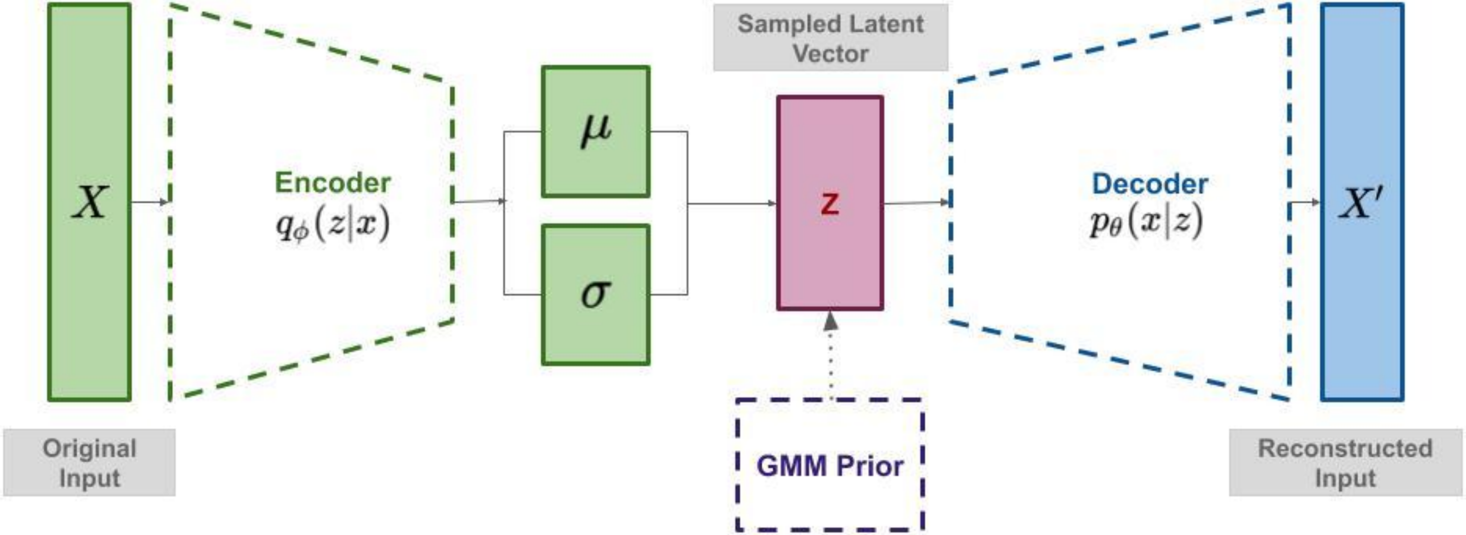
NVAE: A Deep Hierarchical Variational Autoencoder

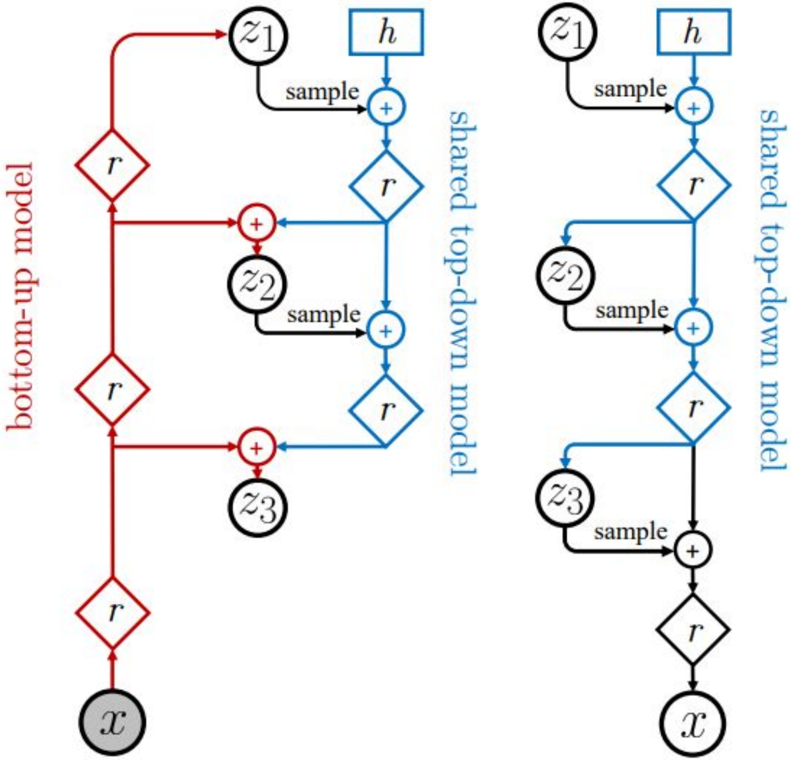
Dmitry Gradoboev, 171.

