Disentanglement in variational autoencoders

Tutor: Alasdair Newson (alasdair.newson@telecom-paris.fr)

13 February 2024

1 Introduction

Variational autoencoders are generative models which allow for synthesis of random examples of complex data. However, given a data point, we would like to be able to modify it in the latent space, for the purpose of image editing. This requires that each code correspond to a single attribute of the image. In other words, that the latent space is *disentangled*.

The aim of this project is to study the supposed disentanglement properties of different versions of the variational autoencoder [1]. Much of current research on generative models attempts to encourage disentanglement in the latent space. Two prominent approaches include:

- BetaVAE [2]
- FactorVAE [3]

The first approach adds a parameter to the standard variational autoencoder formulation to encourage independence in the latent space. The second does this by encouraging a factorised distribution in the latent space.

There is currently no unanimously accepted definition of disentanglement. Both papers propose their definition of disentanglement, which use synthetic data. This project will also aim to study these two definitions and confirm or not the conclusions of the papers. More generally, the project will study disentanglement in the case of complex face data (Celeb-A dataset [4]), using the original VAE versus BetaVAE/FactorVAE. For these data, pre-trained networks should be used (training may be difficult).

2 Tasks of the project

- Study and understand the two approaches to disentanglement BetaVAE and FactorVAE
- Understand and implement the two metrics proposed by these papers
  - Implementation on shapes data
- Study the disentanglement of both methods (in comparison to VAE if possible) on complex face data. This will consist in training or using a pre-trained classifier to analyse the annotated properties of this data (smile etc) and comparing navigation in their respective latent spaces (using a technique such as InterfaceGAN [5])

References

[1] D. P. Kingma and M Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[3] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[5] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *arXiv preprint arXiv:2005.09635*, 2020.