**目 录**

引言 1

1 研究内容完成情况 1

1.1 多模态信息智能理解算法研究 1

1.1.1 Attend and Imagine: Multi-label Image Classification [1] 1

1.1.2 Breaking Winner-takes-all [2] 3

1.2 多模态信息融合算法研究 6

1.2.1 LABIN: Balanced Min Cut for Large-scale Data [3] 6

1.2.2 Online Heterogeneous Transfer Learning by Knowledge Transition [4] 9

1.2.3 Online Heterogeneous Transfer by Hedge Ensemble of Offline and Online Decisions [5] 12

1.3 面向机器人应用的轻量级深度学习算法研究 15

1.3.1 火焰检测 15

1.3.2 人脸检测 16

1.3.3 仪表数字识别 20

1.3.4 语音去噪 21

2 总结 22

参考文献 23

插图清单

图 1 基于注意力机制的CNN-RNN架构示例图 2

图 2 BNID突发新闻图像数据集样例 2

图3 “赢家出局”策略示意图 4

图4 “赢家出局”策略训练算法流程 4

图5 “赢家出局”策略测试算法流程 5

图6 结合群体注意力模块的双流迭代“赢家出局”网络结构示意图 5

图7 LABIN关键算法示例 7

图8 算法时间代价对比 8

图9 从文本源域到图像目标域的知识传递 10

图10 在文本-图片数据集上时间代价对比 12

图11 OHTHE整体框架 14

图12 OHTHE关键算法 14

图 13 室外火焰检测 16

图 14 室内火焰检测 16

图 15 人脸识别流程图 17

图 16 mtcnn模型演示图 18

图 17 活体检测流程图 18

图 18 mobilefacenet流程图 19

图 19 人脸识别实机演示 19

图 20 仪表数字识别整体流程及运行环境 20

图 21 读数区域检测整体结构 20

图 22 评价指标IoU 21

图 23 DNN网络训练流程图 22

图 24 使用DNN降噪前后音频对比图 22

附表清单

[表 1 Pascal VOC 2007数据集实验结果对比 3](#_Toc103806644)

[表2 THUMOS14测试集实验结果对比 6](#_Toc103806645)

[表3 模拟数据集上ACC指标对比 7](#_Toc103806646)

[表4 模拟数据集上NMI指标对比 8](#_Toc103806647)

[表5 真实数据集上ACC指标对比 9](#_Toc103806648)

[表6 真实数据集上NMI指标对比 9](#_Toc103806649)

[表7 在文本-图片数据集上不同算法的错误率（使用线性核） 10](#_Toc103806650)

[表8 在多语言数据集上不同算法的错误率（使用线性核） 11](#_Toc103806651)

[表9 文本-图片数据集上性能对比 15](#_Toc103806652)

[表10 跨语言数据集上性能对比 15](#_Toc103806653)

[表11 视频数据集上性能和时间对比表 15](#_Toc103806654)

[表 12 人脸识别耗时统计表 19](#_Toc103806655)

# 引言

随着视觉感知、语音识别与合成技术以及人工智能技术与产业的发展，智能机器人与人们生活具有越来越密切的联系，人机交互已经获得很大的进步。但是对于多模态信息的智能理解与融合仍是在许多人机交互在真实环境场景应用急需解决的问题。本年度主要研究内容如下：（1）多模态信息智能理解算法研究；（2）多模态信息融合算法研究；（3）面向机器人应用的轻量级深度学习算法研究。

本年度计划撰写相关学术论文6篇，申请相关专利3件。目前已经撰写相关学术论文6篇，申请相关专利4项，完成了本年度计划。

# 研究内容完成情况

## 多模态信息智能理解算法研究

### Attend and Imagine: Multi-label Image Classification [1]

在过去几年中，由于学习理论的发展以及ImageNet等大规模数据集的可用性，单标签图像分类的任务取得了重大进展，其中每个图像仅与一个标签相关。然而，在现实世界中，我们通常看到的是多标签图像，其中每个图像包含多个对象或多个语义标记。例如，来自社交媒体和新闻的图像通常包含丰富的语义标记。

与单标签图像分类相比，多标签分类问题更具挑战性并具有以下特殊点：第一，多个对象可以位于图像中的任何位置，并且每个对象都与一个或多个标签相关联；第二，图像中不同区域的重要性不同，一张多标签图像中的感兴趣区域可能与另一张图像里的感兴趣区域非常不同；第三，对于特定类型的对象（例如人），它们在不同的图像中可能具有不同的大小，这进一步增加了多标签图像学习的难度；第四，由于复杂的图像结构，图像的多个标签可能具有标签依赖性，例如，一艘船可能会与水一起出现。因此，直接将单标签方法应用于多标签图像分类是没有帮助的。

为了解决多标签图像分类问题，人们做了很多工作。传统方法可分为两类：基于问题转换的方法和基于算法自适应的方法。近年来，基于卷积神经网络的方法在这一领域占据主导地位。然而，上述方法均未考虑标签依赖性，因此很难对标签或对象之间的相关性进行建模。最近，递归神经网络（RNN）被证明是机器翻译中序列相关性建模的强大工具，于是一些工作提出使用CNN-RNN方法来充分提取标签之间的相关性，一般的，其遵循机器翻译中的编码器-解码器设计模式：CNN模块始终设置为编码器，将图像编码为固定长度的特征向量，该向量接着被送入RNN模块，并解码成一系列标签。在基于CNN-RNN的方法中，标签可以被视为比句子更密集的序列，因为它们没有句法结构，也没有不必要的限定词。一般来说，在预测特定标签时，整个图像被用作RNN的输入。然而，对于实际的多标签图像，它们通常非常复杂，并且一个特定的标签通常与图像的一个区域相关。因此，直接使用整个图像作为每个标签的输入可能会影响性能。

最近，注意机制已被用于机器翻译、图像字幕和视觉问答并取得了良好效果。特别的，注意机制可以被视为一种特征选择策略，可以用于对有用信息进行关注而忽略冗余。基于上述讨论，本文提出了一种基于注意力机制的CNN-RNN方法来解决多标签分类问题，与传统机器翻译和图像字幕中基于注意力机制的方法不同，我们的模型考虑了注意力区域在每一步的连续变换，框架示例图如图 1所示。具体的，我们遵循机器翻译和图像字幕中流行的编码器-解码器模式。在编码器中，我们从预训练的CNN的浅层卷积层中提取特征。在解码器中，我们首先用提取的特征初始化LSTM。然后，在每个时间步，我们使用注意机制计算一个动态上下文，并将上下文注入LSTM。此外，为了防止LSTM在传播过程中忘记整个图像，我们还通过将图像在CNN中更高层的特征与LSTM合并，从而在每一时间步提醒我们的模型。

