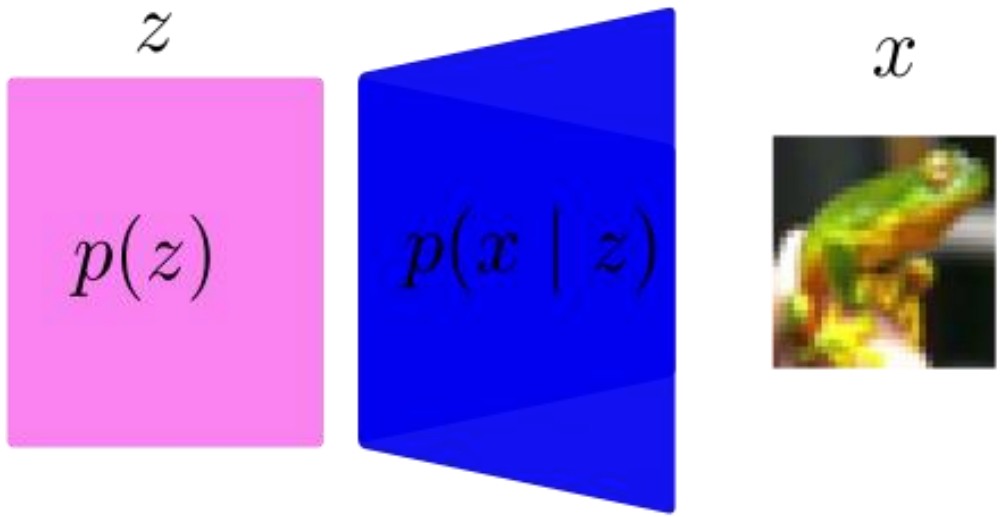
Deep Attentive Variational Inference

Ifigeneia Apostolopoulou, Ian Char, Elan Rosenfeld, Artur Dubrawski

Variational AutoEncoder: A quick review

Generative Model

$$p(x, z) = p(z) \times p(x \mid z)$$



$$p(x) = \int p(x, z) dz$$

Variational AutoEncoder: A quick review

✓Introduce approximate posterior $q(z \mid x)$:

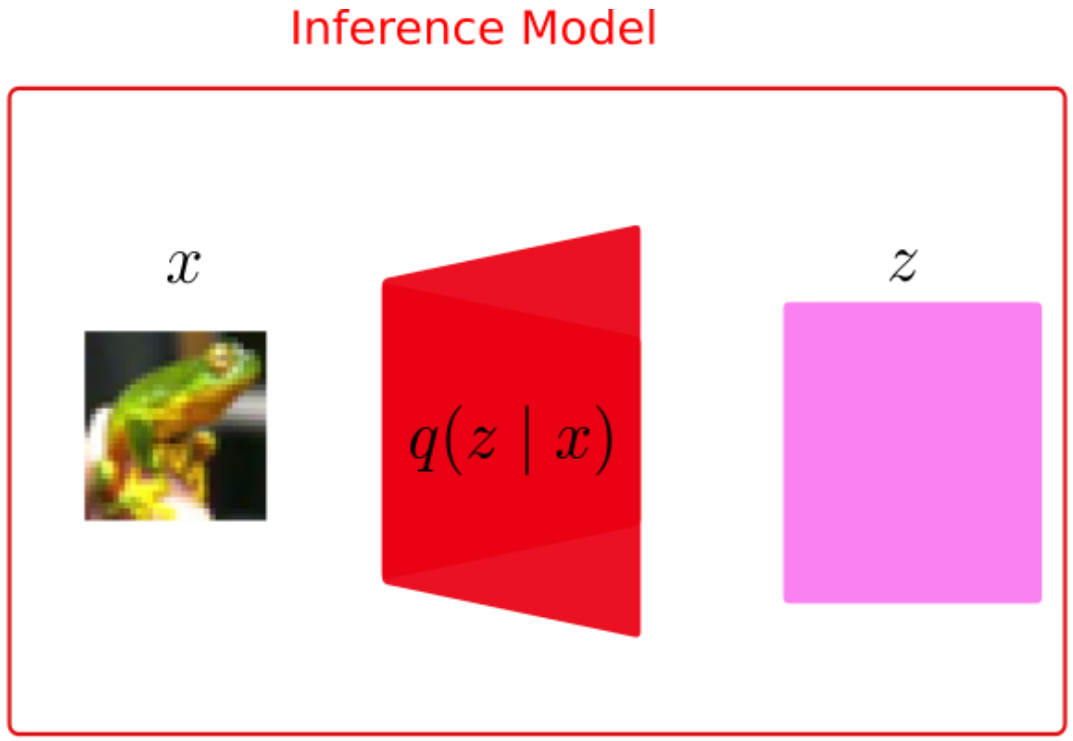
$$\log p(x) = log \int p(x,z)dz = \log \int \frac{q(z \mid x)}{q(z \mid x)} p(x,z)dz$$

✓Apply Jensen's inequality:

$$\log p(x) \geq \int q(z \mid x) \log \frac{p(z)p(x|z)}{q(z \mid x)} dz \geq E_{q(z|x)}[\log p(x \mid z)] - KL(q(z \mid x) \,||\, p(z))$$

Variational AutoEncoder: A quick review

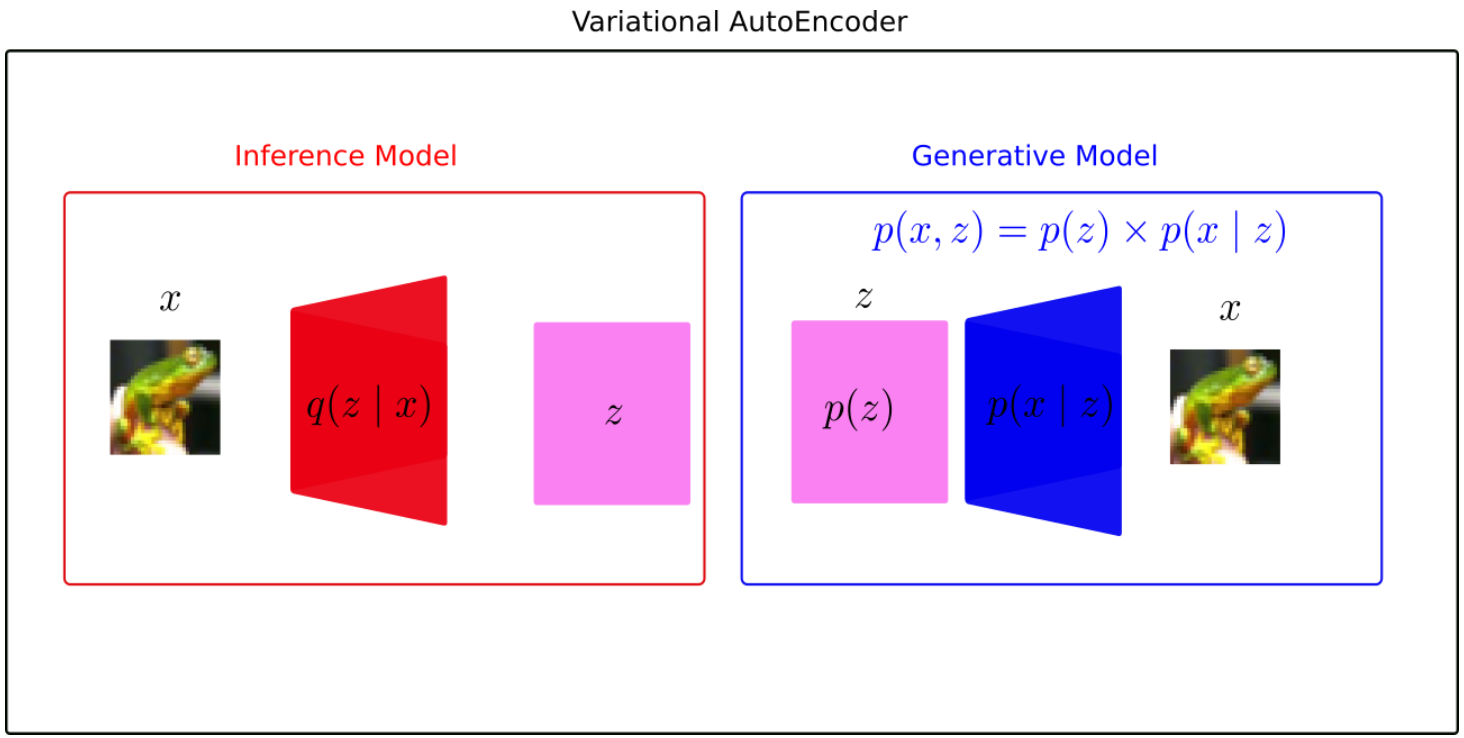
- Approximate the exact posterior.
- Obtain a lower bound of the marginal likelihood.



$$logp(x) \geq \mathbb{E}_{q(z|x)}[logp(x \mid z)] - KL(q(z \mid x) \mid\mid p(z))$$

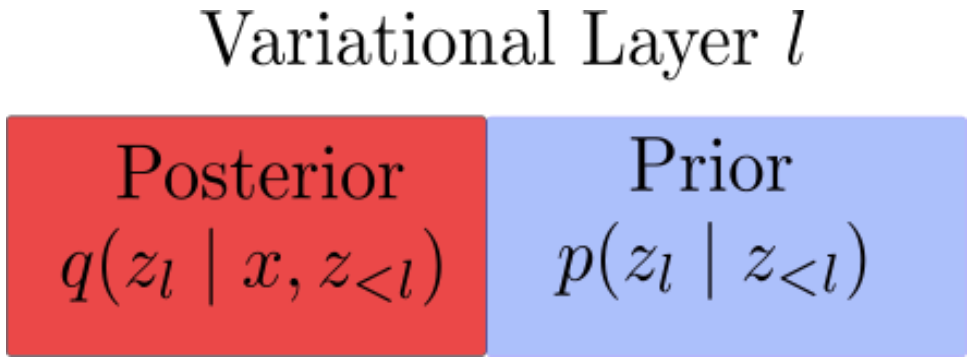
Variational AutoEncoder: A quick review

- Jointly train the inference and generative model.



Hierarchical VAEs

The latent variables are generated in blocks. Each block is generated by a layer in a hierarchy.

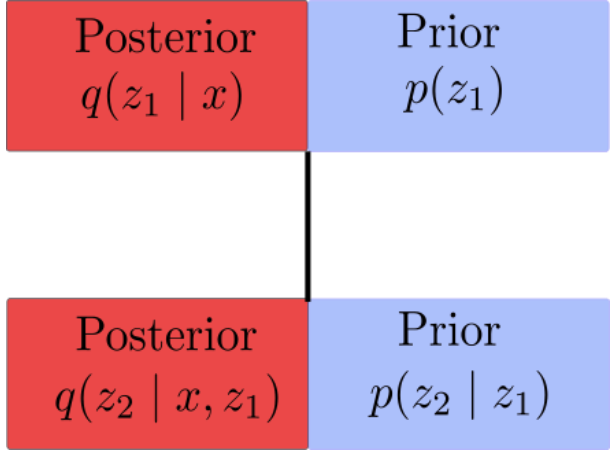


Hierarchical VAEs

- Stack multiple layers of latent variables.

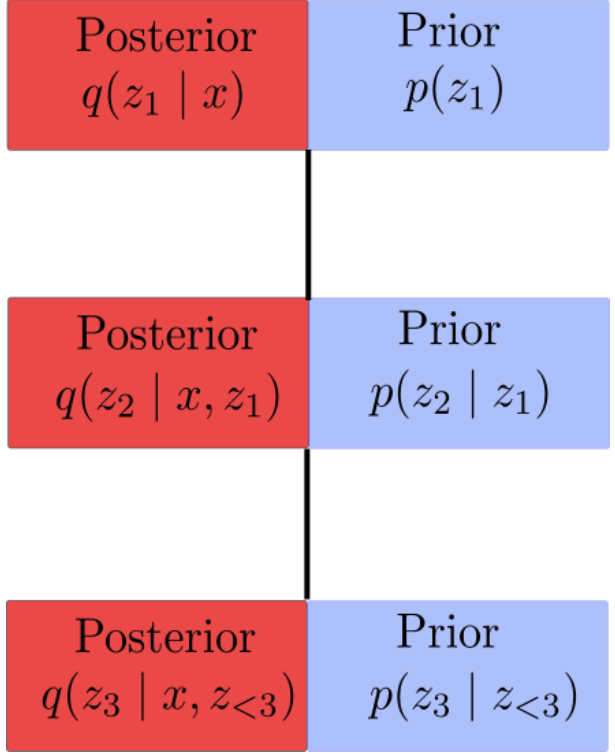
Hierarchical VAEs

- Stack multiple layers of latent variables.



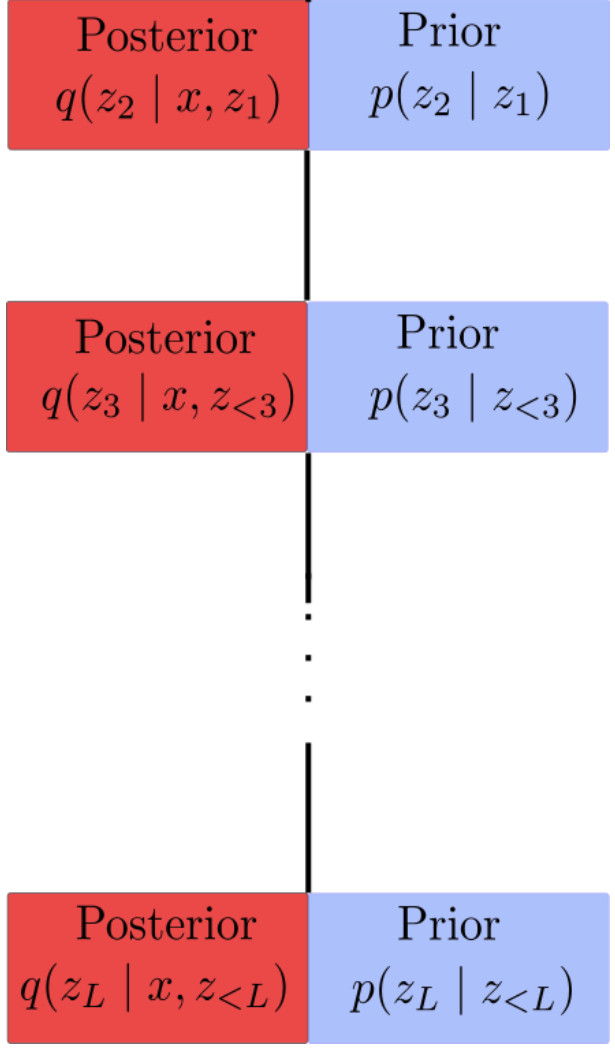
Hierarchical VAEs

- Stack multiple layers of latent variables.



