多模态视觉数据处理综述

摘要：视觉数据来自于不同的设备，不同的获取条件可以看作是不同的模态，比如红外图像、可见光图像和遥感图像。一种模态就是我们获取视觉数据的一种形式，当我们处理的信息具有多种形式时，就是多模态处理。多模态视觉处理旨在建立能够处理多模态视觉图像的模型。多模态视觉数据处理是一个跨学科，持续发展且影响力在逐渐增强的领域。本文旨在综述多模态视觉数据处理的方法，以及分类，也就是：多模态表示学习、模态间映射、对齐、融合以及协同学习。

1. 引言

我们所生活的世界本身就是多模态的，我们看到的物体，听到的声音，感受到的纹理，闻到的气味，尝到的味道都是不同的模态[1]。除了主观感受上的多模态性，由于视觉传感器的不同特性，视觉数据的多模态通常也具有多模态特性。

多模态意味着数据的异质性，多模态视觉数据处理领域面对着独特的挑战。从多模态数据中学习，为我们提供了一个更好的理解数据方式，能够实现信息的互补同时去除冗余。本文将多模态视觉处理面临的挑战分成五个部分，这五个部分是多模态视觉处理面对的核心问题，也是目前研究的热点问题，同时需要更多的研究来促进其发展。

1) 表示学习

多模态图像数据处理面对的首要问题是将多模态数据的互补信息和冗余信息提取出来，将多模态数据中蕴含的语义信息抽值为实值向量。多模态图像数据的异质性使得建立多模态图像数据的表示成为一项有挑战的工作。比如红外图像与可见光图像具有不同的表示方法。

2) 模态间映射

如何在多模态图像数据间映射关系是多模态图像数据面临的第二大挑战。模态间映射是将给定实体的情况下，获取其它模态转换在该实体下的表示过程。多模态数据不仅有异质性，同时模态间关系也是主观开放的。比如：一张图片有不同的方式来表示它，但是可能没有最优的模态间映射。

3) 对齐

对齐是多模态数据处理面对的第三大挑战，其旨在发现多模态数据的不同子部件间的相互关系，从而使得学习到的多模态模型更加准确。比如：对于红外与可见光的同一人的两张或者多张人脸图像在同一坐标系下达到像素级的对齐[2]。

4) 融合

多模态视觉数据处理中面对的第四大挑战是将两个模态甚至是很多的模态数据进行结合，然后进行预测任务。比如：医学图像中CT图对于骨骼结构能够获得比较好的显示效果，而MRI则对软组织结构的获取更有优势，对这两种模态进行融合可以帮助医生进行诊断以及制定治疗计划[3]。

5) 协同学习

多模态视觉数据面对的第五大挑战是在不同模态间进行知识的迁移。在样本数量少、同时标注数量有限以及标注数据不可靠的情况下，通过不同模态间的知识迁移能够提高模型的准确率。

本文按照Baltrušaitis等人[1]提出的多模态机器学习所面对挑战分成五类，将多模态视觉数据处理面对的挑战同样分为五类：表示学习、模态间映射、对齐、融合和协同学习。第2章将对表示学习的方法在多模态视觉数据中的应用进行综述，第3章将多模态图像数据映射常见的使用方法分为两种，基于实例的和基于生成模型的。第4章探讨了多模态视觉数据的对齐问题，按照显性对齐和隐形对齐进行分别讨论。第5章对融合的方法进行综述，分成与模型无关的方法和基于模型的方法。第6章将协同学习分为平行数据与非平行数据进行分别探讨。最后，对本文内容进行总结，总结多模态视觉数据处理的方法，以及本文的主要工作。本文的主要工作在对多模态视觉数据面临的挑战进行分类，同时综述了各类挑战下的代表性工作。

2. 多模态表示学习

将多模态数据表示为可以计算的模型并不是一件容易的事情，按照Bengio等人[4]的划分，将“特征”和“表示”看作是同一表达，都是将同一实体表述为向量或者是张量。多模态表示学习就是使用来自多个实体的信息进行表示。多模态表示学习面对着许多困难：数据的异质性、不同程度的噪声水平、数据丢失的问题。

从卷积神经网络在视觉目标的分类任务中的优异表现[5]，可以看出表示学习对于多模态视觉数据的处理中起到了非常关键的作用。Benigo等人[4]指出好的表示学习应当具有如下的特性：平滑性、时空一致性、稀疏性以及良好的聚类特性。Srivastava等人[6]指出不同模态间在表示空间下具有相似性，在某些模态数据缺乏的情况下，在基于观测的多模态数据训练可以提高模型精度。

在过去的十几年中，单模态的表示已经有了广泛的研究[7]，同时，表示特征也从原先的手工设计特征演变成为基于数据驱动的特征，后者常与深度深度相关也被称为深度特征。比如在21世纪伊始，图像的表示常用一种手工设计特征，尺度不变变换(Scale Invariant Feature, SIFT)[8]。现在我们对于图像的表示常常会使用基于数据驱动的方法，比如卷积神经网络(Convolutional Neural Network, CNN)[5]。

下面将多模态的表示分成两类，一类是联合表示，另一类是协同表示，如图1.1所示。联合表示是将多个单一模态的信息集中在同样的表示空间中，而对等表示是将单模态的数据分别表示，但是模态之间存在着约束。

联合表示在数学上可以表示为：

 (2-1)

其中，是单模态到模态的映射，可以使深度神经网络，受限玻尔兹曼机，或者是递归神经网络。协同表示的数学表达如下：

 (2-2)