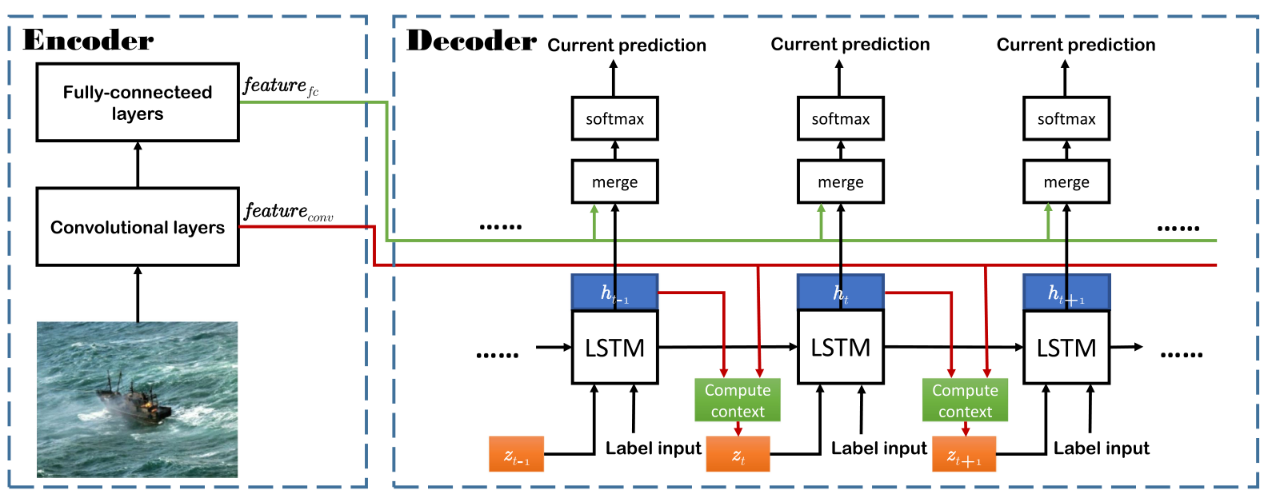


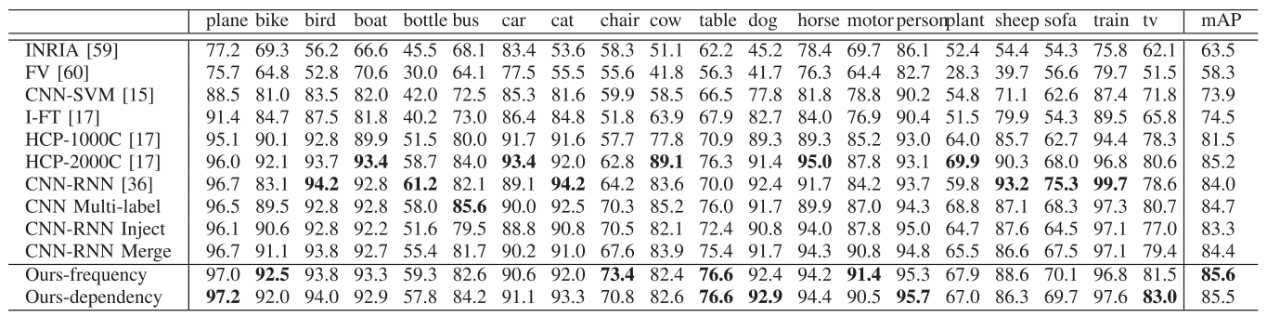
图 1 基于注意力机制的CNN-RNN架构示例图



图 2 BNID突发新闻图像数据集样例

在本文中，我们构建了一个新的图像数据集BNID（突发新闻图像数据集），部分数据样例如图 2所示，它的图像都来自突发新闻。在实验部分，我们在自己的数据集BNID上验证了模型的有效性，此外我们也在两个流行的多标签数据集上评估了我们的模型，即Pascal VOC 2007和MS-COCO。我们将我们的模型与几种最先进的多标签分类方法I-FT[6]、 HCP-1000C/HCP-2000C[6]和CNN-RNN[7]进行了比较，实验结果表明该方法在三个数据集上均取得了较好的性能，对比实验结果如表 1所示，例如，在Pascal VOC 2007数据集中达到了先进的85.6% mAP。

表 1 Pascal VOC 2007数据集实验结果对比



### Breaking Winner-takes-all [2]

近近年来，动作定位在视频检索、视频监控等领域体现了广袤的潜在应用空间，得到越来越多的关注。动作定位对于视频理解非常重要，特别是对于未修剪的视频，因为它们通常包含动作实例和背景，而我们只对动作实例感兴趣。给定一段视频，动作定位旨在同时解决两个问题：识别视频中每个动作实例的开始时间和结束时间；识别每个动作实例的类别。

早期的工作主要基于滑动窗口法结合手工制作的特征对生成的候选提案进行分类。近年来，借助基于深度学习的图像或视频分析方法，动作定位取得了很大进展。然而，上述的方法均为全监督动作定位方法，即严重依赖时间注释（每个动作实例的类别以及时间流）。显然，一帧一帧人工手动标记动作类别不仅费力费时，而且对于不同的注释员来说标注都存在一定的主观性，从而导致注释存在严重偏差，在消耗大量人工、时间成本下也无法保证效果。

弱监督学习旨在利用较弱、易获取的注释解决高级任务，在静态图像领域取得了广泛的研究。以语义分割为例，其目的是预测给定图像中每个像素的类别，由于全监督语义分割使用的像素级别注释是非常难以标注的，所以研究人员试图通过只使用较弱的图像级的类别标签来完成语义分割。基于上述思想，我们在视频中的定位问题中利用弱监督学习的技术来减弱对昂贵注释的依赖。

具体地，[8]提出的对抗性消除方法在弱监督语义分割领域取得了最先进的成果，该方法对图像中属于目标类的区域进行迭代选择，并下一次迭代中将所选区域从图像中移除，最后得到完整分割结果。

