

西安交通大学

硕士学位论文

深度表示学习在时序、图像与视频数据上的应用初探

学位申请人：刘仕琪

指导教师：孟德宇 教授

学科名称：数学

2020 年 06 月

Primary Applications of Deep Representation Learning in Time Series, Image and Video Data

A thesis submitted to
Xi'an Jiaotong University
in partial fulfillment of the requirements
for the degree of
Master of Science

By
Shiqi Liu
Supervisor: Prof. Deyu Meng
Mathematics
June 2020

硕士学位论文答辩委员会

深度表示学习在时序、图像与视频数据上的应用初探

答辩人：刘仕琪

答辩委员会委员：

西安交通大学教授： 孟德宇 孟德宇

西安交通大学教授： 李继成 李继成

西安交通大学教授： 张春霞 张春霞

西安交通大学副教授： 惠永昌 惠永昌

西安交通大学副教授： 朱学虎 朱学虎

答辩时间：2020年5月29日

答辩地点：西安交通大学数学楼423会议室

摘要

大数据时代的到来带来了机器学习基本模式的颠覆性改变。传统人为抽取特征的方式已无法满足具有多样化、复杂化、海量化应用目标的现实需求。深度学习由于其端到端的数据驱动自适应抽取特征模式，逐渐成为应用的主流，引起了学术界与工业界的广泛关注。基于此研究现状，本文针对时间序列、图像、视频等典型数据类型的代表性深度学习方法进行介绍，并尝试在对相应深度学习方法的特性进行深入理解的基础上，将其改造应用于水文时间序列处理、图像因子学习、脑电数据情感预测等应用问题。论文包括以下研究内容。

首先，我们利用深度长短记忆网络（LSTM）进行水文站日径流数据预测。利用 LSTM 的长短时记忆特性，深度 LSTM 在宜昌水文站取得了好的成果。我们进一步比较了深度 LSTM 和回馈神经网络（BPNN）在汛期于宜昌水文站的预测结果，其涵盖了相对系统误差，相对标准差，和相对误差范围等指标，结果验证了深度 LSTM 比 BPNN 在汛期预测的结果要准确。为了验证所提方法的有效性，我们在寸滩，万县，奉节和宜昌水文站获取的数据上对比了 LSTM 模型和深度 LSTM 模型的结果。结果显示两个模型都适用于日径流预测，并且深度 LSTM 预测获得了更好的实验表现。

其次，针对深度表示学习的主要模型之一，变分自编码器（VAE），我们聚焦于监督 VAE 发掘的生成因子。VAE 被用于学习图像独立低维表示，但是却面临一些预先指定的因子被忽略的问题。我们认为输入和学得表示的每个因子的互信息，是一个用来发现有影响力的因子必要的指标。同时，我们深入研究了 VAE 的目标函数，说明了其倾向于诱导超过数据本质维度时因子互信息的稀疏性，从而产生一些没有影响力的生成因子。这些没有影响力的生成因子对数据重建能力起到近似能够忽略的作用。我们展示了互信息会影响 VAE 重建误差的下界和后续分类任务，并进而提出了一个算法来计算 VAE 的互信息指标且证明了其一致性。在 MNIST、CelebA 和 DEAP 数据集上的实验结果展现了互信息能帮助确定有影响力的生成因子。并且一些生成因子具有可解释性，可以对后续生成和分类任务产生有益的帮助。

最后，我们聚焦于深度表示学习在脑电视频数据情感识别上的应用。通过刻画 VAE 和 LSTM 的脑电情感深度表示学习架构，方法性能达到了与目前前沿方法可比的性能。同时模型还可以监测学得的特征形态。

综合上述三个部分，深度表示学习在时序，图像以及视频数据上获得良好的表现性能，且能够在一些已知的应用领域，例如水文预测，图像因子学习，脑电情感识别达到国际前沿。因此，本文对于这些应用领域的拓展，以及深度学习自身的方法论层面，均具有显著的研究与应用启发。

关 键 词: 深度学习；表示学习；长短记忆网络；日径流预测；变分自编码器；互信息；生成模型；脑电情感预测

论文类型: 应用研究

ABSTRACT

The advent of the era of big data has brought about a drastic change in the basic model of machine learning. The traditional method of artificial feature extraction has been unable to meet the actual needs of diversified, complicated, and quantified application goals. Deep learning has gradually become the mainstream of applications due to its end-to-end data-driven adaptive feature extraction, which has attracted widespread attention in academia and industry. Based on the current research status, this article introduces representative deep learning methods for typical data types such as time series, images, and videos. Based on a deep understanding of the characteristics of the corresponding deep learning methods, it is applied to application problems such as hydrological time series processing, image factor learning, and emotional prediction of EEG data. The paper includes the following research contents.

First of all, we use the deep long short-term memory network (LSTM) to predict the daily runoff data of hydrological stations. Using the long short-term memory characteristics of LSTM, the deep LSTM achieves good results in Yichang Hydrological Station. We further compare the relative systematic error, relative standard deviation and relative error range of prediction results of LSTM and backpropagation neural network(BPNN) in flood season of Yi Chang, and the comparison results verify the superiority of the deep LSTM method over BPNN in predicting daily runoff in flood season. In order to verify the effectiveness of the proposed method, we compared the results of LSTM models and deep LSTM models on the data obtained from Cuntan, Wanxian, Fengjie and Yichang hydrological stations. The results show that both models are suitable for daily runoff data, and deep LSTM models achieve better prediction results.

The second part is about one of the main models of deep representation learning, the variational autoencoder (VAE). We focus on supervising the factors extracted by VAE. VAE is used to learn the independent low-dimensional representation of images, but it faces the problem that some pre-specified factors are ignored. We assert that the mutual information of the input and each learned factor of the representation plays a necessary indicator to discover the influential factors. At the same time, we delve into the objective function of VAE, which shows that it tends to induce the sparsity of mutual information of factors when it exceeds the essential dimension of the data, resulting in some non-influential factors which can be ignored and which have negligible data reconstruction capabilities. We show that mutual information can affect the lower bound of VAE reconstruction errors and down-stream classification tasks. We propose an algorithm to calculate the mutual information indicator of VAE and prove its consistency. Experimental results on the MNIST, CelebA, and DEAP datasets show that mutual

information helps determine influential generating factors. Besides some generation factors are interpretable, which can help down-stream generation and classification tasks.

At last, we focus on deep representation learning applied to EEG video data emotion recognition. By characterizing the VAE and LSTM-based EEG emotion representation learning architecture, the method reaches the latest level. At the same time, the model can also monitor the learned feature shapes.

Summarizing the contents of the above three parts, the deep representation learning achieves good performance in time series, image, and video data and can reach the international frontier in some known application fields, such as hydrological prediction, image factor learning, and EEG emotion recognition. Therefore, this paper has significant research and application inspiration for the methodological expansion of these application fields and the methodological level of deep learning itself.

KEY WORDS: Deep learning; Representation learning; Long short-term memory network; Daily runoff prediction; Variational autoencoder; Mutual information; Generative model; EEG emotion recognition

TYPE OF THESIS: Application Research

目 录

摘要.....	I
ABSTRACT	III
第一章 深度表示学习绪论	1
1.1 深度表示学习的背景	1
1.1.1 深度表示学习的第一次复苏.....	1
1.1.2 深度表示学习的第二次复苏.....	2
1.2 文章架构	3
第二章 时序序列的深度表示学习.....	4
2.1 循环网络与长短记忆网络.....	4
2.2 深度长短记忆网络用于日径流量预测	6
2.2.1 日径流时间序列预测的意义.....	6
2.2.2 日径流时间序列预测的背景和相关工作	7
2.2.3 深度长短记忆网络用于日径流时间序列预测的合适性	8
2.2.4 深度长短记忆网络日径流模型构造.....	8
2.2.5 模型目标函数	10
2.2.6 应用实例及分析	10
2.2.7 结论	13
第三章 图像数据的深度表示学习.....	15
3.1 变分自编码器及生成模型的背景	15
3.2 变分自编码器	16
3.2.1 生成过程	16
3.2.2 分类过程	17
3.2.3 求解方法	17
3.3 变分自编码器用于发掘数据变化的功能.....	17
3.4 变分自编码器发掘有影响的因子的能力	21
3.4.1 互信息作为一个指标的必要性	22
3.4.2 相关工作	29
3.4.3 实验结果	29
3.5 因子的等价特性.....	33
3.6 总结	34

第四章 视频数据的深度表示学习.....	36
4.1 脑电情感识别的意义.....	36
4.2 脑电信号的简介.....	37
4.3 脑电情感识别的背景和相关研究	37
4.4 DEAP 数据集介绍	38
4.5 脑电数据情感识别模型	39
4.5.1 β -变分自编码部分.....	39
4.5.2 长短记忆网络部分	39
4.5.3 图模型部分	39
4.6 脑电数据情感识别实验结果	42
4.7 脑电数据情感识别结论	42
第五章 结论与展望.....	43
5.1 结论	43
5.2 展望	44
致 谢.....	46
参考文献	47
攻读学位期间取得的研究成果	52
声 明	

CONTENTS

ABSTRACT (Chinese)	I
ABSTRACT (English)	III
1 Deep Representation Learning Preface	1
1.1 Background of Deep Representation Learning.....	1
1.1.1 The First Renaissance of Deep Representation Learning	1
1.1.2 The Second Renaissance of Deep Representation Learning	2
1.2 Paper Structure.....	3
2 Deep Represent Learning for Time Series.....	4
2.1 RNN and LSTM	4
2.2 Using Deep LSTM Networks for Daily Runoff Prediction	6
2.2.1 The Significance of Daily Runoff Time Series Prediction	6
2.2.2 Background and Related Research on Daily Runoff Time Series Prediction..	7
2.2.3 Applicability of Deep LSTM Networks for Daily Runoff Prediction	8
2.2.4 Construction of Deep LSTM Networks for Daily Runoff Prediction.....	8
2.2.5 Objective	10
2.2.6 Examples and Analysis	10
2.2.7 Conclusion	13
3 Deep Representation Learning for Image	15
3.1 Background of Variational Autoencoders and Generative Models	15
3.2 Variational Autoencoders.....	16
3.2.1 Generation	16
3.2.2 Classification	17
3.2.3 The Solving Approach	17
3.3 The Function of Variational Autoencoder to Discover the Data Variation	17
3.4 The Ability of Variational Autoencoder to Discover Influential Factors	21
3.4.1 The Necessity of Mutual Information as An Indicator.....	22
3.4.2 Related Works	29
3.4.3 Experimental Results	29
3.5 Equivalence of Factors	33
3.6 Conclusion	34

4 Deep Represent Learning for Video.....	36
4.1 The Significance of EEG Emotion Recognition	36
4.2 Introduction to EEG Signals	37
4.3 Background and Related Researches on EEG Emotion Recognition.....	37
4.4 Introduction of DEAP Dataset.....	38
4.5 Emotion Recognition Models for EEG Data	39
4.5.1 β -VAE Part	39
4.5.2 LSTM Part	39
4.5.3 Graph Model Part	39
4.6 Results of Emotion Recognition of EEG Data	42
4.7 Conclusion of Emotional Recognition of EEG Data	42
5 Conclusion and Perspective	43
5.1 Conclusion	43
5.2 Perspective	44
Acknowledgements.....	46
References	47
Achievements	52
Declarations	

1 深度表示学习绪论

我们生活在大数据的时代，科学、技术、工程和人们的日常生活都在产生着数以 PB 级和 EB 级的数据。如果想要利用好这些数据来为特定目标服务，就需要很好的建模出它们之间的复杂的依赖关系。早期人们利用特征工程 (Feature Engineering) 的方法为数据抽取特征，然后投入后续任务之中的方法已经不再能够满足现在的需求，而以深度神经网络（多隐含层的神经网络模型）为代表的端对端 (end-to-end) 的、不需要人为手工提取特征的、在大数据上进行学习的方式正占据主要潮流。

1.1 深度表示学习的背景

我们将深度学习 (Deep Learning) 和表示学习 (Representation Learning) 等利用深度神经网络的学习数据表征用于后续任务的学习范式，统称作深度表示学习 (Deep Representation Learning)。而深度表示学习的历史就是指神经网络发展的历史。

深度神经网络起源于感知机。感知机是由 Frank Rosenblatt 教授在 1957 年在康奈尔航空实验室 (Cornell Aeronautical Laboratory) 时所发明的一种人工神经网络，被视为最简单的前向神经网络，是一种线性分类器。它能够把输入实数值向量 x 经过线性变换和激活函数映射到一个二元值 $f(x)$ 上。它还能够很好的解决线性可分的问题，然而 1969 年 Marvin Lee Minsky 在感知机 [1] 一书当中指出单层感知机不能拟合异或 (XOR)，以及不能解决线性不可分问题，并且多层感知机没有有效的训练方法。因此人工神经网络的第一次寒冬降临。

1.1.1 深度表示学习的第一次复苏

在深度表示学习的第一次复苏过程中，人们重新考虑有多个隐藏层的神经网络。于 1986 年，David Rumelhart, Geoffrey Hinton 和 Ronald Williams 合著的《Learning representations by back-propagating errors》[2] 打开了研究的进路，指出可以用反向传播解决多个隐层的学习问题。而《Learning internal representations by error propagation-hiton》[3] 特别探讨了感知机一书当中的问题，正是这两篇提出的构想让人们理解了如何解决多层神经网络的训练问题。另一个伟大的数学发现提振了人们对多层神经网络的信心，多层神经网络是普适函数拟合器 [4]。

在 1989 年，Yann LeCun 的团队验证了一个反向传播的杰出的应用，即反向传播应用于手写邮编识别 [5]。他的研究成果在 90 年代中期被成功应用到支票读取，他的采访和谈话中经常提到这一系统在那段时间读取了美国 10% 到 20% 的支票。此时，神经网络此时在无监督学习领域也开始发展，涌现出一些新的神经网络模

型，例如自编码器 [6]，自组织神经网络 [7]，玻尔兹曼机 [8] 等等。同时在强化学习领域也出现了广泛的应用 [9][10]。

天有不测风云，好景也没有长久下去。虽然反向传播在卷积神经网络中应用效果不错，但是其在深度的网络当中效果并不好。深度网络不容易训练，其回传的梯度积累起来要么过小，要么过大，这个现象被称作梯度消失和梯度的爆炸 [11]。为此 Jürgen Schmidhuber 及 Sepp Hochreiter 在 1997 年引进长短记忆网络 [12]，当中有一个遗忘门一定程度上起到管理回传梯度的作用。而且此时电脑不够快，神经网络还较为粗糙。当时重新兴起一个叫做支持向量机 [13] 的方法。而支持向量机的发展优于神经网络的发展。1995 年，Yann LeCun 的在手写数字识别学习算法的比较 [14] 一文中说明 SVM 比现有的神经网络算法好，或者至少水平一样。故此，机器学习社群对于神经网络的热度逐渐减退。此时，第二次寒冬来临。

1.1.2 深度表示学习的第二次复苏

第一个因素，在加拿大政府的研究资助支持下，Geoffrey Hinton 计划用深度学习来重新命名神经网络领域。2006 年，Geoffrey Hinton、Simon Osindero 与 Yee-Whye Teh 的一篇论文《A fast learning algorithm for deep belief nets》[15] 重燃人们对深度学习的兴趣。当中提出一种利用逐层训练多层受限玻尔兹曼机的方法。第二个因素，计算机算力的提升，利用图形处理器（GPU）计算，原来以周记的工作量可以按天来记 [16]。第三个因素，神经网络中选择的非线性的激活函数对性能影响很大，而原来的激活函数并不是最好的选择：Yann LeCun, Geoffrey Hinton, Yoshua Bengio 三组分别独立地都发现利用 Relu 激活函数即 $f(x) = \max(x, 0)$ 更好。它能够一定程度上缓解梯度消失的问题。有了上述的利好因素，深度学习开始重新复苏。

近年来在图像领域，2012 年，Geoffrey Hinton, Ilya Sutskever 和 Alex Krizhevsky[17] 将深度卷积神经网络架构 AlexNet 带到了 Imagenet 目标识别比赛中去达到了 10.8% 准确率的提升，高出使用传统的目标识别方法 41%；在语音领域演讲识别任务中，基于深度网络的模型更是超过最优的传统高斯混合模型，将识别错误率下降了 30%[18]（自 27.4% 到 18.5% 在 RT03S 数据集上）等；并且在引入变分自编码器 [19, 20] 的深度网络中，深度的因子可以无监督地刻画数据的一些内在变化，例如人脸数据上人的肤色变化，性别变化等等。在视频领域，Carl Vondrick, Hamed Pirsiavash 和 Antonio Torralba 利用生成对抗网络生成短视频 [21]，并且实现视频内容的分类 [22]。特别是在脑电领域，Pouya Bashivan 等人利用 LSTM-CNN 架构结合脑电数据来预测人的工作记忆 [23]。除此之外，深度表示学习正在包括 AlphaGo 下棋 [24]、唇语识别 [25]、图像生成 [26]、图像分类、艺术品和风格模仿 [27]、电子游戏 [28]、阅读理解 [29]、医疗病灶检测 [30] 等诸多领域达到或者超过人类水平。

然而，除了在语音领域这类时间序列任务上，深度表示学习能够提升其他时间序列任务例如水文预测任务的效果吗？变分自编码器经常会学得没有用的因子，怎么发现有影响的生成因子来实现更好的表示学习呢？在更复杂的视频类别数据上

(时间序列 \times 图像数据) 引入变分自编码器的深度表示学习能够取得怎样的效果呢?

1.2 文章架构

本文将初步从水文时间序列开始, 其次是图像因子学习, 以及脑电(视频)数据学习展开对于深度表示学习的有效性的研究, 并特别会探究变分自编码器模型的特点。这包括发掘其有影响的生成因子(即学得的重要表示), 以及这些因子的特性。我们将变分自编码器引入到脑电(视频)数据的表示学习任务中。后续章节分为4个部分。

第二章 主要针对时序数据深度学习方法进行介绍, 并构建典型时序深度学习网络, 深度长短时记忆网络, 对日径流量预测任务进行应用。具体地, 文章首先说明了深度长短记忆网络适合于日径流时间序列预测, 然后文章介绍了循环神经网络和长短记忆网络及深度长短记忆网络的日径流模型构造, 文章给出了预测的目标函数, 并分析了实验: 1. 设计了增加日径流时间序列训练长度, 来观察预测精度是否提高。2. 在相同的序列长度条件下, 预测结果同传统的回馈神经网络(BPNN)预测结果比较, 看看其预测结果是否更好。3. 在不同地区, 考察深度长短记忆网络模型是否普遍的优于长短记忆网络模型。最后进行了总结。

第三章 针对图像数据深度学习方法进行介绍, 并针对典型网络方法, 变分自编码器, 进行深入挖掘, 分析其因子作用, 并通过实验验证分析结果合理性。具体地, 文章介绍了变分编码器的生成过程、分类过程以及求解方法, 展示了变分自编码器用于发掘数据的变化, 接着阐述了如何发掘变分自编码器有影响力的生成因子: 首先说明了互信息是其中一个必要的指标, 并进行了被忽略因子的分析, 介绍了相关的工作, 给出了有影响力生成因子发掘实验、发掘因子的生成能力实验、发掘因子的分类能力实验。最后文章讨论了因子的等价特性, 并进行了总结。

第四章 针对视频数据深度学习方法展开, 特别讨论变分自编码器与长短时记忆网络在脑电数据情感预测的模型和实验。具体地, 首先介绍了脑电数据情感能识别模型, 其分为变分自编码器部分、长短记忆网络部分和图模型部分, 然后进行了对比实验比较了自编码器与长短记忆网络模型、卷积神经网络与长短记忆网络模型以及利用功率谱密度的支持向量机模型和变分自编码器与长短记忆网络模型, 最后进行了总结。

第五章 对全文进行总结以及展望。

2 时序序列的深度表示学习

时间序列是指按照统一指标，且按其发生的时间顺序排列得到的数列 $\{x^i\}_{i=1}^T$ 。而时间序列一般问题是利用之前的变化来预测未来的数列（即利用 x^1 预测 x^2 ，利用 x^1, x^2 预测 x^3, \dots ，利用 x^1, \dots, x^{T-1} 预测 x^T ）。其概率分布为

$$p(x^2, \dots, x^T | x^1) = p(x^2 | x^1)p(x^3 | x^2, x^1) \cdots p(x^T | x^1, \dots, x^{T-1}). \quad (2-1)$$

传统的 M 阶马尔可夫模型假定时间序列数据在条件上 M 个时刻的元素，是与剩下其他时刻数据独立的（即，

$$p(x^2, \dots, x^T | x^1) = p(x^2 | x^1)p(x^3 | x^2, x^1) \cdots p(x^T | x^{T-m}, \dots, x^{T-1}). \quad (2-2)$$

