# Deep Metric Learning with Triplet Loss and Variational Autoencoder

Haque Ishfaq, Ruishan Liu

## Abstract

We propose a novel structure to learn embedding in variational autoencoder (VAE) by incorporating deep metric learning. The features are learned by a triplet loss on the mean vectors of VAE. We show that VAE has a good performance and a high metric accuracy is achieved at the same time.

## Introduction

Recently deep metric learning has emerged as a superior method for representation learning. For extreme classification problem, where the number of categories is enormous, traditional classification methods are essentially useless. Triplet network learns feature embedding by optimizing the relative distance between the samples from the same classes and dissimilar classes.

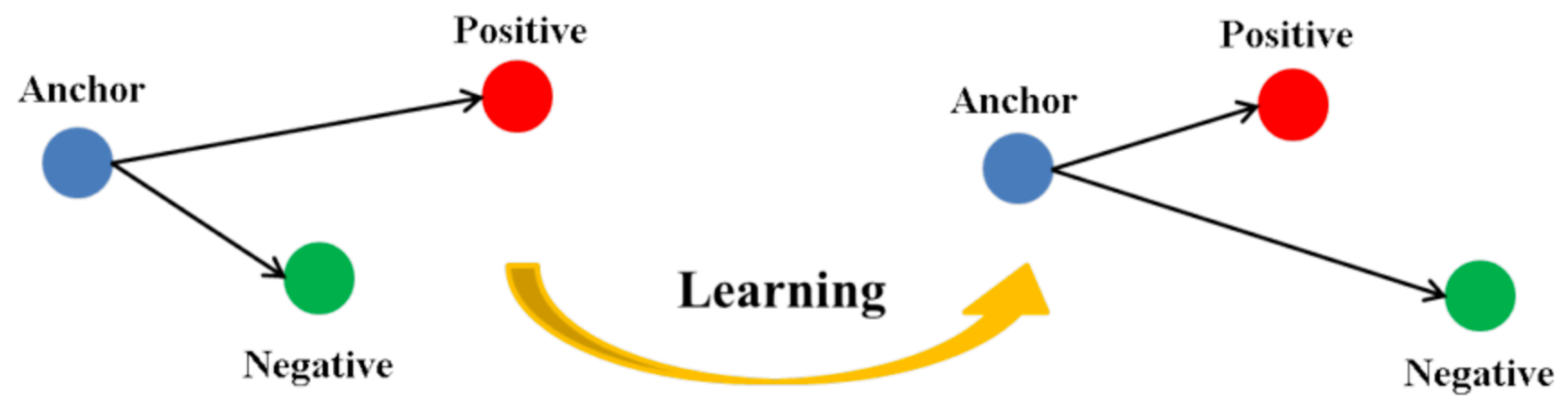


Figure.1 Triplet example before and after learning

## Triplet Loss

The triplet loss is trained on a series of triplets $\{x_a, x_p, x_n\}$ where $x_a$ and $x_p$ come from the same class and $x_n$ comes from a different class. The goal of the triplet loss is to keep $x_a$ closer to $x_p$ than $x_n$. The triplet loss is formulated as:

$$L_{trp} = \sum_{a,p,n}^{N} \left[ \|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + \alpha_{margin} \right]_+$$

## Our Model

The loss of our model consists of two parts: the triplet loss and the VAE losses for each of the triplet.

