# Proposal - Learning Disentagled Representations with Reference-based Variational Autoencoders

Manjulata Chivukula

March 7, 2019

## 1 Introduction

The goal of my project is to understand the role of variational autoencoder in learning disentangled representations in data specifically in image data. Further as a part of the project I would like to study the utilization of disentangled representations for improving accuracy of image classification tasks (supervised as well as semi-supervised task), feature learning, conditional image generation, and attribute transfer.

## 2 Probabilistic Model

The probabilistic model that will be used in the project is the **variational autoencoder**. Specifically, the project will focus on utilizing the reference-based variational autoencoder model described in the paper [1]. Reference-based variational autoencoder described in the paper uses generative process for learning specific set of generative factors called target factors (e) (like facial expression) and common factors (z) (like pose, illumination, identity etc.) from the input image distribution (x) in an adversarial unsupervised learning setting.

**Evaluation of results:** Similar to the experiments performed in the paper [1] the project will be evaluated using the MNIST [2] data (that contains two sets of images i.e., reference set and unlabeled set), qualitative and quantitative evaluation of the algorithm will be performed. In addition, the project will evaluate the performance of the algorithm by computing the accuracy of the downstream linear classifier that utilizes the learned feature vectors.

## 3 Implementation Details

Following are some objectives that will be used to realize the goal of my project:

- Implement variational autoencoder described in the paper [3];
- Study and understand the implementation for beta-VAE model described in the paper [4];
- Implement VAE-GAN model, to understand the use of GAN network with VAE;

- Study and implement the algorithm described in the paper: [1], to disentangle latent distributions; the code for the algorithm has not been released by the authors, hence will need to be implemented from scratch;
- Extend the algorithm to perform classification on MNIST digit recognition data as a supervised learning task;
- Further extend the algorithm for classification of image data as a semi-supervised learning task (with some labelled and some unlabelled);

The dataset that will be used for the task is the MNIST digit recognition dataset [2]. The dataset is available to download for free. As described in the paper synthetic dataset will be generated on half of the images using transformations and other half will be unmodified. This will be two sets of images reference and unlabelled dataset that will be used to train the model. Implementation of deep learning models will run on Google Colaboratory notebook that use the free Tesla K80 GPU. The code will be implemented using Pytorch deep learning framework.

## References

[1] Adria Ruiz, Oriol Martínez, Xavier Binefa, and Jakob Verbeek. Learning disentangled representations with reference-based variational autoencoders. *CoRR*, abs/1901.08534, 2019.

[2] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[4] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $\beta$-vae. *CoRR*, abs/1804.03599, 2018.