Deep Attentive Variational Inference

Ifigeneia Apostolopoulou, Ian Char, Elan Rosenfeld, Artur Dubrawski

Motivation

Very Deep VAE architectures favor only proximate dependencies in the latent space **limiting long-range conditional dependencies.**

In practice, the factorizations may not hold:

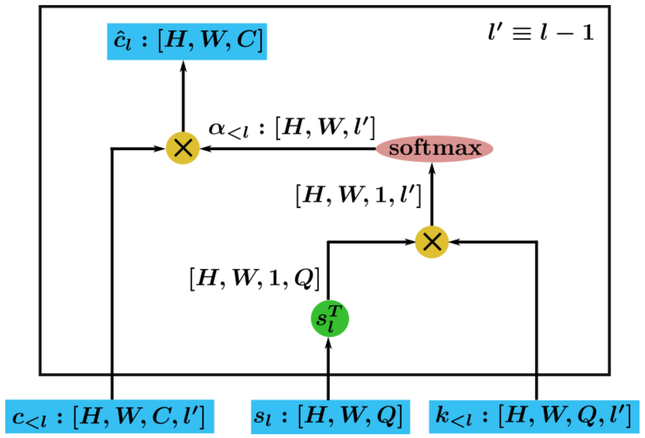
$$p(z)! = \prod p(z_l|z_{<l}), \qquad q(z|x)! = \prod q(z_l|x, z_{<l}).$$

Main Idea

To build more expressive distributions $p(z_l|z_{<l})$ and $q(z_l|x, z_{<l})$ that learn to attend to latent and observed features most critical to inference.

Depth-wise Attention

- $c_{<l}$: $H \times W$ independent pixel sequences of C −dimensional features of length $l-1$.
- $\alpha_{<l} = \{\alpha_{m\to l}\}$: the attention scores.
- $\alpha_{m\to l}(i,j)$: how important is the $m-th$ term at pixel (i,j) ?



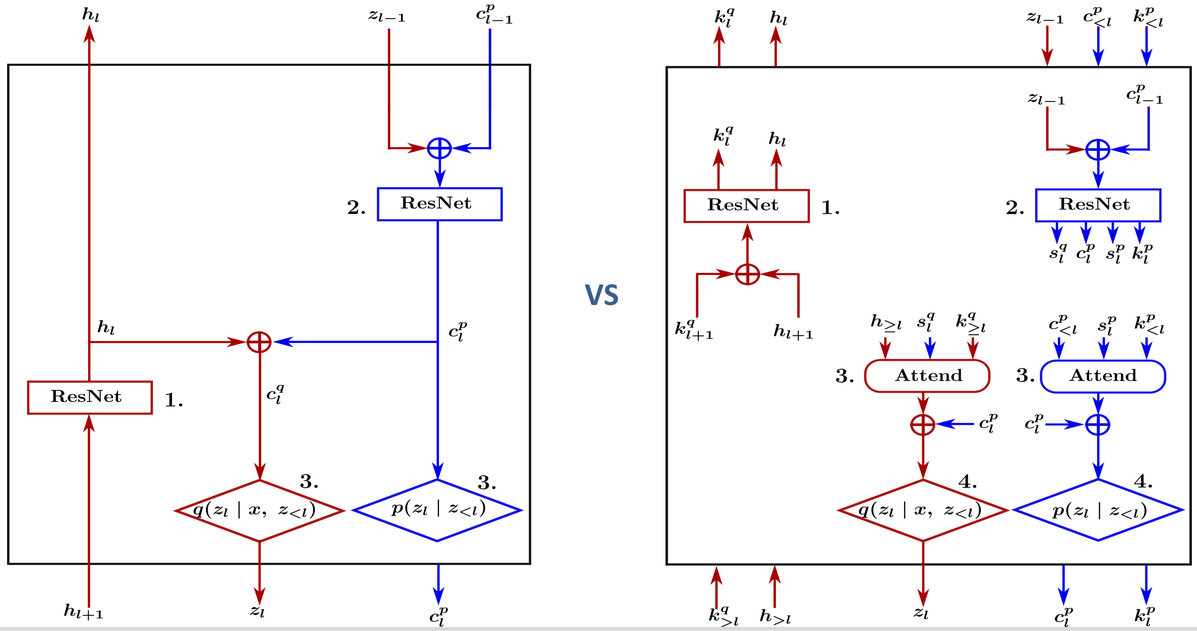
Method

Inference Path: constructs the posterior $q(z_l|x, z_{<l})$ Generative Path: constructs the prior $p(z_l|\, z_{<l})$

A **strongly connected variational layer** l receives:

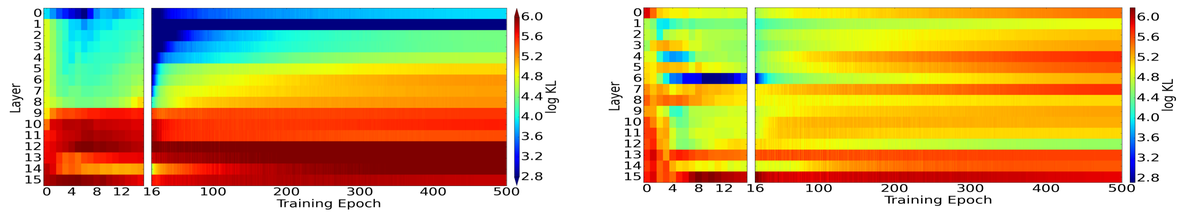
- Stochastic context $c^p_{<l}$ from all layers above in the hierarchy during the top-down pass (steps 2-4).
- Deterministic features $h_{>l}$ of x from all layers below in the hierarchy during the bottom-up pass (step 1).

The Attend modules (step 3) decide which parts of the features from earlier layers are important for inference.

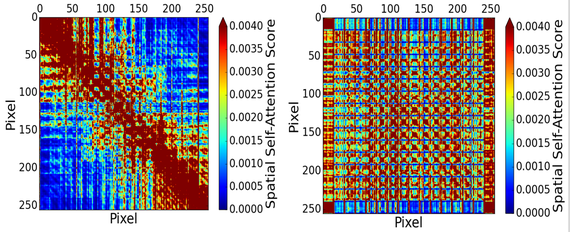


Depth-wise Attention keeps the upper variational layers active (KL >>0) and mitigates posterior collapse.

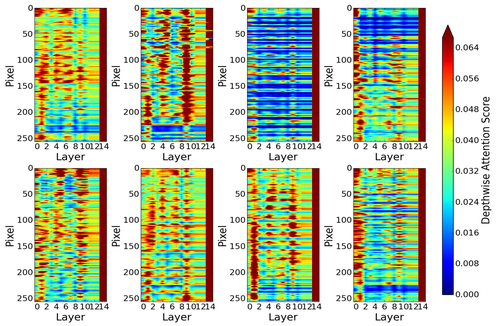
Locally connected variational layer vs Strongly connected variational layer (ours)



Spatial Attention Patterns



Depth-wise Attention Patterns



Experiments

- SOTA results on binary and natural images compared to hierarchical VAEs (BIVA, NVAE, Very Deep VAE,…).
- Significantly **fewer layers** needed.
- Significantly **less** training & inference **time.**

Discussion

- First **attention-driven** framework for variational inference in **deep probabilistic models.**
- Attention for **better utilization of** the **latent space.**
- **Factorized** intra-layer and inter-layer **attention** operations.
- **Sparse** and highly structured **attention patterns.**

iapostol@andrew.cmu.edu