中国科学技术大学

# 博士学位论文


## 社交多媒体数据语义理解和关联表达

作者姓名：　　吴岳

学科专业：信号与信息处理

导师姓名：　俞能海 教授

完成时间：二〇一七年五月

University of Science and Technology of China

A dissertation for doctor's degree


# Social Media Data Semantic Understanding and Associative Expression


| | |
|---|---|
| Author's Name: | Yue Wu |
| Speciality: | Signal and Information Processing |
| Supervisor: | Prof. Nenghai Yu |
| Finished Time: | May, 2017 |

# 中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____　　签字日期：_____

# 中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定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，可以将学位论文编入《中国学位论文全文数据库》等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、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

□公开　□保密（____ 年）

作者签名：_____　　导师签名：_____

签字日期：_____　　签字日期：_____

# 摘　要

近几年来，智能手机及其他智能移动设备呈现出了爆发式的增长与普及。高清摄像头，大容量存储，和高速的网络连接为用户创造了极其便利的拍摄和分享条件。用户几乎可以在任何时间任何地点拍摄图片或者视频，并将它们分享到社交网络或存储到云端服务器，产生了海量的社交多媒体数据。然而这些数据都以碎片化的形式存在于社交媒体上，缺乏智能的工具或服务将他们组织起来，并根据用户需求选取并呈现给用户，用户也很难快速准确地搜索到他们需要的数据。因此，如何充分挖掘并有效利用社交多媒体数据成为了当前重要的研究问题。

本论文对社交多媒体数据的语义理解和关联表达做了深入的研究，目标是实现一个能够理解社交多媒体数据，并根据用户需求选取有关联的数据，以丰富的表达形式呈现给用户的系统。由于社交多媒体数据的语义概念丰富多样，无法对每个语义收集数据并标注，语义理解首先需要解决标注难的问题。其次，由于社交多媒体数据的规模庞大，语义理解需要解决处理慢的问题。社交多媒体数据关联表达是基于对社交多媒体数据语义理解的结果，根据用户个性化的需求选取有关联的数据，并以丰富的表达形式呈现给用户。本文分别从图片和视频的角度，研究了关联表达的具体应用。语义理解和关联表达构成了挖掘和利用社交多媒体数据相对完整的框架。本文对相关的问题做了深入研究，取得了以下成果：

(1) 对于语义理解标注难的问题，提出一种弱监督深度相关反馈学习算法，直接从弱标注的社交多媒体数据中学习语义理解模型。传统深度学习算法对于训练数据中的标注噪音十分敏感，本论文基于感知连续性，利用数据在特征空间上的关系，使得不同的训练数据在训练过程中有不同的贡献加权，从而抑制噪音标注的影响。为了加速模型训练的速度，进一步对相关反馈网络进行了简化和近似，降低了模型训练的复杂度。与已有算法相比，本文提出的相关反馈神经网络具有更好的噪声鲁棒性。

(2) 对于语义理解处理慢的问题，提出了一种从大规模高维数据中选取特征的高效算法。利用二阶在线学习算法，基于特征的置信度进行特征选择，并利用最大最小堆的结构提出了快速特征选取算法。由于训练过程中置信度的单调递特性，进一步提出了快速二阶在线特征选取算法，将算法的复杂度降低为于非零特征数目成正比。相比于已有算法，算法的时间开销减小了几倍至几十倍。

(3) 简化深度模型，基于二阶特征选择算法提出了基于在线特征选取的深度神经网络模型简化算法。算法对卷积层输出的每个通道的神经元增加了权重层，在权重层上进行特征选取，从而将三维的卷积核组稀疏优化问题转化为一维特征选取问题。利用二阶在线特征选取算法，在不损失模型准确率的情况下极大地减少了模型的参数。

(4) 对于图片关联表达问题，提出了一个基于主题的个人照片集故事化关联表达系统——Monet。系统首先根据照片的时间和位置信息对照片集进行事件检测。其次，根据照片的质量，多样性和事件的均衡性选取一部分代表性的图片。然后，利用弱监督的深度学习算法对代表性照片进行内容分析，并利用在线特征选取算法选取最能鉴别照片语义的特征子集。系统设计了 17 个主题风格，利用照片特征对照片赋予不同的主题。最后，通过不同主题风格的可计算的视频编辑语法，对照片进行动画特效处理以及音乐的匹配，最终生成具有关联表达能力的视频呈现给用户。

(5) 对于视频关联表达问题，提出了全自动的移动多摄像头视频自动剪辑系统——MoVieUp。我们邀请了专业的视频编辑人员探讨可计算的视频编辑语法。自动剪辑系统首先对音频流进行质量评估，在最少切换次数下选取高质量的音频流，拼接成单一音频流。对于视频流，首先将多摄像头视频进行语义分割，得到视频子镜头，其次对这些子镜头的视觉质量，运动，以及相互之间的多样性进行评估，最终在保证镜头运动一致性的前提下，最大化质量和多样性，选取视频镜头。对于镜头切换时机的选取，则根据音频的节奏以及语义特性，对切换频率进行匹配。系统最后将单一的音频流和单一的视频流进行混流，得到最终剪辑好的视频呈现给用户。

**关键词：**社交多媒体数据　语义理解　弱监督深度学习　特征选取　模型简化　关联表达　照片集编辑　视频自动剪辑

# ABSTRACT

Recent years have witnessed the explosive growth and popularity of mobile devices. The high resolution cameras, large storage, and fast network connection of mobile devices have founded the superior conditions for capturing and sharing. Users can capture photos or videos and share them to social networks or clouds at almost anytime and anywhere. Up to now, the amount of social media data has increased to a huge scale. However, these data exist in a fragmented way on social media, lacking intelligent services to organize them. Neither can social media provide data according to personalized user needs, nor can users search for the required data efficiently and effectively. As a result, how to exploit and utilize the large scale social media data has become an important problem.

This thesis probes into the semantic understanding and associative expression of social media data. The aim is to implement an intelligent system that can understand, select, and show social media data in an expressive way. Due to the wide range of semantics, it's hard to collect and label data for every semantic tag. Semantic understanding should solve the difficulty of labelling. Besides, it needs to accelerate the processing speed due to the large data scale. Based on the semantic understanding, associative expression selects and shows social media data in an expressive way according to personalized user needs. We studies associative expression from photo and video aspects. Semantic understanding and associative expression compose a relatively complete framework for mining and utilizing social media data. This thesis conducts a deep research on the related problems with the following achievements:

(1) For the difficulty of labelling, we propose a weakly supervised deep learning algorithm with relevance feedback to learn from the weakly labelled social media data directly. Traditional deep learning algorithms are sensitive to the label noises in training data. Our algorithm is based on the perceptual consistency to attenuate the sensitiveness. It utilizes the correlation in the feature space so that different training samples contribute differently. To speed up training, we further simplifies the algorithm to reduce the training complexity. Compared to existing algorithms, our relevance feedback algorithm shows better robustness to label noises.

(2) For the processing speed, we propose a large scale high dimensional second-order online feature selection algorithm. Based on the second-order online learning algorithms, we select features according to the confidence of features. We propose fast algorithms with the Max/Min heap. Due to the monotonous increasing property of confidence, we further propose the fast second-order online feature selection algorithm which reduces the complexity to be linear to the number of non-zero features. Compared to existing algorithms, the training time cost is less by orders of magnitude.

(3) For model simplification, we propose an online deep model simplification algorithm based on online feature selection. The algorithm adds a new weighting layer for each channel of the output feature maps of the convolutional layer. As a result, the traditional group sparsity problem on the 3D convolutional kernels is transformed into the feature selection problem on a 1D weighting vector. Levering the second-order online feature selection algorithm, model parameters are reduced significantly with little impact on accuracy.

(4) For associative photo expression, we propose a theme-based personal photo storytelling system—Monet. First, the system detects events in personal photos according to the time and location information. It then selects a representative photo subset according to photo quality, diversity, and balance of events. After that, we use the weakly supervised relevance feedback algorithm to analyze the content of the representative photos. Online feature selection algorithm is applied to extract the most distinctive features. With these features, each photo is assigned to one of the 17 theme styles designed for the system. Finally, a fancy video with animation and motion effects is generated and aligned with a music according to the computational filming grammars of each theme style.

(5) For associative video expression, we propose an automatic mobile multi-camera video mashup system—MoVieUp. We invites professional video editors to exploit computational filming grammars. First, the system assesses quality of the audio streams. Under the less switching principle, it selects high quality audio segments and stitches them into a single audio stream. For video streams, they are first segmented into subshots. The system then evaluates the quality, motion, and diversity of the subshots. To select video shots, we maximize the quality and diversity under the condition of motion consistency. The switching points of video shots are

detected according to the tempo and semantics of audio. Finally, the system multiplexes the audio and video streams to generate the well-edited video.

**Key Words:** Social Media Data, Semantic Understanding, Weakly Supervised Deep Learning, Model Simplification, Feature Selection, Associative Expression, Photo Storytelling, Video Mashup

# 目 录

摘要 ····· I
Abstract ····· III
第 1 章 绪论 ····· 1
1.1 社交多媒体数据研究意义 ····· 1
1.2 社交多媒体数据关键问题 ····· 3
1.3 本文的主要工作 ····· 4
1.4 本文的主要创新点 ····· 7
1.5 本文的结构安排 ····· 8
第 2 章 国内外研究现状和工作基础 ····· 9
2.1 弱监督学习 ····· 9
2.1.1 数据去噪 ····· 9
2.1.2 鲁棒噪音模型 ····· 10
2.2 特征选取 ····· 12
2.2.1 批处理方法 ····· 12
2.2.2 在线特征选取 ····· 13
2.3 模型简化 ····· 14
2.4 社交多媒体数据的关联表达 ····· 15
2.4.1 照片集关联表达 ····· 16
2.4.2 移动多摄像头视频关联表达 ····· 17
第 3 章 弱监督社交多媒体数据语义理解 ····· 21
3.1 弱监督目标识别问题建模 ····· 21
3.2 相关反馈弱监督深度神经网络 ····· 22
3.2.1 经典卷积神经网络 ····· 22
3.2.2 相关反馈卷积神经网络 ····· 23
3.2.3 相关反馈分析 ····· 27
3.3 实验结果和评估 ····· 27
3.3.1 目标识别 ····· 27
3.4 小结 ····· 30

第 4 章 大规模社交多媒体数据快速处理 ····· 31
4.1 二阶在线特征选取 ····· 31
4.2 置信度加权二阶在线特征选取 ····· 32
4.3 二阶在线特征选取快速算法 ····· 35
4.3.1 一阶快速特征选取算法 ····· 37
4.3.2 二阶快速特征选取算法 ····· 37
4.3.3 复杂度分析 ····· 39
4.4 二阶多类在线特征选取 ····· 41
4.5 二阶在线特征选取实验评估 ····· 42
4.5.1 实验设置 ····· 42
4.5.2 人工数据集实验评估 ····· 43
4.5.3 大规模真实数据集实验评估 ····· 46
4.5.4 图片检索中的应用 ····· 48
4.6 深度卷积神经网络模型简化 ····· 48
4.6.1 深度卷积网络模型简化建模 ····· 48
4.6.2 基于在线学习的模型简化 ····· 48
4.6.3 实验结果和评估 ····· 48
4.7 小结 ····· 48
第 5 章 照片集关联表达 ····· 49
5.1 照片集关联表达系统框架 ····· 49
5.2 照片集事件检测 ····· 49
5.3 照片集照片筛选 ····· 49
5.4 照片集风格选取 ····· 49
5.5 照片集故事合成 ····· 49
5.6 实验结果 ····· 49
5.7 小结 ····· 49
第 6 章 移动多摄像头视频自动剪辑 ····· 51
6.1 可计算视频编辑语法 ····· 51
6.2 移动多摄像头视频自动剪辑系统框架 ····· 51
6.3 音频剪辑 ····· 51
6.4 镜头切换点检测 ····· 51
6.5 视频镜头选取 ····· 51
6.6 实验结果 ····· 51
6.7 小结 ····· 51

第 7 章 总结与展望 ····· 53
7.1 本文总结 ····· 53
7.2 研究工作展望 ····· 53
参考文献 ····· 55
致谢 ····· 65
在读期间发表的学术论文与取得的研究成果 ····· 67

# 图目录

1.1 社交多媒体数据的产生和利用现状 ····· 2
1.2 社交多媒体数据的研究内容和相互关联 ····· 5

3.1 数据标注噪音类型: (a) 完全随机噪音; (b) 随机噪音; (c) 非随机噪音··· 21
3.2 深度卷积神经网络的结构图 ····· 23
3.3 相关反馈卷积神经网络 ····· 25
3.4 训练数据对于梯度的贡献曲线，横坐标为数据与其他数据的距离 ···· 27
3.5 参数 $\alpha$ 对于卷积神经网络分类性能的影响 ····· 29
3.6 不同噪音程度下不同方法的准确率相对于无噪声的下降程度比较 ···· 29

4.1 合成数据集 $X_1$ 和 $X_2$ 上测试准确率和特征数目之间的关系 ····· 44
4.2 Time cost versus number of selected features on synthetic datasets $\mathcal{X}_1$ and $\mathcal{X}_2$ ····· 45
4.3 “news” 和 “rcv1” 数据集上测试准确率和训练时间与选取特征个数之间的关系 ····· 47

# 表目录

2.1 照片集关联表达系统比较 ····· 17
2.2 视频自动剪辑系统比较 ····· 18

3.1 不同算法在 CIFAR-10 数据集上不同噪音条件下的准确率比较····· 30
3.2 不同算法在 VOC2007 数据集上不同噪音条件下的不同类别平均准确率比较 ····· 30

4.1 合成数据信息（“K”,“M”,“B” 分别代表千，百万，十亿）····· 44
4.2 SOFS 算法可伸缩性评测 ····· 45
4.3 大规模真实数据集信息 ····· 46
4.4 大规模高维数据集评测结果（$\rho$ 是选取的特征比例）····· 46

# 算法索引

4.1 PET：截断感知机算法框架 . . . . . 33
4.2 Truncate: 截断函数 . . . . . 33
4.3 FOFS：一阶在线特征选取算法 . . . . . 34
4.4 二阶在线特征选取的算法框架 . . . . . 36
4.5 快速一阶在线特征选取算法 . . . . . 38
4.6 SOFS：快速二阶在线特征选取算法 . . . . . 40

# 第 1 章　绪论

本章首先介绍社交多媒体数据的含义以及对于社交多媒体数据的研究意义，并由此引出社交多媒体数据的研究中存在的关键问题。然后介绍本文的主要工作及创新点，最后介绍全文的结构安排。

## 1.1　社交多媒体数据研究意义

近几年来，智能手机及其他智能移动设备呈现出了爆发式的增长与普及。高清摄像头，大容量存储，和高速的网络连接为用户创造了极其便利的拍摄条件，从而创造了海量的个人多媒体数据。用户几乎可以在任何时间任何地点拍摄图片或者视频，并将它们分享到社交网络或存储到云端服务器。据统计，截止到 2014 年产生了大约 2.7 万亿的用户图片。到 2017 年，该数字将会增长到约 4.9 万亿[1]。这类由个人用户产生的多媒体数据具有明显的社交性特点：

1. 用户通常在社交活动中拍摄视频或者图片，同一时间段内其他用户也会拍摄与之相关的内容；
2. 大量的个人多媒体数据被用户分享到 Flickr，Instagram，YouTube，美拍，优酷等社交网站，这些数据本身包含的时间、地点等信息与其他用户分享的个人多媒体数据产生关联。

因此，本文将这类用户数据统称为**社交多媒体数据**。

相比于传统的多媒体数据，个人用户产生的多媒体数据具有以下特点:

1. **质量不确定**：由于个人用户拍摄技巧比较业余，拍摄时的光线限制，相机的快速移动或者抖动，场景快速变换等原因，个人多媒体数据通常伴有抖动，散焦，过度曝光/欠曝光，模糊，遮挡以及拍摄出无意义的视频或图片等问题，影响到个人多媒体数据的后期浏览体验。
2. **内容冗余**：用户在拍照时通常采取多次拍摄的方式来获得最理想的拍照效果。一般情况下，用户会在短时间内选择最好的照片或视频分享到社交网络，却很少删除其他内容非常相似的数据，使得个人多媒体数据中有大量的冗余。这些冗余数据带来三个方面的主要问题：1）需要耗费大量的时间和精力去整理；2）占据了大量的存储资源；3）增加了用户查找数据的难度。

3. **多样性**：用户拍摄的时间、地点和环境比较随机，拍摄的角度，内容具有很大的不确定性，使得个人多媒体数据的内容表现出多样化的特点。个人多媒体数据相关的服务，不仅需要去除冗余的用户数据，同时也需要最大程度的保留用户数据的多样性。

4. **故事性**：用户的拍摄行为并不是完全随机的，而是选择性地记录对于他们有意义的时刻和场景。相同时间段内相近地点拍摄的个人多媒体数据，浓缩了用户在一段时间内的足迹，以及经历的活动。而完整的个人多媒体数据则记录了很长一段时间内发生在用户身上的故事。然而，当前的个人多媒体数据服务还不能很好的将这些记录的原始片段，整理成高观赏性的纪录片。

5. **具有位置和时间信息**：现代移动设备拍摄时一般都能检测到拍摄的时间、地点等信息，并将这些信息存储在图像或者视频的文件头中。这些时间和地点信息记录了拍摄时的时空上下文关系，使得数据之间能够产生关联。基于用户历史拍摄的时间、地点和具体内容，可以更好的挖掘出发生在用户身上的故事。

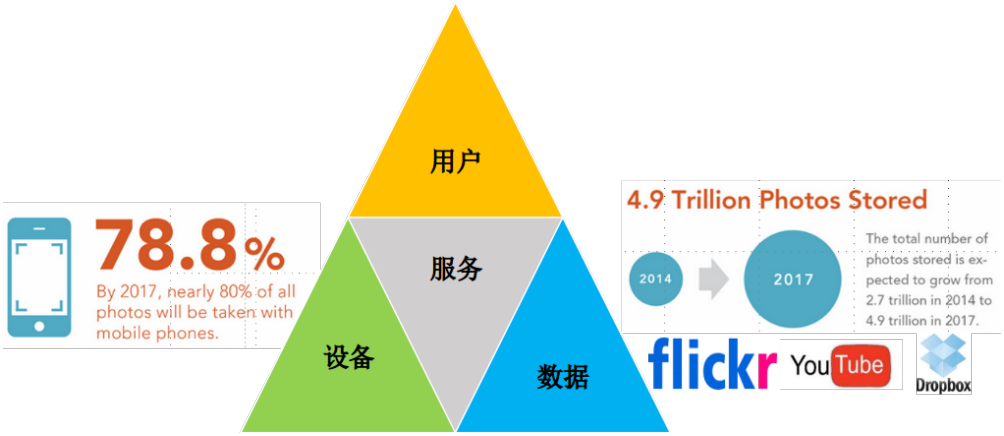



**图 1.1**　社交多媒体数据的产生和利用现状

然而，由于缺乏智能的社交多媒体数据建模和表达服务，这些海量的数据所包含的信息并没有得到充分的挖掘和利用。如图 1.1所示，一方面，社交多媒体数据呈现出并保持着爆炸式的增长趋势；另一方面，这些记录了珍贵记忆的用户数据被大量的存在本地或者云端磁盘上，却很少再次被用户浏览。社交多媒体数据中所蕴含的故事性和社交性，并没有带来更好的用户体验。随着人们对于多媒体内容品质和个性化需求的提高，越来越多的研究人员和工业界研究机构投入更多的精力到社交多媒体数据中。此外，由于社交多媒体数据的数量庞大，质量不确定、内容冗余、多样，并带有丰富的社交信息，对相关的处理算法提出了更

复杂的要求。因此，对于社交多媒体数据的研究已成为计算机视觉领域的研究热点，相关的成果既有利于推动计算机视觉以及多媒体领域相关课题的创新，对于用户体验的提升以及工业界的发展都也具有重要的应用价值和现实意义。