然而受限于其参数随着 M 的增加指数级别增长，这使得这类模型没有办法建模较长时间的依赖性。

2.1 循环网络与长短记忆网络

与马尔可夫模型相比之下，循环神经网络理想情况下，可以建模不定长时间的依赖关系。假设

$$x^t = f^{t-1}(x^{t-1}, \dots, x^1) + \varepsilon. \quad (2-3)$$

其中

$$f^{t-1}(x^{t-1}, \dots, x^1) = Wh^t = WA(x^{t-1}, h^{t-1}), \quad (2-4)$$

$$h^t = A(x^{t-1}, h^{t-1}) = A(x^{t-1}, A(x^{t-2}, h^{t-2})) = \cdots = G(x^{t-1}, x^{t-2}, \dots, x^0, h^0), \quad (2-5)$$

$\varepsilon \sim N(0, \sigma^2 I)$, $h^0 = 0$. A 为循环神经网络，它输入了上一次的序列值 x^{t-1} 和隐藏状态 h^{t-1} ，输出下一个时刻的隐藏状态 h^t 。 W 为线性变换的矩阵。循环网络的整体展开结构由图2-1 所示。

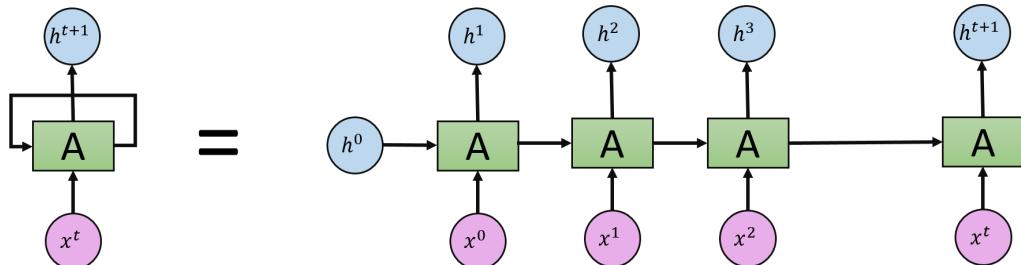


图 2-1 循环网络示意图

从中可以看出，循环神经网络 (RNN) 比传统的前向神经网络多了状态量的传

递和处理过程。其中，网络会对前面的信息进行处理记忆，并应用于当前输出的计算中，即隐藏层里的节点不再无连接，而是有连接的，同时隐藏层的输入不仅含有输入层的输出，还包括上一时刻隐藏层的输出。

那么目标函数可以变成如下形式，

$$\begin{aligned}
 \ln p(x^2, \dots, x^T | x^1) &= \ln p(x^2 | x^1) + \dots + \ln p(x^T | x^1, \dots, x^{T-1}) \\
 &\propto \|x^2 - f^1(x^1)\|^2 + \|x^3 - f^2(x^1, x^2)\|^2 \\
 &\quad + \dots + \|x^T - f^{T-1}(x^1, x^2, \dots, x^T)\|^2 \\
 &= \sum_{i=1}^{T-1} \|x^{i+1} - WA(x^i, h^i)\|^2.
 \end{aligned} \tag{2-6}$$

理论上，这样的结构能使得 RNN 可以处理任意长度的序列数据，然而实际上由于种种原因，训练这样的普通 RNN 结构难以捕捉序列数据中一些时间间隔很长的输入输出间的依赖关系，由此，有关的研究人员引入了长短时记忆 (Long Short-Term Memory) 来弥补这个缺陷。长短时记忆 (Long Short-Term Memory) 通过引入输入门、输出门和遗忘门来控制隐藏状态量中信息的累积速度，并可以选择遗忘之前隐藏状态量中累积的信息。这允许网络可以学习何时遗忘历史信息、何时用新信息更新记忆单元。相关研究已经表明，这种网络结构在处理长时间序列依赖关系问题非常有效。一般把长短时记忆 (Long Short-Term Memory) 网络简称为 LSTM 网络或长短记忆网络。

单个 LSTM 模型的结构如下图 2-2 所示。

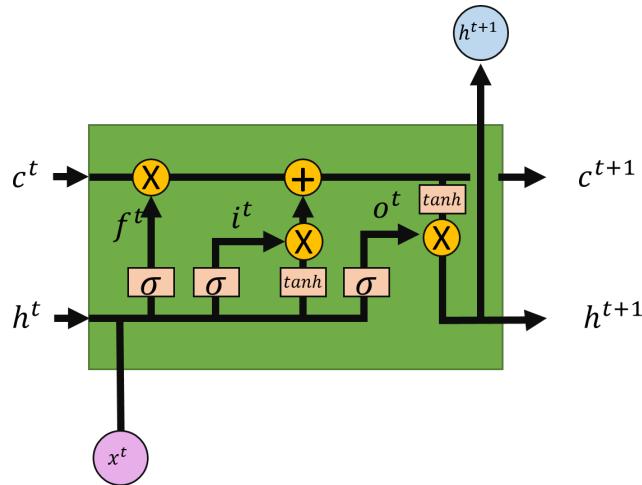


图 2-2 LSTM 循环网络示意图

状态初始化 $\mathbf{h}^0 = \mathbf{c}^0 = 0$, 输入: x^t , 门的计算:

$$\mathbf{f}^t = \sigma(W^f[\mathbf{h}^{ttr} x^{ttr}]^{tr} + \mathbf{b}^f), \tag{2-7}$$

$$\mathbf{i}^t = \sigma(W^i[\mathbf{h}^{ttr} x^{ttr}]^{tr} + \mathbf{b}^i), \tag{2-8}$$

$$\mathbf{o}^t = \sigma(W^o[\mathbf{h}^{ttr} x^{ttr}]^{tr} + \mathbf{b}^o), \tag{2-9}$$

状态更新:

$$\tilde{\mathbf{c}}^t = \tanh(W^c[\mathbf{h}^{ttr}x^{tr}]^{tr} + \mathbf{b}^c), \quad (2-10)$$

$$\mathbf{c}^{t+1} = \mathbf{f}^t \odot \mathbf{c}^t + \mathbf{i}^t \odot \tilde{\mathbf{c}}^t, \quad (2-11)$$

$$\mathbf{h}^{t+1} = \mathbf{o}^t \odot \tanh(\mathbf{c}^t), \quad (2-12)$$

输出:

$$y^{t+1} = W^y\mathbf{h}^{t+1} + b^y. \quad (2-13)$$

其中 $\mathbf{i}^t, \mathbf{f}^t, \mathbf{o}^t, x^t, \mathbf{c}^t, \mathbf{h}^t$ 分别是 LSTM 模型的输入门, 遗忘门, 输出门, 输入向量, 隐藏状态向量(记忆和输出部分)。 W^*, \mathbf{b}^* 是待学习的网络中的权重矩阵和偏差变量。 σ 表示逐元素(element-wise)的 sigmoid 激活函数, \tanh 表示逐元素的反正切激活函数, \odot 表示哈达玛乘积(Hadamard Product, 逐元素乘积)。 \mathbf{h}^{tr} 表示 \mathbf{h} 的转置。可以看到隐藏状态 c^{t+1} 由输入门和遗忘门来管控, 当遗忘门等于 1, 输入门等于 0, 那么内容 c^{t+1} 将会保留与 c^t 一样的值, 使得过去的信息能够长期在隐藏状态中保留。这使得长短记忆网络能够将较长的依存关系保存下来。

2.2 深度长短记忆网络用于日径流量预测

2.2.1 日径流时间序列预测的意义

水资源是人类的生命之源, 其有效的促成了人类文明的快速发展, 深刻地影响着人类社会的经济发展和国民生活。世界地球上总的水量大约 13.86 亿 km^3 , 然而可利用的淡水总量仅占其 2.53% , 而其中还又有 68.7% 的淡水以冰雪形式存在, 其主要分布在地球的南极北极地区和中低维度的高山上, 30% 存在于地下 [31]。人类可以方便且直接利用的淡水资源主要来自于沼泽、河流、湖泊(水库) [32], 其水量仅占地球水量的 0.014% 合计 19 万 km^3 [33]。这部分占比很小的水资源却与我们人类息息相关。由于我国人口众多, 尽管我国地域幅员辽阔、水资源总量大, 但是我国人均的可利用的水资源却少之又少。另一方面, 防洪是人类一直面临的重要挑战。洪水可能是多种自然灾害(例如地震、热带风暴、火山爆发、雪灾)中造成整体经济损失最严重的自然灾害之一。Aon Benfield 在 2016 年发表的报告显示, 洪水灾害已持续四年成为全球范围内致经济损失最严重的自然灾害之一, 其所造成的直接间接经济损失高达 620 亿美元。特别是那些修建在低洼地带的基础设施和功能区域, 当洪水来临时将会蒙受巨大的经济损失和人员生命安全损失。我国的水资源在空间和时间上分布也不均匀, 年内和年间水量变化也非常之大。

径流是指某一特定地区内部的地底或者地表自由流动的水流, 其具有时序特性且具有非线性。径流预测意在对历史的水文数据和气象数据进行采集分析, 从而对特定区域地底或者地表自由流动的水流趋势进行预测。日径流量是洪水猛烈程度重要特征之一, 对日径流的预测在水文领域非常重要, 对于防洪、制定生产的计

划、抗旱、发电、金融分析、水资源的综合利用和规划管理等起着关键的作用：径流量预测是研究水质水环境等问题的研究基础，有助于在气候和环境不断恶化、工农业污染、人类随意开采等复杂因素下，解决水生态受损、水环境质量差等问题；从国家角度来讲，日径流预测有助于帮助国家的生态建设和可持续发展；有效的预测径流流量，更是有助于减少和防护洪水灾害，且可以更好的利用水资源，以降低其带来的灾难损失；日径流量同时也可以为工业农业用水提供参考，帮助规划工业农业用水的调度和使用，从而使得经济效益最大化；水环境质量管理、水库调度也需要精准的日径流预测；日径流同时对水利发电行业有着一定参考作用。

2.2.2 日径流时间序列预测的背景和相关工作

有关水文时间序列的研究最早可以追溯古埃及王国，古埃及王国的尼罗河水位的涨落变化被古埃及人记录，从那时起日径流建模是水文领域的一大挑战。至今相关模型非常广泛，从纯粹的数据驱动模型到基于物理的认知的模型都有存在。然而基于物理和空间显示来表示水文过程的模型需要消耗大量的计算代价和大量的有必要的输入数据 [34]。这使得径流预测在操作中往往少有采用精细的物理建模的模型而是采用集成模型。而常规应用的操作中运用到集成模型的子模型，往往是基于简单的物理和认知的模型或者是数据驱动的模型。所以导致数据驱动建模的概念（例如神经网络、基于混沌的方法、基于回归的方法以及基于数据机理的模型）发展起来并且在本文中被探索。

人工神经网络非常适合拟合复杂的系统。早在 1990 年代，人们就开始使用人工神经网络来预测径流数据 [35]。之后，很多研究应用人工神经网络建模径流过程例如 [36]。在水文领域人工神经网络也有很多的应用：例如基于人工神经网络的湘江最大洪峰流量的中长期预测 [37]、基于神经网络的长江三峡年最大洪峰流量长期预报、人工神经网络峰值识别理论及其在洪水预报中的应用 [38]。然而，人工神经网络存在着缺点。人工神经网络会使得序列的时序信息受到损失，因为人工神经网络一次的输入有限，并且没有对之前的输入的信息有记录的功能。循环神经网络作为一种新的形式的神经网络有一个隐藏状态，能够描述之前输入的时间序列的信息，这使得循环神经网络能够克服人工神经网络的不足，并用于径流数据的预测 [39]。

然而传统的循环网络学习较长时间的依存关系存在着一些问题 [12]，而这样的依存关系在水文日径流预测中关键且非常重要。[12] 提出了一种特别设定的网络结构，并称之为长短时记忆（Long Short-Term Memory）。这种网络通过引入输入门、输出门和遗忘门，使得距离当前较长时间的信息，在遗忘门关闭的情况下可以持续保留，这使得之能够克服传统循环网络不能学习较长时间的依存关系的问题。

这些年深度学习开始吸引人们的注意。随着图形处理器单位（GPU）的快速发展以及大型的数据集的出现，使得深度学习在各个领域的成功。其中，最成功的深度学习应用在计算机视觉 [17]，演讲识别和自然语言处理等，很少有将深度学习引

入水文领域。目前已知[40]也采取了深度学习的长短记忆网络架构，但其网络的具体宽度以及应用地点范围及应用对象（为降水日径流，本文研究日径流）不同。水文领域，人们用深度学习多层感知机预测洪水[41]、用深度学习长短记忆网络做降水量仿真[42]、用深度学习长短记忆网络进行大规模水文建模[43]、用长短记忆网络进行农业区域水位深度的预测[44]。很多人研究使用深度学习模型来替代传统模型，他们的结果是深度学习模型往往更出色。大体上来说，深度学习方法在水科学和水文领域只有近年来才成为讨论的焦点。这也留给我们利用深度长短记忆网络刻画日径流数据好的机会。

2.2.3 深度长短记忆网络用于日径流时间序列预测的合适性

首先，由混沌理论的研究表明，自然状态下日径流数据中存有不同时间长度拟周期的依赖关系[45]。而选用较长时间输入的（未特殊设计的）全连接前向网络来学习这种相互的关系是低效的[36]，并且依赖关系还可能涉及多个时间间隔期，难以构造这样的前向神经网络输入。而长短记忆网络的结构适于学习这种有较长依存关系的时间序列数据。其次，由于日径流数据的非线性系统关系复杂，而深度学习能够赋予了长短记忆网络的隐藏状态更强的表示能力，所以我们采取两层的长短记忆网络结构，这有助于更好的拟合这种复杂的关系。最后，也最重要的是，日径流时间序列的数据量较为充足，这能够缓解深度学习模型的过拟合现象。

2.2.4 深度长短记忆网络日径流模型构造

本文采用含两层长短记忆网络隐含层和一层全连接前向输出网络的结构，构建一个深度长短记忆网络模型见图2-3。我们记时间次序 $t = 0, \dots, T$ ，模型隐层的深度序号 $d = 0, 1$ 为堆叠的长短记忆网络层，每层节点数为 128。状态初始化 $\mathbf{h}^{0d} = \mathbf{c}^{0d} = 0$ ，输入： $x^{t0} = x^t, \mathbf{x}^{t1} = \mathbf{h}^{t+10}$ ，门的计算：

$$\mathbf{f}^{td} = \sigma(W^{fd}[\mathbf{h}^{tdtr} \mathbf{x}^{tdtr}]^{tr} + \mathbf{b}^{fd}), \quad (2-14)$$

$$\mathbf{i}^{td} = \sigma(W^{id}[\mathbf{h}^{tdtr} \mathbf{x}^{tdtr}]^{tr} + \mathbf{b}^{id}), \quad (2-15)$$

$$\mathbf{o}^{td} = \sigma(W^{od}[\mathbf{h}^{tdtr} \mathbf{x}^{tdtr}]^{tr} + \mathbf{b}^{od}), \quad (2-16)$$

状态更新：

$$\tilde{\mathbf{c}}^{td} = \tanh(W^{cd}[\mathbf{h}^{tdtr} \mathbf{x}^{tdtr}]^{tr} + \mathbf{b}^{cd}), \quad (2-17)$$

$$\mathbf{c}^{t+1d} = \mathbf{f}^{td} \odot \mathbf{c}^{td} + \mathbf{i}^{td} \odot \tilde{\mathbf{c}}^{td}, \quad (2-18)$$

$$\mathbf{h}^{t+1d} = \mathbf{o}^{td} \odot \tanh(\mathbf{c}^{td}), \quad (2-19)$$

输出：

$$y^{t+1} = W^y \mathbf{h}^{t+11} + b^y. \quad (2-20)$$

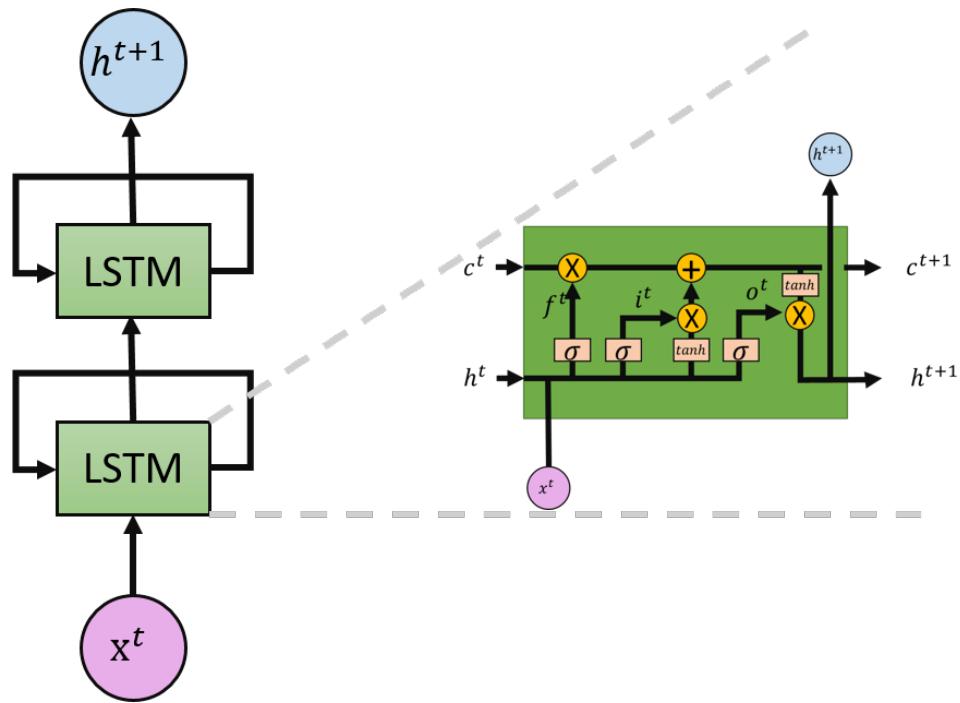


图 2-3 构建的深度 LSTM 网络模型示意图。

其中 $f^{td}, i^{td}, o^{td}, x^{td}, c^{td}, h^{td}$ 分别是两层 LSTM 模型的遗忘门，输入门，输出门，输入向量，隐藏状态向量（记忆和输出部分） $(d=0,1)$ 。 W^{*d}, b^{*d} 是待学习的网络中的权重矩阵和偏差变量 $(d=0,1)$ 。 σ 表示逐元素（element-wise）的 sigmoid 激活函数， \tanh 表示逐元素的反正切激活函数， \odot 表示哈达玛乘积（Hadamard Product，逐元素乘积）。

深度长短记忆网络模型展开图，如图2-4所示，具有两层循环连接的结构。这使得其隐藏状态有着更强的表示能力，更倾向于能够刻画出自径流数据复杂的非线性依存关系。

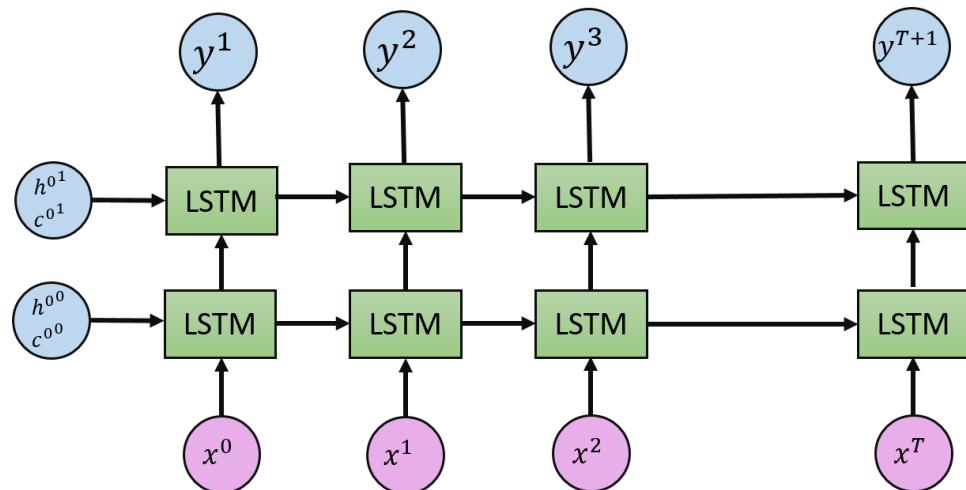


图 2-4 深度长短记忆网络模型展开示意图。

2.2.5 模型目标函数

假设平均流量数据由 $\{x^0, \dots, x^T\}$ 来表示, 给定初始隐藏状态 $\{h^{0d}, c^{0d}\}$ ($d=0,1$), 那么经过两层 LSTM 网络, 依据最大似然的推导2-6可得目标函数,

$$\min_{all \ the \ W^*, \ b^*} \sum_{i=1}^T \|y^i - x^i\|^2. \quad (2-21)$$

2.2.6 应用实例及分析

我们选择长江宜昌水文站的日平均流量序列资料作为研究。该站位于长江上游的宜昌市, 设立于 1877 年 4 月, 有集水面积 1,005,500 平方公里, 其距离河口 1,855 公里。该站观测资料序列较长, 文中使用了 1945 年 10 月 1 日开始到 1962 年 12 月 31 日为止的逐日平均流量资料进行分析, 我们将 1945-1959 作为 LSTM 神经网络训练学习资料, 1960 年到 1962 年的资料作预测对比。这个时间段葛洲坝还未受人工干预, 因此对于分析极为有利。大江大河汛期的预测结果十分重要, 所以我们对比时段选择了宜昌水文站 1960 年的 7 月、1961 年 8 月和 1962 年 9 月的资料, 来进行预测结果比较。

