Learning Disentangled Representations in Face Images for Face Attributes Manipulations

Rajan Gyawali

(rgkg2)

Student ID: 14418115

May 10, 2023

MOTIVATION

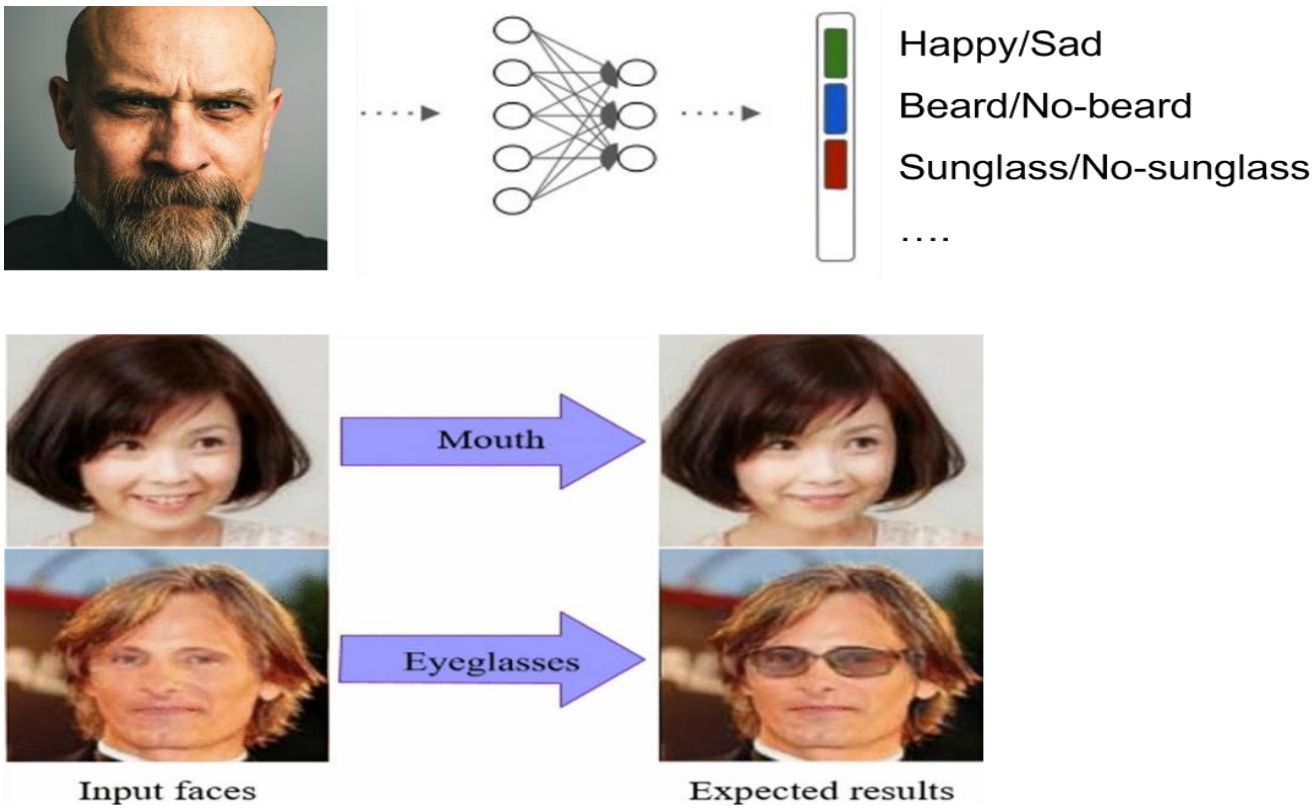

Figure 1: (Top) Representation of face with different attributes
(Bottom): Manipulation of face attributes by altering disentangled representations

Traditional Machine Learning:

- Features are highly correlated and difficult to separate.
- Difficulty in interpretation.

Applications:

- In computer vision, helps identifying attributes of the images for better generalization.
- In robotics, helps understand the different factors in the scene and aids more controlled actions.

Disentangled Representations:

- Learning each attribute in a separate representation.
- Enables machine to understand the data in a more interpretable and human like way.

