分 类 号_____ 学号 M201772537
学校代码 10487 密级_____

華中科技大學

硕 士 学 位 论 文

基于深度学习的
医学影像分类方法研究

学位申请人： 王阳
学 科 专 业： 控制科学与工程
指 导 教 师： 伍冬睿 教授
答 辩 日 期： 2020 年 5 月 27 日

A Thesis Submitted in Partial Fulfillment of the Requirements
For the Degree of Master of Engineering

Research on Medical Image Classification ~~Based on Deep Learning~~

Candidate : Yang Wang
Major : Control Science and Engineering
Supervisor : Prof. Dongrui Wu

Huazhong University of Science & Technology

Wuhan 430074, P. R. China

May, 2020

独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保密□，在 ____ 年解密后适用本授权书。

　　不保密 □。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘　要

医学影像技术为大量疾病的临床诊断和治疗提供了关键的信息。在实际操作中，对医学影像数据的解读通常由具有丰富实践经验的专业医生人工进行，该过程繁琐、工作量大、耗费时间，且严重依赖于专家的主观经验，不同专家之间也存在评判一致性不高的问题。计算机辅助诊断技术的引入可以极大地提高医生的效率。现有的基于人工特征和传统机器学习方法的研究，通常需要较强的医学先验知识和繁琐的特征工程，限制了算法性能的上限。深度学习技术可以提供一种端到端的解决方案，能够从数据中自发地学得具有良好判别能力和泛化能力的多层次抽象特征，显著提高了预测结果的精度，研究深度学习在医学影像分析中的应用具有重要意义。

本文以最基础也最重要的诊断分类为例，分别基于脑电图（EEG）、磁共振成像（MRI）和延时视频（time-lapse）这三种具有较大差异的医学影像数据，对不同的医学诊断问题进行研究，展示了不同的深度学习技术和用法。通过对深度学习方法进行不同程度上的改进和创新，本文提出的方法在对应研究的医学问题上均达到了了业界领先的水平。希望能够抛砖引玉，为深度学习在医疗影像数据中的应用提供比较多样和全面的视角。本文的主要研究内容如下：

(1) 基于EEG和深度学习的睡眠状态检测。睡眠状态的检测有利于睡眠相关疾病的诊断。目前现有的深度学习方法大多基于原始的EEG信号，具有计算量大、算法性能一般的缺点。本文将单通道EEG信号通过短时傅立叶变换转换成时频图后，从自然语言处理中获取灵感，提出了两种新颖的深度学习模型用于睡眠状态的检测。基于卷积神经网络（CNN）的C-CNN模型紧凑且高效，在计算代价和模型性能之间达到了很好的平衡；基于注意力机制和双向长短时记忆的Attention模型则可以获得更好的性能，领先于现有的方法。此外，通过将代价敏感学习整合到模型训练过程中，解决了睡眠状态中存在的严重类别不平衡问题，保证每个睡眠状态均有较高的召回率，取得了更好的平衡分类正确率。

(2) 基于MRI和CNN的鼻咽癌诊断支持系统。鼻咽癌是中国东南部、台湾、香港、马来西亚和新加坡等地区最常见的头颈癌，本文为鼻咽癌的诊断提供了一套可视化的诊断支持系统，且可以在不同分辨率不同型号的MRI设备下运行。首先，该套系统通过开操作和大津阈值法对MRI切片进行自适应的分割和裁剪，提取切片中有效的大脑部分，同时解决了跨分辨率跨系统的问题；然后使用修改后的残差网络对不同MRI切片进行处理提取特征，利用提出的可视化方法快速定位恶性肿瘤可能存在的切片和区域；最后对所有切片的高层次抽象特征进行整合，给出最终的鼻咽癌阳性概率。整套系统在为医生提供可疑肿瘤位置标注的同时，鼻咽癌阳性诊断的ROC曲线下面积AUC指标也达到0.994。

(3) 基于time-lapse和多任务学习的胚胎早期发育阶段分类。在体外人工受孕的治疗过程中，对胚胎的早期发育阶段进行准确地检测，可以为评估胚胎质量提供宝贵的信息，有利于受孕的成功。本文提出了一种具有动态规划的多任务深度学习框架（MTDL-DP），用于胚胎早期发育阶段的分类。它首先基于视频中相邻帧具有大量互补信息的特性，利用多任务学习和相邻帧为时延视频的每一个图片帧生成多个预测结果；然后通过集成思想对这些预测进行整合，赋予当前帧一个胚胎发育阶段；最后使用动态规划进行后处理，优化整个视频的发育阶段序列，使得最终预测的发育阶段序列非递减，且搬运距离损失最小。通过提出的MTDL-DP算法，本文将胚胎早期发育阶段分类的精度提高了3.1%。

.

关键词： 深度学习　　注意力机制　　脑电图　　磁共振成像　　延时视频　　睡眠状态检测　　鼻咽癌　　胚胎发育阶段分类

Abstract

Medical imaging technology provides critical information for ~~the~~ clinical diagnosis and treatment of a large number of diseases. In practice, the interpretation of medical image data is usually performed manually by ~~professional~~ doctors ~~with rich practical experience.~~ The process is tedious, ~~heavy~~, time-consuming, and heavily dependent on the subjective experience of experts. There is also a problem of judgment consistency among different experts. The introduction of computer-aided diagnosis technology can greatly improve the efficiency of doctors. Existing research based on artificial features and traditional machine learning methods usually requires strong medical prior knowledge and tedious feature engineering, which limits the upper limit of ~~algorithm~~ performance. Deep learning ~~technology~~ provides an end-to-end solution that can spontaneously learn the multi-level abstract features with good discrimination and generalization capabilities from the data, which significantly improves the accuracy of prediction results, so it quickly dominates various fields.

This ~~paper~~ ~~takes the most basic and important diagnostic classification as an example.~~ Based on three different medical image ~~data~~, including electroencephalography (EEG), magnetic resonance imaging (MRI), and time-lapse video, we study different medical diagnostic problems, and demonstrate different deep learning techniques and usages. Through various degrees of improvement and innovation in deep learning methods, our proposed method has reached the industry-leading level in the corresponding medical research problems. ~~Hoping~~ to provide a comprehensive perspective for the application of deep learning in medical imaging data. The main research contents of this paper are as follows:

(1) Sleep state classification based on EEG and deep learning. Scoring of sleep stages plays an important role in the diagnosis of sleep-related diseases. At present, most of the existing deep learning methods are based on the original EEG signals, which requires a large amount of calculation and the algorithm performance is poor. In this paper, we convert the single-channel EEG signal into a time-frequency map by short-time Fourier transform, and obtain inspiration from natural language processing. Two novel deep learning approaches are proposed for the detection of sleep states. The C-CNN model based on convolutional neural network (CNN) is compact and efficient, achieving a good balance between computational cost and model performance; the Attention model based on attention mechanism and bidirectional long-term memory can get better performance, ~~ahead of existing methods.~~ In addition, we integrate cost-sensitive learning into the model training process, which solves the serious class imbalance problem in sleep states, ensures that each sleep state has a higher

recall rate, and achieves a better balance classification accuracy.

(2) Nasopharyngeal carcinoma diagnosis support system based on MRI and CNN. Nasopharyngeal carcinoma ~~has a high incidence~~ in China and other Southeast Asian regions. This paper provides a set of visual aided decision-making techniques for the diagnosis of nasopharyngeal carcinoma and can be run under different resolutions and MRI equipment. It first performs adaptive segmentation and cropping of MRI slices to extract the effective brain parts through the open operation and Otsu threshold method, simultaneously solving the problem of different resolution and MRI equipment. ~~And then uses~~ the modified residual network to extract features from slices, the proposed visualization method is also used to quickly locate the slices and regions where malignant tumors may exist. Finally, the high-level abstract features of different slices are integrated to give the final positive probability of nasopharyngeal carcinoma. The whole set of techniques ~~not only~~ provides doctors with the suspicious location of tumors, ~~but also~~ the area under the ROC curve (AUC) for nasopharyngeal carcinoma positive diagnosis ~~achieves~~ 0.994.

(3) Multi-task deep learning with dynamic programming for embryo early development stage classification from time-lapse videos. During the treatment of in-vitro fertilization, accurate detection of the early developmental stages of the embryo can provide valuable information for the quality assessment of embryo, which is beneficial to the success of conception. This ~~paper~~ proposes a multi-task deep learning framework with dynamic programming (MTDL-DP) for the classification of embryo early development. Firstly, based on the characteristic that adjacent frames in the video have lots of complementary information, it uses multi-task learning and adjacent frames to generate multiple predictions for each frame in the time-lapse video. ~~Then~~ integrates these predictions through ensemble learning to give the current frame one embryo development stage. Finally, dynamic programming post-processing is used to optimize the predicted embryo development stage sequence of the entire video, so that the final sequence is non-decreasing, and the loss of earth-mover distance is minimal. Through the proposed MTDL-DP approach, this ~~paper~~ has improved the accuracy of embryo early development stage classification by 3.1%.

.

Key words: Deep Learning Attention Mechanism EEG MRI Time-lapse Sleep Stage Classification Nasopharyngeal Carcinoma Embryo Development Stage Classification

目　录

1 绪论

1.1 研究背景及意义

在过去的几十年中，计算机断层扫描（Computed Tomography，CT）、脑电图（Electroencephalogram，EEG）、磁共振成像（Magnetic Resonance Imaging，MRI）、正电子发射断层扫描、超声波成像和X 射线等医学影像技术已经被广泛用于大量疾病的早期检测、诊断和治疗中[1]。这些改进的观测工具极大地提高了临床诊断治疗的效率，也推动了科学研究的发展。在提供了与病情相关的重要信息的同时，也揭示了生物学和疾病之间的未知秘密。随着医学影像技术的快速发展和大数据人工智能时代的到来，如何从大量的医学影像数据中挖掘有用信息，并加以分析应用，对于疾病的诊断治疗和科学研究的发展均具有重大的意义。

医学影像的研究分析涉及到光学成像、数字图像处理、物理建模、临床医学和人工智能等多个方面。长期以来，如何实现医学影像的自动化处理和分析、提高算法预测性能，一直都是计算机科学领域的研究热点。但在实际应用中，由于病理学上的巨大差异，以及临床对预测结果精度和可靠性有极高要求，现有的计算机辅助技术难以落地。因此，在当前临床环境下，医学影像的解释主要由具有丰富实践经验的放射学家和医生等人类专家完成。人工处理的过程十分繁、工作量大，需要耗费较多的时间。评估结果的质量主要依赖于人类专家的主观经验，考虑到医生可能的疲劳、实践经验不一以及医疗资源的匮乏等因素，如何保证评估结果的客观性和可靠性是一个极大的挑战。因此，为了提高临床医生和研究人员的工作效率、改善落后地区的医疗水平、提高临床诊断的可靠性，仍然迫切需要合适的全自动医学影像分析算法。

在深度学习得到广泛青睐之前，主要的医学影像分析算法是基于特征工程和传统机器学习技术的。传统的的机器学习技术难以直接处理原始的自然数据，因此构建一个机器学习系统通常需要包含以下步骤：对原始数据进行复杂的预处理、设计和提取特征、选取合适的特征、训练恰当的模型。其中最关键的工作是设计能够很好地描述数据内在规律和模式的特征。一般而言，设计有意义且任务相关的特征需要较强的先验知识，只能由具有目标领域知识的专家进行，这使得非相关专家难以利用机器学习技术进行自己的研究。而且，受制于经验，专家设计的特征也可能存在一些未知的局限性，限制了算法性能的上限。有一些研究[2,3] 尝试基于预定义的字典从训练集中学习稀疏表示，并证明了使用此类方法在医学影像分析中进行特征表示和特征选择的有效性。然而，上述工作中使用的稀疏表达和字典学习也只能从数据学得一些比较浅层的信息模式和内在规律，限制了它们的表达能力。

近年来，随着数据量以及计算资源的飞速增长，基于深度学习的方法得到了广

泛的应用。深度学习是机器学习方法中一个新兴的领域，与传统机器学习方法不同的是，深度学习技术通过将特征工程整合到学习过程中。这使得特征工程的重担从人类专家转移到了计算机上，既克服了人工设计特征存在的问题，也提供了一种端到端的解决方案。图 1-1 展示了传统机器学习算法和深度学习算法整个训练流程的对比，深度学习算法的训练只需要进行模型的调整，整个流程显得更加简洁和流畅。在对原始数据进行少量的预处理后，深度学习可以自动挖掘输入和输出之间的关系，通过多层的特征转换，自发地学得复杂的具有优秀判别能力和泛化能力的信息表达，显著提高了系统性能。事实证明，深度学习非常善于挖掘高维数据中的复杂结构，在很多人工智能领域难以攻克的问题上取得了重大进展，在计算机视觉[4]、音频识别[5]、自然语言处理[6] 等领域得到了广泛应用，引发了深度学习的研究浪潮。

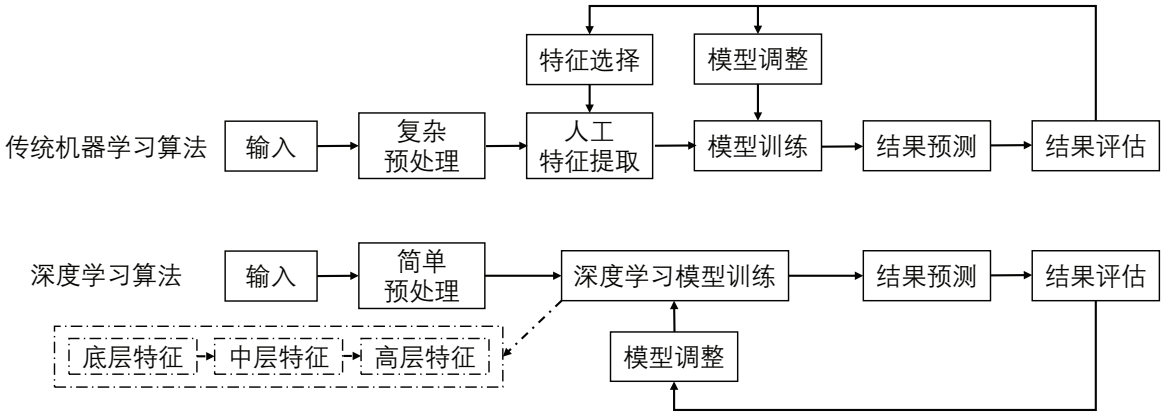


图 1-1 传统机器学习和深度学习算法训练流程的对比

不同于在其它领域上取得的惊人成就，深度学习在医学影像上的应用还相对有限。与传统的图像相比，医学影像具有其领域的局限性。除了同为二维图像但仍然具有显著差异的超声波和X光等观测手段外，医学领域中还存在着大量像MRI和CT一样的三维数据，以及类似时间序列的生理信号（例如EEG和心电图等），难以直接复制深度学习在其它领域的应用经验。虽然还存在着较大的困难，但将深度学习应用于医学影像分析中，开发高性能的计算机辅助诊断系统，已是大势所趋。这有利于提高医生的诊断效率，补充落后地区的医疗资源，提前疾病的诊断时间，挽救无数患者的生命。其次，让医生从繁琐冗杂的诊断工作中解放出来，将更多的注意力和时间放在跟病人的沟通上，也有助于缓解医患冲突的问题。随着人民物质生活水平的提高和健康意识的日益增加，对于医学影像分析中最基础且最重要的诊断分类问题，研究如何将深度学习应用其中，提高自动化诊断系统的性能，可以带来巨大的社会价值和经济效益，具有重大意义。

1.2 深度学习方法概述

深度学习是一种具有多层表达的表达学习方法，通过反向传播算法调整模型权重，其能够自动从数据中挖掘输入输出之间的关系，在很多问题上已经表现出优异的性能。深度学习模型一般由多层简单的线性或非线性模块组成，每层模块将一个级别的输入转换成更高层次的抽象表达，经过足够多的此类转换后，可以学得非常复杂的函数[7]，从而产生具有优秀判别能力和泛化能力的鲁棒特征。以分类任务为例，其高层表达会放大对类别区分非常重要的方面，而抑制不相关的变化。在其中最关键的一点是，这些深度学习模块的功能不是由人类工程师设计的，而是通过通用的学习过程从数据中自发地学得的。这将特征工程的重担从人类专家转移到了计算机上，从而使得机器学习领域中不具有相关背景的专家也能有效地使用深度学习进行相关领域的研究和应用，引领了很多跨界研究。发展至今，绝大多数的深度学习模型都遵守着从上面整理出的两条核心原则：1）由多层简单的线性或非线性处理模块堆叠而成；2）通过反向传播算法（链式法则）更新模型权重。

在本节中，将对深度学习的发展历史进行简单地回顾，并对最基础的几种深度学习模型进行介绍。

1.2.1 深度学习发展简史

虽然深度学习像是最近几年才兴起的技术，但是它所基于的神经网络技术已经被研究了近百年，最早可以追溯到1943年。如今的深度学习，可以视为加强加深版的神经网络。但它在21世纪初期的时候并不流行，导致其看起来像是一门新的技术。神经网络的发展历经波折，大致可以分为三个阶段。

从神经两个字就可以看出，早期的神经网络尝试模拟人类大脑神经元的运作机制。最早的神经网络模型是由神经生理学家McCulloch和逻辑学家Pitts在1943年提出的，他们模拟大脑神经元的结构提出了MP模型[8]。但实际上人类对大脑的运作机制并没有足够的了解，此时的模型还比较简陋，其本质上是对输入的线性加权和。对于n个输入，MP模型会通过n个权重来计算输入的加权和，通过检验函数的正负得到0或者1的输出。输入的权重由人根据经验设置，比较麻烦。1958年Rosenblatt 对此进行了改进，提出了可以根据数据学习特征权重的感知机模型[9]。然而，这些MP和感知机模型都存在比较明显的局限性，只能处理简单的线性可分问题，连异或这种问题都无法解决。此后，神经网络的研究陷入长达十几年的衰退期。

在20世纪80年代，伴随着联结主义潮流，迎来了神经网络的第二次研究浪潮，其中最重要的成就是分布式表达和反向传播算法。分布式表达认为，每个系统的每一个输入都应该由多个特征表示，且每个特征都应该参与多个可能输入的表示。例如对于n种颜色m种型号的汽车，相比对$n \times m$个组合使用单独的神经元来激活，使用n个神经元描述颜色，m个神经元描述型号，可以使用更少的参数，且

通过对颜色和型号进行解耦，可以从更多的样本中学习特征，不再仅限定于特定的组合。分布式表达增强了模型的表达能力，使得神经网络从宽度转向深度发展。1986年Rumelhart等人提出反向传播算法[10]用于调整网络权重的大小，成功改善了神经网络的训练过程，得到了迅速普及。至今为止，反向传播算法仍然是训练深度学习模型的主要方法。同时期以支持向量机为代表的传统机器学习算法也取得了突破性进展，在很多重要任务上均取得很好的效果，超过了神经网络的结果。此时的神经网络仍然受到数据量和计算资源的限制，这导致了神经网络研究的第二次停滞。

到2010年左右，随着互联网的发展、传感器成本的降低，获得海量数据变得简单。图形处理器GPU和云计算的出现，也使得计算机性能得到了进一步的提升。至此，神经网络对数据量和计算资源的需求得到了解决，迎来了新的发展。在ILSVRC图像分类挑战赛中，传统计算机视觉方法在ImageNet数据集上的最低top5错误率为26.2%，2012年Krizhevsky等人提出的深度卷积神经网络AlexNet[11]将其降低到了16.44%，以大幅度的优势取得了冠军。从此，深度学习作为深层神经网络的代名词为世人所熟知。2013年之后的ILSVRC 就基本只有深度学习方法参加了，2016年Kaiming等人提出的残差网络ResNet[4]进一步将top5错误率降低惊人的3.57%，到现在为止ResNet仍在学术界和工业界广泛采用。随着Caffe、Tensorflow和Pytorch等开源框架的出现，深度学习的热度越来越高，也由最开始的计算机视觉领域扩展到机器学习的各个领域，开启了近些年来的研究浪潮。

1.2.2 人工神经网络

人工神经网络（Artificial Neural Network，ANN）是最著名的深度学习模型之一，有着悠久的研究历史。19世纪60年代提出的单层感知机就属于ANN的范畴，其是一种仅有输入层和输出层的特殊ANN。然而单层感知机仅能处理一些简单的线性可分问题，对异或等非线性问题则无法解决。随着数据量和计算资源的增加，ANN通过对处理层进行堆叠加深以及引入非线性变换，极大地提高了网络的处理能力，扩展了ANN的使用范围。

ANN通常由多层具有大量计算单元（神经元）的全连接层堆叠组成。相邻两层上的神经元互相密集地连接在一起，这也是这些处理层被称为全连接层的原因。神经元会对输入进行一些简单地变换计算，通过在神经元的计算中引入Tanh、Sigmoid和整流线性单元（ReLU）等非线性激活函数，也使得ANN具有非常复杂的非线性变换能力。图 1-2 展示了一个神经元的计算过程。对于输入数据$\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$，神经元进行如下计算：

$$h(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + b) = f(\sum_i w_i x_i + b) \tag{1.1}$$

其中$\boldsymbol{w} = (w_1, w_2, ..., w_n)^T$为权重向量，$b$为偏置项，函数$f$为激活函数。从上可知，

神经元的计算其实比较简单。只是对输入数据的特征进行简单地加权和后，通过激活函数进行非线性变换，输出最终的标量结果。

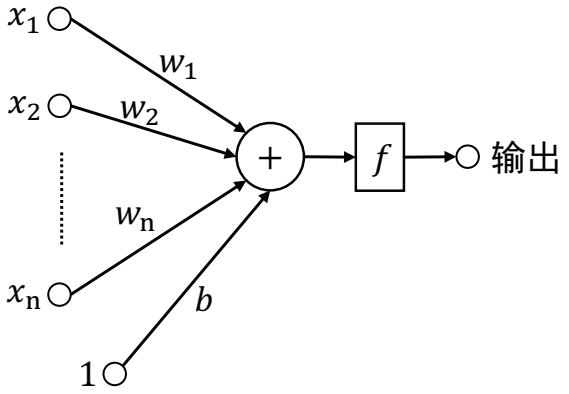


图 1-2 神经元

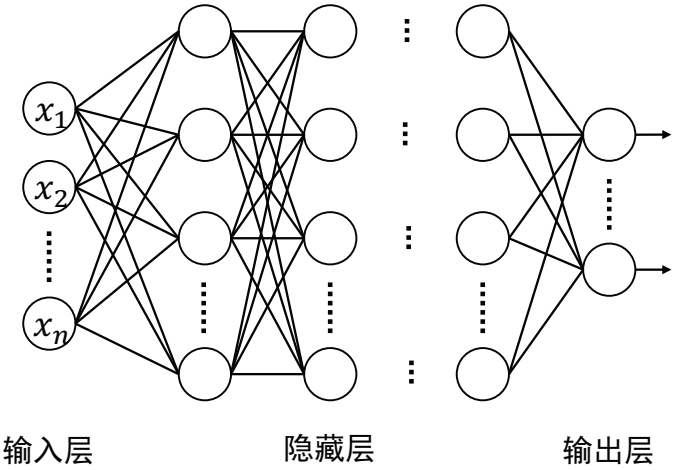


图 1-3 人工神经网络

ANN的典型结构如图 1-3 所示，除输入层外，其外的节点均为神经元。输入层的节点代表数据的特征，经过一层或者多层的隐藏层，逐层对特征进行变换，最后由输出层对这些高层特征进行整合，给出网络最终的预测结果。在反向传播算法的监督下，通过简单地堆叠神经元，ANN可以自发地调整权重对输入数据的特征进行多层次的非线性变化，达到较高的预测精度。人工神经网络入和输出之间的转换关系可以由式 1.2 表示。

$$h(\boldsymbol{x}) = f\left(\boldsymbol{W}^T f\left(\boldsymbol{W}^T \cdots f(\boldsymbol{W}^T \boldsymbol{x} + \boldsymbol{b})\right) + \boldsymbol{b}\right) \tag{1.2}$$

其中$\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$为全连接层的权重矩阵，$\boldsymbol{w}_i$为该全连接层中第$i$个神经元的权重向量，偏置向量$\boldsymbol{b} = (b_1, b_2, ..., b_n)$同理。