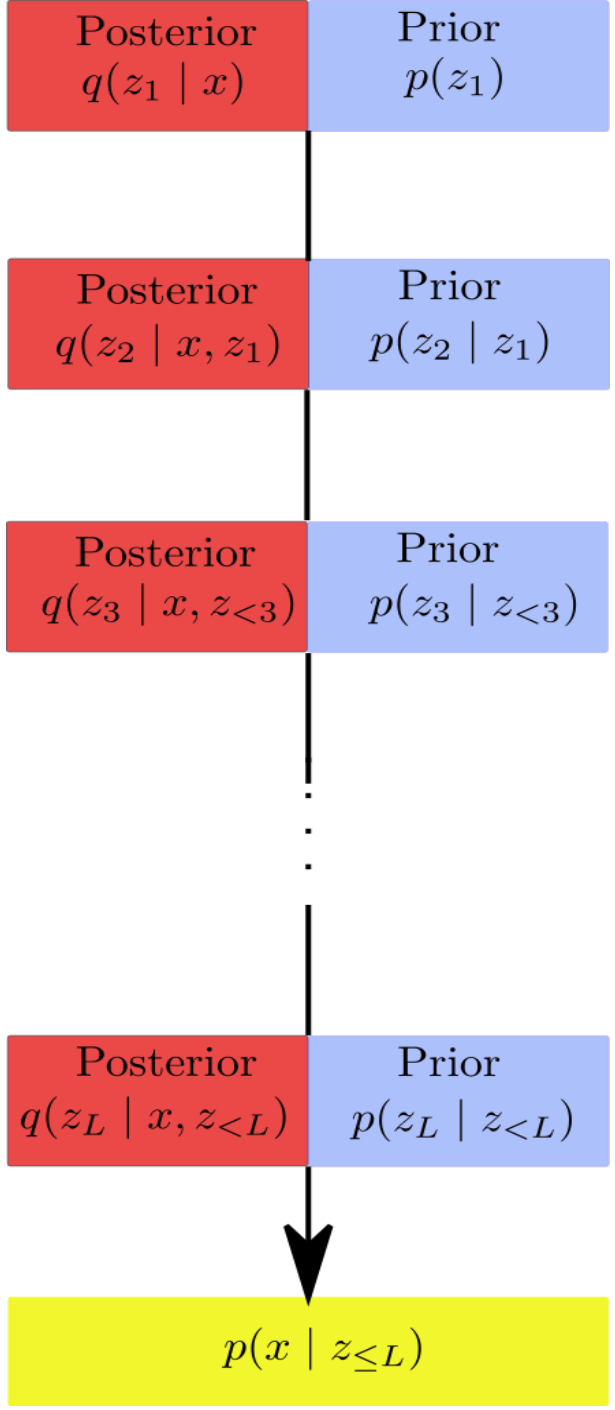
Hierarchical VAEs

- Stack multiple layers of latent variables.



Hierarchical VAEs

- The conditional likelihood at the end receives context and latent samples from the last layer in the hierarchy and generates image x.

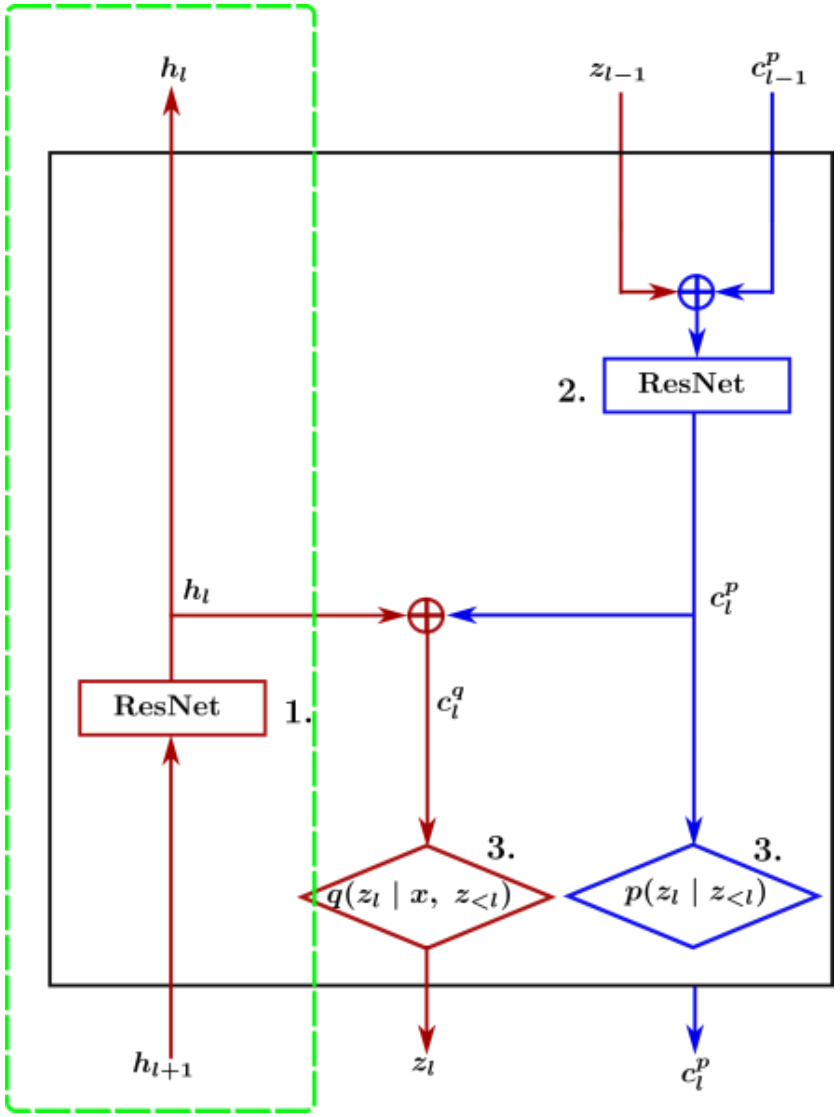


Bidirectional Inference

- Inference is a two-stage process.
- Phase 1: Bottom-up pass.
- Phase 2: Top-down pass.

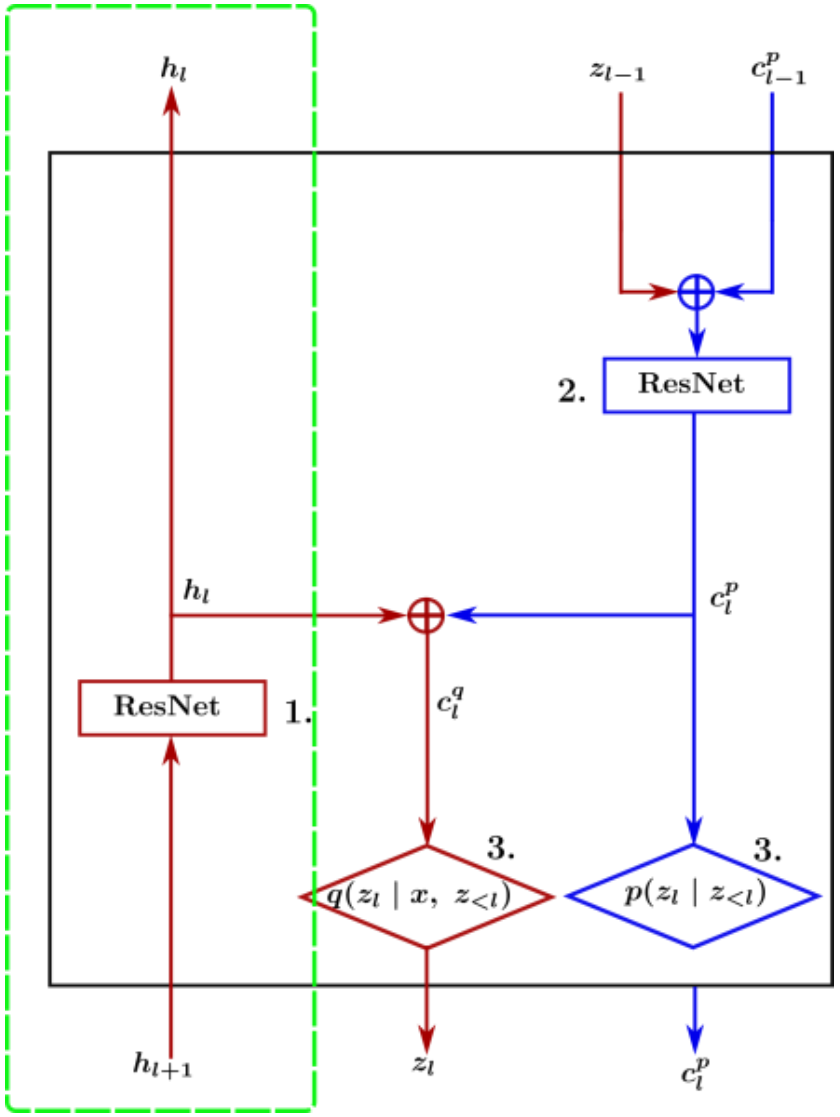
Bidirectional Inference

- During the *bottom-up pass,* deterministic features h_l of data x are computed.



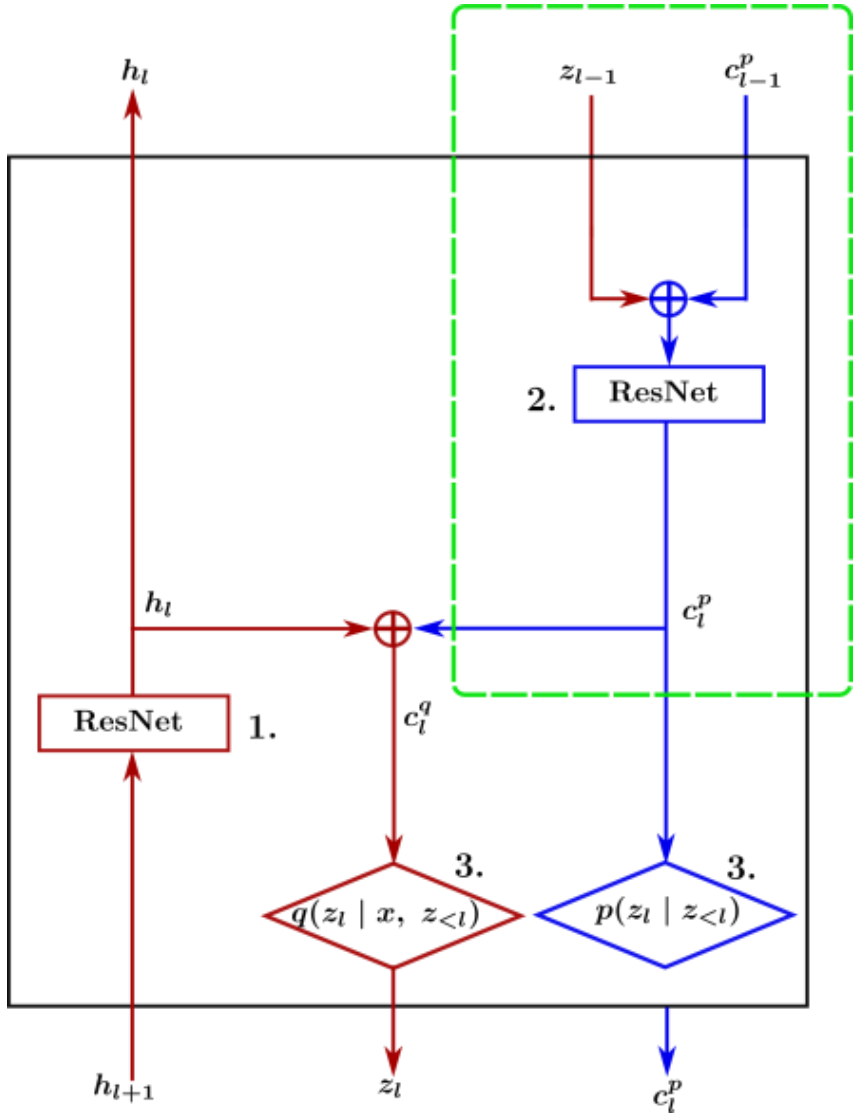
Bidirectional Inference

- $h_l \leftarrow T_l^q(h_{l+1})$.
- $h_{L+1} \equiv x$.
- T_l^q is a block of ResNet cells.



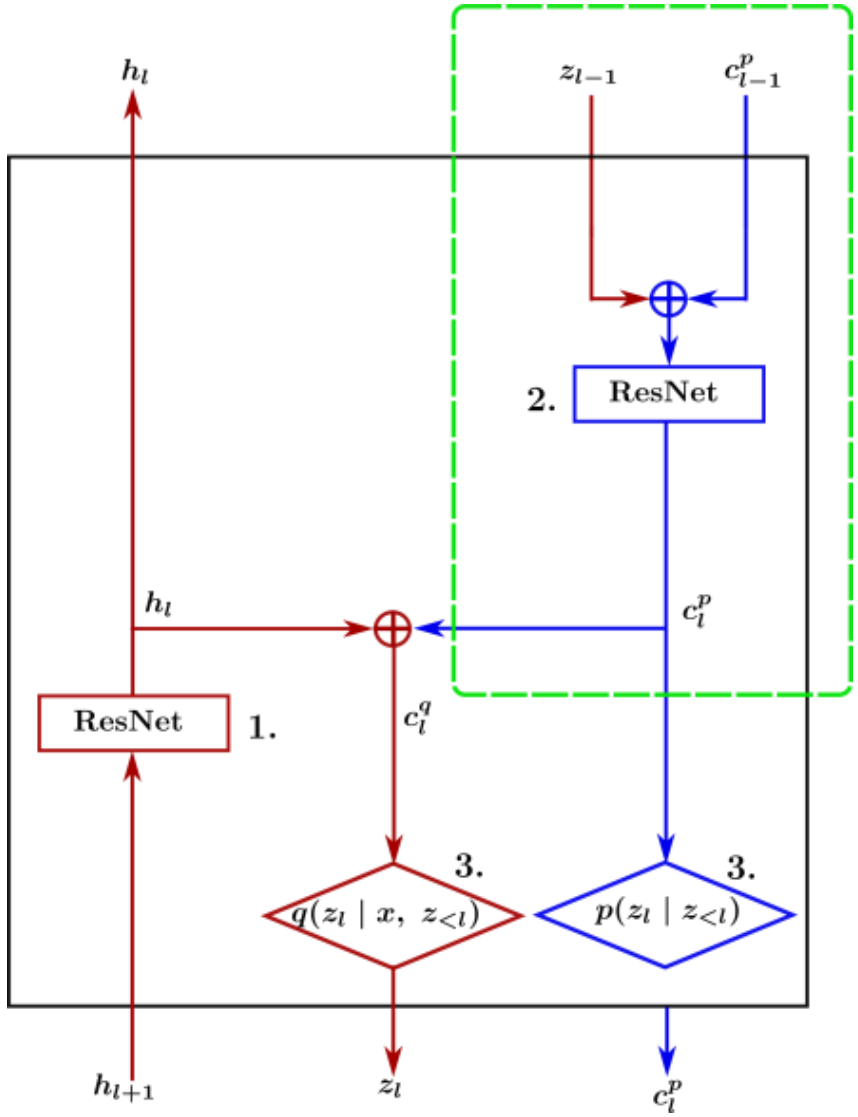
Bidirectional Inference

- During the *top-down pass,* a stochastic context c_l^p is computed from the latent sample $z_{l-1}$and stochastic context c_{l-1}^p of the previous layer.



