**An Attentive Survey of Attention Models**

摘要

注意模型已成为神经网络中的一个重要概念，并在不同的应用领域得到了广泛的研究。这项调查提供了一个结构化和全面的发展，在建模注意。特别是，我们提出了一种分类法，将现有的技术分成连贯的类别。我们回顾了注意力被纳入的不同神经结构，并展示了注意力如何提高神经模型的可解释性。最后，我们讨论了一些应用程序，其中建模关注具有重大影响。我们希望这项调查能为注意力模型提供一个简洁的介绍，并在开发应用方法的同时指导实践者。

1介绍

注意模型（AM），最初是为机器翻译引入的【Bahdanau等人，2014年】，现在已经成为神经网络文献中的一个主要概念。人工智能（AI）作为神经结构的重要组成部分，在自然语言处理、统计学习、语音和计算机视觉等领域有着广泛的应用。

注意力背后的直觉可以用人类的生物系统得到最好的解释。例如，我们的视觉处理系统倾向于有选择地将注意力集中在图像的某些部分，而忽略其他不相关的信息，从而有助于感知[Xu等人，2015]。同样，在涉及语言、言语或视觉的几个问题中，输入的某些部分可能比其他部分更相关。例如，在翻译和摘要任务中，只有输入序列中的某些单词可能与预测下一个单词相关。同样，在图像字幕问题中，输入图像的某些区域可能更适合于在字幕中生成下一个单词。AM通过允许模型动态地关注输入中有助于有效执行手头任务的某些部分，从而融入了关联的概念。使用AM对Yelp评论进行情绪分类的示例如图1所示【Yang等人，2016年】。在这个例子中，AM了解到在五个句子中，第一个和第三个句子更相关。

此外，这些句子中的“美味”和“令人惊奇”两个词更有意义地决定了评论的情绪。

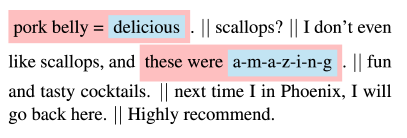


图1:Yelp评论中情感分类的注意建模示例。数据来源【Yang等人，2016年】

神经网络建模关注度的迅速提高主要有三个原因。首先，这些模型现在已经成为机器翻译、问答、情感分析、词性标注、选区解析和对话系统等多项任务的最新技术[Young等人，2018年]。其次，除了在主要任务上提高性能之外，它们还提供了其他一些优势。它们被广泛用于提高神经网络的可解释性，而神经网络又被认为是黑箱模型。这是一个显著的好处，主要是因为人们对影响人类生活的应用程序中机器学习模型的公平性、问责制和透明度越来越感兴趣。第三，它们有助于克服递归神经网络（RNN）的一些挑战，例如随着输入长度的增加性能下降（第2节）和输入的顺序处理导致的计算效率低下（第4.3节）。因此，在这项工作中，我们旨在提供一个简短，但全面的调查注意建模。

组织：我们简要解释了Bahdanau等人提出的AM。[2014]在第2节中，在第3节中描述我们的分类。然后，我们分别在第4节和第5节中讨论了使用AM的关键神经结构和注意力如何促进神经网络的可解释性。最后，我们在第6节中介绍了注意力广泛应用的应用，并在第7节中总结了本文。

相关调查：有一些关注计算机视觉的调查【Wang和Tax，2016年】和图表【Lee等人，2018年】。另一个类似的工作是由Galassi等人。[2019]，但我们进一步整合了可访问的分类法、关键架构和应用程序，以及AM的可解释性方面。我们希望，我们的贡献不仅将促进对AM的更广泛理解，还将帮助人工智能开发人员和工程师为他们的应用领域确定正确的方法。

**2 Attention Model**

序列到序列模型包括一个编码器架构[Cho等人，2014b]，如图2（A）所示。编码器是一个RNN，它接受标记的输入序列，其中T是输入序列的长度，并将其编码为固定长度向量。解码器也是一个RNN，它将单个固定长度向量hT作为其输入，并生成输出序列逐个标记，其中T0是输出序列的长度。在每个位置t，ht和st分别表示编码器和解码器的隐藏状态。