1.2.3 卷积神经网络

有很多数据格式是以多维数组的形式出现的：包括语言在内的一维信号序列；二维的图像和音频频谱图；三维的视频和体积图像等。ANN通常被用于处理非结构化数据，对于以这些以多维数组形式呈现的结构化数据ANN往往无能为力。这是由ANN稠密连接的特性决定地，计算效率比较低。在ANN失宠的时期里，卷积神经网络（Convolutional Neural Network，CNN）取了许多实际的成功，并且近年来在计算机视觉领域受到了广泛采用。在CNN的背后，隐藏着基于信号的天然属性提出的四个关键思想：局部连接、权重共享、池化和多层堆叠。

最早可以通过反向传播算法训练的CNN是由LeNet等人在1990年提出的LeNet-5[12]，被用于进行低分率的手写数字图像识别。典型的CNN结构如图 1-4 所示，其前几个阶段通常由卷积层和池化层交替组成，每层的输出被称为特征图（Feature Map），最终的特征图经向量化展开后，由全连接层整合给出预测结果。

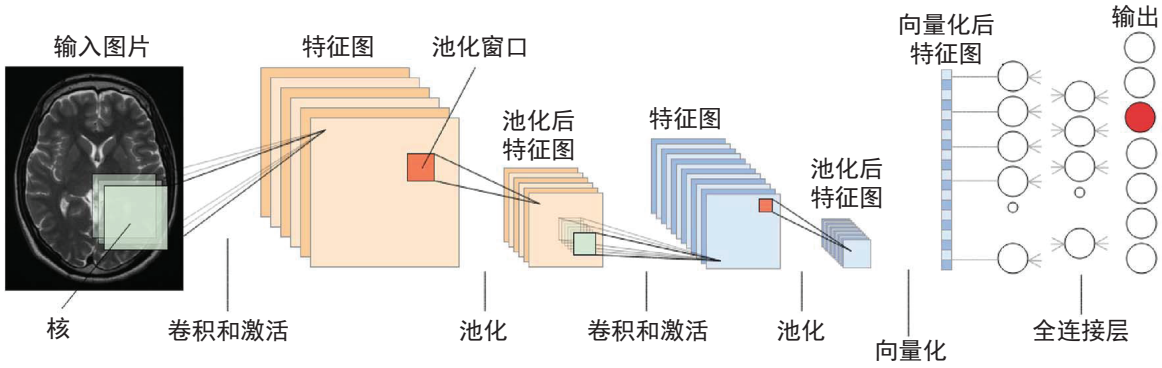


图 1-4 典型的CNN结构

卷积层包含多个具有可训练参数的卷积核（kernel），这些卷积核也可以称为过滤器（filter）。以二维卷积层为例，可令卷积核尺寸为$k \times m$。该卷积核会在输入的数据或者特征图上，以一定的步长依次移动（权重共享），对与它局部相连的数据计算加权和组成卷积层的输出（局部连接）。图 2-5 展示了一个尺寸为2×2、步长为1的卷积核进行卷积计算的过程。卷积层之所以这样设计是有两个原因的。首先，对于结构化数据（例如图像），数据中的值通常具有很强的局部相关性，形成了易于检测的独特局部模式，因此卷积核在进行计算时是设计为局部连接的。其次，图像和其它信号的局部统计模式对于位置是不敏感的。即，如果某模式出现在图像中的某一部分，那么它也可以出现在图像中的任何其它位置。因此，可以使用具有相同权重的卷积核在数组上移动，从任何位置中检测某个相同的模式。在数学上，卷积层在输入上的这种操作被称为离散卷积，这也是卷积层的命名原因。卷积层后，往往紧接着如ReLU类似的激活函数，进行非线性处理。

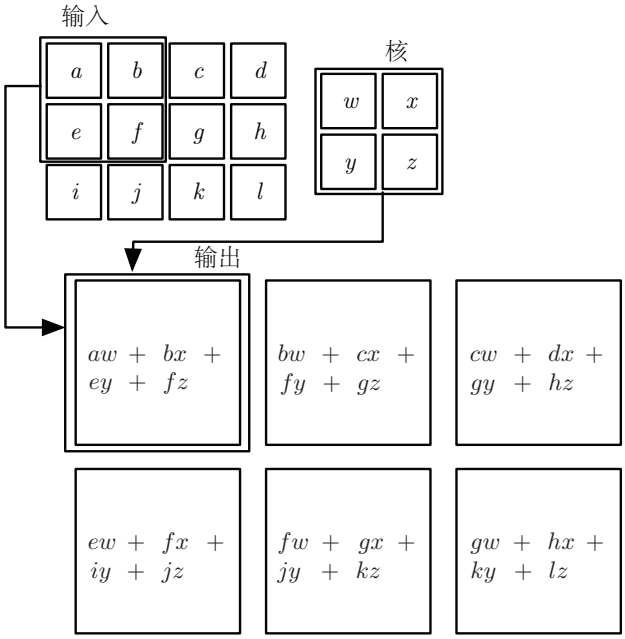


图 1-5 卷积运算示意图

卷积层的作用是从上一层输入中检测局部特征模式，而池化层的作用则是将语义相似的特征合并为一个特征。对于组成某个图案的特征，其位置通常可能是略有变化的。因此，对这些特征的位置进行粗粒化的处理，可以更可靠地检测该图案。池化层的计算模式与卷积层类似，但将卷积核替换成了无训练参数的操作。常用的池化层有最大池化和平均池化层。对于最大池化，其会取局部连接的特征图中的最大值作为输出，平均池化则是取均值。与卷积层不同是，池化层的步长通常比较大，这样不仅可以减小特征图的尺寸减少计算量，也可以引入小幅度的移位和失真，从而创造了位移不变性。当上一层中的元素位置和外观发现变化时，池化层可以减小这些变化，提高特征的鲁棒性。

很多自然信号都是具有层次化结构的，这是进行多层堆叠的原因。深度神经网络利用此特性，通过对较低层的特征进行组合获得高层的抽象特征。以图像为例，边缘的局部组合形成了图案，而图案组合成了部分结构，这些结构最终组成了图像中的对象。由音素和音节组成的音频，由单词和句子组成的文本，均具有类似的层次化结构。通过对卷积层、非线性激活函数和池化层进行多层堆叠，最后紧随着全连接层预测结果，组成了一个典型的CNN结构。CNN的参数同ANN一样，可以通过反向传播算法进行更新。

1.2.4 循环神经网络

对于音频、语言等涉及时序输入的任务，最适合的模型是循环神经网络（Recurrent Neural Network，RNN）。RNN是一种特殊的神经网络，具有很强的长

时依赖捕获能力。RNN使用相同的参数对不同时间步上的输入进行处理，在其隐藏单元中维护着一个“状态向量”，该向量隐藏着该输入序列过去所有元素的历史信息。图 1-6 中展示了一个随时间展开后的RNN层。

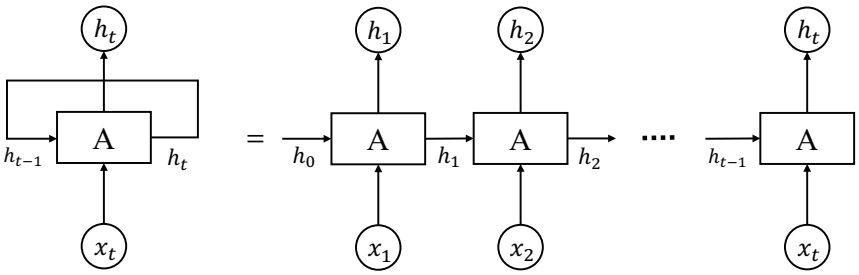


图 1-6 循环神经网络按时间展开后的结构

对于给定输入序列$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{T'})$，RNN从$t = 1$ to T'通过迭代下式计算隐藏层状态$\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{T'})$:

$$\boldsymbol{h}_t = f(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}) \tag{1.3}$$

其中$\boldsymbol{W}_x$和$\boldsymbol{W}_h$为RNN层的权重矩阵，$\boldsymbol{b}$为偏置向量，f是激活函数。从上式可以发现，RNN的当前隐藏层状态$\boldsymbol{h}_t$不仅跟当前的输入$\boldsymbol{x}_t$有关，也受到之前的隐藏层状态$\boldsymbol{h}_{t-1}$的影响。

将RNN在时间维展开后，可以视作非常深的前馈网络，其中的所有层共享相同的权重。尽管RNN为了学习长时依赖关系特意设计了这样的结构，但理论和实验均表明，它还是很难长时间地存储信息。对于这个问题，一个主要的改进思路是显式地使用增强的记忆单元。长短时记忆[13]和门控循环网络[14] 等研究使用记忆单元和门机制来有选择地存储和遗忘信息，显著提高了捕获长时依赖信息的效率。跟RNN相关的工作，最高可以追溯到1982年Hopfield教授的研究[15]。发展至今，经过结构和训练方式上的改进，RNN已经在机器翻译[6]、图像标题生成[16]和问答[17]等自然语言处理相关问题上长期占据着支配地位。

1.3 本文的工作和贡献

计算机视觉的发展在一定程度上促进了深度学习在医学影像分析中的发展，在诊断分类[18]、图像分割[19]、影像配准[20]、多模态融合[21]、标注[22]、计算机辅助诊断和预后以及病灶检测[23]等子领域均有一定的应用。但是这些研究还比较粗糙，更多的只是利用了深度学习优秀的特征提取能力，没有关注到医学影像和传统图像在数据格式上的差异，注入医学领域的专业知识。

本文以最基础也是最重要的诊断分类为例，分别基于脑电图（EEG）、磁共振成

像（MRI）和时延视频（Time-lapse）等不同类型的医疗影像数据，对三种不同的医学诊断问题进行了探究。图 1-7 展示了这三种数据的示意图。其中EEG是一种时间序列数据，大脑神经元在进行活动时会产生生物电流，通过放置在大脑表皮的电极，可以采集到EEG信号；而MRI是一种三维的图像数据，在使用MRI设备对人体进行扫描后，可以生成高精度的结构信息；Time-lapse则是在体外胚胎培养中常用的技术，其会以较短的时间间隔拍摄胚胎图片，实时记录胚胎的发育过程，得到的延时视频有利于医生对胚胎的发育状况进行评估。

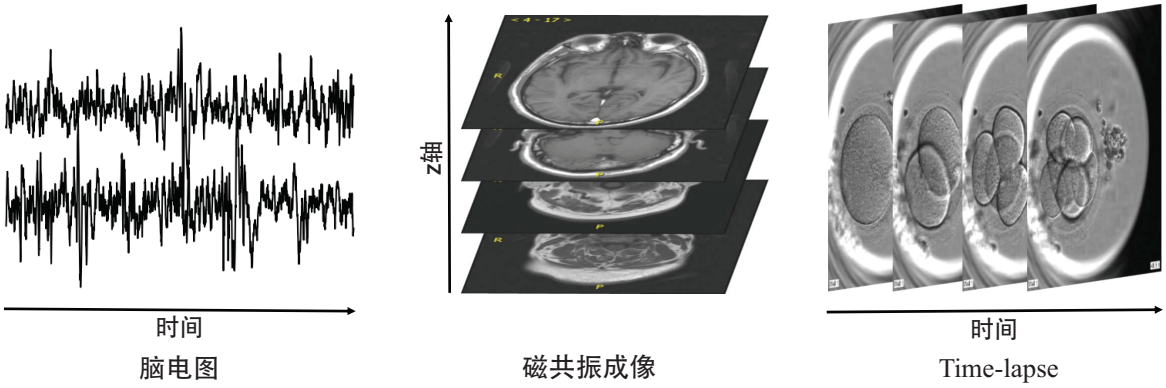


图 1-7 数据格式示意图

本文对这三种具有较大差异的数据进行了探讨，从不同的角度出发，展示了不同的深度学习技术和用法。通过对深度学习方法进行不同程度上的改进和创新，本文提出的方法在对应研究的医学问题上均达到了了业界领先的水平。希望能够抛砖引玉，为深度学习在医疗影像数据中的应用提供比较多样和全面的视角。

本文的主要研究内容如下，其均为本文作者在硕士研究生期间的研究内容，大部分工作已经在正式会议或期刊上发表：

(1) 基于EEG和深度学习的睡眠状态检测。睡眠状态的检测有利于睡眠相关疾病的诊断。目前现有的深度学习方法大多基于原始的EEG信号，具有计算量大、算法性能一般的缺点。本文将单通道EEG信号通过短时傅立叶变换转换成时频图后，从自然语言处理中获取灵感，提出了两种新颖的深度学习模型用于睡眠状态检测。基于卷积神经网络的C-CNN 模型即紧凑又高效，在计算代价和模型性能之间达到了很好的平衡；基于注意力机制和双向长短时记忆的Attention模型在需要更长训练时间的代价下，可以获得更好的性能，领先于现有的方法。此外，通过将代价敏感学习整合到模型训练过程中，解决了睡眠状态中存在的严重类别不平衡问题，保证每个睡眠状态均有较高的召回率，取得更好的平衡分类正确率。

(2) 基于MRI和CNN的鼻咽癌诊断支持系统。鼻咽癌是中国东南部、台湾、香港、马来西亚和新加坡等地区最常见的头颈癌，本文为鼻咽癌的诊断提供了一套

可视化的诊断支持技术，且可以在不同分辨率不同型号的MRI设备下运行。其首先通过开操作和大津阈值法对MRI切片进行自适应的分割和裁剪，提取切片中有效的大脑部分，同时解决了跨分辨率跨系统的问题；然后使用修改后的残差网络对不同MRI切片进行处理提取特征，利用提出的可视化方法快速定位恶性肿瘤可能存在的切片和区域；最后对所有切片的高层次抽象特征进行整合，给出最终的鼻咽癌阳性概率。整套系统在为医生提供可疑肿瘤位置标注的同时，鼻咽癌阳性诊断的ROC曲线下面积AUC指标也达到0.994。

(3) 基于time-lapse和多任务学习的胚胎早期发育阶段分类。在体外人工受孕的治疗过程中，对胚胎的早期发育阶段进行准确地检测，可以为评估胚胎质量提供宝贵的信息，有利于受孕的成功。本文提出了一种具有动态规划的多任务深度学习框架（MTDL-DP），用于胚胎早期发育阶段的分类。它首先基于视频中相邻帧具有大量互补信息的特性，利用多任务学习和相邻帧为时延视频的每一个图片帧生成多个预测结果；然后通过集成思想对这些预测进行整合，赋予当前帧一个胚胎发育阶段；最后使用动态规划进行后处理，优化整个视频的发育阶段序列，使得最终预测的发育阶段序列非递减，且搬运距离损失最小。通过提出的MTDL-DP算法，本文将胚胎早期发育阶段分类的精度提高了3.1%

1.4 本文的组织结构

本论文的组织安排如下：

第一章是绪论。介绍了将深度学习应用于医学影像数据分析中的研究意义、深度学习的发展历史以及基础模型，并阐述了本文的主要研究内容。

第二章介绍了睡眠状态检测的相关现状和面临的问题，并提出了基于脑电图和深度学习的睡眠状态检测算法。

第三章基于磁共振成像结合卷积神经网络和可视化技术，提出了一套针对鼻咽癌的诊断支持技术；

第四章基于time-lapse技术对人工体外受孕中的早期胚胎发育阶段分类进行了研究，通过多任务学习、集成思想和动态规划显著提高了现有算法的精度；

第五章为本文的总结与展望。对全文的工作进行了总结回顾，并且探讨了将深度学习应用在医学影像中存在的一些问题。

2 基于EEG和深度学习的睡眠状态检测算法

睡眠状态检测在睡眠相关疾病的诊断中起着重要的作用。本章介绍了睡眠状态检测的相关现状和面临的问题，并进行了改进。在将单通道EEG信号通过短时傅立叶变换转换成时频图后，从自然语言处理中获取灵感，提出了两种新颖的深度学习模型用于睡眠状态检测，并在扩展的Sleep-EDF数据集上验证了本文提出方法的有效性。基于卷积神经网络的C-CNN 模型即紧凑又高效，在计算代价和模型性能之间达到了很好的平衡；基于注意力机制和双向长短时记忆的Attention 模型在需要更长训练时间的代价下，可以获得更好的性能，领先于现有的方法。此外，通过将代价敏感学习整合到模型训练过程中，解决了睡眠状态中存在的严重类别不平衡问题，保证每个睡眠状态均具有较高的召回率，取得了更好的平衡分类正确率。

2.1 引言和相关工作

每个人的一生中，有几乎三分之一的时间是在睡眠中度过的。在睡眠期间，身体的大多数系统都处于合成代谢状态，这有助于免疫、神经、骨骼和肌肉等系统的恢复。因此，睡眠在人体健康中起着至关重要的左右。

然而大多数人存在睡眠问题，在美国至少有10%的人口饱受睡眠疾病的困扰[24]。对睡眠状态进行合适的评价，有利于诊断睡眠疾病和追踪治疗效果[25]。多导睡眠图（Polysomnography，PSG）是睡眠质量评定的金标准。它通常需要被试者穿戴各种传感器以记录多种生理信号，包括脑电图（Electroencephalogram, EEG）、眼电图（Electrooculography，EOG）、肌电图（Electromyogram, EMG）、呼吸速率等。将PSG记录分割成多个30s的片段（epoch），其后睡眠专家会根据一定的协议视觉检查这些生理信号，如传统的Rechtschaffen and Kales（R&K）标准[26]，或者由美国睡眠医学学会（American Academy of Sleep Medicine, AASM）[27]提出的标准。这个人工打分的过程很耗时，并且在不同的专家之间打分的结果会不一致。有研究表明在不同是专家之间，打分结果的一致性仅有82.6%[28]，因此迫切需要全自动的睡眠状态分类系统。

现在已经有不少基于人工特征的机器学习方法被应用于睡眠状态分类[29,30]。其步骤一般如下：预处理去除伪影和噪声，特征提取和选择以获得具有区分性的特征，最后使用机器学习方法训练一个分类器。这些打分系统的性能严重依赖于人工特征的质量，通常受到特征设计者的经验限制，难以达到最优。

相比于传统的基于特征工程和机器学习的方法，深度学习提供了一个端到端的解决方案，可以自动挖掘原始输入和输出之间的关系。深度学习已经在很多的应用中获得了巨大的成功，包括图像处理[31]、视频分析[32]、自然语言处理[6,33]等。同样

地，自动编码器[34]、卷积神经网络（Convolutional Neural Network，CNN）[35–37]和循环神经网络（Recurrent Neural Network，RNN）[38,39]等深度学习的方法也被引入到睡眠状态检测中。

还有一些研究尝试利用睡眠状态的转移规则[35,37–40]，睡眠专家在实践中也会利用这种规则，参考当前状态来决定下一个可能的状态。以R&K标准为例，如果当前的状态是S3，那么下个睡眠状态只可能出现在S2、S3和S4中。通过利用这种转移规则，检测系统的性能可以得到进一步的提高。然而，它通常需要系统使用多个相邻的epoch作为输入，这会增加运行时间和计算代价。

全自动睡眠状态分类系统的输入信号既可以是单模态，也可以是多模态的。虽然使用多模态信号通常可以达到更好的性能，但需要被试穿戴众多的传感器，这对长时间的睡眠监测来说，很不友好：由于过多的传感器，穿戴和连接保持接触良好很花时间，且不舒适易导致部分被试入睡困难。随着人们对睡眠质量意识的增强，对使用移动脑电图设备的自动睡眠状态系统也越来越感兴趣[41,42]，这些设备的计算资源通常都是有限的。为了便宜、方便和适合在现实生活中进行长时间的睡眠质量监测，本文计划只使用单通道EEG，且不使用转移规则，这样可以便于嵌入移动设备中。需要注意的是，本文提出的方法也可以作为其它研究的基础，轻松加入转移规则、扩展到多通道多模态的输入。

此外，不同的睡眠状态的状态出现的概率是不等的，如果不对该类别不平衡现象进行特殊的处理，少数类可能会被算法忽略掉。大多数基于深度学习的睡眠状态分类方法会试图对数据进行重采样（过采样或者下采样），从而使得所有的睡眠状态具有相同的样本数[34–40]。然而，重采样会改变原始数据的分布，可能导致其它的问题。例如，过采样增加了样本的总数量，导致更多的计算代价；下采样会随机的去掉部分样本，因此会损失部分信息。在本文中，通过使用代价敏感学习来解决类别不平衡的问题[43]，可以简单的理解为，对少数类样本赋予更大的损失权重，使得它们训练过程中不会被倾轧忽视。

2.2 方法

2.2.1 数据集

在本章的实验中，使用了PhysioNet[44]上的扩展Sleep-EDF数据集[45]，更准确的说，使用了该数据中的SC（Sleep Cassette）部分。包括了20名被试，10名男性和10名女性，年龄在25到14岁之间。其中19名被试具有连续两天的PSG记录，还有一名被试仅有一天。为了保持一致性，我们只使用了具有两天记录的19名被试。每个PSG记录包含了EEG（来自Fpz-Cz和Pz-Cz对）、垂直EOC、颏下巴EMG 和呼吸速率，其中EEG 和MEG的采样率均为100Hz。每个30s的epoch 会由相关专家根据R&M标准分成一下几类：运动时间（movement time，M）、清醒（wakefulness，

W）、快速眼动（rapid eye movement, REM）和非快速眼动（non-REM），其中non-REM又可以分为状态1 （S1）、状态2 （S2）、状态3（S3）和状态（S4）。本文仅考虑W、REM 和S1至S4共六个状态，且仅使用EEG中的一个电极对信号。不同睡眠状态的epoch数量如表 2.1 所示。

表 2.1 各个睡眠状态的EEG epoch的数量和比例.