Benefits:

- Interpretability
- Control
- Generalization
- Robustness

TECHNICAL DESCRIPTION

Dataset: Labelled Faces in the Wild (LFW)

- Database of face photographs
- 13,233 images of 5,749 people
- 73 attributes
- 80% training, 20% validation

Figure 2: Sample Images

Algorithm: Variational Autoencoder (VAE) and Beta-VAE

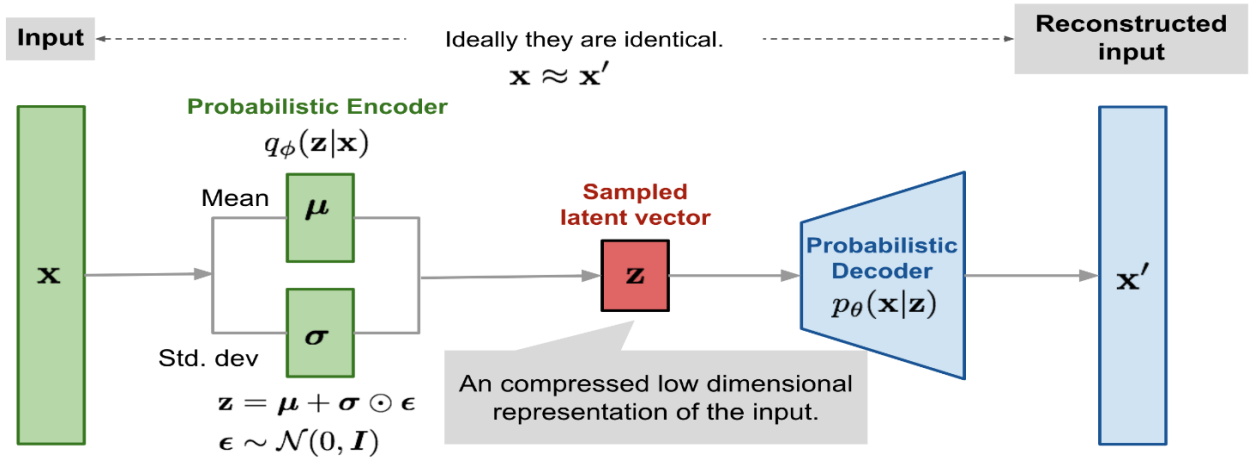

Figure 3: Variational Autoencoder

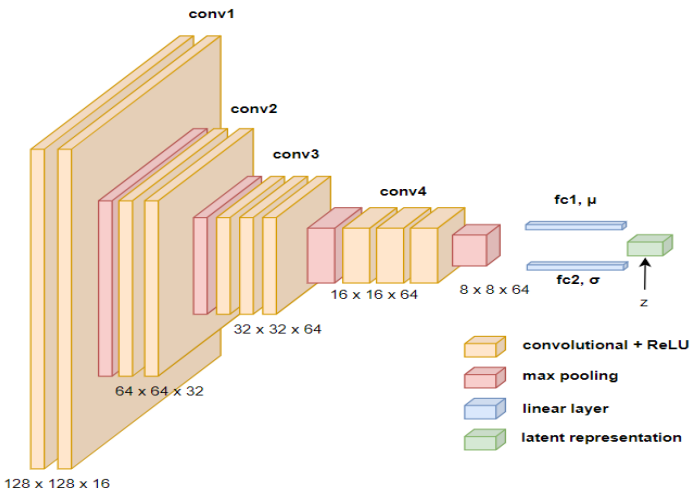

Figure 4: Architecture of Encoder

Different than Autoencoder?

- Probabilistic model
- Training is regularized to avoid overfitting.
- Latent space has good properties that enable generative process.
- Encoded as a distribution over the latent space rather than a single point.

TECHNICAL DESCRIPTION

VAE LOSS

- It consists of reconstruction loss and KL divergence term.

$$\text{Total Loss } (\mathcal{L}) = \mathcal{L}_1 + \mathcal{L}_2$$

- $\mathcal{L}_1$ is the reconstruction loss and is calculated as a binary cross entropy loss.

$$\mathcal{L}_1\,(x, x') = \sum (x * \log\sigma(x') + (1 - x) * \log\sigma(1 - x'))$$

- $\mathcal{L}_2$ is the KL divergence loss and is represented as:

$$\mathcal{L}_2 = D_{KL}\,[N(\mu, \sigma) \,||\, N(0, 1)] = -0.5 * \sum (1 + \log(\sigma^2) - \mu^2 - \sigma^2)$$

$$\text{where } D_{KL}\,[p(x) \,||\, q(x)] = \sum_{x \in X} p(x) log\left(\frac{p(x)}{q(x)}\right)$$

For Beta-VAE, the loss is represented as,

$$\mathcal{L} = \mathcal{L}_1 + \beta * \mathcal{L}_2$$

where β is a hyperparameter that controls the degree of disentanglement in the learned latent representations of the data.

