第1周汇报

——张溢炉

1. 《Deep learning》总结
2. 主要内容

本文对深度学习进行了概述，依次介绍了监督学习、用来训练多层网络的反向传播、卷积神经网络及其在图像处理的应用、分布式表示和语言处理、循环神经网络、深度学习的发展方向。

1. 简单总结
2. 深度学习的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。深度学习的优缺点如下：

* 优点：相较于传统机器学习，深度学习不需要工程师或专家手动提取特征，是端对端的学习方法，应用比较简便。随着算力、数据的驱动以及算法的优化，深度学习的优势会越来越明显。
* 缺点：计算量大；需要的数据集大；其为黑箱模型，内部原理复杂。

1. 监督学习：利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程。该过程是根据预测输出和实际输出构建目标函数，然后利用相关算法（如梯度下降、牛顿方法等）使目标函数最大化或最小化来调节模型内参。从文中获得的要点如下：

* 使用传统的机器学习很容易受无关因素干扰，也较难分辨细微差别。而深度学习能较好的解决这些问题。
* 常用的目标函数是预测输出与实际输出的方差。
* 监督学习中常用的调参方法是随机梯度下降法。

1. 常用来训练多层网络的反向传播是随机梯度下降法，它是链式求导的实际应用。GPU使得随机梯度下降在深度学习中应用方便，运算速度更快。实践也证明，深度学习很少陷入一些专家担心的局部最小和鞍点问题。
2. 卷积神经网络常用于处理多维数组形式的数据，如图片等。卷积神经网络的前几层一般由卷积层和池化层组成。其中，卷积层的作用是检测来自前一层的特征的局部连接，池化层的作用是将语义相似的特征合并为一个。另外，卷积神经网络在计算机视觉中取得很好的效果。很多公司也在研发卷积神经网络芯片来提高图像处理的实时性，使其能应用于智能手机、相机和自动驾驶中。
3. 深度学习理论表明，与不使用分布式表示的经典学习算法相比，深度网具有两种不同的指数优势。首先，学习分布式表示可以推广到新的学习特征值的组合超出训练期间所见的值。 其次，在深层网络中组合表示层带来了另一种潜在指数优势。在神经网络中，上下文的每个单词常用n维向量来表示，词向量已被广泛应用于自然语言处理。
4. 循环神经网络常用于处理序列数据，如语音、语言等。其缺点是不具备长时记忆功能。因而，有人提出了长短期记忆网络、神经图灵机和记忆网络。
5. 深度学习未来的几个可能发展方向是无监督学习、深度强化学习和自然语言处理。
6. 其他

经过看书、找资料和老师的指导，对自然语言处理有了初步了解，总结如下：

1. 转变做工程的思维，要把心思投入到科研中；
2. 先看自然语言处理方面概述性文章，找到感兴趣方向后，再重点看感兴趣方向文章，注意方向顶会和牛人的文章，下表能较好的展示自然语言处理的几个方向；另外，常见NLP任务有：机器翻译、问答、对话系统、词性标注、句法分析、命名实体识别、情感分析、上下文嵌入、信息检索、信息抽取、知识工程、自然语言生成、推荐系统等；



1. 以英文论文为中心展开学习，弄懂论文中不理解的地方，也可以找有代码的论文，复现论文模型，加深理解；另外，看论文时总结别人的写作方法，提高自己的写作能力，为以后出论文做准备。

