Bit-Swap

Recursive Bits-Back Coding for Lossless Compression with Hierarchical Latent Variables

Friso H. Kingma Pieter Abbeel Jonathan Ho

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley UNIVERSITY OF CALIFORNIA

Compression with Hierarchical Latent Variable Models

Abstract

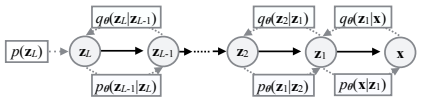
Previous work:
- BB-ANS makes a practical compression scheme with the use of latent variable models.
- Bitrate approx. equal to the ELBO of a VAE.
- Practical for one latent layer
- Impractical for multiple latent layers

Our contribution:
- Making it practical for a hierarchical latent variable model
- Recursively applying the bits-back argument

Hierarchical Latent Variable Model

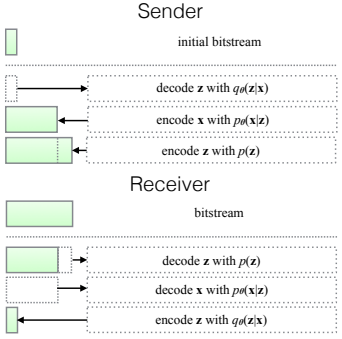
- **VAE** with multiple latent layers
- Sampling process of generative model and inference model obeys **Markov chain**

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_1)p_{\boldsymbol{\theta}}(\mathbf{z}_1)\mathrm{d}\mathbf{z}_1$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}_1) = \int p_{\boldsymbol{\theta}}(\mathbf{z}_1|\mathbf{z}_2)p_{\boldsymbol{\theta}}(\mathbf{z}_2)\mathrm{d}\mathbf{z}_2$$
$$\vdots$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}_{L-1}) = \int p_{\boldsymbol{\theta}}(\mathbf{z}_{L-1}|\mathbf{z}_L)p(\mathbf{z}_L)\mathrm{d}\mathbf{z}_L,$$



Bits-Back Coding

BB-ANS on **one layer latent variable model**:
- sender uses initial bits and receiver recovers these initial bits
- net bitrate approx. equal to ELBO



Naive: Sampling

Achieves

- A compression rate approx. equal to

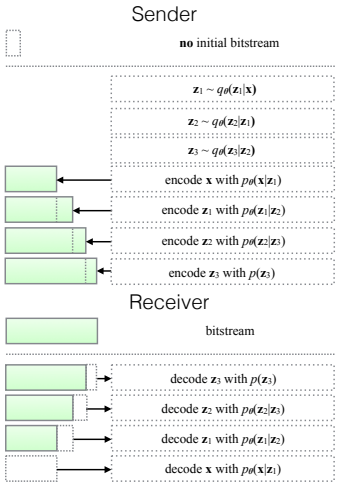
$$N_{\text{net}}^{\text{sampling}} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L})]$$

How

- Sender samples $\mathbf{z}_{1:L} \sim q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})$
- Sender encodes $\mathbf{x}$ using $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_1)$
- Sender encodes $\mathbf{z}_{1:L}$ as well using $p_{\boldsymbol{\theta}}(\mathbf{z}_1|\mathbf{z}_2)$, $p_{\boldsymbol{\theta}}(\mathbf{z}_2|\mathbf{z}_3), \ldots, p(\mathbf{z}_L)$
- Sent the full bitstream over to the receiver.
- Receiver decodes in reverse order

Limitations

After decoding $\mathbf{z}_{1:L}$ and $\mathbf{x}$, the goal of lossless compression is reached: the receiver fully recovered the data $\mathbf{x}$. However, the **additionally recovered latent variables $\mathbf{z}_{1:L}$ are redundant after retaining $\mathbf{x}$.**



Townsend et al. 2019: BB-ANS

Achieves

- A compression rate approx. equal to

$$N_{\text{total}}^{\text{BBANS}} = N_{\text{init}}^{\text{BBANS}} + \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})}[\log q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L})]$$

- **Net number of bits** added to the bitstream is approx. **equal to the negative ELBO**

$$N_{\text{total}}^{\text{BBANS}} - N_{\text{init}}^{\text{BBANS}} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})}[\log q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L})] = -\mathcal{L}(\boldsymbol{\theta}) \text{ (ELBO)}$$