受此启发，我们提出了一种新的“赢家出局”迭代优胜劣汰策略。对于输入视频，我们均匀的将其裁剪成多个片段，通常，某个片段会包含一些与视频标签强相关的视频帧，从而导致网络过度偏向此具有判别力的视频帧，阻止对其他相关动作进行定位，我们称之为“赢家通吃”问题。为解决该问题，“赢家出局”策略迭代定位视频中最有区别性的片段，并将它们从训练片段集中移除。每次迭代根据分类网络学习到的类别信息给每个片段分配一个类别特定的重要性分数，我们选择重要性分数最大的视频作为“赢家”，并将该片段从视频中删除以“出局”，然后对分类网络再次训练，使网络挖掘并定位其他辨别力较弱的片段，直至定位到视频中的所有动作实例。“赢家出局”策略示意图如图3所示。

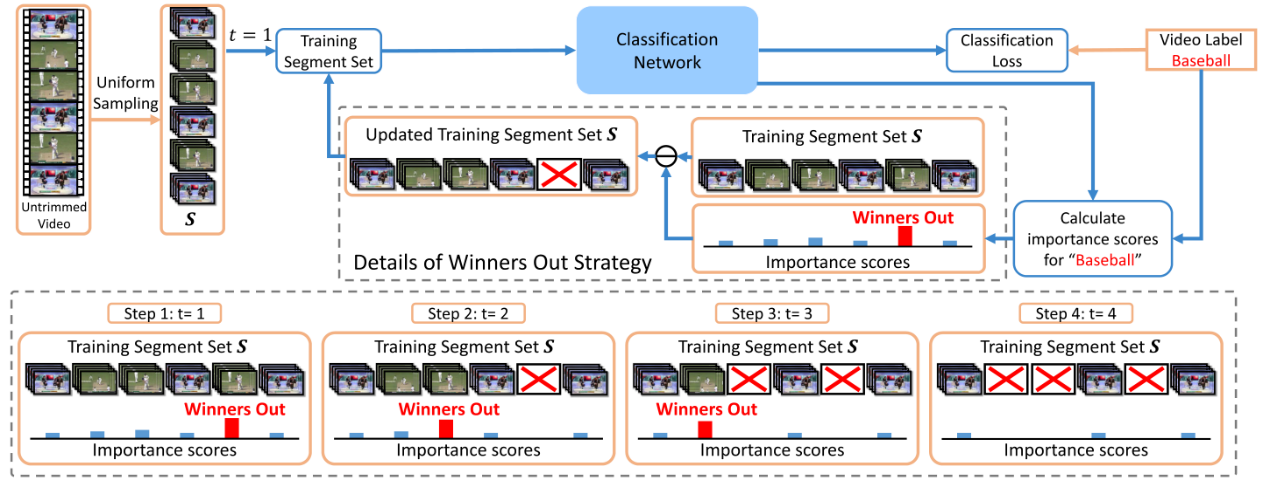


图3 “赢家出局”策略示意图

“赢家出局”策略的训练算法流程和测试算法流程如图4，图5所示。

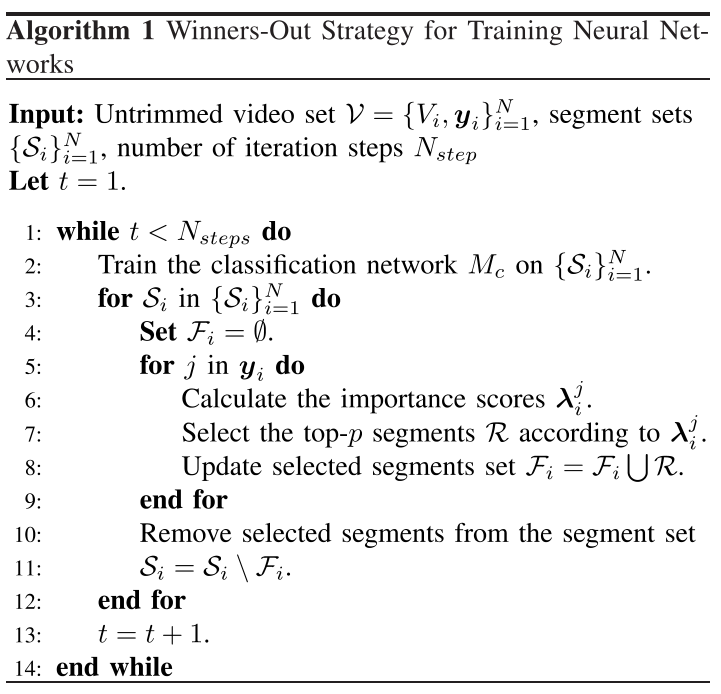


图4 “赢家出局”策略训练算法流程

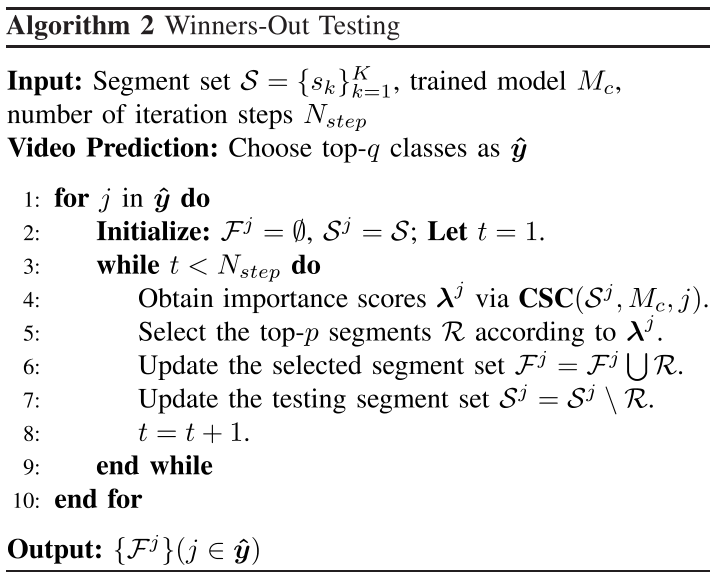


图5 “赢家出局”策略测试算法流程

同时，图像和视频不同模态信息之间的巨大领域差异也导致以下挑战：（1）视频相比图像具有更高的复杂度。图像往往在分类网络的最后一层卷积层上加入全局平均池化层，然而，由于视频远比图像复杂，平均池化层是平等对待每个片段的，它没有选择有区别的片段来表示视频，所以无法输出良好的视频分类特征。为此，我们探索了注意力机制来解决这一问题，我们给每个片段级特征分配一个注意力权重，然后进行加权求和从而得到最终的全局特征。再利用增强的特征结合“赢家出局”策略训练分类网络。（2）如果考虑到注意力机制，“赢家出局”策略中选择要移除的片段的标准需要特别制定。要选择去除哪个片段，我们首先要分析每个片段的重要性。“赢家出局”网络结构示意图如图6所示。