	W	S1	S2	S3	S4	REM	Total
数量	70,450	2,731	17,302	3,307	2,249	7,545	103,600
比例（%）	68.00	2.65	16.70	3.19	2.17	7.28	100.00

2.2.2 算法流程图

本章所提出算法的流程图如图 2-1 所示。对于每一个30s的EEG epoch，其首先执行预处理，将它转换成一个时频图$\boldsymbol{X}$，然后将其送入特征提取组件提取特多层次的特征，不同的特征提取组件设计将在下文中进行详细地介绍。提取器后紧跟着dropout[46]以避免过拟合，然后使用具有128个神经元的全连接层去整合这些特征，由Softmax层给出最终的预测结果。

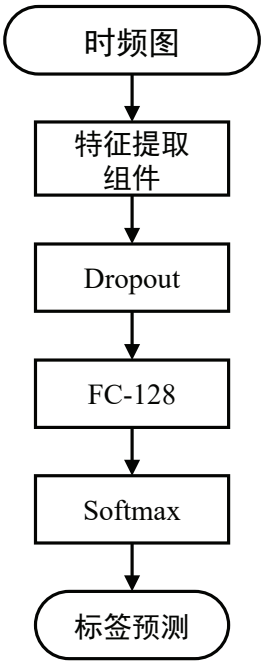


图 2-1 所提出算法的流程图

2.2.3 预处理

考虑到直接对原始EEG数据进行处理计算量会比较大，以及睡眠状态主要跟频域特征有关，对原始的EEG信号进行了如下的预处理。

对于给定的30s epoch，首先使用短时傅立叶变换（Short-Time Fourier Transform, STFT）将其转换成时频图，其横轴为时间、纵轴为频率。STFT使用了Hamming窗口，窗口大小为1s，步长为0.5s。之后将功率谱密度的单位转换成dBs。因此，每一个30s的epoch会被转换成一个时频图$\boldsymbol{X} \in R^{T \times F}$，其中$T$ (time) 为59，F (frequency) 为51。一个转换好的时频图如图 2-2 所示。需要注意的是，因为EEG信号的采样频率为100Hz，由奈奎斯特采样定理可知，信号可恢复的最大频率为50Hz。

最后，执行z-score标准化，使得所有特征的均值为0，方差为1，该步骤有利于模型训练的收敛。

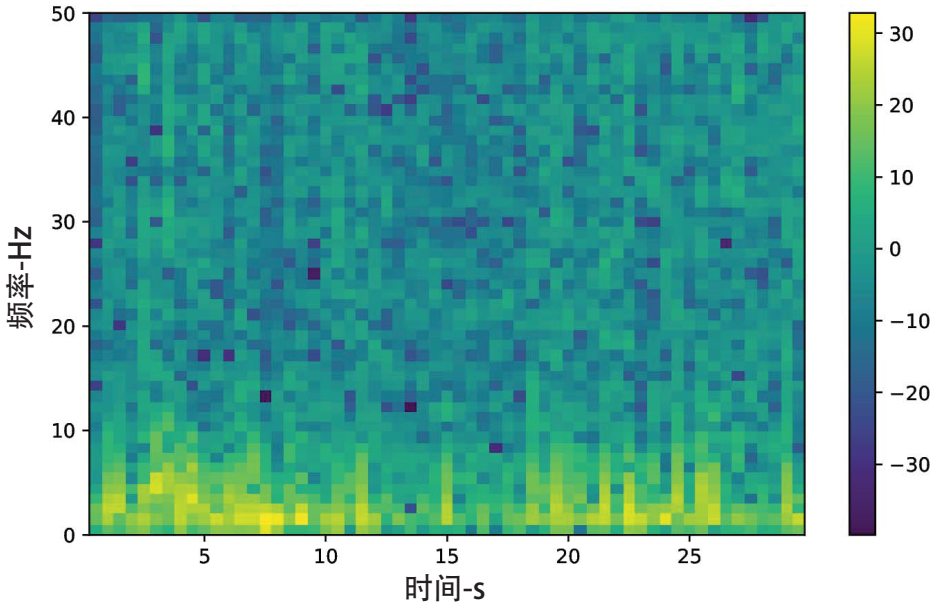


图 2-2 时频图

2.2.4 基础特征提取组件

本文采用的基准特征提取组间类似于文献[33]中的CNN-static (CNN-s)。该模型最初被提出用于自然语言处理，因为其简单的结构和有竞争力的表现，现在已经被广泛用于其它相关的领域。

图 2-3 展示了该模型的主要部分。在自然语言处理中，通常使用一个预训练过的固定维度的词向量来代表一个词。词向量是一个低维的稠密向量，其一般通过在大规模的语料库上进行训练得到，可以代表词与词之间的相对关系。将句子中的词依次替换成对应的词向量，构成该句子的特征表达，以该词向量序列作为模型的输

入。经过对比可以发现，这里的词向量序列和转换EEG信号后得到的时频图很相似：词向量对应时频图中的频率向量，词向量的顺序对应频率向量的时间顺序。因此，本文采用文献[33]中的CNN-s作为基准特征提取组件。

图 2-3 中的$Conv1D(3\times F-128)$代表具有128个通道（特征）的一维卷积层，其卷积核大小为$3\times F$，其中F为时频图中频率向量的长度。在卷积操过程中，由于该卷积核是在时频图的时间维度上移动的，因为也被称为时间维卷积。

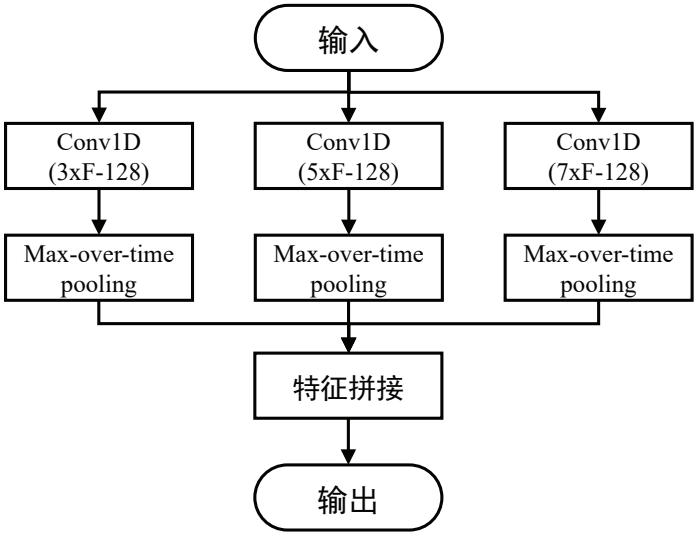


图 2-3 基准特征提取组件A-CNN

使用不同的卷积核大小可以获得不同大小的感受野。我们在初步实验中尝试一些不同卷积核尺寸，获得相似的结果，这也侧面说明了模型的稳定性。最终选择的卷积核尺寸为$3\times F$、$5\times F$和$7\times F$。卷积步长为1，使用受限线性单元（rectified linear unit, ReLU）作为激活函数。

对于经过卷积获得的特征图，执行了一个步长为1的时间维上最大池化操作（max-over-time pooling）。即对于一个特定的卷积通道，在时间维度上选择最大的值作为最终的特征。这里的想法是，对于每一个卷积通道，只期望捕获最重要的特征（最大值）。最后，将不同尺度的特征拼接在一起，用于后续的处理。

2.2.5 提出的特征提取组件

2.2.5.1 CNN模型

本文提出一个新的特征提取组件以强化A-CNN，如图 2-4 所示。在该模组件中，只使用了大小为$3\times F$和3×1的卷积核，而不是使用$5\times F$ 和$7\times F$大卷积核（感受野）。这个想法来自Simonyan等人的研究[31]。通过在一个$3\times F$的卷积层上堆叠一个3×1的卷积层，可以获得$5\times F$大小的等效感受野，图形解释见图 2-5。相似的，

在$3 \times F$ 卷积层上堆叠两个3×1 的卷积层可以获得$7 \times F$的等效感受野。

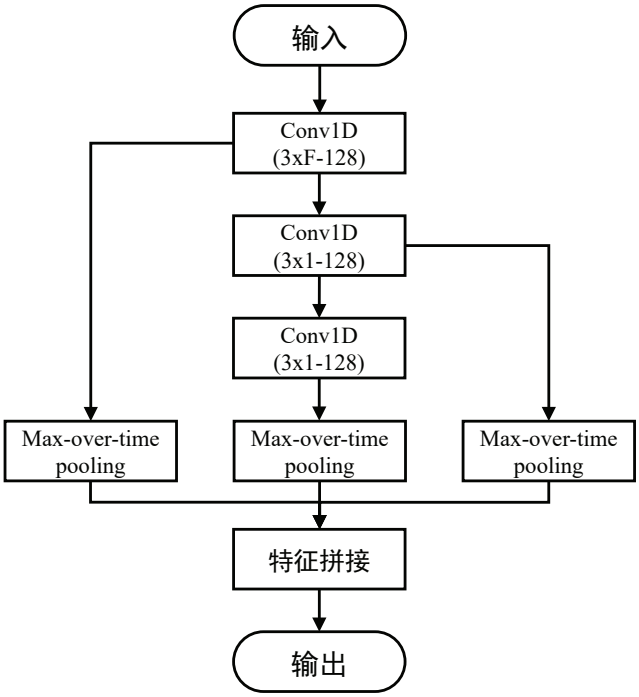


图 2-4 提出的C-CNN特征提取组件

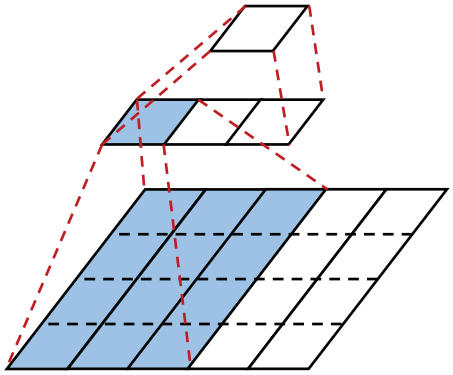


图 2-5 堆叠一个$3 \times F$和3×1的卷积层等效于一个$5 \times F$的卷积层

堆叠三个小卷积核的卷积层而不是直接使用一个7×F的卷积层，至少可以提供一下两个益处：首先，实际上使用了三个ReLU层而不是一个，这样可以获得更多的非线性；其次，需要的参数量更少，使得模型更加紧凑。假设卷积层的输入输出通道数均为C，那么一个单独的$7 \times F$ 的卷积层所需参数为$7 \times F \times C^2$，然而堆叠卷积层所需的参数为$3 \times (F+1+1) \times C^2$，仅为前者的43%。

更进一步的，为了获得不同尺度和层次的特征，我们直接将中间层的特征用于

最终的预测。这样也可以直接向低层直接推动梯度，让这些层立即起作用，使得训练过程更加稳定和收敛更快。在每个卷积层后都使用了批标准化[47]，其被证实能够处理内部协变量偏移（Internal Covariate Shift），使得训练更加容易。因此，获得最终了图 2-4 所示的特征提取组件C-CNN。

2.2.5.2 Attention模型

除了CNN外，我们也在睡眠状态分类中应用了具有注意力机制的双向长短时记忆（Long Short-Term Memory，LSTM）[13]，提出了如图 2-6 所示的Attention模型。注意力机制[6]已经在时间序列和自然语言处理中获得了广泛的应用。

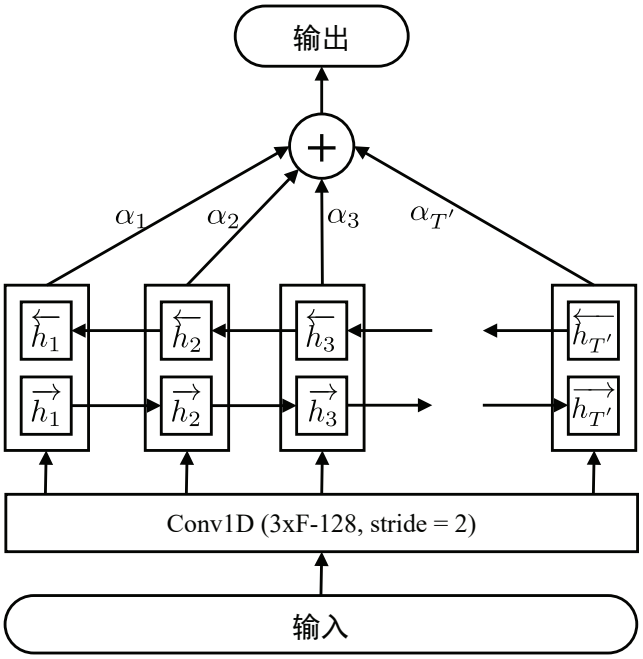


图 2-6 提出的Attention特征提取组件

RNN具有很强地捕获长时信息依赖的能力，考虑到EEG数据和神经反应的时间动态特性，使用RNN对大脑活动的时间演变进行建模是一个很合理的选择。因此，很自然的可以将RNN应用在基于EEG的睡眠状态检测中。给定输入序列$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{T'})$，RNN从$t = 1$ to T'通过迭代下式计算隐藏层状态$\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{T'})$:

$$\boldsymbol{h}_t = f(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}) \tag{2.1}$$

其中$\boldsymbol{W}_x$和$\boldsymbol{W}_h$为RNN层的权重矩阵，$\boldsymbol{b}$为偏置向量，f激活函数，它们均在不同的时间步上共享。因此，RNN的当前隐藏层状态$\boldsymbol{h}_t$不仅跟当前的输入$\boldsymbol{x}_t$有关，也受到之前的隐藏层状态$\boldsymbol{h}_{t-1}$的影响。

长短时记忆（Long Short-Term Memory，LSMT）[13]是RNN的一种推广，它是一种具有增强记忆单元的RNN，由Hochreiter & Schmidhuber在1997年提出。具体而言，LSTM加入了记忆单元，该单元具有内部记忆状态和输入、输出以及遗忘门控。LSTM的隐藏层函数由以下的方程组计算：

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \tag{2.2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \tag{2.3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-t} + \boldsymbol{b}_c) \tag{2.4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o) \tag{2.5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t tanh(\boldsymbol{c}_t) \tag{2.6}$$

其中σ是logistic sigmoid函数，而LSTM层组件中的输入门、遗忘门、输出门和记忆单元向量分别用$\boldsymbol{i}$、$\boldsymbol{f}$、$\boldsymbol{o}$和$\boldsymbol{c}$表示（具体细节可以参见原始文献[13]）。

LSTM通过使用这些门机制和记忆单元来有选择地存储和遗忘信息，因此可以更有效地捕获长时依赖关系。在本文中，我们更进一步地使用了具有注意力机制的双向LSTM。据我们所知，目前还没有研究将其引入到睡眠状态分类当中。

当训练一个单向的LSTM时，存在无法使用未来输入信息的缺点。然而，双向的LSTM[48]可以解决这个问题。它由前向和后向的LSTMs的组成，分别在正的时间方向和负的时间方向上同时训练。拼接前向的隐藏层状态$\overrightarrow{\boldsymbol{h}_t}$和后向$\overleftarrow{\boldsymbol{h}_t}$（$\overleftarrow{\boldsymbol{h}_t}$包含了未来的输入信息），可以获得双向LSTM的隐藏层状态$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}]$。在本文中使用的LSTM 单元数为128。对于注意力机制，其已经被各种应用广泛采用，例如机器翻译[6]、图像标题生成[16]、问答[17]等。当与LSTM结合时，其输出，语义向量$\boldsymbol{c}$为LSTM的隐藏层状态$\boldsymbol{h}_t$在不同时间步上的加权和：

$$\boldsymbol{c} = \sum_{t=1}^{T'} \alpha_t \boldsymbol{h}_t \tag{2.7}$$

每个隐藏层状态$\boldsymbol{h}_t$的权重α_t计算如下：

$$\alpha_t = \frac{exp(e_t)}{\sum_{t=1}^{T'} exp(e_t)} \tag{2.8}$$

其中e_t可以通过一个以$\boldsymbol{h}_t$作为输入的可训练的全连接层获得。

此外，为了避免在LSTM训练中容易出现的梯度消失或爆炸问题，我们首先在输入上进行了步长为2的一维卷积，以期减少序列在时间维上的长度。这样既可以大大减少训练的时间，也可以缓解梯度消失或爆炸的现象。

2.2.6 网络训练

网络的训练目标是最小化交叉熵损失函数。我们使用了Adam优化器[49]训练模

型，学习率为10^{-3}。每在训练集上完整训练一遍，就会对学习率应用一次0.95的衰减系数。批大小batch size设置为128。为了减少过拟合，在全连接层前面使用了失活率为0.5 的dropout。此外，使用早停来决定何时停止模型的训练。即，当模型在一个随机选取的验证集上的正确率不再提升时，停止训练。本文模型通过tensorflow实现，并在单个Nvidia GeForce GTX 1080 GPU上进行训练。需要注意的是，所提出模型的所有超参数均由经验给出，没有进行精细调整。也许对超参数进行网格或随机搜索可以进一步的提高性能。

2.2.7 类别不平衡

当在类别不平衡的数据上训练模型时，得到的模型通常会倾向于将样本预测为多数类，而不是样本数更少的少数类，例如S1睡眠状态。因为在扩展的Sleep-EDF数据集中存在严重的类别不平衡（见表 2.1），我们在损失函数中加入了一个类别不平衡权重用于重新调整预测错误的权重，这也被称为代价敏感学习[43]。整体的原则是根据对应类别的样本数量，样本数越少，给与其的权重越大。因此，对于每一个类别，其权重计算如下：

$$w_l = \frac{max\{N_l\}_{l=1}^{L}}{N_l} \tag{2.9}$$

其中N_l为类别l的样本数，L为类别数。

这样做的初衷是希望模型对样本数量少的类别给与同样多的注意（具有更多样本的类别在模型训练过程中，会被模型更多的看到），确保每个类别都具有相似的预测正确率（召回率），而不是仅仅关注于总的正确率。除了R&M标准，我们也有根据AASM手册合并状态S3和S4，得到一个五分类任务。更进一步的，我们合并S1和S2，得到一个更简单的任务，以期在不同的任务难度下，评价我们提出的模型。

2.3 实验和结果

本章节将介绍模型性能的评价指标和实验的具体设置。

2.3.1 实验设置

为了评价本文提出的模型性能，我们使用召回率（recall，RE），其也被称为敏感度，总的正确率（ACC）和Cohen's Kappa系数（Kappa）作为性能指标。其次假设一个极端情况，对于一个二分类问题，其中90%为正类，10%为负类。简单地将所有样本均预测成正类即可获得90% 的正确率，此时所有的负类均预测错误，这明显是不科学的。考虑到Sleep-EDF数据集中存在严重的类别不平衡，我们提出了一个类

别不平衡指标，即平衡分类正确率（Balanced Classification Accuracy，BCA）。RE、ACC和BCA的计算方式如下：

$$RE_l = \frac{TP_l}{N_l} \tag{2.10}$$

$$ACC = \frac{\sum_{l=1}^{L} TP_l}{N} \tag{2.11}$$

$$BCA = \frac{1}{L}\sum_{l=1}^{L}\frac{TP_l}{N_l} = \frac{\sum_{l=1}^{L} RE_l}{L} \tag{2.12}$$

其中TP_l代表类别 l 被预测正确的样本数量（true positives），L为类别数，N 为总的样本数，N_l为类别 l 中的样本数。

实际上RE_l代表了类别为l的样本中，被预测正确的比例；而ACC代表所有的样本中，被预测正确的比例。仅增加ACC会趋向于将难以明确预测的样本预测为具有更多样本的多数类。我们希望所有的类别都具有相似的预测正确率（即召回率），因此我们使用了BCA作为指标，即不同类别召回率的均值。

此外，由于不同被试的EEG信号存在明显的差别，为了证明我们提出的方法可以在被试内和被试间均可学得一致的优良特征，可以用于即插即用的系统，我们对所有的实验都计划使用留一被试交叉验证。即对于每一次训练，将某一个被试作为测试集，剩余的数据作为训练集和验证集。这个过程会被执行多次，使得每个被试都能够轮流做为一次测试集，取不同被试的平均值作为最终的结果。

2.3.2 不同CNN模型的结果

本节中，在R&K标准的六分类任务上，比较了不同CNN特征提取组件的性能，平均ACC和BCA结果见表 2.2所示。

表 2.2 不同CNN模型在19个被试上的平均性能

特征提取组件	ACC	BCA
A-CNN	87.37	68.80
C-CNN	**88.19**	**70.95**

从表格中可以看出，由于良好的结构设计，基准组件A-CNN已经获得了比较高的分类正确率。使用小感受野卷积层堆叠组成的C-CNN在ACC和BCA均取得了进一步的提升，这主要得益于它更深的网络层数以及更强的非线性拟合能力。此外C-CNN的参数量更少，在计算代价和模型性能之间达到了比较好的平衡。

2.3.3 类别不平衡权重的影响

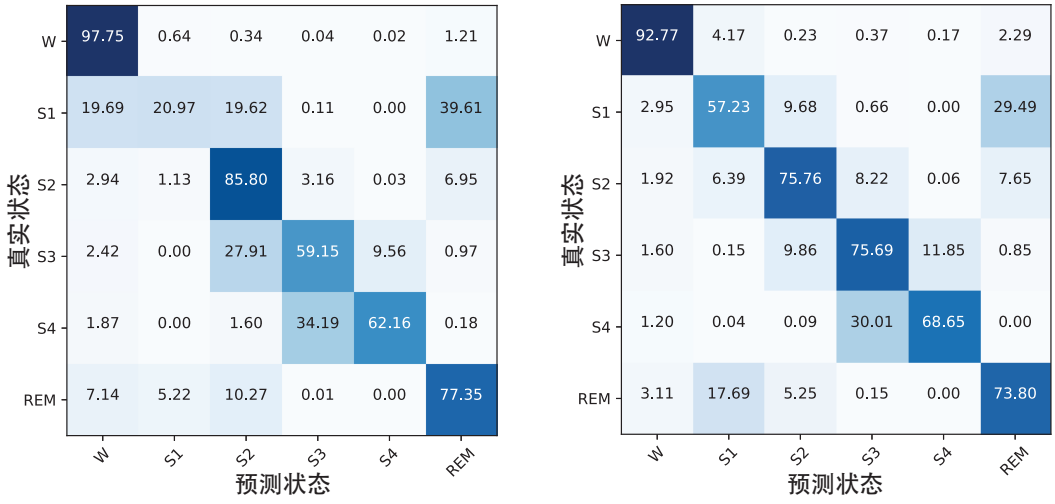
在该小节中，我们探索了类别不平衡权重对性能指标的影响。这个是一个六

分类的任务，使用效果更好的C-CNN作为特征提取组件。在这两个实验中，唯一的区别为是否在损失函数中使用类别不平衡权重。具体结果见表 2.3。其中$Mean$的结果为不同被试指标的均值，而$Summary$指的是将所有被试的预测结果组合在一起之后，再计算的结果。需要注意的是，由于存在一个被试不具有S4状态，这导致了$Mean$ 的BCA会稍低于$Summary$。

表 2.3 类别不平衡权重的作用

六分类	$Mean$		$Summary$		
	ACC	BCA	ACC	BCA	Kappa
加权重	88.19	**70.95**	88.24	**73.21**	77.44
无权重	**90.94**	67.94	**90.98**	70.18	**81.97**

从实验比较中可以发现，类别不平衡权重对ACC和BCA具有明显的影响。在损失函数中应用权重后，BCA指标获得了显著增加，同时ACC和Kappa 略有降低，这主要是由于具有大多数样本的多数类的预测正确率有所降低导致的。当数据集严重类别不平衡时，即使某些类只具有少数的样本，我们也希望它和其它的类别都具有相似的召回率（即更高的BCA）。



(a) 无权重　　(b) 加权重

图 2-7 六分类任务下C-CNN模型的混淆矩阵

此外，为了更好地理解所有睡眠状态的召回率和不同状态之间互相错误分类的具体情况，我们将不同被试的预测结果进行整合，计算了预测结果的混淆矩阵，见图 2-7。当未在损失函数中采用类别不平衡权重时，可以从图 2-7(a) 中发现，相邻的睡眠状态之间很容易被误分类错误，且样本数更少的睡眠状态会被误

分类成样本数更多的相邻状态。例如接近20%的S1状态被分别预测成样本数更多的W和S2，28%的S3则被误分类成了S2。此外，S1的预测准确率非常低，只有21%，几乎40%的S1被误分类成了REM（REM状态介于M和S1之间）。不同睡眠状态的召回率差距悬殊。在应用类别不平衡权重后，从图 2-7(b) 中可以看到该状况得到了缓解。在多数类W和S2 预测准确率稍微降低的代价下，S1 和S3的召回率得到了显著提高。因此，类别不平衡权重的使用可以使得不同类别的召回率更加平衡。在一些特殊场景中，例如希望S1 状态的预测尽可能准确，可以在损失函数中给S1样本更大的权重。需要注意的是，即使采用了权重，在S1和REM状态以及S3和S4之间仍然存在比较大量的误分类，表明它们在某种程度上是比较相似的。这也符合现实中R&M标准的情况，因此后面的AASM指南将S3和S4合并成了一个睡眠状态。

2.3.4 C-CNN和Attention模型的比较

卷积神经网络在图像处理上具有良好的表现，而循环神经网络则更擅长于处理时间序列任务。在本节中我们进一步地比较了C-CNN和Attention特征提取组件在多个任务的性能，具体结果展示在表 2.4 中。

表 2.4 C-CNN和Attention模型的比较

	六分类		五分类		四分类	
Model	ACC	BCA	ACC	BCA	ACC	BCA
C-CNN	88.19	70.95	90.23	77.09	91.03	85.94
Attention	**88.43**	**73.58**	**90.27**	**79.50**	**91.91**	**87.56**

可以看到Attention模型在ACC和BCA总是均好于C-CNN，这意味着它学得了更好的特征表达，能很好地描述数据内在的规律和模式，说明了Attention模型的有效性。随着任务变简单，即类别数量变少时，Attention模型的在BCA指标上的优势也在逐渐减少。此外，需要注意的是，由于LSTM本身的特性，无法在工程上并行实现，Attention模型需要的训练时间会多于C-CNN。

图 2-8 中展示了在六分类任务下，C-CNN和Attention模型在19个被试上的性能表现。从结果中可以看出，即使不同的被试之间的EEG信号可能存在比较明显的差距，我们的模型仍然可以学得对被试内和被试间的差异保持一致的表达，在所有的被试上均表现良好。