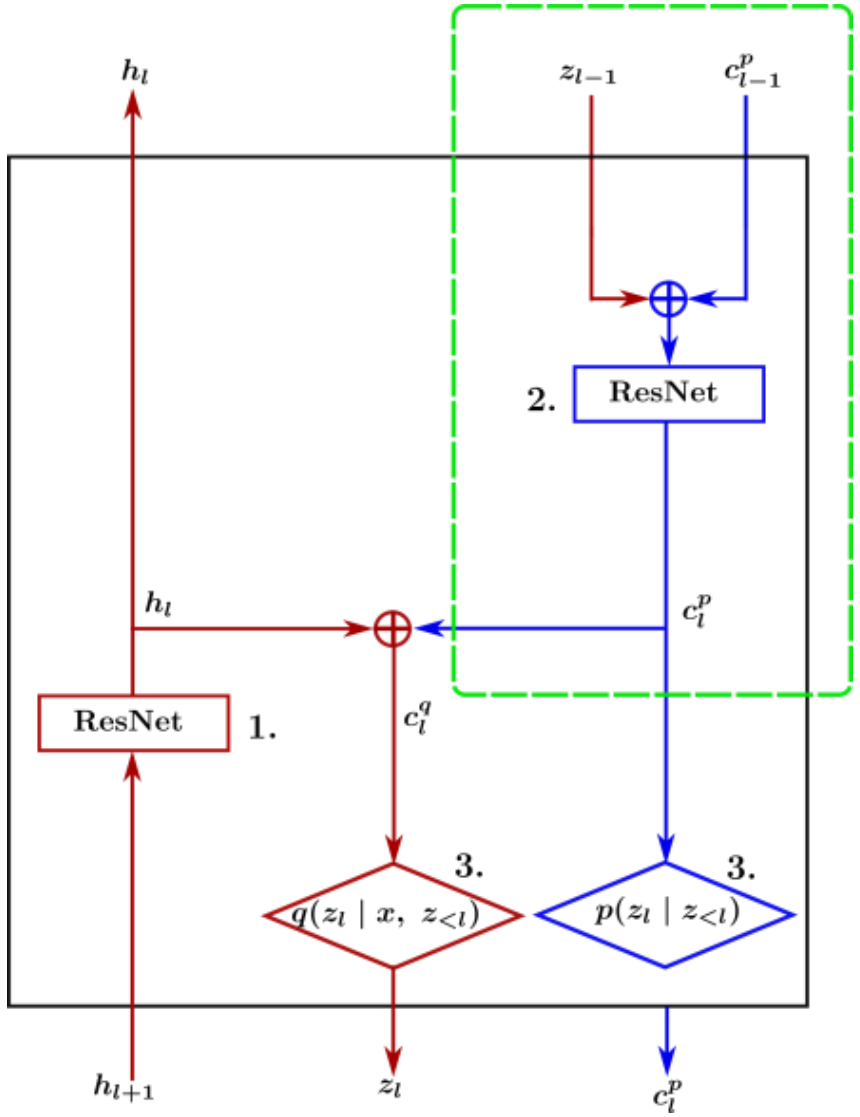
Bidirectional Inference

- $c_l^p \leftarrow T_l^p(z_{l-1} \oplus c_{l-1}^p)$.
- c_0^p is a constant.
- T_l^p is a block of ResNet cells.



Bidirectional Inference

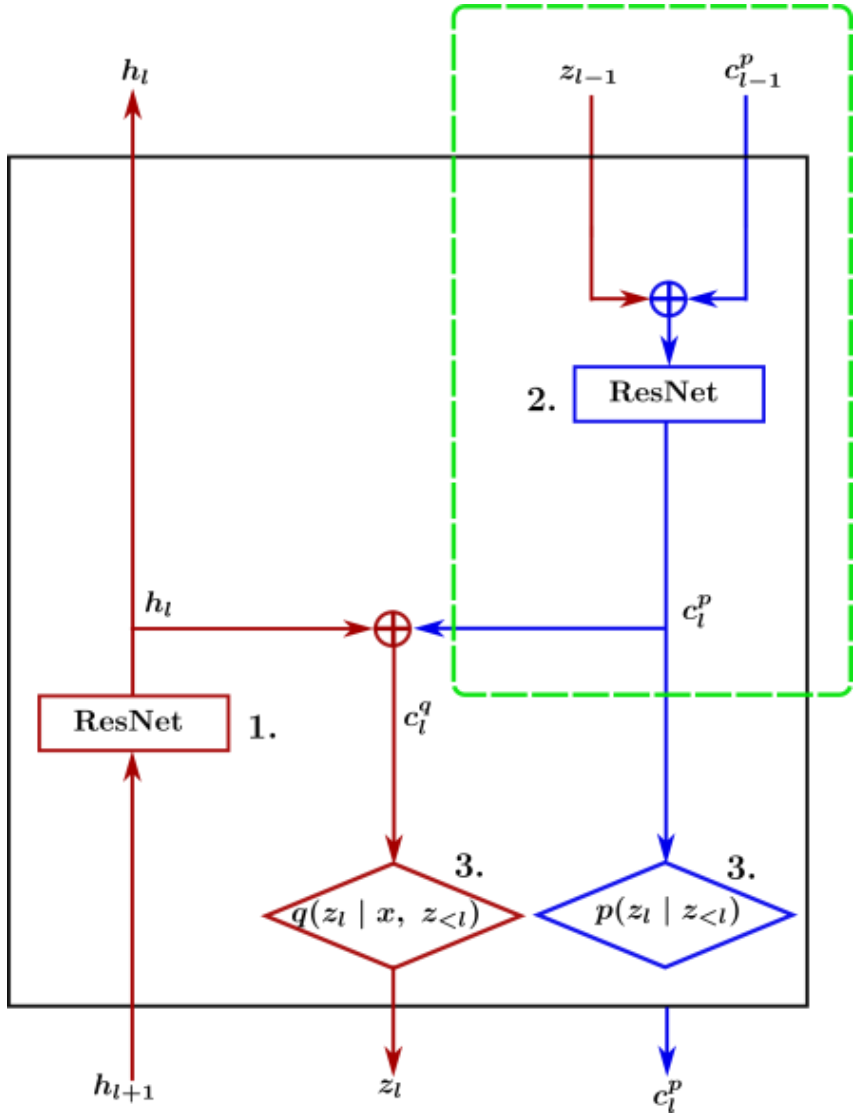
- $c_l^p \leftarrow T_l^p(z_{l-1} \oplus c_{l-1}^p)$.
- c_l^p carries information from earlier latent samples $z_{<l}$.
- It is a representation for the conditioning factor of the prior.



Bidirectional Inference

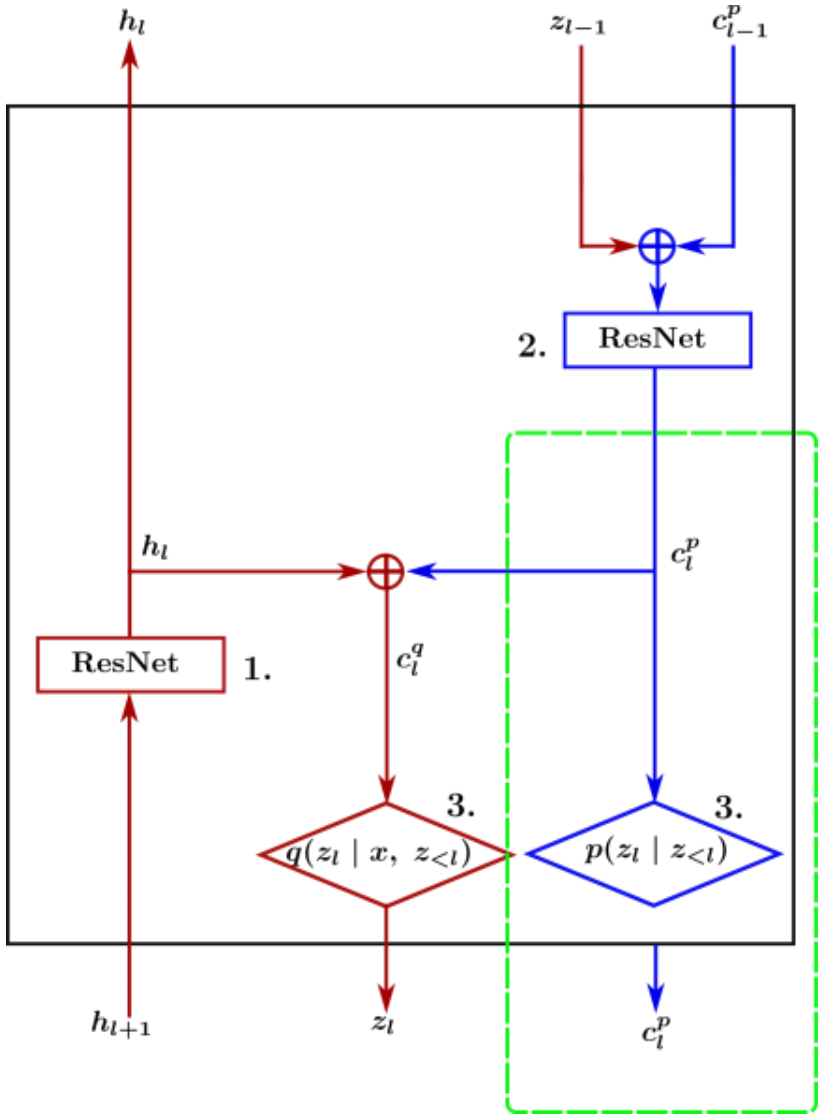
- c_l^p carries information from earlier latent samples $z_{<l}$.
- Due to c_l^p, a strongly connected factorization of the prior is achieved:

$$p(z) = p(z_1) \times \prod p(\, z_l \mid z_{<l}\,).$$



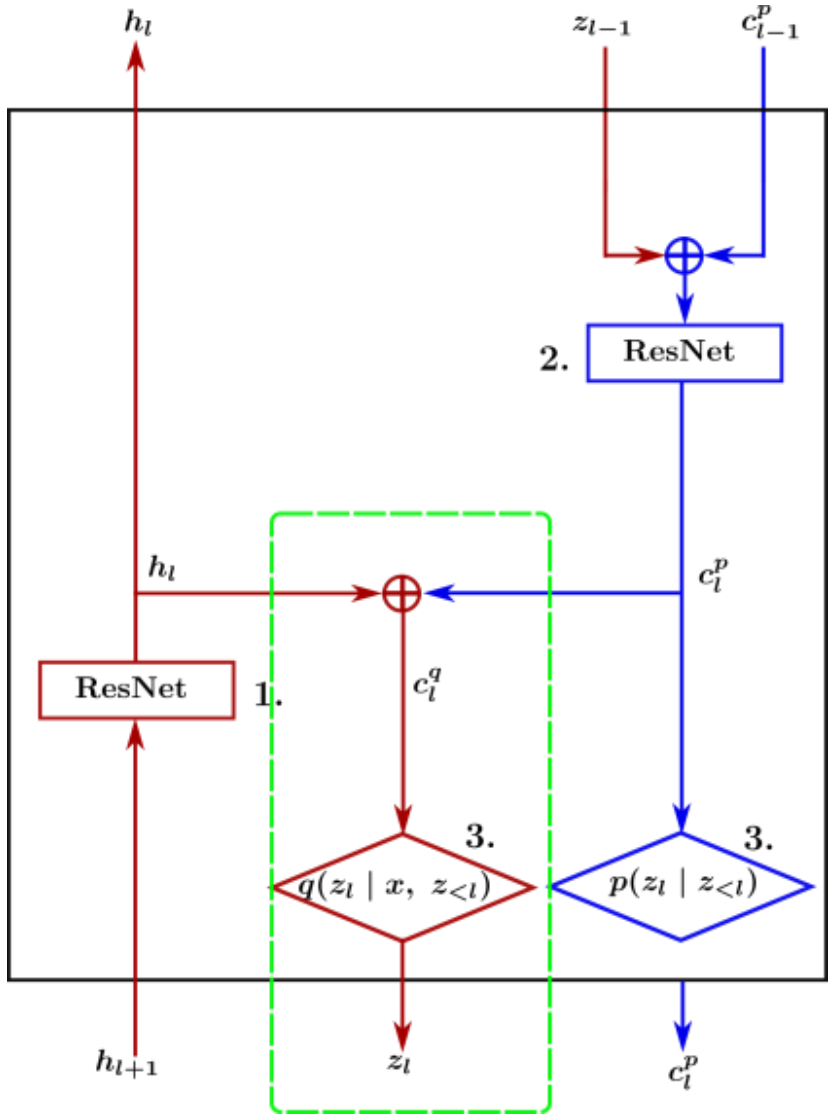
Bidirectional Inference

- This context c_l^p feeds the network that is responsible for generating the parameters of the prior.



Bidirectional Inference

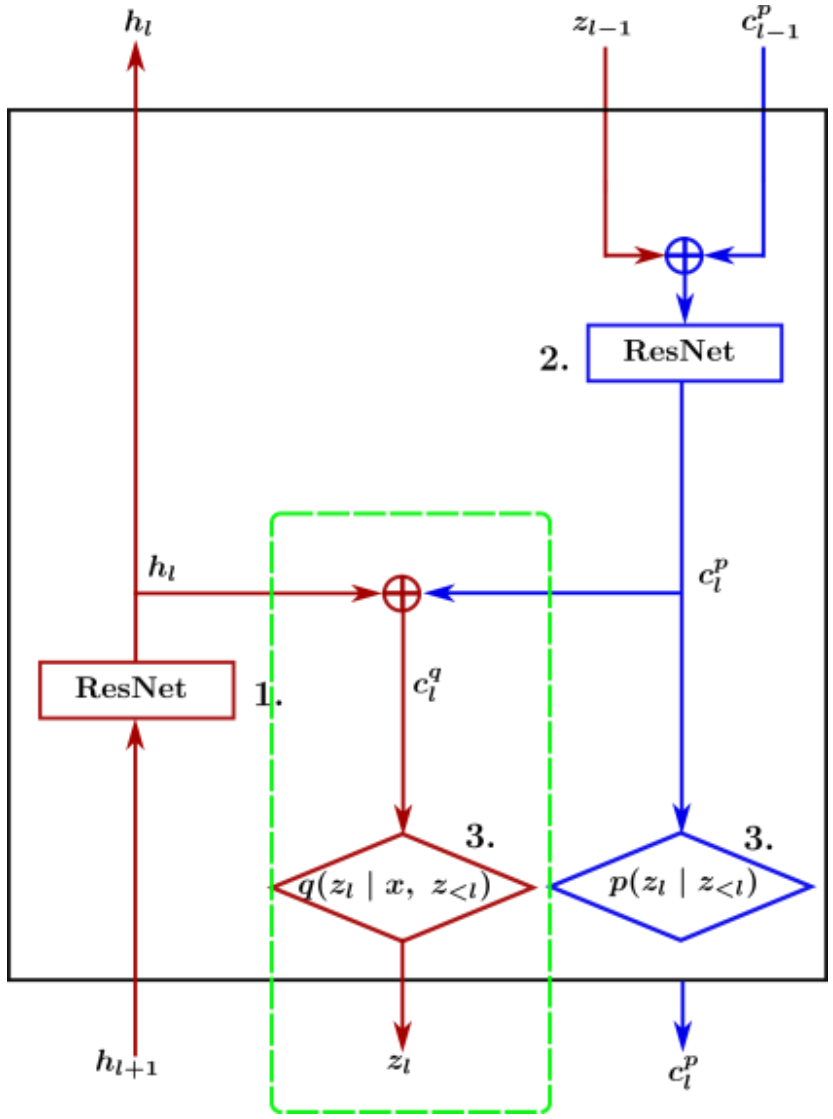
- The deterministic and the stochastic features are merged to give the context c_l^q of the posterior.
- $c_l^q \leftarrow h_l \oplus c_l^p$.