在[8]中，作者使用CAM[9]分析每个像素与目类别之间的相关性，在CAM中每个像素的相关性只取决于特征向量激活值和最后一个全连通层的权值，若直接应用在我们的方法中会忽略注意力权重，使重要性分数不够精确。为了解决这个问题，我们提出了一种类别区分的定位技术，即特定于类别的分数计算。在这种技术中，每个片段的重要性分数与它的注意力权重有关，同时，我们还使用了片段特征，以及最后一层全连接层连接视频级别特征和动作类别的权重。因此，该技术很好地利用了在训练阶段学习到的类别信息，为选择判别能力最强的动作实例提供了更好的参考。

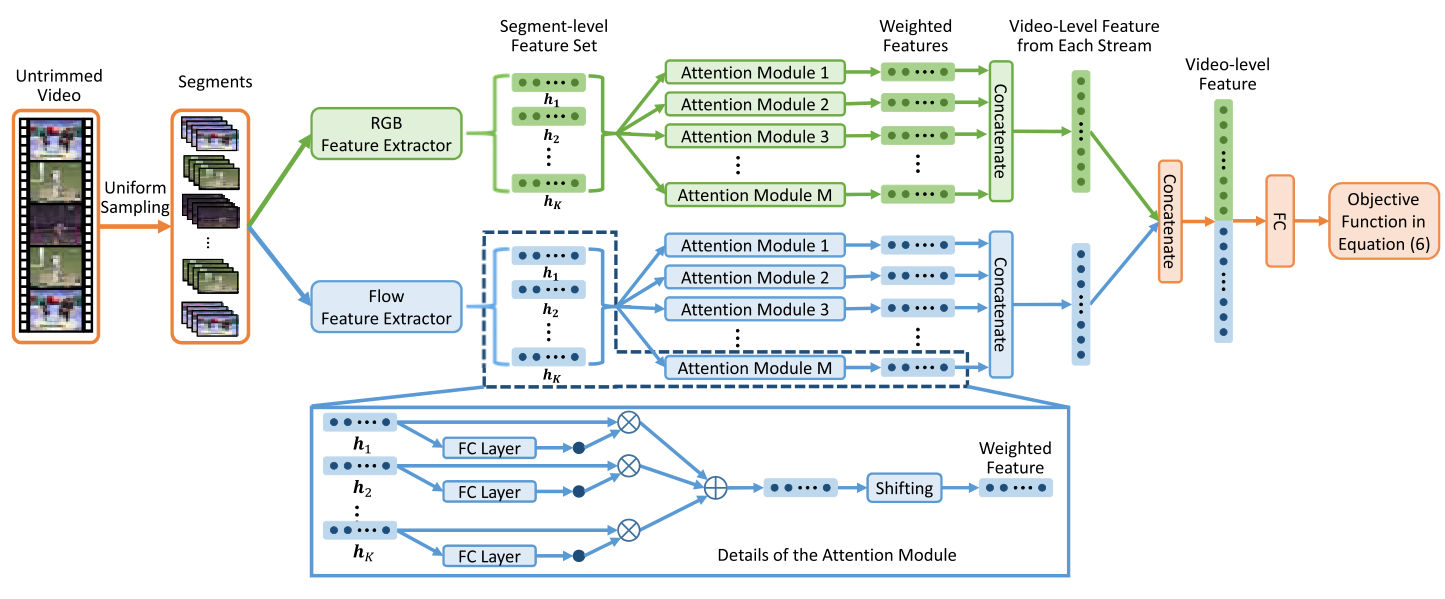
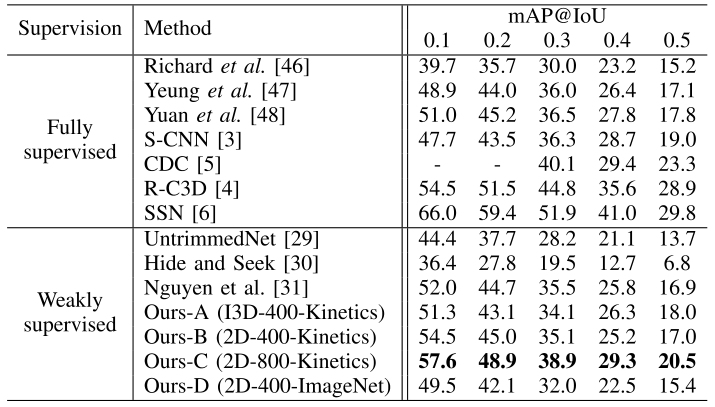


图6 结合群体注意力模块的双流迭代“赢家出局”网络结构示意图

实验部分，我们在广泛使用的基线数据集THUMOS14、ActivityNet1.3上测试了我们的方法。大量的实验结果表明，我们的方法在ActivityNet1.3上优于最先进的弱监督方法，并在THUMOS14上将mAP从16.9%提高到20.5%。值得注意的是，即使是弱视频级监督，我们的方法也能达到与使用全帧级监督的方法相当的准确性。

表2 THUMOS14测试集实验结果对比



## 多模态信息融合算法研究

### LABIN: Balanced Min Cut for Large-scale Data [3]

聚类是一个基础的研究领域，近十多年来出现了许多聚类分析的算法，包括谱聚类，子空间聚类，多视角聚类，协同聚类等。其中，谱聚类算法由于易于实现以及效果较好，受到了较大关注。这些方法在许多领域发挥了其作用，例如图像聚类[10]，图像分割**错误!未找到引用源。**等。

谱聚类算法通常先对相似度矩阵进行特征值分解，然后使用k-means或谱旋转等方法来从特征向量中获取最终的聚类结果。此外，由于特征值分解的时间复杂度较高，谱聚类通常具有O（n^3）的时间复杂度，其中n为对象的数目。为了解决这一可扩展性问题，研究者们已提出了三类方法。第一类致力于减少特征值分解步骤的时间代价[12]。第二类方法使用预处理来减少数据大小，然后在减小的数据集上进行谱聚类[13]。第三类方法使用基于anchor的策略来构建二分相似矩阵，然后对二分相似矩阵而非全相似矩阵进行聚类[14]。最近，[10]提出了一种新的谱聚类算法，SBMC，该算法中提出了一个SBMC模型，隐式使用了平衡正则化来获取更平衡的数据划分，并且该模型可以同时最小化图割和在聚类间平衡划分。然而，上述这些方法都没法很好地适应大规模数据，或者对输入数据有一定的要求，具有一定局限性。