传统编码器-解码器的挑战：这种传统编码器-解码器框架有两个众所周知的挑战。首先，编码器必须将所有的输入信息压缩成一个固定长度的向量hT，然后传递给解码器。使用单个固定长度向量压缩长而详细的输入序列可能导致信息丢失[Cho等人，2014a]。其次，它无法对输入和输出序列之间的对齐进行建模，这是结构化输出任务（如翻译或摘要）的一个重要方面[Young等人，2018年]。直观地说，按顺序到顺序的任务，每个输出标记都会受到输入序列某些特定部分的影响。然而，解码器缺乏在生成每个输出令牌时有选择地关注相关输入令牌的任何机制。

关键思想：AM的目的是通过允许解码器访问整个编码的输入序列来减轻这些挑战。中心思想是诱导输入序列上的注意权重，以对存在相关信息的位置集进行优先排序，以生成下一个输出令牌。

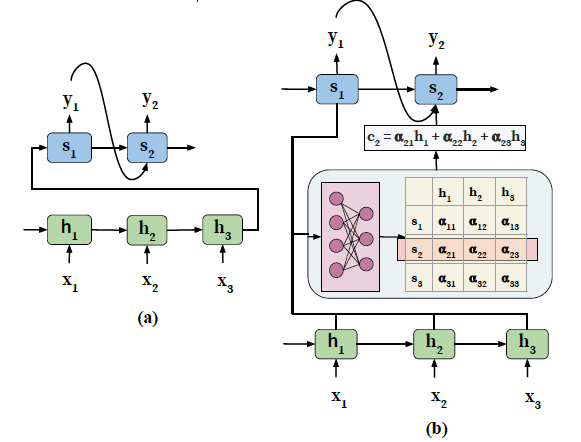


图2：编码器-解码器架构：（a）传统（b）带注意模型

注意事项的用法：注意事项对应的编解码架构如图2（b）所示。架构中的注意块负责自动学习注意权重，它捕获（编码器隐藏状态，我们称之为候选状态）和（解码器隐藏状态，我们称之为查询状态）之间的相关性。然后，这些注意权重用于构建上下文向量c，该上下文向量c作为输入传递给解码器。在每个解码位置j，上下文向量cj是编码器的所有隐藏状态及其对应的注意权重的加权和，即。这个额外的上下文向量是解码器访问整个输入序列并关注输入序列中相关位置的机制。

学习注意权重：注意权重是通过在体系结构中加入一个额外的前馈神经网络来学习的。该前馈网络学习特定的注意权重作为两个状态hi（候选状态）和（查询状态）的函数，这两个状态被神经网络作为输入。此外，该前馈网络与该体系结构的编码器-解码器组件联合训练。

**3注意分类**

我们将注意力分为四大类，如表1所示，阐明每一类中的不同注意类型。我们要强调，这些类别并非相互排斥。注意可以作为多个类别的组合来使用，例如，杨等人使用了多层次、自我和软注意的组合[2016年]。因此，人们可以把这些范畴看作是一个维度，在运用注意力进行感兴趣的应用时，可以沿着这个维度来考虑注意力。为了使这一概念易于理解，我们提供了一份关键技术论文清单，并在表2中详细说明了拟议方法中使用的多种注意类型。

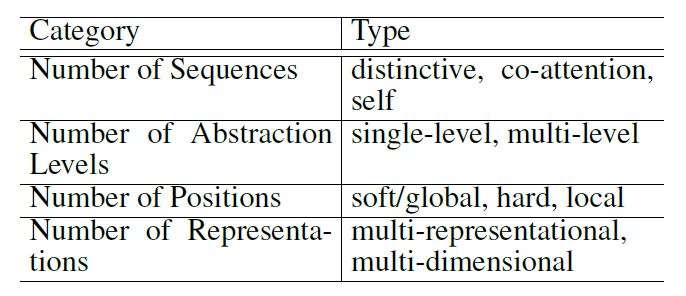


表1：每个类别内的注意类别和类型。

**3.1 Number of sequences**

到目前为止，我们只考虑了涉及单个输入和相应输出序列的情况。当候选状态和查询状态分别属于两个不同的输入和输出序列时，使用这种我们称为独特distinctive的注意。大多数用于翻译的注意模型【Bahdanau等人，2014年】、摘要【Rush等人，2015年】、图像字幕【Xu等人，2015年】和语音识别【Chan等人，2016年】属于独特的注意类型。