第2周汇报

——张溢炉

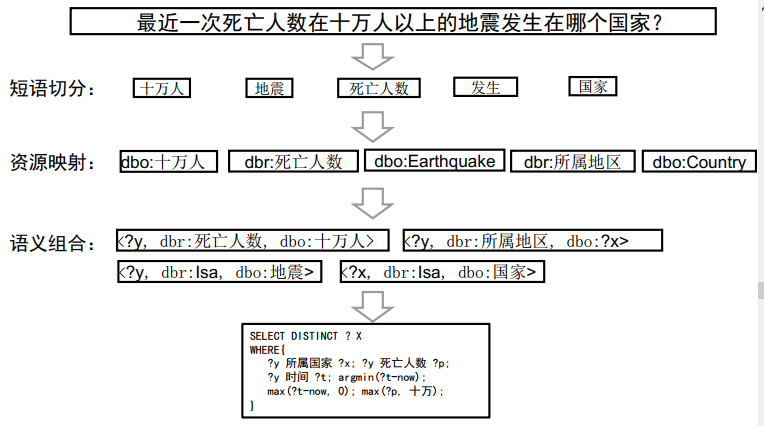
1. 《深度问答技术》总结
2. 内容概述

本文是第三届中文知识图谱研讨会的会议记录，会议的主题是深度问答技术。会议的主要内容有深度问答中的知识表示、关键技术、资源整合和对未来的展望。

1. 总结及理解
2. 涉及人工智能难点问题：自然语言理解、知识表示、知识推理。
3. 发展历程：图灵测试、专家系统、问答式检索、基于知识图谱的深度问答、社区问答和阅读理解式问答。知识图谱本质上是语义网络的知识库。从实际应用的角度出发其实可以简单地把知识图谱理解成多关系图。由节点（Vertex）和边（Edge）构成，一般包含多种类型的节点和多种类型的边。
4. 答案的来源：结构化知识（知识库、抽取事实）、非结构化知识（文本）。结构化知识一般只需要简单预处理即可以作为后续AI系统的输入，但非结构化知识一般需要借助于自然语言处理等技术来提取出结构化信息。
5. 知识表示：符号表示（one-hot）、分布表示。分布表示能将自然语言表示为计算机擅长处理的数值计算。
6. 基于符号语义的深度问答：基于复述的语义匹配（事实库问答）、基于句法树的语义匹配（文本库问答）、基于篇章语义图的语义演算（篇章阅读理解）。
7. 事实问答步骤：问句复述（将问句以计算机易于理解的方式复述）->问句解析（将问句映射到事实库）->查询转换（事实的不同表示）->执行查询，回答问题。
8. 符号表示：准确度好，覆盖度低，泛化能力弱，不方便扩展，推理效率低，用户可理解，流水线方式；

分布表示：准确度弱，覆盖度好，泛化能力强，容易扩展和融合到其他系统，效率相对高，不可解释，端到端方式。

1. 关键技术：

* 知识抽取：包括实体识别、关系识别、实体链接、事件识别。通过基于卷积神经网络的实体关系分类方法可实现抽取实体间语义关系的文本表示。
* 问句语义解析：通过短语切分、资源映射和语义组合可实现自然语言问句到结构化查询语句的映射，如下图所示。规则和模板驱动的传统方法无法处理大规模链接数据上的真实问句。歧义严重是面临的难点，传统pipeline策略可逐步解决每一步的歧义问题，但会造成错误的积累与传递。可尝试采用联合提取的方法解决歧义问题，如基于马尔科夫逻辑网络的联合消歧。联合消歧巧妙运用了语句的潜在联系。
* 推理规则学习：通过推理规则学习可对未知知识进行补全，如基于高斯分布表示学习的补全方法、基于动态映射矩阵的补全方法（可表示知识的多样性）。

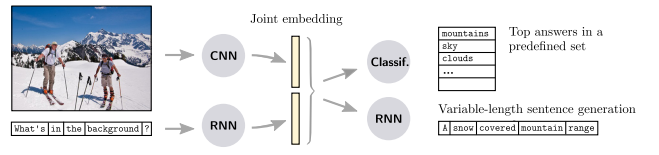
1. 通过基于表示学习对知识图谱、抽取事实和文本统一学习，不仅能够进行知识补充， 还可以学习语言符号与知识符号间的关联、匹配和推理。
2. 启发