为了解决上述提到的问题，我们提出了LABIN算法，将SBMC拓展到大规模数据。为了减小输入相似度矩阵的大小，我们使用了基于anchor的策略来构建二分相似矩阵，然后在二分相似矩阵而非原始矩阵上进行计算。我们还使用了ISR算法来优化新的模型，并且提出了一种高效的方法来进行特征值分解，该方法能够达到线性的时间复杂度。对比原有的SBMC，我们的方法显著减少了空间复杂度，并且由于使用了ISR，具有更好的鲁棒性。对比已有的基于anchor聚类算法，我们的方法可以从由噪声或孤立样本组成的复杂数据集中发现高质量的聚类结构。

算法2为我们提出的关键算法，即LABIN。给定一个数据集，首先构建样本-锚相似度矩阵，然后迭代更新s和Y直到收敛。在新算法中，我们需要 O(r1(nm2 + r2nc2)) 时间来迭代求解 s 和 Y，其中 r1 是更新 s 的迭代次数，r2 是更新 Y 的平均迭代次数。这里，由于有限的解空间，Y 的离散解收敛速度非常快，因此 r2 通常非常小。 因此，LABIN 的时间复杂度为 O(nm2) (m c)，与时间复杂度为 O(n3) 的传统谱聚类方法相比，计算量显着降低。

文本

描述已自动生成

图7 LABIN关键算法示例

我们首先在模拟数据集上进行实验。我们模拟生成了11个2维数据集。如表3，表4所示，我们使用准确率（ACC）和正则化互信息（NMI）来评价聚类性能。我们提出的模型大部分数据集上达到了最好的ACC和NMI。

表3 模拟数据集上ACC指标对比

表格

描述已自动生成

表4 模拟数据集上NMI指标对比

表格

描述已自动生成

除此之外，如图8所示，在时间代价方面，我们的算法也要优于大部分现有模型，并且随着数据集大小增长，时间代价呈线性增加。

图表, 折线图

描述已自动生成

图8 算法时间代价对比

在真实世界的数据集上，我们在8个大规模数据集上进行了对比实验，在所有这8个数据集上，我们都实现了LABIN。实验结果见表5和表6，可以看到，我们的算法在5个数据集上超过了所有的方法，具体来说，在CA28数据集上，LABIN相比第二的KASP方法，在NMI上平均提升了接近11%。在PA25数据集上，LABIN相比第二的KASP，在ACC上平均提升了接近15%。

表5 真实数据集上ACC指标对比

表格

描述已自动生成

表6 真实数据集上NMI指标对比

表格

描述已自动生成

### Online Heterogeneous Transfer Learning by Knowledge Transition [4]

迁移学习致力于利用源域知识来增强目标域的学习性能，其中源域和目标域的数据分布或表示可能是不同的。在过去的十年中，许多迁移学习的算法被提出并且被证明在许多应用上是有效的，例如文本分类[15]，情感分析[16]，图像分类[17]，推荐系统[18]等。

现有的大部分迁移学习的方法专注于同构的设定，即源域和目标域有共同的特征空间。近来，许多研究者[19][22]已经开始考虑异构迁移学习（HTL）的设定。区别与同构迁移学习，异构迁移学习中源域和目标域有着不同的特征空间。例如，目标域的学习任务可以是使用SIFT特征来进行图像分类，并且只有少量的标注数据，源域数据可以是一些有标签的文本文档的词袋特征。异构迁移学习是非常具有挑战性的，原因在于源域和目标域在特征空间没有交集。一些学者提出了使用传递迁移学习来解决这一困难[23][27]，即利用一些辅助的中间数据来连接两个异构的特征空间，这样源域数据的知识可以通过中间域传递到目标域。

大多数现有研究专注于离线(批次)的学习形式，即假设目标域的训练数据是可以随时获取的。然而，这种假设在真实世界中有时候可能是不成立的，目标域数据可能是顺序到达的，例如，在线社交平台或者股价平台。在这些应用中，需要使用在线学习的方法来处理数据系统，并且即时反馈。因此，在目标域执行一个批次学习算法对于这种在线学习的场景的不合适的，因此需要开发一个高效的在线学习系统。现有的针对异构迁移学习的在线学习算法的主要缺陷在于，源域的特征空间必须是目标域的一个子集。

为了解决上述提到的问题，我们提出了一个新的学习算法OHKT,该算法通过一个包含共现数据的中间域，将知识从源域传递到目标域。图9展示了我们方法的整体框架。

图示

描述已自动生成

图9 从文本源域到图像目标域的知识传递

我们首先在标注的源域数据上训练一个分类器来分类无标签的共现数据。随后，在这些带伪标签的共现数据上，我们再训练一个分类器来协助目标域的学习任务。这样，从源域数据上抽取的标签数据通过共现数据传递到了目标域数据。针对在线学习，我们提出了一个涉及共现数据和目标域数据的在线学习模型，并且设计了一个算法来解决出现的优化问题。

对于文本-图片数据集，我们使用NUS-WIDE数据集来生成在线分类任务。我们随机挑选了10个类别，构建了45个二分类任务。对每个任务，随机挑选600张图片作为目标域数据，1200个标注文本作为源域数据，1600个无标注文本-图片对作为共现数据。对于多语言数据集，我们使用了Multi-Language数据集来构建在线学习任务，我们构建了90个二分类任务。我们将我们提出的方法与6个基线方法进行比较，使用的评价指标为错误率。在文本-图片数据集上，我们的方法平均超过了第二好的方法14%，在多语言数据集上，我们的方法平均超过了第二好的方法27%。总的实验结果如表7,表8所示。

表7 在文本-图片数据集上不同算法的错误率（使用线性核）

表格

描述已自动生成

表8 在多语言数据集上不同算法的错误率（使用线性核）

表格

描述已自动生成

除此之外，我们通过实验验证了提出的算法在文本图像数据集的代表性任务上的效率。由于 HET-PA 和 PCA-PA 算法是基于 PA 进行的，我们只报告 PA 的运行时间作为基准。我们使用线性内核来测试算法，并在图10中报告代表性任务的运行时间。

图表, 折线图

描述已自动生成

图10 在文本-图片数据集上时间代价对比

我们看到 PA 是比较方法中最有效的方法，因为它不使用任何辅助信息。由于 SVM 的定期训练，HTLIC 需要更多时间。 OHKT 需要比 PA 更多的时间来使用在同现数据上训练的分类器来更新分类器。总的来说，与 PA 相比，OHKT 的在线计算时间相当，并且可以达到更好的性能。可以从其他任务中得出类似的观察结果。