其中每个模态都有对应的映射函数(如上和)，然后将后面的结构投影到协同空间。常见的协同的方法有，最小化余弦距离[9]，最大化相关性[10]。

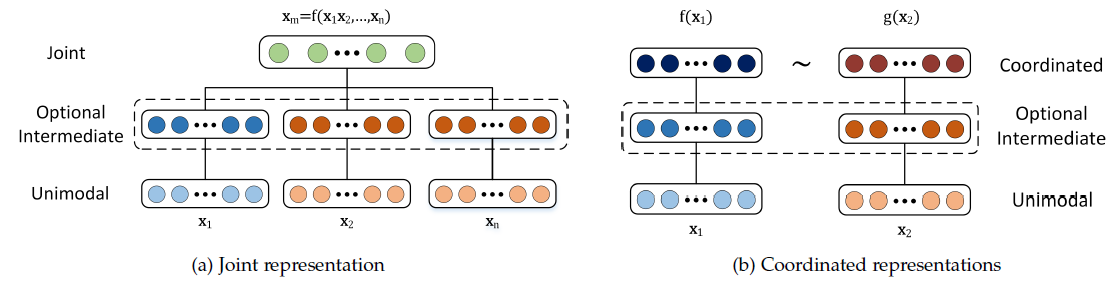


图2.1 联合表示与协同表示

2.1 联合表示

联合表示是将多模态信息映射到联合的多模态空间中，如公式2-1所示。联合表示在多模态图像数据的训练和推断过程中使用广泛。最简单的联合表示是将多模态特征进行串联，也被叫做早期融合[11]。本节将多模态视觉数据的表示方法分为两类，第一类是神经网络，第二类是概率图模型。

1) 神经网络

神经网络已经成为非常重要的多模态视觉数据表示的重要方法，下面我们将详细的描述神经网络是如何进行多模态联合表示的。

一般来说，神经网络是由多个层级联构成，每个层之后都会进行非线性激活。神经网络要成为表示的方法，首先要针对特定任务(比如目标检测)进行训练。由于深度神经网络的多层特性，每个级联层所表示的特征都是抽象的，所以我们通常使用最后一层或者倒数第二层特征作为数据表示。使用神经网络建立多模态联合表示时，每个模态都需要经过独立的神经网络的层，然后在隐含层中将多模态数据映射到联合空间中[12]。多模态联合表达通过多个隐含层，也可以直接用于预测。这样的模型可以端对端的训练，同时进行表示以及执行相应的任务。这是因为使用神经网络进行表示学习时，与多模态前期融合是一致的。

因为神经网络的训练需要大量的有标签数据，但是高质量的标签数据并不是容易获得的，所以使用无监督数据(比如自编码器[13])和相关领域的有监督数据进行预训练[14]来获得多模态表示成为比较流行的方式。

Ngiam等人[11]提出将自编码器(Autoencoder)的概念扩展到多模态数据处理中，他们使用自编码器对每个模态进行表示，然后再将两个模态融合，后面再使用一个自编码器进行解码。Ding等人[15]提出将多模态人脸数据进行融合，然后利用自编码器得到多模态表示。

利用神经网络联合表示的主要优势在于有标签数据不足时，神经网络可以利用无标签数据进行预训练。另外，如果对于特定的任务，无标签数据具有一般性，同时不需要最优的表示，我们可以微调表示结果来适应特定任务[16]。利用神经网络进行联合表示学习比较难以解决数据缺失的问题，虽然有写学者提出了一些改进的方法[11][16]。另外，深度神经网络比较难以训练[17]，但是作为正在迅速发展的领域，一些技术方法在不断的发展完善中，比如：正则化[18]、批标准化[19]和自适应梯度算法[20]。

2) 概率图模型

概率图模型通过使用隐随机变量来进行表示学习，本节我们将阐述概率图模型在多模态联合表示中的应用。概率图模型中的表示学习有深度玻尔兹曼机(Deep Boltzmann Machine, DBM)[21]，其使用受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)作为一部分模块 [22]。类似于神经网络的级联结构，深度玻尔兹曼机同样是将数据抽象到更高的水平来获得表示。区别于神经网络，DBM不需要有监督的数据来进行训练。因为图模型的数据表示形式是概率的，虽然可以将图模型改变成神经网络，但是这样就失去了图模型的生成特性[21]。

Srivastava和Salakhutdinov提出将多模态深度置信网络[23]和多模态深度玻尔兹曼机[24]用于多模态表示。Ouyang等人利用多视角数据利用多模态深度玻尔兹曼机进行人体姿态检测[12]。

概率图模型最大的优势是其生成特性，能够处理数据丢失的问题，即使是某一模态丢失，我们仍然能够根据其它模态生成一些样本。深度玻尔兹曼机这样的概率图模型最大的劣势是难以训练，需要巨大的计算成本，使用近似变分的方法可以进行近似的训练。

2.2 协同表示

不同于联合表示，协同表示则是另外一种表示方法。联合表示是将多模态映射到共同的空间中，协同表示则是对不同模态进行学习，通过限制条件得到模态间的协同关系。我们首先叙述通过相似性矩阵来度量模态间的协同关系，然后论述模态间的结构化的关系。

1) 相似性模型

相似性模型是在协同空间中最小化模态间距离。比如，用来表示红外图像的狗，以及可见光图像的狗，我们希望能最小化两种模态狗的图像的距离，使得狗之间的图像的距离要小于狗与汽车之间的距离[25]。

由于神经网络在表示学习上的突出表现，已经成为一种流行的建立协同表示的方法。神经网络的优势在于能够以端对端的方式学习到模态间的相互关系。Frome等人提出了深度视觉语义嵌入模型，利用内积与排序损失函数，最小化两种模态协同表示的距离[25]。Pan等人使用了相似的方法用在视频与文本上[26]。