- Disentangled representation separate the facial attributes such as shape of nose, eyes position into different dimensions of the latent space.
- The model learns to generate new faces combining these features.
- Beta controls the trade-off between disentanglement and reconstruction accuracy.
- Higher Beta results more disentangled representations.

Manipulating Face Attributes

- Z_s = vector representing average latent representations of the smiling people.
- Z_n = vector representing average latent representations of the non-smiling people.
- The difference vector is $\Delta Z = Z_s - Z_n$
- Thus, a non-smiling face can be changed to smiling using the relation,

$$Z_{new} = Z_{original} + \alpha * \Delta Z$$

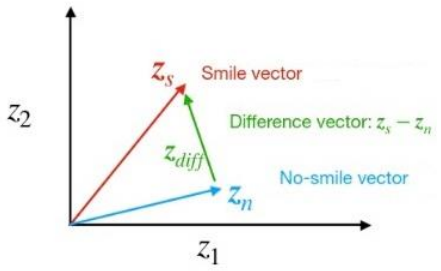

Figure 5: Latent Vectors

RESULTS

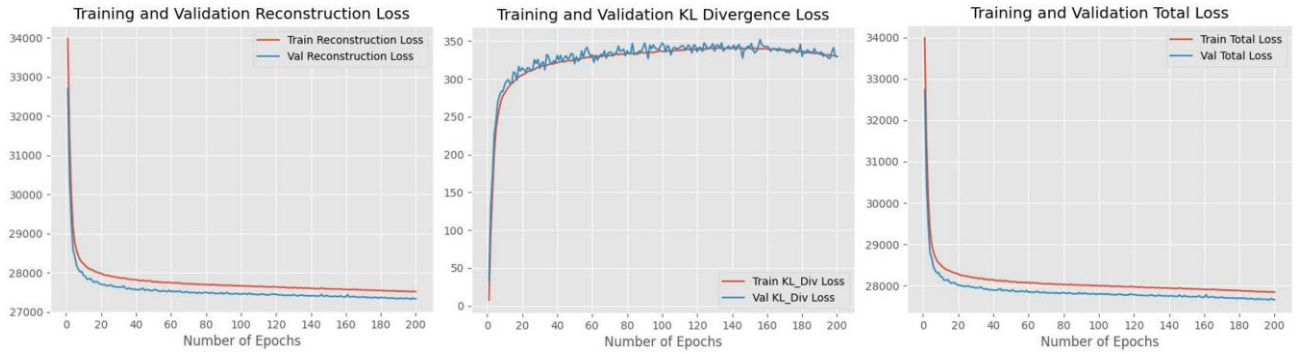

Figure 6: Reconstruction loss, KL Divergence and Total Loss for VAE

- **Normal VAE with no trade off between reconstruction loss and KL divergence loss**

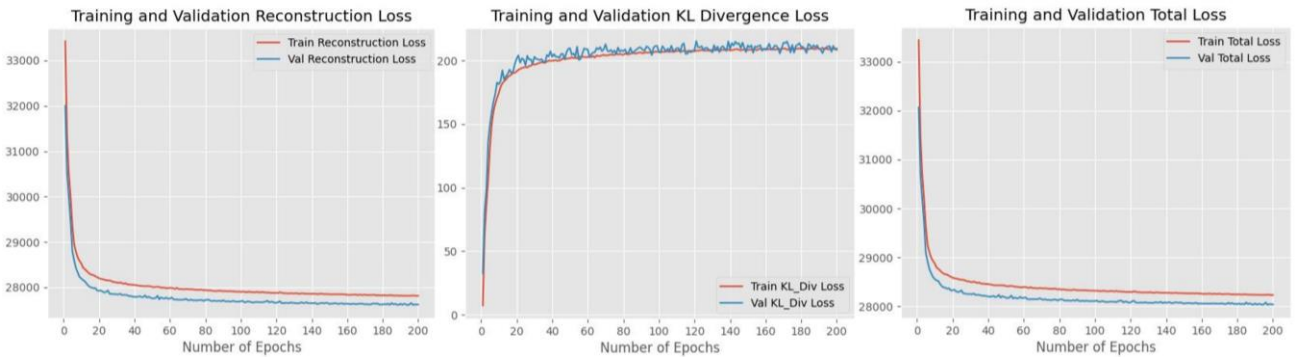

Figure 7: Reconstruction loss, KL Divergence and Total Loss for β-VAE (β = 2)