## 1.2　社交多媒体数据关键问题

社交多媒体数据的研究涉及内容的语义理解以及应用层的关联表达两个大的方面。其中，语义理解面临的主要问题包括：

- 解决标注难的问题。机器学习算法可以归纳为有监督学习，半监督学习，和无监督学习三个类别。有监督学习通常能够获得最好的分类效果。然而，它依赖大量准确的标注数据，在当前大数据的背景下有监督学习的应用成本十分高昂。据了解，目前最大的有标注图片数据集 ImageNet[2] 花费了大约 25,000 名用户一年左右的时间完成标注。尽管如此，ImageNet 仅包含 22,000 个类别，与现实应用中的语料库相差甚远（如 WordNet [1]）。与之相反，半监督学习和无监督的学习不需要大量的标注，但是分类的效果与有监督学习还有明显的差距。基于以上问题，越来越多的研究人员将注意力放在了弱监督学习上 (Weakly Supervised Learning)。弱监督学习是指从标注不完备、不精确的大规模噪音数据中, 充分挖掘有价值的信息，滤除或抑制错误信息，从而达到学习模型的目的。因此，弱监督学习能够充分利用大规模有噪音的标注数据，解决有监督学习标注难和无监督学习性能差的问题。

- 解决处理慢的问题。社交多媒体数据的规模十分庞大，对于模型的复杂度以及硬件的计算水平都提出了很高的要求。此外，移动设备的计算能力、存储空间以及电池容量依然有限，提高社交多媒体数据的处理速度对于提升移动端的用户体验至关重要。为了提升社交多媒体数据的处理速度，一方面可以利用特征选取 (Feature Selection) 的方法减少特征提取的种类和数目。图片或视频的内容既包括传统的全局特征, 局部特征，也包括近年来提出的深度神经网络不同层产生的特征。对于不同的任务，有些特征具有很强的表现力，有些特征则十分冗余。因此，选取对具体任务最紧凑、最具有表征能力的特征作为数据内容的表达, 可以减少特征提取的种类和数目，从而提高处理速度。另一方面，可以对特征提取过程中用到的模型进行简化，减少每种特征提取的时间开销。例如，近年来深度学习网络在图片识别，物体检测等领域取得了非常好的效果，但是网络的深度和参数数目也

[1] http://wordnet.princeton.edu/

在不断增加，如何在不影响模型准确度的情况下简化深度网络模型已经成为了当前的热点问题。

社交多媒体数据关联表达是指根据用户个性化的需求，从社交多媒体数据中选择有关联的数据，并以一定的表达形式将这些关联的数据呈现给用户。它面临的主要问题包括：

- 内容的组织和选取。当前，社交多媒体数据以碎片化的形式，根据不同用户按照拍摄或上传的时间顺序存储在云端服务器上，社交网站以及搜索引擎根据根据用户提供的标签对数据进行索引，查询和检索。然而，社交多媒体数据存在质量不确定，内容冗余多样以及故事性等特点，高效的内容组织方式需要理解数据之间的关联性和故事性，从时间，位置，用户，内容，关联性等多个维度对数据进行组织。此外，针对用户个性化的需求，给用户返回最具有代表性的数据，同时返回低质量，重复的数据。例如，xxx 系统。。

- 可计算的编辑语法。内容的组织和选取只是将数据高效地组织在一起，并选取最能满足用户需求的数据。然而在实际应用中，需要在原始的社交多媒体数据的基础上以一种新的、富有艺术美感的形式重新呈现给用户。例如，Magisto [2] 系统能够给用户的图片和视频加上丰富的视频特效，并将视频和音乐的节奏进行匹配，生成一段类似专业编辑人员编辑的具有丰富表现力的音乐视频。专业的编辑人员在视频编辑中根据素材的内容以及需要表达的效果选取与之相适应的素材和特效对视频进行编辑，对于计算机，如何将专业编辑人员在编辑中运用到的规则和语法转化成可计算的规则和算法是社交多媒体关联表达面临的一个主要问题。由于表现形式以及编辑语法的多样性，挖掘和应用可计算的编辑语法也具有非常大的挑战性。

## 1.3　本文的主要工作

针对章节 1.2中提到的关键问题，本文分别对社交多媒体数据的语义理解和关联表达做了深入的研究，构成了社交多媒体数据挖掘和利用的一个相对完整的框架。图 1.2给出了本文研究的具体内容以及相顾志坚的关联。针对语义理解标准难的问题，弱监督深度学习可以直接从带有不准确标注的社交多媒体数据中学习语义分类模型，理解数据的内容。结合传统计算机视觉方法和弱监督深度学习得到的特征，特征选取针对具体的任务选取最紧凑，最有代表性的特征自己

[2] http://www.magisto.com

对数据内容进行表达，从而减少特征提取的种类和数目，加快大规模社交多媒体数据处理的速度。此外，弱监督深度学习需要耗费大量的计算资源，本论文结合特征提取算法，提出了深度模型简化算法，减少深度神经网络的计算时间和参数个数。对于社交多媒体数据的关联表达，本文从照片集关联表达和多摄像头视频自动剪辑两个方面做了具体的应用研究。其中，照片集关联表达对照片集中的事件进行分析检测，根据照片的质量和关联选取有代表性的照片，通过可计算的视频编辑语法，对照片集记性故事化的表达。移动多摄像头自动剪辑则将同一时间段同意地点不同用户拍摄的多摄像头视频在时间上进行同步，通过可计算的视频编辑于法，选取镜头和录音，将多摄像头视频剪辑成单一高质量的音视频流。

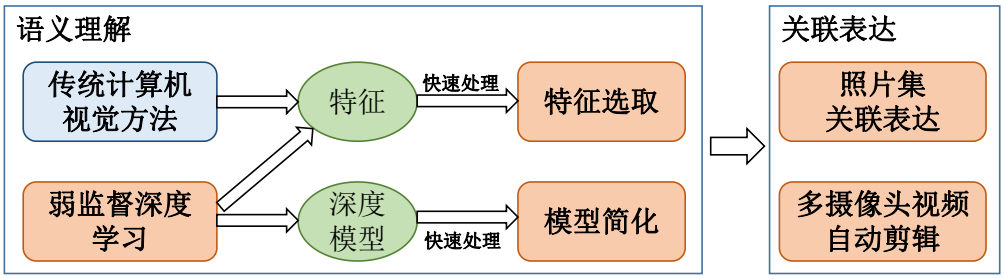



**图 1.2**　社交多媒体数据的研究内容和相互关联

**注**：橙色部分表示本文的研究工作

基于以上研究点，本论文具体的研究内容包括：

(1) 社交多媒体数据弱监督深度学习的算法研究：针对任意的图像类别，从互联网抓取弱标注数据，提出具有抗噪效果的图片分类模型。传统深度学习算法对于噪声的敏感性是由于所有的数据在学习过程中具有相同的权重。本论文提出的方法的基本假设是：同一类别下，正确标注的图片由于语义上关联，在特征空间上比较接近，而错误标注的图片则与其他图片有很大的差异性。因此，我们可以利用图片在特征空间上的关系，使得不同的训练数据在训练过程中有不同的贡献加权，从而使得在特征空间上“孤立”的图片具有较小的权重，特征空间上密集区域的图片具有较大的权重，我们称之为相关反馈。该思想可以通过在学习中进行特征空间的低秩分解实现。为了加速模型训练的速度，我们进一步通过一系列的推导和近似，最终降低模型的复杂度，用于大规模社交多媒体数据的模型训练。

(2) 大规模高维特征选取算法研究：多媒体数据特征表述不仅包括高层次的卷积神经网络特征，还包含低层次全局特征（如颜色特征，纹理特征），局部特征（如 SIFT，SURF）以及用来通过局部特征描述整体视觉信息的词袋特征等。实际应用中需要根据需求选取对目标任务最有用的特征子集，这对于大规模

社交多媒体数据的处理速度以及移动设备有限的计算能力和内存空间尤其重要。此外，去除特定任务不相干的特征，还可以提高特征的表达能力。本论文提出从大规模高维数据中选取特征的高效算法。已有稀疏在线特征选取算法的复杂度与特征的维度成正比，本论文利用二阶在线学习算法，基于特征的置信度进行特征选择，并利用最大最小堆结构提出快速特征选取算法。由于置信度的单调递减特性，本文进一步提出了快速二阶在线特征选择算法，将算法的复杂度降为与非零特征数目成正比。

(3) 深度学习模型简化算法研究：深度卷积神经网络的深度和模型参数通常比较大，例如经典的 VGG-16 网络包含超过 200M 的模型参数。大量的模型参数意味着在实际应用中需要大量的计算资源和时间，极大地限制了深度神经网络在大规模图片检索和图片识别等任务上的应用。此外，深度网络在移动设备上的应用已经成为一种趋势。由于移动设备计算能力的限制，在不影响模型准确度的条件下简化深度网络模型已经成为迫切的需要。本论文提出一种基于在线特征选取的模型简化算法。算法主要针对卷积层进行简化，对每一个卷积层增加一个卷积核的权重层。初始条件下所有卷积核的权重为 1，在学习过程中对权重进行更新，每次更新后利用稀疏在线特征选取算法将部分权重设为 0，保留剩余的权重。区别传统方法，稀疏在线学习的方法可以在训练过程中动态的调整需要保留的卷积核，从而减小模型简化对模型参数的影响。

(4) 照片集关联表达算法研究：照片集的关联表达首先对照片集进行事件检测，找出用户所拍摄的不同事件。其次，利用弱监督的深度学习算法进行照片的内容分析和模型简化。由于深度网络的每一层都是对照片不同层次的语义表达，本文将这些特征拼接到一起组成高维特征，并利用本文提出的在线特征选取算法选取最能鉴别照片语义的特征子集，构成照片内容的最终表达。为了达到更好的表达效果，不同风格的照片需要采用不同的编辑技巧。本文设计了 17 个设计风格，并利用弱监督深度神经网络从网络上抓取训练图片，得到设计风格的分类器，将照片集中的事件分配到不同的风格中。最后，通过可计算的视频编辑语法，对照片进行动画特效处理以及音乐的匹配，最终生成具有关联表达能力的视频呈现给用户。

(5) 移动多摄像头视频自动剪辑算法研究：多摄像头视频是指在同一时间段，同一地点由不同摄像头拍摄的时间上有重叠的一组视频。多摄像头视频是不同的人从不同的角度对相同事件的记录。本文提出一个全自动的移动多摄像头视频自动剪辑系统。我们首先邀请专业的视频编辑人员探讨可计算的视频编辑语法。根据这些语法，自动剪辑系统首先对音频流进行质量评估，在保证

尽可能减少音频流切换次数的条件下选取高质量的音频流，形成最终的单一音频流。对于视频流，首先将多摄像头视频进行语义分割，得到视频子镜头，其次对这些子镜头的视觉质量，运动，以及相互之间的多样性进行评估，最终在保证镜头运动一致性的前提下，最大化质量和多样性，选取视频镜头。对于镜头切换时机的选取，则根据音频的节奏以及语义特性，对切换频率进行匹配。系统最后将单一的音频流和单一的视频流进行混流，得到最终剪辑好的视频呈现给用户。

## 1.4　本文的主要创新点

本论文的主要创新点有以下几点：

- 针对社交多媒体数据标准难的问题，提出一种可以从互联网数据中学习图像类别的深度神经网络学习算法，摆脱了对大量标注数据的依赖。该算法利用数据本身之间的关联，使得不同的数据在模型训练中有不同的贡献加权。同时，通过一系列的推导和简化，最终的模型复杂度低，对于训练大量数据有重要意义。

- 针对大规模社交多媒体数据处理慢的问题，提出了从大规模高维数据中选取特征的高效算法。相比于已有的批处理算法和在线学习算法，该算法大大降低了特征选择的时间复杂度，同时，选取的特征具有和传统算法选出的特征差别不大甚至更好的表述能力。

- 针对深度模型处理慢，耗费计算资源的问题，提出了简化深度卷积神经网络的高效算法，将传统的多维数据组稀疏问题转化为经典的一维特征选取问题，在不影响模型准确率的情况下极大减少了模型参数。

- 提出了全新的照片集关联表达算法——Monet，从个人照片集中自动检测事件，选取有代表意义的图片，选取合适的视频编辑风格，并将设计师设计的视频编辑语法应用到照片集的编辑中。

- 提出移动多摄像头视频自动剪辑系统——MoVieUp，自动编辑移动多摄像头视频并生成一段综合音频视频流的方法。该方法首次考虑了音频流的剪辑，并且首次系统的讨论视频编辑理论在自动化方法中的应用。

## 1.5　本文的结构安排

本论文主要研究社交多媒体数据的语义理解和关联表达中的几个关键问题：弱监督深度学习，特征提取，模型简化，以及照片集关联表达和移动多摄像头自动剪辑。各章节内容安排如下：

- 第 2 章从弱监督学习，特征提取，模型简化，和关联表达四个方面介绍本论文相关工作的研究现状和工作基础。在弱监督学习方面，主要介绍了传统的弱监督学习方法和近年来热点研究的弱监督深度学习方法；在特征提取方面，主要回顾了传统的批处理方法，用于解决大规模流数据的在线学习方法，和在线特征提取算法；在模型简化方面，主要介绍与深度卷积神经网络相关的模型简化工作；在关联表达方面，主要介绍学术界和工业界关于社交多媒体数据的研究和代表性应用产品。

- 第 3 章重点介绍弱监督深度相关反馈网络的具体设计思路和实现方法，简化模型训练的理论证明和近似策略，以及相应的实验结果。

- 第 4 章介绍大规模社交多美数据快速处理的方法。首先介绍在线稀疏学习方法的基本模型，并介绍本文提出的置信度加权在线稀疏学习和置信度加权在线特征选取方法。本文从大规模标准数据集上验证算法的有效性，并在图片检索的实际任务中验证算法的可行性。其次，该部分介绍了深度学习模型简化的问题建模和具体算法, 以及相应的实验结果。

- 第 5 章介绍照片集关联表达系统，包括照片集中事件检测方法，代表性图片选取算法及相应的质量评估，多样性评价，和时间均衡性考量。在选取的代表性图片上，系统基于弱监督学习算法设计了不同风格的分类器，将照片集中的事件分配到不同的风格中，利用针对不同风格设计的可计算的编辑语法，对照片赋予丰富的特效，生成具有表现力的视频，重现照片集中的场景和故事。

- 第 6 章介绍移动多摄像头自动剪辑系统。系统首先介绍针对移动多摄像头视频自动剪辑的可计算的视频编辑语法，并提出相应的系统框架。该部分详细介绍了音频的质量评估和剪辑方法，基于音频节奏和语义内容的视频切换点检测算法和基于镜头质量，多样性，和运动连续性的镜头选取算法。最后，通过实验证明了系统各个部分的有效性和整体的用户体验。

- 第 7 章对全文进行总结，以及未来可以进一步开展或改进的工作。

# 第 2 章　国内外研究现状和工作基础

随着智能手机及其他移动设备的普及，社交多媒体数据的规模也呈现出了爆发式的增长。对于社交多媒体数据的挖掘和利用已经成为了当前研究的热点问题，相关的成果既有利于推动计算机视觉以及多媒体领域相关课题的创新，对于用户体验的提升以及工业界的发展都也具有重要的应用价值和现实意义。

本文主要研究社交多媒体数据的包括包括语义理解和关联表达两个方面，涉及弱监督深度学习，特征提取，模型简化，照片集的关联表达和移动多摄像头视频自动剪辑几个关键问题。本章内容对这些关键问题的研究现状和工作基础做详细回顾。首先回顾弱监督学习，总结近年来弱监督分类和弱监督深度神经网络的发展情况；然后针对社交多媒体数据的特征提取，回顾传统的批处理方法，解决大规模流数据的在线学习方法，和在线特征提取算法，以及他们在解决大规模社交多媒体数据特征提取问题中的问题和不足；在模型简化方面，主要介绍与深度卷积神经网络相关的模型简化工作；在关联表达方面，从照片集和移动多摄像头视频两个角度分别介绍学术界和工业界的研究成果和代表性应用产品，指出他们在社交多媒体数据关联表达中的不足，引出本文提出的照片集关联表达系统和移动多摄像头自动剪辑系统。

## 2.1　弱监督学习

弱监督学习是指数据标注不完备或者包含噪音条件下的分类问题。本文主要针对弱监督图像分类问题做工作总结和算法创新。具体的方法可以分为两个方面：数据去噪（Data Cleaning) 和鲁棒噪音分类模型 (Noise Robust Models)。

### 2.1.1　数据去噪

数据去噪是指找出并移除可能错误标注的数据。数据去噪的优点在于它不依赖于目标任务的模型和训练方法。但是数据去噪也面临区别噪音数据和异常数据的难题[3]。数据去噪方法会导致两种类型的错误：正确标注的样本被误判为错误标注并被丢弃；错误标注的样本被漏判。

经典的数据去噪方法只基于图片本身的视觉特征。Brodley 等人提出的交叉过滤法[4] 采用类似于交叉验证的思路，将训练数据分成 $n$ 等份，并选择 $m$ 种分类算法（称为过滤算法）。对于每份数据，在剩下的 $n-1$ 份数据上训练 $m$ 个分

类模型。然后用得到的 $m$ 个分类器对这部分数据进行预测，最后通过一定的过滤算法判定错误标注的数据并将之移除。最大边界难分类样本学习通过迭代地学习正负类别分界面，保留难分类样本，移除易分类数据，从而更新正负样本集的方法来达到去除噪音数据的目的[5]。

交叉过滤法和最大边界难分类样本学习算法直接利用弱监督的数据标注，属于有监督学习方法。研究人员也提出了一些半监督的噪音数据去除方法。半监督学习方法首先需要人工标注部分数据作为种子数据，记为 $\mathcal{L}=\{(\mathbf{x}_i,y_i)\}_{i=1}^{m}$。剩下的数据记为 $\mathcal{U}=\{\mathbf{z}_j\}_{j=1}^{n}$。通常情况下，$m \ll n$。核平均算法的核心思想是通过对未标注的数据进行加权，使得加权后的数据分布与有标注数据的分布相同[6,7]。直推式支持向量机 (Transductive Support Vector Machine, TSVM) [7,8] 同时根据有标注的数据 $\mathcal{L}$ 和未标注的数据 $\mathcal{U}$ 寻找分界面，并约束最多有 $r$ 个未标注数据被判定为正类样本。

此外，由于社交多媒体数据具有丰富的上下文（context）信息，这些上下文信息也从侧面反映了图片的语义内容，因此可以用来辅助数据去噪。Schroff 等人提出了一种基于图片在网页中的文本上下文信息以及图片本身的视觉特征对图片进行重排序的算法[9]。算法首先利用图片附近的文本上下文学习标注是否正确的后验概率并进行重排序。然后从基于文本的重排序中选择前 $n_+$ 个图片作为正样本，再从所有其他类别的图片中选择 $n_-$ 个图片作为负样本。对图片提取视觉特征以后，训练一个 SVM 分类器。社交主动学习（Social Active Learning）首先用视觉特征训练分类器，并评价数据对于提升分类效果的信息量，同时利用文本特征评价每个数据标注可靠的置信度，综合考虑信息量和置信度选取数据，迭代地训练模型并选取正确标注的数据[10]。

可以看出，数据去噪方法分成两步进行：特征提取和数据去噪的模型训练，对于噪声的判别也依赖于人为假设和定义的准则，这些因素导致在真实的社交多媒体数据上，数据去噪方法的效果受多重因素影响，很难达到理想的效果。因此，我们需要提出一些端到端的学习方法 (End-to-End Learning)，并利用数据之间的关联达到弱监督学习的目的。

### 2.1.2　鲁棒噪音模型

鲁棒噪音模型是指本身对噪音数据不敏感或具有抑制作用的模型。在经验风险最小化（ERM）规则和给定的损失函数下，如果模型发生分类错误的概率保持不变并且与标注噪音无关，则认为模型是对标注噪音鲁棒的。研究人员讨论了在特定条件下，理论上是否存在对标注噪音完全鲁棒的模型[11]。例如，$0-1$ 损失函数在均匀的标注噪音或者能达到 0 错误率的情况下是噪音鲁棒的[12,13]。

Beigman 等人讨论了非随机噪音模型下的鲁棒模型[14], 最小平方差损失函数在均匀标注噪音下同样是噪音鲁棒的。其他常见的损失函数，如指数损失函数，对数损失函数，以及 Hinge Loss 均不是噪音鲁棒的。换句话说，大部分常见的机器学习算法都不是完全噪音鲁棒的，但可以在一定程度上减小错误标注的影响, 提升模型的鲁棒性。

对于经典的 SVM 分类器，Bunescu 等人提出了稀疏多实例学习算法用于解决训练样本中的噪音数据问题[15,16]。多实例学习将一组包含正样本包的有噪音数据称为一个正样本包，将一组负样本数据称为负样本包，同时假设正样本包中至少包含一个真实的正样本，负样本包中全是负样本。通过对正负样本包采用不同的约束条件和惩罚稀疏，达到抑制噪音影响的目的。