一个共同注意模型同时处理多个输入序列，共同学习它们的注意权重，以捕捉这些输入之间的相互作用。Lu等人。【2016】采用共同注意模型进行视觉问答。作者认为，除了在输入图像上建立视觉注意模型外，建立问题注意模型也很重要，因为问题文本中的所有单词对问题的答案并不同样重要。此外，基于注意的图像表示用于引导问题注意，反之亦然，这本质上有助于同时检测问题中的关键短语和与答案相关的图像的相应区域。

相反，对于文本分类和推荐等任务，输入是一个序列，而输出不是一个序列。在这个场景中，注意可以用于学习相同输入序列中每个令牌的输入序列中的相关令牌。换句话说，对于这类注意，查询和候选状态属于同一序列。为此，杨等人提出了自我注意，又称内注意。[2016年]。

**3.2抽象层数量**

在最一般的情况下，注意力权重只针对原始输入序列计算。这种类型的注意可以称为single-level。另一方面，注意力可以按顺序应用于输入序列的多个抽象层次。较低抽象级别的输出（上下文向量）成为较高抽象级别的查询状态。此外，基于权值是自上而下学习的[Zhao和Zhang，2018年]（从较高抽象层次到较低层次）还是自下而上学习的[Yang等人，2016年]，使用多层次注意的模型可以进一步分类。

我们举例说明了这一类别中的一个关键示例，该示例在两个不同的抽象层次（即单词级和句子级）使用注意模型进行文档分类任务[Yang等人，2016]。这个模型被称为“层次注意模型”（HAM），因为它捕捉了文档的自然层次结构，即文档由句子组成，句子由单词组成。多层次注意允许HAM提取句子中重要的单词和文档中重要的句子，如下所示。首先建立了一种基于注意的句子表示方法，并将一级注意应用于嵌入向量序列。然后，它使用第二级注意来聚合这些句子表示形式，以形成文档的表示形式。文档的最终表示形式用作分类任务的特征向量。

请注意，第3.1节中描述的共同注意工作[Lu等人，2016]也属于多层次类别，它在三个层次上共同注意图像和问题：单词层次、短语层次和问题层次。图3描述了共同注意和多层次注意的结合。

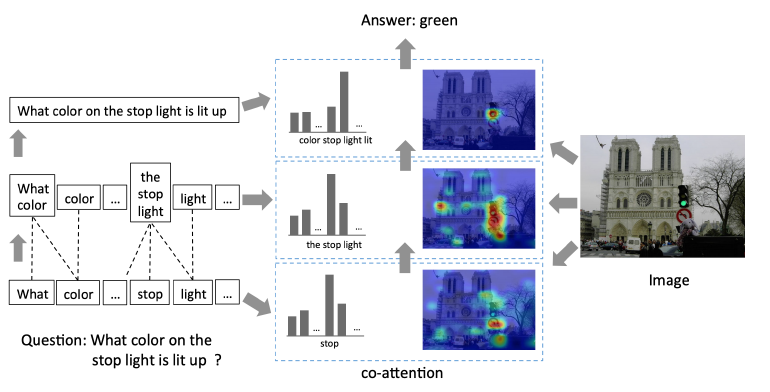


图3:Lu等人提出的AM。[2016]针对视觉问答任务，该任务是共同关注（视觉和文本）和多层次（单词级、短语级和问题级）关注的结合。

Zhao和Zhang[2018]建议使用“注意via注意”，也使用多层次注意（字符在较低层次，单词在较高层次），但以自上而下的方式学习注意权重。

**3.3 Number of positions**

在第三类中，差异来自于计算注意功能的输入序列的位置。Bahdanau等人引入的注意。[2014]也被称为软关注。顾名思义，它使用输入序列所有隐藏状态的加权平均值来构建上下文向量。软权值方法的使用使得神经网络能够通过反向传播进行有效的学习，但也会导致二次计算代价。

Xu等人。[2015]提出了一个硬注意模型，其中上下文向量是根据输入序列中随机采样的隐藏状态计算的。这是通过注意权重参数化的多努利分布来实现的。硬注意模型有利于降低计算成本，但在输入的每个位置进行硬决策，使得得到的框架不可微，难以优化。因此，为了克服这一局限性，文献中提出了变分学习方法和策略梯度方法。

