(G_0^p, \alpha_0^p, \kappa_0)$
- $G_0^p = \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\alpha_0^p = 1$
- Bounded by κ_0 samples

Posterior

$$F = \sum_{i=1}^{n+1} \pi_i \delta_{z_i}$$
$$\boldsymbol{\pi} \sim Dir(\alpha_1^q, \ldots, \alpha_{n+1}^q)$$
$$\boldsymbol{z}_i \sim G_i^q$$

- **Bounded Dirichlet Process** posterior $q(F \mid x)$
- $F \sim BDP(G_0^q, \alpha_0^q, n{+}1)$
- $G_0^q = \sum_{i=1}^{n+1} \frac{\alpha_i^q}{\alpha_0^q} G_i^q$ and $\alpha_0^q = \sum_{i=1}^{n+1} \alpha_i^q$
- $G_i^q = \mathcal{N}(\boldsymbol{\mu}_i^q, \boldsymbol{I}(\boldsymbol{\sigma}_i^q)^2)$ and $G_{n+1}^q, \alpha_{n+1}^q = G_0^p, \alpha_0^p$
- For n inputs, $n{+}1$ mixture components

Variational Information Bottleneck Loss

The VIB loss maximises the log-likelihood of the observation x, where x is the input text:

$$\log(p(x)) \geq \underbrace{\mathbb{E}_{q(F|x)} \log(p(x \mid F))}_{L_R} - \underbrace{D_{KL}(q(F \mid x) \,||\, p(F))}_{\approx L_D + L_G}$$

- L_R - Reconstruction loss
- L_D - Kullback-Liebler divergence for Dirichlet weights $\boldsymbol{\pi}$
- L_G - Kullback-Liebler divergence for Gaussian vectors $\boldsymbol{Z}$

Denoising Attention

Regroup the scaled dot product attention function such that all operations are in $\boldsymbol{Z}$ space:

$$\begin{aligned} Attention(u', \boldsymbol{Z}) &= \text{softmax}\left(\frac{(\boldsymbol{u}'\boldsymbol{W}^Q)(\boldsymbol{Z}\boldsymbol{W}^K)^T}{\sqrt{d}}\right)\boldsymbol{Z}\boldsymbol{W}^V \\ &= \text{softmax}\left(\frac{\boldsymbol{u}\boldsymbol{Z}^T}{\sqrt{d}}\right)\boldsymbol{Z}\boldsymbol{W}^V \\ &= Attn(\boldsymbol{u}, \boldsymbol{Z})\boldsymbol{W}^V \end{aligned}$$

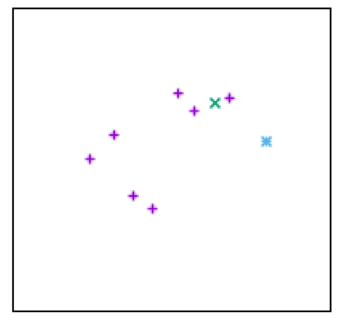

Figure 1. Standard attention.

where $\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{W}^Q(\boldsymbol{W}^K)^T$ is projected into the space of $\boldsymbol{Z}$. We interpret $\boldsymbol{Z}$ as specifying a probability distribution over a vector space, and define a function over such probability distributions which, when given any vector(s) $\boldsymbol{u}$, always returns the result vector(s) $Attn(\boldsymbol{u}, \boldsymbol{Z})$.

$$Attn(\boldsymbol{u}, \boldsymbol{Z}) = \text{softmax}\left(\tfrac{1}{\sqrt{d}}\boldsymbol{u}\boldsymbol{Z}^T\right)\boldsymbol{Z}$$

$$F_{\boldsymbol{Z}} = \sum_{i=1}^{n} \frac{\exp(\frac{1}{2\sqrt{d}}||\boldsymbol{z}_i||^2)}{\sum_{i=1}^{n}\exp(\frac{1}{2\sqrt{d}}||\boldsymbol{z}_i||^2)} \delta_{z_i}$$

$$DAttn(u \,;\, F_{\boldsymbol{Z}}) = \int_{\boldsymbol{v}} \frac{f(\boldsymbol{v})\, g(\boldsymbol{u}; \boldsymbol{v}, \sqrt{d}\boldsymbol{I})}{\int_{\boldsymbol{v}} f(\boldsymbol{v})\, g(\boldsymbol{u} \,;\, \boldsymbol{v}, \sqrt{d}\boldsymbol{I})\, d\boldsymbol{v}} \boldsymbol{v}\, d\boldsymbol{v}$$