逻辑回归是一种经典的概率统计分类模型。相比于其他分类模型（如 SVM），逻辑回归的能够输出样本是否为某个类别的概率，因此得到了十分广泛的应用。然而，逻辑回归算法对于训练数据中的噪音标注十分敏感。Feng 等人提出了一种基于次高斯分布的鲁棒逻辑回归算法[17]。根据理论分析结果，算法首先去除范数大于一定阈值的数据，在剩余的数据中仅最大化前 $n_1$ 数据的标注和预测结果之间的相关度。在随机噪音 (Noise at Random, NAR) 假设下，研究人员提出用隐藏变量对数据的真实类别，错误标注的转移概率以及他们与观测到的标注之间的关系进行建模，并通过 EM 算法优化目标结果[18,19]。

近年来，深度卷积神经网络在图片识别上获得了巨大的成功。卷积神经网络最早于 1998 年被 Lecun 等人运用在了文字识别上[20]。随着显卡能力的提升，，得益于显卡设备的快速发展，卷积神经网络的深度、宽度以及结构不断改进，深度卷积神经网络的识别能力和应用范围都得到了巨大的提升[21–26]。

然而，深度卷积神经网络需要大量的训练数据，对于数据标注的质量有很大的依赖性。如何利用容易获得的大量弱监督数据训练深度卷积神经网络成为了近年来的研究热点之一。自举深度神经网路是 2014 年谷歌研究团队提出的利用图片特征之间的相似性监督网路学习，抑制错误标注影响的学习方法[27]。文章认为，如果图片的特征之间具有相似性，那么预测的结果同样也应该比较类似，并称之为感知连续性。在随机噪声的假设下，在神经网络中加入全连接隐藏层表示真实的类别，为了达到感知连续性的约束，论文提出引入类自适应编码器的方法训练网络。此外，还可以通过约束隐藏层输出的熵最小达到软自举优化的目的, 以及约束隐藏层概率最大类别的熵最小达到硬自举的目的。Sukhbaatar 等人则通过约束隐藏层参数的迹来达到感知连续性的优化目标[28]。Xiao 等人则综合考虑了标注可能遇到的随机噪音和非随机噪音，通过隐藏变量对不同噪音类型下的概率进行建模，用两个神经网络分别学习噪音类型和真实的标注，通过 EM 算法学习整个模型的参数[29]。以上方法都基于特定的噪音模型，Azadi 等人提出

一种辅助图片正则项 (Auxiliary Image Regularizer, AIR) 的方法[30]，根据数据的特征结构，通过组约束的方法使得只有部分数据具有响应，从而在训练数据中识别出有用的辅助数据，更好地训练神经网络。某种程度上，可以认为辅助图片正则项方法是在训练数据中寻找最近邻数据，减少深度模型对噪音数据的拟合。

传统的鲁棒噪音模型方法仍然基于特征提取和模型训练两个步骤，具有章节 2.1.1提到的局限性。现有的弱监督深度学习方法基于特定的噪音模型，难以处理真实的场景，或建模过于复杂，难以训练。AIR 方法利用图片之间的相似性抑制噪音，但组稀疏优化提高了网络训练的难度和时间开销。因此，我们提出一种新的利用数据在特征空间的关联性进行相关反馈的弱监督深度神经网络，并对网络进行简化和近似，提高模型的适用性和实用性。

## 2.2　特征选取

图像特征表述不仅包括高层次的卷积神经网络特征，还包含低层次全局特征（如颜色特征[31]，边缘特征[31]，纹理特征[32]），局部特征（如 SIFT[33]，SURF[34]）以及用来通过局部特征描述整体视觉信息的词袋方法等[35]。实际应用中需要根据需求选取对目标任务最有用的特征子集，这对于计算能力，内存，和电量都十分受限的移动设备尤其重要。此外，去除特定任务不相干的特征，还可以提高特征的表达能力。特征选取在机器学习和数据挖掘领域得到了广泛的研究，可以分成两类：批处理方法和在线特征选取。

### 2.2.1　批处理方法

批处理方法是指每次迭代都需要考虑所有的训练数据，可以分为三个类别：

- 过滤法: 过滤法分析特征之间的关联，距离，交互信息熵等，选取最有代表意义的特征子集[36–38]。Yang 等通过分析指出，传统的过滤法存在单调性的问题，不同大小的特征子集之间存在单调的包含关系，这种包含关系在实际情况中并不成立[39]。他们对特征之间的联系进行建模，提出了一种多核学习的方法。

- 包装法：包装法使用预先定义好的分类器去评价选取的特征子集的性能[40]。这类方法迭代的选取不同的子集，并得到该子集在对应分类器上评价指标，虽然能够获得该分类器上最好的特征子集，但是计算的开销也十分巨大，因此对于该类方法的研究相对较少。

- 嵌入法：嵌入法将特征选取与模型训练进行融合，是一种在高效的过滤法

和高准确率的包装法之间综合平衡的方案[41,42]。

### 2.2.2　在线特征选取

批处理方法的缺点在于需要将所有训练数据都加载到内存中，对于目前的大量高维数据，这类方法的局限性十分明显。同时，批处理方法要求数据预先已经全部存在，实际场景中，存在大量的流媒体数据。因而近年来，随着数据量的增大和维度的增加，大量的工作转向了在线学习。最早的在线学习算法是 1958 年提出的感知机算法。2006 年，Crammer 等人在感知机算法上增加约束，使得模型在每次迭代后在当前数据上都能获得正确的分类结果[43]。考虑到批处理学习算法中，二阶海森矩阵能够显著提高算法的收敛速度，Crammer 等人假设模型参数服从一个高斯分布，用协方差矩阵表示当前模型对于参数的不确定性，提出了置信度加权的在线学习算法 CW[44]。该算法每次迭代时约束更新后的模型以一定概率在当前的数据上获得正确的预测结果。由于该算法假设数据的标注都是正确的，在实际场景中的效果会收到影响。自适应的置信度加权在线学习算法则降低了对于噪音数据的敏感性[45]。

近几年，在线学习被应用到特征选取上。Langford 等人提出的稀疏在线学习算法，在感知机模型上增加了模型参数的 L1 范数作为正则项，获得稀疏的模型[46]。Duchi 等提出的 FOBOS 算法将稀疏在线学习分成两步，第一步是正常的在线学习，第二步优化目标使得模型尽可能接近第一步得到的参数，同时对应的 L1 范数最小[47]。另一种思路则是优化模型在主空间和对偶空间的距离，利用模型的 L1 范数达到稀疏模型的目的而提出的 RDA 算法[48] 。RDA 算法在高稀疏条件下往往能获得更好的特征表达能力。受置信度加权等二阶算法的启发，Duchi 等利用梯度的协方差构建二阶信息，提出了自适应的二阶 FOBOS 算法和 RDA 算法。

基于 L1 范数的稀疏在线学习算法是针对特征选取的一种“软”约束。参数设定与目标特征数目之间没有确定的关联。基于 L0 范数的在线特征选取也得到了研究人员的关注。Wu 等提出的在线特征流学习能够在每次迭代后返回一个模型和它选取的特征子集[49]。该算法假设每次获得所有数据的某个特征，不同特征按照时序被送到算法中。另一种更普遍的应用场景下，算法每次可以获取一个数据的部分或者所有特征，不同数据按照时序到达算法。Huang 等提出一种无监督的在线特征选取算法处理这类数据[50]。Wang 等则利用有监督的数据, 根据权重向量的绝对值进行在线特征选取[51]。

批处理方法的主要问题在于不具有可伸缩性 (Scalability)，以及对于流数据不具有很好的应对能力。基于 L1 范式的在线稀疏学习方法对于特征的数目不能

做直接的约束，在实际应用中需要根据不同的数据调整参数达到选取预期特征数目的目的。当前的在线特征选取算法根据权重向量绝对值选取特征的做法与批处理方法的效果还有较大差距，并且算法的复杂度较高。因此，本论文提出了大规模高维特征选取算法，不仅具有与批处理方法相近的准确率，也极大地减小了计算复杂度，对于处理大规模社交多媒体数据具有非常大的应用价值。

## 2.3　模型简化

近年来，深度卷积神经网络在图片识别，物体检测等领域获得了巨大的成功，为了进一步提高网络的表征能力，研究人员不断改进网络的深度[22]，宽度[52]以及拓扑结构[23,26,53]。然而，大量的网络参数也要求大量的时间开销和计算资源，也极大地限制了深度网络在计算能力、存储空间和电池续航受限的移动设备上的应用。如何在不影响网络性能的情况下减少网络参数成为了当前的研究热点问题。深度网络模型简化相关的工作可以分为三个类别：矩阵分解、量化以及稀疏优化。

矩阵分解利利用参数之间的相关性，对参数矩阵进行低秩分解，从而减少网络的参数个数。Denil 等人提出将参数矩阵分解成两个低秩矩阵的乘积，将其中一个矩阵作为特征空间的一组基，并提出了基向量字典的构建方法[54]。Denton 等人提出在网络的预测阶段，用奇异值分解 (Singular Vector Decomposition, SVD) 对参数矩阵进行分解，如果参数矩阵的奇异值迅速下降，则参数矩阵能够被前最大 $t$ 个奇异值及对应的奇异向量很好地近似[55]。Rigamonti 等人提出的方法在空间上对每个通道的卷积核用秩为 1 的矩阵进行近似，从而减小计算量[56]，Jaderberg 等人在空间分解的基础上进一步利用通道之间的冗余信息，将原始的卷积操作分解成两步卷积运算[57]。Ioannou 等人和 Tai 等人进一步改进并扩展低秩分解方法，将其用于更大的深度网络[58,59]。Mamalet 等人将卷积核分解为秩为 1 的向量乘积，并与后续的池化 (Pooling) 操作冗合为一层卷积运算，从而减少运算量[60]。

量化是指利用较少的比特数表示网路参数，从而减少模型的大小以及乘法运算的复杂度。目前常用的网络参数都采用 32 比特的浮点型数据，研究表明，可以利用更少的比特数来表示每个参数。例如，Hwang 等人和 Arora 等人提出仅用 $+1, -1, 0$ 三个数值表示网络参数训练卷积神经网络[61,62]。Courbariaux 等人和 Rastegari 等人则进一步提出用二个数值表示网路参数[63,64]。Gong 等人提出用向量量化的方法对全连接层的参数进行量化[65]。针对卷积层的向量量化则在 Wu 等人提出的 Q-CNN 得到研究和应用[66]。Anwar 等人用最小平方差的方法量化网络[67]。Chen 等人利用哈希函数随机将网络参数分组，达到量化的目的[68]。

当前网络的参数矩阵是密集矩阵，稀疏优化的目标是使得最终的参数矩阵稀疏，从而达到模型简化的目的。区别于参数矩阵低秩分解，Liu 等人提出对卷积核进行稀疏分解，并提出了高效的系数矩阵相乘算法[69]。受 L1 或 L2 约束的启发，Han 等人提出重复交替进行删除神经元之间连接和重新训练精简后网络[70,71]。然而，这些稀疏方法产生的稀疏性不是结构化的，运算时会导致无规则的内存访问，不能带来实际的运算加速。Li 等人根据卷积核的绝对值之和去除部分卷积核，从而达到运算的加速[72]。Murray 和 Chiang 运用结构化稀疏方法约束隐藏层神经元的个数[73]。Anwar 等提出了卷积核，通道，以及卷积核内部的结构化稀疏方法[74]。他们还提出用粒子滤波器 (Particle Filter) 衡量网络连接的重要性，从而优化网络结构。Wen 等人系统讨论了结构化的稀疏算法，从卷积核，通道，卷积核形状，深度四个方面对网络进行进行结构化的约束，不仅达到了减少网络参数的目的，还获得了实际运算速度上的提升。Hu 等人通过研究发现大网络的部分神经元的响应大部分情况下为 0，且与网络的输入信号无关。因此，他们通过分析网络神经元在大数据集上的响应去除部分神经元[75]。这些结构化的方法稀疏后的网络，有连接的神经元之间在通道上仍然是密集连接，Soravit 等人提出了对通道稀疏连接的方法，在保持其他结构化方法同样计算速度的情况下获得了很好的效果[76]。

以上介绍的模型简化方法，虽然取得了一定的效果，但同时也存在很多的问题。矩阵分解方法对于全连接层以及大卷积操作具有非常好的效果，然而最新的网络更倾向于使用更少的全连接层，并通过级联小卷积核的方法达到大卷积核同样的感知野，不仅减少了运算量，还提高了网络的表征能力[24]。量化方法需要特定的硬件或软件库的支持才能显著提高运算的速度。非结构化的稀疏优化方法对于减少参数数目作用比较明显，对于计算速度的提升十分有限。结构化的方法一般基于参数的绝对值决定参数的重要性，具有一阶在线学习方法同样的缺陷，而组稀疏优化的方法则增加了网络优化的难度。为此，有必要提出一种新的模型简化方法，既能保证简化后网络的表征能力，实际提升网络的运算效率，减少参数规模，又易于优化，提高模型简化的可操作性。

## 2.4　社交多媒体数据的关联表达

本节从照片集关联表达和视频关联表达两个方面回顾社交多媒体数据关联表达相关的工作。

### 2.4.1　照片集关联表达

照片集的关联表达涉及事件检测 (Event Detection)，关键照片选取，和照片故事表达。通常，用户照片包含拍摄时的时间戳和位置信息，我们可以利用这些信息将照片集分成不同的事件。Platt 等人提出用一个小时或者自适应的时间间隔作为相邻事件之间的时间间隔[77]。Graham 等人扩展了该方法，使用事件聚类的类内内拍照频率和类间时间间隔调整已有的事件划分[78]。Gargi 提出将拍摄频率的急速增加的时间点作为时间的起点，将长时间间隔没有拍摄作为事件的终点[79]。Matthew 等人将可信度，动态规划或者贝叶斯信息准则 (Bayes Information Criterion, BIC) 运用到照片的相似度矩阵，从而检测事件的边缘位置[80]。一般来说，事件检测问题可以表示为一个聚类问题。Loui 和 Svakis 提出用 2 类的 K-means 聚类算法将照片分组，并检查照片之间颜色的相似性改进聚类[81]。Gong 等人利用层次聚合聚类算法将照片分配到不同的聚类中心[82]。Platt 等人用隐马尔科夫模型聚类[77]。Mei 等人在时间，位置以及内容特征上利用混合高斯模型解决事件检测问题[83]。Xu 等人进一步利用纹理和深度特征改进了这个算法[84]。

近年来许多研究工作和产品相继出现，用以解决关键照片选取问题。在学术界，关键照片选取主要依赖照片的代表性[80,83–85]。Cooper 等人将事件中第一张照片作为关键照片。Mei 等人选择具有最大后验概率的照片[83]。Chu 等人提出在照片的聚类中，根据近似图片对之间的相互关系决定关键照片[85]。Xu 等人则根据事件的重要性引入了照片的受欢迎程度 (popularity) 以及时间内部的相似度决定关键图片[84]。

照片故事表达一直以来受到了工业界和学术界共同的关注。例如，Magisto[1]和 Animoto[2]是两个可以根据用户提供的照片产生音乐视频 (Music Video) 的在线服务。然而，它们依赖用户主动选取和提供的照片，不能直接从照片集中总结并整理出故事呈现给用户。此外，用户需要手动指定音乐视频编辑的风格。其他的在线服务，如 Microsoft Onedrive[3]，Google[4]可以在一定程度上对照片集进行事件检测和照片选取，并且缺乏对数据的重新表达。在学术界，Hua 等人提出的 Photo2Video 系统是从照片产生视频的先驱性工作[86]，利用相继运动将静态照片转换成运动片段，最终通过转场效果并与节奏匹配生成最终视频。然而该系统没有设计不同的编辑风格，所采用的编辑效果也比较单一, 其他系统比如 Tiling SlideShow 系统将照片和背景音乐同步，并以瓷砖式幻灯片的方法播放[87]。Kuo 等人提出的 Sewing Photos 系统专注于解决在播放照片幻灯片时，给照片之间分

[1] http://magisto.com
[2] http://animoto.com
[3] https://onedrive.live.com
[4] https://plus.google.com

配平滑的转场效果[88]。Sewing Photos 和 Tiling SlideShow 也存在 Photo2Video 同样的编辑效果单一的问题。Yang 等人提出了一种从照片中自动生成有吸引力的版面设计的算法。

以上系统没有系统对照片集进行总结整理，在表达时很少运用丰富的视频制作特效和视频编辑风格和编辑语法，因而对于照片故事的表现能力十分有限。因此，我们需要提出一个能够对照片集进行事件挖掘和关键图片选取，并运用专业的编辑语法对故事进行再现表达，提供更为有效的照片集关联表达方法。在表格 2.1中，我们比较了现有照片集关联表达系统和本文提出的 Monet 系统之间的差异。

**表 2.1**　照片集关联表达系统比较

| | Magisto | Animoto | Google+ | Monet (this paper) |
|---|---|---|---|---|
| 相机运动分析 | + | - | - | + |
| 视频分析 | - | - | - | + |
| 人脸检测识别 | + | - | + | + |
| 场景分析 | + | - | + | + |
| 物体识别 | - | - | + | + |
| 音乐分析 | + | - | - | + |
| 照片分组和选取 | - | - | + | + |
| 设计风格 | + | - | - | + |
| 色彩调整 | + | + | + | + |
| 社交和云存储 | - | - | + | + |

### 2.4.2　移动多摄像头视频关联表达

移动多摄像头视频是指在同一个事件中，有多个移动摄像头从多个角度拍摄的，时间上有重叠的一组视频[89]。随着智能设备的普及与性能的提升，移动多摄像头视频的自动剪辑成为了近年来的热点问题。

移动多摄像头视频的自动剪辑的方法可以总结为三个类别：基于规则的[90]，基于优化的[89]，和基于学习的[91,92]。基于规则的方法模仿专业视频人员的编辑过程。然而，移动多摄像头视频的自动剪辑过程更类似于用户的选择倾向，而并非固定的编辑规则。Shrestha 等人提出了从视频质量，多样性和切换点的合适度进行镜头选取的优化算法[89]。然而实际系统中，论文并没有提出切实可行的切换点合适度的评价方法，仅仅考虑了视频质量和多样性。在视频质量评估中，没有考虑到移动视频中的倾斜和遮挡问题。此外，作者通过贪心算法解优化方程，

仅获得了目标方程的局部最优解。Saini 等人提出 Jiku Director 用于解决在线视频剪辑问题[91,92]。他们通过学习隐马尔科夫模型 (HMM) 用于镜头选取和确定镜头长度。然而，通过这种方式学习到的规则与内容无关，而实际中，镜头角度和长度的选取都是和内容密切相关的，并且受运动强度，音乐的节奏，以及其他因素的影响[93]。此外，该系统由于不能准确判断视频的角度，因而不能做到全自动的视频剪辑，尤其是在模型的训练阶段，需要人工的干预。Arev 等人最近提出了从多摄像头视频的自动编辑系统[94]，但是该系统十分依赖场景的三维重建，不适用于移动多摄像头。

以上系统均没有考虑音频的剪辑，而完整的视频是有音频流和视频流两部分构成的，高质量的音频流对于提升用户体验具有十分重要的作用。在表格 2.2中，我们比较了现有移动多摄像头视频系统和本文提出的 MoVieUp 系统之间的差异。

**表 2.2**　视频自动剪辑系统比较

| | MoVieUp (this paper) | VD[89] | Jiku[92] |
|---|---|---|---|
| diversity | Yes | Yes | Yes |
| shakiness | Yes | Yes | Yes |
| tilt | Yes | No | Yes |
| occlusion | Yes | No | Yes |
| audio mashup | Yes | No | No |
| cut point | Audio+Video | Manual | Learning |

**注**：VD is short for Virtual Director[89]. Transition matrix for cut points learnt by Jiku Director is the same to all videos, thus not content-based.