2.3.5 与现有方法的对比

我们也将本文提出的方法和现有的研究进行比较，并通过实验证明我们方法的优越性。具体结果见表 2.5。

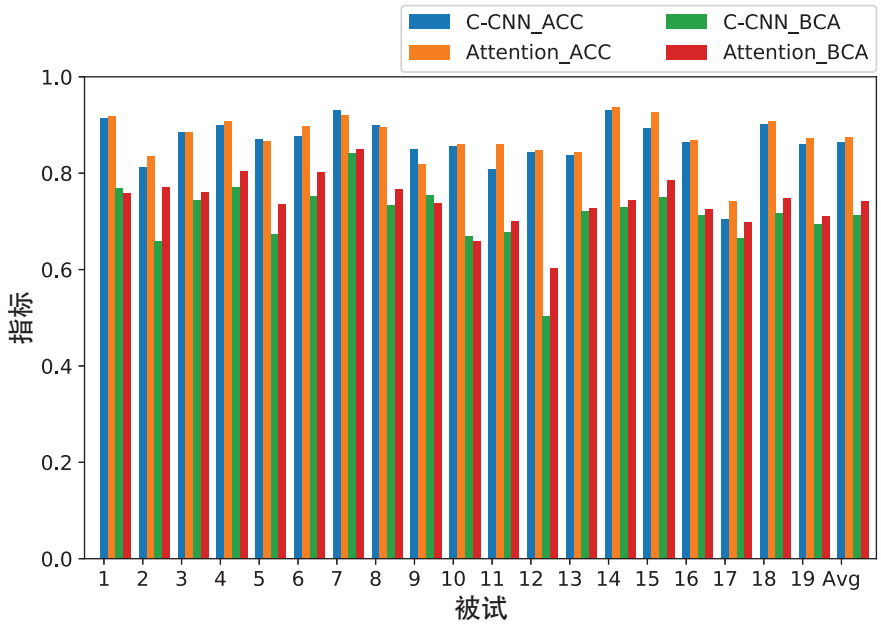


图 2-8 六分类任务下不同被试上的指标

Boostani等人的综述研究[30]在扩展Sleep-EDF数据集上比较了五个基于人工特征的方法[29,50–53]。其中文献[29]获得其最好的性能，该方法使用连续小波变换熵作为特征，使用随机森林作为分类器，获得87.06%的ACC和71.10%的BCA（BCA基于文献中给出的混淆矩阵计算）。和基于人工特征的传统方法相比，我们提出的C-CNN提供了一套端到端的解决方案，且几乎不需要先验知识。对于五分类任务，分别在ACC和BCA上取得了3.17% (从87.06%到90.23%)和6.12% (从71.10%到77.22%)的提升。此外，在不使用类别不平衡权重的情况下，我们的方法可以将ACC进一步提高到91.88%，且提供了可以接受的BCA（73.83%）。

表 2.5 和现有方法进行对比

		六分类					五分类				
		Mean[1]		*Summary*[2]			*Mean*[1]		*Summary*[2]		
		ACC	BCA	ACC	BCA	Kappa	ACC	BCA	ACC	BCA	Kappa
本文提出C-CNN	加权重	88.19	70.95	88.24	73.21	77.44	90.23	77.09	90.26	77.22	81.07
	无权重	**90.94**	67.94	**90.98**	70.18	**81.97**	**91.84**	73.67	**91.88**	73.83	**83.71**
本文提出Attention	加权重	88.43	**73.58**	88.47	**75.89**	77.99	90.27	**79.50**	90.30	**79.77**	81.26
Review[30]							87.06		71.10		
DeepSleepNet[38]							90.01	79.39	90.04	79.75	80.86

[1] *Mean*为不同被试结果的均值。
[2] *Summary*为将所有被试预测结果整合后，计算得到的结果。

DeepSLeepNet[38]是目前最优的睡眠状态检测模型。该模型分成两部分，第一部分使用CNN进行表达的学习，第二部分使用了相邻的多个EEG epoch和双向LSTM去学习不同睡眠状态的转换规则。训练也分成两步，在通过重采样获得的类别平衡数据集上，先只对CNN部分进行有监督的预训练；然后基于序列训练集，微调整个模型，即使用相邻的EEG epoch训练双向LSTM部分。为了比较，我们在本文的实验中复现了该方法的第一部分。与DeepSleepNet相比，我们的模型更加简单，不需要特殊的处理。C-CNN获得了稍微更高的ACC以及差不多的BCA，且需要的训练时间大概只需要DeepSleepNet的一半；而Attention模型在所有指标上均超过了DeepSleepNet，展现了我们所提出方法的优越性。

2.3.6 训练样本数量对C-CNN模型性能的影响

此外，为了验证训练集样本数量对所提出的C-CNN模型性能的影响，我们在六分类任务下进行了一个粗略的实验。将数据集完全打乱之后，将其中20%的数据作为测试集，并在整个实验过程中保持不变。然后从剩余数据中获取不同比例的数据作为训练集，并将训练集中的10%作为验证集。每次训练重复10次，取在测试集上的平均表现作为最终结果。不同训练集大小下的实验结果如图 2-9 所示。

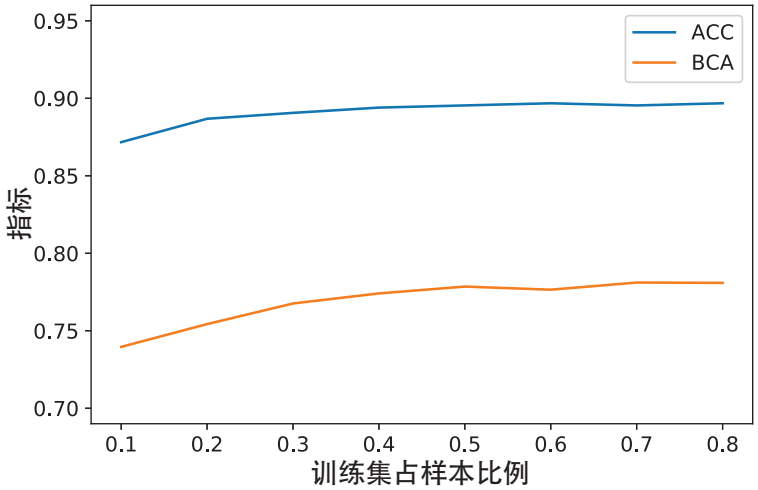


图 2-9 训练集大小对性能的影响

我们注意到，即使使用了80%的数据作为训练集，也仅比只使用10%数据时的ACC和BCA分别高3%和4%。这意味着，即使只有少量数据时，我们提出的C-CNN模型也可以实现令人满意的性能。当然，在另一方面，更多数据也可以进一步的提高模型性能。

2.4 本章小结

本章对睡眠状态检测的方法进行了分析，从自然语言处理中获取灵感，提出了两种新颖的深度学习模型用于睡眠状态检测：C-CNN使用了小感受野的卷积核和多层次的特征，在效率和性能间取得比较好的权衡；基于注意力机制和双向LSTM的模型可以获得更好的性能，领先于现有的方法。此外，通过将代价敏感学习整合到模型训练过程中，解决了睡眠状态中存在的严重类别不平衡问题，保证每个睡眠状态均具有较高的召回率，取得了更好的平衡分类正确率。在扩展的Sleep-EDF上的实验验证了本章所提出方法的有效性，优于现有的基于人工特征的方法和目前最优的深度学习模型。

3 基于MRI和CNN的鼻咽癌决策支持系统

鼻咽癌是中国东南部、台湾、香港、马来西亚和新加坡等地区最常见的头颈癌。在本章中，作者为鼻咽癌的诊断提供了一套可视化的决策支持系统。其首先通过开操作和大津阈值法对MRI切片进行自适应的分割和裁剪，提取切片中有效的大脑部分，同时解决了跨分辨率跨系统的问题；然后使用修改后的残差网络对MRI的不同切片进行处理，利用提出的可视化方法快速定位恶性肿瘤可能存在的切片和区域；最后对所有MRI切片的高层次特征进行整合，给出最终的鼻咽癌阳性概率。整套系统在为医生提供可疑肿瘤位置标注的同时，鼻咽癌阳性诊断的ROC曲线下面积AUC指标也达到0.994。

3.1 引言和相关工作

鼻咽癌（Nasopharyngeal Carcinoma，NPC）是一种发生在鼻咽腔或者上咽喉部的恶性肿瘤，是中国东南部、台湾、香港、马来西亚和新加坡等地区最常见的头颈癌[54]。在中国，NPC的发病率约为每10万人中有40人，香港为25人，马来西亚华人中为27人。而在美国和欧洲，发病率仅为10万分之1[55,56]。通常使用化疗或者放射疗法对鼻咽癌进行治疗。医学影像可以在临床诊断和治疗中起着重要作用，提供肿瘤区域、位置、大小、区域强度、体积以及病情程度等有用信息。常用的有内窥镜、计算机发射断层扫描（Computed Tomography，CT）、正电子断层扫描（Positron Emission Tomography，PET）和磁共振成像（Magnetic Resonance Imaging, MRI）等影像技术。在临床中，NPC的影像诊断主要还是人为的，完全取决于医生和放射科专家的经验和主观判断。然而，NPC具有复杂且不规则的结构表现，即使专家也难以可靠地进行诊断。

目前有一些研究利用机器学习技术来辅助NPC的诊断。Mohammed等人[57]提出了一种基于内窥镜的辅助诊断技术，他们利用局部二值模式、灰度共生矩阵、方向梯度直方图和分型维度等特征，训练了一个多层的神经网络用于诊断NPC。Chuang等人[58]成功地将CNN应用于NPC肿瘤病理切片的活检中，取得了比较好的结果。Wu等人[59]则是利用PET-CT双模态数据，提出了一种分阶段的NPC诊断方法。其首先基于PET和CT的图像特征以及解剖先验知识提取候选病变区域，然后使用支持向量机进行NPC 的分类。Zhao等人[60]提出了一种具有辅助路径的全卷积神经网络，实现了在双模态PET-CT图像上的NPC肿瘤区域分割，为放射治疗中的定位提供了极大的方便。

在众多的医学影像技术中，MRI具有非侵入、高效的优点。MRI可以展现鼻咽中的解剖学结构，包括咽隐窝和鼻咽深处的组织细节。它对软组织、咽后淋巴结

的转移、颅底侵犯、神经周围浸润等方面的信号变化比较敏感，且可以在临床中指导内镜下可疑部位的活检[61]。MRI在其它医疗领域也得到了广泛应用[18,19,62–64]。Korolev等人[18]基于3D MRI数据，将扩展到3D的VGGNet[31]和残差网络[4]应用到阿尔茨海默病和轻度认知障碍的诊断中。Pinaya等人[62]则利用深度置信网络从MRI图像中提取特征，用于诊断精神分裂症病人。此外，MRI还被应用于注意缺陷多动障碍[63]、中风[64]和多发性硬化症[19]等问题中。

已经有不少研究利用MRI和机器学习技术进行NPC的辅助治疗，但大多数集中在利用MRI的高空间分辨率特性进行NPC肿瘤区域的分割[65–67]，用于放射疗法的定位，几乎没有文献使用MRI直接进行NPC的诊断。先进行NPC的诊断，然后再进行恶性肿瘤区域的分割才是一个更符合直觉的流程。Wu等人[68]做了尝试，他们先是使用Unsharp Mask锐化算法处理MRI图像，以增强图像边缘；为了减少计算时间并提高效率，需要由医生在图像中指定感兴趣的鼻咽区域；在那之后使用直方图均衡去除噪声和大津阈值法自适应的取阈值分割提取鼻咽肿瘤区域；最后根据肥大和对称分布等特性，从肿瘤区域提取纹理和几何特征，训练基于神经模糊的Adaboost模型以识别是良性肿瘤还是恶性肿瘤。

上述流程繁琐，且需要很强的医学相关先验知识。本文利用深度学习的优点，基于MRI提出了一套端到端的鼻咽癌可视化辅助诊断系统。只需要少量的预处理，也可以用在不同分辨率的MRI设备中。该系统除了根据被试的MRI数据给出患NPC的概率外，其可视化技术也可以协助医生快速定位恶性肿瘤存在的MRI切片和区域，显著提高了诊断效率。

3.2 方法

在本节中，将介绍使用的数据集，以及本章提出的鼻咽癌辅助决策支持系统的处理流程。

3.2.1 数据集

本章使用的MRI数据集来自华中科技大学同济医学院附属协和医院，并由一名专业的医生进行标注。它包括了526名被试的数据，其中326名被试患有鼻咽癌，其余200名为正常被试。MRI数据一般沿着轴状位进行多个切片的扫描（一张切片对应一张图像），最终组成整个大脑的3D结构扫描图。图 3-1 中展示了某症状比较明显的鼻咽癌患者的所有轴状位MRI图像，其中间几张图像可以发现较为明显的组织肥大，鼻咽结构受到了挤压。由于鼻咽癌的判定对轴向扫描密度要求不高，因此该数据相邻切片间的空间间隔比较大。此外，得到的数据来源于多个不同型号的MRI设备，具有208×256 到640×640 等多种分辨率，不同被试具有的切片数量在14到35之间，主要集中在15到21张。在本文中，仅使用了MRI的轴状位T1结构像。

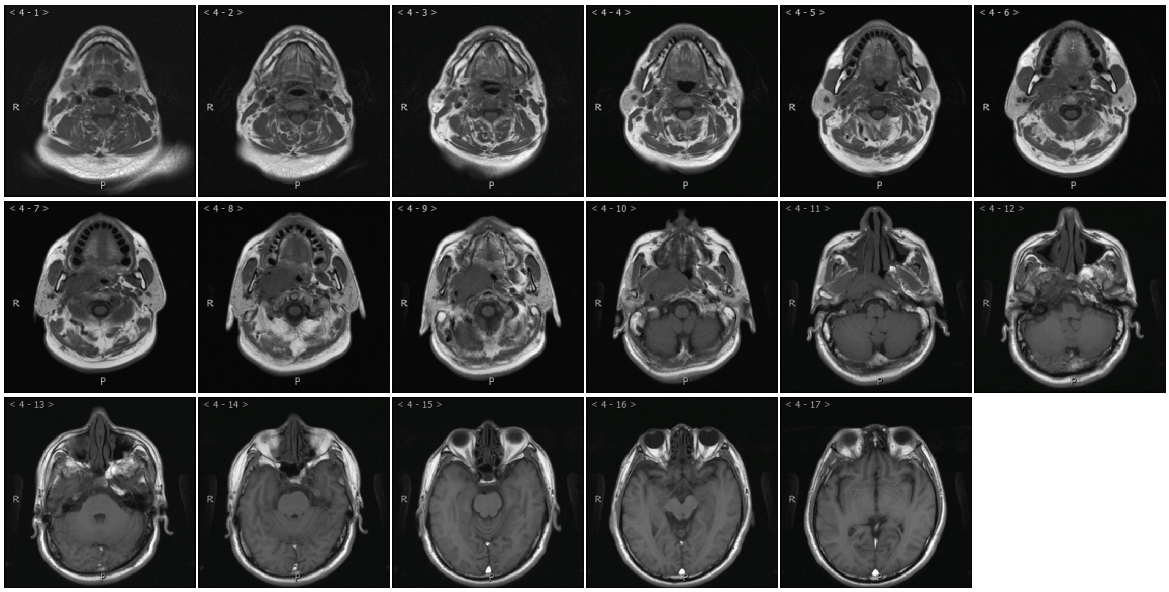


图 3-1 某鼻咽癌患者的轴状位MRI数据

考虑到数据集中MRI轴向扫描的密度比较低，不适用于直接使用3D卷积进行处理，因此分成了两个阶段来完成鼻咽癌的诊断。首先是判断一张MRI图像中是否具有鼻咽癌区域；然后在前者的基础上，根据一个被试的所有图像综合判断该被试是否患有鼻咽癌。在进行数据标注时，对于一个患者的所有MRI图像，可以观察出鼻咽癌的图像被标注为正样本，其余的标注为负样本；对于正常被试，所有的图像均被标注为负样本。总共得到9708张MRI图像，其中正样本为1470张，负样本为8238，大致比例为1：5.6（由于该标注分成了多次标注，可能存在判断标准不一致的情况，所以也许有一定比例的负样本被标注成了正样本）。

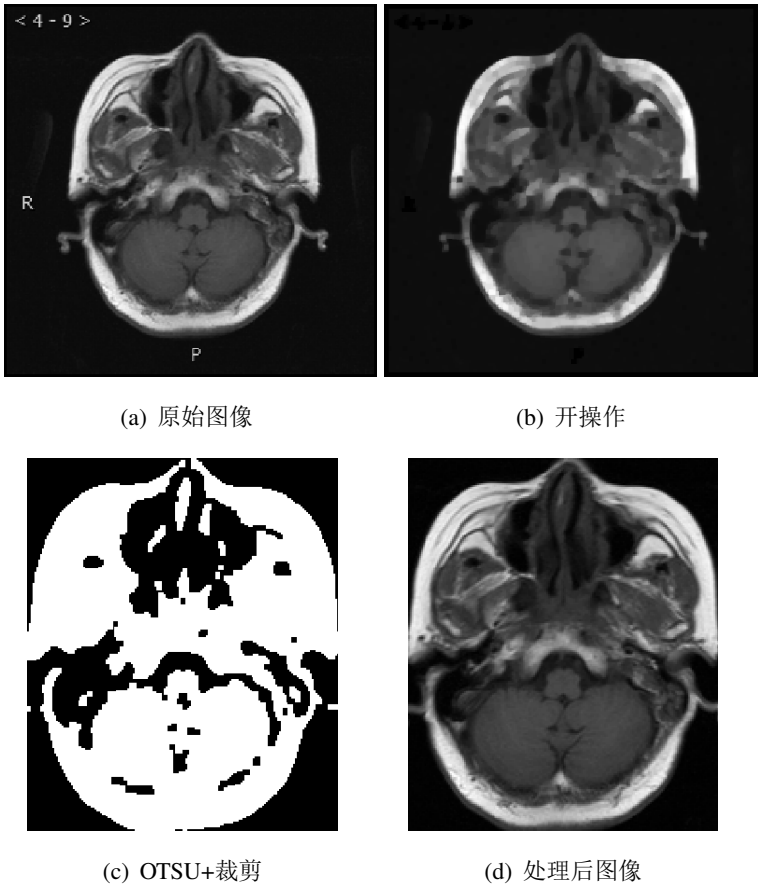
3.2.2 预处理

从图 3-1 中的MRI数据中，我们可以观察到原始图像具有一些无用的编号和字母，而且图像的边缘存在着大量的黑色区域，这些区域无法提供任何有用的信息。为了提高图像中有意义信息的占比，为之后的缩放操作变相提高空间分辨率，我们对其进行了一定的预处理。

首先，对图像进行了开操作（即先腐蚀后膨胀）以去除那些无用的编号和字幕，也去掉了一些细小的组织；其次，我们使用大津阈值法（OTSU）算法[69]对图像进行了二值化，其自适应的选择阈值对图像中的前后景进行分割，使得前后景之间的类间方差最大；最后，对二值化后的图像进行了连通域分析，裁剪以保留具有最大连通域的区域，处理完成的最终结果见图 3-2(d)。从图 3-2 中可以看出，图 3-2(a) 所示的原始图像，在经过一系列的预处理后，原始MRI图像中的编号、字母以及大量

黑边已经被完全去除，只保留了中间有用的大脑结构区域，显著提高了MRI图像中有效信息的占比。

此外，虽然不同MRI设备扫描数据的边缘黑色区域大小是不同的，但我们在这里采用的预处理是自适应的。其可以自适应地处理这种差异，保证不同MRI设备的数据在送入后续算法进行处理时是一致的，解决了跨分辨率跨系统的问题。



(a) 原始图像　　(b) 开操作

(c) OTSU+裁剪　　(d) 处理后图像

图 3-2 MRI图片预处理流程示意

3.2.3 图像级别模型

残差网络ResNet[4]由何凯明等人在2016年提出，在图像处理中得到了广泛的应用。ResNet考虑到深层网络训练难，出现退化的现象，巧妙地应用恒等映射，为梯度的反向传播提供了直接的通路。在本文中，我们采用了ResNet作为主干网络。

考虑到鼻咽癌的数据量偏少，为了避免过拟合，我们使用了参数量较少的ResNet18，并根据可视化的需求进行了一些改动。具体结构如表 3.1 所示，除了conv_fc外，所有的卷积层后面都跟随着BN层[47]和ReLU激活函数。改动的具体细节为：1）网络输入尺寸由原来的224×224修改为225×225，以获得更好的边缘对

齐和空间分辨率（NPC的判定主要依赖于结构信息）；2）加入了dropout2d防止过拟合，与普通的dropout相比，dropout2d会以一定概率随机地抛弃一些通道，更有利于消除特征的冗余；3）直接使用1×1卷积预测结果，然后使用global average pool对预测结果进行整合，经过sigmoid输出最终的概率。这样的组合有利于进行可视化，可以粗略知道输入图像中哪部分对最终预测的贡献最大。

表 3.1 修改后的ResNet18

layer name	18-layer	output size
conv1	7×7, 64, stride 2	$113 \times 113 \times 64$
res2	3×3 max pool, stride 2	$57 \times 57 \times 64$
	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 2$	
res3	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 2$	$29 \times 29 \times 128$
res4	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 2$	$15 \times 15 \times 128$
res5	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix} \times 2$	$7 \times 7 \times 2048$
dropout2d	drop rate 0.5	$8 \times 8 \times 512$
conv_fc	1×1, 1	$8 \times 8 \times 1$
global average pool + sigmoid		$1 \times 1 \times 1$

3.2.4 可视化

图 3-3 为图像级别模型对输入图像进行处理的示意图。为了判断输入图像中的哪些区域对模型预测起到重要作用，我们提出了一种可视化的方法，这样有利于迅速定位可能存在鼻咽癌的区域。该可视化方法主要思路来自2016年Zhou等人的工作[70]，我们做了一些修改使得推导更加严谨和便于理解。

对于给定的输入，令$\boldsymbol{f}$代表ResNet18网络中res5输出的特征图，$\boldsymbol{f_c}(x,y)$代表在空间位置(x,y)处通道c的激活。则对于位置(x,y)，经conv_fc层处理后的输出$\mathbf{F}(x,y) = \sum_c \boldsymbol{w}_c \boldsymbol{f}_c(x,y) + b$，其中$\boldsymbol{w}_c$代表conv_fc层对通道$c$的权重，$b$为偏置。经过全局平均池化后得到$sigmoid$ 的输入$\boldsymbol{S} = \frac{1}{N}\sum_{x,y}\mathbf{F}(x,y)$，其中$N$是特征图$\boldsymbol{f}$中的cell数量。因此，

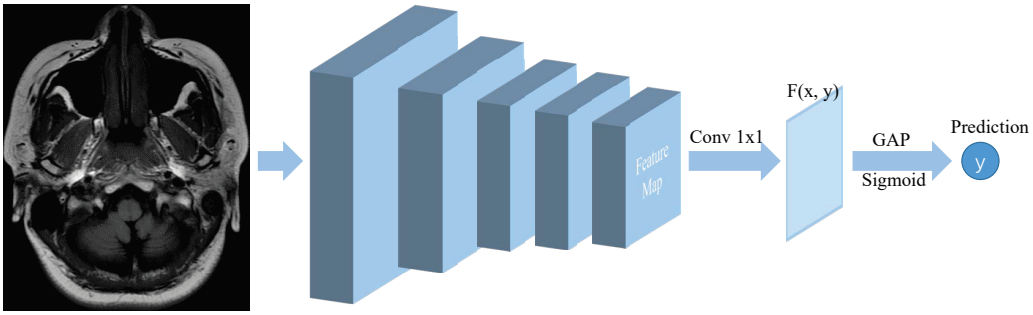


图 3-3 图像级别模型

最终的预测概率$P = sigmoid(S)$。

对上述结果整理得：

$$P = sigmoid\left(\frac{1}{N}\sum_{x,y}\mathbf{F}(x,y)\right) \tag{3.1}$$

$$= sigmoid\left(\frac{1}{N}\sum_{x,y}\left(\sum_{c}\mathbf{w}_c\mathbf{f}_c(x,y)+b\right)\right) \tag{3.2}$$

因此，对于给定的一张图像，我们可以明确的知道$\mathbf{F}(x,y)$代表了位置(x,y)对最终预测结果的贡献度。直觉上，我们可以认为每个cell都是由其感受野对应的视觉模式所激活得。通将**F**上采样到原始的图像分辨率，就可以找出哪些区域对最终预测结果起到的贡献最大。在图 3-4 中，展示了某个被试部分MRI 图像的可视化结果，其中颜色从深蓝到红代表对预测概率为阳性的贡献越来越大。

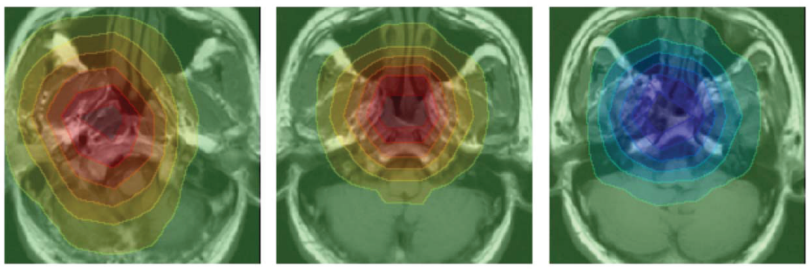


图 3-4 某被试部分MRI图像的可视化结果

在实际操作中，鼻咽癌的判定主要还是依赖于鼻咽区域的结构信息。从图 3-4 可视化的结果中，我们可以发现对预测结果起着主要作用是大脑鼻咽部的某个区域，即图中的红色区域。该可视化图可以在医生判定鼻咽癌时提供辅助支持，快速定位可能存在异常的区域，提高诊断效率。