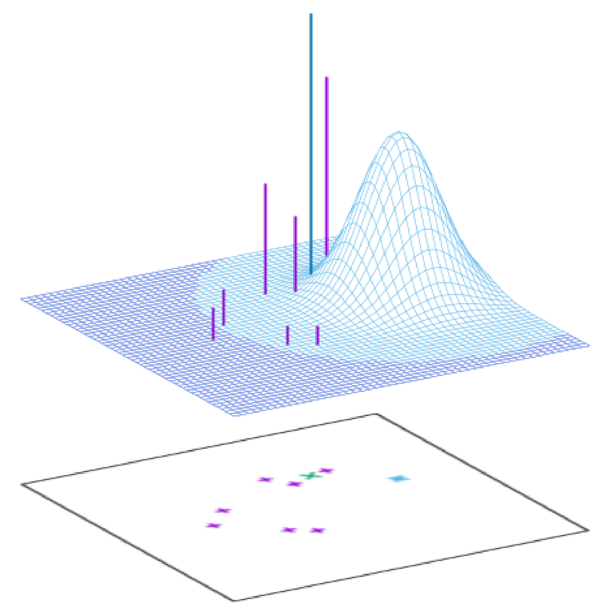

Figure 2. Denoising attention.

where δ_{z_i} is an impulse distribution at $\boldsymbol{z}_i$, $f(\cdot)$ is the probability density function for distribution F, and $g(\boldsymbol{u};\, \boldsymbol{v}, \sqrt{d}\boldsymbol{I})$ is a multivariate Gaussian function with diagonal variance of $\sqrt{d}$. Figure 2 shows that $DAttn(u \,;\, F_{\boldsymbol{Z}})$ 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.

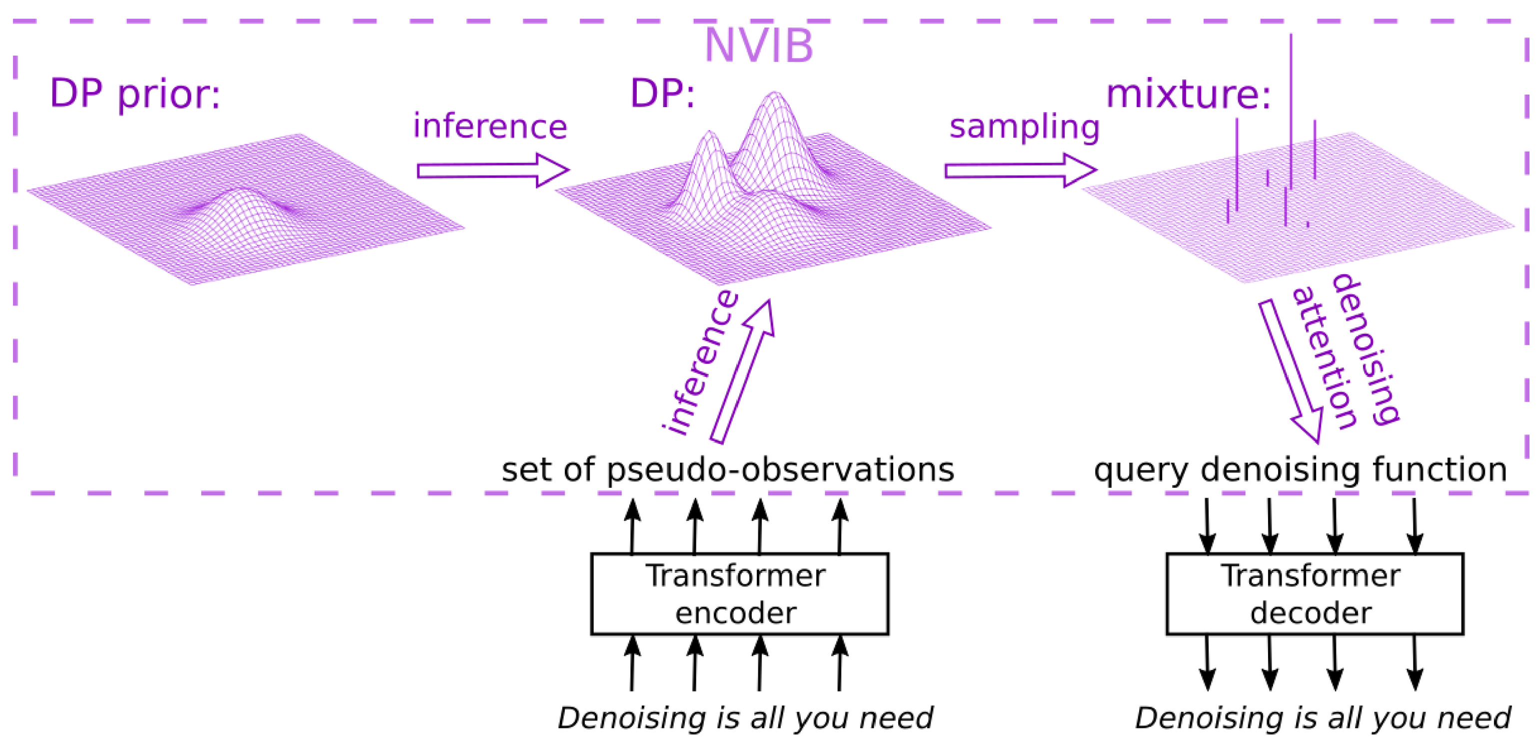

Figure 3. Nonparametric VAE model, with its NVIB layer.

