# Adversarial Autoencoders 对抗自编码

# Abstract

In this paper, we propose the “adversarial autoencoder” (AAE), which is a probabilistic autoencoder that uses the recently proposed generative adversarial networks (GAN) to perform variational inference by matching the aggregated posterior of the hidden code vector of the autoencoder with an arbitrary prior distribution. Matching the aggregated posterior to the prior ensures that generating from any part of prior space results in meaningful samples.

As a result, the decoder of the adversarial autoencoder learns a deep generative model that maps the imposed prior to the data distribution.

We show how the adversarial autoencoder can be used in applications such as semi-supervised classification, disentangling style and content of images, unsupervised clustering, dimensionality reduction and data visualization.

We performed experiments on MNIST, Street View House Numbers and Toronto Face datasets and show that adversarial autoencoders achieve competitive results in generative modeling and semi-supervised classification tasks.

在本文中，我们提出了“对抗性自动编码器（adversarial encoder）”（AAE），它是一种概率自动编码器并使用GAN来实现变分推断（variational interference），通过匹配自编码器隐藏代码矢量的aggregated posterior和任意的先验分布。这样可以产生有意义的样本。

AAE的解码器学习一个深层的生成模型，它可以将先验映射到数据分布。

我们展示了如何应用AAE，比如半监督分类(semi-supervised classification),disentangling style和图像内容，半监督聚类，维度降低和数据可视化。

我们应用MNIST，Street View House Numbers和Toronto Face datasets做实验，展示AAE有应用于生成模型和半监督分类任务上具有竞争力的结果。

# 1.Introduction

Building scalable generative models to capture rich distributions such as audio, images or video is

one of the central challenges of machine learning. Until recently, deep generative models, such as Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) were trained primarily by MCMC-based algorithms [Hinton et al., 2006, Salakhutdinov and Hinton, 2009]. In these approaches the MCMC methods compute the gradient of log-likelihood which becomes more imprecise as training progresses. This is because samples from the Markov Chains are unable to mix between modes fast enough. In recent years, generative models have been developed that may be trained via direct back-propagation and avoid the difficulties that come with MCMC training. For example, variational autoencoders (VAE) [Kingma and Welling, 2014, Rezende et al., 2014] or importance weighted autoencoders [Burda et al., 2015] use a recognition network to predict the posterior distribution over the latent variables, generative adversarial networks (GAN) [Goodfellow et al., 2014] use an adversarial training procedure to directly shape the output distribution of the network via back-propagation and generative moment matching networks (GMMN) [Li et al., 2015] use a moment matching cost function to learn the data distribution.

构建可扩展的生成模型以捕获丰富的分布，如音频，图像或视频机器学习的核心挑战之一。

直到最近，深度生成模型如RBM，DBNs和DBMs主要通过基于MCMC算法训练。

在这些方法中，MCMC方法计算似然函数的梯度，似然函数在训练过程中会变得更加不精确。这是因为马尔科夫链的样本无法在模式之间快速混合。

近年来，生成模型可以直接通过back-propagation训练并且避免MCMC训练带来的困难。例如VAE或重要性加权自编码器（importance weighted autoencoders）使用一个识别网络去预测潜在变量的后验分布，GAN使用对抗性训练过程并通过back-propagation和GMMN（使用moment matching cost function）来直接生成网络的输出分布。

In this paper, we propose a general approach, called an adversarial autoencoder (AAE) that can turn an autoencoder into a generative model. In our model, an autoencoder is trained with dual objectives – a traditional reconstruction error criterion, and an adversarial training criterion [Goodfellow et al., 2014] that matches the aggregated posterior distribution of the latent representation of the autoencoder to an arbitrary prior distribution. We show that this training criterion has a strong connection to VAE training. The result of the training is that the encoder learns to convert the data distribution to the prior distribution, while the decoder learns a deep generative model that maps the imposed prior to the data distribution.

在本文中，我们提出了一种通用方法，称为AAE，可以将自动编码器变为生成模型。

在我们的模型中，自编码器采用双重目标进行训练——传统的重建误差标准和对抗性训练标准，他们使自编码器的潜在表示的合计后验分布与任意的先验分布匹配。

我们展示了这种训练标准与VAE训练有很强的联系。训练的结果是编码器学会将数据分布转换为先验分布，而解码器学习一个深度生成模型，该模型使先验分布映射到数据分布。

# 2.Adversarial Autoencoders

## 2.2 Relationship to GANs and GMMNs

In the original generative adversarial networks (GAN) paper [Goodfellow et al., 2014], GANs were used to impose the data distribution at the pixel level on the output layer of a neural network. Adversarial autoencoders, however, rely on the autoencoder training to capture the data distribution. In adversarial training procedure of our method, a much simpler distribution (e.g., Gaussian as opposed to the data distribution) is imposed in a much lower dimensional space (e.g., 20 as opposed to 1000) which results in a better test-likelihood as is discussed in Section 3.

最初的GAN用于将像素级上的数据分布强加在神经网络的输出层上。然而，AAE依靠自动编码器训练来捕获数据分布。在对抗性训练过程中，一个更简单的分布强加于一个低维空间上，这将带来更好地test-likelihhod(chap.3).

Generative moment matching networks (GMMN) [Li et al., 2015] use the maximum mean discrepancy (MMD) objective to shape the distribution of the output layer of a neural network. TheMMDobjective can be interpreted as minimizing the distance between all moments of the model distribution and the data distribution. It has been shown that GMMNs can be combined with pre-trained dropout autoencoders to achieve better likelihood results (GMMN+AE). Our adversarial autoencoder also relies on the autoencoder to capture the data distribution. However, the main difference of our work with GMMN+AE is that the adversarial training procedure of our method acts as a regularizer that shapes the code distribution while training the autoencoder from scratch; whereas, the GMMN+AE model first trains a standard dropout autoencoder and then fits a distribution in the code space of the pre-trained network. In Section 3, we will show that the test-likelihood achieved by the joint training scheme of adversarial autoencoders outperforms

GMMN使用MMD来塑造神经网络输出层的分布。MMD的目标可以解释为最小化距离(模型分布的所有moments和数据分布之间的距离)。GMMN可以与pre-trained dropout AE相结合（GMMN+AE）来达到更好的likelihood结果。我们的AAE也依靠自动编码器来捕获数据分布。然而，我们与GMMN + AE工作的主要区别在于对抗训练程序相当于正则化器，可以从头开始训练AE的同时塑造code分布；然而GMMN+AE首先训练一个标准的dropout AE,然后在预训练网络的代码空间拟合分布。

## 2.3 Incorporating Label Information in the Adversarial Regularization

合并标签消息（对抗正则化）

In the scenarios where data is labeled, we can incorporate the label information in the adversarial

training stage to better shape the distribution of the hidden code. In this section, we describe how to leverage partial or complete label information to regularize the latent representation of the autoencoder more heavily. To demonstrate this architecture we return to Figure 2b in which the adversarial autoencoder is fit to a mixture of 10 2-D Gaussians. We now aim to force each mode of the mixture of Gaussian distribution to represent a single label of MNIST.

在标记数据的场景中，我们可以将标签信息合并到对抗训练阶段，以更好塑造hidden code的分布。在本节中，我们将介绍如何更好地利用将部分或完整的标签信息是latent representation正则化。为了演示这种架构，我们返回图2b中,AAE已经拟合10个2D高斯分布的混合模型。 我们现在的目标是混合高斯分布模型的每个mode表示MNIST的单个标签。

# 6.Unsupervised Clustering with Adversarial Autoencoders