Vanilla VAE



X
Original Input
Encoder
$q_\phi(z|x)$
μ
σ
Sampled Latent Vector
z
GMM Prior
Decoder
$p_\theta(x|z)$
X'
Reconstructed Input

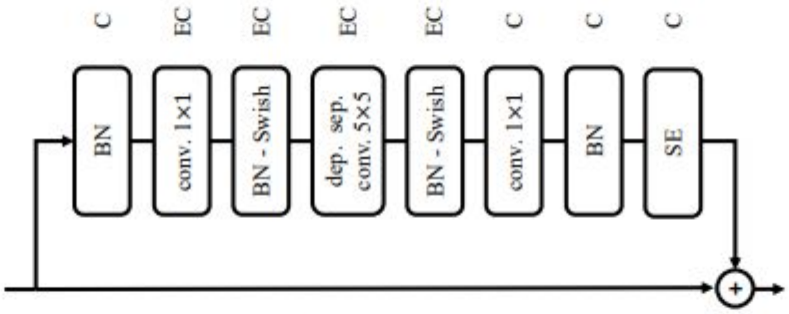
Hierarchical VAE



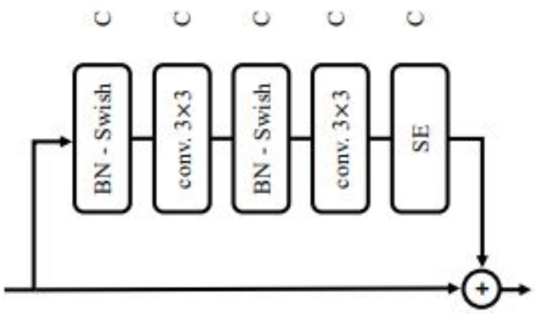
h - trainable parameter
r - residual neural network

z_1 - 8x8
z_2 - 16x16
…
z_n - 128x128

Residual Cells



(a) Residual Cell for NVAE Generative Model



(b) Residual Cell for NVAE Encoder

Depthwise convolutions for generative part.

BN hurts performance for vanilla VAE, but it's "hacked" in NVAE.

Swish activation: $f(u) = \frac{u}{1+e^{-u}}$

Squeeze and Excitation: channel-wise gating layer.

Depthwise convolution

Memory requirement.