1) 结构化协同空间

除了在模态表示上对相似性进行学术，结构协同空间模型同时对模态表示的结构信息进行约束。结构化的约束主要体现在哈希编码(Hashing)，跨模态检索，以及图像字幕的生成。

结构化的协同空间主要用于跨模态哈希编码上，也就是将高维数据压缩成紧密的二维编码，对于相似的物体有着相近的二维编码[27]。跨模态哈希编码的出现时为了实现跨模态检索出现的[12]。哈希编码对于多模态数据有三点要求：要求二进制表示必须位数可控；来自不同模态的相同对象要求相似的哈希编码；在空间中必须保留相似性。Jiang和Li利用端对端深度学习的方法学习了图像和对应的句子之间的二进制空间[30]。Li等人利用卷积神经网络建立了多模态遥感图像的二进制编码空间[31]。

典型相关性(Canonical Correlation Analysis, CCA)是结构化协同空间的特殊例子[32]。CCA能够计算出线性映射，同时使两个模态间的相关性最大，并且对两个模态进行正交限制。CCA模型广泛地用于跨模态检索[33]，视觉数据的融合[34]。核典型相关性分析(Kernel Canonical Correlation Analysis, KCCA)利用核函数实现多模态到协同空间的非线性映射[35]。为了处理数据规模大的情况，深度典型相关性分析(Deep Canonical Correlation Analysis, DCCA)被引入到多模态表示中[36]。

3.模态间映射

从一个模态映射到另外一个模态是多模态视觉数据处理研究的重要部分。也就是对于某一实体给定某个模态表示生成另一模态表示的方法就是模态映射。比如，我们现在有一张红外下人的图片，我们想得到人的可见光图像，这样就是从红外这一模态到可见光模态的映射[37]。

虽然多模态映射的方法有很多，而且是与模态特性相关，但是多模态映射同样具有非常多的共通点。我们将多模态映射分为两类，一类是基于个例，另一类是基于生成方法的。基于个例的方法是用词典学习在多模态之间映射。基于生成的方法则是建立一个生成模型实现多模态映射。二者的不同之处类似于非参数模型与参数模型，如图3.1所示。

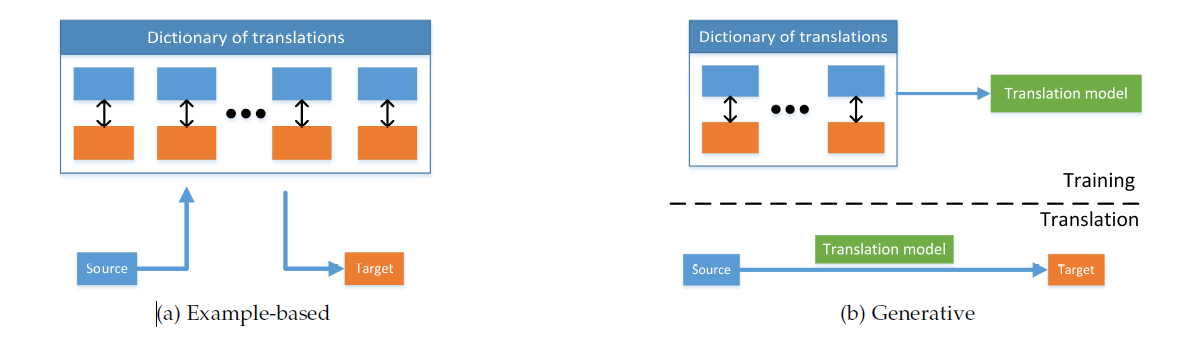


图3.1 基于个例的模态间映射与基于生成的模态间映射

3.1 基于个例的模态间映射

基于个例的模态间映射受限于训练样本的缺乏，也就是所谓的词典(如图3.1)。我们将词典分为两类，一类是基于检索的模型，另一类是基于组合的模型。基于检索的模型直接使用而且不修改映射，而基于组合的模型需要更多的模态间映射规则，这些规则基于大量的检索实例。

1) 基于检索的模型

基于检索的模型是最简单的多模态映射形式，它们将词典中最靠近的样本作为检索结果。检索可以在单模态空间中进行，也可以在两个模态的语义空间中进行。

给定一个需要映射的模态实例，单模态检索就是在源模态词典中找到最接近的模态实例，比如利用图像的视觉特征。这样的方法广泛的应用于生成图像的文字描述中[38]。Yagcioglu等人利用基于卷积神经网络(Conolutional Neural Network，CNN)的图像表示方法来检索视觉相似图像用于自适应紧邻选择[39]。单一模态的检索方法的优势在于当我们进行单模态检索时，我们可以获得比较好的单模态表示。但是单模态的映射通常需要后续的排序处理，这就表明了单模态空间的相似性并不能很好的完成模态间映射的任务。

可以替代的方案就是我们在模态检索中采用多模态中间语义空间进行相似性比较。早期的手工设计语义空间由Farhadi等人提出[40]。Socher等人通过学习协同表示句子和CNN视觉特征来进行模态间映射[41]。

在语义空间的多模态检索往往比在单模态上的检索要表现更好，这反映出语义空间是对多模态的高度表达，同时也是对检索的优化结果。但是，语义空间的检索需要构建语义空间，同时也需要大量的训练样本。

2) 基于组合的模型

基于组合的模型将检索更加进了一步，不同于从词典中检索实例，基于组合的模型则将检索的实例组合起来获得更好的映射。基于组合的检索的动机在于不同模态的图像具有相同的可以探索的结构，多数组合规则是手工设计的或者是基于启发式的。