通过本文让我对深度问答有了一定的了解。要想让计算机像人一样进行问答，利用传统学习方法手工选择特征很难做到，只能利用深度学习来自动选择特征。

1. 《Visual question answering A survey of methods and datasets》总结
2. 内容概要

本文是对图像问答（Visual question answering，VQA）的概述。首先介绍了一下图像问答，再从动机、方法、性能和局限介绍了VQA的4个主要方法：联合嵌入方法、注意力机制、组合模型、使用外部知识库的模型。然后按时间线介绍了典型的数据库及性能评估方法。之后介绍了注释图像的方法，最后对VQA作了一些讨论、指出了其发展方向并作了简单总结。

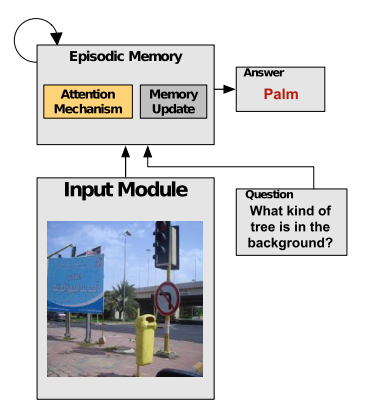
1. 总结及理解
2. VQA有机结合了计算机视觉和自然语言处理（NLP）。它与其它计算机视觉任务最大的区别是问题没有预先设定，是在运行时设定的。图像问答与文本问答有很大相似之处，文本问答的很多方法也能用在图像问答上。图像问答相较于文本问答，将输入的文本换成了图像，这就带来了很多挑战，比如维度变高、噪声变多、图像中包含现实生活的很多信息。相较于图像描述，图像问答可能涉及不包含在图像中的常识。
3. 图像问答用于解决不同形式、开放性的问题。
4. 联合嵌入方法：从文中 图中可看出，大多数VQA模型都要用到联合嵌入方法。

* 动机：使用CNN和RNN模型将图像和问题嵌入到同一个低维空间，方便后续处理。
* 方法：Neural-Image-QA（如下图）、Multimodal QA（mQA）、dynamic parameter layer (DPPnet)、MCB pooling、multimodal residual learning (MRN)等。
* 性能和局限：随着方法的改进，性能不断提高，也说明方法还有待继续改进。

1. 注意力机制

* 动机：找出与问题相关的重要局部特征，用重要的局部特征替换全部特征，减少模型处理的维度，同时减小预测层输入的无关信息或噪声。
* 方法：word-guided attention、question-guided attentionmap(QAM)、stacked attention networks(SAN)、multi-hopimage attention scheme(SMem)、hierarchical co-attention model(HieCoAtt)、Multimodal Compact Bilinear pooling(MCB)等。HieCoAtt会同时关注图像和问题，而在它之前方法仅仅关注图像。
* 性能和局限：注意力机制在所有VQA数据集使用时的准确度都能得到提升，但面对2个选项（是/否）的问题表现很差，可能是这种问题需要更多推理。

1. 组合模型（Compositional models）:到目前为止，所讨论的方法表现出与用于提取图像和句子的表示的CNN和RNN的整体性质相关的限制，可以通过模块化结构的人工神经网络解决这一问题，它为特定所需功能（如记忆功能、特定类型的推理）设计不同的模块。它的两个特别模型如下：

* 神经模型网络（Neural Module Networks，NMN)
* 动机：充分利用问题的语言构成结构，针对每个问题实例即时组合合适模型解决问题。
* 方法：使用NLP中使用的问题的语义解析工具，面对不同的问题组合不同的模块，其中各模块之间相互独立。
* 性能和局限：它通常在具有一定组成结构的问题上表现较好，但在简单问题上表现不好。
* 动态记忆网络（Dynamic Memory Networks，DMN）
* 动机：通过多次传递对数据的多个部分之间的交互进行建模来解决需要复杂逻辑推理的任务。
* 方法：如下图所示，DMN一般由输入模块、问题模块、注意力机制和记忆更新模块组成。
* 性能和局限：与NMN方法相比，在2个选项（是/否）问题上表现差不多，在计数问题上表现差一些，但在其他问题上明显优于NMN方法。