- The encoder estimates the posterior *psuedo-observation* parameters given an input text x.
- 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 Kullback-Leibler divergence.

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.

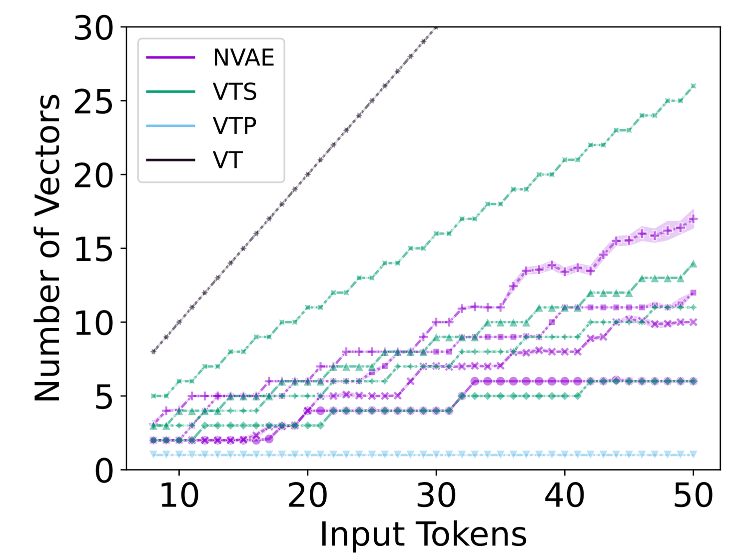

Figure 4. Number of latent vectors vs input tokens.

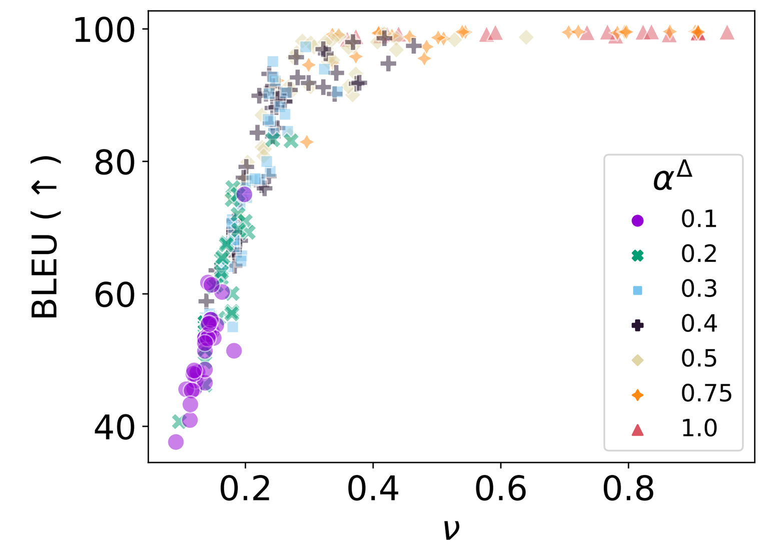

Figure 5. BLEU vs proportion of retained vectors ν.

- Figure 4: the NVAE models regularise across text with varying input lengths.
- Figure 5: the conditional prior hyperparameter α^Δ controls latent vector sparsity.

Smooth Interpolations

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

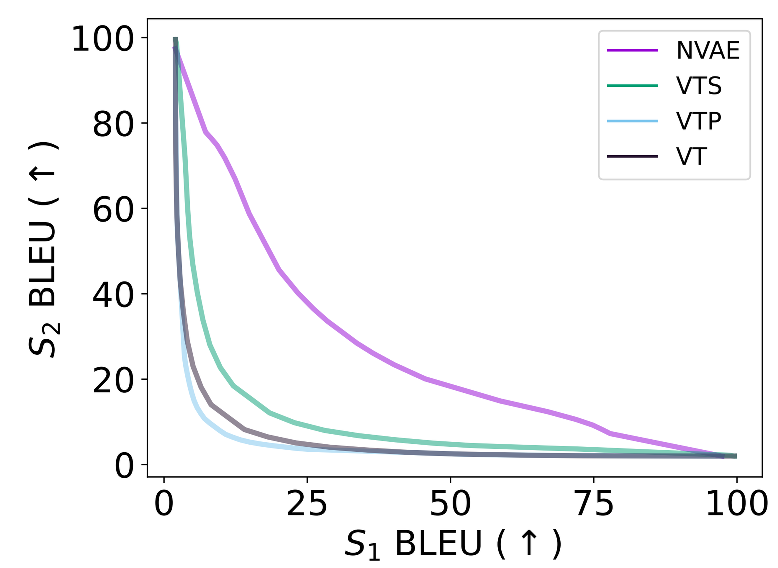

Figure 6. Interpolation BLEU between sentences S_1 versus S_2.

- Figure 6: the NVIB regulariser in NVAE provides more fluent and smoother interpolations.

Future Work

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

Video   Repository

fabio.fehr@idiap.ch & james.henderson@idiap.ch