Bidirectional Inference

- Due to c_l^p, a strongly connected factorization of the posterior is achieved:

$$q(z \mid x) = q(z_1 \mid x) \times \prod q(z_l \mid x, z_{<l}).$$

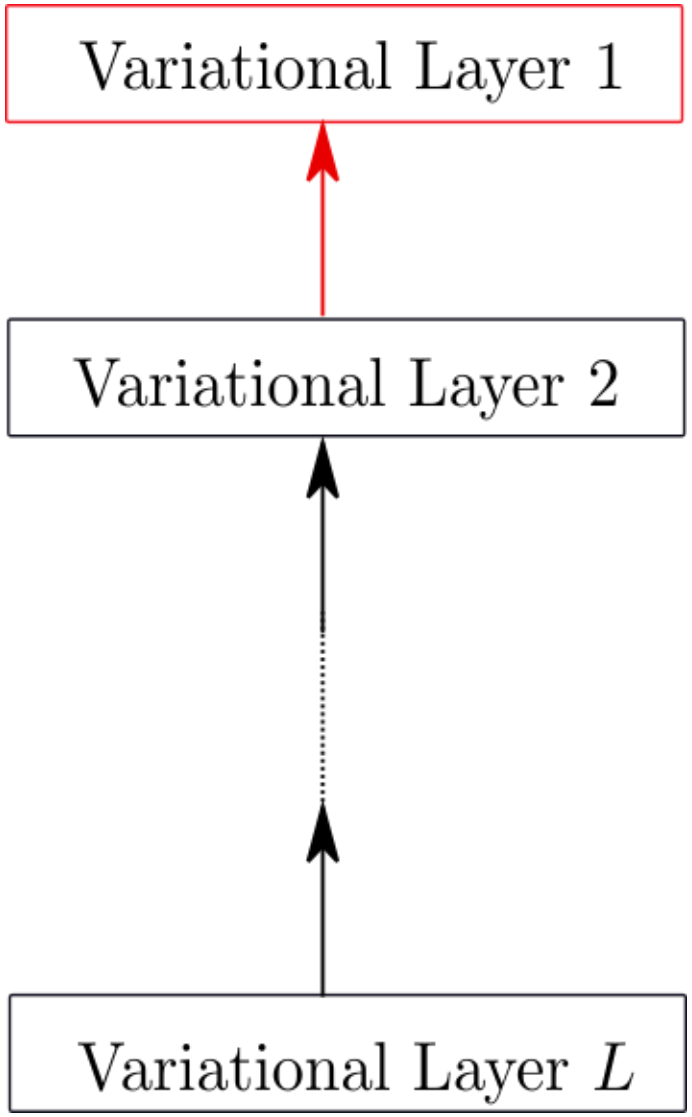


Problem: Locality in Deep VAEs

- During inference, a variational layer is connected only with the immediately adjacent variational layer in the architecture.

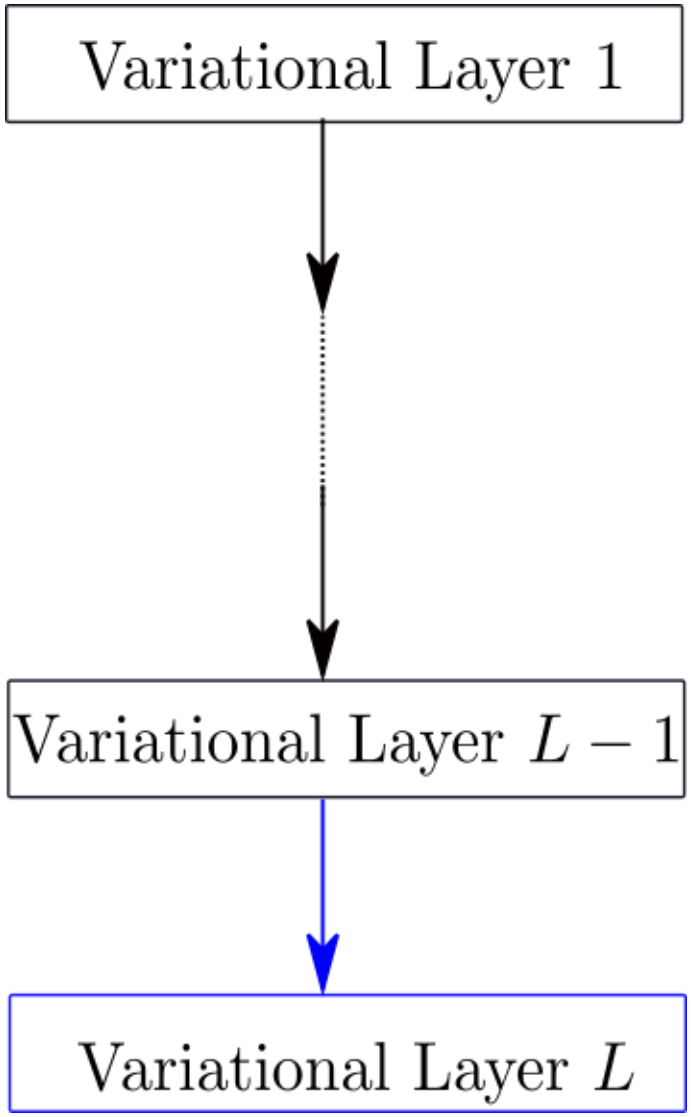
Problem: Locality in Deep VAEs

- During the bottom-up pass, a layer is connected only with the layer below in the hierarchy.



Problem: Locality in Deep VAEs

- During the top-down pass, a layer is connected only with the layer above in the hierarchy.



Problem: Locality in Deep VAEs

- Current hierarchies may overlook long-range latent or deterministic features.

Problem: Locality in Deep VAEs

- Current hierarchies may overlook long-range latent or deterministic features.
- The conditional dependency between z_l and $z_{<l-1}$, in practice may not be respected.
  - For $L = 30$ layers, z_{30} is far away from $z_1, z_2, \ldots$

Problem: Locality in Deep VAEs

- Current hierarchies may overlook long-range latent or deterministic features.
- The conditional dependency between z_l and $z_{<l-1}$, in practice may not be respected.
  - For $L = 30$ layers, z_{30} is far away from $z_1, z_2, \ldots$
- The factorizations may no longer hold in practice:
  - $p(z) = p(z_1) \times \prod p(z_l \mid z_{<l})$.
  - $q(z \mid x) = q(z_1 \mid x) \times \prod q(z_l \mid x, z_{<l})$.

Problem: Locality in Deep VAEs

- Current hierarchies may overlook long-range latent or deterministic features.
- This problem is usually compensated by adding more layers!

Problem: Locality in Deep VAEs

- Performance of NVAE on CIFAR-10 for a different number of layers.
- The predictive gains diminish as depth increases.

Table 1: $-\log p(x)$ **for varying depth** L **(bits/dim).**

Depth (L)	bits/dim ↓	$\Delta(\cdot)\%$
2	3.50	–
4	3.26	−6.8
8	3.06	−6.1
16	2.96	−3.2
30	2.91	−1.7

Idea: Strongly Connected Layers

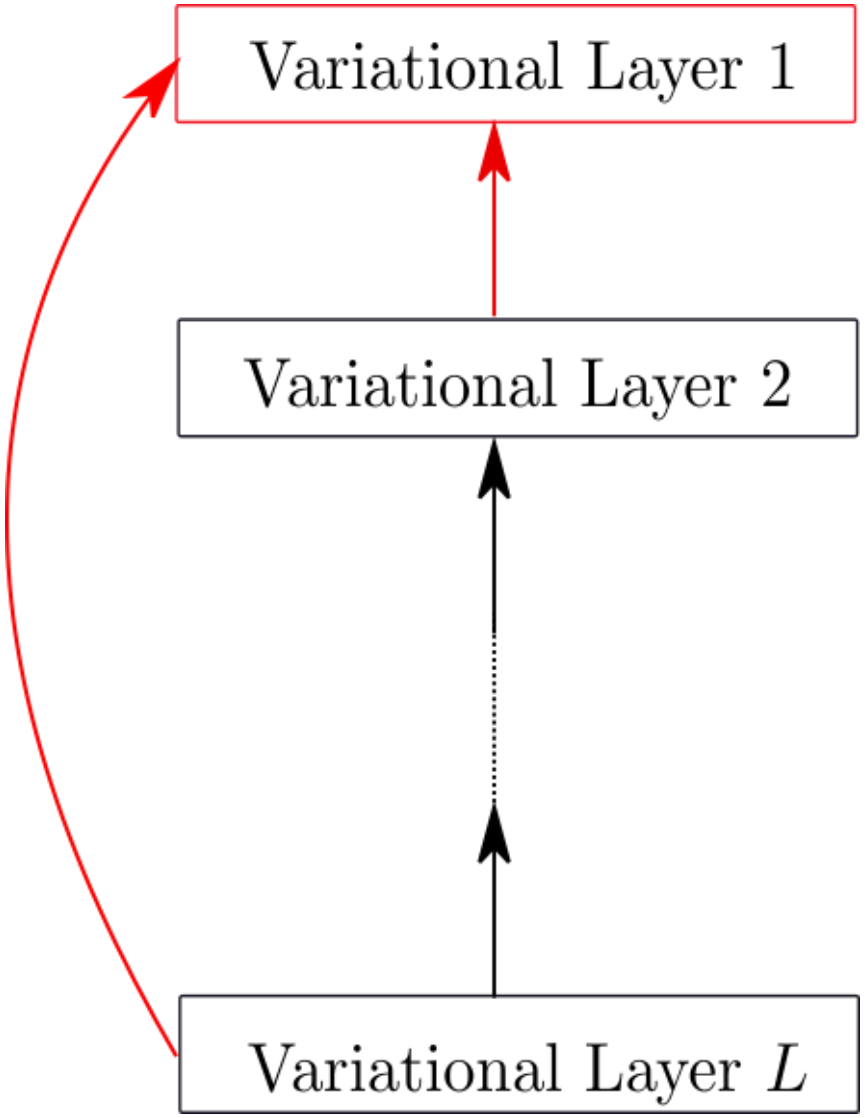
- We enforce couplings between layers.

Idea: Strongly Connected Layers

- We enforce couplings between layers.
- We allow the layer to *dynamically* decide which parts of the contexts are critical to inference.

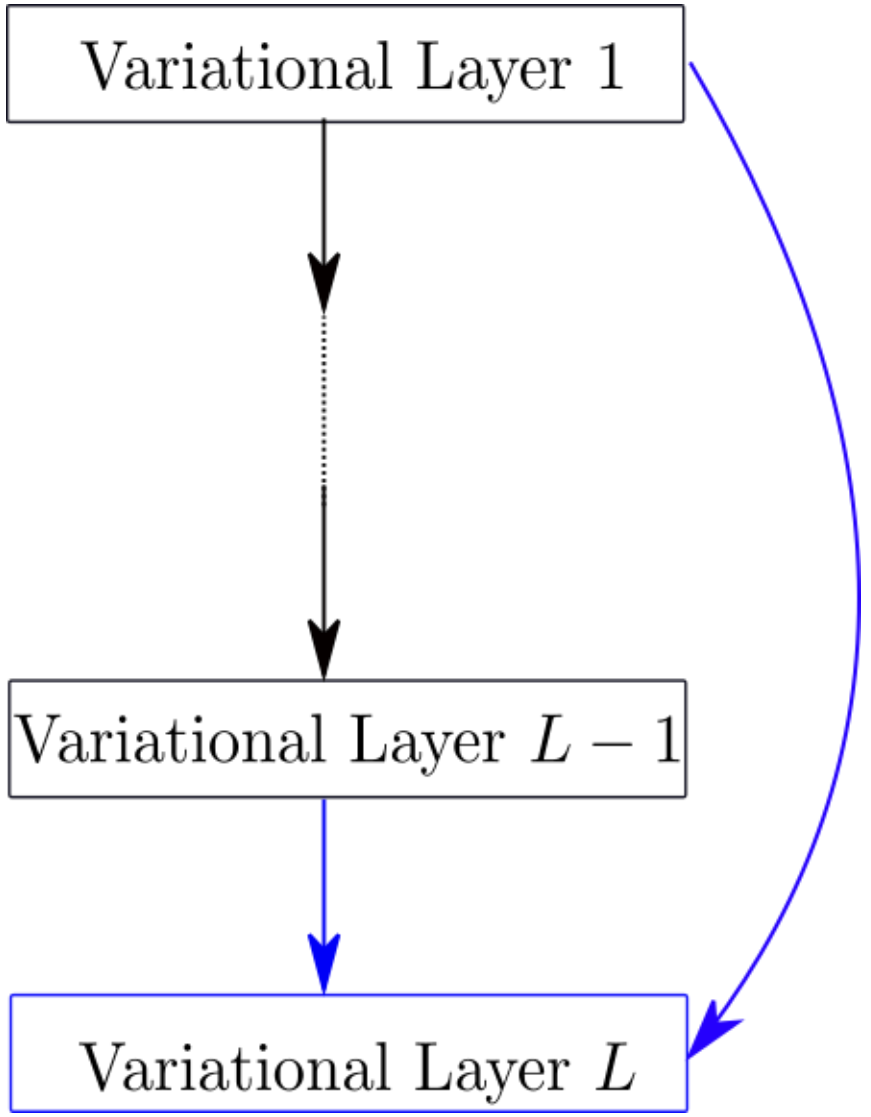
Idea: Strongly Connected Layers

- During the bottom-up pass, a layer is connected with *all* layers below in the hierarchy.



Idea: Strongly Connected Layers

- During the top-down pass, a layer is connected with *all* layers above in the hierarchy.

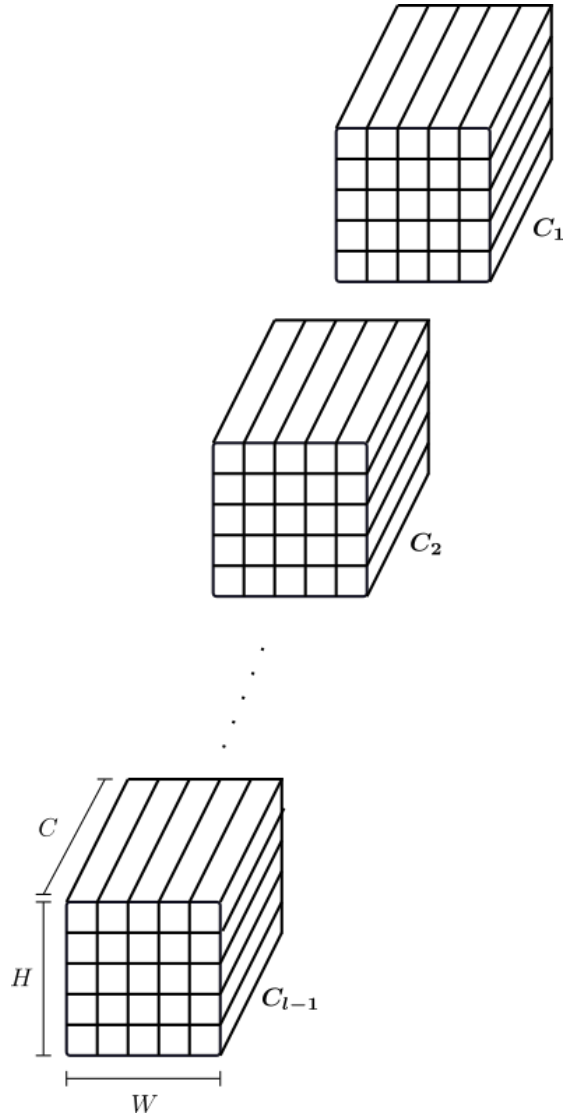


Depth-Wise Attention

- The technical tool that let us realize the strong couplings between layers.