### Online Heterogeneous Transfer by Hedge Ensemble of Offline and Online Decisions [5]

迁移学习致力于通过利用源域的知识来提升目标域的学习性能。迁移学习已经被广泛地应用到目标域数据有限或者获取代价昂贵等场景。在过去的十多年内，许多的迁移学习方法被提出，大多数这些方法都专注于同构设定，即源域和目标域有共同的特征空间。

最近，一种被称为异构迁移学习的模式被提出。不同于传统的同构迁移学习，异构迁移学习的源域数据和目标域数据来自两个不同的特征空间。异构迁移学习已经被应用到许多计算机视觉和机器学习的任务上。例如，对于图像分类任务，通常只有有限的标注图片，但有大量的有标签的文本作为辅助信息。而图片和文本通常是用不同的特征表示的。因此，如何利用这些标签文本来增强预测性能是一个非常具有挑战性的任务。

一种解决办法是通过收集一些与图片相关的文本作为共现数据，例如图片标题或者包含图像的文档中的文本。在这种情况下，知识可以从文本数据顺利地传输到图片，从而提升图像分类的性能。值得注意的是，相比图像数据昂贵的标注成本，无标签的文本-图片共现数据可以很容易地获得。尽管许多方法已经验证了上述思路的有效性[28] [29]，大多数现有的工作专注于离线/批次学习问题，它们假设目标域的训练样本可以马上获得。然而，在现实世界中，这种假设可能不成立，目标域的训练样本可能是顺序到达的，例如社交网络场景。为了解决这一缺陷，我们专注于在线异构迁移学习，在我们的场景下，目标域数据是以一种在线方式到达的，而源域数据则是被离线收集和标注的，并且给定共现数据作为辅助信息。在线异构迁移学习有两个主要挑战。首先是源域和特征域的特征空间可能是完全不同的，其次是目标域的标注数据是很少的，很难建立一个精确的对应关系图来连接源域和目标域。

为了解决上述挑战，我们提出了一种新的技术OHTHE，通过同时利用源域和目标域知识来增强在目标域的学习性能。以图像分类任务为例，OHTHE的整体框架如图11所示。为了连接源域和目标域，通过同时使用源数据和无标签的共现数据，我们构建了一个基于异构相似性的离线决策函数。我们使用了一个对冲加权策略来同时利用离线和在线决策函数，以增强学习性能。具体来说，OHTHE根据预测值和决策值的差值来动态调整两种决策函数的比重。我们的关键算法如图12所示，其中，根据映射函数φ和损失函数ψ的不同，我们将算法分为多个版本：OHTHE-01,OHTHE-CUBE,OHTHE-SIG。

图示

描述已自动生成

图11 OHTHE整体框架

文本

描述已自动生成

图12 OHTHE关键算法

在文本-图片数据集上，我们使用了从FLICKR上收集的NUS-WIDE数据集。在跨语言数据集上，我们使用了[30]提供的数据集，包含了原始的文档和对应的翻译版本。在视频数据集上，我们基于[31]构建了一个大规模的二分类任务。

如表9所示，我们的算法在文本-图片数据集的大多数任务上，HTLIC-PA和OHTHE-SIG取得了更好的成绩，优于PA，这表明从源域转移知识对目标域任务有版主，在某些任务上OHTHE-SIG优于HTLIC-PA，并且在其他任务上，OHTHE-SIG与HTLIC-PA不相上下，证明了我们异构知识迁移方法的有效性。如表10所示，在跨语言数据集上，OHTHE-CUBE在大多数任务上都优于PA。如表11所示，在视频数据集上，我们对比了不同方法的运行时间以及错误率，在运行时间上，我们的方法由于引入了额外的计算相似度来寻找最邻近的样本，因此运行时间出于中等水平。在平均错误率上，OHTHE-SIG超过了所有方法，达到了最优的性能。

表9 文本-图片数据集上性能对比

表格

描述已自动生成

表10 跨语言数据集上性能对比

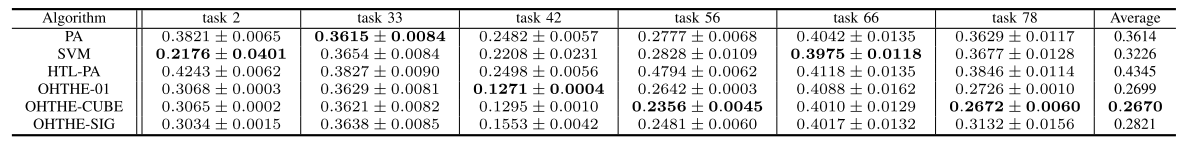


表11 视频数据集上性能和时间对比表

表格

描述已自动生成

## 面向机器人应用的轻量级深度学习算法研究

### 火焰检测

随着人类社会的发展进步，火灾成为了不可忽视的多发性问题。火灾一次次展示了其对于人员伤亡和财产的巨大破坏性。火灾的预防和检测一直是人类与火灾斗争过程中关注的焦点。近些年来，随着数字通信技术、计算机技术的发展, 数字图像处理技术获得了广泛的应用。因此，人们提出了利用图像处理技术进行火灾的监控与报警。

现存的火焰检测算法中主要有以下2类：（1） 利用图像通道、颜色直方图及帧差法等图像处理技术进行火焰识别检测，但当图像背景复杂、目标形变较大时，该类方法的抗干扰能力较低，容易出现误差。（2）利用支持向量机和BP神经网络等机器学习方法进行火焰识别检测，但是需要对火焰设定大量的物理特征并进行提取，过程较为繁琐且火焰形态多变，对于不同的特征关注点，在不同数据集上所得的准确度可能会降低。

由于利用深度学习进行目标检测已经广泛应用于其他领域，故我们将基于深度学习的目标检测框架Faster R-CNN用于空中目标检测。深度卷积神经网络提取的是图像的颜色、边缘和轮廓等特征，所以针对火焰这种颜色、 边缘和轮廓特征比较突出的目标，其算法的优越性可以得到充分发挥。针对小样本利用迁移学习策略实现高效学习，可有效降低复杂背景、光照变化以及目标形态多变等因素对于检测的影响，进一步提高火焰检测的稳定性和准确性。

为进一步提高抗干扰性能，（1）我们使用颜色抖动等数据增强方法来屏蔽由于不同燃烧材料的光谱信息差异大所产生的影响。（2）面对现实生活存在许多与火焰颜色相近的物体，我们使用更多的负样本数据进行训练。（3）在需要全屏检测/屏蔽区域时，我们使用一个屏蔽矩阵与网络输出做“与”操作，得到最终输出。