移动多摄像头自动剪辑还与视频编辑相关，包括视频摘要 (Video Summarization)，相机选取，以及家庭或音乐视频编辑。视频摘要与视频剪辑的共同点在于它们都要最大化有信息内容的部分。Sundaram 等人提出了从可计算的镜头中生成快速概览的实用框架[95]。该论文将视觉编辑语法运用到镜头编辑中（选取，缩放，时长，顺序等）[96]，对于本文的工作具有很大的借鉴意义。

可计算镜头的检测通常基于人类记忆的一个仿真模型[95]。相机选取在演讲和会议等诸多特定场景都得到了广泛的研究，通常可以通过识别演讲者或者检测人脸来选取需要展示的相机内容[97,98]。Ranjan 等人和 Zhang 等人提出的系统中用跟踪和基于音频的定位来选择相机。以上系统都可以归结为基于音频的方法。在移动多摄像头视频自动剪辑中，音频不是唯一的关注点，音频定位和人脸检测在嘈杂的拍摄环境以及低视觉质量的条件下的应用能力十分有限。

此外，还有大量关于家庭视频或音乐视频编辑的工作。Hua 等人提出了自动

家庭视频编辑系统 AVE，从一系列的家庭视频中提取一部分最精彩的镜头[99]。他们提出了两套规则分别保证对原来视频的代表性，以及音频和视频之间的协调性。类似的方法被拓展到自动音乐视频编辑，该系统分析视频的时序结构和音频的节奏并进行匹配[93]。然而，由于移动多摄像头视频需要进行时间上的同步，并且需要保证内容上的质量和多样性，这些系统不能直接用于移动多摄像头视频自动编辑。

# 第 3 章　弱监督社交多媒体数据语义理解

弱监督社交多媒体数据语义理解是指在训练数据标注不准确的条件下的语义理解。弱监督社交多媒体数据语义理解是分析挖掘利用社交多媒体数据的基础，本文后续部分的特征提取和关联表达都是基于弱监督社交多媒体数据予以理解的结果。本章主要研究弱监督目标识别问题。

## 3.1　弱监督目标识别问题建模

在目标识别问题中，给定 $n$ 个训练数据 $X = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$，$\mathbf{x} \in \mathbb{R}^d$ 是训练数据的特征表示，$y \in \mathcal{Y}$ 是数据的类别，$|\mathcal{Y}| = K$ 是数据的类别空间，类别数目为 $K$。$y$ 的取值通常为离散的整数值。本文仅讨论单类别分类问题，即每个训练数据有且仅有一个类别。

传统的目标识别问题假设数据的标注 $y$ 是准确无误的，在社交多媒体数据中实际获得的标注 $y$ 和真实的数据类别 $z$ 存在不一致的情况。在统计学上，用一个二值的随机变量 $E$ 表示是否存在标注噪音。数据 $X$, 真实标注 $Z$, 实际观测到的标注 $Y$，和随机变量 $E$ 之间存在图 3.1 所示三种关系[100]。

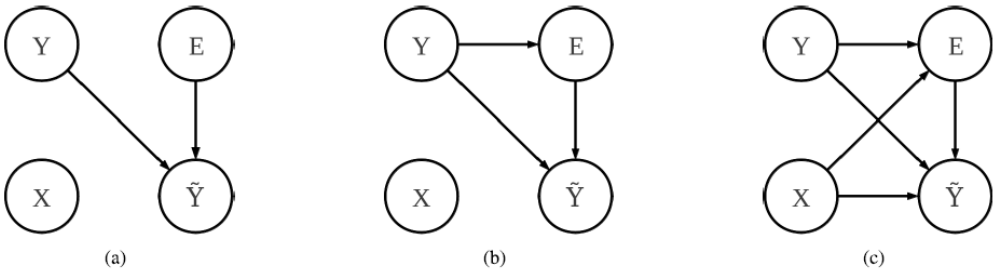



**图 3.1**　数据标注噪音类型: (a) 完全随机噪音; (b) 随机噪音; (c) 非随机噪音

- **完全随机噪音 (NCAR)**: 噪音 $E$ 独立于其他随机变量，包括真实的标注 $Z$。

- **随机噪音 (NAR)**: 噪音 $E$ 独立于数据 $X$, 但依赖真实的标注 $Z$。该模型允许非对称噪音, 即某些类别的数据更有可能出现标注噪音。随机噪音可等价

地表示为一个标注矩阵或转移矩阵：

$$\gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{K1} & \cdots & \gamma_{KK} \end{pmatrix} = \begin{pmatrix} \mathrm{P}(Y=1|Z=1) & \cdots & \mathrm{P}(Y=K|Z=1) \\ \vdots & \ddots & \vdots \\ \mathrm{P}(Y=1|Z=K) & \cdots & \mathrm{P}(Y=K|Z=K) \end{pmatrix} \tag{3.1}$$

- **非随机噪音 (NNAR)**: 上述两种噪音类型均假设标注噪音对于同一类别下的所有样本具有相同的影响。然而在实际场景中，上述假设不一定成立。例如，当样本与其他类别样本之间的距离较近时，更有可能发生错误标注。此外，样本分布密度较低的区域标注的可靠性也比其他区域低。在图 3.1(c) 所示的模型中，标注噪音 $E$ 同时依赖于数据 $X$ 和真实类别 $Y$，错误标注更有可能出现在某些类别和数据空间的某些区域。非随机噪音是最有普适性的噪音类型。例如，分界面附近和低样本密度分布区域的标注噪音只能用非随机噪音来建模。

## 3.2　相关反馈弱监督深度神经网络

传统弱监督目标识别算法需要人工设计一系列的特征，如全局特征 (HOG)，局部特征 (SIFT, SURF) 等。这些特征的表达能力直接影响到图片分类的效果，不仅提高了研究人员设计特征的难度，也制约了目标识别效果的提高。近年来，深度卷积神经网络在图片识别上获得了巨大的成功。2012 年，Alex 等人将深度卷积神经网络应用到了百万级规模的 ImageNet 图像识别任务上，提出了 AlexNet 网络模型，通过 5 层的卷积神经网络，直接从原始的图片像素中提取从浅层语义到深层语义的特征，然后用 3 层的全连接神经网络作为分类器[21]。深度学习的优点在于不需要人工设计图片特征，网络通过反向传播的方式同时学习特征提取和分类器。然而，经典的深度卷积神经网络对于数据标注的质量有很大的依赖性，如何利用大规模弱监督社交多媒体数据训练深度卷积神经网络成为了近年来的研究热点之一。

### 3.2.1　经典卷积神经网络

如图 3.2所示，经典深度卷积神经网络的前几层通常是卷积层，最后几层为全连接层，不同的任务可能会采用不同深度的网络。最后一个全连接层的输出作为柔性最大传递函数（Softmax）分类器的输入，得到在所有类别上的概率分布。假设训练数据为 $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$，$\mathbf{x}_i$ 表示第 i 个图片，$N$ 是图片总数目。用 $\mathbf{y} = [y_1, \ldots, y_N] \in \mathbb{R}^N$ 表示训练数据的标注向量，$y_i \in [0, K-1]$ 是第

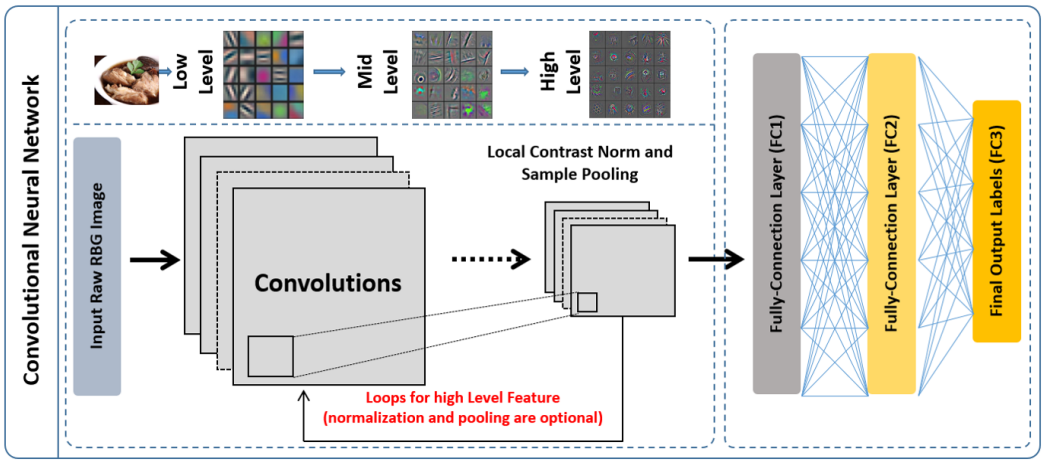



**图 3.2**　深度卷积神经网络的结构图

i 个图片的标注类别, $K$ 是类别数目。假设网络的层数为 $M$，网络参数表示为 $W = \{W^1, \ldots, W^M\}, W^m \in \mathbb{R}^{d_{m-1} \times d_m}$。数据在第 $m$ 层的特征图 (Feature map) 表示为 $Z^m(X) = [\mathbf{z}^m(\mathbf{x}_1), \ldots, \mathbf{z}^m(\mathbf{x}_N)]^T \in \mathbb{R}^{N \times d_m}$。

通常，卷积神经网络的损失函数是柔性最大传递函数的负对数似然函数和权重衰减项 (weight decay) 之和：

$$\mathcal{L}(W; X, \mathbf{y}) = -\frac{1}{N}\left[\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; W)\right] + \frac{\beta}{2}\|W\|_F, \tag{3.2}$$

其中，$\beta$ 是权重衰减项的系数。上述损失函数相对于最后一层特征图 $Z^M$ 以及参数 $W^M$ 的梯度为：

$$\begin{aligned}
Z^M &= Z^{M-1}W^M && (3.3)\\
\frac{\partial \mathcal{L}(W; X, Y)}{\partial \mathbf{z}^M(\mathbf{x}_i)} &= -\frac{1}{N}\Big(\mathbf{1}_{y_i}(\mathbf{z}^M(\mathbf{x}_i)) - \mathbf{p}(\mathbf{z}^M(\mathbf{x}_i))\Big) && (3.4)\\
\frac{\partial \mathcal{L}(W; X, Y)}{\partial W^M} &= (Z^{M-1})^T \frac{\partial \mathcal{L}(W; X, Y)}{\partial \mathbf{z}^M(\mathbf{x}_i)} && (3.5)\\
&= -\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}^{M-1}(x_i)\Big(\mathbf{1}_{y_i}(\mathbf{z}^M(\mathbf{x}_i)) - \mathbf{p}(\mathbf{z}^M(\mathbf{x}_i))\Big)^T && (3.6)
\end{aligned}$$

其他层参数的梯度可以通过方向传播 (Back Propagation) 得到[20]。从上述公式可以看出，错误的标注会导致参数梯度计算错误，并被反向传播，使得经典的卷积神经网络对于标注噪音十分敏感。

### 3.2.2　相关反馈卷积神经网络

目前已有部分工作研究弱监督深度学习，然而现有方法大多基于特定的噪音模型，难以处理真实的场景，或建模过于复杂，难以训练。2014 年谷歌研究

团队提出的利用图片特征之间的相似性监督网路学习，抑制错误标注影响的学习方法[27]。文章认为，如果图片的特征之间具有相似性，那么预测的结果同样也应该比较类似，并称之为感知连续性, 并将感知连续性应用到随机噪声假设下的网络学习中。本文在感知连续性的基础上提出了不依赖于特定噪声类型的相关反馈卷积神经网络。

基于感知连续性，本文方法的基本假设是正确标注样本在图像空间具有相似性，它们的特征表示在特征空间也具有相似性，而错误标注的样本则不具有这种相似性。因此，可以利用特征之间的相关性作为反馈，使得网络训练过程中不同数据在模型训练中发挥不同的作用。

为了表示特征之间的关系，我们将网络最后一层的特征转换为能反映特征之间相互关系的关联特征表示 (Affinity Representation)。类似于 Belkin 等人提出的最近邻系统[101]，我们定义如下相似度矩阵 $S \in R^{N\times N}$：

$$S_{ij} = \begin{cases} \exp\{-\frac{\|\mathbf{z}^M(\mathbf{x}_i)-\mathbf{z}^M(\mathbf{x}_j)\|^2}{\gamma^2}\} & y_i = y_j \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

$\gamma$ 是尺度因子。为了更好地反映相似度矩阵的局部结构，我们用一个对角矩阵 $D$ 对相似度矩阵进行正则化，$D_{ii} = \sum_{j=1}^{N} S_{ij}$。训练数据最终的特征表示为 $\Psi(X;W) = [\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_N)] = D^{-1}S$，矩阵 $\Psi(X;W)$ 的每一列包含了数据 $\mathbf{x}_i$ 的特征和其他数据特征之间的关系。

假设理想情况下, 不受噪音影响的模型参数为 $W^*$，则噪音鲁棒学习算法应该尽量优化 $W$ 使其逼近 $W^*$。该优化目标可以通过最小化学习到的特征表示 $\Psi(X;W)$ 和理想情况下的特征表示 $\Psi(X;W^*)$ 之间的差值 $E_n$ 得到。$E_n$ 是由于噪音标注引起的特征表示的误差。换句话说，可以认为 $\Psi(X;W)$ 是理想特征与一个加性噪声之和：

$$\Psi(X;W) = \Psi(X;W^*) + E_n \tag{3.8}$$

根据方程 (3.8)以及低秩理论[102]，我们假设 $\Psi(X;W^*)$ 是低秩矩阵：

$$\text{rank}(\Psi(X;W)) > \text{rank}(\Psi(X;W^*)) \tag{3.9}$$

在社交多媒体数据上，假设类别数目足够多，错误的标注来自于同一训练数据集上的其他类别，并假定训练数据的特征最多有 $K$ 个模式，$\Psi(X;W^*)$ 的秩等于类别数目 $K$，因此，$\Psi(X;W^*)$ 可以通过如下优化方程得到：

$$\begin{aligned} \min_{\Psi(X;W^*)} & \|\Psi(X;W) - \Psi(X;W^*)\|_F, \\ \text{s.t.} \quad & \text{rank}(\Psi(X;W^*)) = K \end{aligned} \tag{3.10}$$

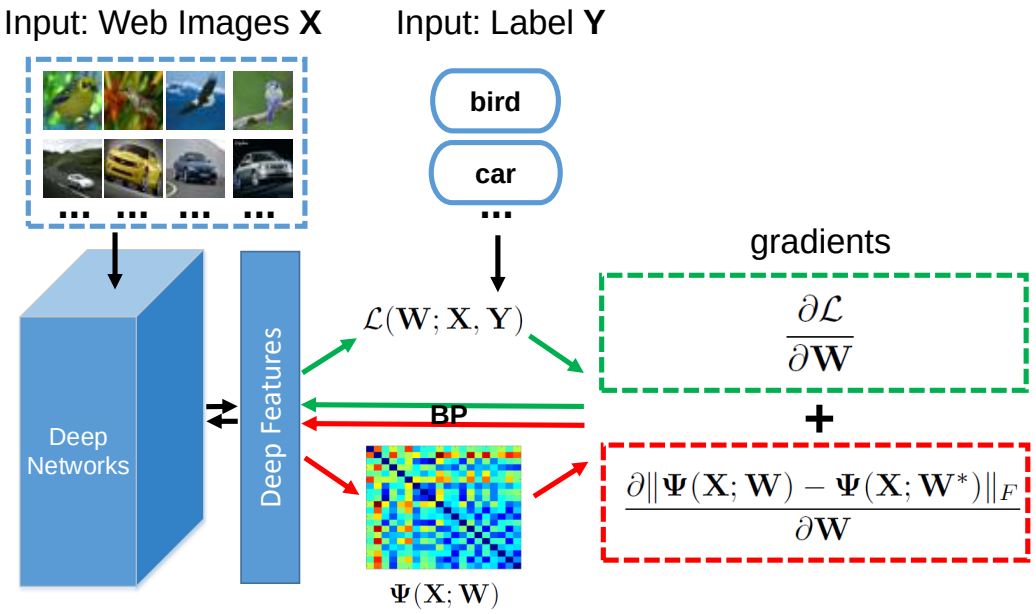



**图 3.3**　相关反馈卷积神经网络

如图 3.3所示，相关反馈卷积神经网络通过特征矩阵的额重构误差，利用训练数据之间的感知连续性，抑制训练过程中噪音的影响。然而，由于方程 (3.10)带来的计算量，上述方法极大增加了优化过程的时间开销。此外，该方法优化过程分两步：解优化方程 (3.10)和反向传播。为了避免该方法呢的缺点，本文进一步提出了改进的快速算法。快速算法基于下面的命题：

**命题 3.1**　令 $L = D - S, H^* \in R^{N\times K}$ 由 $\Psi(X;W)$ 的 K 个最大的特征值对应的特征向量构成，可以得到：1）方程 (3.10)，即 $\Psi(X;W)$ 的秩为 $K$ 的最佳近似由特征向量矩阵 $H^*$ 唯一决定；2）$H^*$ 也是下列优化方程的最优解：

$$\min_{H} \mathrm{tr}[H^T L H] \quad \text{s.t.} \quad H^T H = I. \tag{3.11}$$

由于方程 (3.10)和方程 (3.11)均在 $H^*$ 取得最优值，可以认为方程 (3.10)和方程 (3.11) 作为惩罚项是等价。

**证明**　命题 3.1可以通过如下三个定理得到。不失一般性，假设 $\mathrm{rank}(\Psi(X;W)) = r$。矩阵的秩为 $K$ 的最小重构误差矩阵可以通过 *Eckart-Young-Mirsky* 定理[103] 得到：

**定理 3.2** (Eckart-Young-Mirsky)　对秩为 $r$ 的矩阵 $P \in \mathbb{R}^{m\times n}$ 进行奇异值分解 (Singular Value Decomposition, SVD) 得到 $P = U\Sigma V^T, U^T U = I, V^T V = I$，如果 $K < r$，则有：

$$\underset{\substack{\hat{P}\in\mathbb{R}^{m\times n}\\ \mathrm{rank}(\hat{P})=K}}{\arg\min} \|P - \hat{P}\|_F = U\hat{\Sigma}V^T,$$

其中 $\hat{\Sigma}$ 是包含 $P$ 的前 $K$ 个最大奇异值的对角矩阵。

此外，如果矩阵 $P$ 是实对称矩阵，它的奇异值和特征值之间具有如下定理所示关系：

**定理 3.3**　对实对称矩阵 $P$ 特征值分解 (EVD) 得到 $P = Q\Lambda Q^T, Q^TQ = I, \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N).\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ 是矩阵 $P$ 的特征值。则有

$$Q = U$$

因此，根据定理 3.2和定理 3.3，实对称矩阵 $\Psi(X;W)$ 的秩为 $K$ 的最小重构误差矩阵由 $\Psi(X;W)$ 的前 $K$ 个最大特征值对应的特征向量对应的矩阵构成。

根据 Rayleigh[104] 可以得到如下定理：

**定理 3.4**　令 $H^* = [\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*] = \arg\min_H \text{tr}[H^TLH]$，并且 $H^TH = I$，则最优解 $H^*$ 可以通过求解如下泛化特征值分解问题得到：

$$L\mathbf{h}_i = (1 - \lambda_i)D\mathbf{h}_i,$$

其中 $\{1 - \lambda_i^* | i = 1, \ldots, K\}$ 是矩阵 $\Psi(X;W)$ 的前 K 个最大特征值，$H^* = \{\mathbf{h}_i^* | i = 1, \ldots, K\}$ 是对应的特征向量。

由于方程 (3.10)和方程 (3.11)均在 $H^*$ 取得最优值，可以认为方程 (3.10)和方程 (3.11) 作为惩罚项是等价。 □

通过以上命题和证明可以发现，方程 (3.10)的最优解可以通过方程 (3.11)中最小迹得到。因此，我们提出将最小迹的优化目标引入到经典的神经网络中，最终的噪音鲁棒深度网络目标方程为：

$$\widetilde{\mathcal{L}} = \mathcal{L}(W; X, Y) + \alpha\text{tr}[H^TLH]. \tag{3.12}$$

上述优化方程仍然需要对特征矩阵进行特征值分解，本文提出了进一步的近似方法。首先构建标注矩阵 $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \{0,1\}^{N\times K}$，每一列 $\mathbf{y}_i \in \{0,1\}^{K\times 1}$ 表示数据 $\mathbf{x}_i$ 的标注向量，并且只有 $y_i$ 位置为非零值。由于 $Y$ 矩阵是在语义空间上对数据的表述，$H$ 是在特征空间上的主成分表述，根据感知连续性，$H$ 可以通过如下优化方程近似得到[105,106]：

$$\min_H \|HH^T - YY^T\|_F^2. \tag{3.13}$$

为了满足 $H$ 矩阵的正交性，一个合理的近似解为 $H = Y(Y^TY)^{-\frac{1}{2}}$。

### 3.2.3　相关反馈分析

为了验证上述方法有效性，我们从梯度的角度分析相关反馈对于噪声的抑制能力。定义第 $M-1$ 层特征之间的距离为：

TBD.