- **Beta VAE (Beta=2) with more emphasis on disentanglement**

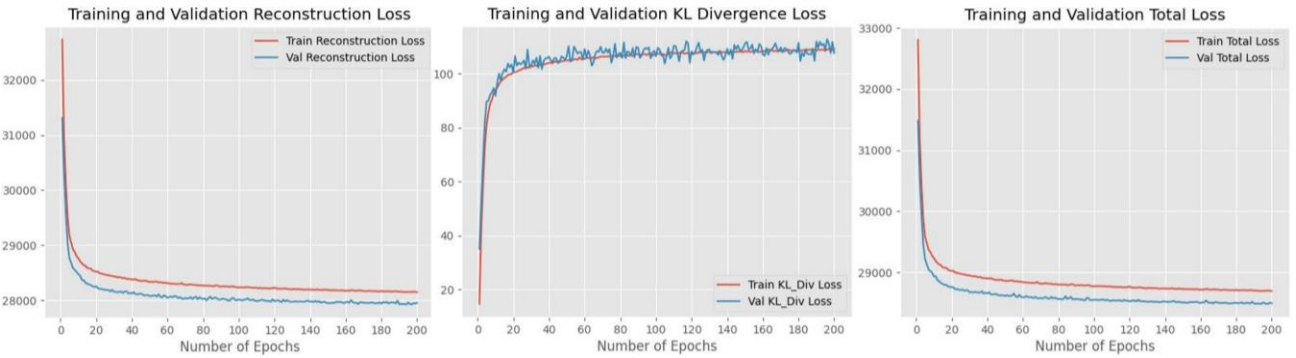

Figure 8: Reconstruction loss, KL Divergence and Total Loss for β-VAE (β = 5)

- **Increasing the value of Beta to 5, more emphasis on disentanglement increasing the reconstruction loss**