3.2.5 被试级别模型

由于每个被试的MRI数据包含多张图像，自然地，我们可以整合多张图像的信息，给出更为可靠的结果。主要的思路是使用一个CNN去提取这些图像的特征，该CNN在不同的图像上共享，之后可以通过不同的整合策略对这些特征进行处理，利用整合后的特征输出最终的预测结果。图 3-5 中，展示了我们设计的几种特征融合策略。在这里，我们使用之前的图像级别模型作为特征提取器，取res5的输出作为特征，在图中用具有文字C的蓝色矩形表示。图中的GAP代表全局平均池化，1D max pool为一维最大池化，1D Conv为一维卷积，Attention为注意力机制，FC为全连接层和$sigmoid$函数，用于输出最最终的预测结果。

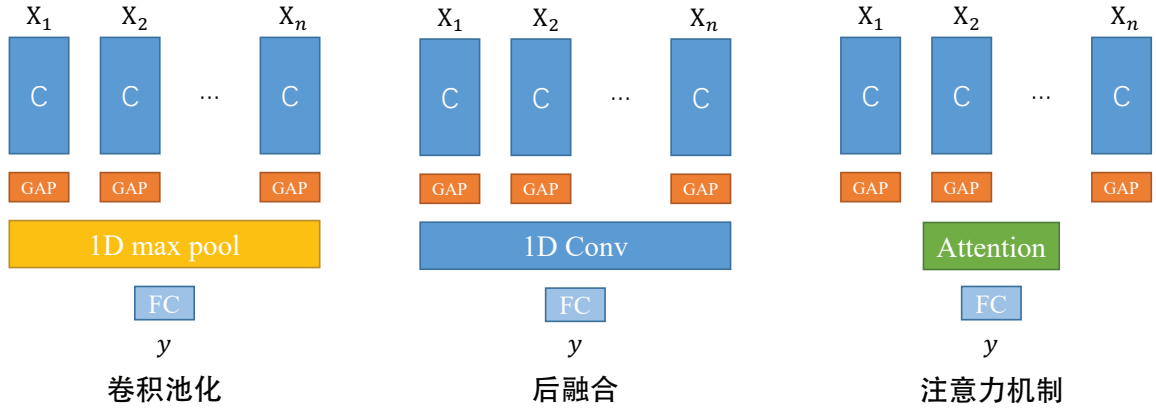


图 3-5 不同的特征融合策略

3.2.5.1 卷积池化

卷积池化方案利用了一维最大池化来整合不同MRI切片的特征。该池化层的步长为1，核大小为n（n为一个被试具有的MRI切片数）。使用一维最大池化背后的动机是，对于每一种特征（或通道），我们仅选取所有图像中的最大值作为特征的取值。该方案也有一定的缺点，因为只是单纯地取最大值，可能对输入过于敏感，容易造成误检，假阳率过高。

3.2.5.2 后融合

与卷积池化中只是简单地选取每个特征的最大值不同，在后融合中，我们使用了一维卷积去捕获相邻图像特征之间的空间信息（即 Z 轴）。卷积层背后的设计思想局部连接和权重共享，天然地适合处理这里的层次化特征。卷积核大小设置为5，步长取2，通道数为128。将卷积层的输出展开后，送入全连接层做最终的预测。

3.2.5.3 注意力机制

在对被试进行MRI扫描时，是沿着特定轴向按照某个方向一次性扫描的。这自然会导致每张图像包含的鼻咽区域不同，甚至某些图像完全不含鼻咽区域。因此虽然一个被试具有多张MRI图像，但每张图像所含的有用信息是不同的，对于不同的图像应当给予不同的权重。直观上的，我们可以对包含鼻咽区域的图像给予更大的权重。然而如何判断是否包含鼻咽区域，以及鼻咽区域的大小，也比较困难。除了根据人的先验设置权重外，我们也可以使用注意力机制。注意力机制已经在多种领域得到了广泛应用，如机器翻译[6]、图像标题生成[16]、视频问答[17]、睡眠状态检测[71]等。它可以根据不同特征，自动地计算应该赋予该特征的权重，将这些特征的加权和作为最终的特征表达。

令被试的第i张MRI图像$\boldsymbol{X_i}$经C和GAP后的特征为$\boldsymbol{h_i}$，经过注意力机制后，最终的特征表达$\boldsymbol{c}$是不同图像特征的加权和：

$$\boldsymbol{c} = \sum_{i=1}^{n} \alpha_i \boldsymbol{h}_i \tag{3.3}$$

每张图片的特征$\boldsymbol{h}_i$的注意力权重α_i由下式计算：

$$\alpha_i = \frac{exp(e_i)}{\sum_{i=1}^{n} exp(e_i)} \tag{3.4}$$

上式的e_i可以通过一个输入为$\boldsymbol{h}_i$的可训练网络得到。在本文中，我们使用了两层的全连接层。

3.2.6 训练和前向过程

数据划分：我们以被试为单位，使用分层采样，按照60%/20%/20%的比例将数据集划分成训练集、验证集和测试集。使用分层采样的原因是为了使得不同的数据集中正常被试和患病被试的比例保持基本一致，维持在0.615比1左右。我们将提出的算法在训练集上训练，在验证集上调整超参数和选择模型，最后在测试集上比较不同实验设置下模型的结果。

训练优化：由于我们使用$sigmoid$函数输出模型最终的预测概率，因此采用了二元交叉熵作为损失函数。使用Nesterov SGD[72] 对模型进行优化，momentum 为0.9，weight decay 为1e-3，训练50个epoch。此外，考虑到数据集存在较为严重的类别不平衡，我们在损失函数中对正样本加了权重。对于图像级别模型，batch size设置为32，学习率为5e-4，每20个epoch将学习率衰减10倍，正样本权重为5；对于被试级别模型，batch size为16，学习率为1e-4，无学习率衰减，正样本权重为0.5。需要注意的是，我们使用了图像级别模型训练后的参数，去初始化被试级别模型中的特征提取部分C，这有利于模型的性能和加速收敛。实验表明没有利用预训练参数进行初始化时，即使将被试级别模型训练200个epcoh，也未必收敛且性能一般。

数据增强：鉴于医疗数据采集的困难，使用数据增强可以很好的防止模型在小数据集上过拟合。在将样本送入模型之前，所有的图像均等比例将较短边缩放到256，在训练时会从缩放后的图像中，随机裁剪225 × 225的区域，然后以50%的概率随机进行水平和垂直翻转；在前向过程中（即测试时），仅执行缩放和中间裁剪225 × 225 的区域。对于被试级别模型，由于有多种采集设备，每个被试具有的MRI图像数量可能不同。因此在训练时，会随机地选取连续的n张图片构成一个样本；在前向过程时，则选择中间的连续n张图片。统计被试具有MRI图片数量的直方图后，将n设置为13。

在训练过程中，通过监测验证集上的AUC指标，使用了早停策略决定何时停止训练。我们进行了一些简单的超参数搜索，选择在验证集上表现最好的参数作为最后的模型，然后在测试集上进行测试，给出最终的实验结果。我们使用Pytorch实现所提出的模型，训练设备为一张英伟达GeForce GTX 1080显卡。

3.3 实验和结果

在本节中，将对模型性能的评价指标和实验的具体设置进行介绍。

3.3.1 性能指标

为了评价所提出模型的性能，我们使用了正确率ACC、sensitivity（敏感度，也被称为召回率和真阳率）和specificity（特异性，也称为真阴率）。此外考虑到数据集存在较为严重的类别不平衡，我们还采用了AUC和平衡分类正确率BCA作为指标。ACC、敏感度、特异性和BCA定义如下：

$$ACC = \frac{TP + TN}{N} \tag{3.5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.7}$$

$$BCA = \frac{Sensitivity + Specificity}{2} \tag{3.8}$$

其中TP、TN、FN、FP和N分别为真阳、真阴、假阴、假阳和总样本数。真阳代表鼻咽癌患者中有多少人被预测为鼻咽癌阳性；真阴代表健康被试中有多少人被预测为鼻咽癌阴性；假阴代表鼻咽癌患者被预测为鼻咽癌阴性的人数；假阳则是健康被试被预测为鼻咽癌阳性的人数。

AUC是医学问题中常用的二分类指标，其值为接受者操作特性曲线ROC下的面积。由于ROC曲线的横纵坐标轴分别是假阳率和真阳率，因此不会受到阳性和阴性

类别不平衡的影响。此外，我们希望患病和正常被试均具有较高的预测正确率，因此采用了BCA，它在这里是敏感度和特异性的均值。当ACC和BCA之和最大时，取此时的阈值计算所有指标。

3.3.2 图片级别模型结果

由于本文中使用的鼻咽癌MRI数据来源比较复杂，存在分辨率不同、图像中有大量空白和来自多种MRI设备等问题，因此在 3.2.2 节中，我们引入了一些自适应的预处理方法对MRI图像进行处理，以期解决这些问题。表 3.2 展现了是否进行预处理对图片级别模型性能的影响。

表 3.2 是否预处理对结果的影响

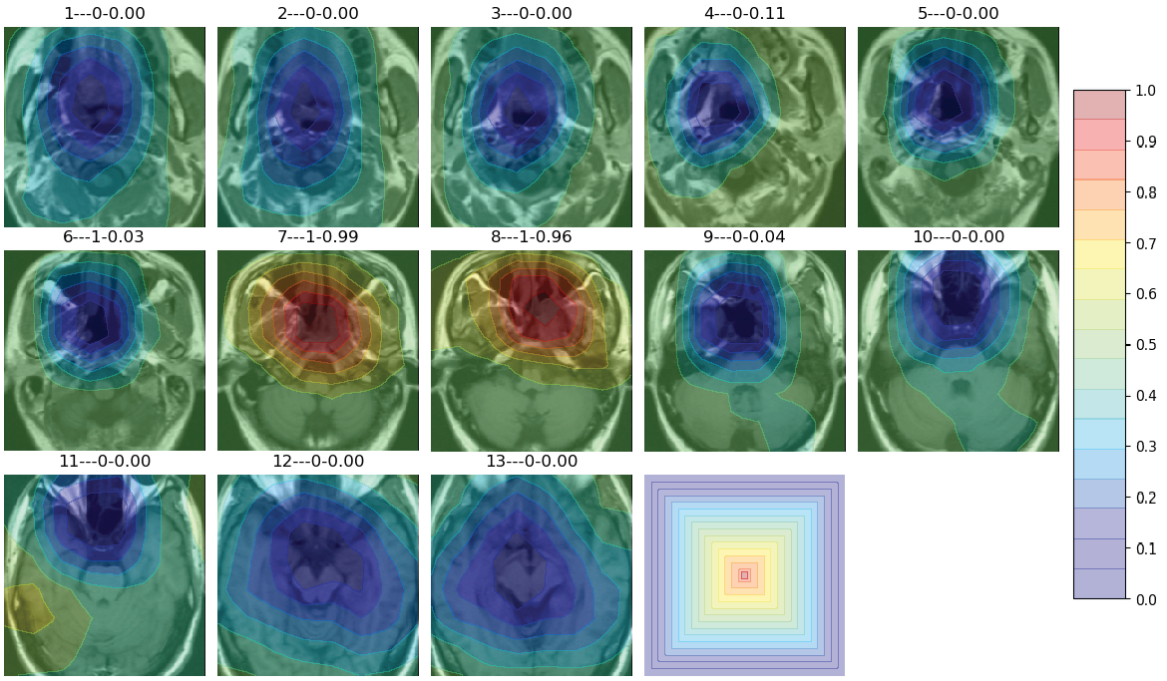
	AUC	ACC (%)	BCA (%)	Sensitivity (%)	Specificity (%)
无预处理	0.942	90.42	90.30	89.83	90.77
预处理	**0.972**	**92.76**	**92.44**	**91.99**	**92.90**

从表中可以看到，在未进行预处理时，模型已经取得了可以接受的结果。AUC达到了0.942，ACC、BCA、敏感度和特异性等指标均超过或接近90%。在通过开操作、OTSU等预处理操作后，性能得到了显著得提升。分别在AUC、ACC、BCA、敏感度和特异性上提高了0.03（从0.942到0.972）、2.3%（从90.42%到92.76%）、2.14%（从90.30%到92.44%）、2.16%（从91.99%到89.83%）和2.17%（从92.90%到90.77%）。

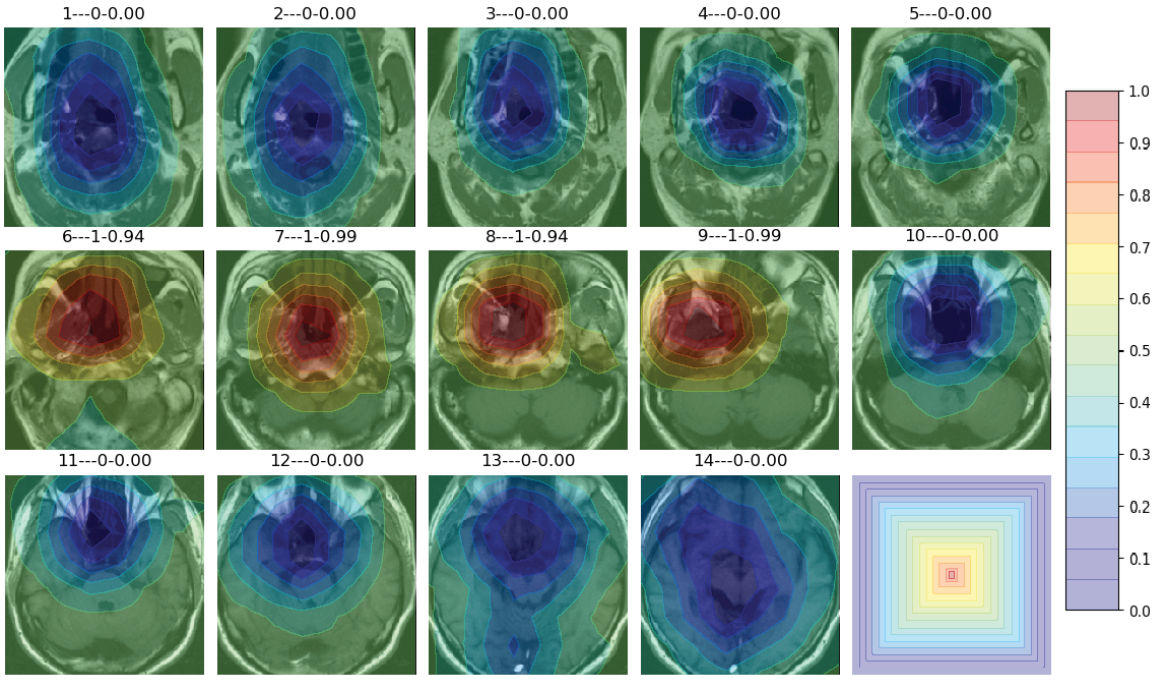
3.3.3 可视化结果

在 3.2.4 节中，我们提出了一种可视化技术，通过展示图像中不同位置对最终预测结果的贡献程度，可以快速定位可疑的鼻咽癌位置。图 3-6 展示了某两个被试所有MRI切片的图像级别模型预测概率和可视化结果。最右边的颜色条表示图中不同颜色对最终预测鼻咽癌阳性的贡献程度。每个小图的上方的数字表示“图像序号—真实标签(1为阳性)-模型预测为阳性概率”。

从这两位被试的可视化结果中可以发现：可观察出鼻咽癌阳性的MRI切片均被图像级别模型以高达94%以上的概率指出；对于图像中鼻咽癌阳性的可疑区域，也通过高贡献度的深红色等高线进行标注。对于其它的MRI图像，虽然同样包含鼻咽区域，但是在该鼻咽区域无法发现恶性肿瘤，所以图中是以低贡献度的深蓝色标注出来的。通过这样的可视化图，可以协助医生快速定位可疑的鼻咽癌区域，显著提高诊断效率。



(a)



(b)

图 3-6 某两个被试所有MRI图片的可视化结果

3.3.4 被试级别模型结果

在图像级别模型中，我们仅根据被试的某张图像，判断该图中是否可观察出鼻咽癌。但实际上，每个被试均具有多张大脑不同位置的MRI结构图，通过综合不同位置的特征信息，可以给出更为稳定和准确的预测结果。因此被试级别模型先通过图像级别模型对被试的不同位置的MRI图片提取高层次抽象特征，然后以一定的融合方法整合这些特征表达给出最终的预测结果。在 3.2.5 节中，我们提出了一些不同的特征融合策略。

值得注意的是，在我们的实验中发现，对图像级别模型进行预训练非常重要。利用图像级别模型训练后的参数，对被试级别模型的特征提取部分参数进行初始化，可以显著加快训练的收敛和提高模型最终的性能。

表 3.3 展示了未进行预训练时的各种特征融合方法在测试集上的表现。由于未进行预训练时各个模型收敛的比较缓慢，表中展示的是在训练集上训练200个epoch时的测试结果。注意力机制和后融合方法大概在150个epoch左右在验证集上收敛；卷积池化在75个epoch 时在训练集上的AUC已经达到了0.99，200个epoch时才在验证集上收敛，卷积池化融合方案的过拟合现象比较严重。从表 3.3 中可以看出，未预训练时，注意力机制融合特征的效果最好，后融合其次，卷积池化效果最差，与其它特征融合方法存在明显差距，因此后续实验未再考虑卷积池化。

表 3.3 未预训练时，被试级别模型结果

	AUC	ACC (%)	BCA (%)	Sensitivity (%)	Specificity (%)
卷积池化	0.952	86.21	85.08	90.02	80.14
后融合	0.959	89.52	88.45	**92.57**	84.33
注意力机制	**0.972**	**92.76**	**92.44**	90.05	**99.73**

表 3.4 展示了进行预训练做初始化时，采用不同特征融合方法下，被试级别模型的结果，以及和图像级别模型结果的对比。可以发现，当融合被试不同位置的MRI图像信息之后，与图像级别模型相比，被试级别模型的所有指标均有明显的提高。尤其是特异性指标得到了显著提升（+5%以上），这意味着显著提高了真阴率，可以降低正常人被误诊的风险。

表 3.4 预训练时，被试级别模型结果

	AUC	ACC (%)	BCA (%)	Sensitivity (%)	Specificity (%)
被试级别模型	0.972	92.76	92.44	91.99	92.90
后融合	0.991	94.85	95.43	92.44	**98.42**
注意力机制	**0.994**	**95.68**	**96.01**	**93.72**	98.30

在被试级别模型中，注意力机制特征融合方案的性能整体上稍高于后融合方法，仅在特异性上低0.12%，模型预测的敏感度和特异性分别达到了93.72%和98.30%。当有需要时，也可以通过降低$sigmoid$的阈值提高敏感度（即鼻咽癌患者确诊率），但特异性也会对应地降低，这需要做好权衡。

图 3-7 展示了使用注意力机制对不同MRI切片特征进行融合时，最终模型的接受者特性曲线（Receiver Operating Characteristic，ROC）。我们提出的模型表现的十分优异，ROC曲线下面积AUC达到了惊人的0.9942，与 1 仅差0.0048，这表明了我们所提出方法的有效性。

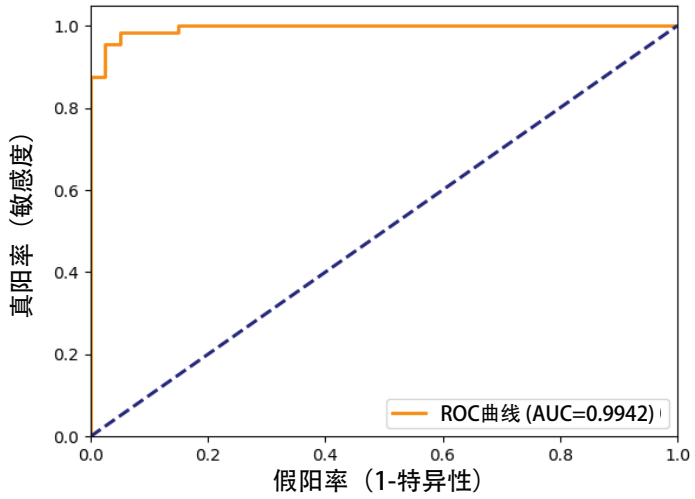


图 3-7 注意力机制模型的ROC曲线

3.4 本章小结

鼻咽癌在我国以及东南亚地区具有较高的发病率。在本章中，采用华中科技大学附属协和医院提供的MRI数据，为鼻咽癌的诊断提供了一套决策支持系统，用于辅助医生进行诊断。其首先通过开操作和大津阈值法对MRI切片进行自适应的分割和裁剪，提取切片中有效的大脑部分，同时解决了跨分辨率跨系统的问题；然后使用修改后的残差网络对MRI 的不同切片进行处理，利用提出的可视化方法快速定位恶性肿瘤可能存在的切片和区域；最后对所有切片的高层次特征进行整合，给出最终的鼻咽癌阳性概率，ROC曲线下面积AUC 指标达到0.994。此外，该系统还可以通过可视化技术标注疑似区域，能够显著降低医生的负担，提高诊断效率。

4 基于time-lapse和多任务学习的胚胎早期发育阶段分类

在进行体外人工受孕时，明确胚胎早期发育过程的不同阶段，可以为胚胎学家提供评估胚胎质量的宝贵信息，对受孕的成功至关重用。为了对胚胎的不同阶段进行准确地分类，本章提出了一种具有动态规划的多任务深度学习框架（MTDL-DP）。它首先基于视频中相邻帧具有大量互补信息的特性，利用多任务学习和相邻帧为时延视频的每一个图片帧生成多个预测结果；然后通过集成思想对这些预测结果进行整合，赋予当前帧一个胚胎发育阶段；最后使用动态规划进行后处理，优化整个视频的发育阶段序列，使得最终预测的发育阶段序列非递减，且搬运距离损失最小。通过本章提出的MTDL-DP算法，将胚胎早期发育阶段分类的精度提高了3.1%。

4.1 引言和相关工作

体外人工受孕（In-vitro Fertilization, IVF）[73–75]是治疗不孕症的常见技术。该过程涉及到收集卵泡、体外受精和体外培养。其中，胚胎的培养、选择和移植是IVF成功的关键步骤[76,77]。在胚胎发育过程中，胚胎的形态[78]和动力学特征[79]与移植的成功高度相关。

Time-lapse技术已经被广泛用于各种生殖医学中心，监测胚胎的培养过程。Time-lapse设备会以较短的时间间隔拍摄胚胎图片，实时记录胚胎的发育过程。因此，在该过程中，对每个胚胎均会产生大量的时间序列图片数据，组成了延时视频。在胚胎选择的最终阶段，胚胎学家会审查胚胎的整个发育过程，对它们进行评分和排序。部分使用time-lapse 设备进行的研究表明，通过分析人类胚胎卵裂早期的形态动力学特征，可以提高胚胎移植的成功率[80–84]，且这些特性对移植的最终结果具有统计学意义[79]。

目前分析time-lapse视频的方法还比较少[80,85–90]。由于time-lapse技术的固有缺陷，拍摄时不同高度的细胞会重叠在一起。当卵裂的细胞数量超过八个时，即使是经验丰富的胚胎学家，也难以在单张time-lapse图片中准确地计算出细胞数量。因此，大多数的研究均集中在胚胎的早期发育阶段。Wong等人[80]确认了几个关键参数，这些参数可以在time-lapse系统中准确预测四细胞阶段胚泡的形成，并且他们采用了基于顺序蒙特卡洛方法的概率模型来监控这些参数和跟踪细胞数量。Wang等人[85]提出了一种多层次的胚胎阶段分类方法，该方法结合人工设计和自动学得的胚胎特征，确定time-lapse视频中的细胞数量。Conaghan等人[86]使用了专门的图像分析软件Eeva（Early Embryo Viability Assessment，早期胚胎生存力评估）来跟踪一细胞到四细胞阶段的细胞分裂，此外Eeva可以通过暗场分析来提高图片对比度。他们的实验表明，Eeva Test可以显著提高胚胎学家的能力，识别出可发展为可用胚泡的胚

胎。还有很多其它的研究[91–94]也使用了Eeva软件进行胚胎的选择，但是没有提供使用Eeva的算法细节。Jonaitis等人[87]比较了神经网络、支持向量机和最近邻分类用于检测细胞分裂时间的性能。Khan等人[90]使用了深度卷积神经网络预测细胞的数量，此外也有通过语义分割提取time-lapse图片中的细胞区域[88]。Ng等人[89]结合了后融合网络和动态规划，预测胚胎发育阶段，并取得了比单帧模型更好的性能。