最终场景检测下，我们使用1920\*1080，f/1.8，30fps的摄像头进行录像，分别在室外与室内环境进行了测试，室外环境为晴朗白天，非阳光直射，拍摄距离为摄像头与火焰垂直高度2m，水平距离3m，火焰大小约15\*10cm；室内环境为日光灯位于火焰上方，四周无直射光源，拍摄距离为摄像头与火焰垂直距离1.5cm，水平距离1.5m，火焰大小约1.5\*1.5cm，可视结果如图 13，图 14所示。

树林间的小路

描述已自动生成

图 13 室外火焰检测

屏幕上有字

描述已自动生成

图 14 室内火焰检测

### 人脸检测

人脸识别过程主要分为三步，流程图如图 15所示：

（1）对摄像头拍摄图片进行人脸区域检测和人脸标志点检测，选取面积最大的人脸图像。（2）对选取到的图像进行活体检测。（3）可信度大于阈值的及认定为活体，进一步进行人脸验证，与人脸数据库内图片进行对比，相似度大于阈值则匹配成功。

图示

描述已自动生成

图 15 人脸识别流程图

在人脸检测中，我们使用mtcnn模型来完成人脸区域和人脸标志点检测这两项任务，如图 16所示，其过程分为下面几个阶段：

（1）预处理，图像中不同人的脸大小不同，通过缩放图片更加准确地检测到人脸区域，扩大检测距离的范围；

（2）P-Net，生成一系列候选框，初步筛选并优化候选框；

（3）R-Net，进一步筛选候选框，对高度重合的候选框进行合并；

（4）O-Net，输出检测到人脸框的坐标和人脸标志坐标；

（5）选取面积最大的人脸图像，进入活体检测阶段。

图形用户界面, 应用程序, 日程表

描述已自动生成

图 16 mtcnn模型演示图

下一阶段中，我们LBP+SVM来进行活体检测，流程图如图 17所示，主要过程分为下面3步：

（1）将人脸RGB图像转换到HSV颜色空间，应用LBP算子凸显纹理变化， 统计HSV直方图，拼接HSV每个通道的直方图构成颜色纹理特征；

（2）利用svm算法进行二分类，计算检测人脸图像真实性的可信度；

（3）当可信度大于阈值就判断为活体，进入人脸验证阶段。

图示

描述已自动生成

图 17 活体检测流程图

对于已经判断为活体的图片我们使用mobilefacenet进行人脸验证，流程如图 18所示，主要分为下面5步过过程：

（1）为每个检测到的人脸图像缩放为112\*112；

（2）统计分析人脸图像的像素值分布，调整过亮或过暗人脸部位的像素值，减弱亮度的影响；

（3）通过特定的网络结构和卷积操作将人脸图像转换为128维的特征向量；

（4）与人脸数据库中每个人脸的 特征向量计算余弦相似度；

（5）选取最大相似度所对应的人脸作为预测结果，其大于阈值则识别成功。

图示, 日程表

中度可信度描述已自动生成

图 18 mobilefacenet流程图

最后效果展示中，我们分别使用了真实的人脸和用手机展示的人脸进行测试，测试其检测精度与活体检测的效果，结果如图 19所示，左图为真实人脸，右图为手机展示的人脸。

人的脸被修图

低可信度描述已自动生成

图 19 人脸识别实机演示

本次测试中一共录入10张人脸图片，每张图片消耗8KB内存，与库内的图片进行对比时一共消耗时长为0.017ms。每个阶段所消耗的时长如表 12所示。

表 12 人脸识别耗时统计表

|  |  |
| --- | --- |
| 阶段 | 耗时/ms |
| 人脸检测 | 120-150 |
| 活体检测 | 75 |
| 人脸验证（10人） | 130-140 |
| 总耗时 | 325-365 |

### 仪表数字识别

本工作的主要目的是为了对仪表上的数字进行识别，如图 20所示，其整体流程主要分为4步：

（1）是由摄像头对水表进行拍摄，获取原图；

（2）使用单片机对图片进行粗略的裁剪；

（3）对读数区域进行分割；

（4）对分割好区域的图片进行数字识别。

手机屏幕截图

中度可信度描述已自动生成

图 20 仪表数字识别整体流程及运行环境

数据源的获取为网络爬虫和自行拍摄的图片，使用单片机或嵌入式设备对图片的粗略裁剪。裁剪后图片使用MobieleNetV3模型进行读数区域的检测，检测后获得目标区域的坐标后，再进行剪裁压缩，所得准确的读数区域图片，读取区域检测整理结构如图 21所示。

图示

描述已自动生成

图 21 读数区域检测整体结构

在读数区域检测的测试中使用IoU作为评价指标，如图 22所示。在无后处理情况下可以达到平均IoU为0.87，在使用后处理后可以进一步达到平均IoU为0.97。聚类后的数字识别准确率为0.995。

形状

中度可信度描述已自动生成

图 22 评价指标IoU

### 语音去噪

针对语音去噪使用了减谱法和深度学习法。其中减谱法算法原理主要是通过对采集的原始信号进行噪声估计，最后降噪后的信号就是使用采集的信号减去估计的噪声信号。噪声估计主要分为两个步骤：

（1）对初始的静默时间内的信号进行统计，得到估计噪声；

（2）通过发音间隔的静默时间内的信号特征，对估计噪声进行不断修正。

减谱法算法顺序处理整个音频分为下面几步：

（1）对前5帧进行快速傅里叶变换（FFT），将这5帧的振幅的绝对值求平均，得到频域内噪声的估计振幅，记为n；

（2）对之后每一帧i的时域信号进行FFT，转换为频域信号；

（3）减去估计噪声n，得到预测信号；

（3）如果的信噪比小于阈值，则认为当前帧为噪音，更新；

（4）对预测信号进行IFFT，还原成时域信号；

（5）对相邻帧重叠部分相加，还原语音信号。（6）拼接得到最终输出。

对于使用深度学习进行语音降噪，我们尝试了两种方法，第一种是使用SEGAN网络，该网络是用U-NET结构的生成网络和VGG结构的判别网络训练GAN，但训练测试出效果很差，故不采用。第二种是使用DNN+维纳滤波的方法进行降噪，先利用DNN消除声音信号中的噪音，再用传统的维纳滤波进一步去噪，DNN网络训练流程如图 23所示。