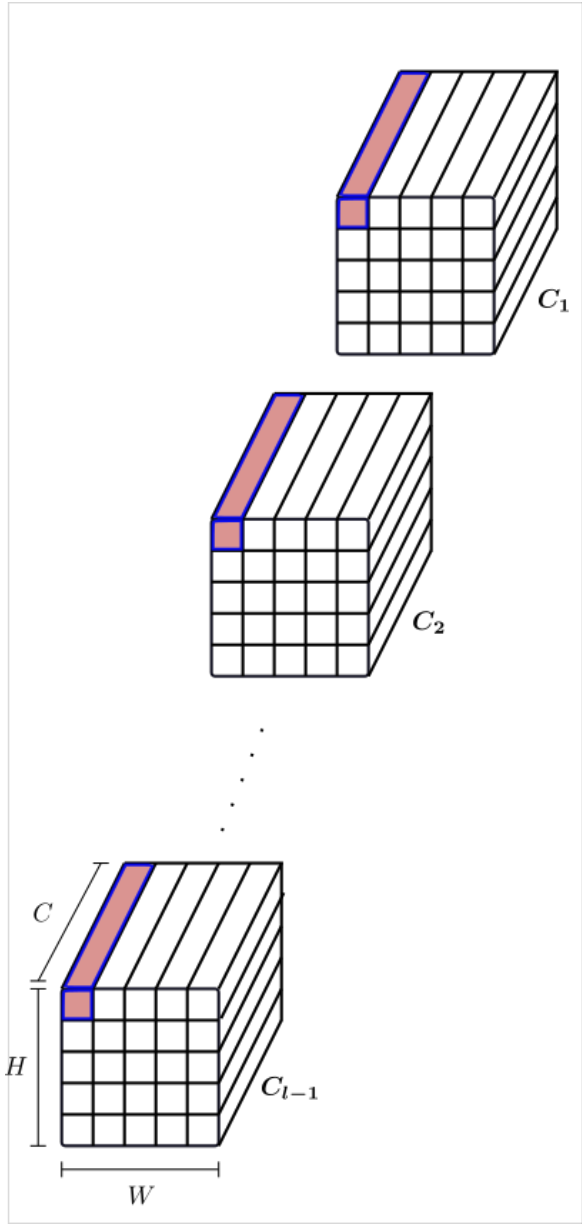
Depth-Wise Attention

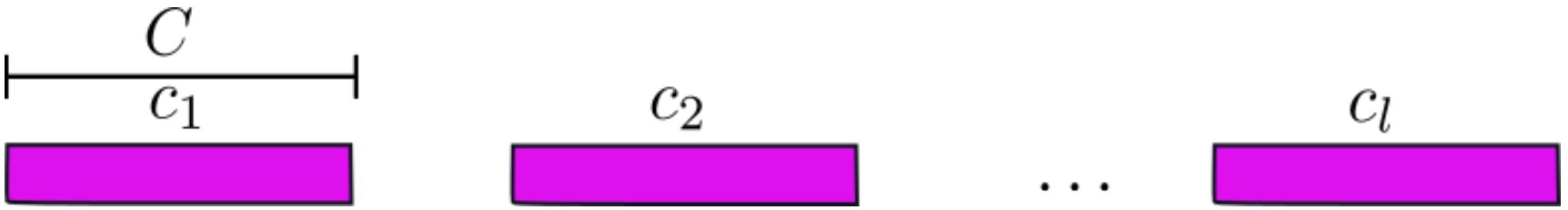
- **Problem**: We must be able to handle *long* sequences of *large* 3D context tensors.



Depth-Wise Attention

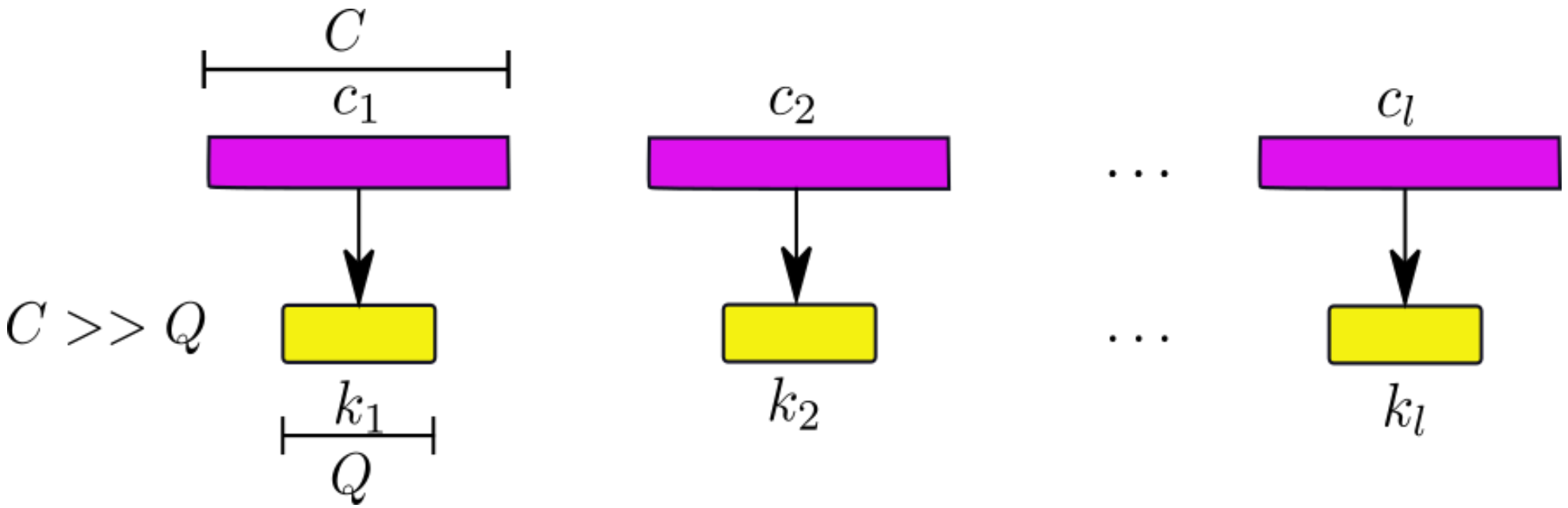
- **Solution**: Handle $H \times W$ pixel sequences of C— dimensional features independently.



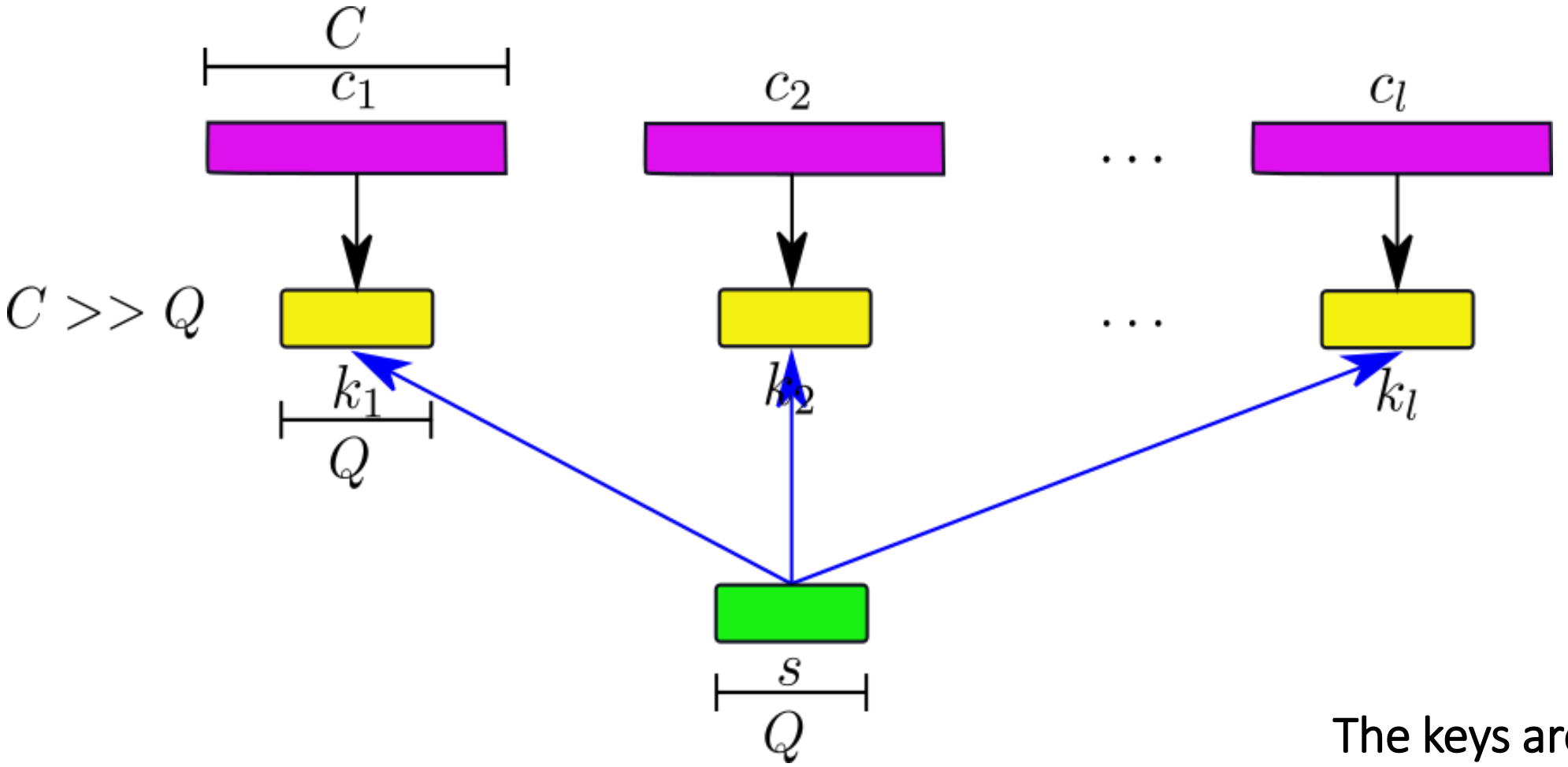


The sequence of contexts for each pixel is processed independently from the rest.

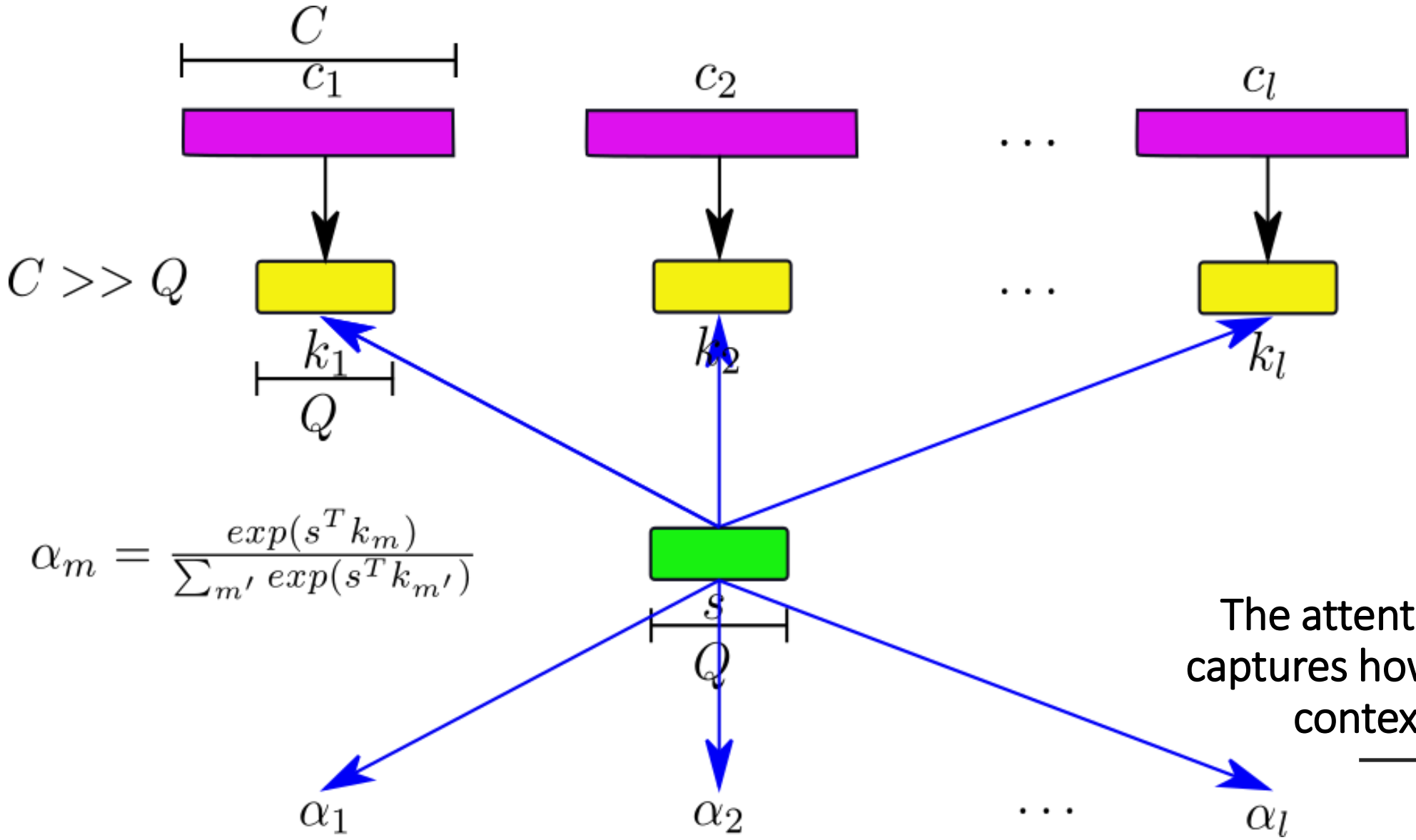
1/5



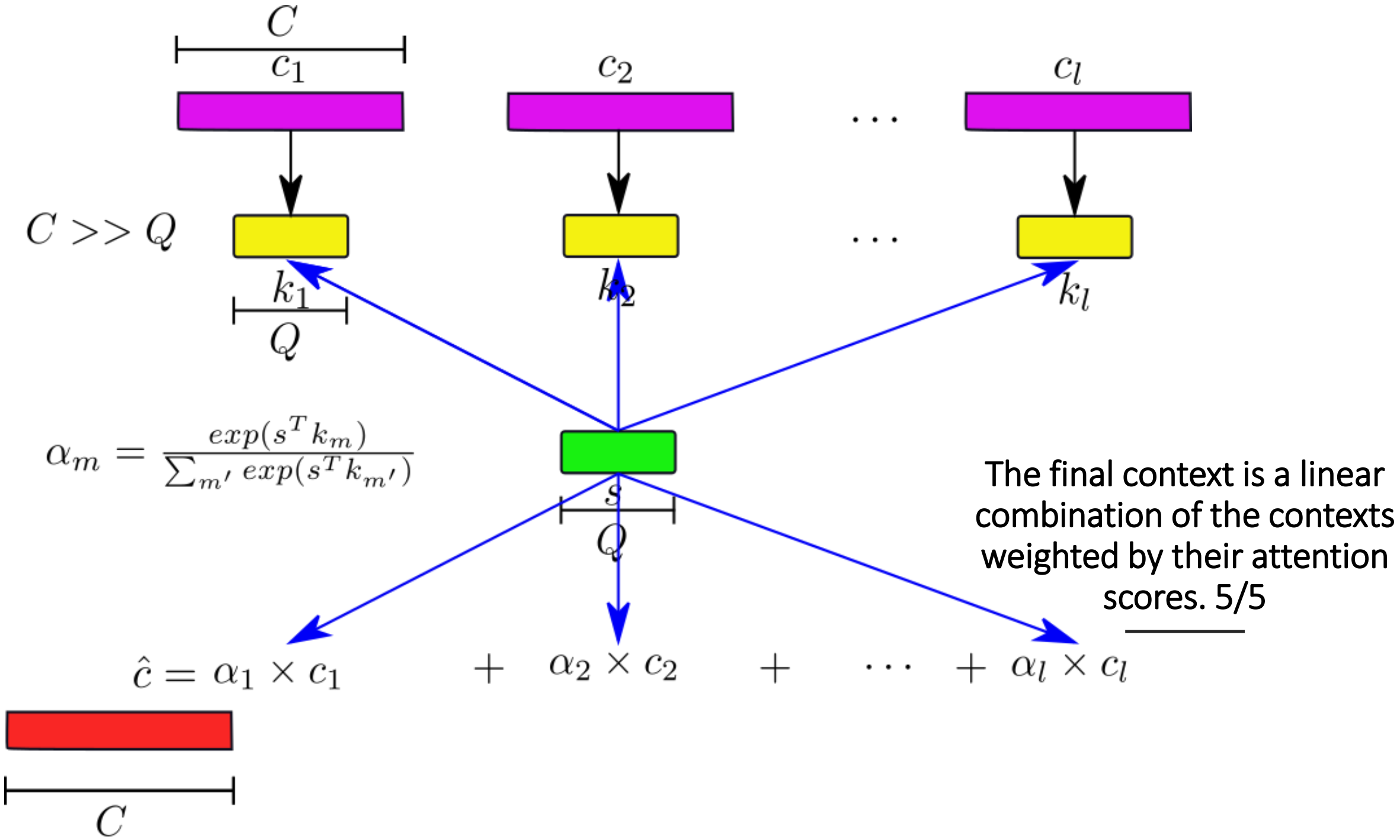
Each context is represented by a key of lower dimension.