基于个例的多模态映射的最大问题在于词典:词典的模型巨大并且词典的推断过程缓慢(尽管利用哈希编码可以改善这一问题)。另外基于个例的映射得到某个可以理解的准确的映射是不现实的，因为词典中并不总是包含源个例，除非执行的任务非常的简单，或者是词典的规模非常的巨大。

3.2 基于生成的方法

基于生成的方法用于多模态映射需要建立多模态映射模型，利用给定源单模态映射到其它模态。这样的方法需要对两个模态充分的理解，所以是非常具有挑战的。

我们将基于生成的方法分成三类：基于语法、编码解码。基于语法的模型是将任务简化然后使用语法来选择目标域。编码解码模型将源模态进行编码到隐表示，然后解码器利用这个隐表示生成目标模态。

1) 基于语法的模型

基于语法的模型需要与训练的语法规则来生成特定的模态。通过从源模态检测高层级的概念，比如图像中的物体，视频中的动作。将这些检测合并起来，然后利用预训练的语法规则得到目标模态。

一些基于语法的方法依靠图模型来生成目标模态。Yang等人利用图像中提取出来的视觉特征来预测句子，然后利用隐马尔科夫模型(Hidden Markov Model, HMM)对句子风格修改[42]。Zitnick等人提出使用条件随机场来生成抽象视觉场景[43]。因为语法的存在，基于与训练的语法模型方法生成模态更有逻辑性。但是语法也限制了生成的多样性，同时需要分离的模型和训练数据。

2) 编码解码模型

基于端对端训练的编码解码模型是目前最流行的多模态映射技术。利用编码器对源模态进行编码表示，然后通过解码器进行解码生成目标模态，编码和解码过程通过一次完成。编码解码模型首先被用于机器翻译中[44],[45]，随后又将编码解码结构用于多模态图像的分割[46],[47]。

编码解码结构首先需要对源模态进行编码，不同的模态有着不同的方法。图像通常采用卷积网络进行编码[5],[48]，同时其是视频的表示方法。对于任务的特殊条件，手工设计的特征仍然有着应用[49]。

解码是利用编码的结果来对目标模态进行生成的过程。虽然基于神经网络的编码解码模型广泛的以用于利用句子生成图像[50]，利用图像生成另外模态的图像[51]。基于神经网络的编码解码模型取得了比较大的成功，但是仍然有着许多开放的问题。Devlin等人认为神经网络很有可能是通过记住训练样本而不是学习到理解图像中的场景。基于这样的观察，K紧邻(K-nearest Neighbor)模型和神经网络在生成方式相类似的。

4. 对齐

我们将多模态对齐定义为确认两种或者是更多模态中子部件的相关性。比如在红外与可见光图像中，将同一物体的两种模态图像进行对齐。这对多模态的进一步处理很重要，同时也是多模态的检索的重要内容。

我们将多模态对齐分为两种，一种是显性对齐，一种是隐性对齐。显性对齐中，我们是直接将模态间的子部件对齐，比如对齐视频与相应的字幕[53]。隐性对齐是作为其它任务的中间过程，比如基于文字描述的图像检索就需要隐性对齐[54]。

4.1 显性对齐

我们将模型的最要目标是对齐的文章分为显性对齐。显性对齐中非常重要的内容是相似性矩阵。大多数的方法依靠于对子部件的相似性矩阵进行测量。我们将显性对齐分为无监督的和有监督的两种方式。无监督的方式对与两种模态间的实例没有直接的对齐标签，而有监督的方式存在标签。

1) 无监督

无监督多模态对齐处理模态间对齐不需要模态间对齐的标签。动态时间规整 (Dynamic Time Warping, DTW)[55]是一种动态编程方法，广泛应用于多视角时间序列中。动态时间规整通过测量模态间的相似性矩阵，然后通过时间调整(插入帧)找到最优的匹配。动态时间规整需要时间来对模态进行比较，同时得到相似性矩阵。通过手工设计相似性矩阵，DTW可以直接用于多模态对齐。Tapaswi利用设计的相似性矩阵来对齐综艺画面和文字简介[56]。

图模型在多模态对齐中也非常流行。Yu和Ballard利用生成图模型来对齐图像中目标和说话的词语[57]。 Noulas等人使用了一个动态贝叶斯网络来对齐说话的人和视频[58]。

动态时间规整以及图模型都需要对齐的模态不能有大的时间跳跃。动态时间规整需要我们自己设计相似性矩阵，图模型则需要专家知识来构建。

2) 有监督

有监督的对齐需要有标签对齐的实例。通过对有标签的样本进行训练来生成获取相似性来对齐模态。

大量的有监督的序列对齐技术是受无监督对齐启发的。Bojanowsk等人提出类似于典型时间规整的方法，但是他们充分利用了对齐实例进行训练[59]。深度学习用于显性对齐也逐渐流行起来[60]。

4.2 隐性对齐

不同显性对齐，阴性对齐是作为其它任务的中间环节。我们将阴性对齐分成两种，一种是基于图模型的，一种是神经网络的方法。

1) 图模型

图模型作为早期的对齐方法，广泛用于机器翻译中。图模型需要人工建立模态间的映射，建立这样的模型需要训练数据或者是人工的定义。

2) 神经网络

通过对齐可以有效地提高模态间映射效果。神经网络是一种比较好的处理模态间映射结构，使用编码解码结构或者是使用跨模态检索。在没有阴性对齐时进行模态间映射，编码器的权重就会比较大，这样就不能对图像很好的总结。

比较流行的解决这样问题的方法是使用注意力机制[61]，注意力机制使得解码器集中于源实例的子部件中。注意力机制将引导解码器更加关注目标子部件需要被映射的部分，比如图片的区域[62]。