1) 提出设想

为了证明深度 LSTM 的优越性和有效性, 我们从三方面进行了考虑。

- 1. 设计了增加日径流时间序列训练长度, 即由 1950-1959 年的 10 年的 LSTM 训练学习长度增加到 1945 年 10 月 1 日 -1959 年 12 月 31 日的 15 年的训练长度, 来观察预测精度是否提高。
- 2. 在相同的序列长度条件下, 预测结果同传统的回馈神经网络 (BPNN) 预测结果比较, 看看其预测结果是否更好。
- 3 在不同地区, 考察深度 LSTM 模型是否普遍的优于 LSTM 模型。

2) 判别标准的选取

为了定量比较预测结果, 水文行业的通常的判别方法被选用 (即, 统计预测结果的相对平均误差 ($m = \sum_{i=1}^T \frac{|y_i - x_i|}{x_i}$), 其绝对值越小说明系统误差越小; 相对标准差 ($(s = \sqrt{\frac{\sum_{i=1}^T \frac{(y_i - x_i)^2}{n-1}}{\sum_{i=1}^T x_i}})$), 其绝对值越小说明预测的结果更集中在均值附近, 意味着泛化能力好; 相对最大正误差 ($Max+ = \max_i(0, \frac{y_i - x_i}{x_i})$)、相对最大负误差 ($Max- = \min_i(0, \frac{y_i - x_i}{x_i})$) 和极差 ($r=Max+-Max-$), 其值越小则长序列训练学习后的预测值偏离真实值的范围更小)。

3) 模型的训练

我们选用 tensorflow 平台实现深度长短记忆网络模型。其中利用了 tensorflow.contrib.rnn.BasicLSTMCell 中默认初始化方法对长短记忆网络门变量和权重变量进行初始化, 同时采用截断单位高斯分布作为输出 W_y 的权重初始化方法。对

时序训练数据进行线性标准化 $\frac{(x-3100)}{63000}$ 归一化于近似 [0,1] 区间，并采用学习率为 0.0001 的 AdamOptimizer 作为梯度优化方法 [46] 来方便模型训练。训练过程中循环网络的损失呈现震荡下降性质。我们采用了早停 (early stopping) 的方法 [6] 来防止模型过拟合：保存训练次数（每隔 600 次保存一遍模型）不同情况下的多组模型，并利用 1960 年 1 月到 6 月数据作为检验集挑选最优的模型。最终我们选取 10 年序列训练次数 (epoch) 为 6000 次的模型，和近似 15 年序列训练次数为 8400 次的模型作为两组最终的模型。对于万县、寸滩和奉节水文站的数据，我们选取了训练次数为 9000 次的模型。

4) 实验结果

本文按上述的方案用短（1950-1959 年）长（1945-1959 年）序列训练后分别做了计算预测。采用 1960 年 7 月和 1961 年 8 月的预测值和实际值的日平均流量过程线对比图，见图 2-5 和图 2-6。1960 年 7 月前半个月是一场小的洪水的退水过程，后半个月是一场中等的洪水过程，而 1961 年 8 月是一场中等和一场较大的复式洪水过程。文中 1962 年 9 月是一场洪水退水过程，本文不列其图。从图 2-5、2-6 中可以看出预测的趋势过程比较好，具体请看下面的误差统计指标表 1 和表 2。

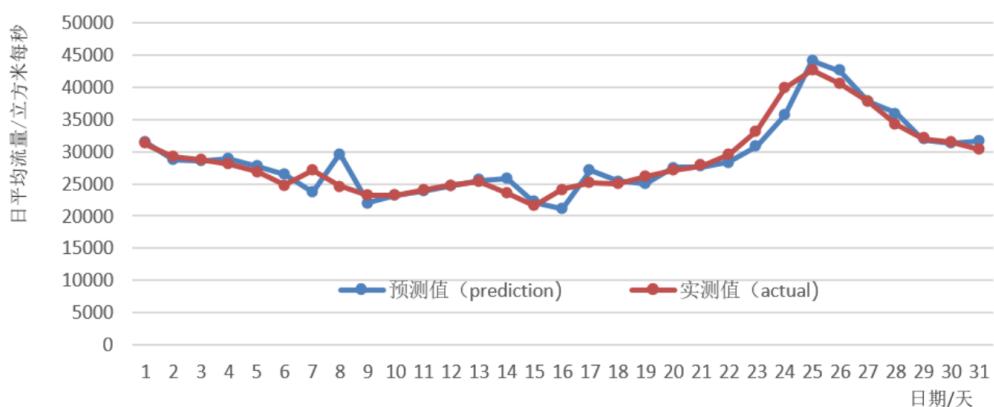


图 2-5 1960 年 7 月实测及预测平均流量对比图。

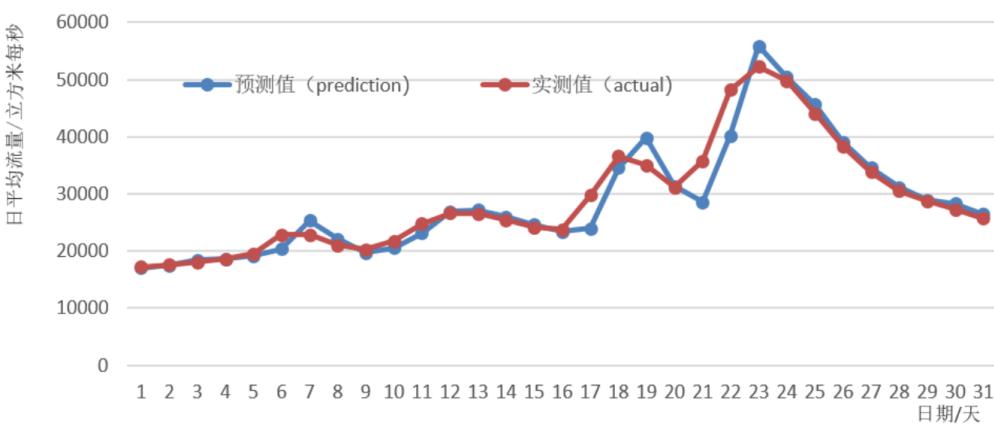


图 2-6 1961 年 8 月实测及预测平均流量对比图。

(1) 设想一

为了验证前面的设想，我们对洪水情期间预测结果统计的各误差指标见下表。从表2-1可以看出，模型预测相对平均误差(m)绝对值都很小，说明预测的系统偏差(bias)很小，所以预测结果都较为准确；长序列训练学习后的预测相对标准差 s 值总体要小些，说明预测的结果更集中在均值附近，意味着泛化能力好；而误差范围(极差 r)越小则长序列训练学习后的预测值偏离真实值的范围更小。因此，综合本表2-1比较结果可以看出用深度长短记忆网络模型的长短期记忆功能被体现，而总体的相对标准差随着训练样本的增加在缩小，更是说明其预测泛化精度在提高。

表 2-1 长短记忆网络预测成果统计表

样本年		1950-1959						1945-1959					
统计量		m	s	MAX+	MAX-	r	m	s	MAX+	MAX-	r		
预测年月	1960.7	0.45	6.51	20.43	-12.99	33.42	0.51	6.65	21.76	-12.69	34.45		
	1961.8	-0.89	7.59	13.74	-20.32	34.06	-0.44	7.32	13.52	-19.83	33.35		
	1962.9	1.08	4.21	9.85	-7.66	17.51	0.81	4.03	10.16	-6.99	17.15		
	平均	0.21	6.10	14.67	-13.66	28.33	0.29	6.00	15.15	-13.17	28.32		

(2) 设想二

表2-2是主要汛期的长短记忆网络(LSTM)与反向传播神经网络(BPNN)神经网络预测对比情况。表中的反向传播神经网络(BPNN)法结合了混沌法，从24个方案中优先出的4-4-1型的反向传播神经网络(BPNN)计算预测的最优结果[36]。表2-2中的 m 值表明反向传播神经网络(BPNN)和深度长短记忆网络预测误差均值都很小，说明预测的系统偏差很小，这也说明预测结果准确度都较高；深度长短记忆网络训练学习后的预测误差统计相对标准差值较比反向传播神经网络(BPNN)法要小很多，这说明总体预测的结果比反向传播神经网络(BPNN)精确度更高；而误差范围极差 r 值则长短记忆网络明显的好很多。因此，比较结果得出，深度长短记忆网络法预测结果明显比反向传播神经网络(BPNN)好。

表 2-2 长短记忆网络(LSTM)与反向传播神经网络(BPNN)预测结果比较表

样本年		1950-1959						1950-1959					
训练预测方法		LSTM						BPNN					
统计量		m	s	MAX+	MAX-	r	m	s	MAX+	MAX-	r		
预测年月	1960.7	0.45	6.51	20.43	-12.99	33.42	0.07	8.54	27.25	-18.28	45.53		
	1961.8	-0.89	7.59	13.74	-20.32	34.06	-1.87	8.13	15.2	-22.42	37.62		
	1962.9	1.08	4.21	9.85	-7.66	17.51	0.04	5.73	12.26	-11.83	24.09		
	平均	0.21	6.10	14.67	-13.66	28.33	-0.59	7.47	18.24	-17.51	35.75		

(3) 设想三

对于宜昌，万县，寸滩和奉节水文站，我们比较了深度长短记忆网络模型和普通长短记忆网络模型。我们采用了宜昌水文站1950年到1958年的数据去预测1960

年到 1962 年的数据。我们利用了万县，寸滩和奉节水文站 1990 年到 2000 年的数据去预测 2001 年的数据。

从表 2-4 和 2-3，我们可以发现深度 LSTM 模型在宜昌、寸滩和万县水文站数据上在更多的指标上优于 LSTM 模型。具体地在宜昌水文站深度长短记忆网络在相对标准差、最大正误差、最大负误差和极差取值上都优于普通长短记忆网络模型；在万县水文站，深度长短记忆网络在相对平均误差、最大正误差、最大负误差和极差取值上都优于普通长短记忆网络模型；在奉节水文站，深度长短记忆网络在预测相对平均误差和相对标准差取值上都优于普通长短记忆网络模型；寸滩水文站深度长短记忆网络在相对平均误差和最大正误差取值上都优于普通长短记忆网络模型。因为相对平均误差在所有位置都很小，这展示了长短记忆网络模型和深度长短记忆网络普遍的适用性。

表 2-3 LSTM 和深度 LSTM 对比结果

水文站	宜昌 (1950-1959/1960-1962)					万县 (1990-2000/2001)				
指标	m	s	MAX+	MAX-	r	m	s	MAX+	MAX-	r
LSTM	0.44	8.96	26.37	-30.26	56.63	0.56	8.99	28.47	-25.57	54.04
深度 LSTM	0.48	8.95	24.77	-30.14	54.91	0.55	9.00	27.37	-25.52	52.89

表 2-4 LSTM 和深度 LSTM 对比结果

水文站	奉节 (1990-2000/2001)					寸滩 (1990-2000/2001)				
指标	m	s	MAX+	MAX-	r	m	s	MAX+	MAX-	r
LSTM	0.80	7.39	15.48	-29.45	44.93	0.33	13.36	36.57	-38.42	74.99
深度 LSTM	0.46	7.14	16.79	-30.29	46.08	0.30	13.40	36.46	-38.53	74.99

2.2.7 结论

本文得到结论：深度长短记忆网络方法用于日径流时间序列是可行的。对于设想一，随着训练序列长度的增加，预测泛化精度进一步提高，极差将进一步缩小。当然，由于 1945 年以前的资料存在中断，本次计算没有继续往前延伸，相信如果系列能继续加长，预测精度提高应该也会更加明显；

对于设想二，深度长短记忆网络预测方法明显比浅层的反向传播神经网络（BPNN）方法预测时间序列的效果好。尽管结合混沌理论方法的 BPNN 是从 24 个方案中优选出来的，但是因为反向传播神经网络（BPNN）的方法只有一个隐含层，没有深度表示学习，所以反向传播神经网络（BPNN）反映出的非线性系统复杂关系的能力有所欠缺。而相反的，长短记忆网络所含有的长的短期记忆功能，赋予了长短记忆网络能够学习更加复杂的深度表示的能力。

对于设想三，在不同地区的实验展现出深度长短记忆网络模型更优于长短记忆网络模型。我们可以发现深度长短记忆网络模型在宜昌、寸滩和万县水文站

数据上在更多的指标上优于长短记忆网络模型。具体地在宜昌水文站深度长短记忆网络在相对标准差、最大正误差、最大负误差和极差取值上都优于普通长短记忆网络模型；在万县水文站，深度长短记忆网络在相对平均误差、最大正误差、最大负误差和极差取值上都优于普通长短记忆网络模型；在奉节水文站，深度长短记忆网络在预测相对平均误差和相对标准差取值上都优于普通长短记忆网络模型；寸滩水文站深度长短记忆网络在相对平均误差和最大正误差取值上都优于普通长短记忆网络模型。这些都体现出深度长短记忆网络模型可以更加清楚的反映出其复杂的内在关系。由于相对平均误差在所有位置都很小，因此这展示了普通长短记忆网络模型和深度长短记忆网络都适用于日径流预测。

3 图像数据的深度表示学习

学习有效的图像低维数据表示在机器学习和相关应用中具有重要意义。高效、本质的低维表示有助于挖掘数据的基础知识，并为后续任务（包括生成、分类和关联）提供服务。早期线性维数约简（主成分分析, Principle Component Analysis）在原始数据分析中得到了广泛的应用。其变种已被应用于人脸识别 [47]。经典线性独立表示（独立成分分析, Independent Component Analysis [48] 在盲源分离中得到了广泛的应用。非线性维数降维（例如自编码器, Autoencoder[6]）能够进一步学习抽象表示，并已用于语义哈希 [49] 和许多需要学习抽象表示的其他任务。最近，一种称为变分自动编码器的新技术 [19][20] 由于具有提取非线性独立表示的能力而引起了研究者的广泛关注。该方法可以进一步建立因果关系模型，表示分离的视觉变化 [50] 和可解释的时间序列变量。该方法可用于以“因子可控”的方式产生多种多样的信号 [51, 52]。相关技术使知识能够通过不同任务之间的共享因素迁移 [53]。

3.1 变分自编码器及生成模型的背景

在大规模的数据集上，变分自编码器用于对有向概率模型当中关于连续隐藏变量和难以计算的后验概率分布，实施有效的学习和推断。其引入了一个推断识别网络模型来表示估计后验分布，并且使用重参数化技巧来进行随机联合优化一个变分下界函数，其包含了生成模型和推断模型的参数。

在变分编码器提出之后，出现了很多提升变分自编码器的生成质量和其解耦合能力的变分自编码器变种。在这些方法中，有很多方法努力改进生成模型和推断模型的结构。这个方向上，典型的工作包含由 [54] 提出的卷积/反卷积结构和 [55] 提出的梯子状层级网络结构。一些其他工作则在变分自编码器的生成过程和推断过程中更进一步。主要工作包含由 [56] 提出的迭代注意力生成推断机制。[57] 提出的正则化流（Normalizing Flow）方法增强了变分自编码器估计后验的表示能力，其相关的一些变种 [58] 也达到了同样的功能。

除了对于变分自编码器直接的模型上的改造，一些其他努力致力于集成生成对抗网络（Generative Adversarial Nets）与变分自编码器。例如 [59] 集成了生成对抗网络和变分自编码器，去获得更好的重建和高层次抽象视觉嵌入特征。[50] 也集成了生成对抗网络和变分自编码器，但是更强调解耦因子的变化。虽然受制于不稳定的训练过程和模式崩塌 [60]，但是在不考虑问题噪声水平的情况下，没有额外设计的生成对抗网络依旧可以学习数据分布。而变分自编码器包含了噪声和理想的干净数据的分解的假设。其还假设了关于因子的先验。

除此之外，人们在正则化因子的分布和因子的生成效果上进行努力。例如 [61] 引入了对抗损失到自编码器的隐藏空间。其在理想情况下可以学习任意隐藏的分布，包含那些解耦的数据分布。由 [62] 提出的 InfoGAN 通过额外的互信息正则引入

了信息最大化原理（Infomax Principle）到生成对抗网络中。这个互信息正则允许生成对抗模型能够进行推断并且引导模型学得一个更好解耦的表示。

近年来，[63] 提出了一种新的变分自编码器变种。该种 β -VAE 架构强化了变分自编码器后验分布和先验分布的 KL 散度约束，并且展现了新奇的解耦表现。这个方法相比于传统的变分自编码器方法已经获得更好的效果，尤其是它能够方便的调整一个在 KL 散度和似然项之间的简单参数 β 以实现一定程度的解耦功能。

3.2 变分自编码器

具体的来说，变分自编码器是一个可规模化的无监督学习算法，它假设输入 \mathbf{x} 由若干独立同分布的高斯随机变量 \mathbf{z} 生成（即， $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(0, I_H)$ ）。因为高斯分布可以连续可逆映射到很多其他分布，所以对于变分编码器的理论分析可以对于其他连续隐变量的 VAE 具有指导作用。它的生成/解码过程为 $p_{\text{dec}}(\mathbf{x}|\mathbf{z})$ 并且推断和编码过程建模为 $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_H(\mathbf{x})))$ 。我们假设它们由以 enc 和 dec 为参数的神经网络参数化。

因子：设 \mathbf{z}_{enc} 是以 $q_{\text{enc}}(\mathbf{z}) = \int q_{\text{enc}}(\mathbf{z}|\mathbf{x}) p_{\text{data}} d\mathbf{x}$ 为分布的因子随机变量，并且一个因子是指 \mathbf{z}_{enc} 的一个维度。

3.2.1 生成过程

在变分自编码器设置中，估计推断被用到了最大化变分下界 $\log p_{\text{dec}}(\mathbf{x}) = \log \int p_{\text{dec}}(\mathbf{x}|\mathbf{z}) p_{\text{dec}}(\mathbf{z}) d\mathbf{z}$ 中，

$$\begin{aligned} L_{\text{rec}} &= \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{z}|\mathbf{x})} \log p_{\text{dec}}(\mathbf{x}|\mathbf{z}) - D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || p_{\text{dec}}(\mathbf{z})) \\ &\leq \log p_{\text{dec}}(\mathbf{x}) \end{aligned} \quad (3-1)$$

等式成立当且仅当

$$D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || p_{\text{dec}}(\mathbf{z}|\mathbf{x})) = 0 \quad (3-2)$$

为了去限制信息容量 [63]，引入 $\beta > 1$ 到目标函数的第二项中：

$$\begin{aligned} \mathcal{L}_{\text{rec}-\beta} &= \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{x}|\mathbf{z})} \log p_{\text{dec}}(\mathbf{x}|\mathbf{z}) - \beta D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || p_{\text{dec}}(\mathbf{z})) \\ &< \log p_{\text{dec}}(\mathbf{x}). \end{aligned} \quad (3-3)$$

在完成训练之后，通过对 $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, I_H)$ 采样或者有目的设置 z ，来利用训练好的 $p_{\text{dec}}(\mathbf{x}|\mathbf{z})$ 生成新样本。

3.2.2 分类过程

$q_{\text{enc}}(\mathbf{z}|\mathbf{x})$ 可以进一步支持后续的任务，例如分类。设 $p_{\text{pre}}(\mathbf{y}|\mathbf{z})$ 为预测过程，并且分类过程目标函数是 $\mathcal{L}_{\text{pre}} = \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{z}|\mathbf{x})} \log p_{\text{pre}}(\mathbf{y}|\mathbf{z})$ 。在实际优化过程中上述损失需要进一步在数据分布上取期望。

3.2.3 求解方法

因为

$$q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_2(\mathbf{x}))), \quad (3-4)$$

那么给定 $\boldsymbol{\mu}$ 和 $\text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_2(\mathbf{x}))$ 我们可以通过从 $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ 采样，然后计算

$$\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_2(\mathbf{x}))^{\frac{1}{2}} \times \boldsymbol{\varepsilon}, \quad (3-5)$$

来实现从 $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$ 中进行采样。同时我们可以将目标函数转化为如下形式，

$$\begin{aligned} L_{\text{rec}} &= \mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)} \log p_{\text{dec}}(\mathbf{x}|\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_2(\mathbf{x}))^{\frac{1}{2}} \times \boldsymbol{\varepsilon}) \\ &\quad - D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || p_{\text{dec}}(\mathbf{z})). \end{aligned} \quad (3-6)$$