RESULTS

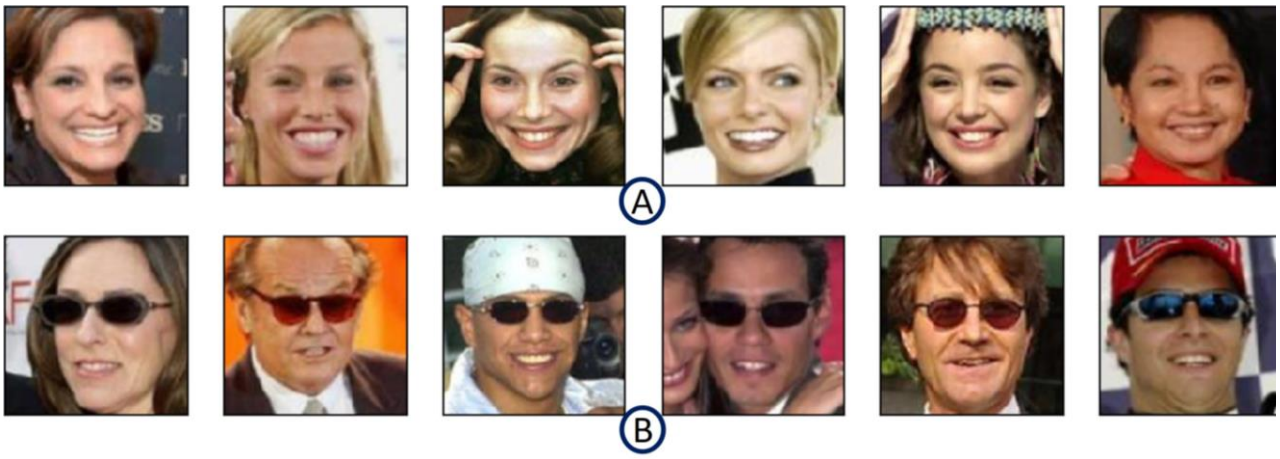

Figure 9: (A): Smiling faces, (B): Faces with sunglass

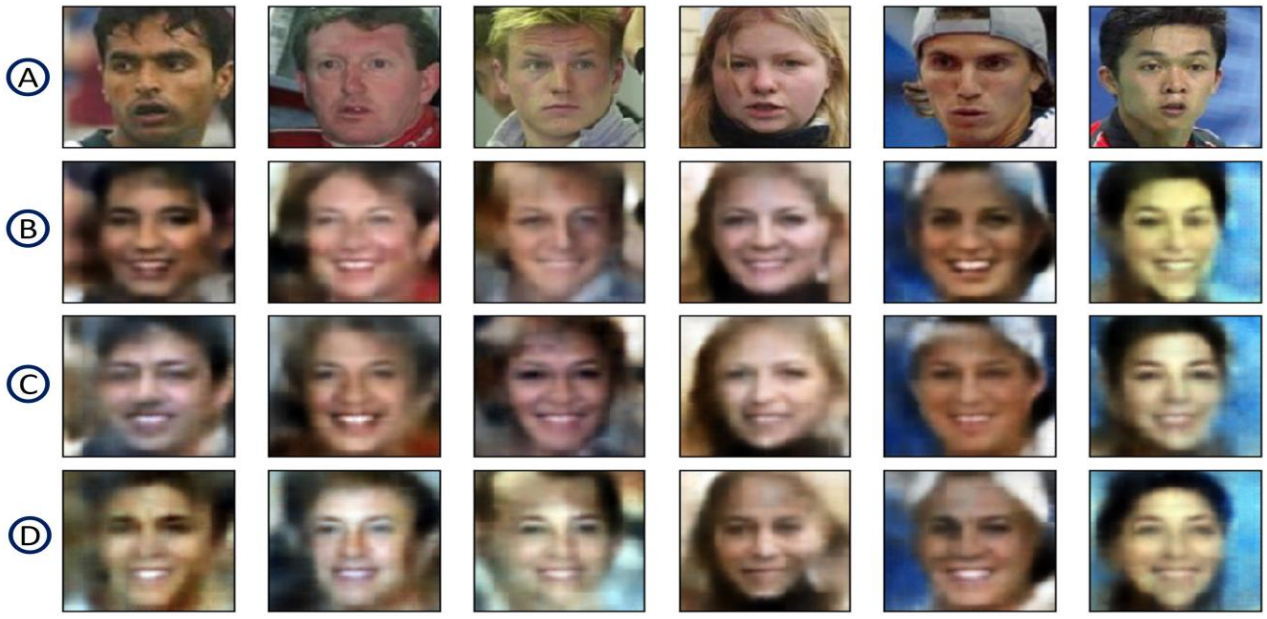

Figure 10: (A): Non-Smiling faces, (B): VAE results, (C): Beta-VAE with β=2, (D) β=5

Figure 11: Results for sunglasses

- In VAE, there is a moderate disentanglement and less reconstruction loss compared to β-VAE
- Increasing the β has caused more disentangled representations
- Setting β to 5 resulted in a loss of reconstruction quality

SUMMARY and FUTURE WORK

- Extension of Vanilla autoencoder and incorporates probabilistic modeling
- Key idea is to learn a latent representation
- A combination of reconstruction loss and Kullback-Leibler (KL) divergence term to optimize the model
- KL divergence encourages the learned latent space to match the desired distribution
- The β parameter controls the weight assigned to the KL divergence term in the VAE's objective function
- Varying the value of β, helps model learning disentangled and interpretable representations in the latent space
- Higher value of β encourages more disentangled representations
- Much higher value of β leads to poor reconstruction

β-VAEs provide a way to balance the trade-off between reconstruction accuracy and disentanglement.

Future Works

1. Conditional VAE
   Have more control over each attribute of the face by applying conditional information.
2. Factor VAE
   - Encourages each dimension of the latent space to capture a single independent factor of variation.
   - Employs a separate discriminator network.
3. Cascade VAE
   - Hierarchical VAE that consists of multiple levels of latent variables.
   - Each level captures increasingly abstract and high-level features.

Overall VAEs have greater applications in generative AI and can be extended to different applications.