1) **NVIDIA APEX**

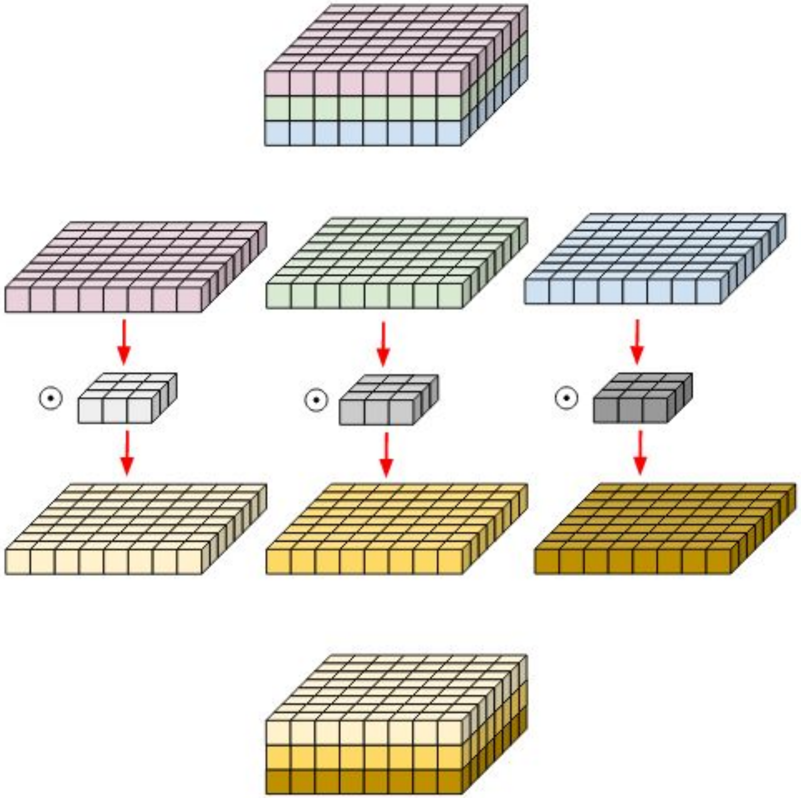
   Mixed precision training

   ~40%

2) **Gradient check-pointing**

   Fuse BN and Swish

   CIFAR-10 ~18%



Taming the Unbounded KL Term

1. Residual Normal Distributions.

   Let $p\left(z_l^i \mid z_{<l}\right) := \mathcal{N}\left(\mu_i\left(z_{<l}\right), \sigma_i\left(z_{<l}\right)\right)$ be a Normal distribution for the ith variable in z_l prior.

   $q\left(z_l^i \mid z_{<l}, x\right) := \mathcal{N}\left(\mu_i\left(z_{<l}\right) + \Delta\mu_i\left(z_{<l}, x\right), \sigma_i\left(z_{<l}\right) \cdot \Delta\sigma_i\left(z_{<l}, x\right)\right)$

   KL term in L_{VAE}: $KL(q(z^i \mid x) \parallel p(z^i)) = \frac{1}{2}\left(\frac{\Delta\mu_i^2}{\sigma_i^2} + \Delta\sigma_i^2 - \log\Delta\sigma_i^2 - 1\right)$

2. Spectral Regularization.

   Formally, we add $L_{SR} = \lambda \sum_i s^{(i)}$ to L_{VAE}, where $s^{(i)}$ is the largest singular value of the ith conv layer.

Quantitative results

The performance is measured in bits/dimension (bpd) for all the datasets but MNIST in which negative log-likelihood in nats is reported (lower is better in all cases).

For large image datasets such as CelebA HQ and FFHQ, NVAE consists of 36 groups of latent variables starting from 8 × 8 dims, scaled up to 128 × 128 dims with two residual cells per latent variable groups.

Method	MNIST 28×28	CIFAR-10 32×32	ImageNet 32×32	CelebA 64×64	CelebA HQ 256×256	FFHQ 256×256
NVAE w/o flow	**78.01**	2.93	-	2.04	-	0.71
NVAE w/ flow	78.19	**2.91**	3.92	**2.03**	**0.70**	**0.69**
VAE Models with an Unconditional Decoder						
BIVA [36]	78.41	3.08	3.96	2.48	-	-
IAF-VAE [4]	79.10	3.11	-	-	-	-
DVAE++ [20]	78.49	3.38	-	-	-	-
Conv Draw [42]	-	3.58	4.40	-	-	-
Flow Models <u>without</u> any Autoregressive Components in the Generative Model						
VFlow [59]	-	2.98	-	-	-	-
ANF [60]	-	3.05	3.92	-	0.72	-
Flow++ [61]	-	3.08	**3.86**	-	-	-
Residual flow [50]	-	3.28	4.01	-	0.99	-
GLOW [62]	-	3.35	4.09	-	1.03	-
Real NVP [63]	-	3.49	4.28	3.02	-	-
VAE and Flow Models with Autoregressive Components in the Generative Model						
δ-VAE [25]	-	2.83	3.77	-	-	-
PixelVAE++ [35]	78.00	2.90	-	-	-	-
VampPrior [64]	78.45	-	-	-	-	-
MAE [65]	77.98	2.95	-	-	-	-
Lossy VAE [66]	78.53	2.95	-	-	-	-
MaCow [67]	-	3.16	-	-	0.67	-
Autoregressive Models						
SPN [68]	-	-	3.85	-	0.61	-
PixelSNAIL [34]	-	2.85	3.80	-	-	-
Image Transformer [69]	-	2.90	3.77	-	-	-
PixelCNN++ [70]	-	2.92	-	-	-	-
PixelRNN [41]	-	3.00	3.86	-	-	-
Gated PixelCNN [71]	-	3.03	3.83	-	-	-

Normalization and Activation Functions

Table 2: Normalization & activation

Functions	$L = 10$	$L = 20$	$L = 40$
WN + ELU	3.36	3.27	3.31
BN + ELU	3.36	3.26	3.22
BN + Swish	**3.34**	**3.23**	**3.16**

Residual Cells

Table 3: Residual cells in NVAE

Bottom-up model	Top-down model	Test (bpd)	Train time (h)	Mem. (GB)
Regular	Regular	3.11	43.3	6.3
Separable	Regular	3.12	49.0	10.6
Regular	Separable	**3.07**	48.0	10.7
Separable	Separable	**3.07**	50.4	14.9

Residual Normal Distributions

Table 4: The impact of residual dist.

