Interpretation of Activation Maps in Generative Modelling

22 July 2021

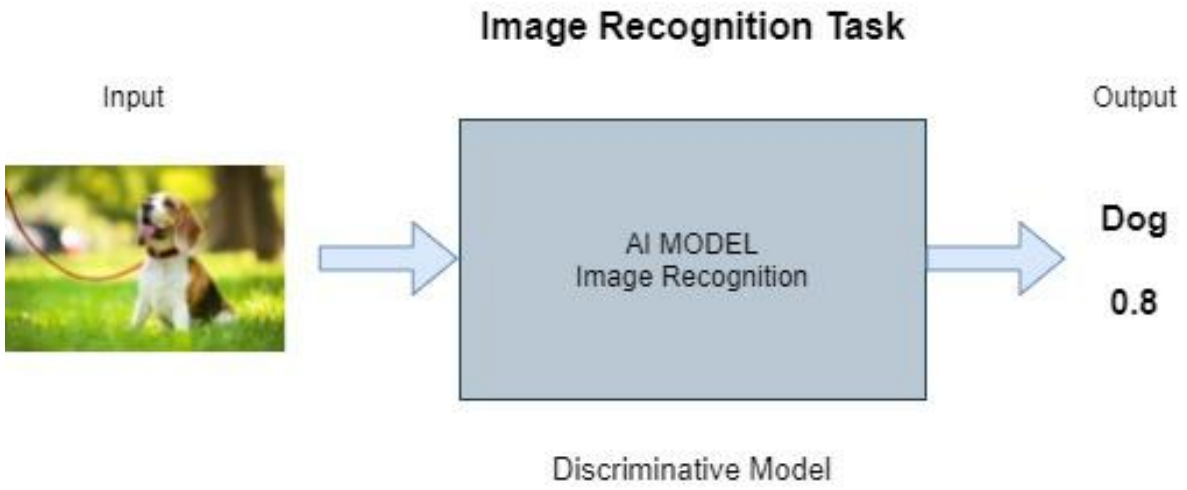
Srijay Kolvekar
MSc Electrical Engineering
Universität Stuttgart

Content

- Motivation
- Objective
- Method
- Outcome

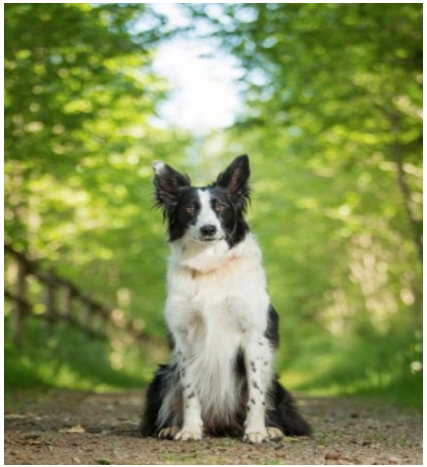
Liu, Wenqian, et al. "Towards visually explaining variational autoencoders." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020