1. 使用外部知识库的模型

* 动机：联合嵌入的方法存在2个局限：只能学到训练集中的知识；神经网络的容量有限。将外部知识库作为由图像读取内容的有效补充，可以有效提高回答超出图像信息的问题的准确率，这里的外部知识包括常识、百科知识等。
* 典型知识库（Knowledge Bases，KB)：DBpedia、Free base、YAGO、OpenIE、NELL、WebChild、ConceptNet。
* 方法：Ahab、FVQA。
* 性能和局限：这种方法在需要超出图像信息的问题表现比较好，但现存的很多数据集中的很多问题都能从图像中找到答案，不需要额外知识。要训练这种网络需要开发新的数据集。

1. 数据集：VQA主要运用的是深度学习算法，同样也需要大数据来驱动，而VQA数据集的标注相对于其他任务要复杂一些，需要考虑问题的类型、问题的长度、答案的形式等，

* 数据集的主要因素：图像的类型，如自然图像、剪贴画、合成图像；问答的形式，如开放性问题、多选项问题；数据集的大小，如图片、问题和表示的不同概念的数量；所需推理的复杂程度；所需超出图像信息的知识。
* 常用数据集
* 自然图像数据集：DAQUAR、COCO-QA、VQA-real、FM-IQA、VQA-real、Visual Genome and Visual7W、Visual Madlibs等。
* 剪贴画数据集：VQA abstract scenes、Balanced dataset等。
* 知识库增强数据集：KB-VQA、FVQA等。
* 其他数据集：Diagrams、Shapes等。
* 写这篇论文时最大的数据集是Visual Genome dataset。
* 问题和答案的生成方式：人工标注、使用模型自动生成。
* 人工标注的数据集会带有一些偏见，如人会倾向于关注一些突出的物体，如桌椅。
* 7W：who, what, where, when, why,how, and which。
* 如何将外部知识库与VQA有机结合还是一个有待解决的问题

1. 结果评估方法：对于答案为选项的问题比较好判断，但对于开放性问题，评估会比较困难，以下是文中提到的一些评估方法：

* 使用字符串匹配简单地测量关于地面实况的准确性。只有完全匹配才被认为是正确的。
* Wu-Palmer similarity (WUPS)：评估分类树中它们共同子序列之间的相似性。 当两个单词之间的相似度超过阈值时，候选答案被认为是正确的。
* 图灵测试。

1. 展望：

* VQA将更好地与知识库相结合。
* 更多计算机视觉和自然语言处理的方法会更好地运用到VQA上。
* 更复杂多样的数据集将会出现。

1. 启发

通读全文后，自己对VQA有了初步的了解。VQA将图像描述、文本问答等计算机视觉和自然语言处理结合起来，CV和NLP的很多方法也能用在VQA上，能在一定程度上代表AI进步，同时也面临巨大挑战。不同类型问题和问题输入及不同类型答案输出需要模型处理更多不同的情况，也意味着需要更大的数据集。而VQA的数据集相比其他数据集标注更复杂，也存在更多偏见和歧义，比如不同的人对相同的问题有不同的答案。个人觉得很难找到单一的模型解决所有问题，需要通过模块化的思想建立串行和并行网络和接入知识库来处理各种不同情况，如通过激活函数的阈值控制不同并行通道的开关来实现各通道的切换，处理不同问题。另外，体会到大数据的量级，如150000个问题，感觉训练集最好尽可能多的包含现实中的各种情况，如各种姿态下的人或猫。VQA方向发展是个新方向，论文更新很快，要关注新的论文。

1. 《中文信息处理前沿技术进展》总结
2. 内容概述

本文首先对自然语言处理（NLP）进行了简单概述。从感知到认知，由运算智能到感知智能，再到认知智能，而NLP正是认知只能的核心，可见NLP的发展前景优越。然后依次介绍了NLP的阶段性发展、所面临的技术挑战及应对之道、行业应用。