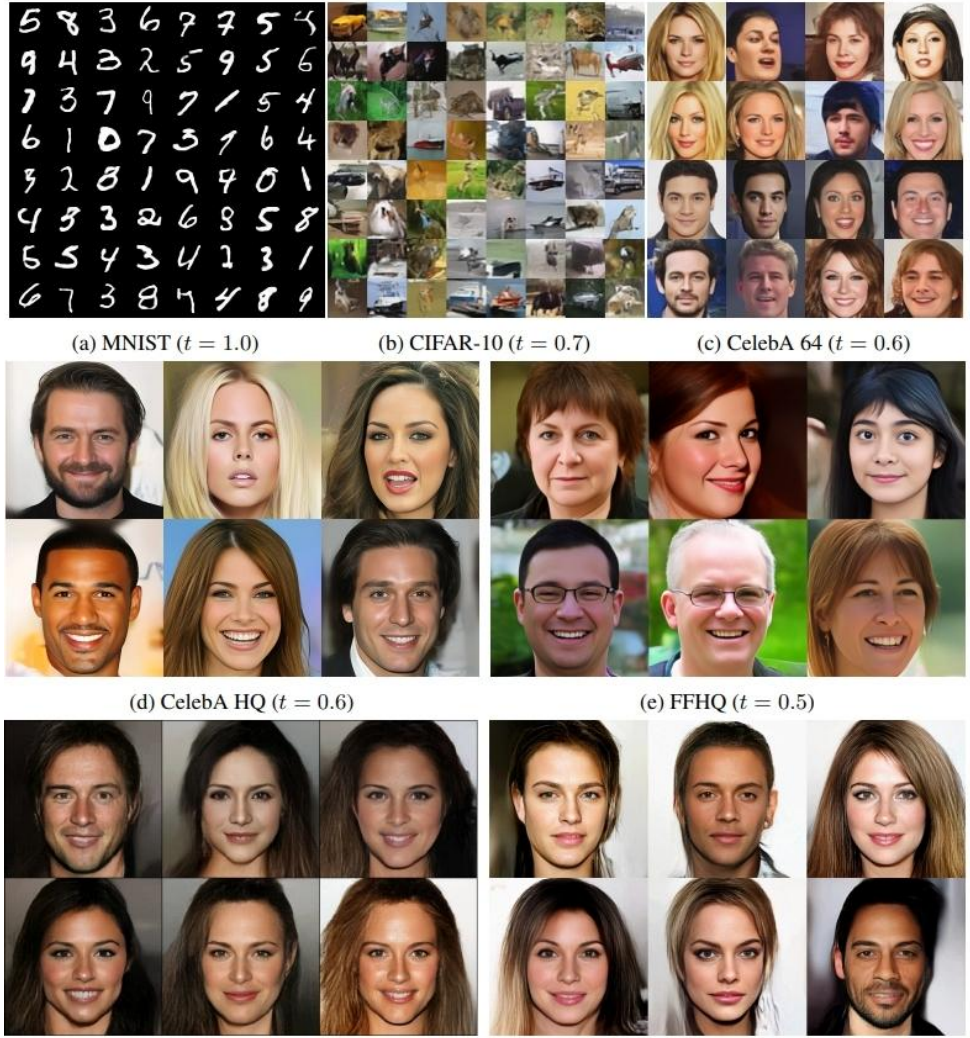
Model	# Act. $\boldsymbol{z}$	Training KL	Rec.	$\mathcal{L}_{\text{VAE}}$	Test LL
w/ Res. Dist.	53	**1.32**	1.80	**3.12**	**3.16**
w/o Res. Dist.	54	1.36	1.80	3.16	3.19

The effect of SR and SE

Table 5: SR & SE

Model	Test NLL
NVAE	**3.16**
NVAE w/o SR	3.18
NVAE w/o SE	3.22

Qualitative Results



(a) MNIST ($t = 1.0$) (b) CIFAR-10 ($t = 0.7$) (c) CelebA 64 ($t = 0.6$)

(d) CelebA HQ ($t = 0.6$) (e) FFHQ ($t = 0.5$)