图片 $\mathbf{x}_i$ 对于梯度的贡献如图 3.4所示。

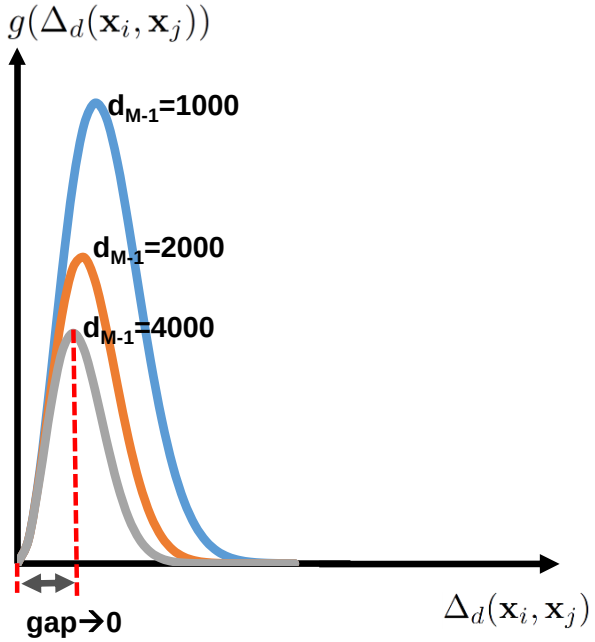



**图 3.4**　训练数据对于梯度的贡献曲线，横坐标为数据与其他数据的距离

## 3.3　实验结果和评估

本节从实验的角度验证本文提出的弱监督相关反馈卷积神经网络的有效性。首先在标准数据集上验证该方法对于噪声标注的鲁棒性，其次，我们将该方法用于真实的社交数据集，验证该方法在图片标注上的有效性。

### 3.3.1　目标识别

**实验数据**：我们在两个公开数据集上分别验证算法对噪声的鲁棒性。一个是 CIFAR-10[107]，包含 10 个类别 60,000 张 $32\times32$ 的彩色图片，其中 50,000 用于训练，10,000 用于测试。为了产生不同噪声比例的训练数据，在每个类别的训练数据上按照不同比例，随机选取图片，并将他们的类别随机替换为数据集中的其他类别，训练数据集的总图片数目保持不变。在我们的实验设置中，训练数据从无噪声到 90% 的噪声均匀取 10 个噪声比例。另一个数据集是 PASCAL

VOC2007[108]，包含 20 个类别总共 9,963 张图片。我们将数据集随机等分成训练数据和测试数据。

**比较基准**：本文提出的相关反馈卷积神经网络称为 RFCNN，并以下四个方法进行比较：

- **CNN**: 经典的卷积神经网络。
- **RPCA+CNN**: 在训练卷积神经网络之前，首先用 RPCA[102] 方法对每个训练数据进行重构，并移除重构误差较大的训练数据，移除的比例和噪音的比例相同。
- **CAE+CNN**: 首先用卷积自动编码器对卷积神经网络的每一层进行预训练，然后微调整个网络，从而减小噪音标注的影响[109]。
- **NL+CNN**: 用全连接层表示噪音转移概率矩阵，和卷积神经网络一起训练[28]。

对于 VOC2007 数据集，我们还与另外两种方法进行比较：

- **Best_VOC**: 用 ImageNet 数据集预训练网络，并在 VOC2007 上微调[110]。
- **Web_HOG**: 通过基于局部的模型和人工设计的特征，在网络图片上训练语义概念表征[111]。

**参数设置**：首先，我们调整公式 (3.2) 中权重衰减项的系数 $\beta$。对于 10% 的噪音比例，该稀疏取 0.004 时网路能达到最好的效果，对于 20% 的噪音比例，取值为 0.008，其他噪音比例下取值为 0.04。该参数设置对于两个数据集都能取得最好的效果。此外，我们按照经验将公式 (3.7)中的 $\gamma$ 参数设为 0.1，使得特征相似度在合理的范围。图 3.5显示了在 CIFAR-10 数据集 20% 噪音下，公式 (3.12)中不同 $\alpha$ 取值对于网络准确率的影响。我们发现，只有当 $\alpha$ 取值过大时 (比如取 10)，模型的完全丧失了分类能力，对于其他取值，准确率都保持在相对稳定的范围，并在取值为 0.5 时达到最优。此外，我们发现 $\alpha$ 取 0.5 在其他噪音条件下也能取得最好的效果。因此，以下实验 $\alpha$ 均取 0.5。

**实验结果**：表格 3.1显示了在 CIFAR-10 数据集上不同噪音程度下不同算法的分类准确率比较。本文提出的算法在所有条件下都达到了最好的实验结果，甚至在无噪声数据集上，我们的算法也比经典的卷积神经网络取得了略好的准确率。我们发现，在 30% 噪音数据下，经典卷积神经网络的准确率下降了将近 20%。相比之下，本文提出的算法值下降了 10%，表现出了对噪音数据很强的鲁棒性。此外，我们发现数据预处理方法 RPCA+CNN 在噪音比例小于 50% 时的

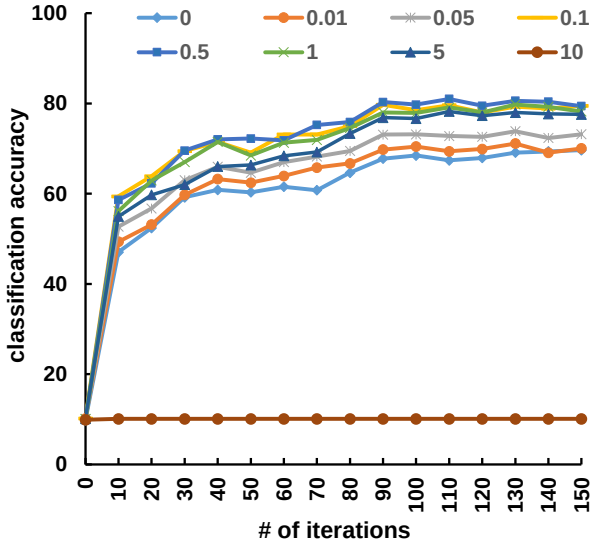



**图 3.5**　参数 $\alpha$ 对于卷积神经网络分类性能的影响

准确率要好于经典的卷积神经网络，当有更多的噪音数据时，RPCA+CNN 的效果则比经典 CNN 要差。这个现象的原因在于当噪音数据增多时，数据预处理移除正确数据的风险也随之增大，导致在最终的训练数据中噪音数据的比例增加。CAE+CNN 和 NL+CNN 算法的性能十分接近，在 30% 噪音比例下，准确率分别下降 17.0% 和 15.9%。CAE+CNN 虽然能够解决区域级的噪声（背景噪声），但对于样本集噪声 (如本文中的标注错误)，它的鲁棒性则比较有限。对于 NL+CNN，我们的实验证明仅仅在网路上增加一层噪音适应层并不能噪音鲁棒性。相反，我们的方法可以抑制噪音在所有层的影响。图 3.6反映了在不同噪音程度下不同方法分类准确率相对于无噪声下降程度的比较。

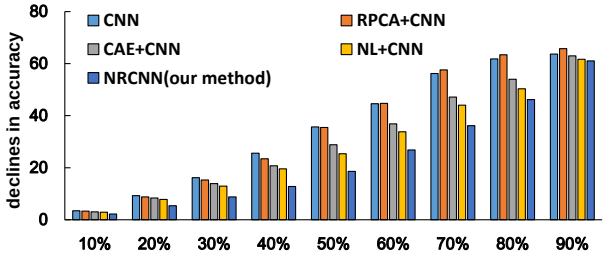



**图 3.6**　不同噪音程度下不同方法的准确率相对于无噪声的下降程度比较

在 PASCAL VOC2007 数据集上，我们使用 AlexNet 网络[21] 首先在 ImageNet 数据集上预训练，然后在网络数据上微调网络参数。为了获取网络数据，我们将数据集的每个类别作为查询词抓取搜索引擎的返回结果，并滤除重复的图片。我们收集了两个训练数据集，第一个数据集中正负样本的数目和 VOC2007 中相同，在该数据集上的实验我们记为 CNN(Web) 和 RFCNN(Web)。第二个数据集中我们将正样本的数目增加到 VOC2007 的 4 倍，并将该数据集上的方法记

**表 3.1**　不同算法在 CIFAR-10 数据集上不同噪音条件下的准确率比较

| 噪声比例 / 算法 | 无噪声 | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|
| CNN | 81.24 | 77.79 | 71.97 | 65.09 | 55.65 | 45.60 | 36.65 | 25.02 | 19.46 | 17.55 |
| RPCA+CNN | 81.24 | 77.94 | 72.44 | 65.94 | 57.82 | 45.77 | 36.55 | 23.68 | 17.85 | 15.49 |
| CAE+CNN | 81.55 | 78.54 | 73.19 | 67.69 | 60.83 | 52.71 | 44.71 | 34.39 | 27.54 | 18.61 |
| NL+CNN | 81.16 | 78.28 | 73.36 | 68.26 | 61.63 | 55.83 | 47.33 | 37.12 | 30.81 | 19.49 |
| RFCNN | **81.60** | **79.39** | **76.21** | **72.81** | **68.79** | **63.01** | **54.78** | **45.48** | **35.43** | **20.56** |

为 CNN(Webx4) 和 RFCNN(Webx4)。根据我们统计的结果，两个数据集上的噪音比例分别是 20% 和 40%。不同方法在 VOC2007 测试数据集上的平均准确率 (Average Precision) 如表 3.2所示。可以发现：

- CNN(Web) 相比 Web_HOG 具有十分明显的提升，证明了深度学习网络比利用人工设计的特征训练的分类模型具有更强的噪声鲁棒性。
- RFCNN(Webx4)　取得了比在大量标注好的数据集上训练的网络更好的效果。

**表 3.2**　不同算法在 VOC2007 数据集上不同噪音条件下的不同类别平均准确率比较

| 类别 / 算法 | plane | bike | bird | boat | btl | bus | car | cat | chr | cow | tab | dog | horse | moto | pers | plnt | shp | sfa | train | tv | **mAP** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Best_VOC | 88.5 | **81.5** | **87.9** | **82.0** | 47.5 | 75.5 | **90.1** | 87.2 | 61.6 | 75.7 | **67.3** | 85.5 | 83.5 | 80.0 | **95.6** | 60.8 | 76.8 | **58.0** | **90.4** | **77.9** | 77.7 |
| Web_HOG | 68.5 | 48.2 | 47.3 | 55.7 | 40.0 | 56.3 | 60.1 | 64.1 | 43.6 | 59.2 | 32.9 | 46.5 | 56.2 | 62.4 | 41.3 | 29.6 | 41.4 | 35.6 | 68.9 | 35.5 | 49.6 |
| CNN(Web) | 84.1 | 68.8 | 77.1 | 73.0 | 63.0 | 74.2 | 74.3 | 79.2 | 61.8 | 73.8 | 48.9 | 79.5 | 81.0 | 82.1 | 48.4 | 57.9 | 72.0 | 31.6 | 83.4 | 64.7 | 68.9 |
| CNN(Webx4) | 85.4 | 69.4 | 77.1 | 74.5 | 63.7 | 74.7 | 75.0 | 81.6 | 62.3 | 75.7 | 53.3 | 80.2 | 83.8 | 84.6 | 50.7 | 58.9 | 75.9 | 41.0 | 84.5 | 69.1 | 71.1 |
| RFCNN(Web) | 85.8 | 69.7 | 77.4 | 75.1 | 63.8 | 75.8 | 75.6 | 82.7 | 62.7 | 76.9 | 53.5 | 80.6 | 84.7 | 84.9 | 49.2 | 59.1 | 76.0 | 50.8 | 84.8 | 69.2 | 71.9 |
| RFCNN(Webx4) | **91.3** | 75.2 | 83.3 | 81.5 | **70.2** | **81.3** | 80.6 | **88.3** | **67.0** | **82.5** | 60.0 | **86.3** | **90.0** | **90.3** | 75.8 | **64.8** | **81.0** | 57.8 | 89.9 | 74.9 | **78.6** |

## 3.4　小结

# 第 4 章　大规模社交多媒体数据快速处理

社交多媒体数据由于规模庞大，需要耗费大量的计算资源和时间进行处理。本章主要从特征选取和模型简化的角度讨论大规模社交多媒体数据的快速处理。

本章首先介绍用于大规模高维数据的特征选取算法。对于社交多媒体数据的特征描述不仅包括高层次的卷积神经网络特征, 还包含低层次全局特征 (如颜色特征, 纹理特征), 局部特征 (如 SIFT,SURF) 以及用来通过局部特征描述整体视觉信息的词袋特征等。实际应用中需要根据需求选取对目标任务最有用的特征子集, 这对于大规模社交多媒体数据的处理速度以及移动设备有限的计算能力和内存空间尤其重要。此外, 去除特定任务不相干的特征, 还可以提高特征的表达能力。

其次，本章介绍用于深度卷积神经网络的模型简化算法。深度卷积神经网络在很多计算机视觉领域都表现出很好的性能，然而网络的深度和模型参数通常比较大, 例如经典的 VGG-16 网络包含超过 200M 的模型参数。大量的模型参数意味着在实际应用中需要大量的计算资源和时间, 极大地限制了深度神经网络在大规模图片检索和图片识别等任务上的应用。此外, 深度网络在移动设备上的应用已经成为一种趋势。由于移动设备计算能力的限制, 在不影响模型准确度的条件下简化深度网络模型已经成为迫切的需要。

## 4.1　二阶在线特征选取

特征选取是指从数据中移除不相关或者冗余特征的过程。在当前大数据的背景下，特征选取已经成为了十分重要的一项技术，并在多个领域尤其是高维数据的场景下获得了广泛应用[112,113]。尽管特征选取已经被广泛的研究，大部分的算法都属于批处理学习。批处理学习的主要问题在于它需要将整个数据集加载到内存中，由于计算机的内存容量无法跟上数据规模，这对于实际问题中大规模高维数据显然不具有可伸缩性。此外，批处理方法假设所有的训练数据和特征在训练前已经给定，而实际场景中需要处理的往往是流数据，并可能伴随有新的特征出现。为了克服已有算法的这些问题，近年来有部分工作研究在线特征选取[39,49,51]。在线学习的优点在于算法每次迭代只处理一个数据，从而达到很高的可伸缩性，并且可以很好地应对数据中模式和特征的变化。然而，目前已有的在线特征选取算法复杂度仍然过高，模型的准确率与批处理方法也有不小差距。

因此，文本提出了二阶在线特征选取算法，不仅对于大规模高维数据具有很高的可伸缩性和学习效率，算法的准确率也与批处理方法十分接近。

不失一般性，我们首先研究二分类问题，并在 4.4章节将算法扩展到多类问题。令 $\{(\mathbf{x}^t, y^t)|t = 1, \ldots, T\}$ 表示训练过程中依次收到的数据，每个数据 $\mathbf{x}^t \in \mathbb{R}^d$ 是一个 $d$ 维的向量，数据类别 $y^t \in \{+1, -1\}$。在线学习算法将学习一个相同维度的分类器 $\mathbf{w} \in \mathbb{R}^d$。在时刻 $t$，算法接收到数据 $\mathbf{x}^t$，并基于当前的模型参数 $\mathbf{w}^t$ 预测它的类别 $\hat{y} \in \{+1, -1\}$。

$$\hat{y}^t = \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t) \tag{4.1}$$

预测完以后，算法将得到真实的类别 $y^t$，从而衡量在数据 $(\mathbf{x}^t, y^t)$ 上的损失函数 $\ell(\mathbf{w}^t)$。损失函数通常为真实类别和预测结果之间的差值的函数。在每轮迭代的最后，算法根据特定的规则更新模型参数 $\mathbf{w}^t$。例如，在线梯度下降算法更新的规则为：

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta^t y^t \mathbf{x}^t \tag{4.2}$$

$\eta^t$ 是时刻 $t$ 的学习率 (learning rate)。根据不同的更新规则，在线学习算法可以分为两类：

- 一阶算法：本质上是梯度下降算法[43]；
- 二阶算法：挖掘输入数据的几何特点[45] 或构建目标方程的近似海森矩阵[114]；

在线特征选取需要选取权重向量 $\mathbf{w}^t$ 中相对较小的一部分元素，并将其他元素设为 0。换句话说，我们对 $\mathbf{w}$ 的 $L_0$ 范式施加如下约束：

$$\|\mathbf{w}\|_0 \leq B, \|\mathbf{w}\|_0 = \sum_{i=1}^{d} w_i^0 \tag{4.3}$$

$B$ 是预先定义的常数，因此，最多只有 $\mathbf{x}$ 的 $B$ 个特征被用来做预测。

## 4.2　置信度加权二阶在线特征选取

在线特征选取最直接的一种做法是运用截断感知机算法 (Perceptron with Truncation, PET)[51]。具体来说，分类器在每次迭代首先根据 $\mathbf{w}^t$ 预测类别 $\hat{y}^t$。如果 $\hat{y}^t$ 是正确的，则 $\mathbf{w}^{t+1} = \mathbf{w}^t$；否则，分类器根据感知机规则更新 $\mathbf{w}^t$：$\hat{\mathbf{w}}^{t+1} = \mathbf{w}^t + \eta y^t \mathbf{x}^t$。更新后的参数进一步被保留绝对值最大的 $B$ 个元素，其他元素的值设为 0. 截断后的分类器参数 $\mathbf{w}^{t+1}$ 将被用于下一次迭代数据的预测。算法 4.1显示了 PET 算法的框架，算法 4.2显示了在线特征选取的截断函数。

**input :** $B$ - 需要选取的特征个数，$\eta$ - 学习率
**output:** 权重向量 $\mathbf{w}^T$

初始化 $\mathbf{w}^1 = \mathbf{0}$;
**for** $i \leftarrow 1$ **to** $T$ **do**
    接收到数据 $\mathbf{x}^t \in \mathbb{R}^d$，预测类别 $\hat{y}^t = \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t$;
    接收真实类别 $y^t$;
    计算损失函数 $\ell(\mathbf{w}^t)$;
    **if** $\ell(\mathbf{w}^t) > 0$ **then**
        $\hat{\mathbf{w}}^{t+1} = \mathbf{w}^t + \eta y^t \mathbf{x}^t$;
        $\mathbf{w}^{t+1} = \text{Truncate}(\hat{\mathbf{w}}^{t+1}, B)$;
    **end**
**end**

**算法 4.1:** PET：截断感知机算法框架

**input :** $\hat{\mathbf{w}}$ - 权重向量，$B$ - 需要选取的特征个数
**output:** 截断的权重向量 $\mathbf{w}$

$\mathbf{w} = \hat{\mathbf{w}}$;
**if** $\|\hat{\mathbf{w}}\|_0 > B$ **then**
    除了 $\mathbf{w}$ 的绝对值最大的 $B$ 个元素，其他元素全部设为 0;
**end**

**算法 4.2:** Truncate: 截断函数

根据 Wang 等人的分析，上述算法在实际应用中并不总能取得很好的效果。它不能保证被截断的参数足够小，因而不能保证很小的错误率。因此，他们在截断之前运用稀疏投影，提出了一阶在线特征提取算法 (FOFS)。FOFS 算法保证了每轮迭代分类器参数 $\mathbf{w}^t$ 都限制在一个 $L_1$ 范式约束的超体内部。算法显示了 FOFS 算法的细节。

**input :** $B$ - 需要选取的特征个数，$\eta$ - 学习率，$\lambda$ - 正则化参数

**output:** 截断的权重向量 $\mathbf{w}$

1 $\tilde{\mathbf{w}}^{t+1} = (1-\lambda\eta)\mathbf{w}^t + \eta y^t \mathbf{x}^t$;

2 $\hat{\mathbf{w}}^{t+1} = \min\{1, \frac{\frac{1}{\sqrt{\lambda}}}{\|\tilde{\mathbf{w}}^{t+1}\|_2}\}\tilde{\mathbf{w}}^{t+1}$;

3 $\mathbf{w}^{t+1} = \text{Truncate}(\hat{\mathbf{w}}^{t+1}, B)$;

**算法 4.3:** FOFS：一阶在线特征选取算法

一般来说，一阶在线特征选取算法的复杂度和特征维度成正比。对于超高维度数据，算法的速度会比较慢。此外，当输入数据的不同维度特征不在同一个尺度时，一阶算法可能会移除有价值的特征。如公式 (4.1)所示，预测结果不仅依赖于权重向量，同时也依赖于输入数据。即使 $|w_i| < |w_j|$，并不能保证 $w_i * E(x_i) < w_j * E(x_j)$，$E(x_i)$ 是 $x_i$ 的期望。为了克服一阶算法的局限性，我们探索了二阶在线学习最新发展，提出了二阶在线特征提取算法 (SOFS)。