多任务学习已经被成功地应用于自然语言处理、音频识别和计算机视觉等领域中[95]。多任务学习潜在的思想是通过在相关的任务中共享表达，从而使得训练得到的模型可以具有更好的泛化能力。在time-lapse视频中，相邻图像帧之间通常会具有较强的相关性，其胚胎发育状态也是相似的。为了利用好这种互信息，本章提出了一种具有动态规划的多任务深度学习方法MTDL-DP（multi-task deep learning with dynamic programming）。它首先基于视频中相邻帧具有大量互补信息的特性，利用多任务学习和相邻帧的信息为时延视频的每个图像帧生成多个预测结果，再通过集成思想对这些预测进行整合，赋予当前帧一个胚胎发育阶段；然后使用动态规划优化整个视频的发育阶段序列，使得最后输出的发育阶段序列非递减，且搬运距离损失最小。通过提出的多任务集成和动态规划后处理，本文将胚胎发育阶段分类的精度提高了3.1%。

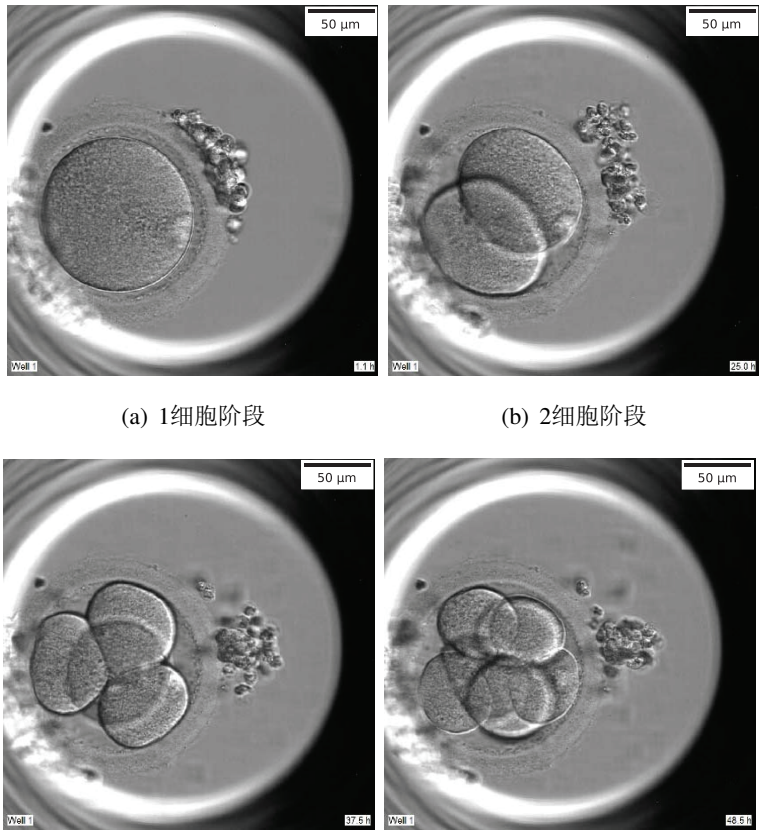
4.2 方法框架

本节将基于time-lapse视频，介绍用于胚胎早期发育阶段分类的四种框架。首先会描述本章中使用的数据集和基准模型框架，然后将基准模型扩展到*多对一*、*一对多*和*多对多*MTDL框架。

4.2.1 数据集

在本章的研究中，使用的数据集来自华中科技大学同济医学院生殖医学中心。它包含从孵化器中提取的170个time-lapse视频。使用的time-lapse显微镜设备为*Embryoscope+*，相邻两帧之间的时间间隔为10分钟。视频中的每一帧均为分辨率800×800的灰度图，图片的左下角为视频帧编号，右下角的时间标志，记录了受精后的时间，样例如图 4-1 所示。在显微镜视野中，胚胎被一些颗粒细胞包裹。帧的右上角为显示细胞尺寸的比例尺。视频在受精后的2个小时内开始记录，在受精后的140个小时结束。本文只使用了每个视频的前$N = 350$帧，并对这些帧人工标记了胚胎发育阶段。因此，该数据集总共具有$170 \times 350 = 59500$张标记的帧图片。

和该文献[89]一样，本文主要关注于胚胎发育早期的前六个阶段，其中包含了初始阶段（tStart），男性和女性原核的出现与消失（tPNf），以及分裂到2 到4+ 细胞的阶段（t2、t3、t4、t4+）。图 4-2 展示了数据集中不同胚胎发育阶段的图片数量，需要注意的是t3阶段在该数据中很少出现。



(a) 1细胞阶段　　(b) 2细胞阶段

(c) 4细胞阶段　　(d) 4+细胞阶段

图 4-1 Time-lapse延时视频部分帧示意

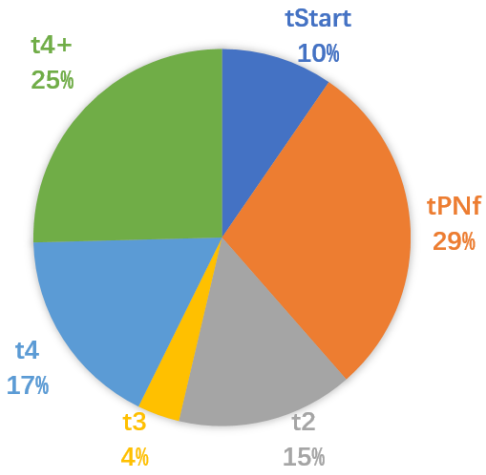


图 4-2 不同胚胎发育阶段的帧比例

4.2.2 一对一基准分类框架

令$\mathbf{x}_n$代表time-lapse视频中第n帧。对于图像分类，一个标准的*一对一*分类框架是学习一个映射：

$$f_0 : \mathbf{x}_n \mapsto y_n \in L, \tag{4.1}$$

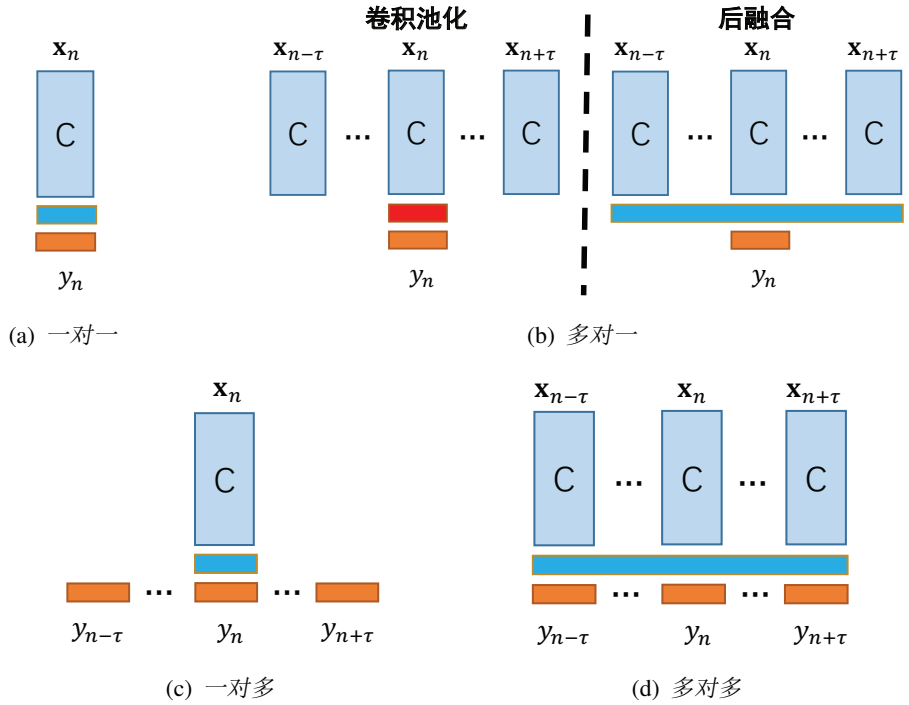
其中y_n是$\mathbf{x}_n$的发育阶段标签，L是胚胎发育阶段的标签集。

本文采用ResNet50[4]作为基准模型处理视频中的图像帧，其结构如表 4.1 所示。ResNet获得了2015 ImageNet分类竞赛的冠军，并在图像领域中得到广泛应用。Time-lapse视频中原始的800×800图像被缩放到224×224作为输入。模型使用在ImageNet数据集[96]上预训练的参数进行初始化，这有利于缓解模型在小数据集上过拟合。

表 4.1 ResNet50基准模型

layer name	50-layer	output size
conv1	7×7, 64, stride 2	$112 \times 112 \times 64$
res2	3×3 max pool, stride 2	$56 \times 56 \times 256$
	$\begin{bmatrix} 1 \times 1,\ 64 \\ 3 \times 3,\ 64 \\ 1 \times 1,\ 256 \end{bmatrix} \times 3$	
res3	$\begin{bmatrix} 1 \times 1,\ 128 \\ 3 \times 3,\ 128 \\ 1 \times 1,\ 512 \end{bmatrix} \times 4$	$28 \times 28 \times 512$
res4	$\begin{bmatrix} 1 \times 1,\ 256 \\ 3 \times 3,\ 256 \\ 1 \times 1,\ 1024 \end{bmatrix} \times 6$	$14 \times 14 \times 1024$
res5	$\begin{bmatrix} 1 \times 1,\ 512 \\ 3 \times 3,\ 512 \\ 1 \times 1,\ 2048 \end{bmatrix} \times 3$	$7 \times 7 \times 2048$
global average pool		$1 \times 1 \times 2048$
fc		$1 \times 1 \times 1024$
softmax		$1 \times 1 \times 6$

当能够同时使用当前帧和相邻帧的信息时，标准的*一对一*分类框架就可以拓展到*多对一*、*一对多*和*多对多*MTDL分类框架，具体的示意图如图 4-3 所示。图中的C代表ResNet50基准模型，用于处理独立的图像帧提取特征；蓝色和红色矩形分别代表展开层和最大池化层；橙色矩形代表全连接和softmax层，用于整合特征输出最终的预测结果。



(a) *一对一*　(b) *多对一*

(c) *一对多*　(d) *多对多*

图 4-3 不同的分类框架

4.2.3 多对一MTDL框架

图 4-3(b) 展示了两种不同的*多对一*MTDL框架，其被广泛应用于视频理解领域中[97–99]。由于一个视频具有多张图像，但一般只有一个标签，因此可以在综合考虑所有的图像后，再给出最终的预测结果。与*一对一*相比，*多对一*框架可以更好得利用输入的上下文信息。

*多对一*执行如下的映射:

$$f_1 : (\mathbf{x}_{n-\tau}, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{n+\tau}) \mapsto y_n \in L. \tag{4.2}$$

除了当前帧$\mathbf{x}_n$外，还会取它的前τ帧和τ帧，实际的输入上下文窗口大小为$2\tau+1$。

融合相邻的$2\tau+1$个图像帧中的时间域信息，通常有两种方法：卷积池化[100]（Conv Pooling）和后融合[98]（Late Fusion）。

4.2.3.1 卷积池化

这是一种先卷积然后进行时域池化的特征融合方法，常常被应用于视频分类中。其会使用共享的CNN对输入的图像帧提取特征，然后在时域上进行最大池化对特征进行合并。池化后的特征被送入全连接层经softmax给出最终的预测结果。该方法的主要优点在于，卷积层输出的多个帧的空间信息，可以在进行时间域最大池化的时候得到保留。Ng等人[100]使用120帧的AlexNet模型[11]，在Sports-1M数据上进行了实验，表明卷积池化优于其它的特征池化方法。

4.2.3.2 后融合

在特征后融合策略中，输入上下文窗口中的所有图像帧都会通过同一个卷积层（CNN网络）进行编码。经过卷积后的特征被拼接在一起，使用全连接层生成最终的分类结果。在实际操作时，可以是对输入上下文窗口中的部分帧的特征拼接[98]，也可以是对窗口中的所有帧进行拼接[89]。先前Ng等人的研究[89]表明，在使用time-lapse视频预测胚胎的形态动力学时，使用15帧的后融合网络优于早融合策略（Early Fusion）。

4.2.4 一对多MTDL框架

如图 4-3(c) 所示，*一对多*意味着对于输入的一个图像帧会给出多个输出结果。除了预测当前帧的胚胎发育阶段外，*一对多*的MTDL架构还需要预测相邻帧的标签。这种同时进行多个相似任务的结构，在深度学习中被称为多任务学习。一种常用的多任务策略是对特征提取部分的参数进行硬参数共享[95]，其具体形式如图 4-4 所示。特征提取层（本文中为ResNet50）的参数在不同的任务上共享，但是用于输出预测结果的全连接层（特定任务层），需要各自训练。

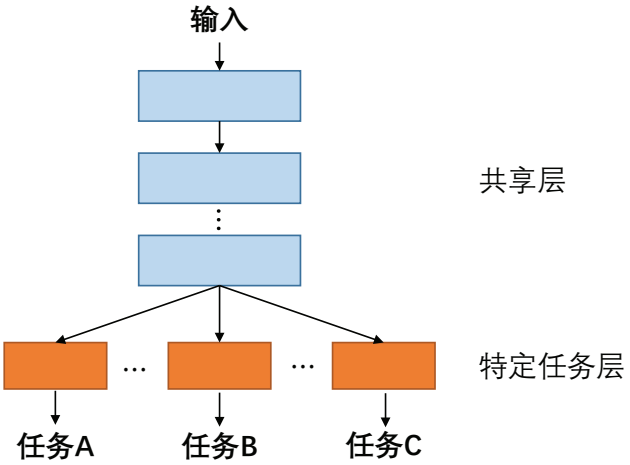


图 4-4 MTDL中的硬参数共享

在*一对多*中，当前帧$\mathbf{x}_n$被用于预测以帧$\mathbf{x}_n$为中心的$2\tau+1$帧的胚胎发育阶段，即其学习如下的一对多映射：

$$f_2 : \mathbf{x}_n \mapsto (y_{n-\tau}, \ldots, y_{n+\tau}) \in \boldsymbol{L}^{2\tau+1}. \tag{4.3}$$

因为本章研究的是一个多分类任务，所以当前帧$\mathbf{x}_n$对于帧$\mathbf{x}_t$的胚胎发育阶段预测结果为一概率向量$\hat{\boldsymbol{p}}_t(\boldsymbol{x}_n) \in \mathbb{R}^{|L|\times 1}$，$t \in [n-\tau, n+\tau]$

从另一个角度看，对于帧$\mathbf{x}_n$，除其本身外，相邻的2τ个图像帧$\mathbf{x}_t$也会对它给出预测标签。通过整合这$2\tau+1$个预测结果，可以给出更加稳定和可靠的预测结果。具体的融合策略，将在下文中进行详细讲解。

此外，由于帧$\mathbf{x}_n$涉及到$2\tau+1$的输出，所以帧$\mathbf{x}_n$在训练时的损失为所有涉及输出的损失之和：

$$\ell(\mathbf{x}_n) = \sum_{t=n-\tau}^{n+\tau} w_t \cdot \ell(y_t, \hat{\boldsymbol{p}}_t(\mathbf{x}_n)), \tag{4.4}$$

上式中的w_t是对第 t 个输出的权重，y_t是帧$\mathbf{x}_t$的真实标签。在本文中，$w_t = 1$，使用交叉熵作为损失函数。对第 t 个输出的交叉熵损失定义如下：

$$\ell(y_t, \hat{\boldsymbol{p}}_t(\mathbf{x}_n)) = -\log\left(\hat{p}_{t,y_t}(\mathbf{x}_n)\right), \tag{4.5}$$

其中$\hat{p}_{t,y_t}(\mathbf{x}_n)$是$\hat{\boldsymbol{p}}_t(\mathbf{x}_n)$的第$y_t$个元素。

4.2.5 多对多MTDL框架

多对多 MTDL框架可以认为是*一对多*和*多对一*框架的融合。

多对多 框架的结构如图 4-3(d) 所示。其以多个图像帧作为输入，预测所有输入帧的胚胎发育阶段标签。和*多对一*框架类似，使用共享的CNN处理每个输入的图像帧，利用特征后融合策略对所有特征进行整合；*一对多*框架相同，单独训练用于输出预测结果的全连接层参数。

4.3 具有DP的多任务深度学习算法

在本节中，将对本章提出的具有动态规划后处理的多任务深度学习算法（MTDL-DP）进行介绍。

4.3.1 MTDL中的集成学习

在 4.2.4节中有提到，一个多任务框架会具有多个输出。对于一个特定图像帧，最简单的方法就是取网络中间的输出结果作为最终的预测结果。但更合适的方法是集成学习[101]，通过集成思想对所有的预测结果进行整合，可以保证预测结果的可靠性，提高算法的最终性能。在本文中，考虑了两种常见的概率集成策略：*加法均值*和*乘法均值*。

令$\hat{\boldsymbol{p}}_n(\mathbf{x}_t)$为图像帧$\mathbf{x}_t$对帧$\mathbf{x}_n$给出的预测概率向量，其中$t \in [n-\tau, n+\tau]$，则多任务框架预测结果的示意图如图 4-5 所示。

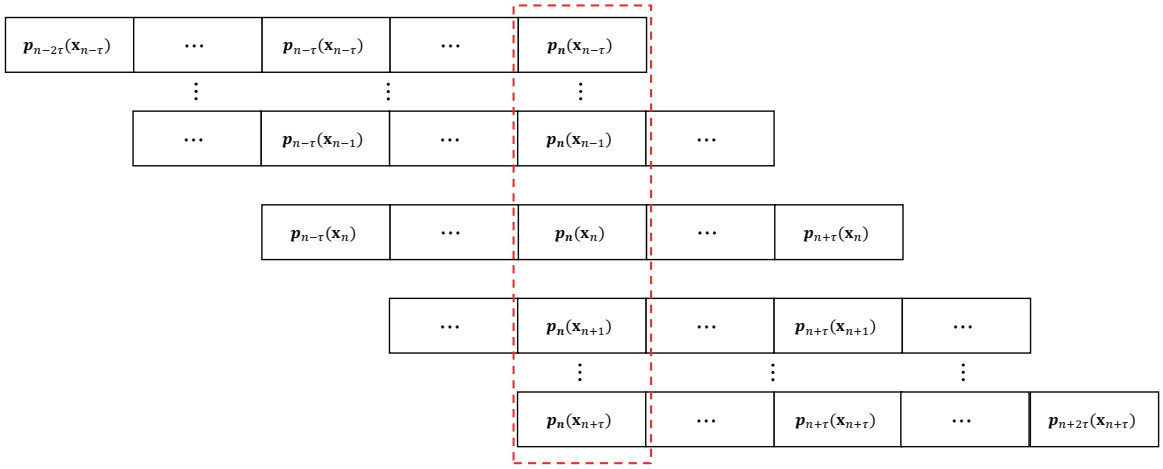


图 4-5 多任务框架的预测结果示意图

当使用*加法均值*策略时，图像帧$\mathbf{x}_n$集成后的预测概率向量$\hat{\boldsymbol{p}}_n$为：

$$\hat{\boldsymbol{p}}_n = \frac{1}{2\tau+1} \sum_{t=n-\tau}^{n+\tau} \hat{\boldsymbol{p}}_n(\mathbf{x}_t). \tag{4.6}$$

如果使用*乘法均值*策略则：

$$\hat{\boldsymbol{p}}_n = \frac{1}{2\tau+1} \prod_{t=n-\tau}^{n+\tau} \hat{\boldsymbol{p}}_n(\mathbf{x}_t). \tag{4.7}$$

由于$\hat{\boldsymbol{p}}_n(\mathbf{x}_t)$是向量，上面的加法和乘法均为元素级别的操作。

最后通过概率最大化，得到图像帧$\mathbf{x}_n$最终的分类标签$\hat{y}_n$:

$$\hat{y}_n = \arg\max_{1\le l\le |L|} \hat{p}_{n,l}, \tag{4.8}$$

其中$\hat{p}_{n,l}$是$\hat{\boldsymbol{p}}_n$的第 l 个元素，$|L|$为总类别数。

4.3.2 DP后处理

在胚胎发育的过程中，分裂细胞的数量几乎总是非递减的，即只会保持不变或者增加[102]。一般不会从分裂成4细胞阶段的阶段回退到3细胞阶段，也不会从2细胞的阶段回退到1细胞阶段。如果不进行特殊处理，直接使用MTDL的预测结果作为输出，则无法实现这样的约束效果，且会有较多的毛刺。因此，我们使用DP对整个视频的发育阶预测段序列进行了后处理，以满足非递减约束。

每个视频具有多张图片帧，对应的真实发育阶段标签$\{y_n\}_{n=1}^N$可组成一个序列。当帧$\mathbf{x}_n$作为输入时，MTDL算法输出的预测概率向量$\hat{\boldsymbol{p}}_n = [p_{n,1}, ..., p_{n,|L|}]^T$，其中$\hat{p}_{n,l}$为帧$\mathbf{x}_n$为发育阶段 l 时的概率。

对整个视频而言，有输出概率矩阵$\hat{\boldsymbol{P}} = [\hat{\boldsymbol{p}}_1, ..., \hat{\boldsymbol{p}}_N]$，定义$E(\hat{\boldsymbol{y}}, \hat{\boldsymbol{P}})$为预测结果是$\hat{\boldsymbol{y}} = \{\hat{y}_n\}_{n=1}^N$时的损失函数。总的损失函数应当是视频中所有帧的损失之和，即有$E(\hat{\boldsymbol{y}}, \hat{\boldsymbol{P}}) = \sum_{n=1}^N e(\hat{y}_n, \hat{\boldsymbol{p}}_n)$。当优化预测结果$\hat{\boldsymbol{y}} = \{\hat{y}_n\}_{n=1}^N$，使其满足单调性约束$\hat{y}_{n+1} \ge \hat{y}_n,\ \forall n$，且最总的损失函数$E(\hat{\boldsymbol{y}}, \hat{\boldsymbol{P}})$最小时，获得结果即为该视频最优的胚胎发育阶段预测序列。

在这里考虑两种常见的损失函数[89]：交叉熵（Cross Entropy，CE）和搬运距离（Earth Mover' Distance，EMD）。

交叉熵损失的定义为：

$$e_{CE}(\hat{y}_n, \hat{\boldsymbol{p}}_n) = -\log(\hat{p}_{n,y_n}). \tag{4.9}$$

搬运距离损失计算方法如下：

$$e_{EMD}(\hat{y}_n, \hat{\boldsymbol{p}}_n) = -\sum_{l=1}^{|L|} \hat{p}_{n,l} |\hat{y}_n - l|. \tag{4.10}$$

综上可知，整个视频最终的胚胎发育阶段预测序列$\hat{\boldsymbol{y}}^* = \{\hat{y}_n\}_{n=1}^N$可通过以下的最小化约束获得：

$$\begin{aligned} \hat{\boldsymbol{y}}^* = & \arg\min_{\hat{\boldsymbol{y}}=\{\hat{y}_n\}_{n=1}^N} \sum_{n=1}^N e(\hat{y}_n, \hat{\boldsymbol{p}}_n) \\ \text{s.t.} \quad & \hat{y}_{n+1} \ge \hat{y}_n,\ \forall n. \end{aligned} \tag{4.11}$$

上式可以通过DP后处理实现，详见算法 1。

Algorithm 1: 动态规划（DP）后处理伪代码

Input: N，time-lapse视频中图像帧的数量;
L，胚胎发育阶段标签集;
$\hat{\boldsymbol{P}} = [\hat{\boldsymbol{p}}_1, ..., \hat{\boldsymbol{p}}_N] \in \mathbb{R}^{|L| \times N}$，对整个视频共$N$帧图像，MTDL模型输出的预测概率矩阵。

Output: $\hat{\boldsymbol{y}}^*$，优化后的发育阶段序列。

令$e(l, \hat{\boldsymbol{p}}_n) = 0$，且$E(l, \hat{\boldsymbol{p}}_n) = 0, \forall l \in [1, |L|], \forall n \in [1, N]$;

for $n = 1, ..., N$ **do**
  for $\hat{y} = 1, ..., |L|$ **do**
    计算式 (4.10) 中的$e(\hat{y}, \hat{\boldsymbol{p}}_n)$;
  end
end

for $n = 2, ..., N$ **do**
  for $\hat{y} = 1, ..., |L|$ **do**
    $E(\hat{y}, \hat{\boldsymbol{p}}_n) = e(\hat{y}, \hat{\boldsymbol{p}}_n) + \min_{1 \leq l \leq \hat{y}} E(l, \hat{\boldsymbol{p}}_{n-1})$;
  end
end