The keys are queried by feature $\boldsymbol{s}$. 3/5



The attention score $\boldsymbol{\alpha_l}$ captures how important is context $\boldsymbol{c_l}$. 4/5



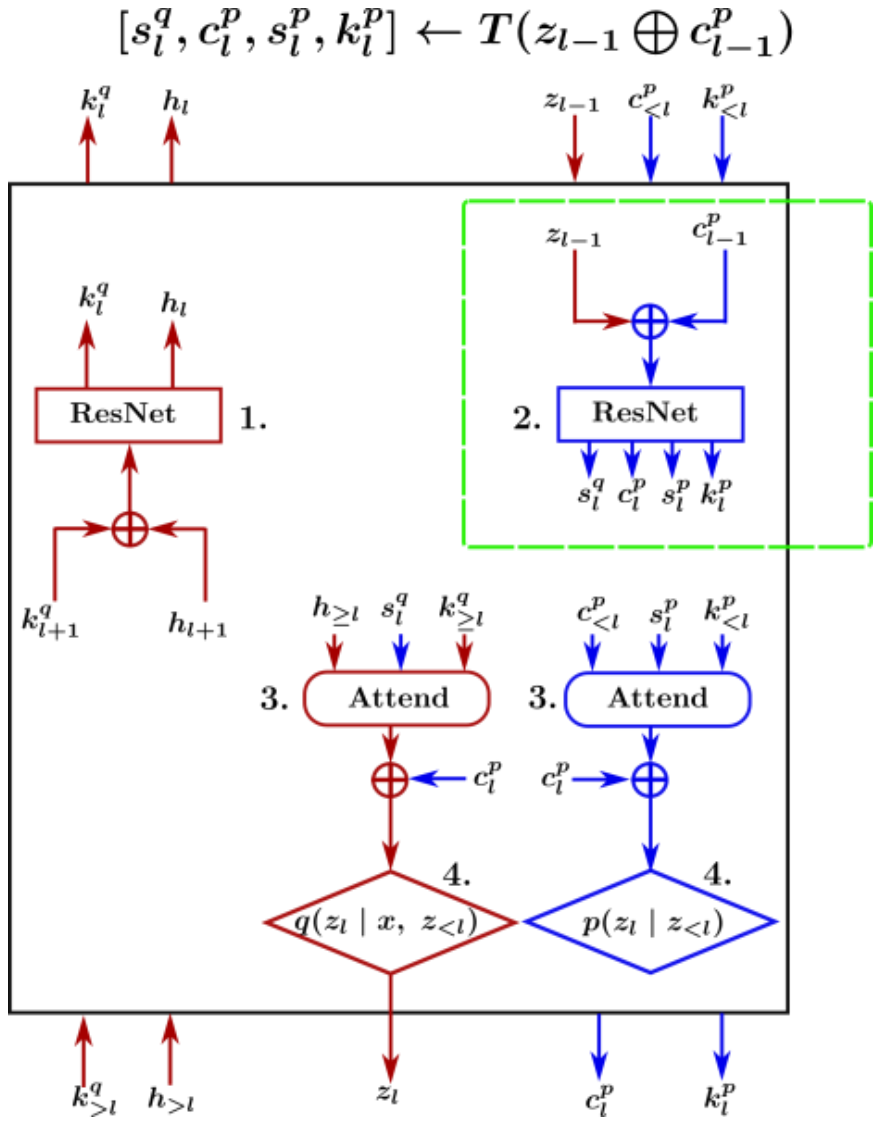
The final context is a linear combination of the contexts weighted by their attention scores. 5/5

Attentive VAE

- Each layer attends to context provided by *all* previous layers when forming its prior and posterior beliefs.

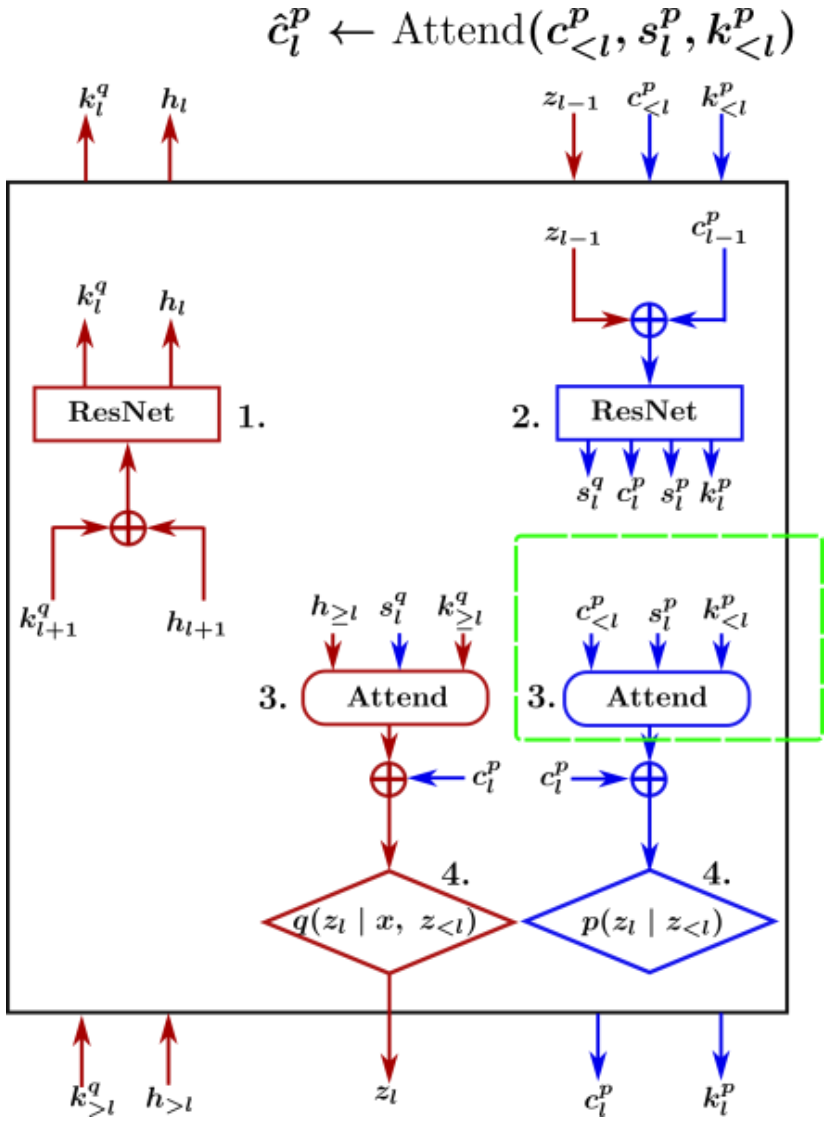
Attentive VAE

- The layer generates contexts c, keys k, and queries s.



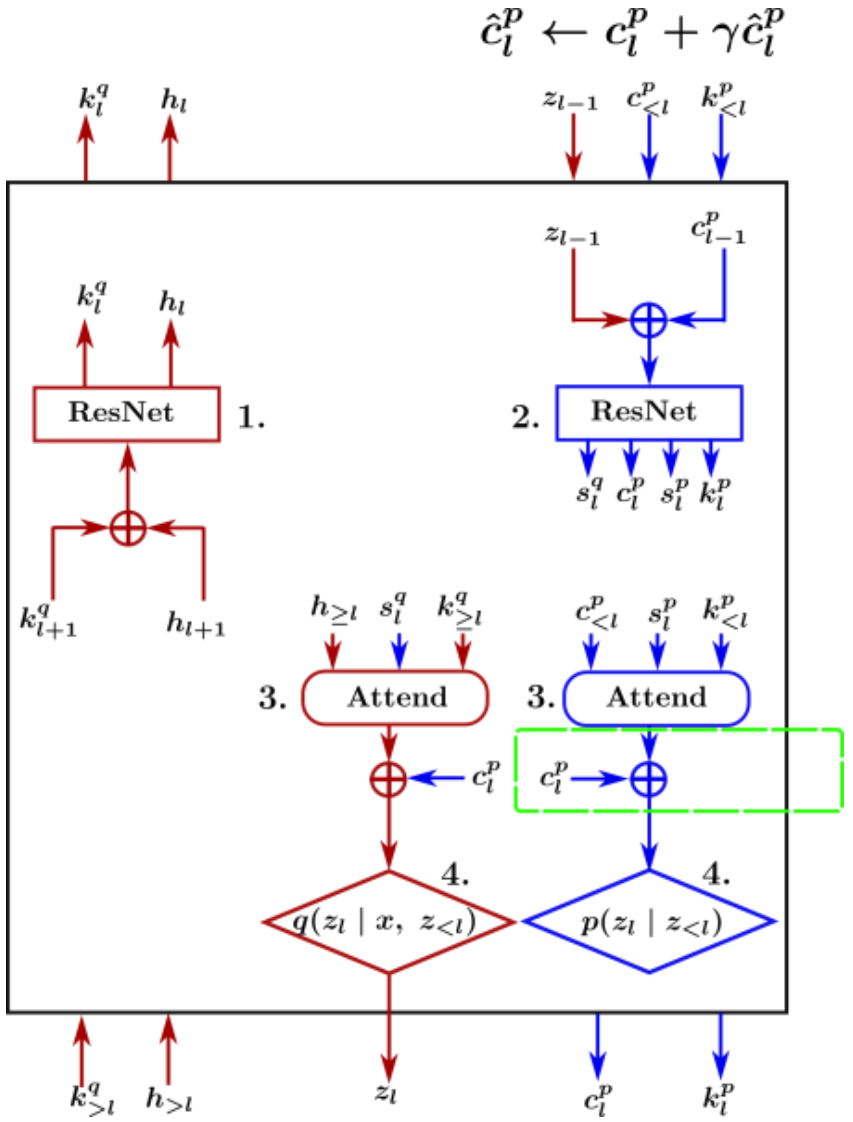
Attentive VAE

- The generative model attends to stochastic contexts of layers above in the hierarchy $c^p_{<l}$ according to their keys $k^p_{<l}$ and its query s^p_l.



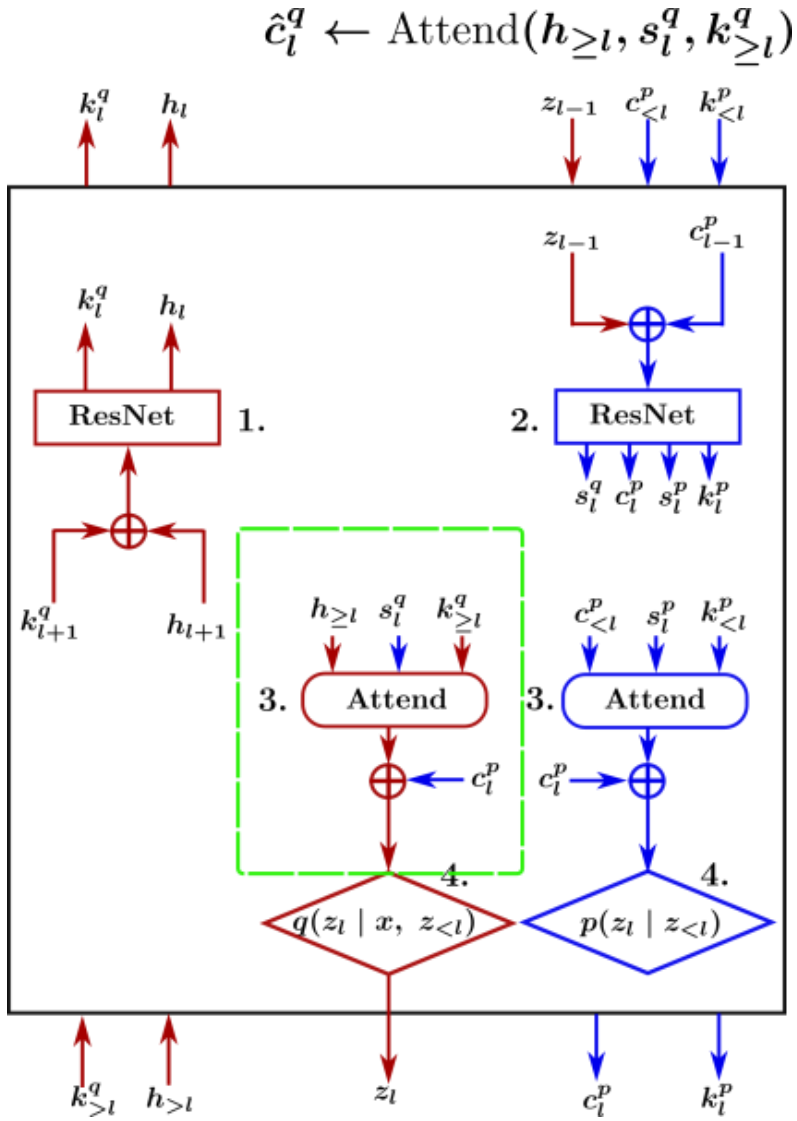
Attentive VAE

- Initially, we let the layer rely only on the local context c_l^p via a residual connection.



Attentive VAE

- A similar procedure is applied to the inference model.