二阶置信度加权算法[115] 假设线性分类器的权重向量服从高斯分布 $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$。权重的置信度通过协方差矩阵 $\Sigma$ 的对角元素表示，对角元素 $\Sigma_{jj}$ 越小，权重 $\mathbf{w}_j$ 的均值的置信度越高。在观察到数据之前，所有权重有共同的置信度或不确定性。在训练过程中，给定一个观察到的训练数据 $(\mathbf{x}^t, y)$，置信度加权算法更新权重使得在当前数据 $\mathbf{x}^t$ 上做出正确预测的概率大于一个阈值 $\tau$。同时，算法尽量保持与更新前的权重分布相同。置信度加权算法可以表示为下面的优化问题：

$$\begin{aligned}(\hat{\boldsymbol{\mu}}^{t+1}, \Sigma^{t+1}) &= \arg\min_{\boldsymbol{\mu},\Sigma} \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) \\ s.t. \quad & \Pr_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}[y^t(\mathbf{w}\cdot\mathbf{x}^t) \geq 0] \geq \tau, \end{aligned} \tag{4.4}$$

$\mathrm{D}_{\mathrm{KL}}(*,*)$ 是 Kullback-Leibler(KL) 距离。两个高斯分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 和 $\mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)$ 的 KL 距离定义为：

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) &= \frac{1}{2}\log\frac{\det\Sigma^t}{\det\Sigma} + \frac{1}{2}\mathrm{Tr}((\Sigma^t)^{-1}\Sigma) \\ &+ \frac{1}{2}(\boldsymbol{\mu}^t - \boldsymbol{\mu})^T(\Sigma^t)^{-1}(\boldsymbol{\mu}^t - \boldsymbol{\mu}) - \frac{d}{2}. \end{aligned} \tag{4.5}$$

公式(4.4)中的约束可以重新表达为：$y^t(\boldsymbol{\mu}\cdot\mathbf{x}^t)\geq\phi\sqrt{(\mathbf{x}^t)^T\Sigma\mathbf{x}^t}$，$\phi=\Phi^{-1}(\tau)$（$\Phi$)是高斯分布的累积函数)。研究人员提出了多种方法解公式 (4.4)中的优化问题。本文采用能够对每个训练数据的预测函数进行自适应正则化的 AROW 算法[45]。研究和实验表明，该算法对于训练数据中的噪音具有更好的鲁棒性。AROW 算法的目标方程为：

$$(\hat{\boldsymbol{\mu}}^{t+1},\Sigma^{t+1})=\arg\min_{\boldsymbol{\mu},\Sigma}\left\{\mathrm{D_{KL}}(\mathcal{N}(\boldsymbol{\mu},\Sigma),\mathcal{N}(\boldsymbol{\mu}^t,\Sigma^t))+\frac{1}{2\gamma}\ell^t(\boldsymbol{\mu})+\frac{1}{2\gamma}(\mathbf{x}^t)^T\Sigma\mathbf{x}^t\right\},\tag{4.6}$$

$\gamma>0$ 是正则化参数。$\ell^t(\boldsymbol{\mu})$ 是平方铰链损失函数：

$$\ell^t(\boldsymbol{\mu})=\max(0,1-y^t(\boldsymbol{\mu}\cdot\mathbf{x}^t))^2\tag{4.7}$$

方程 (4.6)存在如下闭合解：

$$\begin{aligned}\beta^t&=\tfrac{1}{(\mathbf{x}^t)^T\Sigma^t\mathbf{x}^t+\gamma}\quad\mathbf{g}^t=-2\max(0,1-y^t(\boldsymbol{\mu}^t\cdot\mathbf{x}^t)y^t\mathbf{x}^t\\\hat{\boldsymbol{\mu}}^{t+1}&=\boldsymbol{\mu}^t-\tfrac{1}{2}\beta^t\Sigma^t\mathbf{g}^t\quad(\Sigma^{t+1})^{-1}=(\Sigma^t)^{-1}+\tfrac{\mathrm{diag}(\mathbf{x}^t(\mathbf{x}^t)^T)}{\gamma}\end{aligned}\tag{4.8}$$

需要注意的是，本文提出的 SOFS 算法仅仅计算和考虑协方差矩阵 $\Sigma$ 的对角元素。从效率的角度，维持完整的协方差矩阵需要 $O(d^2)$ 的内存空间和 $O(d^2)$ 的计算复杂度，这对于大规模超高维数据是不切实际的。从学习能力的角度，相关研究工作也表明在数据量足够的情况下，对角协方差矩阵可以获得比完整协方差矩阵更好的性能，原因在于在学习初始阶段完整协方差矩阵算法适应数据之间相互依赖的能力，当数据不可分时同业也使得它在逼近最佳权重向量时过度拟合噪音[116]。

不同于一阶在线特征选取算法基于权重向量的绝对值大小决定特征的重要性，本文提出的二阶在线特征选取算法 (SOFS) 的核心思想是利用二阶信息保留 B 个置信度最高的特征。具体来说，在线学习过程中，当在训练数据 $(\mathbf{x}^t,y^t)$ 的损失函数不为 0 时，我们仅更新前 B 个最小协方差 $\Sigma_{jj}$ 对应的 B 个最确信的权重，剩余权重全部设为 0。算法显示了本文提出的 SOFS 算法框架。

## 4.3　二阶在线特征选取快速算法

目前已有在线特征选取算法的一个普遍问题在于高计算复杂度。具体来说，在线特征选取的一个主要时间开销在于从 $d$ 维数组（FOFS 算法中的权重绝对值向量和 SOFS 算法中的最小 B 个元素）中选取最大会最小的 $B$ 个元素。本文提出一个基于最小堆的 FOFS 和 PET 算法的高效可伸缩方法，用以替代在迭代的每一步对整个数组排序[51]。此外，基于类似的最大堆的实现，我们利用 SOFS 算法的单调递减性进一步降低了 SOFS 算法的复杂度。

**input :** $B$ - 需要选取的特征个数，$\gamma$ - 正则化参数
**output:** 权重向量 $\boldsymbol{\mu}^T$ 和对角协方差矩阵 $\Sigma^T$

初始化 $\boldsymbol{\mu}^1 = \mathbf{1}, \Sigma^1 = I$;
**for** $i \leftarrow 1$ **to** $T$ **do**
    接收到数据 $\mathbf{x}^t \in \mathbb{R}^d$，并预测 $\hat{y}^t = \text{sign}(\boldsymbol{\mu}^t \cdot \mathbf{x}^t)$;
    接收到数据的真实类别 $y^t$;
    计算损失函数 $\ell(\boldsymbol{\mu}^t) = \max(0, 1 - y^t(\boldsymbol{\mu}^t \cdot \mathbf{x}^t))^2$;
    **if** $\ell(\boldsymbol{\mu}^t) > 0$ **then**
        根据公式 (4.8)计算 $\beta^t, \mathbf{g}^t$;
        **for** $j \leftarrow 1$ **to** $d$ **do**
            $\hat{\mu}_j^{t+1} = \mu_j^t - \frac{1}{2}\beta^t \Sigma_{jj}^t g_j^t, (\Sigma_{jj}^{t+1})^{-1} = (\Sigma_{jj}^t)^{-1} + \frac{(x_j^t)^2}{\gamma}$;
        **end**
        **for** $j \leftarrow 1$ **to** $d$ **do**
            **if** $\Sigma_{jj}^{t+1}$ 是最小的 $B$ 个元素之一 **then**
                $\mu_j^{t+1} = \hat{\mu}_j^{t+1}$;
            **else**
                $\mu_j^{t+1} = 0$;
            **end**
        **end**
    **end**
**end**

**算法 4.4:** 二阶在线特征选取的算法框架

### 4.3.1　一阶快速特征选取算法

为了从 $d$ 维数组中找出最大的 $B$ 个元素（算法 4.2中的 Truncate 函数），直接的做法是对 $d$ 个元素排序，然后选取前 $B$ 个元素。为了提高计算效率，我们构建了一个最小堆用于存储权重向量 $\mathbf{w}^t$ 的 $B$ 个最大绝对值。学习过程中，当分类器的权重向量发生改变时，通过如下两步更新找出最大的 $B$ 个元素：

- 调整已经存在于堆中的元素的位置，维护最小堆结构。
- 比较不在堆中的每个元素与堆顶元素的大小。如果小于堆顶元素，则将它的值设为 0，否则将堆顶元素替换为当前元素，并调整堆顶元素与子节点的位置，维护最小堆结构，原堆顶元素的值设为 0。

算法 4.5显示了改进的 FOFS 算法的详细步骤。快速 PET 算法的过程与之类似。

为了说明上述算法的正确性，我们需要证明每次迭代以后绝对值最大的 $B$ 个特征仍然在最小堆中。用 $h_1, \ldots, h_d$ 表示堆中特征的位置下标，其他不在堆中特征的下标为 $h_{B+1}, \ldots, h_d$。在第一步中，$w_{h_1}, \ldots, w_{h_B}$ 被重新组织以满足最小堆的条件，我们有下面两个命题：

**命题 4.1**　如果模型更新后 $w_{h_i}, \forall i \in [1, B]$ 仍然在最大的 B 个元素中, 则 $w_{h_i}$ 不会被替换出最小堆；

**命题 4.2**　如果模型更新后 $w_{h_i}, \forall i \in (B, d]$ 在最大的 B 个元素中，则 $w_{h_i}$ 一定会被替换进最小堆。

**证明**　对于命题 4.1，如果 $w_{h_i}$ 不是 $B$ 个最大元素中最小的，则 $w_{h_i}$ 始终不会成为堆顶元素，因而一定不会被替换出最小堆。如果 $w_{h_i}$ 是 $B$ 个最大元素中最小的，则意味着最小堆中元素已经构成了最大的 $B$ 个特征，剩下的 $d - B$ 个特征权重的绝对值均比 $w_{h_i}$ 小。因此仍然不会在第二步过程中被替换出最小堆。对于命题 4.2，我们可以得到 $w_{h_i}$ 是最大的 $B$ 个元素之一时，堆顶元素一定小于 $w_{h_i}$，因此一定会被替换进最小堆。综上所述，本文提出的最小堆结构和更新方法可以找出权重绝对值最大的 $B$ 个特征。　□

### 4.3.2　二阶快速特征选取算法

尽管一阶快速特征选取算法已经避免了排序所有元素，其算法复杂度依然和特征的维度成正比。对于本文提出的二阶特征提取算法，可以进一步利用二阶特征的特殊性将算法复杂度降低为和非零特征的个数成正比，这对于大规模超高维度的稀疏数据具有重大的意义。区别于一阶在线特征选取算法，本文提出的二阶算法具有如下单调递减特性：

**input :** $B$ - 需要选取的特征个数，$\eta$ - 学习率，$\lambda$ - 正则化参数
**output:** 权重向量 $\boldsymbol{\mu}^T$

初始化 $\mathbf{w}^1 = \mathbf{1}, \mathbf{v}^1 = (|w_1^1|, \ldots, |w_d^1|) = \mathbf{0}$，$\mathbf{v}^1$ 上大小为 $B$ 的最小堆 $H$;
**for** $i \leftarrow 1$ **to** $T$ **do**
    接收到数据 $\mathbf{x}^t \in \mathbb{R}^d$，并预测 $\hat{y}^t = \mathrm{sign}(\mathbf{w}^t \cdot \mathbf{x}^t)$;
    接收到数据的真实类别 $y^t$;
    计算损失函数 $\ell(\mathbf{w}^t)$;
    **if** $\ell(\mathbf{w}^t) > 0$ **then**
        $\tilde{\mathbf{w}}^{t+1} = (1 - \lambda\eta)\mathbf{w}^t + \eta y^t \mathbf{x}^t$;
        $\mathbf{w}^{t+1} = \min\{1, \frac{\frac{1}{\sqrt{\lambda}}}{\|\tilde{\mathbf{w}}^{t+1}\|_2}\}\tilde{\mathbf{w}}^{t+1}$;
        $\mathbf{v}^{t+1} = (|w_1^{t+1}|, \ldots, |w_d^{t+1}|)$;
        调整 $H$ 中节点的位置，维护最小堆结构;
        **for** $j \leftarrow 1$ **to** $d, v_j^{t+1} \notin H$ **do**
            **if** $v_j^{t+1} > H_{min}$ **then**
                获取堆顶节点 $H_{min}$，堆顶对应的特征位置记为 $s$;
                $w_s^{t+1} = 0$;
                将堆顶 $H_{min}$ 替换为 $v_j^{t+1}$;
                调整堆顶元素与子节点的位置，维护最小堆结构;
            **else**
                $w_j^{t+1} = 0$;
            **end**
        **end**
    **end**
**end**

**算法 4.5:** 快速一阶在线特征选取算法

**命题 4.3** (单调递减性)　对于公式 (4.8)中的对角协方差矩阵 $\Sigma^t$，对于 $\forall t$ 以及 $\forall j \in [1, d]$，存在 $\Sigma_{jj}^{t+1} \leq \Sigma_{jj}^t$。

命题的正确性可以由 $\mathrm{diag}(\mathbf{x}^t(\mathbf{x}^t)^T)/\gamma$ 始终非负得到。基于上述命题，本文提出二阶在线特征选取的的快速算法。算法维护一个最大堆结构存储当前协方差矩阵的最小 $B$ 个元素。由于协方差矩阵每个元素的单调递减性质，对于每个被更新权重的特征，算法的更新规则为：

- 如果特征已经在最大堆中，算法仅需要比较当前特征与子节点的大小，从而维护最大堆。因为置信度单调递减，更新后一定小于父节点；
- 如果被更新的特征不在最大堆中，则比较其与堆顶的大小，如果小于堆顶，则替换堆顶，并将原堆顶对应的特征权重置为 0，否则将当前特征的权重置为 0。对于没有被更新权重的特征，不需要进行比较，因为堆顶具有单调递减的特性，没有被更新权重的特征的置信度一定大于堆顶。

算法 4.6显示了快速 SOFS 算法的细节。

### 4.3.3　复杂度分析

上述算法显著提高了在线特征选取的效率。本节分析上述算法的计算复杂度。

记权重向量的维度为 $d$，每个数据平均非零特征个数为 $m$，在最差情况下，PET 算法每步迭代需要的计算量为;

- $2m$：计算损失函数，更新权重向量；
- $m$：计算权重向量的绝对值；
- $B \log B$：维护最小堆；
- $(d - B) \log B$：找出最大的 $B$ 个元素，维护最小堆；
- $d - B$：将相应的特征值置为 0.

PET 算法每一步迭代的计算复杂度为 $\{3m + d - B + d \log B\}$。

FOFS 算法与 PET 算法类似，每步迭代需要：

- $2m$：计算损失函数，更新模型；
- $d$：计算权重向量的范数；

**input :** $B$ - 需要选取的特征个数，$\gamma$ - 正则化参数
**output:** 权重向量 $\boldsymbol{\mu}^T$ 和对角协方差矩阵 $\Sigma^T$

初始化 $\boldsymbol{\mu}^1 = \mathbf{1}, \Sigma^1 = I$，$B$ 个 $\Sigma^1$ 元素的最大堆 $H$;
**for** $i \leftarrow 1$ **to** $T$ **do**
    接收到数据 $\mathbf{x}^t \in \mathbb{R}^d$，并预测 $\hat{y}^t = \text{sign}(\boldsymbol{\mu}^t \cdot \mathbf{x}^t)$;
    接收到数据的真实类别 $y^t$;
    计算损失函数 $\ell(\boldsymbol{\mu}^t) = \max(0, 1 - y^t(\boldsymbol{\mu}^t \cdot \mathbf{x}^t))^2$;
    **if** $\ell(\boldsymbol{\mu}^t) > 0$ **then**
        根据公式 (4.8)计算 $\beta^t, \mathbf{g}^t$;
        **for** $j \leftarrow 1$ **to** $d$ **do**
            $\hat{\mu}_j^{t+1} = \mu_j^t - \frac{1}{2}\beta^t \Sigma_{jj}^t g_j^t, (\Sigma_{jj}^{t+1})^{-1} = (\Sigma_{jj}^t)^{-1} + \frac{(x_j^t)^2}{\gamma}$;
            **if** $\Sigma_{jj}^{t+1} \in H$ **then**
                递归调整 $\Sigma_{jj}^{t+1}$ 与它子节点的位置，维护最大堆结构;
            **end**
        **end**
        **for** $j \leftarrow 1$ **to** $d, x_j^t \neq 0, \Sigma_{jj}^{t+1} \notin H$ **do**
            **if** $\Sigma_{jj}^{t+1} < H_{max}$ **then**
                获取堆顶节点 $H_{max}$，堆顶对应的特征位置记为 $s$;
                $\mu_s^{t+1} = 0$;
                将堆顶 $H_{max}$ 替换为 $\Sigma_{jj}^{t+1}$;
                调整堆顶元素与子节点的位置，维护最小堆结构;
            **else**
                $\mu_j^{t+1} = 0$;
            **end**
        **end**
    **end**
**end**

**算法 4.6:** SOFS：快速二阶在线特征选取算法

- $d$：稀疏投影；
- $d$：计算权重向量的绝对向量；
- $B\log B$：维护最小堆；
- $(d-B)\log B$：找出最大的 $B$ 个元素，维护最小堆；
- $d-B$：将相应的特征值向量置为 0。

FOFS 算法单步迭代的整体计算复杂度为 $\{2m+4d-B+d\log B\}$，远高于 PET 算法。

SOFS 算法迭代的复杂度为：

- $3m$：计算损失函数，更新模型和对角协方差矩阵；
- $m\log B$：维护最大堆（只有 $m$ 个值发生改变）；
- $m$：将相应的特征值向量置为 0。

SOFS 算法迭代的复杂度降为 $\{4m+m\log B\}$，当 $m \ll d$ 并且 $B \ll d$ 时，SOFS 算法处理大规模超高维度稀疏数据时具有很高的效率和可伸缩性。在最差情况下 $m \approx d$，SOFS 算法复杂度与 PET 算法接近，但仍然小于 FOFS。

对于空间复杂度，我们只考虑分类器需要的空间占用，不考虑数据加载和存储的内存开销。在我们的实现中，输入数据存储成键值对的稀疏形式，处于效率考虑，模型参数表示成密集向量。PET 和 FOFS 算法要求保存权重向量 $\mathbf{w}$ 和它的绝对值向量，因而空间复杂度为 $O(2d)$。SOFS 算法也需要 $O(2d)$ 的空间复杂度，用来保存权重向量和对角协方差矩阵。因此，SOFS 算法的空间复杂度和一阶在线特征选取算法的空间复杂度相同。

## 4.4　二阶多类在线特征选取

多类问题中，假设共有 $k$ 个类别，每个训练数据的类别为 $y \in \{0,1,\ldots,k-1\}$。我们采用一对多的策略 (one-vs-the-rest) 将二阶在线特征选取算法扩展到多类问题。根据 Crammer 等人的策略[44]，置信度加权模型的分布类似于二分类问题，$\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu},\Sigma), \boldsymbol{\mu} \in \mathbb{R}^{kd}, \Sigma \in \mathbb{R}^{kd\times kd}$。我们引入新的类别相关的特征：

$$\psi(\mathbf{x},i) = [\mathbf{0}^T,\ldots,\mathbf{x}^T,\ldots,\mathbf{0}^T]^T,$$

只有 $\psi(\mathbf{x},i)$ 的第 i 个位置为 $\mathbf{x}$，其他位置为 $\mathbf{0}(\mathbf{0},\mathbf{x} \in \mathbb{R}^d)$。在每次迭代中，分类器接收到训练数据 $\mathbf{x}^t$ 并预测类别 $\hat{y}^t = \arg\max_{i=0}^{k-1} \boldsymbol{\mu}^t \cdot \psi(\mathbf{x},i)$。平方铰链损失函

数为：

$$\ell(\boldsymbol{\mu}^t) = \max(0, 1 - \boldsymbol{\mu}^t \cdot \Delta\psi^t)^2, \tag{4.9}$$

其中 $\Delta\psi^t$ 依赖于多类问题更新的策略。对于最大分数多分类更新：

$$\Delta\psi^t = \psi(\mathbf{x}^t, y^t) - \psi(\mathbf{x}^t, \arg\max_{i=0, i\neq y^t}^{k-1} \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}, i)) \tag{4.10}$$

对于均匀多分类更新：

$$\begin{aligned} \Delta\psi^t &= \sum_{i=0}^{k-1} \alpha_i^t \psi(\mathbf{x}^t, i), \quad \alpha_i^t = \begin{cases} -1/|E^t| & i \in E^t \\ 1 & \text{if} \quad i = y^t \\ 0 & \text{otherwise} \end{cases}, \qquad (4.11) \\ E^t &= \{i \neq y^t : \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}^t, i) \geq \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}^t, y^t)\}, \qquad (4.12) \end{aligned}$$