$k = |L|$;

for $n = N, ..., 1$ **do**
  $\hat{y}_n = \arg\min_{1 \leqslant l \leqslant k} E(l, \hat{\boldsymbol{p}}_n)$;
  if $\hat{y}_n < k$ **then**
    $k = \hat{y}_n$;
  end
end

$\hat{\boldsymbol{y}}^* = \{\hat{y}_n\}_{n=1}^N$;

Return 优化后的发育阶段序列$\hat{\boldsymbol{y}}^*$。

4.3.3 MTDL-DP

本文提出的MTDL-DP算法大致可以分成四步：1）使用*一对一*框架训练基准模型用于后续MTDL框架中模型的初始化；2）构建一个*一对多*或者*多对多* MTDL 框架；3）使用*乘法均值*集成策略整合多任务框架的多个预测结果；4）利用DP和搬运距离损失函数对整个视频的预测序列进行后处理。

在算法 2 中，以*一对多* MTDL框架为例，使用伪代码描述了整个MTDL-DP算法的流程。其中的*一对多*框架可以替换为*多对多* MTDL框架。

Algorithm 2: MTDL-DP

Input: N，time-lapse视频中图像帧的数量;
D，已标注的time-lapse视频数据集;
$\{\mathbf{x}_n\}_{n=1}^{N}$, 已标注的图像帧;
τ, 上下文窗口中要取的左右相邻帧数量。
Output: $\hat{\boldsymbol{y}}^*$，预测的胚胎发育阶段序列。
基于D，使用*一对一*框架训练基准模型 f_0;
初始化*一对多* MTDL模型，其中卷积层参数与 f_0 相同;
在D上微调MTDL模型的全连接层参数;
for $n = 1, ..., N$ **do**
  使用MTDL模型计算$\hat{\boldsymbol{p}}_t(\mathbf{x}_n), t = n-\tau, ..., n+\tau$;
end
for $n = 1, ..., N$ **do**
  利用式 (4.7) 计算$\hat{\boldsymbol{p}}_n$;
  计算式 (4.10) 中的单帧损失$e_{EM}(\hat{y}_n, \hat{\boldsymbol{p}}_n)$;
end
通过算法 1 计算式 (4.11) 中的$\hat{\boldsymbol{y}}^*$;
Return 预测的胚胎发育阶段序列$\hat{\boldsymbol{y}}^*$。

4.4 实验和结果

在本节中，对提出的MTDL-DP算法进行了分解，逐个探究不同组件对模型性能的影响，并通过实验验证了本章提出MTDL-DP算法的优越性。

4.4.1 实验设置

将time-lapse视频数据集按照7：1：2的比例随机选择视频，划分为训练集、验证集和测试集，分别获得为41650 / 5950 / 11900帧胚胎图像。为了使用ResNet50作为基准模型，图像原本的800 × 800分辨率被缩放为224 × 224。在训练时，通过随机旋转一定角度、镜像翻转以及垂直翻转进行数据增强，测试时无数据增强。所有的MTDL框架均使用*一对一*（ResNet50）训练得到的参数进行初始化。然后冻结卷积层参数，仅对新增的全连接层参数进行训练调整。

在模型的训练过程中，使用交叉熵作为损失函数，Adam作为优化器[49]，利用早停策略缓解过拟合。涉及到MTDL-DP算法时，默认使用*乘法均值*策略对预测结果进行集成，在DP后处理中采用EM距离损失。所有的实验均重复五次，取测试集上的均值作为最终的实验结果。在后面的实验中，若无特别申明，默认未使用DP后处理对整个视频的预测序列进行优化。

4.4.2 分类正确率

首先，本文在不同的上下文窗口半径下，探究了不同的MTDL框架，以及是否使用DP后处理对分类正确率的影响。上下文窗口半径 $\tau = \{1, 4, 7\}$ ，对应窗口大小为 $2\tau + 1$。需要注意的是，对于*一对一*框架而言，实际上所有的 τ 均为0，其仅使用了当前帧作为输入。表 4.2 的左半部分展示了未使用DP后处理时的分类正确率。从表中结果可以发现，所有MTDL框架的正确率均高于*一对一*基准模型，这表明在多任务学习中使用相邻图像帧作为输入或者使用对应的标签信息作为监督对算法性能的提升均都有益的。

表 4.2 对于不同的框架和 τ，进行DP后处理前后的分类正确率

框架	方法	无DP时正确率			使用DP时正确率		
		$\tau = 1$	$\tau = 4$	$\tau = 7$	$\tau = 1$	$\tau = 4$	$\tau = 7$
一对一	ResNet50	83.8%	83.8%	83.8%	86.1%	86.1%	86.1%
多对一	卷积池化	84.7%	84.4%	83.8%	85.9%	85.1%	84.5%
	后融合	83.9%	84.6%	85.1%	86.0%	85.2%	85.2%
一对多	本文提出的多任务网络	**85.0%**	85.4%	85.3%	86.5%	85.8%	85.7%
多对多		84.6%	**85.7%**	**85.8%**	**86.6%**	**86.5%**	**86.9%**

使用DP对整个视频的预测序列进行后处理时的分类准确率结果见表 4.2 的右半部分。对于不同的分类器和窗口半径 τ，DP后处理均能提高分类正确率。当 $\tau = 1$时，五个分类器分别提高了2.3%、1.2%、2.1%、1.5%和2.0%。同时，随着 τ 的增加，对大多数分类器而言，DP后处理带来的分类性能提升会明显变少。这是因为 τ 在增加时，大多数分类器的性能也在增加，DP后处理的增益会相应减少。当$\tau = 7$时，*多对多* MTDL框架在所有的方法中取得了最好的分类正确率，86.9%，比基准模型ResNet50高出**3.1%**。

对于*多对一* MTDL框架，无论是否使用DP后处理，当 τ 增加时，特征后融合策略的正确率也在增加，而卷积池化的正确率却在降低。这是符合直觉的。当 τ 增加时，对于不同输入帧的特征，卷积池化的时域池化只是单纯地选取最大值作为特征，容易误检，也忽略了大多数的输入信息。后融合策略则是使用全连接层直接计算所有输入特征的加权和，虽然计算量稍大，但是更加合理。

此外，DP后处理前后，*多对多*和*一对多*框架始终比*多对一* MTDL框架具有更高的分类正确率。这验证了本章所提出的多任务集成框架的优越性。通过多任务网络以图像帧作为输入，输出多个预测标签，然后使用集成策略进行整合，可以使得模型的预测结果更加稳定和可靠。

在进行DP对预测序列进行后处理后，在不同的 τ 下，*一对多*框架分别提高了1.5%、0.4%和0.4%，而*多对一*框架为2.0%、0.8%和1.1%。DP后处理对*多对一*框架的提升始终高于*一对多*框架，这意味着当有足够多的输入输出信息可以利用时，DP后处理的性能提升会更有效率。

4.4.3 均方根误差

除了分类正确率外，本文还计算了真实视频标签序列和模型结果序列的均方根误差（Root Mean Squared Error，RMSE），以探究预测序列的对真实标签的拟合程度。不使用DP进行后处理时的RMSE结果见表 4.3 的左半部分。所有的MTDL框架均比*一对一*框架的RMSE低，这和 4.4.2节的结果类似，再次说明了多任务学习中使用相邻的输入帧和标签信息是有益的。

表 4.3的右半部分展示了使用DP后处理时的结果。可以发现DP后处理对所有的MTDL框架和 τ 均能减少RMSE，这表明DP后处理确实能起到作用，再一次和4.4.2节的结论互相印证。此外，DP后处理后，虽然部分MTDL框架的RMSE仅在 $\tau = 1$时优于*一对一*框架，但*多对多*框架对于不同的 τ 均优于*一对一*框架。

表 4.3 对于不同的分类框架和τ，使用DP后处理前后的RMSE

框架	方法	无DP时RMSE			有DP时RMSE		
		$\tau=1$	$\tau=4$	$\tau=7$	$\tau=1$	$\tau=4$	$\tau=7$
一对一	ResNet50	0.4840	0.4840	0.4840	0.4199	0.4199	0.4199
多对一	卷积池化	0.4728	0.4690	0.4795	0.4066	0.4432	0.4419
	后融合	0.4761	**0.4531**	0.4740	0.4036	0.4254	0.4214
一对多	本文提出的多任务模型	**0.4638**	0.4695	0.4480	**0.3964**	0.4155	0.4260
多对多		0.4752	0.4640	**0.4360**	0.4085	**0.4077**	**0.4083**

4.4.4 训练时间

更进一步的，除了探究模型性能外，本文还研究了不同框架和上下文窗口半径下的模型训练时间，具体见表 4.4。为了保证结果的稳定性，重复进行了五次实验，取训练时间的均值作为表中的结果。DP属于后处理范畴，与模型训练无关，只影响测试时间，对训练时间无影响。

从结果中可以观察到，随着输入长下文窗口的增加，*多对一*和*多对多* MTDL框架的训练时间均会线性增加；而*一对多*框架的训练时间基本上对 τ 不敏感，只会随着上下文窗口半径 τ 的变大微弱增加，且远低于前两种多输入框架。这是因为*多对*

一和*多对多* MTDL框架为多输入，模型训练的时间主要花在对大量输入帧的处理上，而一*对多*框架为单帧输入，这是一*对多* MTDL框架的优势。实际上，该优势也会反映在测试时间上。

表 4.4 不同分类框架和τ的训练时间

框架	方法	训练时间（s）		
		$\tau=1$	$\tau=4$	$\tau=7$
一对一	ResNet50	**2231**	**2231**	**2231**
多对一	卷积池化	5318	15378	29139
	后融合	4892	17390	27534
一对多	本文提出的多任务网络	2246	2265	2542
多对一		5759	16182	27808

4.4.5 不同集成方法对性能的影响

为了对多任务框架的预测结果进行整合，获得更加稳定可靠的输出，本文曾在4.3.1节中介绍了两种不同的集成方法。在本节的实验中，使用了一*对多*和*多对多*这两种多输出MTDL框架构建CNN模型，探究不同集成方法对模型性能的影响。图 4-6展示了使用不同集成方法整合预测结果时的分类正确率和RMSE。

从图中可以发现，不论是分类正确率还是RMSE，*加法均值*和*乘法均值*这两种集成方法均可以取得性能的提升。同时，*乘法均值*的结果会稍好于*加法均值*，这是因为*乘法均值*集成策略更倾向于具有一致性的预测结果。

举例而言，假设一个二分类任务具有两对预测结果，分别是 $(0.1, 0.5)$ 和 $(0.9, 0.5)$。当使用*加法均值*策略时：

$$\frac{1}{2}\left((0.1, 0.5) + (0.9, 0.5)\right) = (0.5, 0.5), \tag{4.12}$$

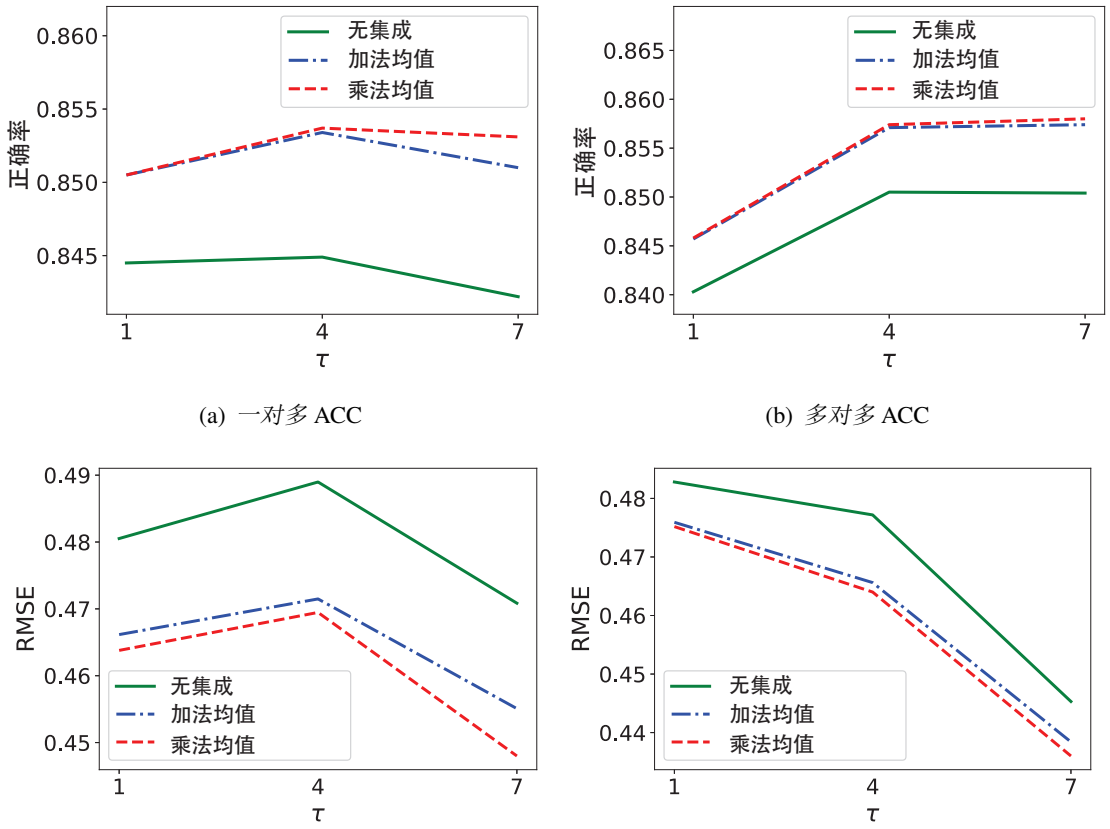
因此两个类别的预测概率均为0.5；而对于*乘法均值*：

$$\sqrt{\left((0.1, 0.5) \times (0.9, 0.5)\right)} = \sqrt{(0.09, 0.25)} = (0.3, 0.5), \tag{4.13}$$

两个类别的概率则分别是0.3和0.5。可以发现*乘法均值*集成策略，会选择这种更加一致且稳定的类别作为预测结果，因此优于*加法均值*策略。

关于窗口半径 τ 的大小对不同多输出MTDL框架的影响。随着 τ 的增加，模型需要预测的图像帧的数量也在增加。对于*多对多*框架，其性能随着 τ 的增加而增加，

且增益逐渐减小，但同时由 4.4.4 节的结论可知，其训练时间也会显著增加，因此需要在性能和时间之间做好权衡。而对于*一对多* MTDL框架，其在$\tau = 4$时，取得了最佳的正确率。当窗口半径 τ 进一步增加时，正确率反而降低。这是因为*一对多*框架需要预测越来越远的相邻图片帧的类别，但它的输入只有当前帧，对于比较远的相邻帧能够包含的信息十分有限，难以进行比较准确的预测。



(a) *一对多* ACC　　(b) *多对多* ACC

(c) *一对多* RMSE　　(d) *多对多* RMSE

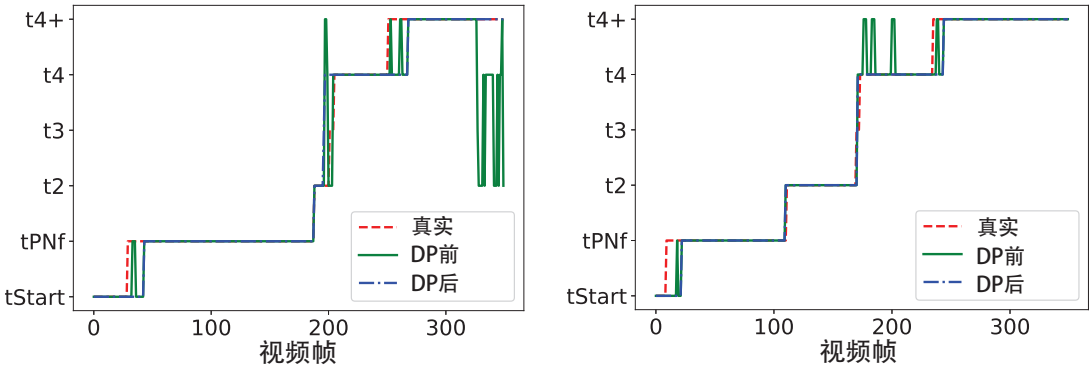
图 4-6 不同集成方法对分类正确率和RMSE的影响

4.4.6 DP后处理以及不同损失函数的影响

在本小结中，探索了是否使用DP后处理优化整个视频的预测序列，以及DP中不同单帧损失函数，对模型性能的影响。默认采用*一对多* MTDL框架，窗口半径$\tau = 1$，且采用*乘法均值*集成策略。

图 4-7 展示了是否进行DP后处理时，某两个time-lapse视频的真实胚胎阶段标签和预测标签序列的对比情况。可以明显地看出，在未使用DP后处理时，模型直接输出的预测结果十分粗糙，有很多毛刺，预测的发育阶段也会出现回退突变的情况。通过加入非递减约束，DP后处理明显平滑了预测的结果，使得最终的输出序列与真

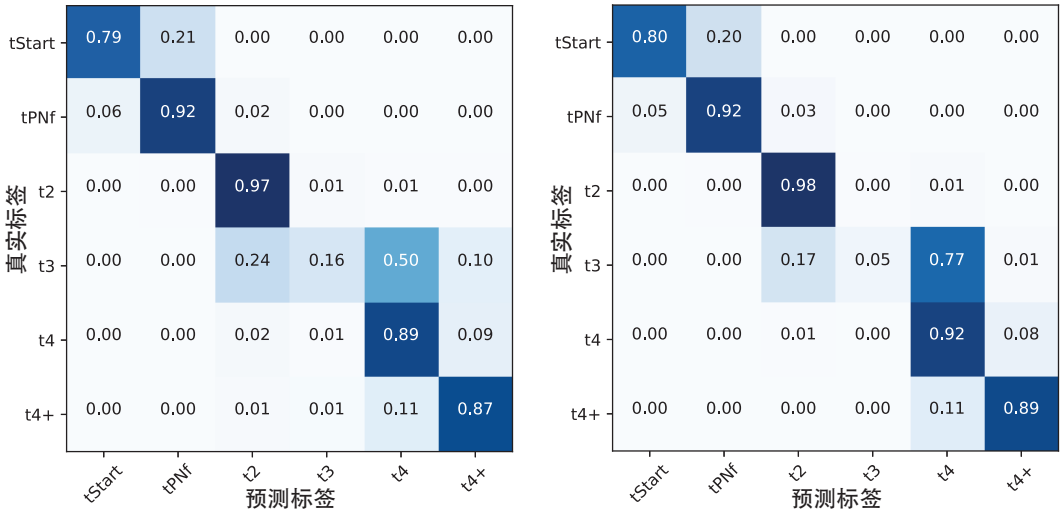
实标签更相近。这是视觉上的结果，数字上的结果可以参见表 4.2。进行DP后处理后，分类正确率从85%升到了86.5%。



(a) 视频 1　　(b) 视频 2

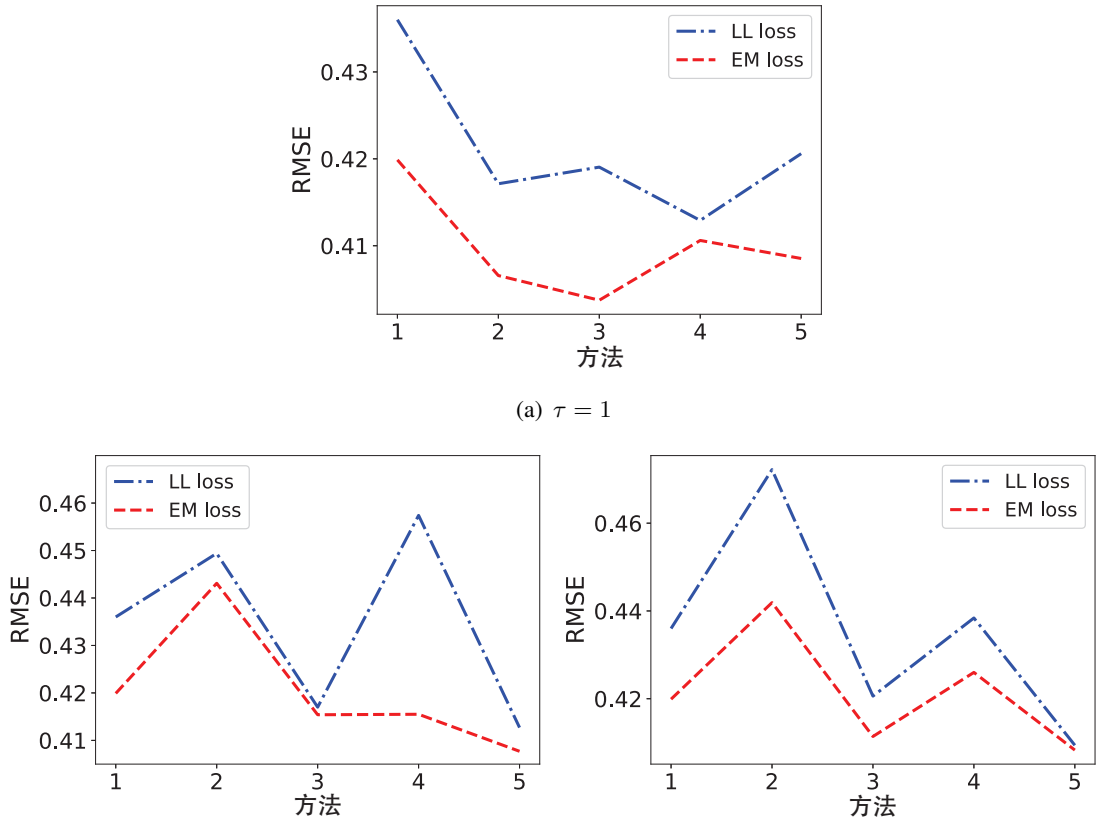
图 4-7 是否使用DP后处理时的预测结果和真实标签序列的对比

图 4-8 则显示了DP后处理在模型输出结果的混淆矩阵上的影响。图中的对角线上数字是每个细胞阶段的召回率（也可以称为敏感度，单位为%）。除了 t3 阶段外，DP后处理可以提高所有细胞发育阶段的召回率。这里面可能的原因有两个：1）在本章的训练集中，t3 阶段的训练样本非常少（见图 4-2），因此它没有得到充分的训练；2）t3的召回率低也可能是因为合子期的多极分裂比较少，该发育阶段仅在12.2%的人类胚胎分裂中出现[103]。



(a) DP后处理前　　(b) DP后处理后

图 4-8 DP后处理对混淆矩阵的影响



(a) $\tau = 1$

(b) $\tau = 4$

(c) $\tau = 7$

图 4-9 DP使用不同单帧损失时的RMSE。横轴上的数字代表了不同的方法，分别对应：*一对一*，*多对一* (卷积池化)，*多对一* (后融合)，*一对多*，*多对多*。

只是通过DP后处理加入发育阶段非递减的约束是不够的。假如将所有的发育阶段均预测为同一阶段，同样满足非递减约束，但是模型性能会很差。因此，在DP后处理时，还需要最小化模型对整个视频输出的概率向量序列和最终预测的类别向量的损失。在 4.3.2 节介绍DP后处理时，提到了两种单帧损失函数：搬运距离（EMD）和交叉熵（CE）。图 4-9 展示了不同窗口半径 τ 和MTDL框架下，采用不同单帧损失对RMSE的影响。从图中结果可知，无论在什么实验设置下，采用EMD损失时的RMSE均小于CE损失。

实际上，CE损失是KL散度在一定条件下的等价形式。KL散度描述了某个分布去拟合另一分布时，自信息差的期望。而且，当两个分布调换位置时，KL散度的值不同。EMD则计算了从一个分布变成另一个分布的最小代价，等价于Wasserstein距离。直观上，可以将每一个分布想象成多堆土，每个位置上均含有一定的土，EMD描述了将某个分布的土堆搬运到另一个分布是需要做的功。两个土堆的距离相距越远，需要做的功越大（功=搬运距离×土的重量）。相比于CE损失，EMD对距离更加敏感，其更适合度量描述两个序列之间的损失。

4.5 本章小结

在进行体外人工受孕时，对胚胎早期的发育阶段进行准确的预估，可以为胚胎学家提供评估胚胎质量的宝贵信息，用于挑选合适的胚胎进行移植，对受孕的成功至关重要。为了利用time-lapse 视频对胚胎的不同发育阶段进行准确地分类，本章提出了一种具有动态规划的多任务深度学习框架（MTDL-DP）。首先，该方法基于视频中相邻帧具有大量互补信息的特性，利用多任务学习和相邻帧为时延视频的每一个图片帧都生成多个预测结果；然后通过集成思想对这些预测结果进行整合，赋予当前帧一个胚胎发育阶段；最后使用动态规划优化整个视频的发育阶段序列，使得最后的发育阶段序列非递减，且搬运距离损失最小。据我们所知，这是第一项将多任务集成应用在胚胎早期发育阶段预测的研究。通过本章提出的MTDL-DP 算法，将胚胎早期发育阶段分类的精度提高了3.1%。