5. 融合

多模态融合是多模态机器学习研究的初始方向，有综述将其分为前端融合、后端融合、以及混合融合[63]。多模态融合是以为了合并多种不同模态的信息来进行预测。自1989年起，多模态融合就是多模态机器学习领域重要的研究内容[64]。

人们对多模态融合的研究兴趣主要来自于三个主要的点。首先，对同一现象获取多种模态的信息可以增强预测的鲁棒性能。第二，使用多模态信息可以对实现信息的互补，在某一模态中缺失的信息可以在其他模态中体现。第三，在某一模态缺失的情况下，多模态系统仍然可以运行，比如在人没有说话的时候辨识人的表情[65]。

多模态融合广泛应用于多模态表情检测[66]，医学图像分析[3]，多媒体行为检测[67]。有一些关于多模态融合的综述文章[68],[63]。这些文章主要综述多模态融合这一人文，比如多媒体分析，信息检索，表情检测。本文旨在综述多模态视觉数据的机器学习方法，以及多模态视觉数据处理面对的挑战。

前面有一些工作用多模态融合来描述所有的多模态算法，我们将融合定义为多模态在预测级的整合，为了获得预测结果的方法。因为深度神经网络的出现，融合和表征的界限变得模糊。

我们将多模态融合分成两类，第一类是与模型无关的方法，也就是不依赖于特定的机器学习方法；第二类是与模型相关的方法，也就是在模型构建过程中明确地使用融合的方法，比如：基于核的方法，图模型，神经网络。

5.1 模型无关的方法

基于模型无关的方法已经有了大量的工作[65]。这些方法可以分为前端融合，末端融合，以及混合融合方法[68]。前端融合通常是将提取出来的特征直接合并，通常是简单的级联起来。末端融合是指每个模态在进行决策之后进行融合，比如分类或者回归任务。混合融合方式是将前端融合以及末端融合的混合结果。与模态无关的融合的优势在于可以使用任意的单模态分类器或者回归方法。

前端融合可以看作是研究人员对于多模态早期融合的尝试，前端融合可以探索每个模态低水平的特征的相关性，使得训练过程相比末端以及混合融合更加容易。

末端融合使用单模态决策结果，然后利用融合机制进行融合，比如平均化[69]，投票机制[70]，对噪声进行加权[71]，信号方差[72]，以及学习到的模型[73]。融合机制使得做决策更加容易，尤其是时出现模态缺失的情况。但是末端融合忽视了模态前端的交互性。

混合融合尝试将前端与末端融合的优势结合在一个框架中，已经成功应用在多媒体事件检测中[74]。

5.2 基于模型的方法

与模型无关的方法很容易使用单模态机器学习的方法实现，之后使用专为多模态设计的技术进行融合。本节我们将描述三种多模态融合的方法：基于核的方法，图模型，以及神经网络。

1) 多核学习

多核学习(Multiple Kernel Learning, MKL)的方法是核支持向量机的扩展(Kernel Support Vector Machine, SVM)，利用不同的核来处理不同的模态[75]。多核学习的方法流行于融合视觉描述特征用于目标检测中[76]，如今被深度学习的方法取代[5]。多核学习的广泛地应用在情感检测[77],[78]，多模态的特征分类[79]。此外Liu等人使用多核学习的方法用于阿尔茨海默病的分类[80]。

除了核选择的灵活性，MKL方法的损失函数式凸函数，可以使用一般的优化算法进行寻优，同时也存在着全局最优解[81]。MKL也可以用于分类或者是回归问题。MKL测试时依赖于支持向量，这样会导致推断的速度较慢，同时需要大量的内存。

1) 图模型

图模型是多模态融合的另外一类方法。本节我们会综述浅层图模型用于多模态融合，以及深层图模型，比如我们在2.1节提到的深度置信网络。

图模型可以分为两种，一种是生成模型(建模联合概率分布)，一种是判别模型(建模条件概率分布)。图模型早期用来进行多模态融合有耦合[82]和因子分解因马尔科夫模型[83]，以及动态贝叶斯网络[84]。

判别模型不考虑建立联合概率，转而建立条件概率密度来获得模型的预测性能。因为预测性能的优异，判别模型如条件随机场(Conditional Random Field, CRF)比生成模型更加流行。条件随机场的具有良好的分割图像和文本的性能[85]。条件随机场发展成为隐条件随机场(Hidden Conditional Random Field)用来建模隐状态[86]，应用在了多模态会议分割中[87]。条件随机场模型还有一些其他的应用，多视角隐条件随机场[88]和隐变量模型[89]。Jiang等人证明多模态隐条件随机场在多媒体分割任务中有着一定的优势[90]。虽然条件随机场的主要任务是分类，有学者将其更改为连续的版本用来做回归任务[91]。

图模型的优势在结构化建模的能力，能够充分利用数据的空间信息和时间信息。同时图模型能够将专家知识融入到模型中，这样使得模型的解释性更强。

2) 神经网络

神经网络目前广泛的应用于多模态融合任务中[11]。神经网络用来融合多源信息进行手势检测[92]，情感分析[93]。浅层的神经网络[94]和深层的神经网络[95],都被研究用来处理多模态融合任务中。

深度神经网络方法最大的优势就是可以从大量数据中进行学习。其次，神经网络允许我们进行端对端的训练和融合。最后，神经网络相比其他非神经网络模型能够学习到复杂的决策边界，这是其它非神经网络模型比较难以处理的。

神经网络最大的缺点在于缺少可解释性，很难能够说明出预测结果的优劣取决于什么，或者说那一个模态对结果去决定性作用。此外，神经网络需要大量的训练样本，以及较多的训练技巧。