Luong等人[2015]在机器翻译任务的背景下，提出了两种注意模型，即局部注意模型和全局注意模型。全局注意模型类似于软注意模型。另一方面，局部注意模型介于软注意和硬注意之间。其关键思想是首先在输入序列中检测出一个注意点或位置，然后在该位置周围选取一个窗口来创建一个局部软注意模型。输入序列中的位置可以设置（单调对齐）或由预测函数学习（预测对齐）。因此，局部注意的优点是在软注意和硬注意、计算效率和窗口内可微性之间提供一个参数权衡。

**3.4 Number of representations**

通常，在大多数应用中使用输入序列的单一特征表示。但是，在某些情况下，使用输入的一个特征表示可能不足以完成下游任务。在这种情况下，一种方法是通过多个特征表示捕获输入的不同方面。注意可以用来给这些不同的表示分配重要性权重，这些表示可以确定最相关的方面，而忽略输入中的噪声和冗余。我们将此模型称为多表示AM，因为它可以确定下游应用程序输入的多个表示的相关性。最终表示是这些多重表示及其注意权重的加权组合。这里要注意的一个好处是，通过检查权重，直接评估哪些嵌入是特定下游任务的首选。

基拉等人。[2018]学习同一输入句子的不同单词嵌入的注意权重，以改进句子表示。同样，Maharjan等人。[2018]利用注意力动态权衡不同特征的图书表现，捕捉词汇、句法、视觉和体裁信息。

基于相似直觉，在多维注意中，引入权重来确定输入嵌入向量各维度的相关性。直觉是，计算向量的每个特征的分数可以选择在任何给定上下文中最能描述令牌特定含义的特征。这对于单词嵌入遇到多义问题的自然语言应用特别有用。这种方法的示例如Lin等人所示。[2017]为了更有效的句子嵌入表征，在Shen等人中。【2018】语言理解问题。

**4 Network Architectures with Attention**

在这一节中，我们描述了三种与注意结合使用的显著神经结构：（1）编码器-解码器框架，（2）将注意力扩展到单个输入序列之外的存储网络，以及（3）利用注意绕过递归模型的序列处理组件的结构。

**4.1 Encoder-Decoder**

最早使用注意力是作为基于RNN的编码器-解码器框架的一部分来编码长输入语句[Bahdanau等人，2014年]。因此，这种体系结构得到了最广泛的应用。

一个有趣的事实是，AM可以接受任何输入表示，并将其减少为在解码步骤中使用的单个固定长度上下文向量。因此，它允许将输入表示与输出分离。人们可以利用这一优势引入混合编码器-解码器，最流行的是卷积神经网络（CNN）作为编码器，RNN或长短期存储器（LSTM）作为解码器。这种体系结构特别适用于许多多模式任务，如图像和视频字幕、可视问答和语音识别。

然而，并非所有输入和输出都是连续数据的问题都可以用上述公式解决（例如排序或旅行推销员问题）。

指针网络【Vinyals等人，2015年】是另一类具有以下两个差异的神经模型，（1）输出是离散的，指向输入序列中的位置（因此称为指针网络），以及（2）输出每一步的目标类的数量取决于输入的长度（因此是可变的）。这不能通过使用传统的编码器-解码器框架来实现，因为输出字典是先验的（例如，在自然语言建模的情况下）。作者利用注意权值来模拟在每个输出位置选择第i个输入符号作为所选符号的概率。该方法可应用于离散优化问题，如旅行售货员问题和排序问题。

**4.2 Memory Networks**

像问答和聊天机器人这样的应用程序需要从事实数据库中的信息中学习的能力。网络的输入是一个知识数据库和一个查询，其中一些事实与查询的关系比其他事实更密切。

端到端内存网络[Sukhbatar等人，2015]通过使用一组内存块来存储事实数据库，并使用注意在内存中每个事实的模型相关性来回答查询来实现这一点。使用注意力还提供了计算优势，使目标连续，并通过反向传播实现端到端的训练。端到端的记忆网络可以被认为是AM的一个推广，在AM中，他们不只是在单个序列上建模注意力，而是在一个大型序列数据库（facts）上建模注意力。

**4.3 Networks without RNNs**

递归架构依赖于编码步骤输入的顺序处理，这导致计算效率低下，因为处理无法并行化[Vaswani等人，2017年]。为了解决这个问题，作者提出了一种由编码器和解码器组成的Transformer结构具有两个子层的相同层的堆栈：位置前馈网络（FFN）层和多头自关注层。