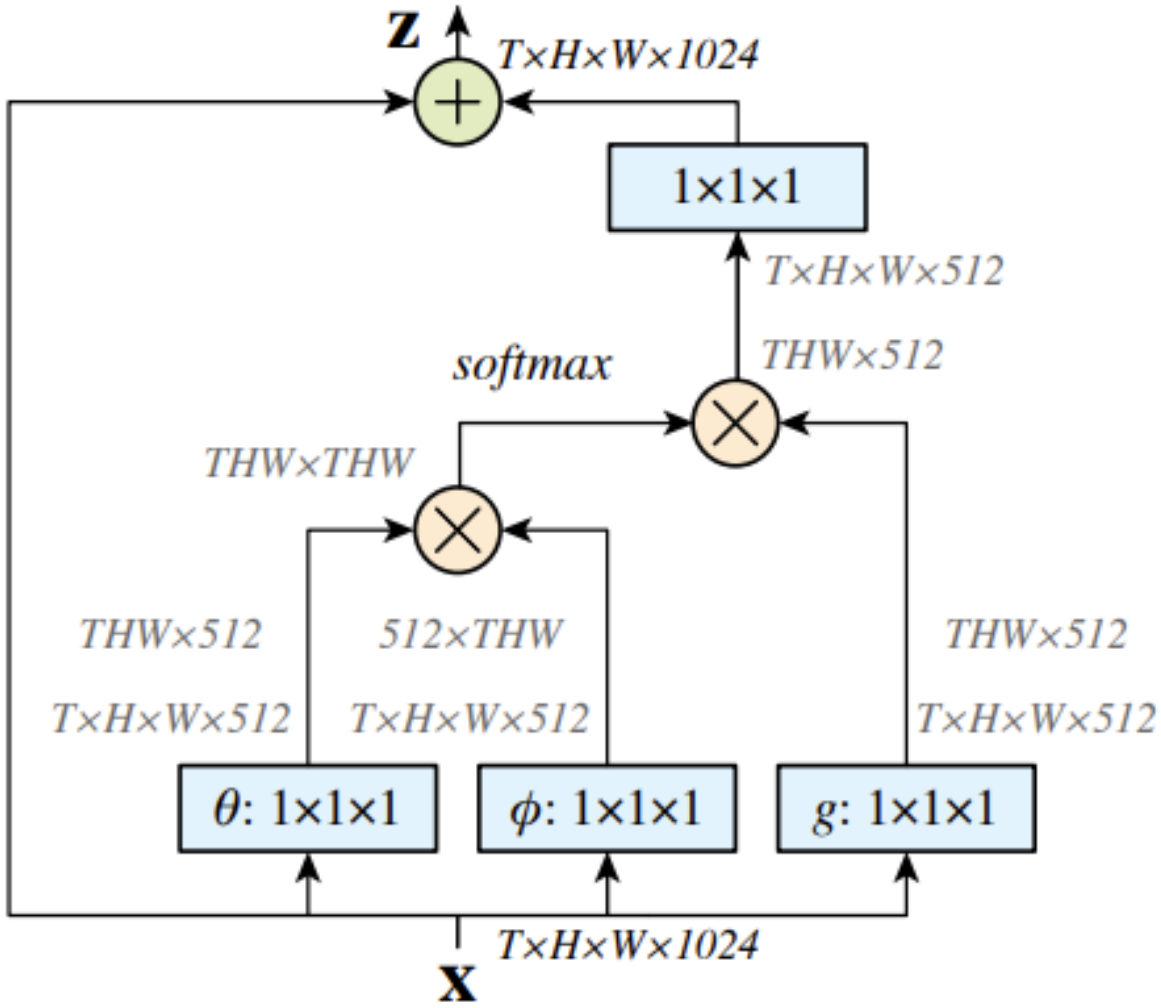


Spatially Non-Local ResNet cells

- We still need to take advantage of latent information that is far away in the spatial domain.
- This occurs at a second stage by interleaving non-local blocks.

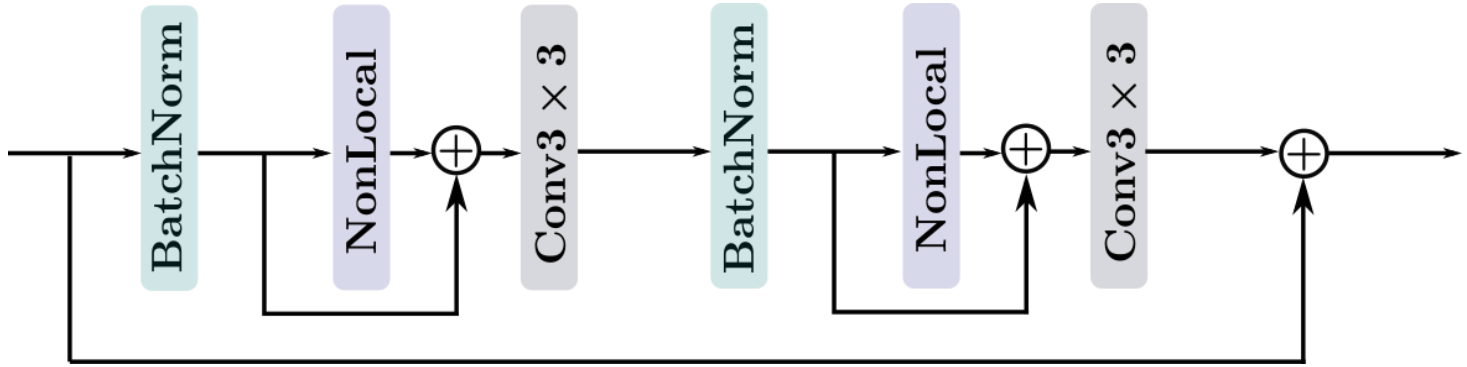
Spatially Non-Local ResNet cells

Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.



Spatially Non-Local ResNet cells

- Non-local operations are interleaved with convolutions to capture inter-pixel long-range interactions in the same layer.



Taming the KL term

Residual variational distributions for training stability.

$$p(z_l \mid z_{<l}) = N(\mu(z_{<l}), \sigma(z_{<l}))$$

$$q(z_l \mid x, z_{<l}) = N(\mu(x, z_{<l}) + \mu(z_{<l}), \sigma(x, z_{<l})\sigma(z_{<l}))$$

Vahdat, Arash, and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder." *Advances in Neural Information Processing Systems* 33 (2020).

Taming the KL term

KL annealing for mitigating posterior collapse.

$$\mathbb{E}_q[\log p(\boldsymbol{x} \mid \boldsymbol{z})] - \beta D_{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})), \quad \beta_0 \leq \beta \leq 1, \ \beta_0 < 1,$$

Sønderby, Casper Kaae, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. "Ladder variational autoencoders." *Advances in neural information processing systems* 29 (2016).

Experiments

Better predictive performance with fewer layers.

Table 3: **CIFAR-10** (Krizhevsky et al., 2009) **performance on the test set.** The marginal log-likelihood is estimated with 100 importance samples. A shallower Attentive VAE outperforms all state-of-the-art VAEs with or without autoregressive components. Attentive VAE performs on par with fully autoregressive generative models. However, it permits fast sampling that requires a single network evaluation per sample as opposed to D, where D the dimension of the data distribution.

Model	VAE	Depth (L)	Autoregressive Decoder	$-\log p(\boldsymbol{x}) \leq$ (bits/dim) $\downarrow$
Attentive VAE (ours) trained for 400 epochs	✓	16	✗	**2.82**
Attentive VAE (ours) trained for 500 epochs	✓	16	✗	**2.81**
Attentive VAE (ours) trained for 900 epochs	✓	16	✗	**2.79**
Very Deep VAE (Child, 2020)	✓	45	✗	2.87
NVAE (Vahdat & Kautz, 2020)	✓	30	✗	2.91
BIVA (Maaløe et al., 2019)	✓	15	✗	3.08
IAF-VAE (Kingma et al., 2016)	✓	12	✗	3.11
δ-VAE (Razavi et al., 2019a)	✓		✓	2.83
PixelVAE++ (Sadeghi et al., 2019)	✓		✓	2.90
Lossy VAE (Chen et al., 2017)	✓		✓	2.95
MAE (Ma et al., 2019)	✓		✓	2.95
PixelCNN++ (Salimans et al., 2017)	✗		✓	2.92
PixelSNAIL (Chen et al., 2018)	✗		✓	2.85
Image Transformer (Parmar et al., 2018)	✗		✓	2.90
Sparse Transformer (Child et al., 2019)	✗		✓	**2.80**

Experiments

Fewer layers decrease training and inference time.

Table 4: **Comparison of the computational requirements for training deep state-of-the-art VAE models.** All models are trained on 32GB V100 GPUs. The additional cost for computing the attention scores is compensated by the smaller number of stochastic layers in the hierarchy without sacrificing the generative capacity of the model, see Table 3.

Model	batch size / GPU	# GPUs	Training Time	Total GPU hours
Attentive VAE (ours), 400 epochs	32	4	68 hours	272
Attentive VAE (ours), 500 epochs	32	4	84 hours	336
Attentive VAE (ours), 900 epochs	32	4	152 hours	608
NVAE	32	8	55 hours	440
Very Deep VAE	32	2	6 days	288

Experiments

Question: Where did this improvement come from?

Experiments

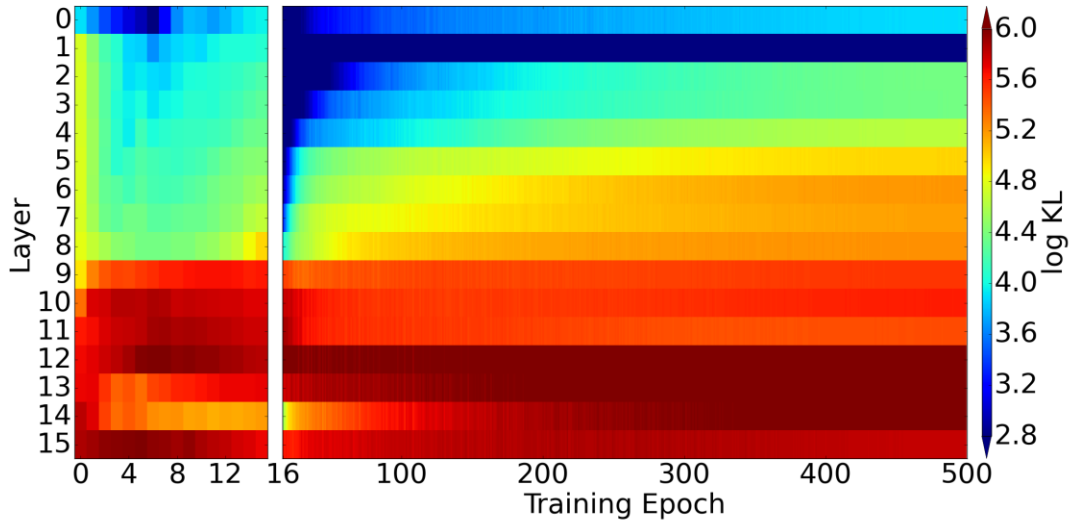
Question: Where did this improvement come from?

Answer: Attention leads to better utilization of the latent space.

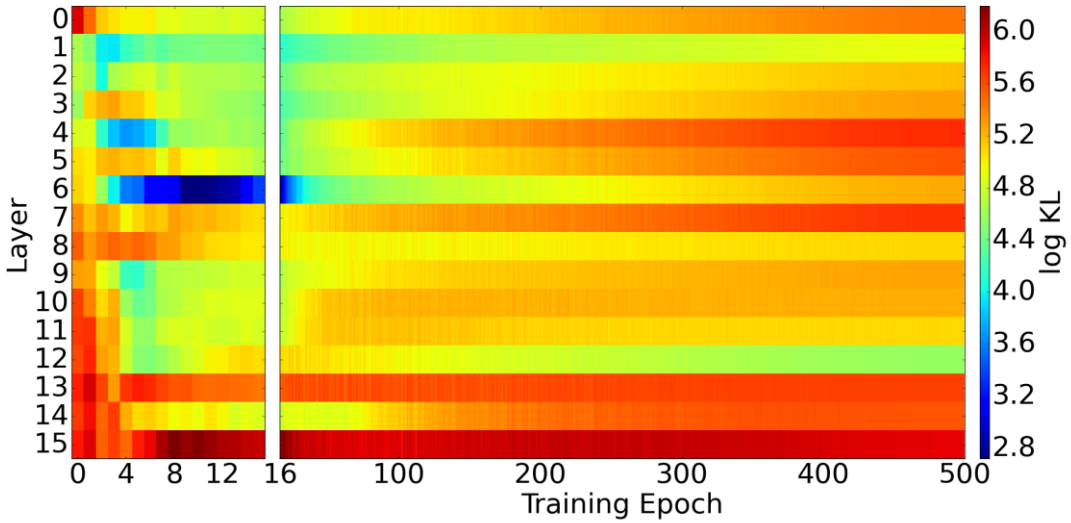
Latent Space Utilization

The additional couplings help mitigate posterior collapse.

Local VAE

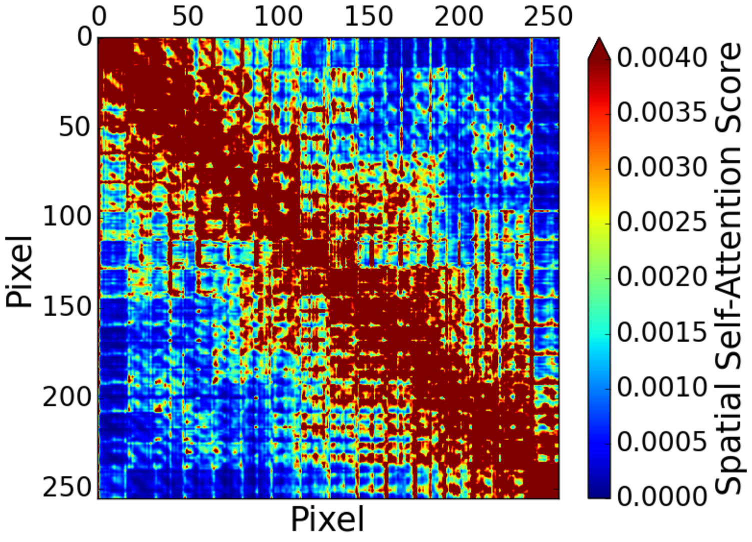


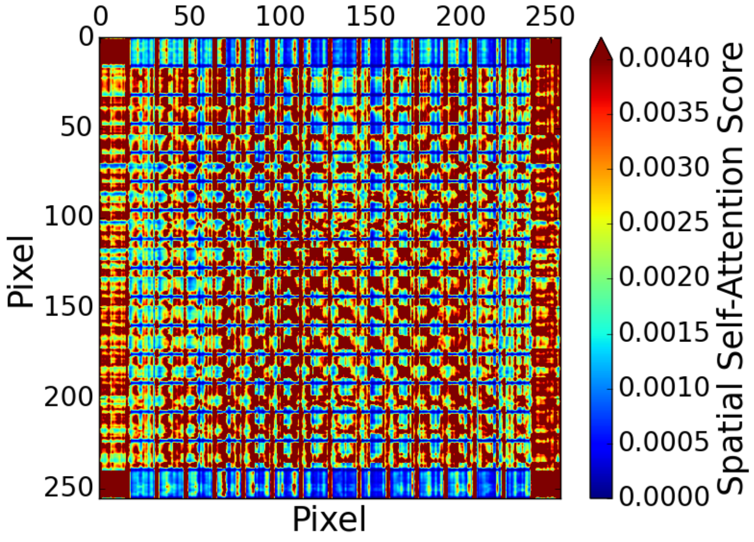
Attentive VAE

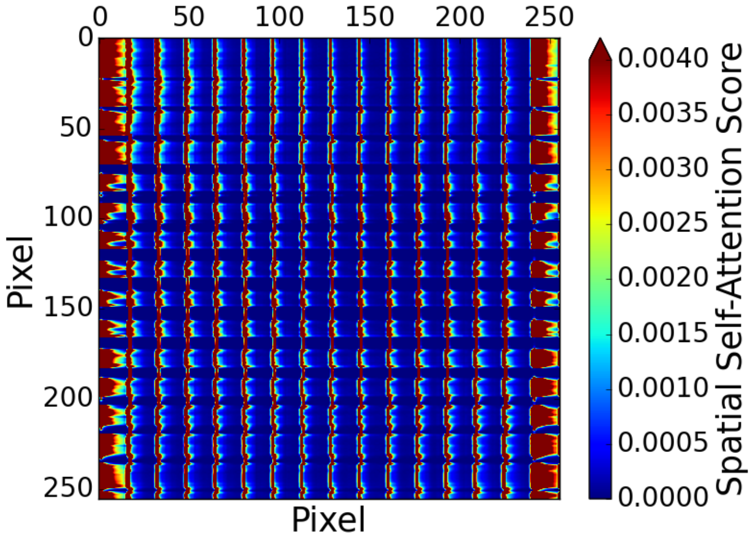


Spatial Attention Patterns

The spatial attention patterns are sparse and highly structured.