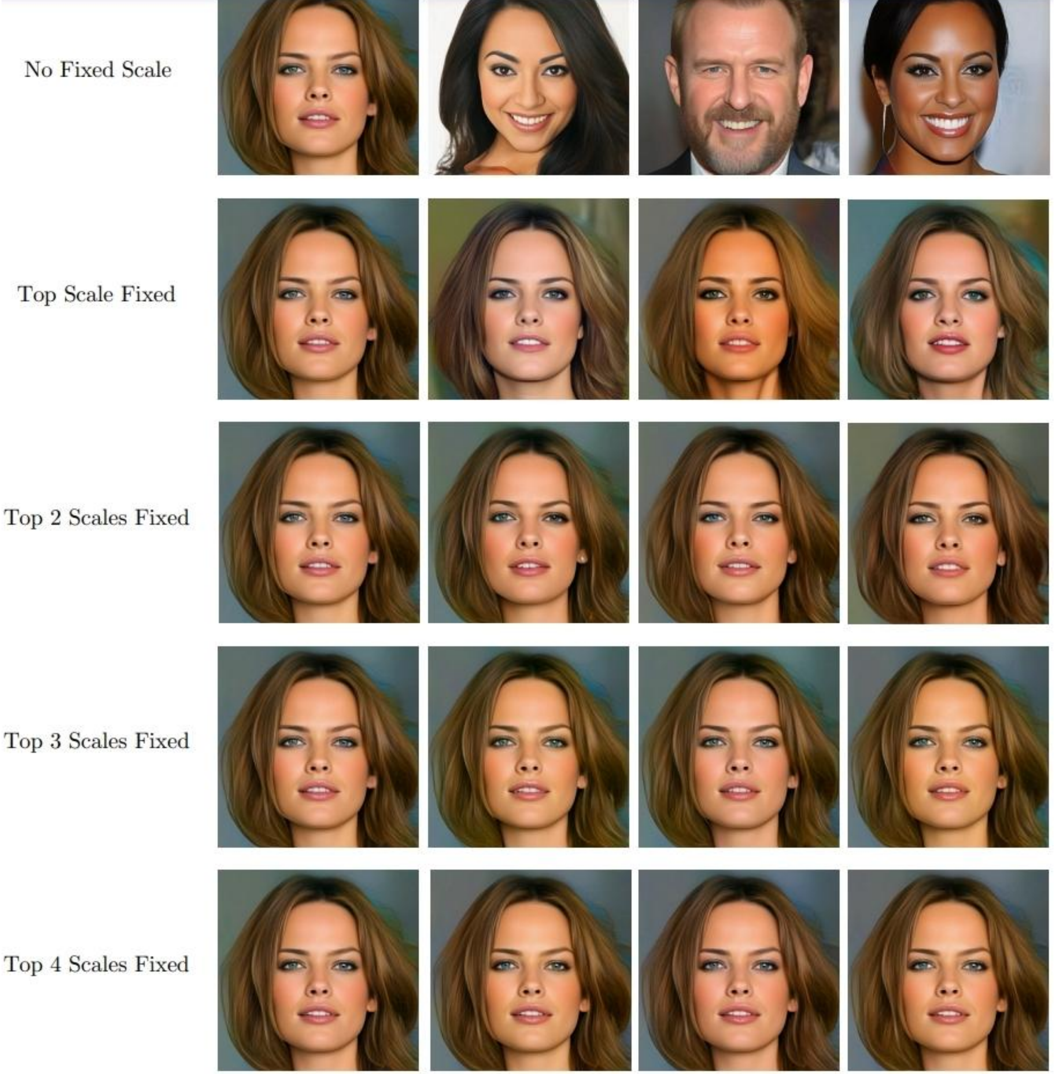
(f) MaCow [67] trained on CelebA HQ ($t = 0.7$) (g) Glow [62] trained on CelebA HQ ($t = 0.7$)






Long-Range Correlations

As we can see, the 20 global long-range correlations are captured mostly at the top of the hierarchy, and the local variations are recorded at the lower groups.



Questions

1. **Why does NVAE need residual cells, what do they look like, and why is each element needed?**
2. **Depthwise convolution: what is it for, what problems arise, how are they solved?**
3. **What is the Residual Normal Distributions approach?**
4. **What is the Spectral Regularisation approach?**

References

https://arxiv.org/abs/2007.03898 - A. Vahdat, J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. 2020