3.3 变分自编码器用于发掘数据变化的功能

图3-2, 图3-3和图3-4呈现了由变分编码器在人脸，手写数字以及脑电图数据中学得的数据变化。

下图3-1为本人脸作为输入得到的变化。包含了人脸男女的变化，脸颜色黄到白的变化，背景黄到蓝的变化，头发颜色黑到白。

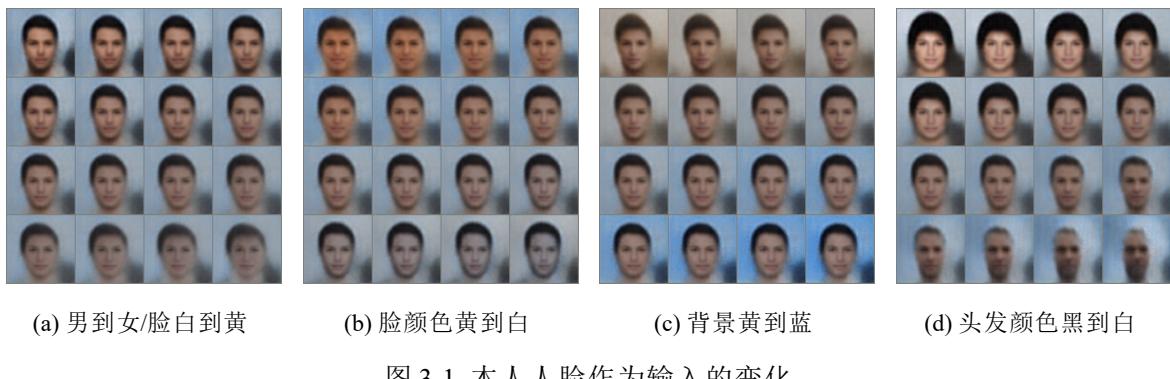
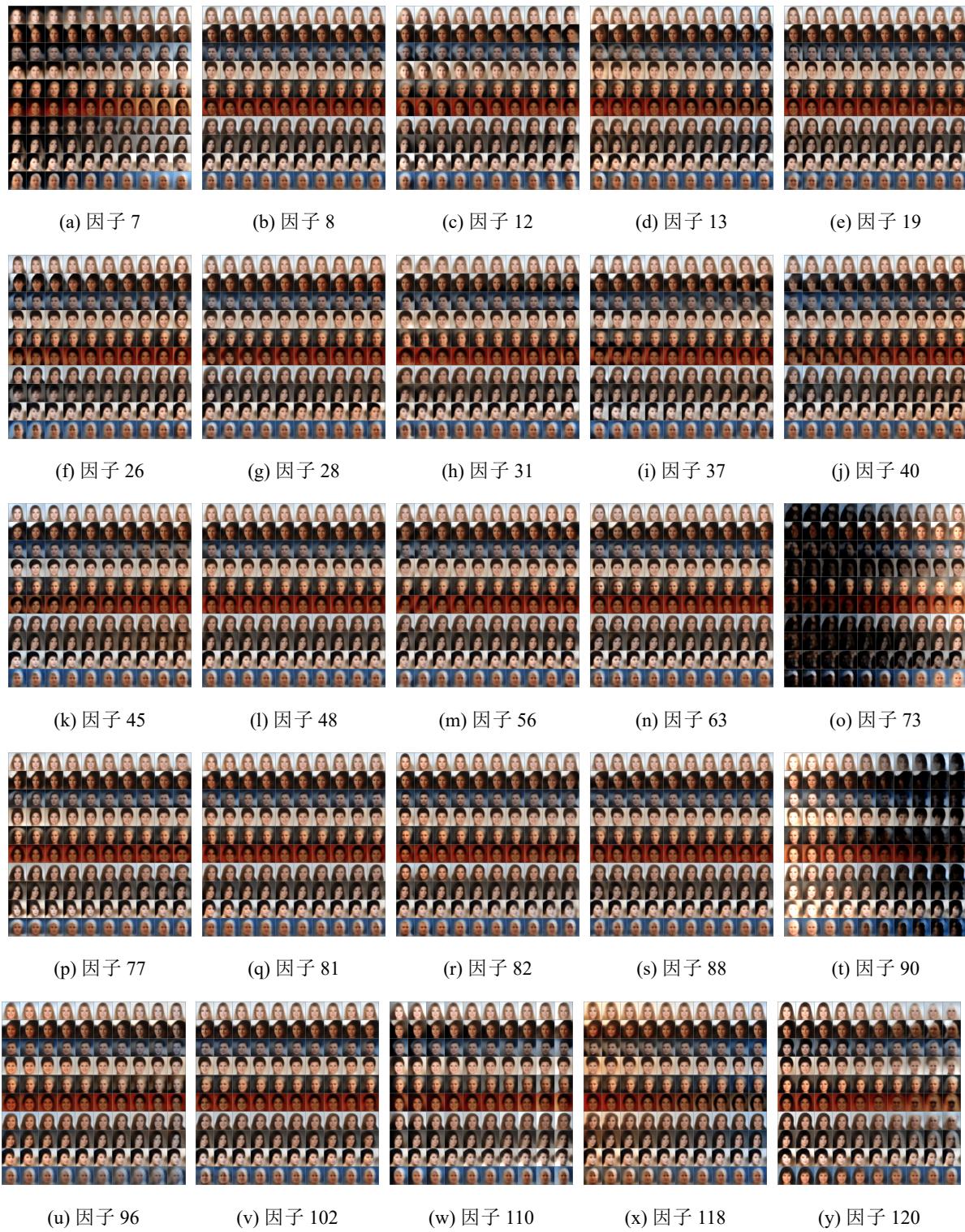
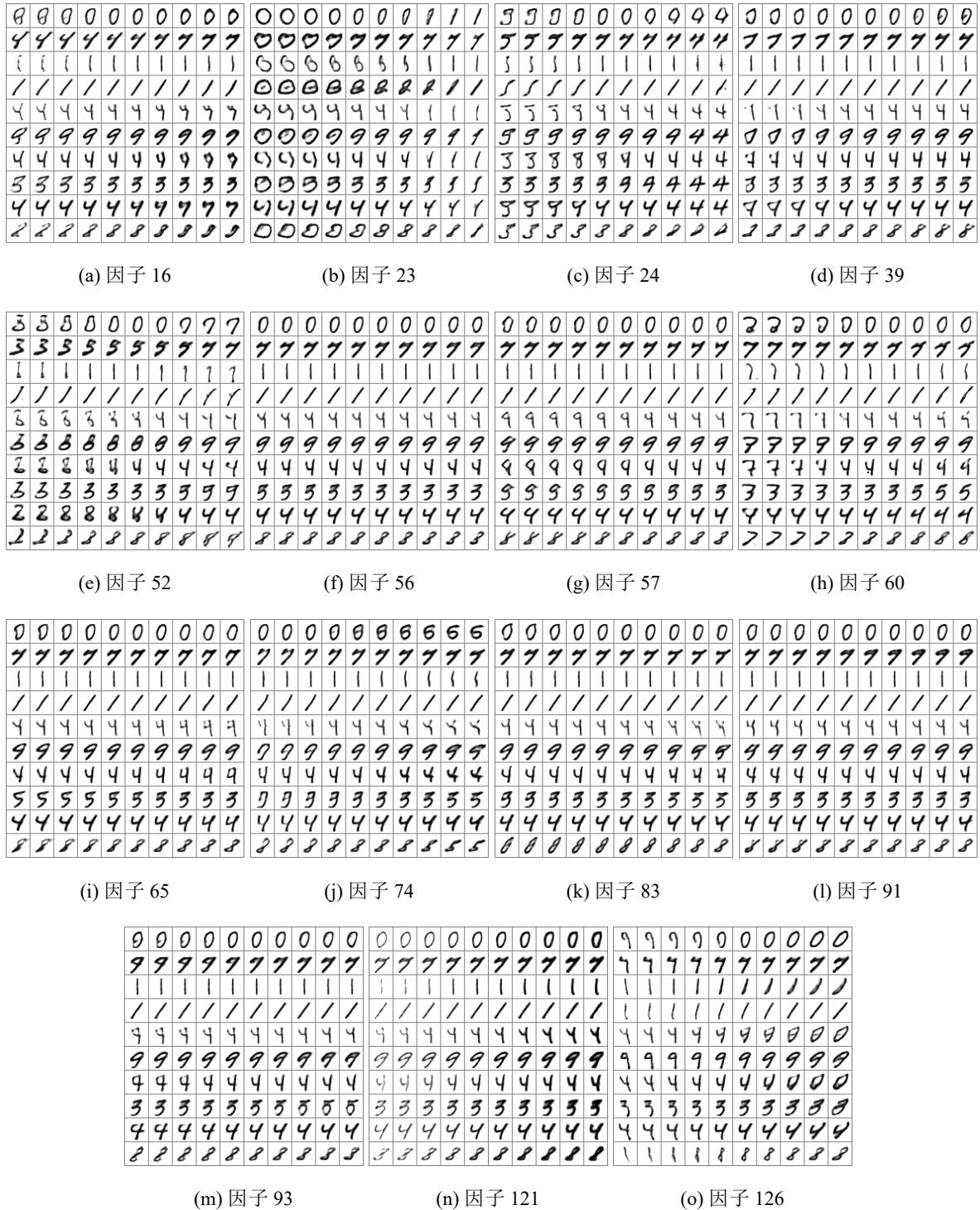


图 3-1 本人人脸作为输入的变化

从图3-2可以发现，背景明暗变化，脸朝向变化，头发方向的变化，头发颜色的变化，饱和度的变化，亮度的变化等等。

图 3-2 名人脸生成因子遍历 [64]。Generating Factors Traversal of $\beta(=40)$ -VAE.

从图3-3可以发现字体粗细的变化，字体倾斜的变化，字体上下的变化等等。

图 3-3 手写数字集生成因子遍历 [64]。Generating Factor Traversal of $\beta(=10)$ -VAE.

从图3-4可以发现不同脑区颜色的变化。

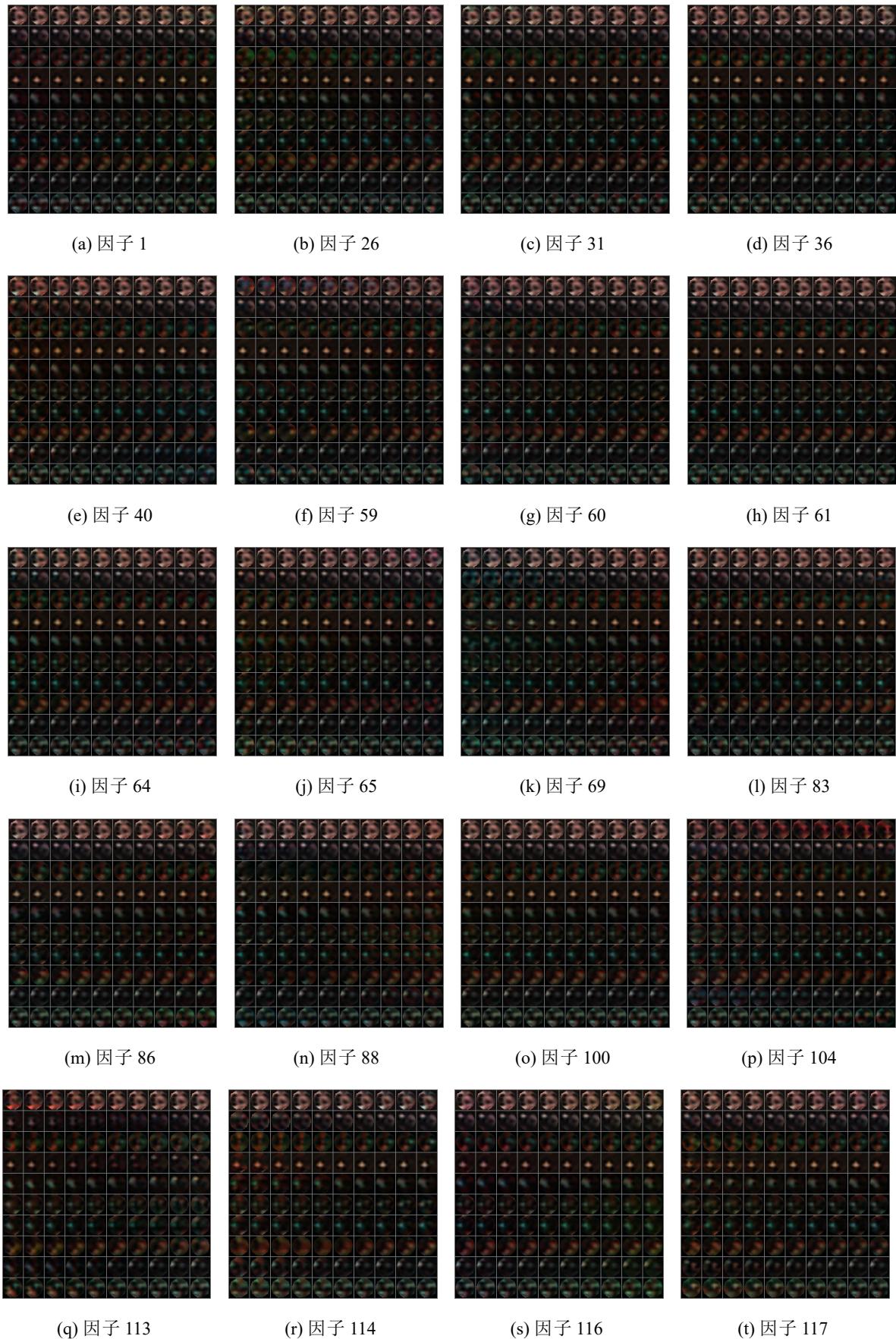


图 3-4 脑电图生成因子遍历 [64]。Generating Factor Traversal of $\beta(=6)$ -VAE.

3.4 变分自编码器发掘有影响的因子的能力

然而尽管变分自编码器已经学得了很多变化的因子，但是很多预先指定的因子是否被使用并不清楚。我们缺乏有效的方法来量化每个已知因子对数据表示的影响。在应用中，有时一些预设因子未被使用^①[6]。并且发现所学因素与原始数据之间的关系必须通过人工干预（视觉观察）。这会导致额外指定因子的浪费，并妨碍后续任务（如生成有意义的图像）的因素选择。此外，一些经典的影响因素确定方法，包括估计各因素的方差，都不具有对变分编码器的效用。因此，确定和监测各因子的影响力就成为这方面研究的一个关键问题。

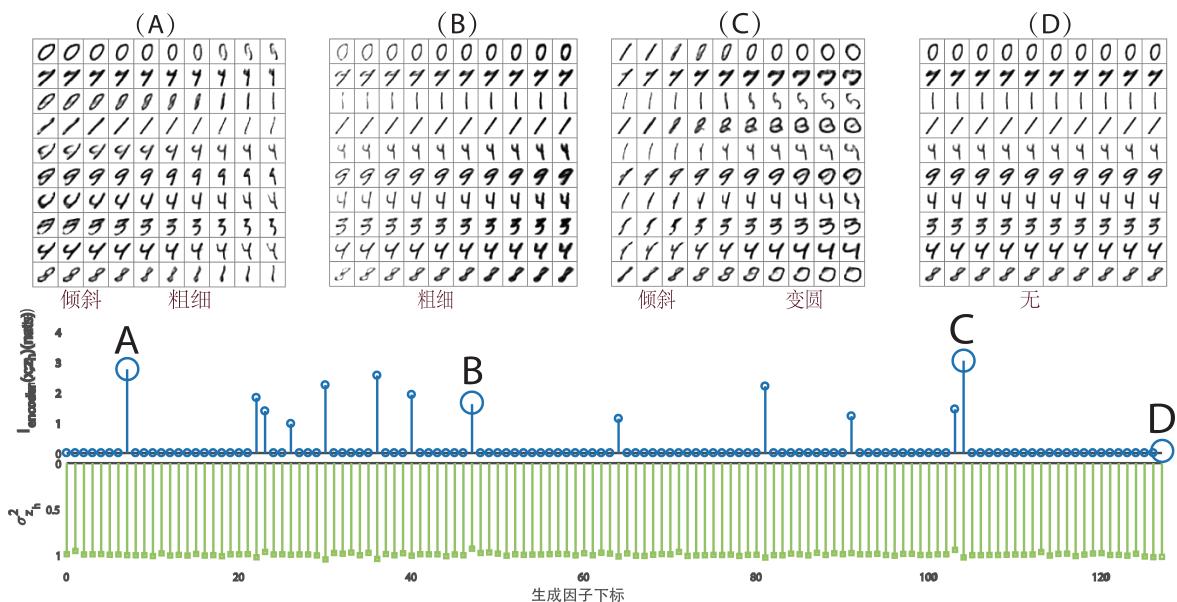


图 3-5 估计的 $I(\mathbf{x}; z_{\text{enc} h})$ 确定了有影响的生成因子 [64]; $I(\mathbf{x}; z_{\text{enc} h})$, $\sigma^2_{z_h}$ 和 MNIST 上定性的有影响生成因子遍历 ($\beta = 10$ -VAE)。向上的脉冲子图表示: 每个因子的 $I(\mathbf{x}; z_{\text{enc} h})$ 。向下的脉冲子图表示每个因子估计的 $\sigma^2_{z_h}$ 。A,B,C 蒙太奇表示: 因子 A,B,C 的遍历。所有有影响生力的成因子遍历见图 (3-3)。蒙太奇 D 是没有影响力的因子的遍历图 (它的互信息很小)。从 4 章蒙太奇可以看出方差并不能确定有影响的生成因子而互信息可以。

为了有效的确定和监督学到的因子，本文做出了如下努力。

- 我们首先采用互信作为衡定各个因子在变分自编码器模型中表示能力的定量指标。并且为了分析指标的合理性，我们从理论上证明了互信息影响变分编码器的重建误差下界以及后续分类任务。
- 我们提出了一个对于变分自编码器所有因子互信息的估计算法，并且证明了它的一致性。
- 我们利用指标在手写数字集 MNIST，名人脸数据集 CelebA 和脑电信号 DEAP 上进行的实验来支撑其有效性。特别是指标发掘了一些有意义且可以理解的因子，以及一些其他的可以忽略的因子。这些因子在泛化和分类任务上的能

^① 在图3-5中蒙太奇 (D) 是未使用因子的典型遍历。

力也得到了验证。

接下来文章结构如下。在章节3.4.1中我们断言互信息是一个必要的指标。具体我们介绍了输入数据与因子之间的互信息。然后我们从数据的互信息和数据的维度角度分析了原因。之后我们讨论了互信息和重建以及分类的联系。我们提出了衡量指标并证明其一致性。在章节3.4.2我们回顾了之前的工作。最后章节3.4.3为实验结果。

3.4.1 互信息作为一个指标的必要性

通过探究为什么因子会被忽略，我们论证互信息是一个发现有影响生成因子的必要指标。下面我们对于变分自编码器忽略因子的情况进行分析。

1) 数据本质低维的特性

变分自编码器的一个目标是学习数据的本质因子但是定理一指出本质因子的维数在连续可逆映射下保持不变。

定理 3.1 (信息守恒): 假设 $\mathbf{z} = (z_1, \dots, z_h)$ 和 $\mathbf{y} = (y_1, \dots, y_P)$ 分别是 H 和 P ($H \neq P$) 个单位独立高斯随机变量, 那么这两个集合的随机变量不能相互为生成因子。即这里不存在连续映射 $f : \mathbb{R}^H \rightarrow \mathbb{R}^P$ 和 $g : \mathbb{R}^P \rightarrow \mathbb{R}^H$ 使得

$$\mathbf{z} = g(\mathbf{y}) \quad \text{and} \quad \mathbf{y} = f(\mathbf{z}). \quad (3-7)$$

证明: 反证法。假设这两个函数存在, 那么其将会互为对方的逆映射, 并且是 \mathbb{R}^H 和 \mathbb{R}^P 的同胚映射。这里我们说 f 是一个同胚映射, 意味着 f 满足下述 3 个条件:

- f 是双射,
- f 连续,
- f 的逆函数也连续。

因为 \mathbb{R}^H 和 \mathbb{R}^P 有不同的同拓扑结构 ($P \neq H$), 同胚映射将不存在。

$$\mathbf{z} = g(\mathbf{y}) = g(f(\mathbf{z})) \quad \forall \mathbf{z} \in \mathbb{R}^H \quad \Rightarrow \quad g \circ f = I_H. \quad (3-8)$$

$$\mathbf{y} = f(\mathbf{z}) = f(g(\mathbf{y})) \quad \forall \mathbf{y} \in \mathbb{R}^P \quad \Rightarrow \quad f \circ g = I_P. \quad (3-9)$$

这表明 g 是 f 的逆函数, 并且 f 是双射。因为 g 和 f 都连续, f 是 \mathbb{R}^H 和 \mathbb{R}^P 之间同胚映射, 这导出了矛盾。 ■

假设理想数据, 记为 \mathbf{x} , 由 \mathbf{y} (包含 P 个单位独立高斯随机变量)生成, 即 $\mathbf{x} = \phi(\mathbf{y})$ 。因子 \mathbf{z} (包含 H 个单位独立高斯随机变量)由同胚映射 $\mathbf{x} = \psi(\mathbf{z})$ 生成了 \mathbf{x} 。那么 $\mathbf{y} = \phi^{-1} \circ \psi(\mathbf{z})$ 且 $\mathbf{z} = \psi^{-1} \circ \phi(\mathbf{y})$ 由于根据信息守恒定理, 必须有 $H=P$ 。这个定理说明, 例如有 10 个高斯因子和 128 个高斯因子那么其不能相互生成对方。类似的, 如果数据由 10 个本质的高斯因子生成, 它从直觉上也不会被变分自编码器推断成