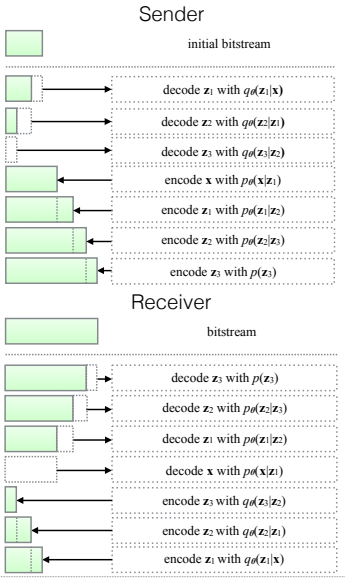
How

Decoding from uniformly random bitstream equals sampling.
- Sender decodes $\mathbf{z}_{1:L}$ from **initial bitstream**
- The receiver later get these "**bits back**".

Limitations

The min. number of  BB-ANS needs **grows linearly with L :**

$$N_{\text{init}}^{\text{BBANS}} = -\log q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x}) = \sum_{i=0}^{L-1} -\log q_{\boldsymbol{\theta}}(\mathbf{z}_{i+1}|\mathbf{z}_i)$$



Our contribution: **Bit-Swap**

Achieves

- Net number of bits added to the bitstream is approx. equal to the negative ELBO:

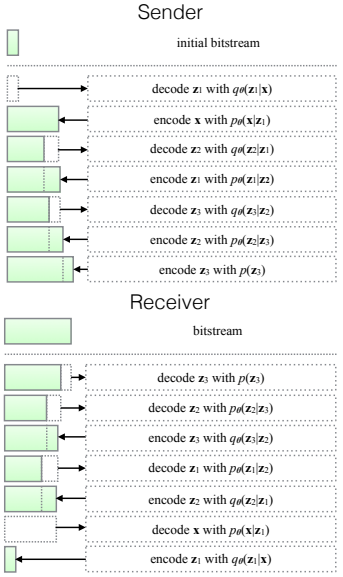
$$N_{\text{total}}^{\text{BitSwap}} - N_{\text{init}}^{\text{BitSwap}} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})}[\log q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L})] = -\mathcal{L}(\boldsymbol{\theta}) \text{ (ELBO)}$$

- Number of **initial bits is bounded:**

$$N_{\text{init}}^{\text{BitSwap}} \leq \sum_{i=0}^{L-1} \max\left(0, \log \frac{p_{\boldsymbol{\theta}}(\mathbf{z}_{i-1}|\mathbf{z}_i)}{q_{\boldsymbol{\theta}}(\mathbf{z}_{i+1}|\mathbf{z}_i)}\right) \leq \sum_{i=0}^{L-1} -\log q_{\boldsymbol{\theta}}(\mathbf{z}_{i+1}|\mathbf{z}_i) = N_{\text{init}}^{\text{BBANS}}$$

How

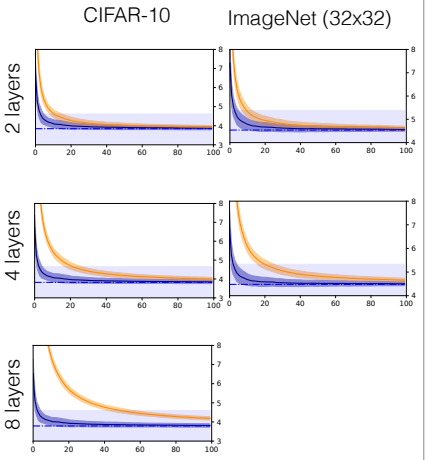
- Interpret HLVM is a multiple nested LVM's
- The "prior" of every layer is again a VAE
- Bit-Swap exploits this nested structure by **recursively applying the bits-back argument**.



Results

Cumulative average compression rate of compressing 100 images in sequence (including initial bits)

= Bit-Swap = BB-ANS



Benchmark compressors

Training: Model trained on random 32x32 patches of Imagenet.
Compression: 100 original ImageNet images cropped to multiples of 32 pixels on each side. Bit-Swap and BB-ANS compressed a grid of 32x32 blocks for each image.

	Compression Rate (bits/dim)
uncompressed	8.00
gzip	5.96
bzip2	5.07
LZMA	5.09
PNG	4.71
WebP	3.66
BB-ANS	3.62
Bit-Swap	**3.51**