图示

描述已自动生成

图 23 DNN网络训练流程图

使用DNN降噪的效果目前具有较好的效果，降噪前与降噪后的音频对比可视化如图 24所示，可以看到添加噪声的音频在经过DNN降噪处理后的音频图与清晰声音的音频图基本一致，且在人耳听觉条件下处理音频与清晰音频表现基本一致。

手机屏幕的截图

描述已自动生成

图 24 使用DNN降噪前后音频对比图

# 总结

本年度主要从多模态信息智能理解算法研究、多模态信息融合算法研究和面向机器人应用的轻量级深度学习算法研究三个方面进行展开，研究内容具有一定的新颖性和理论价值，应用前景十分广阔，推动了深度学习的多个任务例如图像识别、迁移学习、人脸识别等的快速发展。本年度已撰写相关学术论文6篇，已申请相关专利4项，完成了本年度计划。

在研究过程中，总结了很多经验和教训，例如在本年度的研究过程中前期调研不够充分，对现有方法的分析不够全面深入，对后期的方法研究和设计造成了困难。下一年度将继续在这三个主要的研究方向进行研究，吸取经验教训，确保后期能够顺利开展。

# 参考文献

1. Fan Lyu, Qi Wu, Fuyuan Hu, Qingyao Wu, Mingkui Tan, "Attend and Imagine: Multi-label Image Classification with Visual Attention and Recurrent Neural Networks", to appear in IEEE Transactions on Multimedia, 2019.
2. Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, Mingkui Tan, "Breaking Winner-takes-all: Iterative-winners-out Networks for Weakly Supervised Temporal Action Localization", IEEE Transactions on Image Processing, 2019.
3. Xiaojun Chen, Renjie Chen, Qingyao Wu\*, Yixiang Fang, Feiping Nie, Zhexue Huang, "LABIN: Balanced Min Cut for Large-scale Data", IEEE Transactions on Neural Networks and Learning Systems (TNNLS), TNNLS, 2019.
4. Hanrui Wu, Yuguang Yan, Michael Ng, Huaqing Min, Qingyao Wu\*, "Online Heterogeneous Transfer Learning by Knowledge Transition", ACM Transactions on Intelligent Systems and Technology, 2019.
5. Yuguang Yan, Qingyao Wu\*, Mingkui Tan\*, Michael Ng, Huaqing Min, Ivor Tsang, "Online Heterogeneous Transfer by Hedge Ensemble of Offline and Online Decisions", IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 29(7), pp. 3252-3263, 2018.
6. Y . Wei et al., “CNN: Single-label to multi-label,” 2014, arXiv preprint arXiv:1406.5726.
7. J. Wang et al., “CNN-RNN: A unified framework for multi-label image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2285–2294.
8. Y . Wei, J. Feng, X. Liang, M.-M. Cheng, Y . Zhao, and S. Y an, “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1568–1576.
9. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2921–2929.
10. X. Chen, J. Z. Huang, F. Nie, R. Chen, and Q. Wu, “A self-balanced min-cut algorithm for image clustering,” in Proc. Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2080–2088.
11. S. X. Yu and J. Shi, “Multiclass spectral clustering,” in Proc. 9th IEEE Int. Conf. Comput. Vis., vol. 1. Oct. 2003, pp. 313–319.
12. W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, “Parallel spectral clustering in distributed systems,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 568–586, Mar. 2011.
13. D. Yan, L. Huang, and M. I. Jordan, “Fast approximate spectral clustering,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2009, pp. 907–916.
14. X. Chen, F. Nie, J. Z. Huang, and M. Yang, “Scalable normalized cut with improved spectral rotation,” in Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI), 2017, pp. 1518–1524.
15. Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Transferring naive Bayes classifiers for text classification. In Proceedings of the 22nd Association for the Advancement of Artificial Intelligence. AAAI Press, 540–545.
16. Peilin Zhao and Steven C. Hoi. 2010. OTL: A framework of online transfer learning. In International Conference on Machine Learning. 1231–1238.
17. Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 1855–1862.
18. Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 10 (2010), 1345–1359.
19. Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2009. Translated learning: Transfer learning across different feature spaces. In Advances in Neural Information Processing Systems. 353–360.
20. Ben Tan, Etheng Zhong, Michael K. Ng, and Qiang Yang. 2014. Mixed-transfer: Transfer learning over mixed graphs. In Proceedings of the 2014 SIAM International Conference on Data Mining. SIAM, 208–216.
21. Yuguang Yan, Wen Li, Michael K. Ng, Mingkui Tan, Hanrui Wu, Huaqing Min, and Qingyao Wu. 2017. Learning discriminative correlation subspace for heterogeneous domain adaptation. In International Joint Conference on Artificial Intelligence. 3252–3258.
22. Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. 2018. Semi-supervised optimal transport for heterogeneous domain adaptation. In International Joint Conference on Artificial Intelligence. 2969–2975.
23. Michael K. Ng, Qingyao Wu, and Yunming Ye. 2012. Co-transfer learning via joint transition probability graph based method. In Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining, KDD.
24. Ben Tan, Yangqiu Song, Erheng Zhong, and Qiang Yang. 2015. Transitive transfer learning. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1155–1164.
25. Qingyao Wu, Michael K. Ng, and Yunming Ye. 2014. Cotransfer learning using coupled Markov chains with restart. IEEE Intelligent Systems 29, 4 (2014), 26–33.
26. Liu Yang, Liping Jing, Jian Yu, and Michael K. Ng. 2016. Learning transferred weights from co-occurrence data for heterogeneous transfer learning. IEEE Transactions on Neural Networks and Learning Systems 27, 11 (2016), 2187–2200.
27. Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. 2011. Heterogeneous transfer learning for image classification. In Proceedings of the 25th Association for the Advancement of Artificial Intelligence. 1304–1309.
28. L. Yang, L. Jing, J. Yu, and M. K. Ng, “Learning transferred weights from co-occurrence data for heterogeneous transfer learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 11, pp. 2187–2200, Nov. 2016.
29. L. Niu, X. Xu, L. Chen, L. Duan, and D. Xu, “Action and event recognition in videos by learning from heterogeneous Web sources,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 6, pp. 1290–1304, Jun. 2017.
30. M. Amini, N. Usunier, and C. Goutte, “Learning from multiple partially observed views—An application to multilingual text categorization,” in Proc. NIPS, 2009, pp. 28–36.
31. O. Madani, M. Georg, and D. A. Ross, “On using nearly-independent feature families for high precision and confidence,” Mach. Learn., vol. 92, nos. 2–3, pp. 457–477, 2013.