128 个。就算我们预先设定了 128 个生成因子，一些因子可能也会与数据无关。而因子是否与数据有关可以通过互信息来进行判定。

2) 互信息可以反映绝对的统计依存关系

为了定量独立性并且估计哪个因子对于生成过程有影响，我们把 $I(\mathbf{x}; z_{\text{enc}})$ 作为一个合理的指标 [65]。即，

$$I(\mathbf{x}; z_{\text{enc}}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h | \mathbf{x}) || q_{\text{enc}}(z_h)). \quad (3-10)$$

互信息可以反映数据绝对的统计依赖性: $I(\mathbf{x}; z_{\text{enc}}) = 0$ 当且仅当 \mathbf{x} 和 z_{enc} 是独立的。 $I(\mathbf{x}; z_{\text{enc}})$ 越大，那么更多 \mathbf{x} 信息由 z_h 承载，它也更具有表示数据的影响力。

3) 互信息稀疏性

事实上互信息隐式地蕴含在变分自编码器的目标函数上。本文提出的下述定理更进一步的暗示了变分自编码器的目标函数导出了互信息的稀疏性。这也从互信息的角度说明了为什么因子会被忽略。

定理 3.2 (目标函数分解性): 如果 $q_{\text{enc}}(\mathbf{z} | \mathbf{x}) \ll p_{\text{dec}}(\mathbf{z})$ ^②，对于任意 \mathbf{x} ，
 $q_{\text{enc}}(\mathbf{z} | \mathbf{x}) = q_{\text{enc}}(z_1 | \mathbf{x}) \cdots q_{\text{enc}}(z_H | \mathbf{x})$ 和 $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z} | \mathbf{0}, I_H)$ 那么有如下分解：

- 变分自编码器 KL 散度的 L_1 范数表述：

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z} | \mathbf{x}) || p_{\text{dec}}(\mathbf{z})) \\ &= \sum_{h=1}^H \mathbb{E}_{x \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h | \mathbf{x}) || p_{\text{dec}}(z_h)) \\ &= \| \left(\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_1 | \mathbf{x}) || p_{\text{dec}}(z_1)), \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_2 | \mathbf{x}) || p_{\text{dec}}(z_2)), \dots, \right. \right. \\ & \quad \left. \left. \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h | \mathbf{x}) || p_{\text{dec}}(z_h)) \right) \|_1. \end{aligned} \quad (3-11)$$

- L_1 范数单项的进一步分解表达：

$$\begin{aligned} & \mathbb{E}_{x \sim p_{\text{data}}(x)} D_{KL}(q_{\text{enc}}(z_h | x) || p_{\text{dec}}(z_h)) \\ &= I(x; z_{\text{enc}}) + D_{KL}(q_{\text{enc}}(z_h) || p_{\text{dec}}(z_h)). \end{aligned} \quad (3-12)$$

□

证明： L_1 范数表达是显然的。我们证明 L_1 范数单项的进一步分解表达成立：

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h | \mathbf{x}) || p_{\text{dec}}(z_h)) &= \int q_{\text{enc}}(z_h | \mathbf{x}) p_{\text{data}}(\mathbf{x}) \frac{q_{\text{enc}}(z_h | \mathbf{x}) p_{\text{data}}(\mathbf{x})}{p_{\text{dec}}(z_h) p_{\text{data}}(\mathbf{x})} d\mathbf{x} \\ &= \int q_{\text{enc}}(z_h | \mathbf{x}) p_{\text{data}}(\mathbf{x}) \frac{q_{\text{dec}}(z_h | \mathbf{x}) p_{\text{data}}(\mathbf{x})}{q_{\text{enc}}(z_h) p_{\text{data}}(\mathbf{x})} \frac{q_{\text{enc}}(z_h)}{p_{\text{dec}}(z_h)} d\mathbf{x} \end{aligned}$$

^② 即 $q_{\text{enc}}(\mathbf{z} | \mathbf{x})$ 的支撑集被包含于 $p_{\text{dec}}(\mathbf{z})$ 的支撑集。

$$= I(\mathbf{x}; z_{\text{enc}h}) + D_{KL}(q_{\text{enc}}(z_h) || p_{\text{dec}}(z_h)). \quad (3-13)$$

■

这个定理呈现出在公式3-6变分下界的目标函数的第二项可以由 L_1 范数表出。它倾向于诱导 $I(\mathbf{x}; z_{\text{enc}h})$ 和 $D_{KL}(q_{\text{enc}}(z_h) || p_{\text{dec}}(z_h))$ 关于 h 的稀疏性，将非本质的因素维度压缩。 $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h | \mathbf{x}) || p_{\text{dec}}(z_h))$ 的稀疏性事实上诱导和式 $I(\mathbf{x}; z_{\text{enc}h})$ 与 $D_{KL}(q_{\text{enc}}(z_h) || p_{\text{dec}}(z_h))$ 关于 h 的稀疏性（因为两项都是非负的）。对于每一个 0 和式，其两个元素都为零。因此这个正则项倾向于本质的导出互信息 $I(\mathbf{x}; z_{\text{enc}h})$ 的稀疏性。它可以由图3-5和图3-6看出。因此变分自编码器的目标函数倾向于导出互信息的稀疏性，于是因子被忽略的现象出现。一方面，伴随着这个 KL 散度项，就算隐藏因子的数量被设定大了，过拟合的现象也不会发生。另一方面，这也帮助我们获得那些描述数据变化的因子，并且能够利用模型充分的泛化能力，变化一下因子来生成新的数据。

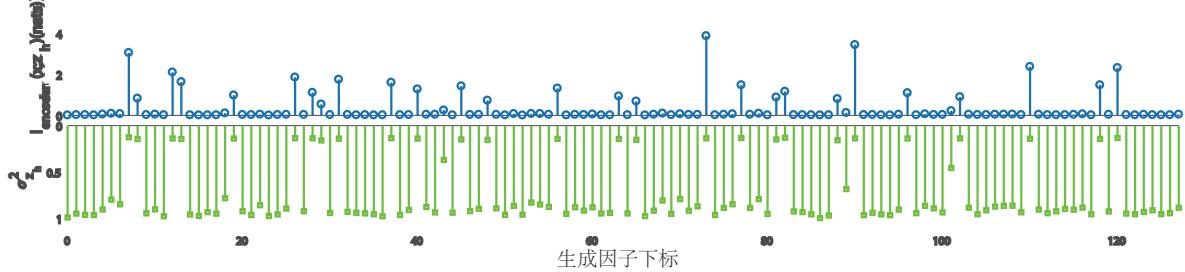
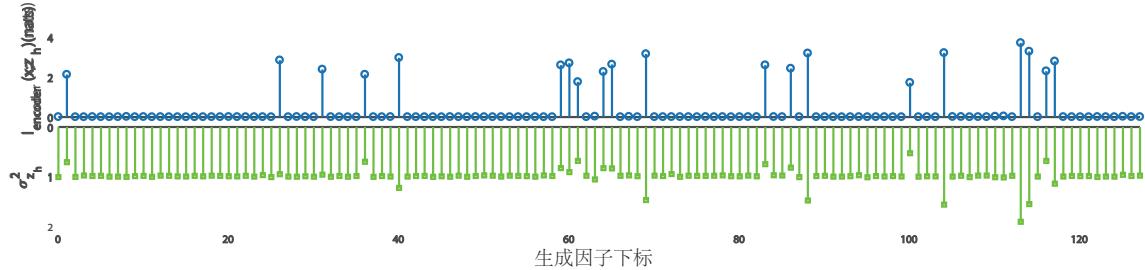
(a)CelebA 上的 $\beta (= 40)$ -VAE 的 $I(\mathbf{x}; z_{\text{enc}h}), \sigma_{z_h}^2$ 绘图。(b)DEAP 上的 $\beta (= 6)$ -VAE 的 $I(\mathbf{x}; z_{\text{enc}h}), \sigma_{z_h}^2$ 绘图。

图 3-6 互信息的稀疏性出现在 CelebA 和 DEAP 数据集上 [64]

顺便提一下，我们提出的下述的定理暗示了可以用 $I(\mathbf{x}; z_h)$ 去估计总的互信息。

定理 3.3 (互信息的分离性): 令 z_1, \dots, z_h 为独立单位高斯分布, z_1, z_2, \dots, z_h 已知 \mathbf{x} 时条件独立。那么

$$\begin{aligned} I(\mathbf{x}; z_1, \dots, z_h) &= \sum_{h=1}^H I(\mathbf{x}; z_h) \\ &= \|(I(\mathbf{x}; z_1), I(\mathbf{x}; z_2), \dots, I(\mathbf{x}; z_h))\|_1. \end{aligned} \quad (3-14)$$

□

证明：

$$\begin{aligned}
 I(\mathbf{x}; z_1, \dots, z_H) &= \int p(z_1, \dots, z_H, \mathbf{x}) \log \frac{p(\mathbf{x}, z_1, \dots, z_H)}{p(z_1, \dots, z_H)p(\mathbf{x})} dz_1 \cdots dz_H d\mathbf{x} \\
 &= \int p(\mathbf{x}, z_1, \dots, z_H) \log \frac{\prod_{h=1}^H p(z_h | \mathbf{x})}{\prod_{h=1}^H p(z_h)} dz_1 \cdots dz_H d\mathbf{x} = \sum_{h=1}^H \int p(\mathbf{x}, z_h) \log \frac{p(z_h | \mathbf{x})}{p(z_h)} dz_h d\mathbf{x} \\
 &= \sum_{h=1}^H I(\mathbf{x}; z_h).
 \end{aligned} \tag{3-15}$$

定理证明完成。 ■

这个定理暗示如果 $q_{\text{enc}}(\mathbf{z})$ 可以分解并且 $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$ 可以分解，那么我们可以使用 $I(\mathbf{x}; z_{\text{enc}_h})$ 去直接估计全部的互信息。

4) 重建和分类理论支撑

根据 [66] 中的定理 8.6.6, 互信息可以提供一个最优重建误差的下界。定理描述如下：

定理 3.4: 假设 \mathbf{x} 有微分熵 $H(\mathbf{x})$, 那么让 $\hat{\mathbf{x}}(\mathbf{z}_{\text{enc}})$ 成为 \mathbf{x} 的估计, 并且给定辅助信息 \mathbf{z}_{enc} , 那么有

$$\mathbb{E}(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z}_{\text{enc}}))^2 \geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}))}. \tag{3-16}$$

因此如果我们设 $p_{\text{dec}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\text{dec}_\mu(\mathbf{z}), \text{dec}_\sigma)$, 那么 $\mathbf{x}_{\text{rec}} = \text{dec}(\mathbf{z}_{\text{enc}})$ 有 $\frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}))}$ 作为重建的下界。令 $\mathbf{z}_{\text{enc}} = [\mathbf{z}_{\text{major}}, \mathbf{z}_{\text{minor}}]$, 让我们只用因子当中 major 集合 $\mathbf{z}_{\text{major}}$ 用于解码: $\mathbf{x}_{\text{rec}} = \text{dec}_\mu(\mathbf{z}_{\text{major}}, \mathbf{0})$ 。结合了 $q_{\text{enc}}(\mathbf{z})$ 可以分解的假设, 那么有互信息的分离结果 $I(\mathbf{x}; \mathbf{z}_{\text{enc}}) = I(\mathbf{x}; \mathbf{z}_{\text{minor}}) + I(\mathbf{x}; \mathbf{z}_{\text{major}})$ 。于是有下述界,

$$\begin{aligned}
 &\mathbb{E}(\mathbf{x} - \mathbf{x}_{\text{rec}})^2 \\
 &\geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{major}}))} \\
 &\geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{major}}))} e^{-2I(\mathbf{x}; \mathbf{z}_{\text{minor}})}.
 \end{aligned} \tag{3-17}$$

这个定理意味着所选的因子蕴含的互信息直接影响最优重建的下界, 并且我们可以选择一些具有高影响力的因子, 以使得在更小重建误差的情况下表示和生成数据。

我们进一步提供一些互信息作为分类指标的理论支撑。假设马尔可夫条件 $\mathbf{y} \rightarrow \mathbf{x} \rightarrow \mathbf{z}_{\text{enc}} \rightarrow \mathbf{y}_{\text{pre}}$ 成立, 依据法诺 (Fano's) 不等式 [66] 和信息处理不等式 [66], 互信息也和分类误差有关联。我们提出了下述推广了的法诺 (Fano's) 不等式定理。

定理 3.5 (法诺 (Fano's) 不等式): 对于任意估计 $\hat{\mathbf{y}}$ 使得 $\mathbf{y} \rightarrow \mathbf{x} \rightarrow \mathbf{z}_{\text{enc}} \rightarrow \mathbf{y}_{\text{pre}}$, $P_e = Pr(\hat{\mathbf{y}} \neq \mathbf{y})$, 那么我们有

$$H(P_e) + P_e \log |\mathcal{Y}| \geq H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) \geq H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}). \quad (3-18)$$

那么这个不等式可以被弱化:

$$1 + P_e \log |\mathcal{Y}| \geq H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) \geq H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}), \quad (3-19)$$

或

$$P_e \geq \frac{H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) - 1}{\log |\mathcal{Y}|} \geq \frac{H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}) - 1}{\log |\mathcal{Y}|}. \quad (3-20)$$

注意到由于信息处理不等式 $I(\mathbf{x}; z_{\text{enc}}) \geq I(\mathbf{y}; z_{\text{enc}})$, 如果 $I(\mathbf{x}; z_{\text{enc}}) = 0 \Rightarrow I(\mathbf{y}; z_{\text{enc}}) = 0$, 那么 h 将不会影响预测。结合假设 $q_{\text{enc}}(\mathbf{z})$ 可以分解, 定理显示由所选的因子所拥有的互信息直接影响分类误差的下界。因此我们可以依据互信息移除一些小影响力的因子而不会显著抬升预测误差的下界。

5) 定性计算指标的算法

为了在实践中去计算 $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$, 我们假设 $q^*(\mathbf{z}) = \mathcal{N}(\mathbf{z} | \mathbf{0}, \text{diag}(\sigma_1^*, \dots, \sigma_H^*))$ 是一个可以对于 $q_{\text{enc}}(\mathbf{z})$ 进行估计的 0 均值高斯分布。于是, 我们可以将需要被估计的指标列出

定义 3.1 ($I(\mathbf{x}; \mathbf{z}_{\text{enc}})$ 的估计: 由全部因子蕴含的互信息):

$$I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z} | \mathbf{x}^m) || q^*(\mathbf{z})). \quad (3-21)$$

这个估计利用了 3.2 推论经验形式的 M 个采样形式。

定义 3.2 ($I(\mathbf{x}; z_{\text{enc}})$ 的估计: 由一个因子蕴含的互信息):

$$I_{\text{est}}(\mathbf{x}; z_{\text{enc}})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(z_{\text{enc}} | \mathbf{x}^m) || q^*(z_{\text{enc}})). \quad (3-22)$$

这个指标定量了一个因子和输入的互信息。

注意到以上指标需要 $q^*(\mathbf{z})$ 的值, 所以我们需要去设计算法计算这项。由定理 3.2 通过最小化等式, 我们知道

$$\begin{aligned} & \min_q \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z} | \mathbf{x}) || q(\mathbf{z})) \\ & \Leftrightarrow \min_q D_{KL}(q_{\text{enc}}(\mathbf{z}) || q(\mathbf{z})) d\mathbf{z}, \end{aligned} \quad (3-23)$$

于是我们可以证明下述结果:

引理 3.1 如果 $q_{\text{enc}}(\mathbf{z} | \mathbf{x}) \ll q^*(\mathbf{z})$ 那么有,

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z} | \mathbf{x}) || q^*(\mathbf{z})) \\ & = I(\mathbf{x}; \mathbf{z}_{\text{enc}}) + D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z})). \end{aligned} \quad (3-24)$$

引理3.1的证明和定理3.2的相似。这个引理显示定义3.1提供了对于编码器信道容量的另一个上界。从经验上来讲，这个估计比目标函数的第二项来得紧很多。 $q^*(z)$ 可以接着由求解下述问题得到，

$$q^*(\mathbf{z}) = \arg \min_q \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q(\mathbf{z})). \quad (3-25)$$

上述最小化问题可以有闭式解^③：

$$\sigma_i^* = \frac{\sum_{m=1}^M \sigma_i(\mathbf{x}^m) + \mu_i^2(\mathbf{x}^m)}{M}. \quad (3-26)$$

下述为本文给出的证明：

注意到假设我们有两个多元正态分布，伴随着均值 μ_0, μ_1 和非奇异的协方差矩阵 Σ_0, Σ_1 ，并且两个分布的维度有同样的维度 H ，那么有 [67]，

$$D_{KL}(\mathcal{N}_0 || \mathcal{N}_1) = \frac{1}{2} (tr(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - H + \ln(\frac{\det \Sigma_1}{\det \Sigma_0})). \quad (3-27)$$

注意到我们已经假设了 $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \text{diag}(\sigma_1, \dots, \sigma_H))$ 。

并且 $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_H(\mathbf{x})))$ 。因此有，

$$\begin{aligned} & \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q(\mathbf{z})) \\ &= \frac{1}{M} \sum_{m=1}^M \frac{1}{2} \sum_{i=1}^H \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)} \\ &= \frac{1}{2M} \sum_{i=1}^H \sum_{m=1}^M \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)}. \end{aligned} \quad (3-28)$$

可以将以上的优化问题分解为以下 H 个子问题：

$$\sigma_i^* = \arg \min_{\sigma_i} \sum_{m=1}^M \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)}, i = 1, \dots, H. \quad (3-29)$$

$$\nabla_{\sigma_i} \left(\sum_{m=1}^M \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)} \right) = \sum_{m=1}^M -\frac{\sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{\sigma_i^2} + \frac{1}{\sigma_i}. \quad (3-30)$$

因为

$$\sum_{m=1}^M -\frac{\sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{\sigma_i^{*2}} + \frac{1}{\sigma_i^*} = 0, \quad (3-31)$$

所以有

$$\sigma_i^* = \frac{\sum_{m=1}^M \sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{M}. \quad (3-32)$$

上述过程被总结并由下述算法呈现以便于计算提出的指标。

^③ 当然也可以直接用梯度下降法求解。

- 1: **输入:** 采样的数据 $\{\mathbf{x}^m\}_{m=1}^M$,
编码网络 $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_H(\mathbf{x})))$.
- 2: **得到:** $q^*(\mathbf{z}) = \arg \min_q \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q^*(\mathbf{z}))$.
- 3: **for** $i = h$ **to** H **do**
- 4: $\sigma_i^* = \frac{\sum_{m=1}^M \sigma_i(\mathbf{x}^m) + \mu_i^2(\mathbf{x}^m)}{M}$.
- 5: **end for**
- 6: **计算:** $I_{\text{est}}(\mathbf{x}; z_{\text{enc}})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(z_{\text{enc}}|\mathbf{x}^m) || q^*(z_{\text{enc}}))$.
- 7: **for** $i = h$ **to** H **do**
- 8: $I_{\text{est}}(\mathbf{x}; z_{\text{enc}})_M = \frac{1}{M} \sum_{m=1}^M \log \frac{\sigma_h^*}{\sigma_h(\mathbf{x}^m)} + \frac{\sigma_h^2(\mathbf{x}^m) + \mu_h^2(\mathbf{x}^m)}{2\sigma_h^{*2}}$.
- 9: **end for**
- 10: **计算:** $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$.
- 11: $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M = \sum_{h=1}^H I_{\text{est}}(\mathbf{x}; z_{\text{enc}})_M$.
- 12: **输出:** $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M, I_{\text{est}}(\mathbf{x}; z_{\text{enc}})_M, q^*(\mathbf{z})$.