Motivation - Discriminative Model

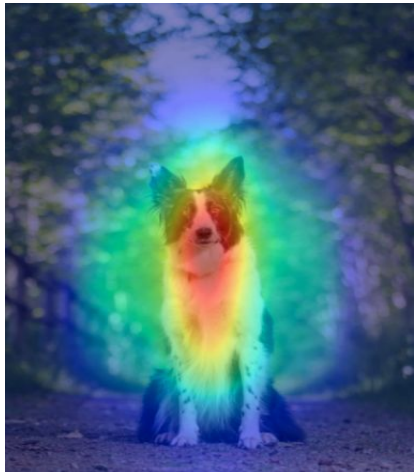


Motivation - Explainable AI

Original Image

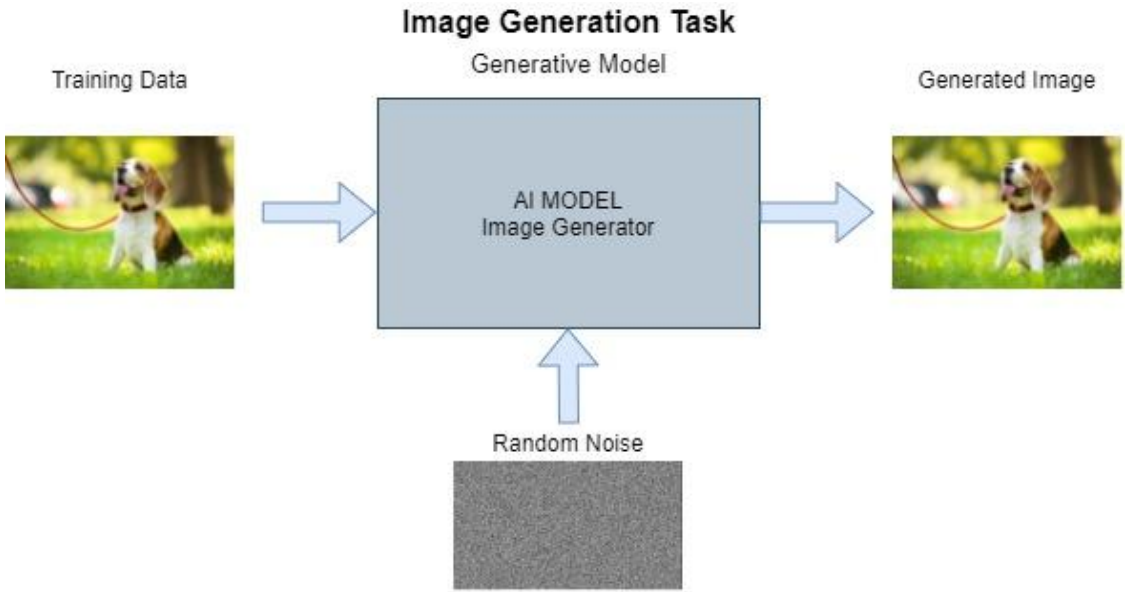


Grad CAM

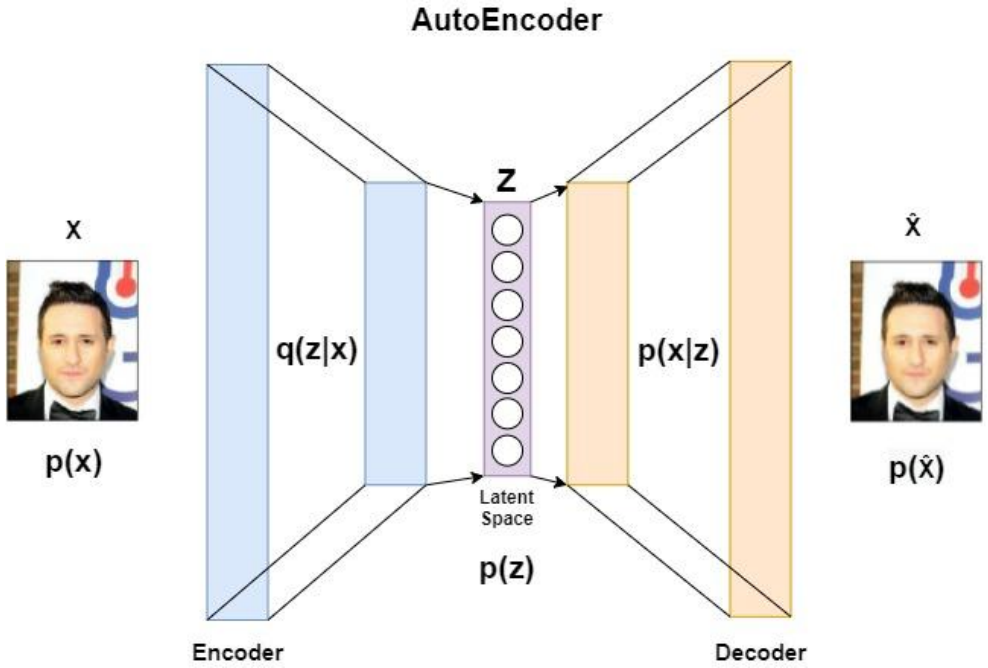


Ref: “Class Activation Explorer”, link

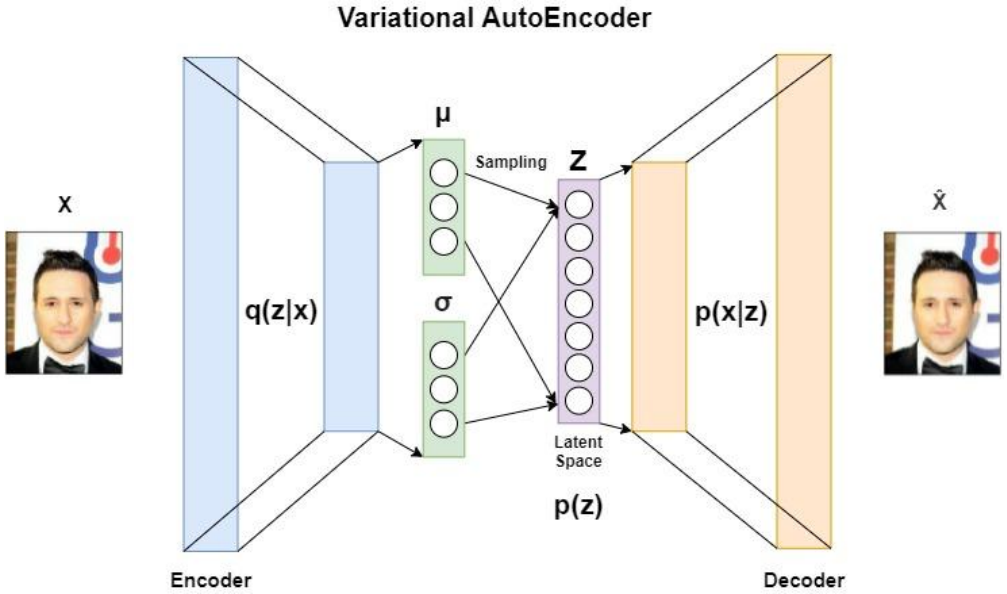
Objective - Generative Model



Method - AutoEncoder (AE)



Method - Variational AutoEncoder (VAE)



$$L = L_r(\mathbf{x}, \hat{\mathbf{x}}) + L_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}), p(\mathbf{z}))$$

Method - VAE Architecture

Model: "Encoder"

Layer (type)	Output Shape	Param #	Connected to
encoder_input (InputLayer)	[(None, 128, 128, 3)	0	
encoder_conv_0 (Conv2D)	(None, 64, 64, 32)	896	encoder_input[0][0]