6. 协同学习

本文中多模态机器学习面临的最后的一个挑战是协同学习，也就是利用某个样本丰富的模态信息对样本不丰富的模态进行补充。协同学习通常与样本标签有限相关：比如缺少标签数据，输入中含有噪声，标签不可靠。协同学习通常出现在训练环节，而不是出现在测试环节。基于训练数据，我们将协同学习分为三种，平行，非平行，以及混合。平行的方法需要训练样本从一个模态到另一个模态直接连接。也就是对于红外和可见光图像，我们要求数据来自同一场景。非对齐的方法不需要不同模态间直接连接。混合的数据，我们将不同模态通过一个中间模态进行过渡。数据平行关系如图6.1。

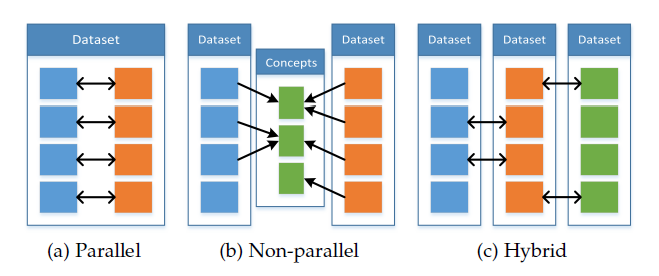


图6.1 数据平行关系

6.1 平行数据

平行数据中进行协同学习，模态都是相同实例的不同模态，比如图像以及对应的文字描述，可见光与对应的红外图像，不同模态的同一患者的脑肿瘤扫描图像。我们将这些算法分为两类，一类是进行协同训练，一类是进行迁移学习。

当我们只有有限的样本数量时，协同训练能够创造出更多的样本[96]。利用基础算法建立在每个模态建立弱分类器，然后利用有标签数据每个模态相互提升无标签数据分类结果。

协同训练目前广泛用于视觉检测器[97]，通过过滤不可靠的样本可以解决模态间的异质性[98]。 虽然协同训练能够产生更多的带有标签的数据，但是也会产生有偏的样本，这样会导致过拟合。

迁移学习是另外一种充分利用协同学习的平行数据的方法，多模态表示学习(见2.1)方法比如多模态深度玻尔兹曼机[6]和多模态自编码器[11]将一个模态的表示迁移到另外一个模态。这样不仅可以获得更好的多模态表示，对比在测试时只有一个模态，多模态的方法比单模态同样有提升。

6.2 非平行数据

依赖于非平行的数据不需要模态有共同的实例，但是需要有着相同的概念。非平行的协同学习方法能够帮助模型有更好的语义理解以及对没有见到的目标进行识别。

迁移学习也可以用于非平行数据，通过使用数据完备性强的模态来对数据质量不好的模态进行迁移。这种迁移的方式也会通过协同多模态表示来实现(如2.1节)。

基础概念中学习语义不仅可以基于语言，也可以基于图像和视频。人类对于语义知识的学习很大程度上取决于感受的信息[99]，许多的概念基于我们的感知系统，而不是仅仅的符号。这样就表明从仅仅从文本信息中进行学习并不是最优的，这也就激发了我们利用其他模态信息的想法。

基础概念在很多任务中有所提升。它表明语言和视觉是互补的信息来源，将它们融合可以提高系统的性能。然而，概念并不是越细致越好，只有当概念与任务是相关的时候[101],[102]，才能起到提升系统性能的作用。

零次学习(Zero Shot Learning, ZSL)是指不从任何观测到的例子获得概念。比如我们对一张猫的图片进行分类，但是我们却没有含有猫标签的图片。这在目标分类任务中是非常重要需要解决的问题，因为很多时候要获得需要目标的训练数据是非常困难的。

有两种零次学习方法，一种是单模态的，一种是多模态的。单模态的零次学习关注与对象的部件中的部分或者属性，比如从已知的视觉属性(颜色，大小，形状)来确定未确定的视觉属性。多模态零次学习是通过对象在可观模态中的辨识来获得不可观模态下对象特性。

6.3 混合数据

在混合数据中，不平行的模态通过一个共同的模态或者是数据集来连接(如图6.1c)。Rajendran等人提出桥连相关神经网络[103]，使用中间模态来学习数据不平行时的多模态表示。除了使用分类的模态进行桥连，也有一些方法使用当前的大规模数据集或者是相关人物来提升系统性能。Socher和Fei-Fei使用大规模的文本预料库来提升图像的分割性能[104]。

7. 总结与展望

多模态机器学习旨在建立能够处理和关联多种模态的信息的模型。本文综述了多模态机器学习的研究进展，并且将当前遇到的挑战分为五类：多模态表示学习，模态间映射，对齐，融合，协同学习。本文旨在对视觉数据的机器学习方法进行综述，但是包含的视觉数据处理的文章还是有限，需要在今后按照这个框架进一步补充。