算法 3-1 互信息估计

本文提出下述定义和定理阐明了指标对于互信息估计的一致性。

定义 3.3 (一致性): 估计值 $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$ 是关于 $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$ 一致的当且仅当: $\forall \varepsilon > 0 \ \forall \delta > 0, \exists N$, 并且存在 $q^*(\mathbf{z})$, $\forall M > N$, 以概率大于 $1 - \delta$, 有

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| < \varepsilon. \quad (3-33)$$

定理 3.6: 当 $D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z}))$ 足够小, 同时 $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q^*(\mathbf{z}))$ 存在, 那么估计值 $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$ 是关于 $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$ 一致的。即, 若 $q^*(\mathbf{z})$ 的选择满足条件 $D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z})) < \varepsilon/2$, 同时 $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q^*(\mathbf{z}))$ 存在, 那么 $\forall \delta > 0, \exists N, \forall M > N$, 以概率大于 $1 - \delta$, 我们有

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| < \varepsilon. \quad (3-34)$$

证明: 令 $\tilde{I}[q^*] = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q^*(\mathbf{z}))$. 由于
 $\{u^m = D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q^*(\mathbf{z})), m \geq 1\}$, $(3-35)$

为独立同分布的随机变量序列, 且 u^m 的期望为 $\tilde{I}[q^*]$ (存在), 根据辛钦大数定律, 我们有 $\forall \delta > 0, \exists N, \forall M > N$, 伴随着概率大于 $1 - \delta$, 我们有

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - \tilde{I}[q^*]| < \varepsilon/2, \quad (3-36)$$

$$\begin{aligned} & |I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| \\ & \leq |I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - \tilde{I}[q^*]| + |\tilde{I}[q^*] - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| \end{aligned}$$

$$< \frac{\varepsilon}{2} + |D_{KL}(q_{\text{enc}}(\mathbf{z})||q^*(\mathbf{z}))| < \varepsilon. \quad (3-37)$$

这个定理显示只要 $q^*(\mathbf{z})$ 能够任意接近 $q_{\text{enc}}(\mathbf{z})$, 并且采样数量足够高, 那么估计值以很高的概率可以任意逼近真的互信息。并且最小化 $D_{KL}(q_{\text{enc}}(\mathbf{z})||q^*(\mathbf{z}))$ 启发了 $q^*(\mathbf{z})$ 的导出.

3.4.2 相关工作

目前并没有太多用设计的指标去衡量变分自编码器有影响的因子的工作。一个普遍并且简单的确定变分自编码器因子影响力的方法是通过直观的视觉 [6], [63] 观察。然而这可能消耗大量劳动力, 来选择接下来任务所需的因子。

在 [68] 工作中, 人们通过画 $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$ 95% 置信度椭球间隔的可视化, 来监督网络的行为。与此同时它也直接反应因子的影响力。然而这种方法还是需要人工去看绘制的图。

在经典的 PCA 中, 人们普遍选择那些有较大方差的因子并且 [52] 暗示因子的方差可能可以指示因子的用途。然而, 方差并不一定可以总是表示因子和数据之间的绝对的统计关系, 这由图3-5 和图3-6中可以发现。

我们的工作更强调互信息, 它传递了数据和因子的绝对统计关系。并且我们利用其为指标去发现有影响的因子。所选因子互信息量与重建以及分类的关系, 支撑了我们对于互信息指标的选择。我们的实验表明设计的指标可以发掘数据表示的有影响的因子。

3.4.3 实验结果

1) 数据集

MNIST 是手写数字数据集 [69]。我们估计了所有上述所学得因子的互信息, 然后利用不同比例的最具影响的因子用于后续生成任务。我们将 70,000 个数据点依据比例 $[0.6 : 0.2 : 0.2]$ 划分成训练, 交叉检验和测试集。估计的互信息和 $q^*(\mathbf{z})$ 通过在 10,000 个测试数据点上计算得到。种子图片选取自测试集。其被用于推断因子值和生成因子遍历。

CelebA[70] 是一个大规模的名人脸及属性的数据集, 我们仅用其图片来做生成因子发掘。我们将 200,000 个数据点依据比例 $[0.8 : 0.1 : 0.1]$ 划分成训练, 交叉检验和测试集。估计的互信息和 $q^*(\mathbf{z})$ 通过在 10,000 个测试数据点上计算得到。种子图片选取自测试集。其被用于推断因子值和生成因子遍历。

DEAP 是由 [71] 提出的一个公开、著名的多模态情感识别数据集。EEG 信号有 32 个通道, 由 32 参与者观看 40 个 63 秒视频记录得到。EEG 数据已经被预处理降采样到 128Hz 以及频段 4-45Hz。通过和 [23] 相同的转换方法, 我们应用快速傅里叶变换于 1 秒内 EEG 信号并将其转化成了图片。这个实验中 alpha (8-13Hz),

beta (13-30Hz) 和 gamma (30-45Hz) 频段用来表示大脑情感出现的活动。接下来利用 Azimuthal Equidistant Projection (AEP) 和 Clough-Tocher scheme 将脑电信号等距映射成 32x32RGB 图像。两个情绪中的维度，清醒和价，被标记为 1-9。我们用 5 作为边界，区分高低两个维度，去生成 4 类别。我们利用转化成的视频序列进行 4 类别情感预测。其中我们选取了有影响力的生成因子，同时这个实验有助于情感相关的因子的导出。

在遍历图片中每一个块对应于在保持其他因子不变情况下单一的因子从 [-3, 3] 范围的变化。每一个行由一个不同的种子图片生成。

2) 网络结构

在三个数据集上的编码网络均采用了卷积神经网络的架构，对于 CelebA 和 DEAP 数据集卷积网络更深，卷积核也更多一些。而解码网络均采用了反卷积网络的架构，并与编码网络对称。优化过程中都采取 Adam 优化算子。隐藏因子数量都设置成了 128。具体结构见表3-1。

表 3-1 网络结构 [64]

数据集	优化子	架构
MNIST	Adam	输入维度 28x28x1
	学习率 $1e - 3$	编码器结构 Conv 32x4x4,32x4x4 (间隔 2). FC 256. ReLU 激活.
	训练周期 200	隐藏因子个数 128
		解码器结构 FC 256. 线性层. 编码器的反卷积逆结构. ReLU 激活. 高斯分布.
CelebA	Adam	输入维度 64x64x3
	学习率 $1e - 4$	编码器结构 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (间隔 2). FC 256. ReLU 激活.
	训练周期 20	隐藏因子个数 128/32
		解码器结构 FC 256. 线性层. 编码器的反卷积逆结构. ReLU 激活. 混合 2-高斯分布.
DEAP	Adam	输入维度 32x32x3
	学习率 $1e - 4$	编码器结构 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (间隔 2). FC 256. ReLU 激活.
	训练周期 300	隐藏因子个数 128/32
		解码器结构 FC 256. 线性层. 编码器的反卷积逆结构. ReLU 激活. 高斯分布.
		LSTM 输入维度 63x128
		循环网络结构 LSTM dim128. 时间间隔 63.
		预测器 FC 4. ReLU 激活.

3) 有影响力生成因子发掘实验

根据图3-5提出的互信息指标有效确定了有影响和没有影响的因子。那些估计互信息小的生成因子没有什么生成作用，而有大互信息值的因子有巨大的生成影响。由此对比，可以观察到传统方法中使用的方差并不能显著指示因子是否被使用。

为了去支撑互信息估计算子的有效性，我们用指标去自动的选择互信息大于0.5的CelebA中的有影响的因子，并挑选了其中3个由图3-7呈现。其中很多因子拥有一些可以理解的变化例如背景颜色，微笑，脸的角度等等。这证明了互信息是一个在变分自编码器设定中自动确定有影响力生成因子的有效指标。

4) 发掘因子的生成能力实验

估计的互信息可以指导利用少数但是有影响力的因子来进行后续的生成任务。我们依据定量的互信息，选取了不同比例的最具有影响力的因子来生成后续图片。因子按照它的互信息值排序并且将其他估计出的无影响力因子被一致的设成0。根据图3-8，我们可以发现通过用前10%的被算法发掘的最具有影响力因子，变分自编码器模型仍然可以生成极为相似于利用所有因子的重建。表3-2呈现了具体的全部信息量和不同比例生成因子的重建误差。最高的10%的因子包含了几乎所有信息，因此他们的重建与使用全部因子有着几乎相同的效果。由信息和重建关联所显示的那样，使用的因子所包含的信息越少，那么最小重建误差的下界被抬得越高。

表 3-2 互信息和重建误差表 [64]

最具互信息因子使用比例 (%)	100	20	10	7	5	3	1	0
$I(\mathbf{x}; \mathbf{z}_{\text{enc}_{\text{used}}})$	24.3	24.3	24.3	19.6	16.5	10.6	5.8	0
平均平方误差	5.6	5.6	5.6	13.4	15.0	27.6	44.4	71.7

5) 发掘因子的分类能力实验

估计的互信息可以进一步利用很少，但是有影响力的因子来指导后续的分类任务。我们依据估计的互信息选择不同比例的最具影响力的因子来去预测情绪。因子按照其互信息大小排序，在后续分类任务中，没有影响力的生成因子被一致的设成0。

表 3-3 互信息和 EEG 情感分类利用 $\beta(= 6)$ -VAE[64]

最具互信息因子使用比例 (%)	100	50	10	7	5	3	1	0
$I(\mathbf{x}; \mathbf{z}_{\text{enc}_{\text{used}}})$	53.8	53.5	38.3	28.0	22.5	13.5	7.0	0
平均测试精度	0.53	0.52	0.46	0.32	0.34	0.29	0.3	0.23

依据表3-3，只利用一半的因子，模型依旧能够拥有相似的预测精度。

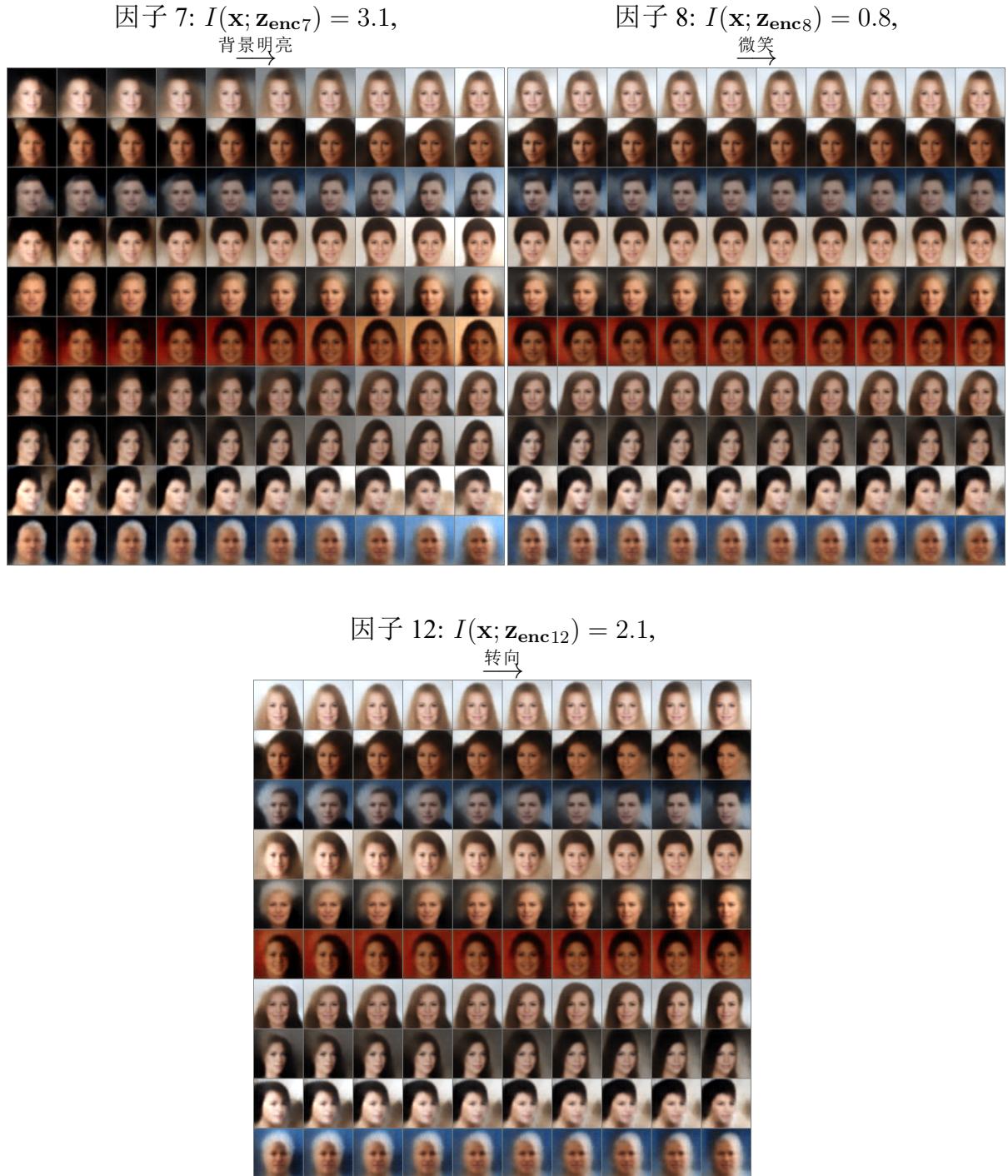


图 3-7 CelebA: $\beta(=40)$ -VAE 生成因子遍历 [64]。我们呈现了由互信息指标确定的前三有影响力的生成来展示。

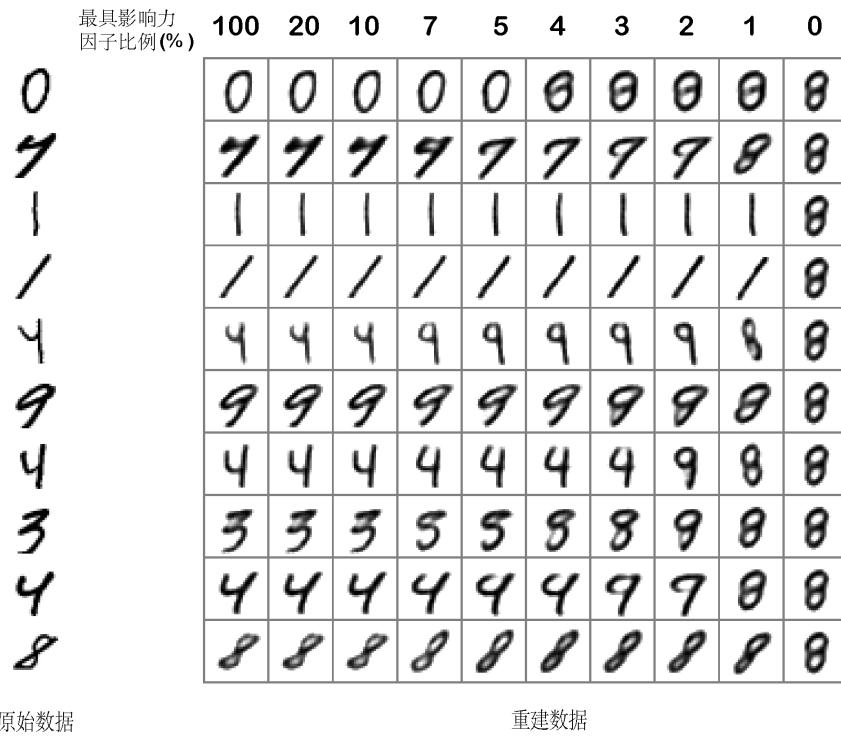


图 3-8 不同比例因子的生成图 [64]

3.5 因子的等价特性

在利用自编码器学习图像表示时，往往会出现一个因子包含多种变化的情况。究竟因子可否一对一的学习数据中的变化呢？答案是因子可以学得的情况在一个与数据相关的正交变换构成等价类当中。我们提出了下述两个定理。

定理 3.7 (高斯因子等价性): 假设 $\mathbf{z} = (z_1, \dots, z_H)$ 是 H 个独立单位高斯随机变量。设 $Q \in \mathbb{R}^{H \times H}$ 为一个正交矩阵，那么 $\mathbf{y} = Q\mathbf{z}$ 也是由 H 独立单位高斯随机变量构成的。并且 \mathbf{z} 和 \mathbf{y} 可以仅需通过一个线性同胚映射相互生成。 \square

证明: 我们只需检验 \mathbf{y} 的均值和方差:

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(Q\mathbf{z}) = Q\mathbb{E}(\mathbf{z}) = 0, \quad (3-38)$$

$$Cov(\mathbf{y}, \mathbf{y}) = QCov(\mathbf{z}, \mathbf{z})Q^{tr} = QIQ^{tr} = I. \quad (3-39)$$

因此, \mathbf{y} 是又一个 H 个独立单位高斯随机变量。因为 $\mathbf{z} = Q^{tr}\mathbf{y}$, \mathbf{z} 和 \mathbf{y} 通过一个线性同胚映射相互生成。 \blacksquare

这个定理说明了这里有一类单位高斯随机变量他们可以相互生成他们的信息守恒。

定理 3.8 (线性高斯因子等价类):

$$[\mathbf{z}] = \{\mathbf{y} | \mathbf{y} = Q\mathbf{z}, \quad Q \in \mathbb{R}^{H \times H} \text{ 为正交变换}\}. \quad (3-40)$$

那么 $\forall \mathbf{y} \in [\mathbf{z}]$, \mathbf{y} 是 H 独立单位高斯随机变量并且通过一个线性变换可以生成 \mathbf{z} 。 \square

这个定理阐明了如果变分自编码器因子有一个线性矩阵乘法的自由度, 那么在等价类中的因子就都可能被学得。

假设一组视觉变化可以被看成是独立高斯变量 (

$$\mathbf{z} = (z_{rotation}, z_{gender}, \dots, z_{with\ glass})^T \quad (3-41)$$

但是学到的却是 $\mathbf{y} = Q\mathbf{z}$ 。

$$y_1 = q_{11}z_{rotation} + \dots + q_{1H}z_{with\ glass}. \quad (3-42)$$

有 $Q^{tr}\mathbf{y} = \mathbf{z}$ 。于是有

$$z_{rotation} = q_{11}y_1 + \dots + q_{H1}y_H, \quad (3-43)$$

\dots ,

$$z_{with\ glass} = q_{1H}y_1 + \dots + q_{HH}y_H. \quad (3-44)$$

那么变化因子 y_1 , 依据 q_{11} 到 q_{1H} 的大小, 视觉变化 $z_{rotation}, z_{gender}, \dots, z_{with\ glass}$ 将以不同程度体现。

3.6 总结

本章节解释了用输入数据和因子的互信息作为指标, 来估计变分自编码器中因子表示图像数据的本质影响。互信息能够反映随机变量间的绝对统计依存关系。变分自编码器的目标函数的第二项和过多预设因子数量, 倾向于诱导互信息的稀疏性。并且帮助诱导变分自编码器的有影响力和没有影响力的因子的出现。我们也证明了互信息涉及重建均值误差的下界和分类的预测误差。我们设立可行的算法去计算指标, 用以估计变分自编码器所有因子的互信息, 并且证明了指标的一致性。实验显示有影响的因子和没有影响的因子, 都能够被自动有效的发现, 有影响力的生成因子一般具有可解释性。例如在手写数字集合 MNIST 数据集上, 有影响力的生成因子对应了数字粗细的变化, 数字倾斜的变化、数字变圆的变化等等。在名人脸数据集 CelebA 上, 有影响力的生成因子, 对应了人脸的肤色从白到黄的变化, 人脸的性别从男到女的变化、人脸的角度的变化、人脸的背景饱和度的变化, 人脸的头发形态的变化, 人脸的背景的颜色变化等等。在脑电数据集合 DEAP 上, 有影响力的生成因子, 对应了不同测量脑区的电信号频率变化和不同测量脑区的电信号能量谱强度变化等等。发掘的因子的可理解性被有效支撑, 并且它们的泛化以及分类能力也被验证。除此之外, 一些和分类相关的变化也得以发现。这实验也

激发了我们利用少量但是有影响力的因子，来进行后续的数据处理任务（例如生成和分类任务）并实现与全部因子利用时有相似的性能，就像 PCA 和 ICA 等的降维能力那样。最后，我们从理论上解释了因子具有等价特性，这说明了变分自编码器很难一对一的学习到数据当中的变化。

4 视频数据的深度表示学习

视频数据是以图像 \mathbf{x}^i 为时间序列元素的数列 $\{\mathbf{x}^i\}_{i=1}^T$ 。我们主要考虑的任务是利用其进行类别判别。由于视频数据过于庞大，故时空上降维压缩的方法对其进行处理是高效的。[72]一文当中比较了卷积神经网络配合长短记忆网络模型的方法，取得了不错的效果。由于变分自编码器更注重学习空间上的特征表示，这使得处理图像的变分自编码器模型和处理时序数据的长短记忆网络模型结合可能是视频数据很好的切入模型。

在前面的章节粗略的提到了变分自编码器与长短记忆网络模型用于脑电数据情感识别这一应用，在这一章节将会详细介绍采用的模型和对比结果。

4.1 脑电情感识别的意义

随着满足于人类生活需要的实际需求的不断发展，对于人类大脑的不断研究在近几十年得到了全面的改善。这些年脑机接口（Brain Computer Interface）和人机交互（Human Computer Interface）快速发展，脑电信号产生了广泛的应用前景。这也使得脑电信号可视化生成变得十分重要。