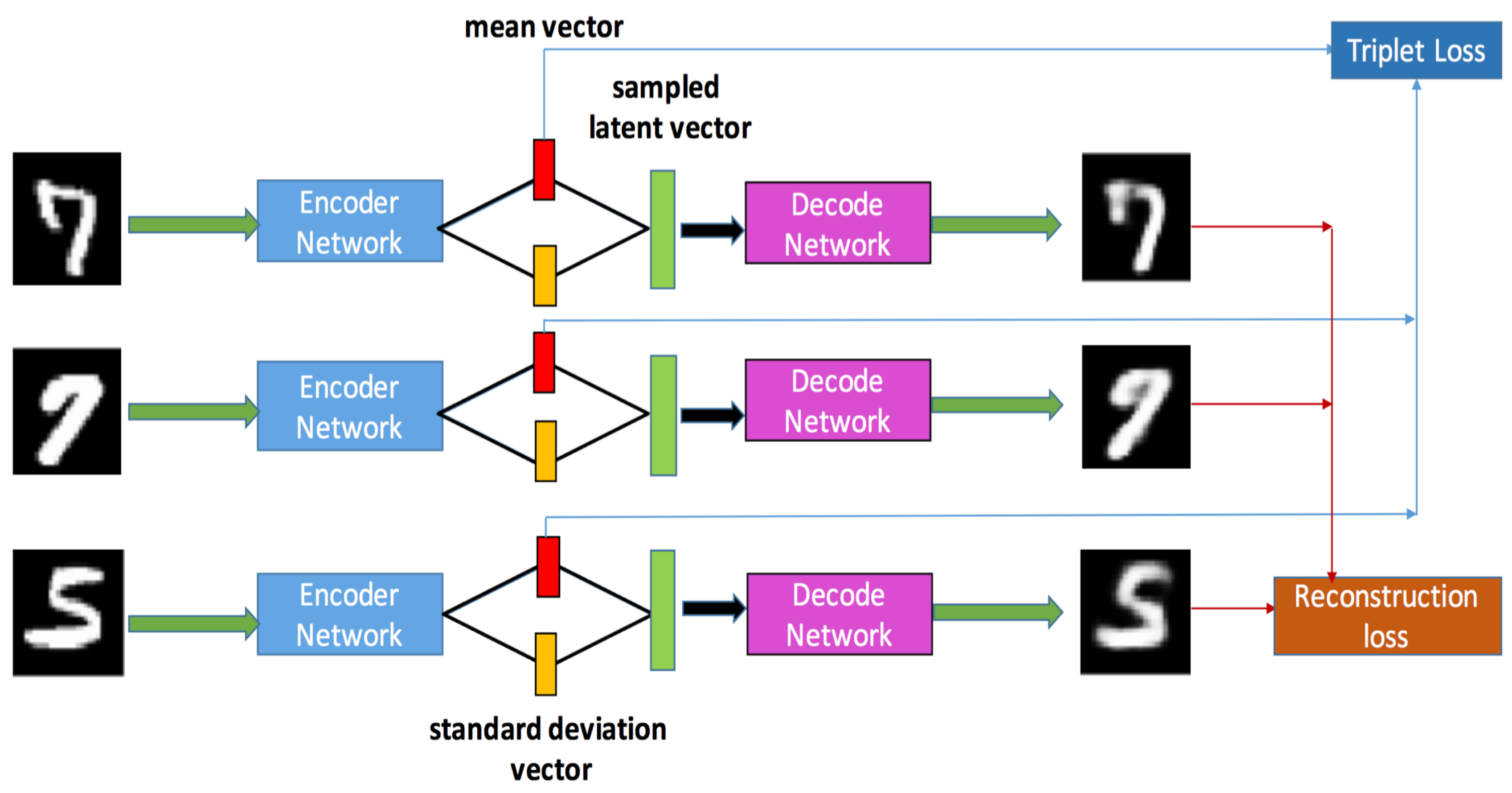


Figure 2: Model Architecture

For illustration, we adopted a simple network structure with two fully connected layers as encoder/ decoder. The dimension of the mean vector and the are both 20.

## Dataset

For our initial experimentation, we use the original MNIST dataset consisting of 60,000 28x28 gray-scale images of handwritten digits 0-9, and a corresponding set of 10000 test images.