5 总结与展望

5.1 总结

医疗影像技术在大量疾病的临床诊断和治疗中起着重要作用，提供了生理信号、疾病区域、病情程度等关键信息。在实际操作中，对医疗影像和数据的解读通常由具有丰富实践经验的专业医生进行，该过程繁琐、工作量大，且严重依赖于专家的经验，这在乡镇医院等医疗力量薄弱处更加堪忧。引入计算机辅助诊断技术可以降低医生的负担，显著提高诊断效率。

目前已经有不少基于传统机器学习技术的方法，但是性能还不够高，且需要复杂的预处理和较强的医学先验知识设计特征。深度学习可以从数据中挖掘有用信息，自动学得具有良好判别能力和泛化能力的高层次抽象特征。在当前大数据、高计算资源的情况下，本文基于先进的深度学习技术，如卷积神经网络、长短时记忆和注意力机制等，为不同的医学问题提供了一套简洁的诊断流程。展现了不同的深度学习技术和用法，希望能抛砖引玉，为深度学习在医疗影像数据中的应用提供比较多样和全面的视野。

在本文中，我们分别基于脑电图（EEG）、磁共振成像（MRI）和时延视频（time-lapse）等不同类型的医疗影像数据，从不同的角度，对三种医学问题进行了探究，所提出的算法均达到了领先的性能。本文的主要研究内容如下，其均为本文作者在硕士研究生期间的工作：

(1) 基于EEG和深度学习的睡眠状态检测。睡眠状态的检测有利于睡眠相关疾病的诊断。本文基于单通道EEG信号，提出了两种新颖的深度学习模型用于睡眠状态检测，基于卷积神经网络的C-CNN模型即紧凑又高效，在计算代价和模型性能之间达到了很好的平衡；基于注意力机制和双向长短时记忆的Attention模型在需要更长训练时间的代价下，可以获得更好的性能，领先于现有的方法。此外，我们将代价敏感学习整合到模型训练过程中，解决了睡眠状态中存在的严重类别不平衡问题，取得更好的平衡分类正确率。

(2) 基于MRI和CNN的鼻咽癌诊断支持系统。鼻咽癌是中国东南部、台湾、香港、马来西亚和新加坡最常见的头颈癌。本文为鼻咽癌的诊断提供了一套可视化的辅助决策技术，且可以在不同分辨率不同型号的MRI设备下运行。其首先通过开操作和大津阈值法对MRI图片进行自适应的分割和裁剪，提取有效的大脑部分，同时解决了跨分辨率跨系统的问题；然后使用修改后的残差网络和我们提出的可视化技术对不同MRI切片进行处理，提取特征和快速定位恶性肿瘤可能存在的切片和区域；最后对所有切片的高层次抽象特征进行整合，给出最终的鼻咽癌阳性概率。在可以为医生提供可疑肿瘤位置标注的同时，对鼻咽癌预测的AUC指标也达到了0.993。

(3) 基于时延视频和多任务学习的胚胎早期发育状态分类。在体外人工受孕的治疗过程中，对胚胎的早期发育过程进行准确地检测，可为评估胚胎质量提供宝贵的信息，有利于受孕的成功。本文提出了一种具有动态规划的多任务深度学习方法。它首先使用多任务学习和集成思想将时延视频的每一帧预分类成一个胚胎发育阶段，然后使用动态规划优化发育阶段序列，使得最后的发育阶段序列非递减，且最小化搬运距离损失。我们是第一项将多任务学习应用于时延视频对胚胎早期发育状态进行分类的研究。

5.2 展望

随着数据量和计算资源的飞速增长，深度学习技术得到了广泛的应用。虽然深度学习在各个领域都占据了支配的地位，但我们在应用到具体医学问题中，发现还存在着一些不足：

(1) 对于数据量的需求过大。深度学习方法本质上是有监督的，需要大量带标签的数据来提高算法性能。但实际上，人类和动物的学习，在很大程度上是无监督：我们是通过观察来发现世界的结构，而不是通过被告知每个物体的名称。医学数据的标注成本是十分昂贵的，利用大量的无标签数据，研究无监督学习[104–107] 和深度学习的结合，从长远来看，会变得更加重要。

(2) 可解释性不够好。虽然深度学习在很多领域取得了喜人的成绩，但是其可解释性还是比较差。深度学习通过多个非线性处理层的堆叠组成计算模型，从数据中学得多层次的抽象特征。其背后没有严格的数学理论支撑，我们难以知道所学得特征的物理意义是什么，为什么效果好。特别是对于严肃的医学问题，如何信任基于无法理解的特征产生的预测结果? 值得庆幸的是，已经有比较多的研究开始关注于深度学习的可解释性和可视化工作[108–111]。

(3) 可迁移性还需要提高。在一个数据集上训练得到表现优秀的深度学习模型，如果直接用在另一个不同分布的数据集上，大概率会表现得很差。这里面涉及到了模型的可迁移性问题，属于迁移学习的范畴[112]。在医学领域中，特别是对于生理信号而言，不同人的生理信号会存在显著的差异性，甚至同一个人在不同时刻的生理信号也会存在差异。在第 2章中，我们基于EEG信号对睡眠状态检测进行了研究。在进行留一被试交叉验证时，虽然我们提出的算法在不同被试间保持了较好的一致性和性能，达到了领先的性能，但不可否认的是，在不同被试上的表现仍然存在一定的差异。为了解决深度学习模型的可迁移性问题，还需要更深入的研究[113,114]。

致　　谢

时光犹如白驹过隙，眨眼之间，硕士研究生的旅程即将走到终点。从大四加入实验室算起，我在这将近四年里度过了人生中一段难忘的岁月，从生活上到学术上均学到了很多。蓦然回首，才发现要感谢的太多太多。

在论文即将完成之际，首先我要感谢我的导师伍冬睿教授。感谢他为我提供了足够的空间施展拳脚，提供实验设备，支持我从事看起来不太靠谱的研究，在不同的方向上探索。伍老师丰富的人生见识、渊博的学识和严谨的治学态度，不仅为我扫除了生活和研究上困扰，也夯实了短暂学术生涯的根基，砥砺我在人生的道路上不断前行。同时也要感谢华中科技大学的栽培，从本科到研究生的这七年，我在喻家山下度过了人生最重要的一段时光，收获了巨大的成长。

我还要感谢实验室的同学们，在一起互相讨论、学习、合作、激励的日子里，我们共同成长。感谢我的室友齐先智、石京磊和肖标。这三年来与你们朝夕相处，感谢你们对我的宽容和帮助。感谢从本科开始就和我在一起的朋友们李炽、赵帅、林雪峰、付波、庞仁坤、谢斌、李梅、杜雅丽、李冰倩、饶振环、袁保建、熊明康、李启雄和张成昱等人，感谢你们在我春风得意马蹄疾时一起看尽繁花，在折戟沉沙时也能够陪在我的身旁给予鼓励。虽然大家走向了不同的地方、不同的道路，但时间和距离并有拉开我们，愿我们的友谊长存。感谢我的家人，你们是我永远的后盾和依靠。

感谢我的同学、朋友和家人们，你们在学习和生活中给予了我非常多的鼓励、支持与陪伴，是你们的帮助使我渡过了这充实、美好的研究生时光。

参考文献

[1] Brody H. Medical imaging. Nature, 2013, 502(7473):S81–S81.

[2] Shao Y, Gao Y, Guo Y, et al. Hierarchical lung field segmentation with joint shape and appearance sparse learning. IEEE Transactions on Medical Imaging, 2014, 33(9):1761–1780.

[3] Yap P, Zhang Y, Shen D. Multi-tissue decomposition of diffusion MRI signals via ℓ_0 sparse-group estimation. IEEETransactions on Image Processing, 2016, 25(9):4340–4353.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June, 2016, 770-778.

[5] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. in: International Conference on Machine Learning (ICML), New York City, June, 2016, 173–182.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. in: International Conference on Learning Representations (ICLR), San Diego, CA, May, 2014.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553):436–444.

[8] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 1943, 5(4):115–133.

[9] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65(6):386.

[10] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Nature, 1986, 323(6088):533–536.

[11] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. in: Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, December, 2012, 1097-1105.

[12] LeCun Y, Boser B E, Denker J S, et al. Handwritten digit recognition with a back-propagation network. in: Advances in Neural Information Processing Systems (NIPS), Denver, CO, November, 1990, 396–404.

[13] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8):1735–1780.

[14] Chung J, Gülçehre Ç, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, 2014, abs/1412.3555.

[15] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 1982, 79(8):2554–2558.

[16] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention. in: International Conference on Machine Learning (ICML), Lille, France, July, 2015, 2048–2057.

[17] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend. in: Advances in Neural Information Processing Systems (NIPS), Montreal, Canada, December, 2015, 1693–1701.

[18] Korolev S, Safiullin A, Belyaev M, et al. Residual and plain convolutional neural networks for 3D brain MRI classification. in: International Symposium on Biomedical Imaging (ISBI), Melbourne, Australia, April, 2017, 835-838.

[19] Brosch T, Tang L Y W, Yoo Y, et al. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Transactions on Medical Imaging, 2016, 35(5):1229–1239.

[20] Wu G, Kim M, Wang Q, et al. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Transactions on Biomedical Engineering, 2015, 63(7):1505–1516.

[21] Suk H I, Lee S W, Shen D, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage, 2014, 101:569–582.

[22] Shin H C, Roberts K, Lu L, et al. Learning to read chest X-Rays: Recurrent neural cascade model for automated image annotation. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June, 2016, 2497–2506.

[23] Dou Q, Chen H, Yu L, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Transactions on Medical Imaging, 2016, 35(5):1182–1195.

[24] Ram S, Seirawan H, Kumar S K, et al. Prevalence and impact of sleep disorders and sleep habits in the United States. Sleep and Breathing, 2010, 14(1):63–70.

[25] Carskadon M A, Rechtschaffen A. Monitoring and staging human sleep. Principles and Practice of Sleep Medicine, 2000, 3:1197–1215.

[26] Rechtschaffen A. A manual of standardized terminology, techniques and scoring systems for sleep stages of human subjects. National Institute of Health, 1968, 1:204–210.

[27] Berry R B, Brooks R, Gamaldo C E, et al. The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine, 2012.

[28] Rosenberg R S, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. Journal of Clinical Sleep Medicine, 2013, 9(1):81–87.

[29] Fraiwan L, Lweesy K, Khasawneh N, et al. Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. Computer Methods and Programs in Biomedicine, 2012, 108(1):10–19.

[30] Boostani R, Karimzadeh F, Nami M. A comparative review on sleep stage classification methods in patients and healthy individuals. Computer Methods and Programs in Biomedicine, 2017, 140:77–91.

[31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. in: International Conference on Learning Representations (ICLR), San Diego, CA, May, 2015, 1–14.

[32] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June, 2015, 4694–4702.

[33] Kim Y. Convolutional neural networks for sentence classification. in: Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October, 2014, 1746–1751.

[34] Tsinalis O, Matthews P M, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Annals of Biomedical Engineering, 2016, 44(5):1587–1597.

[35] Chambon S, Galtier M, Arnal P, et al. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018, 26(4):758–769.

[36] Tsinalis O, Matthews P M, Guo Y, et al. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. CoRR, 2016, abs/1610.01683.

[37] Sors A, Bonnet S, Mirek S, et al. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomedical Signal Processing and Control, 2018, 42:107–114.

[38] Supratak A, Dong H, Wu C, et al. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017, 25(11):1998–2008.

[39] Biswal S, Kulas J, Sun H, et al. SLEEPNET: automated sleep staging system via deep learning. CoRR, 2017, abs/1707.08262.

[40] Phan H, Andreotti F, Cooray N, et al. Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Transactions on Biomedical Engineering, 2018, 66(5):1285–1296.

[41] Debener S, Emkes R, De Vos M, et al. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. Scientific Reports, 2015, 5:16743.

[42] Looney D, Goverdovsky V, Rosenzweig I, et al. Wearable in-ear encephalography sensor for monitoring sleep. Preliminary observations from nap studies. Annals of the American Thoracic Society, 2016, 13(12):2229–2233.

[43] Elkan C. The foundations of cost-sensitive learning. in: International Joint Conference on Artificial Intelligence (IJCAI), Seattle, WA, August, 2001, 973–978.

[44] Goldberger A L, Amaral L A, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 2000, 101(23):e215–e220.

[45] Kemp B, Zwinderman A H, Tuk B, et al. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Transactions on Biomedical Engineering, 2000, 47(9):1185–1194.

[46] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, 2012, abs/1207.0580.

[47] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. in: International Conference on Machine Learning (ICML), Lille, France, July, 2015, 448–456.

[48] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997, 45(11):2673–2681.

[49] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization. in: International Conference on Learning Representations (ICLR), San Diego, CA, May, 2015, 7–9.

[50] Acharya U R, Chua E C P, Chua K C, et al. Analysis and automatic identification of sleep stages using higher order spectra. International Journal of Neural Systems, 2010, 20(06):509–521.

[51] Fraiwan L, Lweesy K, Khasawneh N, et al. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. Methods of information in Medicine, 2010, 49(03):230–237.

[52] Güneş S, Polat K, Yosunkaya Ş. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. Expert Systems with Applications, 2010, 37(12):7922–7928.

[53] Liang S F, Kuo C E, Hu Y H, et al. Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. IEEE Transactions on Instrumentation and Measurement, 2012, 61(6):1649–1657.

[54] Abdul-Kareem S, Baba S, Zubairi Y Z, et al. Prognostic systems for NPC: a comparison of the multi layer perceptron model and the recurrent model. in: International Conference on Neural Information Processing (ICONIP), Singapore, Singapore, November, 2002, 271–275.

[55] Prasad U, Rampal L. Descriptive epidemiology of nasopharyngeal carcinoma in Peninsular Malaysia. Cancer Causes & Control, 1992, 3(2):179–182.

[56] Shanmugaratnam K, Tye C. A study of nasopharyngeal cancer among Singapore Chinese with special reference to migrant status and specific community (dialect group). Journal of Chronic Diseases, 1970, 23(5-6):433–441.

[57] Mohammed M A, Ghani M K A, Arunkumar N a, et al. Decision support system for nasopharyngeal carcinoma discrimination from endoscopic images using artificial neural network. The Journal of Supercomputing, 2020, 76:1086–1104.

[58] Chuang W Y, Chang S H, Yu W H, et al. Successful identification of nasopharyngeal carcinoma in nasopharyngeal biopsies using deep learning. Cancers, 2020, 12(2):507.

[59] Wu B, Khong P L, Chan T. Automatic detection and classification of nasopharyngeal carcinoma on PET/CT with support vector machine. International Journal of Computer Assisted Radiology and Surgery, 2012, 7(4):635–646.

[60] Zhao L, Lu Z, Jiang J, et al. Automatic nasopharyngeal carcinoma segmentation using fully convolutional networks with auxiliary paths on dual-modality PET-CT images. Journal of digital imaging, 2019, 32(3):462–470.

[61] King A D, Vlantis A C, Bhatia K S, et al. Primary nasopharyngeal carcinoma: diagnostic accuracy of MR imaging versus that of endoscopy and endoscopic biopsy. Radiology, 2011, 258(2):531–537.

[62] Pinaya W H, Gadelha A, Doyle O M, et al. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. Scientific reports, 2016, 6:38897.

[63] Han X, Zhong Y, He L, et al. The unsupervised hierarchical convolutional sparse auto-encoder for neuroimaging data classification. in: International Conference on Brain Informatics and Health (ICBIH), London, UK, August, 2015, 156–166.

[64] Schmah T, Hinton G E, Small S L, et al. Generative versus discriminative training of RBMs for classification of fMRI images. in: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, December, 2009, 1409–1416.

[65] Huang W, Chan K L, Zhou J. Region-based nasopharyngeal carcinoma lesion segmentation from MRI using clustering-and classification-based methods with learning. Journal of Digital Imaging, 2013, 26(3):472–482.

[66] Nabizadeh N, Kubat M. Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features. Computers & Electrical Engineering, 2015, 45:286–301.

[67] Ma Z, Zhou S, Wu X, et al. Nasopharyngeal carcinoma segmentation based on enhanced convolutional neural networks using multi-modal metric learning. Physics in Medicine & Biology, 2019, 64(2):025005.

[68] Ming-Chi Wu, Wen-Chi Chin, Ting-Chen Tsan, et al. The benign and malignant recognition system of nasopharynx in MRI image with neural-fuzzy based adaboost classifier. in: International Conference on Information Management (ICIM), London, UK, May, 2016, 47–51.

[69] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1):62–66.

[70] Zhou B, Khosla A, Lapedriza À, et al. Learning deep features for discriminative localization. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June, 2016, 2921–2929.

[71] Wang Y, Wu D. Deep learning for sleep stage classification. in: Chinese Automation Congress (CAC), Xi'an, China, November, 2018, 3833-3838.

[72] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. in: Soviet Mathematics Doklady, 1983, 372–376.

[73] Huang B, Ren X, Wu L, et al. Elevated progesterone levels on the day of oocyte maturation may affect top quality embryo IVF cycles. PLoS One, 2016, 11(1):e0145895.

[74] Huang B, Hu D, Qian K, et al. Is frozen embryo transfer cycle associated with a significantly lower incidence of ectopic pregnancy? an analysis of more than 30,000 cycles. Fertility Sterility, 2014, 102(5):1345–1349.

[75] Huang B, Qian K, Li Z, et al. Neonatal outcomes after early rescue intracytoplasmic sperm injection: an analysis of a 5-year period. Fertility Sterility, 2015, 103(6):1432–1437.

[76] Reproductive Medicine A S, Embryology E S I G. The istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Human Reproduction, 2011, 26(6):1270–1283.

[77] Tomasz B, Rafal K, Wojciech G. Methods of Embryo Scoring in In Vitro Fertilization. Reproductive Biology, 2004, 4(1):5–22.

[78] Holte J, Berglund L, Milton K, et al. Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. Human Reproduction, 2006, 22(2):548–557.

[79] Lemmen J, Agerholm I, Ziebe S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. Reproductive Biomedicine Online, 2008, 17(3):385–391.

[80] Wong C C, Loewke K E, Bossert N L, et al. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. Nature Biotechnology, 2010, 28(10):1115–1121.

[81] Herrero J, Tejera A, Albert C, et al. A time to look back: analysis of morphokinetic characteristics of human embryo development. Fertility Sterility, 2013, 100(6):1602–1609.

[82] Chen A A, Tan L, Suraj V, et al. Biomarkers identified with time-lapse imaging: discovery, validation, and practical application. Fertility Sterility, 2013, 99(4):1035–1043.

[83] Kirkegaard K, Agerholm I E, Ingerslev H J. Time-lapse monitoring as a tool for clinical embryo assessment. Human Reproduction, 2012, 27(5):1277–1285.

[84] Meseguer M, Herrero J, Tejera A, et al. The use of morphokinetics as a predictor of embryo Iimplantation. Human Reproduction, 2011, 26(10):2658–2671.

[85] Wang Y, Moussavi F, Lorenzen P. Automated embryo stage classification in time-lapse microscopy video of early human embryo development. in: Int'l Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Nagoya, Japan, September, 2013, 460–467.

[86] Conaghan J, Chen A A, Willman S P, et al. Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: results from a prospective multicenter trial. Fertility and Sterility, 2013, 100(2):412–419.

[87] Jonaitis D, Raudonis V, Lipnickas A. Application of numerical intelligence methods for the automatic quality grading of an embryo development. International Journal of Computing, 2016, 15(3):177–183.

[88] Khan A, Gould S, Salzmann M. Segmentation of developing human embryo in time-lapse microscopy. in: Int'l Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, April, 2016, 930–934.

[89] Ng N H, McAuley J, Gingold J A, et al. Predicting embryo morphokinetics in videos with late fusion nets & dynamic decoders, May, 2018. https://openreview.net/forum?id=By1QAYkvz.

[90] Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell counting. in: European Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, October, 2016, 339–348.

[91] VerMilyea M D, Tan L, Anthony J T, et al. Computer-automated time-lapse analysis results correlate with embryo implantation and clinical pregnancy: a blinded, multi-centre study. Reproductive Biomedicine Online, 2014, 29(6):729–736.

[92] Diamond M P, Suraj V, Behnke E J, et al. Using the Eeva Test adjunctively to traditional day 3 morphology is informative for consistent embryo assessment within a panel of embryologists with diverse experience. Journal of Assisted Reproduction and Genetics, 2015, 32(1):61–68.

[93] Aparicio-Ruiz B, Basile N, Albalá S P, et al. Automatic time-lapse instrument is superior to single-point morphology observation for selecting viable embryos: retrospective study in oocyte donation. Fertility and Sterility, 2016, 106(6):1379–1385.

[94] Kieslinger D C, De Gheselle S, Lambalk C B, et al. Embryo selection using time-lapse analysis (Early Embryo Viability Assessment) in conjunction with standard morphology: a prospective two-center pilot study. Human Reproduction, 2016, 31(11):2450–2457.

[95] Ruder S. An Overview of Multi-Task Learning in Deep Neural Networks. CoRR, 2017, abs/1706.05098.

[96] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Miami Beach, FL, June, 2009, 248-255.

[97] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, 2012, abs/1212.0402.

[98] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, June, 2014, IEEE, 1725-1732.

[99] Caba Heilbron F, Escorcia V, Ghanem B, et al. Activitynet: a large-scale video benchmark for human activity understanding. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June, 2015, IEEE, 961–970.

[100] Ng J Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification. in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June, 2015, IEEE, 4694-4702.

[101] Zhou Z H. Ensemble methods: foundations and algorithms. Boca Raton, FL: CRC press, 2012.

[102] Liu Y, Chapple V, Roberts P, et al. Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the embryoscope time-lapse video system. Fertility Sterility, 2014, 102(5):1295–1300.

[103] Kalatova B, Jesenska R, Hlinka D, et al. Tripolar mitosis in human cells and embryos: Occurrence, pathophysiology and medical implications. Acta histochemica, 2015, 117(1):111–125.

[104] Kavukcuoglu K, Sermanet P, Boureau Y, et al. Learning convolutional feature hierarchies for visual recognition. in: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, December, 2010, 1090–1098.

[105] Gregor K, LeCun Y. Learning fast approximations of sparse coding. in: International Conference on Machine Learning (ICML), Haifa, Israel, June, 2010, 399–406.

[106] Ranzato M, Mnih V, Susskind J M, et al. Modeling natural images using gated MRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9):2206–2222.

[107] Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models. in: Advances in Neural Information Processing Systems (NIPS), Montreal, Canada, December, 2014, 3581–3589.

[108] Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 2018, 73:1–15.

[109] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. in: European Conference on Computer Vision (ECCV), Zurich,Switzerland, September, 2014, Springer, 818–833.

[110] Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. in: International Conference on Learning Representations (ICLR), New Orleans, LA, May, 2019.

[111] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition. in: International Conference on Machine Learning (ICML), Beijing, China, June, 2014, 647–655.

[112] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345–1359.

[113] Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging, 2016, 35(5):1285–1298.

[114] Tan C, Sun F, Kong T, et al. A survey on deep transfer learning. in: International Conference on Artificial Neural Networks (ICANN), Rhodes, Greece, October, 2018, 270–279.

附录 1　攻读学位期间发表论文目录

[1] **Yang Wang** and Dongrui Wu. Real-Time fMRI-Based Brain Computer Interface: A Review. in: International Conference on Neural Information Processing (**ICONIP**), Guangzhou, China, November. 2017, 833-842.

[2] **Yang Wang** and Dongrui Wu. Deep Learning for Sleep Stage Classification. in: Chinese Automation Congress (**CAC**), Xi'an, China, November. 2018, 3833-3838.

[3] Shuai Zhao, **Yang Wang**, et al. Region Mutual Information Loss for Semantic Segmentation. in: Advances in Neural Information Processing Systems (**NeurIPS**), Vancouver, Canada, December. 2019, 11115-11125.

[4] Liu Z, Huang B, Cui Y, et al. Multi-Task Deep Learning with Dynamic Programming for Embryo Early Development Stage Classification from Time-Lapse Videos. IEEE Access, 2019, 7:122153 – 122163.