情绪产生于人脑，用于保障人类的日常生存和环境适应。其进一步影响人类日常的决策，工作和学习生活。例如悲伤的时候人们容易观察到一些细节的变化而相反开心的时候人们有时会忽略细节。人从平静状态到激动状态会产生脸红、肌肉不听使唤的收缩状态等。这使得情绪识别成为一项有助于电脑或者人类了解他人的生理和心理状态的重要手段。在医疗领域，情感识别有助于监护重症病人以便及时建立诊疗方案。安全监测领域，监控驾驶员的情绪，判断驾驶员是否疲劳驾驶，对于驾驶员调整情绪，放松心情，安全驾驶有着很好的助益。在军事领域，监控士兵的情绪有助于指挥员及时调整作战策略。在教育领域可以帮助辅助判断学生的学习状态，有助于教师进行针对性的指导并改善教学质量等。此外对于一般的人来说，人们也会遭受一些消极的情绪例如抑郁压抑等，能够有效的识别出对象的情绪并帮助其调节能够很好的改善其生活质量。但是通过对于情绪检测和填写一些量表极其容易受到主观因素的影响，而从外在地对于其外表例如表情、语音、手势情感的观察容易受到伪装（例如非常专业的电视剧电影演员能够表演出不同的情绪状态）和不同文明文化的影响。这使得采取具有固有性和不可伪装性的脑电信号来进行客观的情感识别成为不二之选。

现存的情感识别方式时常采取了传统的机器学习的方法或者信号处理的方法，这使得模型不能够很好的学习并处理具有复杂神经结构产生的脑电信号来检测对应的情感。同时传统的模型也不能够具有可视化生成信号的能力。这都是我们的模型所具有的。

4.2 脑电信号的简介

大脑的多个脑区的神经元和突触之间，在不同情感状态下的电位活动具有明显的差异。其通过大脑皮层传送到头皮处，被固定在头皮处的电极记录下来被称为脑电信号（Electroencephalography, EEG）。脑电信号在不同情感状态下会发生相应的变化，这种变化使得脑电信号成为情感分类研究的主要研究对象。因此，在介绍基于脑电信号的情感能识别的相关背景研究前，我们首先会对脑电信号的背景知识进行详细的介绍。

有关脑电信号的研究有着悠久的历史。Richard Canton 于 1875 年首次从裸露的兔子和猿猴的大脑的表面探测到脑电信号。Danilevsky 在 1877 年发表了博士论文，该论文主要研究了动物大脑活动与诱导和自发脑电的关系。Beck 于 1890 年在生理学杂志上，发表了全脑中由闪光和鼓掌刺激引起了慢偶波中断的研究报告。Vladimir 于 1912 年成功地记录到了狗的脑电信号。Hans Berger，德国著名的生理学家、精神病学家，脑电图之父。Hans Berger 在 1929 年首次记录了人的脑电信号，而他使用的设备是自制的。与此同时，他将那些每秒高频出现 10 个周期的波命名为 alpha 波。随后，Berger 的研究不但被 Edgar Adrian 等人通过研究进行证明。而且他们同时在神经活动机制研究方面做出了卓绝的贡献。因此，Edgar Adrian 等人荣获 1932 年的诺贝尔奖 [73]。如今，医学界中众多研究表明，变换的脑电信号是人们对于日常生活的反应。当然，脑电信号也是一种工具，用于对大脑神经活动记录。其本身以及引申得来的事件相关电位（ERP, Event-Related Potential）被大范围的用在类脑研究、人工智能研究、医学、认知科学、心理学等。脑电信号的特点是信号微弱，信噪比低。脑电信号的赋值一般在 $50\mu V$ 左右，交流电 50 Hz 上下的频率会对脑电信号的采集产生一定影响。除此之外，眼球运动产生的电信号、肌肉运动的电信号和心电信号都会对脑电信号产生一定的干扰。脑电信号的产生与众多因素有关，因此，脑电信号会呈现出不平稳且不断变化的特征。

4.3 脑电情感能识别的背景和相关研究

若要进行脑电情感能识别，首先得了解情感的分类，而情感分类又具有多样性。其主要分为两种。第一种是离散型的情感，例如高兴（happiness）、悲伤（sadness）、惊讶（surprise）、恐惧（fear）、愤怒（anger）和厌恶（disgust）或是正面（positive）、负面（negative）和中立（neutral）等。需要注意的是，离散的情感分类只能表达具体的几种情感，具有局限性。为此出现了另一种二维度模型 [74]，其包含感觉良好程度（又称价，valance）以及清醒程度（arousal），而情感平面则是由这两个维度所张成，且平面上的每一个点代表一种情感。在这篇文章中，我们也采取了二维度（感觉良好程度和清醒程度）模型。特此，我们将其离散成高价高清醒、高价低清醒、低价高清醒、低价低清醒 4 类。

现在有许多方法都是利用人脸表情或者语音对人的情绪进行识别，而比较值

得推荐的方法，是使用神经系统产生的情绪信号（包含脑电信号）。目前，有很多已公开的研究，利用了脑电或者心理信号进行情感识别 [75][76][77][78]。极小一部分的工作采用了视频激励的情感，用于数据集的构建和预测，其中 [79] 则使用了对于电影反馈的心理信号去识别情感。这种电影被用于激发观众的 6 种情绪——娱乐、伤心、生气、害怕、失望、惊奇。我们选用的 DEAP 数据集 [71] 同样采取了电影片段激励情感的方法，同时选取了二维度（感觉良好程度和清醒程度）模型。此外，该数据集还测量了该影片观众对于影片的主导性、喜爱程度、熟悉程度等辅助情绪的属性。这些属性由测试者观看影片后，填写量表获得。特别地，[71] 中的 32 电极的脑电信号被提取成 theta (4-8 Hz)，慢波 alpha (8-10 Hz)，alpha (8-12 Hz), beta(12-30 Hz) 和 gamma (30+ Hz) 的能量谱，以及在左右半脑对称电极能量谱差的特征。此外，[71] 利用费舍尔线性判别式进行特征选取，利用高斯朴素贝叶斯进行 4 个类别分类的分类问题。[80] 利用 DEAP 数据集进行基于脑电信号的情感预测，并提取了每个电极的中值、标准差、峰度系数等统计量。而 theta (4-8 Hz)，慢波 alpha (8-10 Hz)，alpha (8-12 Hz), beta(12-30 Hz) 和 gamma (30+ Hz) 的频谱能量以及分型维度等特征也被提取出来。同时，利用最小冗余最大相关性 (minimum-Redundancy Maximum-Relevance) 方法去选择相关的特征。然后，用 SVM 分类器进行高低清醒程度和高低感觉良好程度分类。[81] 采用 DEAP 数据集，利用了 K 近邻方法来进行分类，分为伤心、生气、开心、悲伤四种情绪。[82] 利用快速傅里叶变换，然后使用皮尔森相关系数用于选择特征。他们提出了一个基于贝叶斯理论的概率分类器。[83] 使用分形维度谱探索了实时脑电识别算法。除上述传统方法外，深度信念网络、深度卷积网络这两种深度学习方法也被应用到 EEG 情感识别中 [84][85]。[84] 呈现出增加若干受限玻尔兹曼机层到深度信念网络中，并利用监督预训练后，则可用于学习数据复杂的表示。并且，相对于其他分类器的精确度，有着显著的提升。在 [86][87] 中，卷积神经网络和循环神经网络都被用来提取 EEG 时间序列的表示。

4.4 DEAP 数据集介绍

DEAP 数据集的记录，被使用在基于用户的当前状态来创建自适应音乐的视频推荐系统中。它记录了健康的 32 人，其中有 16 位男性和 16 位女性，且年龄都在 19 岁到 37 岁左右。要求每个受试者观看一分钟长的音乐视频，且每个受试者在每个音乐视频结束后填写含有感觉良好程度、清醒程度、喜欢/不喜欢、主导等选项的自我测试量表。同时，这些人中共有 22 人的前脸部表情被记录下来。EEG 信号和周边的信号被以 512 Hz 的频率进行采样。他们的 EEG 数据被下采样到频率为 128 Hz，然后眼电伪迹被去除，高通滤波器被使用。每一个参与者有 40 个观看的视频 \times 32 个 EEG 通道 \times 8064 个记录。同样地，标签有 40 个观看视频数量 \times 4。

我们通过采用文献 [23] 中相同的方法，采取等距投影 Azimuthal Equidistant Projection (AEP) 和 Clough-Tocher scheme 以及 1 秒短时快速傅里叶变换后提取 alpha

(8-13Hz), beta (13-30Hz) 和 gamma(30-45Hz) 频段的对数能量 (3 个频段中的每一种频段的对数能量对应一种颜色), 再将脑电信号等距映射成 32x32RGB 图像。两个情绪中的维度清醒和感觉良好程度均被标记为 1-9. 我们用 5 作为边界, 区分高低两个维度以生成 4 个类别。最后, 我们利用转化成的视频序列, 进行 4 类别情感预测。数据形状为 $1280 \times 63 \times 32 \times 32 \times 3$, 标签形状为 1280×4 。

4.5 脑电数据情感识别模型

其中 \mathbf{x}^i 为输入序列, \mathbf{z}^i 为因子序列, \mathbf{y} 为预测类别。

4.5.1 β -变分自编码部分

对于每一帧图片, 我们采用 $\beta - VAE$ 处理,

$$\mathcal{L}_{\beta-VAE} = \mathbb{E}_{\mathbf{z} \sim q_{enc}(\mathbf{z}|\mathbf{x})} \ln p_{dec}(\mathbf{x}|\mathbf{z}) - \beta D_{KL}(q_{enc}(\mathbf{z}|\mathbf{x})||p_{dec}(\mathbf{z})). \quad (4-1)$$

4.5.2 长短记忆网络部分

我们引入 $Z = \{\mathbf{z}^1, \dots, \mathbf{z}^M\}$ 来表示经过 $\beta - VAE$ 得到的因子序列, 将其输入给 LSTM,

$$\mathbf{i}^t = \sigma(W^{iz}\mathbf{z}^t + W^{ih}\mathbf{h}^{t-1} + \mathbf{b}^i), \quad (4-2)$$

$$\mathbf{f}^t = \sigma(W^{fz}\mathbf{z}^t + W^{fh}\mathbf{h}^{t-1} + \mathbf{b}^f), \quad (4-3)$$

$$\mathbf{o}^t = \sigma(W^{oz}\mathbf{z}^t + W^{oh}\mathbf{h}^{t-1} + \mathbf{b}^o), \quad (4-4)$$

$$\mathbf{g}^t = \tanh(W^{gz}\mathbf{z}^t + W^{gh}\mathbf{h}^{t-1} + \mathbf{b}^g), \quad (4-5)$$

$$\mathbf{c}^t = \mathbf{f}^t \odot \mathbf{c}^{t-1} + \mathbf{i}^t \odot \mathbf{g}^t, \quad (4-6)$$

$$\mathbf{h}^t = \mathbf{o}^t \odot \tanh(\mathbf{c}^t), \quad (4-7)$$

$$\mathbf{a} = softmax(W^{ah}\mathbf{h}^M + \mathbf{b}^a), \quad (4-8)$$

$$p_{\text{pre}}(\mathbf{y}|Z) = \prod_{i=1}^c a_i(\mathbf{h}^M)^{y_i}. \quad (4-9)$$

其中 $\mathbf{i}^t, \mathbf{f}^t, \mathbf{o}^t, \mathbf{x}^t, \mathbf{c}^t, \mathbf{h}^t$ 分别是 LSTM 模型的输入门, 遗忘门, 输出门, 输入向量, 隐藏状态向量 (记忆和输出部分)。其中 σ 是 sigmoid 函数, \tanh 是双曲正切函数, \odot 是逐点乘积, $softmax(\mathbf{x}) = \frac{(e^{x_1}, \dots, e^{x_n})^{tr}}{\sum_{i=1}^n e^{x_i}}$, W^{*x} 将输入转化成 LSTM 的状态的矩阵, W^{*h} 为隐藏状态之间的转化矩阵并且 \mathbf{b}^* 是偏差。

4.5.3 图模型部分

其中 $\mathbf{x}^i, \mathbf{z}^i, \mathbf{y}$ 之前的独立性关系详见如下图模型 (图 4-1 4-2 4-3)。

我们用下述目标函数作为长短记忆网络部分的目标函数,

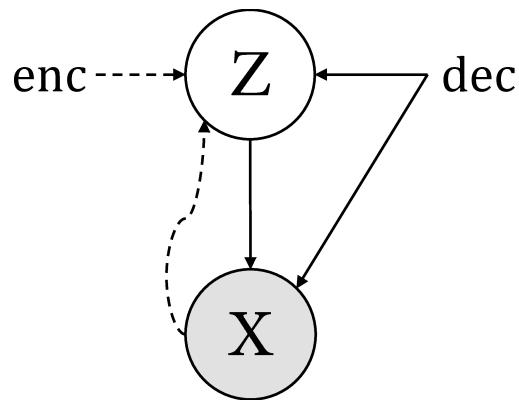


图 4-1 x 和 z 的图模型

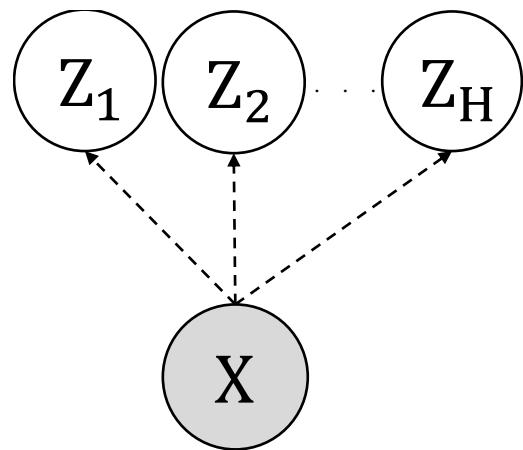


图 4-2 z 给出 x 的条件独立性

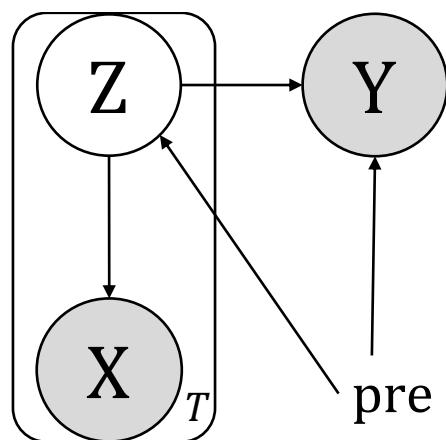


图 4-3 时间序列 x, z 和 y 的图模型

$$\mathcal{L}_{LSTM} = \mathbb{E}_{Z \sim q_{enc}(\mathbf{z}^1|\mathbf{x}^1) \cdots q_{enc}(\mathbf{z}^M|\mathbf{x}^M)} \ln p_{\text{pre}}(\mathbf{y}|Z). \quad (4-10)$$

那么有下述关系:

定理 4.1: 假设有

$$\mathcal{L}_{\beta-VAE_i} = \mathbb{E}_{\mathbf{z}^i \sim q_{enc}(\mathbf{z}|\mathbf{x}^i)} \ln p_{dec}(\mathbf{x}^i|\mathbf{z}^i) - \beta D_{KL}(q_{enc}(\mathbf{z}^i|\mathbf{x}^i) || p_{dec}(\mathbf{z}^i)) \quad i = 1, \dots, M. \quad (4-11)$$

那么如下不等式成立,

$$\mathcal{L}_{LSTM} + \sum_{i=1}^M \mathcal{L}_{\beta-VAE_i} \leq \log p_{\text{dec,pre}}(X, \mathbf{y}). \quad (4-12)$$

证明:

$$\log p_{\text{dec,pre}}(X, \mathbf{y}) \quad (4-13)$$

$$\geq \mathbb{E}_{Z \sim q_{enc}(\mathbf{z}^1|\mathbf{x}^1) \cdots q_{enc}(\mathbf{z}^M|\mathbf{x}^M)} \log p_{\text{dec,pre}}(X, \mathbf{y}|Z) \\ - D_{KL}(q_{enc}(\mathbf{z}^1|\mathbf{x}^1) \cdots q_{enc}(\mathbf{z}^M|\mathbf{x}^M) || p_{dec}(Z)), \quad (4-14)$$

因为

$$\log p_{\text{dec,pre}}(X, \mathbf{y}|Z) = \log p_{\text{dec}}(X|Z) + \log p_{\text{pre}}(\mathbf{y}|Z). \quad (4-15)$$

而

$$\begin{aligned} & \mathbb{E}_{Z \sim q_{enc}(\mathbf{z}^1|\mathbf{x}^1) \cdots q_{enc}(\mathbf{z}^M|\mathbf{x}^M)} \log p_{\text{dec}}(X|Z) - \beta D_{KL}(q_{enc}(\mathbf{z}^1|\mathbf{x}^1) \cdots q_{enc}(\mathbf{z}^M|\mathbf{x}^M) || p_{dec}(Z)) \\ &= \sum_{i=1}^M \mathbb{E}_{\mathbf{z}^i \sim q_{enc}(\mathbf{z}^i|\mathbf{x}^i)} \ln p_{\text{dec}}(\mathbf{x}^i|\mathbf{z}^i) - \beta D_{KL}(q_{enc}(\mathbf{z}^i|\mathbf{x}^i) || p_{dec}(\mathbf{z}^i)) \\ &= \sum_{i=1}^M \mathcal{L}_{\beta-VAE}(\mathbf{x}^i, \mathbf{z}^i). \end{aligned} \quad (4-16)$$

于是证明完毕。 ■

定理说明潜在的优化的目标是抬升最大似然模型的一个下界。

图 4-4 展示了模型处理数据的过程。每一帧图片形式的数据被传送给变分自编码器，以进行生成因子的提取，然后将提取的生成因子交由长短记忆网络，长短记忆网络结合之前的隐变量和当前的输入，将新得到隐变量输出给下一个长短记忆网络，最后一个长短记忆网络，将隐变量通过全连接网络输出所有预测的概率值。

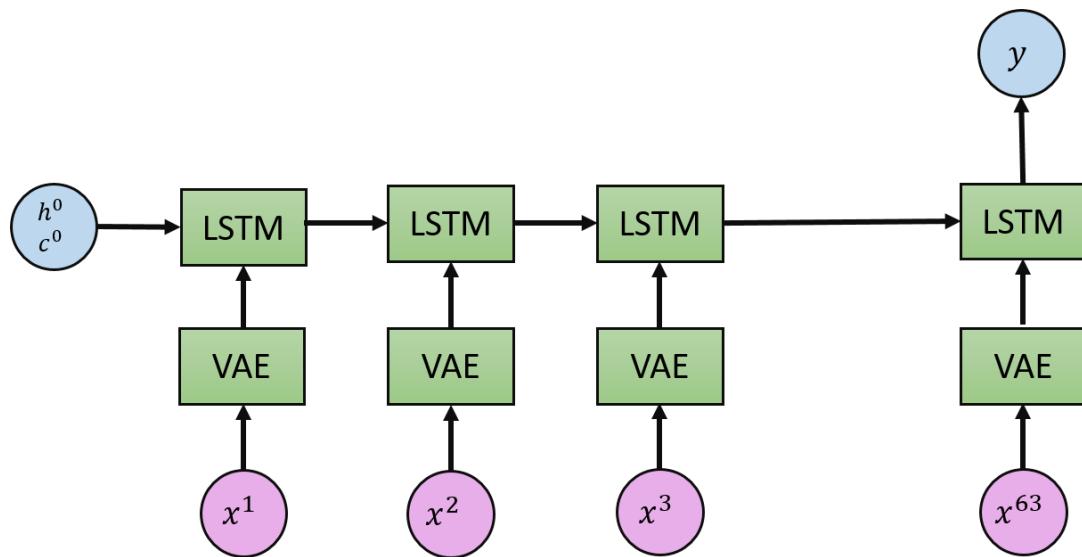


图 4-4 变分自编码器和长短记忆网络构成的脑电情识别展开模型

4.6 脑电数据情感识别实验结果

下表4-1为传统的 SVM 方法，普通的自编码器结合 LSTM，CNN 结合 LSTM 及 $\beta - VAE$ 的对比结果。由此，发现 VAE-LSTM 模型正确率和 CNN-LSTM 模型正确

表 4-1 不同方法脑电情感预测对比表

	方法	测试准确度
传统方法	PSD-SVM	0.39
摘除实验	AE-LSTM	0.52
	CNN-LSTM	0.54
最佳模型	$(\beta = 6)$ VAE-LSTM	0.54

率最高，相比于用抽取特征然后利用 SVM 的方法，该模型的正确率将会远高出 10 个左右的百分点。尽管 VAE-LSTM 模型与作为摘除实验的 CNN-LSTM 模型正确率一样，但 VAE-LSTM 模型具有可视化学得的特征的功能，同时能够适用于半监督学习，所以我们认为最佳的模型的是 VAE-LSTM 模型。

4.7 脑电数据情感识别结论

利用长短记忆网络和变分自编码器结合的模型，能够很好提取视频数据的特征，亦适用于后续分类任务。我们发现目前测试集分类准确率达到 54%，但是训练集分类准确率可以达到 100%，由此，我们估计这是由于 LSTM 模型过拟合造成的结果，如果有更多的成对的数据，也许能取得更好的结果。