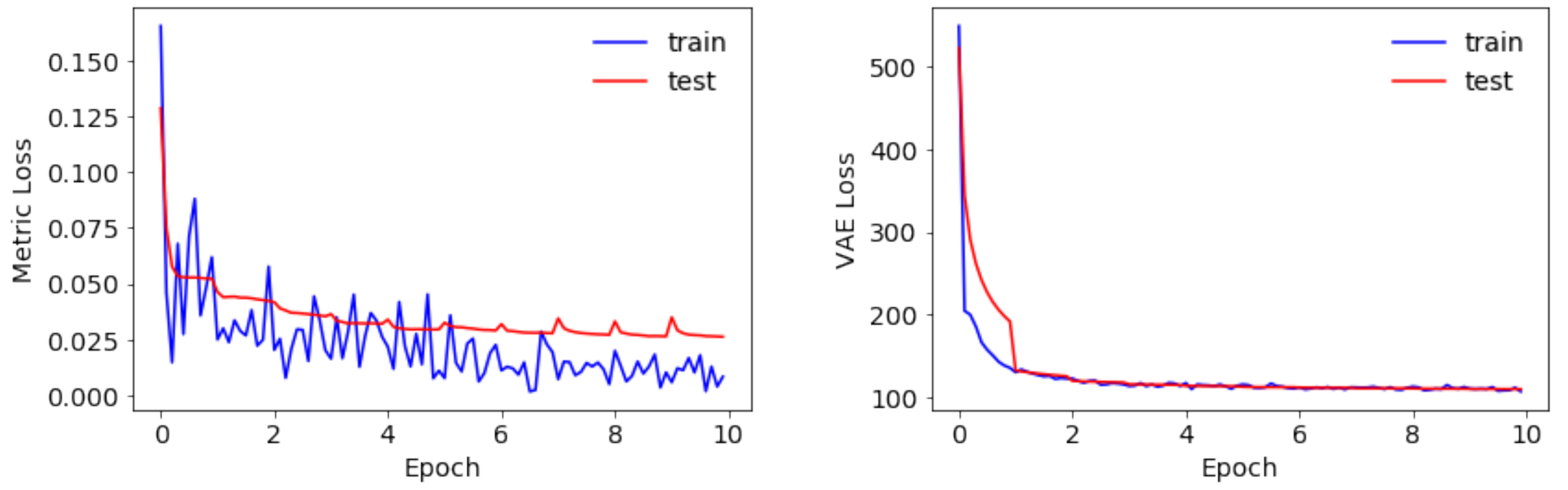
## Conclusion and Future Work

(1) We designed a new architecture motivated from metric learning and VAE and show that it learns high quality embedding on MNIST dataset.
(2) For future work, to further improve our model, we will incorporate conditional similarity network and perform experiments using larger datasets such as CUB200-2011, CARS196.

## Result

We trained our model with shallow network structure for 10 epochs. Both high metric accuracy and good VAE performance are achieved.

Figure 3 plots the learning curve for (a) triplet loss and (b) VAE loss. It is found that both the triplet and VAE suffer a synchronized fast decrease at the start of the training.

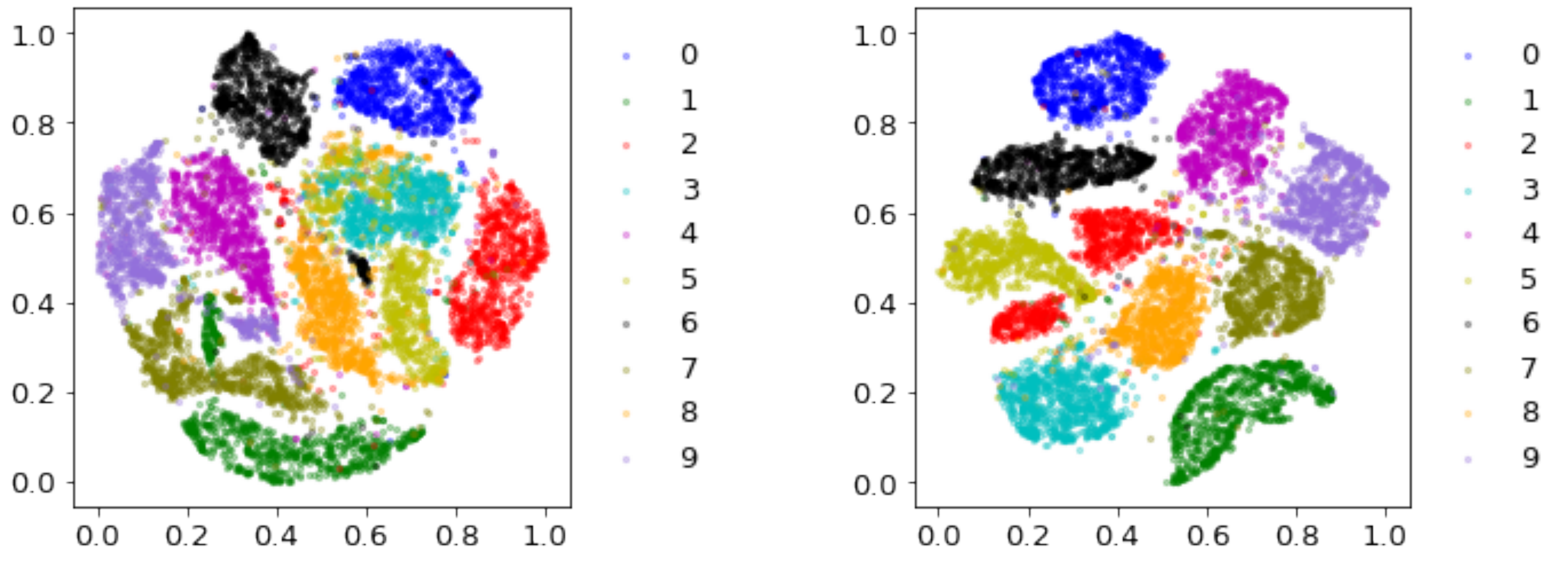


(a) Triplet Loss

(b) Variational autoencoder loss

Figure 3. Learning curve

We carried out t-SNE embedding for the mean vector, which is projected into two dimensional . For traditional vanilla VAE, the metric accuracy is 74.52% and the embeddings from different classes are overlapped at many spots, as shown in Fig. 4(a). In contrast, using our method, the metric accuracy achieves 95.95% and the features are nicely clustered, indicated by Fig. 4(b).



(a) Vanilla VAE
Accuracy = 74.52%

(b) Our structure
Accuracy = 95.95%

Figure 4. T-SNE projection for the mean vector

**Contact**

Haque Ishfaq
MS @ Dept. of Statistics
Stanford University
Email: hmishfaq@stanford.edu

Ruishan Liu
PhD @ Dept. Electrical Engineering
Stanford University
Email: ruishanj@stanford.edu