多分类更新目标方程为：

$$(\hat{\boldsymbol{\mu}}^{t+1}, \Sigma^{t+1}) = \arg\min_{\boldsymbol{\mu}, \Sigma} \left\{ \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) + \frac{1}{2\gamma}\ell(\boldsymbol{\mu}) + \frac{1}{2\gamma}(\Delta\psi^t)^T \Sigma \Delta\psi^t \right\}. \tag{4.13}$$

目标方程的闭合解与公式 (4.8)类似，区别在于将 $y^t\mathbf{x}^t$ 替换成 $\Delta\psi^t$。

我们仍然选取 $B$ 个最确信特征。一对多策略的多类问题中，特征的置信度依赖于 $k$ 个二分类器。第 $j$ 个特征的置信度定义为 $C_j = k - \sum_{i=0}^{k-1} \Sigma_{ij,ij}$。算法仅更新前 $B$ 个最大 $C_j$ 对应的权重参数，其他权重设为 0。算法细节与算法 4.4类似，区别在于将 $y^t\mathbf{x}^t$ 替换成 $\Delta\psi^t$。多类 SOFS 算法的时间复杂度是二分类问题的 $k$ 倍。

多类问题中 $\sum_{i=0}^{k-1} \Sigma_{ij,ij}$ 仍然具有单调递减性：

**命题 4.4** (单调递减性)　对于公式 (4.13)中得到的 $\Sigma^t$，对于 $\forall t$ 和 $\forall j \in [1, d]$，存在 $\sum_{i=0}^{k-1} \Sigma_{ij,ij}^t \leq \sum_{i=0}^{k-1} \Sigma_{ij,ij}^{t+1}$。

因此，快速二分类二阶在线特征选取算法也适用于多类二阶在线特征选取。

## 4.5　二阶在线特征选取实验评估

本节在不同规模的人工数据和真实数据上用实验证明本文提出的二阶在线特征选取算法的有效性。

### 4.5.1　实验设置

对于在线特征选取算法，如果没有显式说明，我们仅在训练数据上学习一轮。我们比较本文提出的算法和目前最好的在线和批处理特征选取算法，包括：

- PET：截断感知机算法，在线特征选取的基准算法[51]；
- FOFS：目前最好的一阶在线稀疏投影特征选取算法[51]；
- mRMR: 最小冗余最大相关特征选取[117]，最好的批处理方法之一以及它的图形处理器并行版本 (GPU-mRMR)[118]；
- Liblinear：用于大规模线性分类的开源库[119]，我们采用了其中的 $l1-SVM$ 算法作为 *Embedded* 特征选取的代表算法。
- FGM：目前最好的批处理 *Embedded* 特征选取方法之一[120]。

对于在线学习方法，我们使用铰链损失作为损失函数。我们使用五重交叉验证找出最优的超参数。对于每一个数据集，在线学习方法随机打乱顺序 10 次并取平均训练结果作为最终结果。对于 Liblinear 中的 $l1-SVM$ 算法，我们调节 $C$ 参数获得不同的特征个数。对于 FGM，为了简单起见我们遵循 Tan 等人论文中的设定将 $C$ 设为 $10$[120]。对于 mRMR，我们首先用它选取特定数目的特征，然后用在线梯度下降算法训练分类器。我们充分利用了在线学习依次处理单个数据的特点，在实现中使用两个并行线程分别处理数据加载和模型训练。

### 4.5.2　人工数据集实验评估

我们仿照 FGM 算法的评估方法，人为产生了三种类型的人工合成数据，分别是 $\mathcal{X}_1 \in R^{100K\times 10K}, \mathcal{X}_2 \in R^{100K\times 20K}, \mathcal{X}_3 \in R^{1M\times 1B}$，用来测试算法的性能，效率以及可伸缩性。三个数据集都用于二分类任务。每个数据从独立同分布的高斯分布 $\mathcal{N}(0,1)$ 中采样得到。为了模拟真实的数据，每个采样得到的数据都是稀疏数据，有效特征维度分别为 100,200，和 500。对于每一个数据，我们随机选取 $\mathcal{X}_1$ 的 200 维，$\mathcal{X}_2$ 的 400 维，和 $\mathcal{X}_3$ 的 500 维作为噪声。为了获得数据的类别，我们从均匀高斯分布 $\mathcal{N}(0,1)$ 中采样得到权重向量 $\mathbf{w}^*$ 作为正确的基准 (groundtruth)，每个数据类别为 $y=\text{sign}(\mathbf{w}^*\cdot\mathbf{x}^*)$，$\mathbf{x}^*$ 是没有噪音特征的数据。合成数据集的详细情况如表 4.1所示。

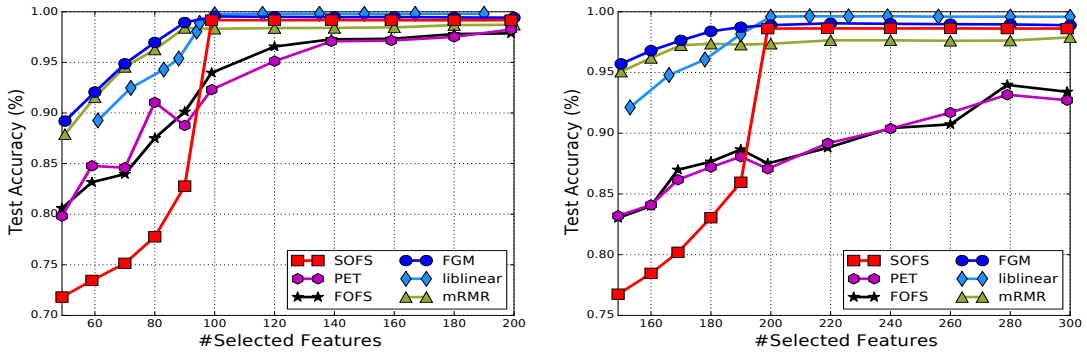
我们首先在 $\mathcal{X}_1$ 和 $\mathcal{X}_2$ 数据集上评估所有的特征选取算法。$\mathcal{X}_3$ 数据集用来测试本文算法的效率和可伸缩性。图 4.1和图 4.2显示了 $\mathcal{X}_1$ 和 $\mathcal{X}_2$ 上的准确率和时间开销。

**准确率**。根据图 4.1中的结果，我们可以总结如下几点发现。首先，当足够多有意义的特征被选取时（$\mathcal{X}_1$ 中的 100 维，$\mathcal{X}_2$ 中的 200 维），SOFS 算法可以达到接近批处理特征选取算法的准确率，而且 liblinear 和 FGM 相对于 SOFS 算法的优势十分有限。其次，当选取的特征数目较少时，批处理算法比在线学习算

**表 4.1**　合成数据信息（“K”,“M”,“B” 分别代表千，百万，十亿）

| DataSet | #Train | #Test | Dim | IDim[a] | NDim[b] | #Feature |
|---|---|---|---|---|---|---|
| $X_1$ | 100K | 10K | 10K | 100 | 200 | 30M |
| $X_2$ | 100K | 10K | 20K | 200 | 400 | 60M |
| $X_3$ | 1M | 100K | 1B | 500 | 500 | 1B |

[a]有效特征维度
[b]噪音特征维度

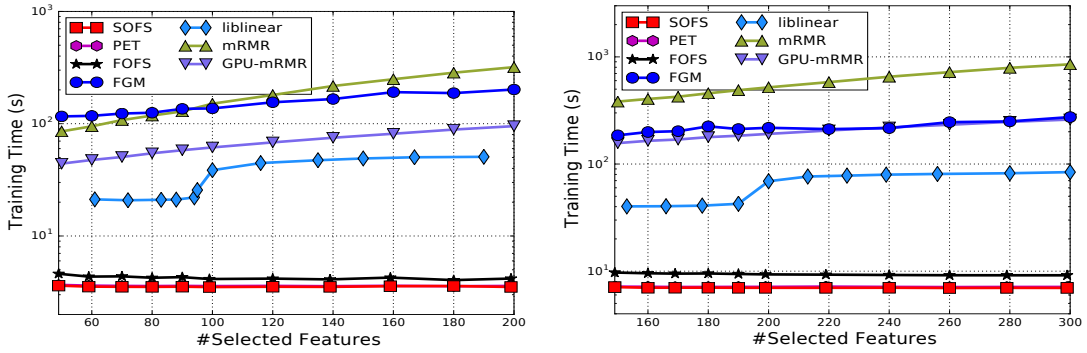



(a) $\mathcal{X}_1$ 数据集　　(b) $\mathcal{X}_2$ 数据集

**图 4.1**　合成数据集 $X_1$ 和 $X_2$ 上测试准确率和特征数目之间的关系

法更好。我们发现这种情况下尤其是 FGM 和 mRMR 比在线学习算法要好很多。SOFS 算法在特征不足是准确率不是很高，然而，随着更多的特征被选取，它的准确率迅速饱和并达到最佳。再次，两个一阶在线特征选取算法表现最差，尤其是在 $\mathcal{X}_2$ 数据集上。仅在特征数目很少时 PET 和 FOFS 算法比 SOFS 算法准确率高。然而，在特征数目足够多时，他们的性能不能达到与批处理算法相当的水平。总结起来，本文提出的算法能有发掘出有意义的特征，并能在特征数目足够多的情况下达到类似于批处理算法的准确率。

**时间开销**。除了测试准确率，训练效率也是实际问题必须考虑的关键问题。图 4.2 显示了各个算法的训练时间开销。一般来说，批处理算法虽然效果较好，但是时间开销远高于在线学习算法。本文提出的 SOFS 算法只需要几秒钟就可以达到批处理相当的准确率。相反，liblinear 需要大约 10 倍的训练时间，FGM 和 mRMR 在 $\mathcal{X}_2$ 数据集上甚至需要 100 倍的训练时间。并行 mRMR 算法相对于非并行算法减少了大约一半的时间。在线特征选取算法中，我们的方法仍然只需要最少的时间。我们发现在这两个数据集上在线特征选取算法的时间开销差别不大，我们将在更大规模和更高维度上评估他们的区别。尽管如此，准确率和时间




(a) dataset $\mathcal{X}_1$　　(b) dataset $\mathcal{X}_2$

**图 4.2**　Time cost versus number of selected features on synthetic datasets $\mathcal{X}_1$ and $\mathcal{X}_2$

开销的比较证明了 SOFS 算法是一个快速有效的在线特征选取算法。

我们继而在大规模超高维度的 $\mathcal{X}_3$ 数据集上测试 SOFS 的可伸缩性，由于已有特征选取算法在 $\mathcal{X}_3$ 上可能耗费几个小时甚至几天才能完成特征选取，我们仅在 $\mathcal{X}_1$，$\mathcal{X}_2$ 和 $\mathcal{X}_3$ 上测试 SOFS 算法是否能够处理增长的维度和规模，特征选取的数目分别固定为 $B = 100, 200, 500$。此外，我们与两个全特征维度上的在线学习基准算法比较，从而验证 SOFS 算法的有效性。两个基准算法分别是在线梯度下降算法 (OGD) 和自适应权重向量正则化算法 (AROW)，结果如表 4.2所示。根

**表 4.2**　SOFS 算法可伸缩性评测

| | Algorithm | $\mathcal{X}_1$ | $\mathcal{X}_2$ | $\mathcal{X}_3$ |
|---|---|---|---|---|
| Time Cost | OGD(s) | 3.58 | 7.06 | 114.82 |
| | AROW(s) | 3.59 | 7.02 | 130.72 |
| | SOFS(s) | **3.52** | **7.00** | **132.94** |
| Accuracy | OGD(%) | 98.44 | 97.83 | 99.39 |
| | AROW(%) | 98.48 | 98.52 | 99.55 |
| | SOFS(%) | **99.17** | **98.62** | **99.56** |
| Model Sparsity | OGD(%) | 0.00 | 0.00 | 83.16 |
| | AROW(%) | 0.00 | 0.00 | 72.22 |
| | SOFS(%) | **99.00** | **99.00** | **99.99** |

据表中的结果可以发现，测试准确率相对于基准算法有所提高，从而验证了移除不相关特征可以提高模型效果。更重要的是，SOFS 只需要少于 1% 的特征就可以达到这个准确率。快速有效的特征选取有如下三个好处：1) 当输入特征是密

集数据时，稀疏的分类器可以显著减少预测时间；2）可以显著减少预测时的内存开销；3) 可以显著减少特征提取的时间。在该数据集上，OGD 和 AROW 算法需要大约 1GB 内存存储分类器（每个权重需要 4 个字节），而 SOFS 算法仅需要 2 KB。在嵌入式系统等内存空间十分有限的条件下，紧凑的分类器更加具有实际意义和经济价值。

此外，我们可以发现随着数据数目和特征维度的增加，SOFS 的训练时间的增加在可接受的范围。在十亿个特征的数据集上，它仅需要 2 分钟多的时间就可以完成模型训练和特征选取。相反，其他特征选取算法陷入维度灾难的问题。例如，PET 算法需要至少 10 个小时从 $\mathcal{X}_3$ 中选取 500 个特征，更不用说其他风复杂的算法。此外，我们特别注意到相比于基准在线学习算法的时间开销，SOFS 并没有引入过多额外的时间开销。原因在于在我们的实现中数据加载和模型训练分为两个线程同时进行，由于三个算法都比较高效，数据加载实际上占据了主要的时间。总结起来，实验中的低训练时间和高准确率表明本文提出的算法能够快速有效地挖掘大规模超高维度数据中的有效特征。的

### 4.5.3　大规模真实数据集实验评估

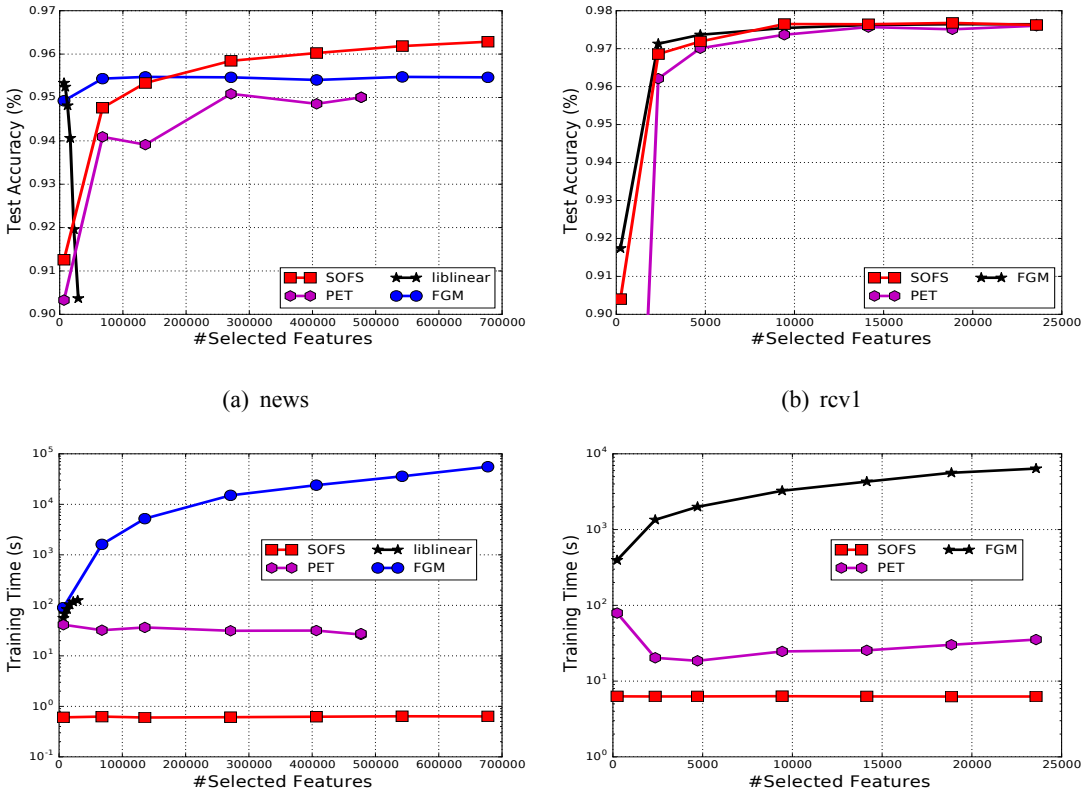
**表 4.3**　大规模真实数据集信息

| 数据集 | 特征维度 | 训练数据个数 | 测试数据个数 | 特征个数 |
|---|---|---|---|---|
| news | 1,355,191 | 10,000 | 9,996 | 5,513,533 |
| rcv1 | 47,152 | 781,265 | 23,149 | 59,155,144 |
| url | 3,231,961 | 2,000,000 | 396,130 | 231,249,028 |

本节在大规模真实数据集上评测 SOFS 算法的性能，采用的数据集如表 4.3所示。第一个数据集 “news” 维度较高，第二个数据集 “rcv1” 规模较大，第三个数据集 “url” 规模和维度均较大。在本实验中为了简单起见，我们仅比较 SOFS 算法，PET 算法（快速），FGM 算法（高效）之间的差异。

**表 4.4**　大规模高维数据集评测结果（$\rho$ 是选取的特征比例）

| Dataset | $\rho$ | 0.005 | 0.05 | 0.1 | 0.2 |
|---|---|---|---|---|---|
| news | PET | 90.33%(41.34s) | 94.09%(32.18s) | 93.91%(36.54s) | 95.08%(31.37s) |
| | SOFS | 91.26%(**0.61s**) | 94.76%(**0.63s**) | 95.33%(**0.60s**) | **95.84**%(**0.61s**) |
| | FGM | **94.92**%(90.10s) | **95.43**%(1610.53s) | **95.47**%(5206.20s) | 95.46%(15055.28s) |
| rcv1 | PET | 73.18%(79.13s) | 96.21%(20.30s) | 97.01%(18.53s) | 97.37%(24.63s) |
| | SOFS | 90.40%(**6.29s**) | 96.86%(**6.27s**) | 97.19%(**6.28s**) | **97.65**%(**6.32s**) |
| | FGM | **91.74**%(394.98s) | **97.13**%(1346.03s) | **97.37**%(1994.78s) | 97.54%(3253.97s) |
| url | PET | 98.15%(1100.28s) | 98.38%(1664.15s) | 98.21%(1528.01s) | 98.21%(1573.35s) |
| | SOFS | **98.32**%(**6.95s**) | **98.74**%(**7.05s**) | **98.92**%(**6.94s**) | **99.18**%(**6.94s**) |




(a) news　　(b) rcv1

(c) news　　(d) rcv1

**图 4.3**　“news” 和 “rcv1” 数据集上测试准确率和训练时间与选取特征个数之间的关系

表 4.4和图 4.3显示了三个算法测试准确度和训练时间之间的比较结果。由于 FGM 算法在 “url” 数据集上训练时间过久，因此表格中缺少相关实验结果。根据表格结果，我们可以发现 SOFS 的性能十分接近甚至要优于 FGM，尤其是当更多的特征被选取的时候。SOFS 和 FGM 算法都比基准算法 PET 要好。对于训练时间，PET 和 SOFS 在 “news” 数据集上的比较结果表明 PET 对于维度更敏感。一个有趣的现象在于 PET 算法在选取 0.5% 的特征时需要耗费更多的时间，原因在于此时 PET 算法收敛的速度过慢，需要反复地更新模型。FGM 算法是计算做复杂的特征选取算法，训练时间通常比在线特征选取算法高一个量级。此外，训练时间随着选取特征数目的增加也迅速增加。根据实验结果，我们可以发现 SOFS 算法在大规模高维数据集上进行特征选取的巨大优势。在某些实际问题中，往往需要在同一个数据集上反复运行多次在线算法使得模型收敛，此时 SOFS 算法的优势将更加明显。

### 4.5.4　图片检索中的应用

## 4.6　深度卷积神经网络模型简化

### 4.6.1　深度卷积网络模型简化建模

### 4.6.2　基于在线学习的模型简化

### 4.6.3　实验结果和评估

## 4.7　小结

# 第 5 章　照片集关联表达

## 5.1　照片集关联表达系统框架

## 5.2　照片集事件检测

## 5.3　照片集照片筛选

## 5.4　照片集风格选取

## 5.5　照片集故事合成

## 5.6　实验结果

## 5.7　小结

# 第 6 章　移动多摄像头视频自动剪辑

6.1　可计算视频编辑语法

6.2　移动多摄像头视频自动剪辑系统框架

6.3　音频剪辑

6.4　镜头切换点检测

6.5　视频镜头选取

6.6　实验结果

6.7　小结

# 第 7 章　总结与展望

## 7.1　本文总结

## 7.2　研究工作展望

# 参考文献


[1] Earth photographed over eight hundred billion mobile phones to take nearly 80%, Online; accessed 2-January-2017[M]. [S.l.]: [s.n.].