参考文献

1. Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.
2. Klare, Brendan, and Anil K. Jain. "Heterogeneous face recognition: Matching nir to visible light images." *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010.
3. James A P, Dasarathy B V. Medical image fusion: A survey of the state of the art[J]. Information Fusion, 2014, 19: 4-19.
4. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.
5. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
6. Srivastava N, Salakhutdinov R R. Multimodal learning with deep boltzmann machines[C]//Advances in neural information processing systems. 2012: 2222-2230.
7. Li Y, Wang S, Tian Q, et al. A survey of recent advances in visual feature detection[J]. Neurocomputing, 2015, 149: 736-751.
8. Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.
9. Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model[C]//Advances in neural information processing systems. 2013: 2121-2129.
10. Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis[C]//International Conference on Machine Learning. 2013: 1247-1255.
11. Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 689-696.
12. Ouyang W, Chu X, Wang X. Multi-source deep learning for human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2329-2336.
13. Hong C, Yu J, Wan J, et al. Multimodal deep autoencoder for human pose recovery[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5659-5670.
14. Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.
15. Ding C, Tao D. Robust face recognition via multimodal deep face representation[J]. IEEE Transactions on Multimedia, 2015, 17(11): 2049-2058.
16. Wang D, Cui P, Ou M, et al. Deep Multimodal Hashing with Orthogonal Regularization[C]//IJCAI. 2015, 367: 2291-2297.
17. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 249-256.
18. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
19. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
20. Kinga D, Adam J B. A method for stochastic optimization[C]//International Conference on Learning Representations (ICLR). 2015, 5.
21. Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 693-700.
22. Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.
23. Srivastava N, Salakhutdinov R. Learning representations for multimodal data with deep belief nets[C]//International conference on machine learning workshop. 2012, 79.
24. Srivastava N, Salakhutdinov R R. Multimodal learning with deep boltzmann machines[C]//Advances in neural information processing systems. 2012: 2222-2230.
25. Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model[C]//Advances in neural information processing systems. 2013: 2121-2129.
26. Pan Y, Mei T, Yao T, et al. Jointly modeling embedding and translation to bridge video and language[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4594-4602.
27. Wang J, Zhang T, Sebe N, et al. A survey on learning to hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 769-790.
28. Kumar S, Udupa R. Learning hash functions for cross-view similarity search[C]//IJCAI proceedings-international joint conference on artificial intelligence. 2011, 22(1): 1360.
29. Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 3594-3601.
30. Q. Jiang and W. Li. Deep Cross-Modal Hashing[C]//*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 3270-3278.
31. Li Y, Zhang Y, Huang X, et al. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018 (99): 1-16.
32. Hotelling H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321-377.
33. Klein B, Lev G, Sadeh G, et al. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation[J]. arXiv preprint arXiv:1411.7399, 2014.
34. Gao L, Qi L, Chen E, et al. Discriminative multiple canonical correlation analysis for information fusion[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1951-1965.
35. Lai P L, Fyfe C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems, 2000, 10(05): 365-377.
36. Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis[C]//International Conference on Machine Learning. 2013: 1247-1255.
37. Faridul H S, Pouli T, Chamaret C, et al. A Survey of Color Mapping and its Applications[J]. Eurographics (State of the Art Reports), 2014, 3.
38. Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs[C]//Advances in neural information processing systems. 2011: 1143-1151.
39. Yagcioglu S, Erdem E, Erdem A, et al. A distributed representation based query expansion approach for image captioning[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015, 2: 106-111.
40. Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.
41. Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association of Computational Linguistics, 2014, 2(1): 207-218.
42. Yang Y, Teo C L, Daumé III H, et al. Corpus-guided sentence generation of natural images[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 444-454.
43. Zitnick C L, Parikh D. Bringing semantics into focus using visual abstraction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3009-3016.
44. Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.
45. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
46. Audebert N, Le Saux B, Lefèvre S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks[C]//Asian Conference on Computer Vision. Springer, Cham, 2016: 180-196. Audebert N, Le Saux B, Lefèvre S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks[C]//Asian Conference on Computer Vision. Springer, Cham, 2016: 180-196.
47. Menze B H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS)[J]. IEEE transactions on medical imaging, 2015, 34(10): 1993.
48. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
49. Rohrbach A, Rohrbach M, Schiele B. The long-short story of movie description[C]//German Conference on Pattern Recognition. Springer, Cham, 2015: 209-221.
50. Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint arXiv:1605.05396, 2016.
51. Huang X, Liu M Y, Belongie S, et al. Multimodal Unsupervised Image-to-Image Translation[J]. arXiv preprint arXiv:1804.04732, 2018.
52. Devlin J, Cheng H, Fang H, et al. Language models for image captioning: The quirks and what works[J]. arXiv preprint arXiv:1505.01809, 2015.
53. Malmaud J, Huang J, Rathod V, et al. What's cookin'? interpreting cooking videos using text, speech and vision[J]. arXiv preprint arXiv:1503.01558, 2015.
54. Karpathy A, Joulin A, Fei-Fei L F. Deep fragment embeddings for bidirectional image sentence mapping[C]//Advances in neural information processing systems. 2014: 1889-1897.
55. Müller M. Dynamic time warping[J]. Information retrieval for music and motion, 2007: 69-84.
56. Tapaswi M, Bäuml M, Stiefelhagen R. Aligning plot synopses to videos for story-based retrieval[J]. International Journal of Multimedia Information Retrieval, 2015, 4(1): 3-16.
57. Yu C, Ballard D H. On the integration of grounding language and learning objects[C]//AAAI. 2004, 4: 488-493.