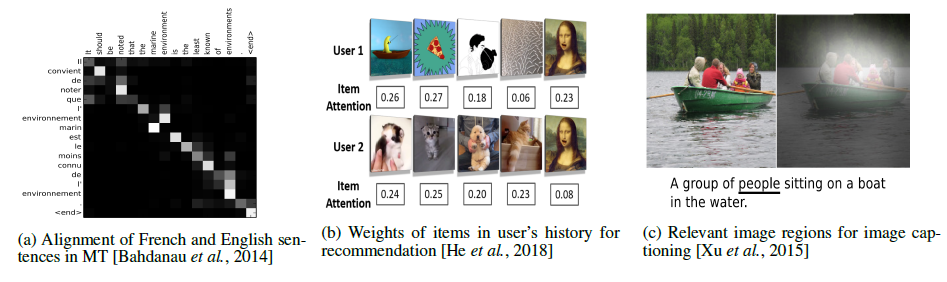


图4：注意力权重可视化的例子。

位置相关的FFN：输入是顺序的，它要求模型利用输入的时间方面，但是捕捉这个位置信息的组件（即RNN/CNN）未使用。为了解释这一点，转换器中的编码器相位使用按位置的FFN为输入序列的每个令牌生成内容嵌入和位置编码。

多头自我注意：在每个子层中使用自我注意来关联标记及其在相同输入序列中的位置。此外，注意力被称为多头（multi-head），因为多个注意力层是平行堆叠的，具有相同输入的不同线性变换。这有助于模型捕获输入的各个方面，并提高其表达能力。

Transformer结构实现了显著的并行处理、更短的训练时间和更高的翻译精度，而不需要任何递归组件，这是一个显著的优点。然而，位置编码只弱地融合了位置信息，可能不适用于对位置变化更敏感的问题。

Shen等人[2018]使用时间卷积和Transformer的自我注意机制来编码位置信息。此外，还有更直接的方法来中断输入的顺序处理。Raffel和Ellis[2015]提出了前馈注意模型，他们使用AM来折叠数据的时间维度，并使用FFN而不是RNN来解决序列数据问题。在这种情况下，AM被用来从可变长度的输入序列中产生一个固定长度的上下文向量，它可以作为FFN的输入馈送。

**5 Attention for Interpretability**

人们对人工智能模型的可解释性非常感兴趣，这是由模型的性能、透明度和公平性所决定的。然而，神经网络，特别是深度学习架构，因其缺乏可解释性而受到批评[Guidotti等人，2018年]。

从可解释性的角度来看，建模关注特别有趣，因为它允许我们直接检查深度学习架构的内部工作。该假设认为，对于序列中每个位置的输出预测，注意权重的大小与特定输入区域的相关性密切相关。这可以通过可视化一组输入和输出对的注意权重来轻松实现。Li等人[2016]将注意力作为解释神经模型内部工作的重要途径之一。

如图4（a）所示，Bahdanau等人[2014]将注意力权重可视化，它清楚地显示法语和英语中句子的自动对齐，尽管不同语言的subjectverbnoun位置不同。特别地，注意模型通过将环境分子与海洋环境正确地对齐来显示非单调对齐。图4（b）显示注意权重有助于识别用户的兴趣。用户1似乎更喜欢“卡通”视频，而用户2更喜欢“动物”视频[他等人，2018年]。最后，徐等人。[2015]提供对图像字幕任务中生成的文本有重大影响的相关图像区域（即具有高注意权重）的大量可视化列表（如图4（c）所示）。

我们还总结了其他一些有趣的发现如下。De Arteaga等人。[2019]探讨职业分类中的性别偏见，并展示在分类任务中受到更多关注的词汇是如何经常被性别化的。Yang等人。[2016]请注意，好的和坏的词的重要性取决于决定评论情绪的语境。作者考察了这些词的注意权重分布，发现它们的分布范围从0到1，这意味着模型捕获了不同的上下文，并为这些词分配了上下文相关的权重。Chan等人。[2016]需要注意的是，在语音识别中，字符输出和音频信号之间的注意能够正确识别音频信号中第一个字符的起始位置，对于声学相似的单词，注意权重是相似的。最后，Kiela等人。[2018]发现多表征注意对GloVe、快速文本词嵌入，特别是低频词手套赋予了更高的权重。