leaky_re_lu (LeakyReLU)	(None, 64, 64, 32)	0	encoder_conv_0[0][0]
encoder_conv_1 (Conv2D)	(None, 32, 32, 64)	18496	leaky_re_lu[0][0]
leaky_re_lu_1 (LeakyReLU)	(None, 32, 32, 64)	0	encoder_conv_1[0][0]
encoder_conv_2 (Conv2D)	(None, 16, 16, 64)	36928	leaky_re_lu_1[0][0]
leaky_re_lu_2 (LeakyReLU)	(None, 16, 16, 64)	0	encoder_conv_2[0][0]
encoder_conv_3 (Conv2D)	(None, 8, 8, 64)	36928	leaky_re_lu_2[0][0]
leaky_re_lu_3 (LeakyReLU)	(None, 8, 8, 64)	0	encoder_conv_3[0][0]
flatten (Flatten)	(None, 4096)	0	leaky_re_lu_3[0][0]
mu (Dense)	(None, 200)	819400	flatten[0][0]
log_var (Dense)	(None, 200)	819400	flatten[0][0]
encoder_output (Lambda)	(None, 200)	0	mu[0][0] log_var[0][0]

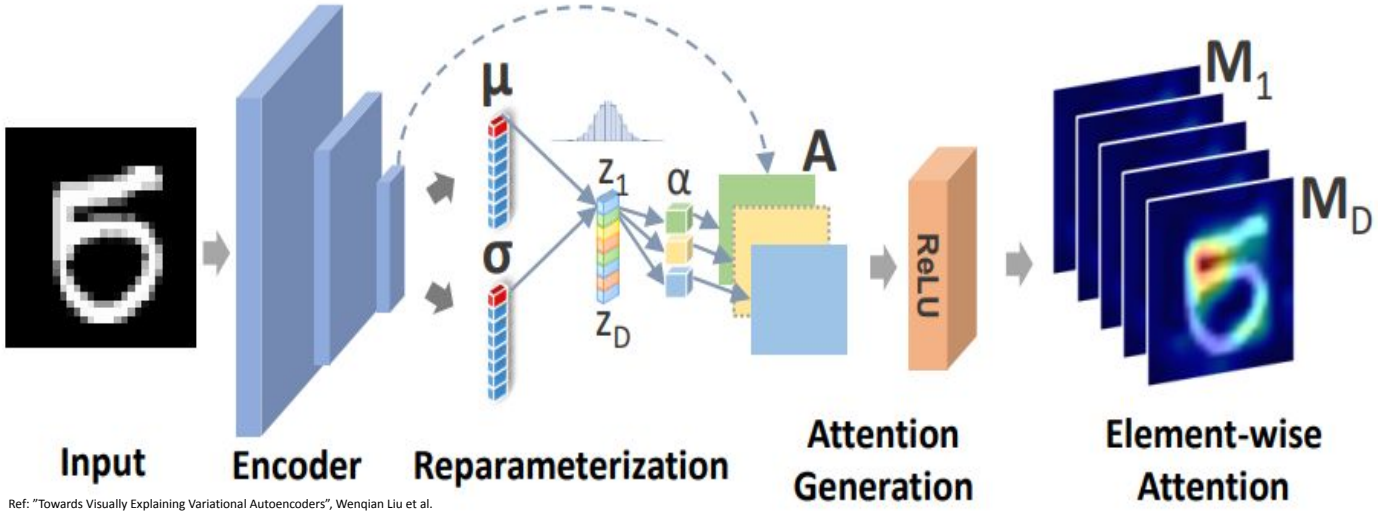
Total params: 1,732,048
Trainable params: 1,732,048
Non-trainable params: 0