58. Naim I, Song Y C, Liu Q, et al. Unsupervised Alignment of Natural Language Instructions with Video Segments[C]//AAAI. 2014: 1558-1564.
59. Bojanowski P, Lajugie R, Bach F, et al. Weakly supervised action labeling in videos under ordering constraints[C]//European Conference on Computer Vision. Springer, Cham, 2014: 628-643.
60. Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 11-20.
61. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
62. Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International conference on machine learning. 2015: 2048-2057.
63. Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE Signal Processing Magazine, 2017, 34(6): 96-108.
64. Yuhas B P, Goldstein M H, Sejnowski T J. Integration of acoustic and visual speech signals using neural networks[J]. IEEE Communications Magazine, 1989, 27(11): 65-71.
65. D'mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys (CSUR), 2015, 47(3): 43.
66. Soleymani M, Pantic M, Pun T. Multimodal emotion recognition in response to videos[J]. IEEE transactions on affective computing, 2012, 3(2): 211-223.
67. Lan Z Z, Bao L, Yu S I, et al. Multimedia classification and event detection using double fusion[J]. Multimedia tools and applications, 2014, 71(1): 333-347.
68. Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia systems, 2010, 16(6): 345-379.
69. Shutova E, Kiela D, Maillard J. Black holes and white rabbits: Metaphor identification with visual features[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 160-170.
70. Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion[C]//Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin, Heidelberg, 2014: 153-162.
71. Evangelopoulos G, Zlatintsi A, Potamianos A, et al. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention[J]. IEEE Transactions on Multimedia, 2013, 15(7): 1553-1568.
72. Glodek M, Tschechne S, Layher G, et al. Multiple classifier systems for the classification of audio-visual emotional states[M]//Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg, 2011: 359-368.
73. Ramirez G A, Baltrušaitis T, Morency L P. Modeling latent discriminative dynamic of multi-dimensional affective signals[M]//Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg, 2011: 396-406.
74. Lan Z Z, Bao L, Yu S I, et al. Multimedia classification and event detection using double fusion[J]. Multimedia tools and applications, 2014, 71(1): 333-347.
75. Liu F, Zhou L, Shen C, et al. Multiple kernel learning in the primal for multimodal Alzheimer's disease classification[J]. IEEE J. Biomedical and Health Informatics, 2014, 18(3): 984-990.
76. Bucak S S, Jin R, Jain A K. Multiple kernel learning for visual object recognition: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1354-1369.
77. Castellano G, Kessous L, Caridakis G. Emotion recognition through multiple modalities: face, body gesture, speech[M]//Affect and emotion in human-computer interaction. Springer, Berlin, Heidelberg, 2008: 92-103.
78. Jaques N, Taylor S, Sano A, et al. Multi-task, multi-kernel learning for estimating individual wellbeing[C]//Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec. 2015, 898.
79. Niazmardi S, Safari A, Homayouni S. A novel multiple kernel learning framework for multiple feature classification[J]. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 2017, 10: 3734-3743.
80. Liu F, Zhou L, Shen C, et al. Multiple kernel learning in the primal for multimodal Alzheimer's disease classification[J]. IEEE J. Biomedical and Health Informatics, 2014, 18(3): 984-990.
81. Gönen M, Alpaydın E. Multiple kernel learning algorithms[J]. Journal of machine learning research, 2011, 12(Jul): 2211-2268.
82. Nefian A V, Liang L, Pi X, et al. A coupled HMM for audio-visual speech recognition[C]//Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. IEEE, 2002, 2: II-2013-II-2016.
83. Ghahramani Z, Jordan M I. Factorial hidden Markov models[C]//Advances in Neural Information Processing Systems. 1996: 472-478.
84. Garg A, Pavlovic V, Rehg J M. Boosted learning in dynamic Bayesian networks for multimodal speaker detection[J]. Proceedings of the IEEE, 2003, 91(9): 1355-1369.
85. Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
86. Quattoni A, Wang S, Morency L P, et al. Hidden conditional random fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(10).
87. Wang S B, Quattoni A, Morency L P, et al. Hidden conditional random fields for gesture recognition[C]//null. IEEE, 2006: 1521-1527.
88. Song Y, Morency L P, Davis R. Multimodal human behavior analysis: learning correlation and interaction across modalities[C]//Proceedings of the 14th ACM international conference on Multimodal interaction. ACM, 2012: 27-30.
89. Song Y, Morency L P, Davis R. Multi-view latent variable discriminative models for action recognition[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2120-2127.
90. Jiang X, Wu F, Zhang Y, et al. The classification of multi-modal data with hidden conditional random field[J]. Pattern Recognition Letters, 2015, 51: 63-69.
91. Qin T, Liu T Y, Zhang X D, et al. Global ranking using continuous conditional random fields[C]//Advances in neural information processing systems. 2009: 1281-1288.
92. Neverova N, Wolf C, Taylor G, et al. Moddrop: adaptive multi-modal gesture recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(8): 1692-1706.
93. Kahou S E, Bouthillier X, Lamblin P, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video[J]. Journal on Multimodal User Interfaces, 2016, 10(2): 99-111.
94. Gao H, Mao J, Zhou J, et al. Are you talking to a machine? dataset and methods for multilingual image question[C]//Advances in neural information processing systems. 2015: 2296-2304.
95. Nojavanasghari B, Gopinath D, Koushik J, et al. Deep multimodal fusion for persuasiveness prediction[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016: 284-288.
96. Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.
97. Levin A, Viola P A, Freund Y. Unsupervised Improvement of Visual Detectors using Co-Training[C]//ICCV. 2003, 1: 626-633.
98. Christoudias C, Urtasun R, Darrell T. Multi-view learning in the presence of view disagreement[J]. arXiv preprint arXiv:1206.3242, 2012.
99. Louwerse M M. Symbol interdependency in symbolic and embodied cognition[J]. Topics in Cognitive Science, 2011, 3(2): 273-302.
100. Barsalou L W. Grounded cognition[J]. Annual review of psychology., 2008, 59: 617-645.
101. Kiela D, Bulat L, Clark S. Grounding semantics in olfactory perception[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015, 2: 231-236.
102. Kiela D, Clark S. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 2461-2470.
103. Rajendran J, Khapra M M, Chandar S, et al. Bridge correlational neural networks for multilingual multimodal representation learning[J]. arXiv preprint arXiv:1510.03519, 2015.
104. Socher R, Fei-Fei L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 966-973.