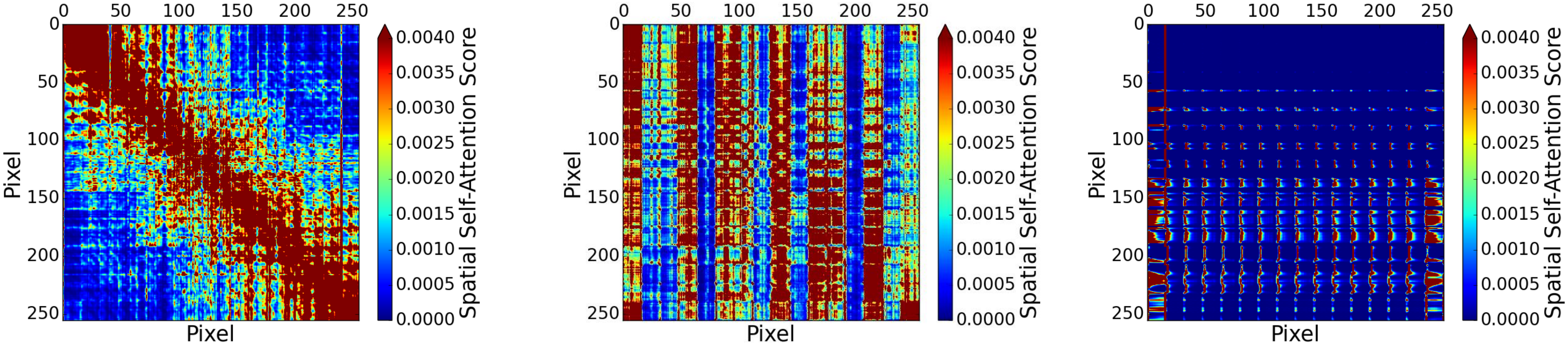




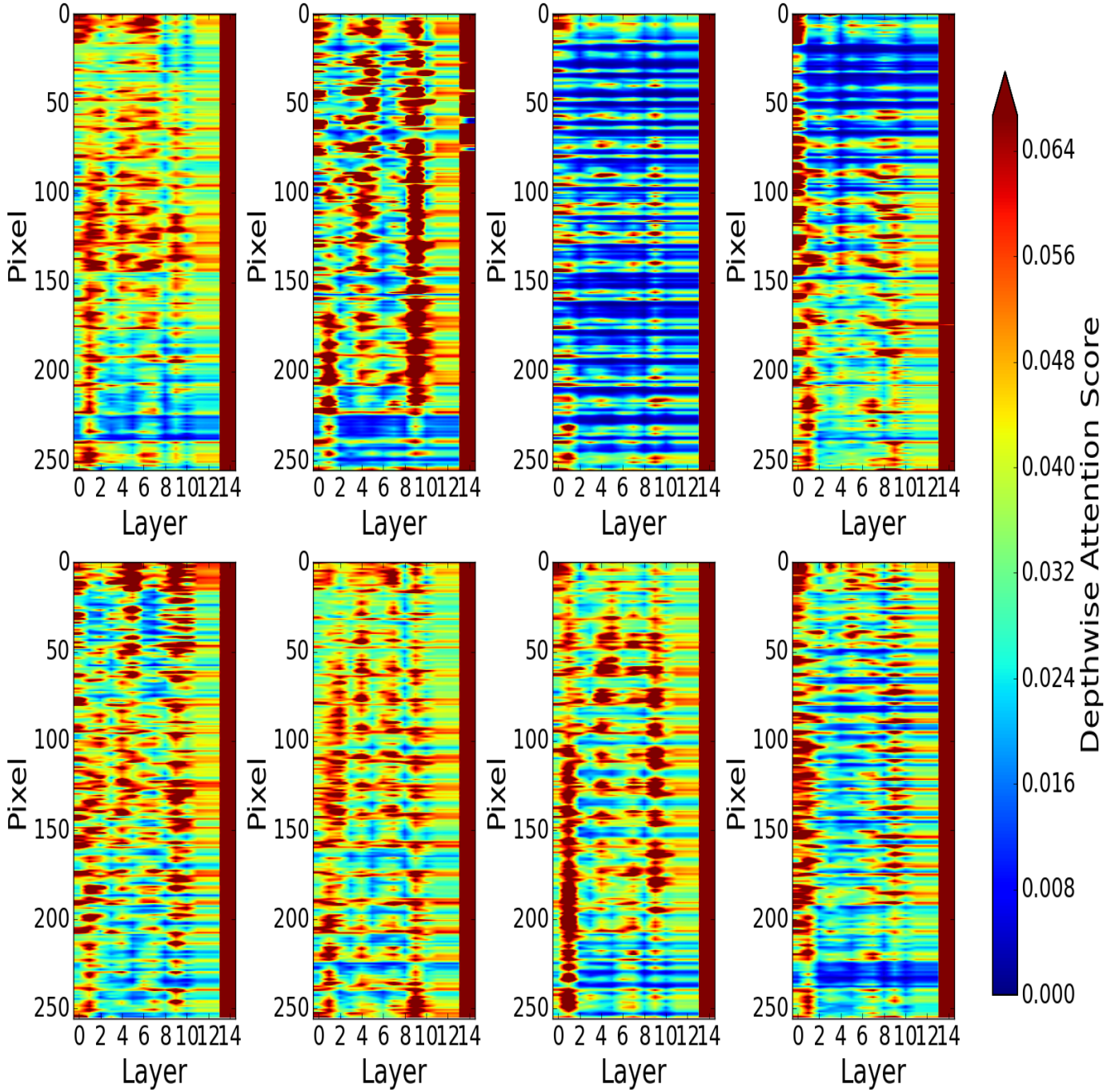


Spatial Attention Patterns

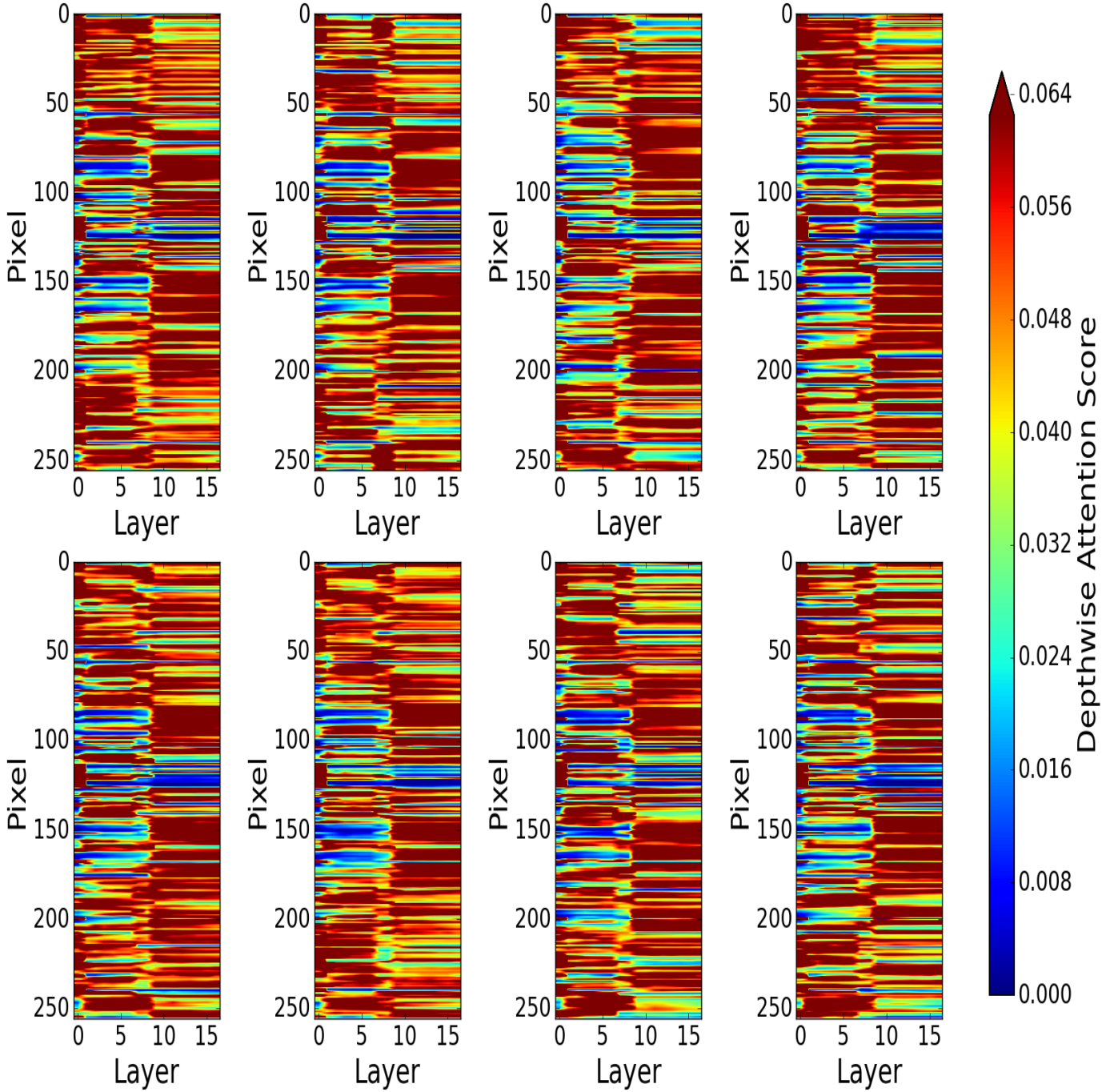
The spatial attention patterns are sparse and highly structured.



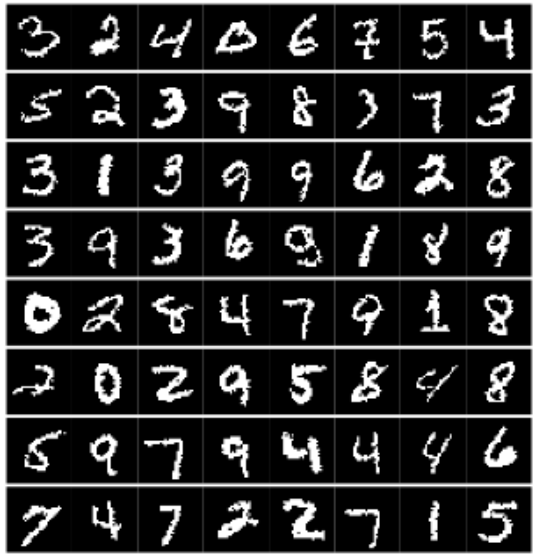
Generative Attention Patterns



Inference Attention Patterns

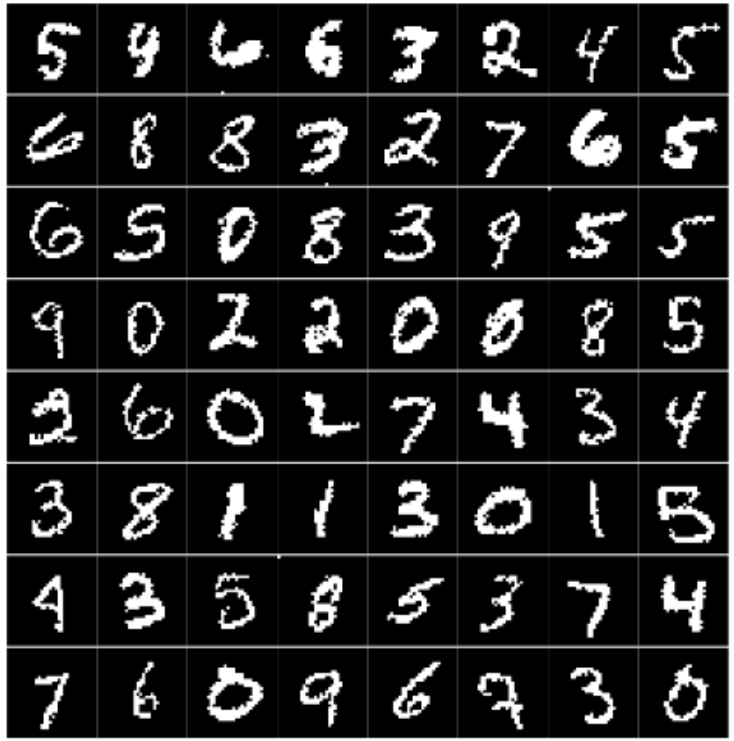


Reconstructed & Novel Images



(a) Original MNIST images.

(b) Reconstructed MNIST images.

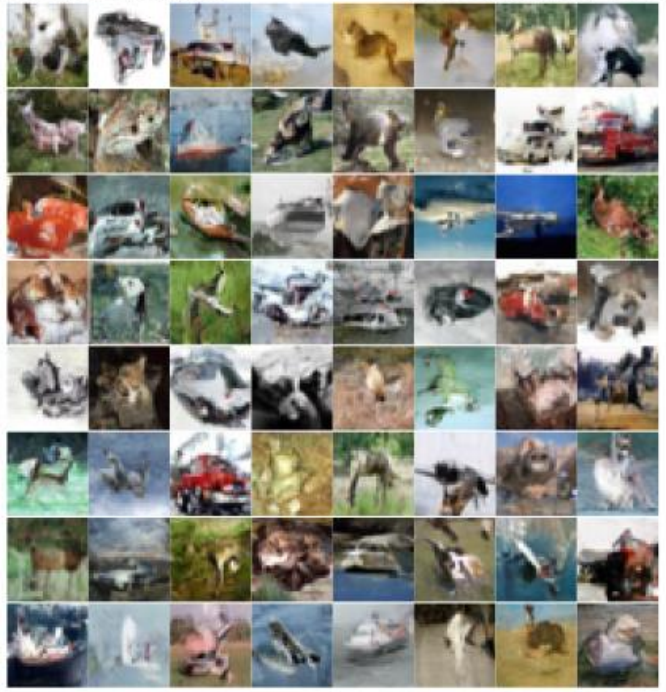


(c) Uncurated generated MNIST images.

Reconstructed & Novel Images

(a) Original CIFAR-10 images.

(b) Reconstructed CIFAR-10 images.



(c) Uncurated generated CIFAR-10 images.

Discussion

- Depth-wise attention on general ResNet architectures for different tasks.

- Efficient attention approximations specific to variational inference.

Thank you!

ifiaposto@gmail.com, iapostol@andrew.cmu.edu