[2] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision. 2015, 115 (3): 211–252.

[3] Danyluk A, Provost F. Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network[C]//Proc. of Tenth International Conference on Machine Learning. [S.l.], 2014: 81–88.

[4] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of Artificial Intelligence Research. 1999, 11: 131–167.

[5] Zhou B, Jagadeesh V, Piramuthu R. ConceptLearner: Discovering visual concepts from weakly labeled image collections[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 1492–1500.

[6] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data[C]//Advances in neural information processing systems. [S.l.], 2006: 601–608.

[7] Vo P D, Ginsca A, Borgne H L, et al. On deep representation learning from noisy web images[J]. arXiv preprint arXiv:1512.04785. 2015.

[8] Sindhwani V, Keerthi S S. Large scale semi-supervised linear svms[C]//ACM, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. [S.l.]: ACM, 2006: 477–484.

[9] Schroff F, Criminisi A, Zisserman A. Harvesting image databases from the web[J]. IEEE transactions on pattern analysis and machine intelligence. 2011, 33 (4): 754–766.

[10] Chatzilari E, Nikolopoulos S, Kompatsiaris Y, et al. Salic: Social active learning for image classification[M]//[S.l.]: IEEE.

[11] Manwani N, Sastry P. Noise tolerance under risk minimization[J]. IEEE transactions on cybernetics. 2013, 43 (3): 1146–1151.

[12] Thathachar M A, Sastry P S. Networks of learning automata: Techniques for online stochastic optimization[M]. [S.l.]: Springer Science & Business Media, 2011.

[13] Sastry P, Nagendra G, Manwani N. A team of continuous-action learning automata for noise-tolerant learning of half-spaces[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2010, 40 (1): 19–28.

[14] Beigman E, Klebanov B B. Learning with annotation noise[C]//Association for Computational Linguistics, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. [S.l.]: Association for Computational Linguistics, 2009: 280–287.

[15] Bunescu R C, Mooney R J. Multiple instance learning for sparse positive bags[C]//ACM, Proceedings of the 24th international conference on Machine learning. [S.l.]: ACM, 2007: 105–112.

[16] Vijayanarasimhan S, Grauman K. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization[C]//IEEE, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. [S.l.]: IEEE, 2008: 1–8.

[17] Feng J, Xu H, Mannor S, et al. Robust logistic regression and classification[C]//Advances in Neural Information Processing Systems. [S.l.], 2014: 253–261.

[18] Izadinia H, Farhadi A, Hertzmann A, et al. Image classification and retrieval from user-supplied tags[J]. arXiv preprint arXiv:1411.6909. 2014.

[19] Izadinia H, Russell B C, Farhadi A, et al. Deep classifiers from image tags in the wild[C]//ACM, Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions. [S.l.]: ACM, 2015: 13–18.

[20] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE. 1998, 86 (11): 2278–2324.

[21] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. [S.l.], 2012: 1097–1105.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556. 2014.

[23] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 1–9.

[24] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 2818–2826.

[25] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[J]. arXiv preprint arXiv:1602.07261. 2016.

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 770–778.

[27] Reed S, Lee H, Anguelov D, et al. Training deep neural networks on noisy labels with boot-

strapping[J]. arXiv preprint arXiv:1412.6596. 2014.

[28] Sukhbaatar S, Bruna J, Paluri M, et al. Training convolutional networks with noisy labels[J]. arXiv preprint arXiv:1406.2080. 2014.

[29] Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 2691–2699.

[30] Azadi S, Feng J, Jegelka S, et al. Auxiliary image regularization for deep cnns with noisy labels[J]. arXiv preprint arXiv:1511.07069. 2015.

[31] Jain A K, Vailaya A. Image retrieval using color and shape[J]. Pattern recognition. 1996, 29 (8): 1233–1244.

[32] Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transactions on pattern analysis and machine intelligence. 1996, 18 (8): 837–842.

[33] Lowe D G. Object recognition from local scale-invariant features[C]//Ieee, Computer vision, 1999. The proceedings of the seventh IEEE international conference on: volume 2. [S.l.]: Ieee, 1999: 1150–1157.

[34] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//Springer, European conference on computer vision. [S.l.]: Springer, 2006: 404–417.

[35] Yang J, Jiang Y G, Hauptmann A G, et al. Evaluating bag-of-visual-words representations in scene classification[C]//ACM, Proceedings of the international workshop on Workshop on multimedia information retrieval. [S.l.]: ACM, 2007: 197–206.

[36] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]//ICML: volume 3. [S.l.], 2003: 856–863.

[37] Jiang F, Sui Y, Zhou L. A relative decision entropy-based feature selection approach[J]. Pattern Recognition. 2015, 48 (7): 2151–2163.

[38] Li F, Zhang Z, Jin C. Feature selection with partition differentiation entropy for large-scale data sets[J]. Information Sciences. 2016, 329: 690–700.

[39] Yang H, Lyu M R, King I. Efficient online learning for multitask feature selection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2013, 7 (2): 6.

[40] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial intelligence. 1997, 97 (1-2): 273–324.

[41] Pappu V, Panagopoulos O P, Xanthopoulos P, et al. Sparse proximal support vector machines for feature selection in high dimensional datasets[J]. Expert Systems with Applications. 2015, 42 (23): 9183–9191.

[42] Le Thi H A, Vo X T, Dinh T P. Feature selection for linear svms under uncertain data: Robust optimization based on difference of convex functions algorithms[J]. Neural Networks. 2014,

59: 36–50.

[43] Crammer K, Dekel O, Keshet J, et al. Online passive-aggressive algorithms[J]. Journal of Machine Learning Research. 2006, 7 (Mar): 551–585.

[44] Crammer K, Dredze M, Kulesza A. Multi-class confidence weighted algorithms[C]//Association for Computational Linguistics, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. [S.l.]: Association for Computational Linguistics, 2009: 496–504.

[45] Crammer K, Kulesza A, Dredze M. Adaptive regularization of weight vectors[C]//Advances in neural information processing systems. [S.l.], 2009: 414–422.

[46] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient[J]. Journal of Machine Learning Research. 2009, 10 (Mar): 777–801.

[47] Duchi J, Singer Y. Efficient online and batch learning using forward backward splitting[J]. Journal of Machine Learning Research. 2009, 10 (Dec): 2899–2934.

[48] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization[J]. Journal of Machine Learning Research. 2010, 11 (Oct): 2543–2596.

[49] Wu X, Yu K, Wang H, et al. Online streaming feature selection[C]//Proceedings of the 27th international conference on machine learning (ICML-10). [S.l.], 2010: 1159–1166.

[50] Huang H, Yoo S, Kasiviswanathan S P. Unsupervised feature selection on data streams[C]//ACM, Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. [S.l.]: ACM, 2015: 1031–1040.

[51] Wang J, Zhao P, Hoi S C, et al. Online feature selection and its applications[J]. IEEE Transactions on Knowledge and Data Engineering. 2014, 26 (3): 698–710.

[52] Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146. 2016.

[53] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387. 2015.

[54] Denil M, Shakibi B, Dinh L, et al. Predicting parameters in deep learning[C]//Advances in Neural Information Processing Systems. [S.l.], 2013: 2148–2156.

[55] Denton E L, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]//Advances in Neural Information Processing Systems. [S.l.], 2014: 1269–1277.

[56] Rigamonti R, Sironi A, Lepetit V, et al. Learning separable filters[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2013: 2754–2761.

[57] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions[J]. arXiv preprint arXiv:1405.3866. 2014.

[58] Ioannou Y, Robertson D, Shotton J, et al. Training cnns with low-rank filters for efficient image classification[J]. arXiv preprint arXiv:1511.06744. 2015.

[59] Tai C, Xiao T, Zhang Y, et al. Convolutional neural networks with low-rank regularization[J]. arXiv preprint arXiv:1511.06067. 2015.

[60] Mamalet F, Garcia C. Simplifying convnets for fast learning[C]//Springer, International Conference on Artificial Neural Networks. [S.l.]: Springer, 2012: 58–65.

[61] Hwang K, Sung W. Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1[C]//IEEE, Signal Processing Systems (SiPS), 2014 IEEE Workshop on. [S.l.]: IEEE, 2014: 1–6.

[62] Arora S, Bhaskara A, Ge R, et al. Provable bounds for learning some deep representations.[C]//ICML. [S.l.], 2014: 584–592.

[63] Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations[C]//Advances in Neural Information Processing Systems. [S.l.], 2015: 3123–3131.

[64] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks[C]//Springer, European Conference on Computer Vision. [S.l.]: Springer, 2016: 525–542.

[65] Gong Y, Liu L, Yang M, et al. Compressing deep convolutional networks using vector quantization[J]. arXiv preprint arXiv:1412.6115. 2014.

[66] Wu J, Leng C, Wang Y, et al. Quantized convolutional neural networks for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 4820–4828.

[67] Anwar S, Hwang K, Sung W. Fixed point optimization of deep convolutional neural networks for object recognition[C]//IEEE, Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. [S.l.]: IEEE, 2015: 1131–1135.

[68] Chen W, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick.[C]//ICML. [S.l.], 2015: 2285–2294.

[69] Liu B, Wang M, Foroosh H, et al. Sparse convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 806–814.

[70] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Advances in Neural Information Processing Systems. [S.l.], 2015: 1135–1143.

[71] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149. 2015.

[72] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710. 2016.

[73] Murray K, Chiang D. Auto-sizing neural networks: With applications to n-gram language models[J]. arXiv preprint arXiv:1508.05051. 2015.

[74] Anwar S, Hwang K, Sung W. Structured pruning of deep convolutional neural networks[J]. arXiv preprint arXiv:1512.08571. 2015.

[75] Hu H, Peng R, Tai Y W, et al. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures[J]. arXiv preprint arXiv:1607.03250. 2016.

[76] Changpinyo S, Sandler M, Zhmoginov A. The power of sparsity in convolutional neural networks[J]. arXiv preprint arXiv:1702.06257. 2017.

[77] Platt J C, Czerwinski M, Field B A. Phototoc: Automatic clustering for browsing personal photographs[C]//IEEE, Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on: volume 1. [S.l.]: IEEE, 2003: 6–10.

[78] Graham A, Garcia-Molina H, Paepcke A, et al. Time as essence for photo browsing through personal digital libraries[C]//ACM, Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. [S.l.]: ACM, 2002: 326–335.

[79] Gargi U. Modeling and clustering of photo capture streams[C]//ACM, Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval. [S.l.]: ACM, 2003: 47–54.

[80] Cooper M, Foote J, Girgensohn A, et al. Temporal event clustering for digital photo collections[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2005, 1 (3): 269–288.

[81] Loui A C, Savakis A E. Automatic image event segmentation and quality screening for albuming applications[C]//IEEE, Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on: volume 2. [S.l.]: IEEE, 2000: 1125–1128.

[82] Gong B, Jain R. Segmenting photo streams in events based on optical metadata[C]//IEEE, Semantic Computing, 2007. ICSC 2007. International Conference on. [S.l.]: IEEE, 2007: 71–78.

[83] Mei T, Wang B, Hua X S, et al. Probabilistic multimodality fusion for event based home photo clustering[C]//IEEE, Multimedia and Expo, 2006 IEEE International Conference on. [S.l.]: IEEE, 2006: 1757–1760.

[84] Shen X, Tian X. Multi-modal and multi-scale photo collection summarization[J]. Multimedia Tools and Applications. 2016, 75 (5): 2527–2541.

[85] Chu W T, Lin C H. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection[C]//ACM, Proceedings of the 16th ACM international conference on Multimedia. [S.l.]: ACM, 2008: 829–832.

[86] Hua X S, Lu L, Zhang H J. Photo2video—a system for automatically converting photographic series into video[J]. IEEE Transactions on circuits and systems for video technology. 2006, 16 (7): 803–819.

[87] Chu W T, Chen J C, Wu J L. Tiling slideshow: an audiovisual presentation method for consumer photos[J]. IEEE MultiMedia. 2007, 14 (3).

[88] Kuo T H, Tsai C Y, Cheng K Y, et al. Sewing photos: Smooth transition between photos[C]//Springer, International Conference on Multimedia Modeling. [S.l.]: Springer, 2011: 73–83.

[89] Shrestha P, de With P H N, Weda H, et al. Automatic mashup generation from multiple-camera concert recordings[C]//ACM Multimedia. [S.l.], 2010: 541-550.

[90] Russell S J, Norvig P. Artificial Intelligence — A Modern Approach[M]. [S.l.]: Pearson Education, 2010: I-XVIII, 1-1132.

[91] Nguyen D T D, Saini M, Nguyen V T, et al. Jiku director: A mobile video mashup system[C]//ACM Multimedia. [S.l.], 2013: 477–478.

[92] Saini M K, Gadde R, Yan S, et al. MoViMash: online mobile video mashup[C]//ACM Multimedia. [S.l.], 2012: 139-148.

[93] Hua X S, Lu L, Zhang H. Automatic music video generation based on temporal pattern analysis[C]//ACM Multimedia. [S.l.], 2004: 472-475.

[94] Arev I, Park H S, Sheikh Y, et al. Automatic editing of footage from multiple social cameras[J/OL]. ACM Trans. Graph. July 2014, 33 (4): 81:1–81:11. http://doi.acm.org/10.1145/2601097.2601198. DOI: 10.1145/2601097.2601198.

[95] Sundaram H, Chang S F. Computable scenes and structures in films[J]. IEEE Transactions on Multimedia. 2002, 4 (4): 482–491.

[96] Sharff S. The Elements of Cinema: Toward a Theory of Cinesthetic Impact[M]. [S.l.]: Columbia University Press, 1982.

[97] Lampi F, Kopf S, Benz M, et al. A virtual camera team for lecture recording[J]. IEEE MultiMedia. 2008, 15 (3): 58-61.

[98] Sumec S. Multi camera automatic video editing[M]. [S.l.]: [s.n.], 2006: 935–945.

[99] Hua X S, Lu L, Zhang H J. Optimization-based automated home video editing system[J]. IEEE Transactions on Circuits and Systems for Video Technology. 2004, 14 (5): 572–583.

[100] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems. 2014, 25 (5): 845–869.

[101] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]//NIPS: volume 14. [S.l.], 2001: 585–591.

[102] Candès E J, Li X, Ma Y, et al. Robust principal component analysis?[J]. Journal of the ACM

(JACM). 2011, 58 (3): 11.

[103] Eckart C, Young G. The approximation of one matrix by another of lower rank[J]. Psychometrika. 1936, 1 (3): 211–218.

[104] Golub G H, Van Loan C F. Matrix computations: volume 3[M]. [S.l.]: JHU Press, 2012.

[105] Yang Y, Shen H T, Ma Z, et al. l2, 1-norm regularized discriminative feature selection for unsupervised learning[C]//IJCAI proceedings-international joint conference on artificial intelligence: volume 22. [S.l.], 2011: 1589.

[106] Ye J, Zhao Z, Wu M. Discriminative k-means for clustering[C]//Advances in neural information processing systems. [S.l.], 2008: 1649–1656.

[107] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[M]//[S.l.]: Citeseer, 2009.

[108] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results[M]. [S.l.]: [s.n.], 2007.

[109] Luo P, Wang X, Tang X. Hierarchical face parsing via deep learning[C]//IEEE, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. [S.l.]: IEEE, 2012: 2480–2487.

[110] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.], 2014: 1717–1724.

[111] Divvala S K, Farhadi A, Guestrin C. Learning everything about anything: Webly-supervised visual concept learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2014: 3270–3277.

[112] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Recent Advances and Emerging Challenges of Feature Selection in the Context of Big Data[J]. Knowledge Based System. September 2015, 86 (C): 33–45.

[113] Zhai Y, Ong Y S, Tsang I. The Emerging ”Big Dimensionality”[J]. Computational Intelligence Magazine, IEEE. Aug 2014, 9 (3): 14-26.

[114] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research. 2011, 12: 2121–2159.

[115] Dredze M, Crammer K, Pereira F. Confidence-weighted linear classification[C]//Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA: ACM, 2008: 264–271.

[116] Ma J, Kulesza A, Dredze M, et al. Exploiting Feature Covariance in High-Dimensional Online Learning[C]//Proceedings of the Artificial Intelligence and Statistics. [S.l.], 2010: 493–500.

[117] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-

dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence. 2005, 27 (8): 1226–1238.

[118] Ramírez-Gallego S, Lastra I, Martínez-Rego D, et al. Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data[J]. International Journal of Intelligent Systems. 2017, 32 (2): 134–152.

[119] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research. 2008, 9: 1871–1874.

[120] Tan M, Tsang I W, Wang L. Towards ultrahigh dimensional feature selection for big data[J]. Journal of Machine Learning Research. 2014, 15 (1): 1371-1429.

[121] Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks[C]//Advances in Neural Information Processing Systems. [S.l.], 2016: 2074–2082.

[122] Yang X, Mei T, Xu Y Q, et al. Automatic generation of visual-textual presentation layout[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2016, 12 (2): 33.

[123] Ranjan A, Henrikson R, Birnholtz J P, et al. Automatic camera control using unobtrusive vision and audio tracking[C]//Graphics Interface. [S.l.], 2010: 47-54.

[124] Zhang C, Rui Y, Crawford J, et al. An automated end-to-end lecture capture and broadcasting system[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP). 2008, 4 (1): 6.

# 致　谢

在研究学习期间，我有幸得到了三位老师的教导，他们是：我的导师，中国科大 XXX 研究员，中科院 X 昆明动物所马老师以及美国犹他大学的 XXX 老师。三位深厚的学术功底，严谨的工作态度和敏锐的科学洞察力使我受益良多。衷心感谢他们多年来给予我的悉心教导和热情帮助。

感谢 XXX 老师在实验方面的指导以及教授的帮助。科大的 XXX 同学和 XXX 同学参与了部分试验工作，在此深表谢意。

# 在读期间发表的学术论文与取得的研究成果

## 已发表论文

1. **Yue Wu**，Steven C.H. Hoi，Tao Mei，and Nenghai Yu. “ Large-scale Online Feature Selection for Ultra-high Dimensional Sparse Data”，ACM Transactions on Knowledge Discovery from Data (accepted).

2. **Yue Wu**, Xu Shen, Tao Mei, Xinmei Tian, Nenghai Yu, and Yong Rui. “Monet: A System for Reliving Your Memories by Theme-Based Photo Storytelling”, *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2206-2216, Nov. 2016 (the first two authors contributed equally).

3. **Yue Wu**，Tao Mei，Ying-Qing Xu，Nenghai Yu，and Shipeng Li. “MoVieUp: Automatic Mobile Video Mashup”，*IEEE Transactions on Cirtuits and Systems for Video Technology*，vol. 25，pp. 1941-1954，Dec 2015.

4. Jianlong Fu, **Yue Wu**, Tao Mei, Jinqiao Wang, Hanqing Lu, and Yong Rui. “Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging”，In Proceedings of the IEEE International Conference on Computer Vision(ICCV)，pp. 1985-1993，2015 (the first two authors are with equal contribution).

5. **Yue Wu**，Shiyang Lu，Tao Mei，Jian Zhang，and Shipeng Li. “Local visual words coding for low bit rate mobile visual search”，In Proceedings of the 20th ACM international conference on Multimedia (MM), pp. 989-992, ACM, 2012.

6. **Yue Wu**, Tao Mei, Nenghai Yu, and Shipeng Li. “Accelerometer-based single-handed video browsing on mobile devices: design and user studies”，In Proceedings of the 4th International Conference on Internet Multimedia Computing and Service (ICIMCS ’12). ACM，New York，NY，USA，157-160.

7. Dayong Wang，Pengcheng Wu，Peilin Zhao，**Yue Wu**，Chunyan Miao，and Steven CH Hoi. “High-dimensional data stream classification via sparse online learning”，In Data Mining (ICDM)，2014 IEEE International Conference on(ICDM)，pp. 1007-1012. IEEE，2014.

## 待发表论文

1. **Yue Wu**, Steven C.H. Hoi, Chenghao Liu, Jing Lu, Doyen Sahoo, and Nenghai Yu. "SOL: A Library for Scalable Online Learning Algorithms", Neurocomputing, 2016(Under review).