5 结论与展望

5.1 结论

第一部分中，我们获得了深度表示学习适用于水文序列预测的结果，所设计的深度 LSTM 方法用于日径流时间序列是可行的，随着训练序列长度的增加，预测精度进一步提高。当然，由于 1945 年以前的资料存在中断，本次计算没有继续往前延伸，相信如果系列能继续加长，预测精度提高应该更明显；深度 LSTM 预测方法明显比浅层的前向网络 BPNN 方法预测时间序列的效果好。尽管结合混沌理论方法的 BPNN 是从 24 个方案中优选出来的，但是由于 BPNN 的方法只一个隐含层，没有深度表示学习，反映出非线性系统复杂关系的欠缺，而此时 LSTM 所含有的长短期记忆功能，赋予了它能够学习更加复杂的深度表示。在不同地区的实验展现出深度 LSTM 模型更优于 LSTM 模型。这体现出深度 LSTM 模型可以反应更加复杂的内在关系。因为相对平均误差在所有位置都很小，这展示了 LSTM 模型和深度 LSTM 普遍的适用性。

第二部分中，我们获得了因子图像表示学习可以学得有意义的生成因子并生成多样的数据以及具有互信息稀疏性的结果，所设计的指标挑选出来的因子，可以很好地服务于后续的分类重建等任务。互信息能够反映随机变量绝对的统计依存关系。变分自编码器的目标函数的第二项和过多预设因子数量倾向于诱导互信息的稀疏性，并且帮助诱导变分自编码器的有影响力和没有影响力因子的出现。我们也证明了互信息涉及重建均值误差的下界和分类的预测误差。我们设立可行的算法去计算指标，用以估计变分自编码器所有因子的互信息，并且证明了指标的一致性。实验显示有影响的因子和没有影响的因子都能够被自动有效的发现。发掘的因子的可理解性被有效支撑，并且它们的泛化以及分类能力也被验证。特别是一些和分类相关的变化也得以发现。这实验也激发了我们利用少量但是有影响力的因子来进行后续的数据处理任务（例如生成和分类）并达到与全部因子利用时相似的性能，就像 PCA 和 ICA 等的降维能力那样。最后我们从理论上解释了因子具有等价特性，这说明了变分自编码器很难一对一的学习到数据当中的变化。

第三部分中，我们获得了 LSTM 和变分自编码器的脑电情感深度表示学习架构达到了最新的水平的结果，所设计的方法同时还可以监督学得的特征形态。利用 LSTM 和变分自编码器结合的模型，能够很好提取视频数据的特征，亦适用于后续分类任务。我们发现目前测试集分类准确率达到 54%，但是训练集分类准确率可以达到 100%，我们估计这是由于 LSTM 模型过拟合造成的结果，如果有更多的成对数据也许能取得更好的结果。

综上所述，深度表示学习在时序、图像以及视频数据上有着充分的发展空间，并在一些已知的领域例如水文预测，图像因子学习生成，脑电情感识别达到了国际

前沿的水平。本文对于这些应用领域的拓展，以及深度学习自身的方法论层面，均具有显著的研究与应用启发。

5.2 展望

我们预测深度表示学习还可以横向的在时序，图像，视频等领域进一步拓展应用前景，例如自然语言理解以及其分支阅读理解、自动文本摘要、智能问答系统、话题推荐、机器翻译、知识库构建、语音技术识别，音频识别，基因数据处理，机器视觉以及其分支目标检测，目标识别，图像生成，图像描述，人脸识别，身份重识别，视频事件监测，视频描述，视频生成，以及混合型数据例如图像和时序文字数据结合（文本生成图像），视频与音频数据结合等领域有前进空间，同时纵向的可以在处理其他形式数据（例如具有图结构数据、点云结构数据、三维空间数据）以及对应的场景得到更进一步的发展。

考虑视频生成，第一种设想的方法是可以使用变分自编码器（Variational Autoencoder）或者自编码器（Autoencoder）的编码器提取图像级别的每个时刻特征，将过去所有时刻的特征输入给长短记忆网络（Long Short Term Memory Network），然后长短记忆网络输出下一个时刻的特征，经过变分自编码器或者自编码器的解码器解码成为生成视频的下一帧。整个模型改变条件上的一些因子，可实现对下一帧的可控制的生成。对于定长的视频，那么可以有第二种设想。第二种设想是直接使用配备了 3D 卷积的变分自编码器，编码器利用 3D 卷积网络抽取输入视频的特征，经过全连接网络得到视频的因子。更改视频的因子，经过全连接网络，再交由反 3D 卷积的解码器，解码成定长的可控制的视频。这样的话，可以预想到学得一些颜色变化的因子，空间位置移动变化的因子，对象速度变化的因子，光线变化的因子等等。

考虑音频生成，预期的结果是可以调整隐藏的生成因子，改变人说话的音色，例如从男声变成女声，从低沉变得尖细等等。从而这样达到类似于变声器的效果。采取的一种方法可以是利用 1D 卷积的变分自编码器网络。利用配备了 1D 卷积的变自编码器的编码器，将音频数据提取时序特征交由全连接网络，最终编码成因子，调整因子达到变换声音特质的效果。将调整的因子交由 1D 反卷积网络解码成变换的声音特质的音频数据。另一种方法是采取长短记忆网络，将音频数据按时序输入给长短记忆网络得到定长表示，将定长表示经过变分自编码器的编码器编码成因子。然后用变分自编码器的解码器将因子解码成定长表示。利用长短记忆网络逐项的输出时序的音频数据。调整其中的因子，达到可控制的音频生成。第二种方式可预期的好处是对于任意长度的音频数据都适用，缺点可能是对于过长的音频数据性能会不好。

考虑视频和音频数据结合的生成，其包含利用视频生成音频数据，利用音频生成视频数据，利用过去的视频和音频生成未来的视频和音频，还有利用因子进行视频音频的生成的主要四个范畴。对于第一种情形，对于成对的音频与视频数据，可

以首先尝试将视频通过变分自编码器结构压缩成因子。再用这些因子解码成音频数据，用真实的音频和生成的音频数据的误差回传梯度。当来一个新的视频数据时，便可以利用整个模型生成未知的音频数据了。对于音频生成视频数据的情形和视频生成音频数据的情形类似。利用过去的视频和音频生成未来的视频和音频，只需利用变分自编码器分别提取视频和音频的因子，然后将因子融合交由长短记忆网络，便可以和视频生成当中的第一设想采取相同的方法，从而实现生成。利用因子进行视频音频的生成，可以采用视频和音频的变分自编码器，将其得到因子融合交给一个新的变分自编码器，得到因子。修改相应的因子从而实现因子可控制的生成。

致 谢

时光如同白驹过隙，转眼之间三年的硕士研究学习生涯即将结束。这个过程中，我不断的学习、成长、收获。这离不开许多人的帮助、包容、鼓励，我对他们表示由衷的感谢。

感谢我的导师孟德宇老师和赵谦老师的悉心指导并为我提供相关资源。孟老师为人亲和，治学严谨、知识渊博。在攻读硕士期间，孟老师还为我提供了优越的实验环境和学习条件，也经常给我创造外出参加学术会议的交流机会，提高了我的科研能力，扩展了我的学术视野，这对于我今后的学习和工作很有益。非常感谢两位老师熬夜陪同修改相关论文《Fitting Data Noise in Variational Autoencoder》。感谢孟老师，在他的指导下《Discovering Influential Factors in Variational Autoencoders》通过多轮同行评审。

感谢刘京鑫的一同在变分自编码器应用在脑电情感识别上的合作。

感谢刘少华、毛红梅、杨文发在水文径流量预测上提供的帮助。

感谢俞韬、马子璐同学对于信息守恒定理的讨论，感谢谢领江，秦瑞同学对于脑电数据的处理。

感谢施宏扬同学帮忙进行论文的语言打磨工作。

感谢审阅本文的老师们，感谢答辩委员会的老师们，感谢老师们对于我论文提出的宝贵修改建议。

感谢各位亲爱的朋友、舍友和同学。感谢他们在学习和生活方面给我提供了并不少帮助，使得我的研究生生活过得丰富多彩。感谢机器学习讨论组的同学们的帮助，感谢他们提出的好点子。

最后，感谢学校提供了这样一个良好的平台为我的工作带来了便利。祝师生们一帆风顺，愿母校蒸蒸日上。

参考文献

- [1] Minsky M, Papert S. Perceptron: an introduction to computational geometry[J]. Expanded edition. Cambridge: The MIT Press. 1969, 19(88):2.
- [2] Rumelhart DE, Hinton GE, Williams RJ, et al. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3):1.
- [3] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation[J]. Readings in Cognitive Science, 1988, 323(6088):399-421.
- [4] Hornik K, Stinchcombe M, White H, et al. Multilayer feedforward networks are universal approximators.[J]. Neural networks, 1989, 2(5):359-366.
- [5] LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4):541-551.
- [6] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge: The MIT press, 2016: 502-525.
- [7] Kohonen T. Exploration of very large databases by self-organizing maps[C]//Proceedings of International Conference on Neural Networks (ICNN'97). Piscataway: IEEE, 1997: PL1-PL6.
- [8] Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for boltzmann machines[J]. Cognitive science, 1985, 9(1):147-169.
- [9] Grossberg S, Merrill JW. A neural network model of adaptively timed reinforcement learning and hippocampal dynamics[J]. Cognitive brain research, 1992, 1(1):3-38.
- [10] Lin CT, Jou CP. Controlling chaos by ga-based reinforcement learning neural network[J]. IEEE Transactions on Neural Networks, 1999, 10(4):846-859.
- [11] Pascanu R, Mikolov T, Bengio Y. Understanding the exploding gradient problem[J]. ArXiv preprint arXiv:1211.5063, 2012.
- [12] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.
- [13] Vapnik V, Chervonenkis A. A note on class of perceptron[J]. Automation and remote control. 1964, 24:1.
- [14] LeCun Y, Jackel L, Bottou L, et al. Comparison of learning algorithms for handwritten digit recognition[C]//International conference on artificial neural networks. Berlin: Springer, 1995, 60: 53-60.
- [15] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7):1527-1554.
- [16] Raina R, Madhavan A, Ng AY. Large-scale deep unsupervised learning using graphics processors[C]// Proceedings of the 26th International Conference on Machine Learning. New York: ACM, 2009: 873-880.
- [17] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2012: 1097-1105.
- [18] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.
- [19] Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models[C]//Proceedings of the 31th International Conference on Machine Learning. New York: ACM, 2014: 1278-1286.

- [20] Kingma DP, Welling M. Auto-encoding variational bayes[J]. ArXiv preprint arXiv:1312.6114, 2013.
- [21] Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2016: 613-621.
- [22] Zhu XZ, Xiong YW, Dai JF, et al. Deep feature flow for video recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2349-2358.
- [23] Bashivan P, Rish I, Yeasin M, et al. Learning representations from eeg with deep recurrent-convolutional neural networks[J]. ArXiv preprint arXiv:1511.06448, 2015.
- [24] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676):354.
- [25] Assael YM, Shillingford B, Whiteson S, et al. Lipnet: End-to-end sentence-level lipreading[J]. ArXiv preprint arXiv:1611.01599, 2016.
- [26] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2014: 2672-2680.
- [27] Gatys LA, Ecker AS, Bethge M. A neural algorithm of artistic style[J]. ArXiv preprint arXiv:1508.06576, 2015.
- [28] Wu B. Hierarchical macro strategy model for moba game ai[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019, 33: 1206-1213.
- [29] Hu MH, Wei FR, Peng YX, et al. Read+ verify: Machine reading comprehension with unanswerable questions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019, 33: 6529-6537.
- [30] Li ZH, Zhang S, Zhang JG, et al. Mvp-net: Multi-view fpn with position-aware attention for deep universal lesion detection[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer 2019: 13-21.
- [31] El-Hinnawi E, Hashmi MH, Yigong H. The state of global water resources[J]. World Environment, 1988, 4:22-25.
- [32] 陈兆丰. 水资源问答 [J]. 陕西水利, 1987, (4):19.
- [33] 戴明龙. 长江上游巨型水库群运行对流域水文情势影响研究 [D]. 武汉: 华中科技大学, 2017.
- [34] Wood EF, Roundy JK, Troy TJ, et al. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth's terrestrial water[J]. Water Resources Research, 2011, 47(5).
- [35] Halff AH, Halff HM, Azmoodeh M. Predicting runoff from rainfall using neural networks[C]//Engineering hydrology. Reston: ASCE, 1993: 760-765.
- [36] 刘少华, 丁贤荣, 毛红梅. 水文时间序列的混沌神经网络预报 [J]. 人民长江, 2002, 33(9): 13-15.
- [37] 胡国华, 宋荷花, 李正最. 基于人工神经网络的湘江最大洪峰流量中, 长期预报 [J]. 长沙交通学院学报, 2008, 24(2):72-77.
- [38] 李鸿雁, 苑希民, 等. 人工神经网络峰值识别理论及其在洪水预报中的应用 [J]. 水力学报, 2002, 6(6):15-20.
- [39] Carriere P, Mohaghegh S, Gaskari R. Performance of a virtual runoff hydrograph system[J]. Journal of Water Resources Planning and Management, 1996, 122(6):421-427.
- [40] Kratzert F, Klotz D, Brenner C, et al. Rainfall-runoff modelling using long short-term memory (lstm) networks[J]. Hydrology and Earth System Sciences, 2018, 22(11):6005-6022.

- [41] Widiasari IR, Nugroho LE, et al. Deep learning multilayer perceptron (mlp) for flood prediction model using wireless sensor network based hydrology time series data mining[C]//2017 International Conference on Innovative and Creative Information Technology (ICITech). Piscataway: IEEE, 2017: 1-5.
- [42] Hu CH, Wu Q, Li H, et al. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation[J]. Water, 2018, 10(11):1543.
- [43] Kratzert F, Klotz D, Shalev G, et al. Benchmarking a catchment-aware long short-term memory network (lstm) for large-scale hydrological modeling[J]. ArXiv preprint arXiv:1907.08456, 2019.
- [44] Zhang JF, Zhu Y, Zhang XP, et al. Developing a long short-term memory (lstm) based model for predicting water table depth in agricultural areas[J]. Journal of hydrology, 2018, 561:918-929.
- [45] 黄国如, 茹孝芳. 流域降雨径流时间序列的混沌识别及其预测研究进展 [J]. 水科学进展, 2004, 15(2):255-260.
- [46] Kingma DP, Ba J. Adam: A method for stochastic optimization[J]. ArXiv preprint arXiv:1412.6980, 2014.
- [47] Yang J, Zhang D, Frangi AF, et al. Two-dimensional pca: a new approach to appearance-based face representation and recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2004, 26(1):131-137.
- [48] Hyvärinen A, Karhunen J, Oja E. Independent component analysis[M]. Hoboken: John Wiley & Sons, 2004.
- [49] Salakhutdinov R, Hinton GE. Semantic hashing[J]. International Journal of Approximate Reasoning, 2009, 50(7):969-978.
- [50] Mathieu MF, Zhao JJ, Zhao J, et al. Disentangling factors of variation in deep representation using adversarial training[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2016: 5047-5055.
- [51] Suzuki M, Nakayama K, Matsuo Y. Joint multimodal learning with deep generative models[J]. ArXiv preprint arXiv:1611.01891, 2016.
- [52] Higgins I, Sonnerat N, Matthey L, et al. Scan: Learning abstract hierarchical compositional visual concepts[J]. ArXiv preprint arXiv:1707.03389, 2017.
- [53] Higgins I, Pal A, Rusu A, et al. Darla: Improving zero-shot transfer in reinforcement learning[C]//Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 1480-1490.
- [54] Kulkarni TD, Whitney WF, Kohli P, et al. Deep convolutional inverse graphics network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2015: 2539-2547.
- [55] Zhao SJ, Song JM, Ermon S. Learning hierarchical features from deep generative models[C]//Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 4091-4099.
- [56] Gregor K, Danihelka I, Graves A, et al. Draw: A recurrent neural network for image generation[J]. ArXiv preprint arXiv:1502.04623, 2015.
- [57] Rezende D, Mohamed S. Variational inference with normalizing flows[C]//Proceedings of the 32nd International Conference on Machine Learning. New York: ACM, 2015: 1530-1538.
- [58] Kingma DP, Salimans T, Jozefowicz R, et al. Improved variational inference with inverse autoregressive flow[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2016: 4743-4751.

- [59] Larsen ABL, Sønderby SK, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric[C]//Proceedings of the 33rd International Conference on Machine Learning. New York: ACM, 2016: 1558-1566.
- [60] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2016: 2234-2242.
- [61] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders[J]. ArXiv preprint arXiv:1511.05644, 2015.
- [62] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: The MIT Press, 2016: 2180-2188.
- [63] Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework[C/OL]//5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings. Amherst: OpenReview.net: 2017 [2020-05-30]. <https://openreview.net/forum?id=Sy2fzU9g1>.
- [64] Liu SQ, Liu JX, Zhao Q, et al. Discovering influential factors in variational autoencoders[J]. Pattern Recognition, 2020, 100:107166.
- [65] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27(8):1226-1238.
- [66] Cover TM, Thomas JA. Elements of information theory[M]. Hoboken: John Wiley & Sons, 2012.
- [67] Duchi J. Derivations for linear algebra and optimization[J]. Berkeley, California, 2007, 3: 2325-5870..
- [68] Alemi AA, Fischer I, Dillon JV, et al. Deep variational information bottleneck[J]. ArXiv preprint arXiv:1612.00410, 2016.
- [69] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
- [70] Liu ZW, Luo P, Wang XG, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 3730-3738.
- [71] Koelstra S, Muhl C, Soleymani M, et al. Deap: A database for emotion analysis; using physiological signals[J]. IEEE Transactions on Affective Computing, 2012, 3(1):18-31.
- [72] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2015: 4694-4702.
- [73] 刘志勇. 基于单通道脑电信号的睡眠分期算法及应用研究 [D]. 哈尔滨: 哈尔滨工业大学, 2017.
- [74] Russell JA. Affective space is bipolar.[J]. Journal of personality and social psychology, 1979, 37(3): 345.
- [75] Healey JA. Wearable and automotive systems for affect recognition from physiology[D]. Cambridge: Massachusetts Institute of Technology, 2000.
- [76] Lang PJ, Greenwald MK, Bradley MM, et al. Looking at pictures: Affective, facial, visceral, and behavioral reactions[J]. Psychophysiology, 1993, 30(3):261-273.
- [77] Kim J, André E. Emotion recognition based on physiological changes in music listening[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(12):2067-2083.

- [78] Wang JJ, Gong YH. Recognition of multiple drivers' emotional state[C]//2008 19th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society, 2008: 1-4.
- [79] Lisetti CL, Nasoz F. Using noninvasive wearable computers to recognize human emotions from physiological signals[J]. EURASIP Journal on Advances in Signal Processing, 2004, 2004(11):929414.
- [80] Atkinson J, Campos D. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers[J]. Expert Systems with Applications, 2016, 47:35-41.
- [81] Jadhav N, Manthalkar R, Joshi Y. Electroencephalography-based emotion recognition using gray-level co-occurrence matrix features[C]//Proceedings of International Conference on Computer Vision and Image Processing. Berlin: Springer, 2017: 335-343.
- [82] Yoon HJ, Chung SY. Eeg-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm[J]. Computers in biology and medicine, 2013, 43(12):2230-2237.
- [83] Liu YS, Sourina O. Eeg-based subject-dependent emotion recognition algorithm using fractal dimension[C]//2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway: IEEE, 2014: 3166-3171.
- [84] Plis SM, Hjelm DR, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study[J]. Frontiers in neuroscience, 2014, 8:229.
- [85] Mirowski P, Madhavan D, LeCun Y, et al. Classification of patterns of eeg synchronization for seizure prediction[J]. Clinical neurophysiology, 2009, 120(11):1927-1940.
- [86] Cecotti H, Graser A. Convolutional neural networks for p300 detection with application to brain-computer interfaces[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(3): 433-445.
- [87] Güler NF, Übeyli ED, Güler I. Recurrent neural networks employing lyapunov exponents for eeg signals classification[J]. Expert systems with applications, 2005, 29(3):506-514.

攻读学位期间取得的研究成果

- [1] Ma ZL, Liu SQ, Meng DY, et al. On Convergence Properties of Implicit Self-paced Objective[J]. Information Sciences, 2018, 462: 132-140.
- [2] Liu SQ, Ma ZL, Meng DY. Understanding Self-Paced Learning under Concave Conjugacy Theory[J]. Communications in Information and Systems, 2018, 18(1): 1-35.
- [3] Liu SQ, Liu JX, Zhao Q, et al. Discovering influential factors in variational autoencoders[J]. Pattern Recognition, 2020, 100: 107166.

声明

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 交回学校授予的学位证书；
2. 学校可在相关媒体上对作者本人的行为进行通报；
3. 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。
4. 本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）：刘仕琪 日期：2020年6月2日

学位论文独创性声明（2）

本人声明：研究生刘仕琪所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 学校可在相关媒体上对本人的失察行为进行通报；
2. 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。
3. 本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：孟德军 日期：2020年6月2日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公開阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）：刘仕琪 日期：2020年6月2日

指导教师（签名）：孟德军 日期：2020年6月2日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)