作为注意力的另一个有趣应用，Lee等人[2017]和Liu等人[2018]提供了一个可视化深层神经网络注意力权重的工具。其目的是解释和扰动注意权重，以便人们可以模拟假设情景，并以交互方式观察预测的变化。

**6 Applications**

注意模型因其直观性、通用性和可解释性而成为研究的热点。注意模型的变体被用来处理不同应用领域集合的独特特征，例如摘要、阅读理解、语言建模、句法分析等。我们讨论了注意建模在三个应用领域的应用：（i）自然语言生成（NLG），（ii）分类和（iii）推荐系统。

NLG任务包括生成自然语言文本作为输出。一些NLG应用程序从合并AM中受益，包括机器翻译（MT）、问答（QA）和多媒体描述（MD）。

机器翻译使用算法将文本或语音从一种语言翻译到另一种语言。针对机器翻译的神经技术中的注意力建模可以更好地对齐不同语言中的句子，这是机器翻译中的一个关键问题。在翻译较长的句子时，注意力模型的优势也变得更加明显【Bahdanau等人，2014年】。包括[Britz et al.，2017]和[Tang et al.，2018]在内的一些研究表明，使用注意力可以提高机器翻译的绩效。

QA问题利用注意力（i）通过关注问题的相关部分来更好地理解问题【Hermann等人，2015年】，（ii）使用内存网络存储大量信息以帮助寻找答案【Sukhbatar等人，2015年】，以及（iii）通过使用共同注意对输入中的多模态进行建模，提高视觉质量保证任务的性能【Lu等人，2016年】。

MD的任务是生成多媒体输入序列（可以是语音、图像和视频）的自然语言文本描述[Cho等人，2015]。与QA类似，这里attention执行在语音输入[Chorowski et al.，2015]或输入图像的相关部分[Xu et al.，2015]中查找相关声学信号的功能，以预测标题中的下一个单词。此外，Li等人。【2017】利用视频的时空结构，将多层次注意力用于视频字幕任务。较低抽象层在帧内提取特定区域，较高抽象层有选择地关注帧的小子集。

文档分类：如第3节前面所述，分类问题主要利用自我关注来构建更有效的文档表示。Yang等人[2016]使用多层次的自我关注，而Lin等人[2017]提出了一个多维和基拉等。【2018】提出一个多表征的自我注意模型。

情绪分析：同样，在情绪分析任务中，自我关注有助于集中在对确定输入情绪重要的单词上。wang等人提出的基于方面的情感分类方法。[2016]和Ma等人。[2018]将方面相关概念的额外知识纳入模型，并利用注意力将概念与内容本身进行适当权衡。情绪分析应用程序还使用了多种架构，如存储网络[Tang等人，2016]和变压器[Ambartsoumian和Popowich，2018；Song等人，2019]。

推荐系统：AMs也被广泛应用于推荐系统中，用于用户分析，即为用户的交互项分配注意权重，以更有效的方式获取长期和短期的兴趣。这是直观的，因为用户的所有交互都与项目的推荐无关，用户的兴趣是短暂的，并且在长时间和短时间内都会发生变化。

多篇论文使用自我关注机制来寻找用户历史上最相关的项目，以改进项目推荐，要么使用协作过滤框架【He等人，2018年；Shuai Yu，2019年】，要么使用编码器架构来进行顺序推荐【Kang和McAuley，2018年；Zhou等人，2018年】。

近年来，人们以新颖的方式关注这一问题，为研究开辟了新的途径。一些有趣的方向包括外部知识库的平滑合并、预训练嵌入和多任务学习、无监督的表征学习、稀疏学习和典型学习，即样本选择。

**7 Conclusion**

在本次调查中，我们讨论了文献中注意力形成的不同方式，并试图通过讨论注意力的分类、使用注意力的关键神经网络架构和已看到显著影响的应用领域，提供各种技术的概述。我们讨论了注意力在神经网络中的结合如何带来显著的性能提高，通过促进可解释性提供了对神经网络内部工作的更深入了解，并且通过消除输入的顺序处理提高了计算效率。我们希望这项调查将有助于更好地了解关于这一主题的研究的不同方向，以及如何将在一个领域开发的技术应用于其他领域。