Model: "Decoder"

Layer (type)	Output Shape	Param #
decoder_input (InputLayer)	[(None, 200)]	0
dense (Dense)	(None, 4096)	823296
reshape (Reshape)	(None, 8, 8, 64)	0
decoder_conv_0 (Conv2DTransp	(None, 16, 16, 64)	36928
leaky_re_lu_4 (LeakyReLU)	(None, 16, 16, 64)	0
decoder_conv_1 (Conv2DTransp	(None, 32, 32, 64)	36928
leaky_re_lu_5 (LeakyReLU)	(None, 32, 32, 64)	0
decoder_conv_2 (Conv2DTransp	(None, 64, 64, 32)	18464
leaky_re_lu_6 (LeakyReLU)	(None, 64, 64, 32)	0
decoder_conv_3 (Conv2DTransp	(None, 128, 128, 3)	867
activation (Activation)	(None, 128, 128, 3)	0

Total params: 916,483
Trainable params: 916,483
Non-trainable params: 0

Method - VAE + Gradient Class activation Maps

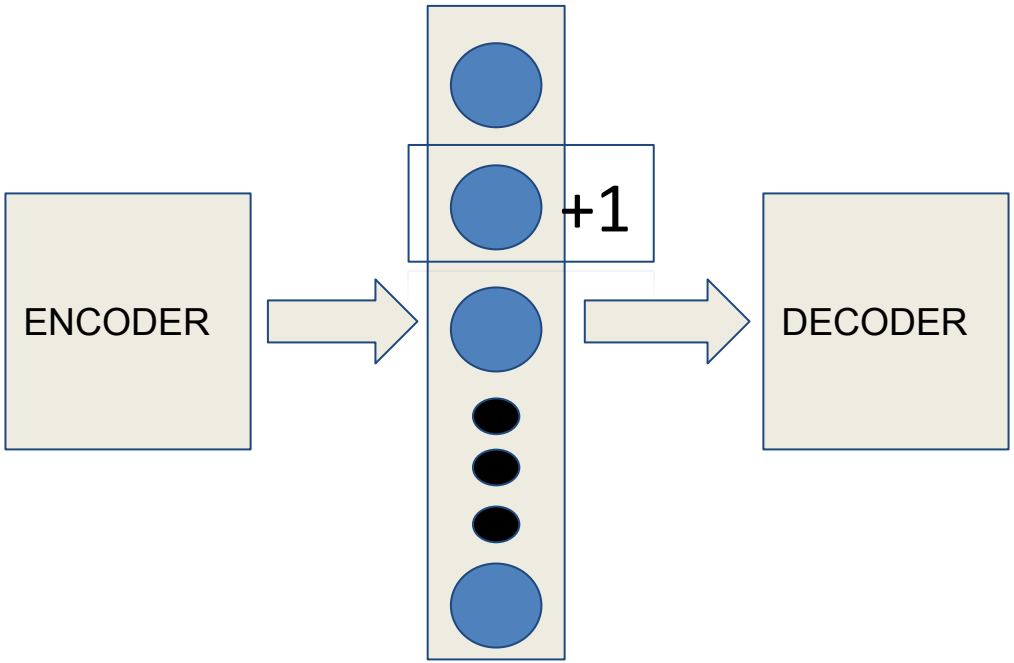


Ref: "Towards Visually Explaining Variational Autoencoders", Wenqian Liu et al.

$$\mathbf{M}^i = \mathrm{ReLU}(\sum_{k=1}^{n} \alpha_k \mathbf{A}_k)$$

$$\alpha_k = \frac{1}{T} \sum_{p=1}^{h} \sum_{q=1}^{w} (\frac{\partial z_i}{\partial A_k^{pq}})$$

Outcome - Visualizing Latent Space



Outcome - Heatmaps

Tweaking the latent space

Latent node: 22

Generated Image **Heatmap**

Input

Latent node: 190

Generated Image **Heatmap**

Code

https://github.com/srijayjk/Computer-Vision/blob/main/Ziess(VAE%2BGradCAM).ipynb