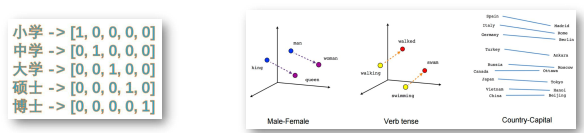
1. 总结及理解
2. 下表有条理、清晰的展现了NLP的一些应用，其中形式=名；语义=名+实；推理=名+实+知；语用=名+实+知+人。另外名为语言符号；实为客观事实、主观事实；知为知识；人为语言使用者。
3. NLP中常用的深度学习模型：循环神经网络、卷积神经网络、递归神经网络。
4. 分布表示较符号表示有了较大改进，由高维变为低维，可以减小训练数据集，可以根据两个词的距离分辨词义相似度。

图 1 符号表示 图2 分布表示

1. 知识图谱的发展将为NLP的发展奠定坚实的基础，将为算法提供大量可靠的先验知识。
2. 模型的预训练相当于把整理好的数据提供给训练模型，可以减小噪声干扰、特征维度等，可大幅提高多任务的准确率。
3. 文本生成的应用意味着计算机在一定程度上拥有了像人一样的创作力，有了思维。
4. 自然语言处理平台的开放很可能像开源软件（如安卓、opencv等）一样极大促进NLP领域的发展。
5. 汉语相较于英语处理更困难。
6. 对话系统从应用到平台化意味着技术的落地，逐步走向成熟。
7. 带标数据不足和常识知识不足是NLP面临的2个主要技术挑战，相信随着NLP的发展，带标数据会越来越多，随着事理图谱和知识图谱的发展，常识不足的问题也会得到解决。
8. 事理图谱和知识图谱的区别：事理图谱研究对象是谓词性事件及关系，谓词一般是动词，而知识图谱研究对象是名词性实体及关系。
9. 智慧医疗将极大提高医院的工作效率，有利于解决看病难等问题，医学大数据和使用经验也将推动NLP的发展；同样智慧教育、智慧司法都可提高工作效率，进一步促进NLP的发展，另外智慧司法也可有效避免徇私枉法等，使法律做到真正意义上的公平、公正、公开。
10. 启发

通过本文对自然语言处理有了初步了解。作为人工智能皇冠上的明珠，NLP的前途无限光明。除了学习NLP相关算法模型外，还应注意分析和理解词、句的含义，将两者有机结合起来

1. 《机器学习在能源与电力系统领域的应用和展望》总结
2. 内容概述

本文介绍了机器学习在能源与电力系统领域的应用和展望，同时，以此为支点，在文中和附录中很好的总结了强化学习（RL）、深度网络（DL）、迁移学习（TL）、平行学习（PL）、混合学习（HL）、对抗学习（AL）、集成学习（EL）。

1. 总结及理解
2. 深度强化学 习 （DRL）作为AI2.0代表在无需先验知识、降低资源消耗、提高训练速度等方面取得了可喜进步,可提高机器对无标签的未知环境的感知能力。
3. AI概念：一般认为AI是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门集控制论、信息论、计算机科学、数理逻辑、神经生理学等学科于一体的新兴交叉学科。其终极目标是让计算机能像人一样思考和行动，并进一步提升人的智能。
4. 推动AI发展的三大关键因素：海量数据资源、超常计算能力、核心算法模型。
5. 人工智能（AI）、机器学习（ML）、深度学习（DL）三者的层次关系是：AI>ML>DL。常见ML算法主要分四大类：有监督式学习、无监督学习、半监督式学习和强化学习。
6. 强化学习（RL）：RL属于主动学习的一种，即本质上是一种在环境中“审时度势”地学习策略的过程，强调如何基于状态而行动，以取得最大化的预期收益。近20年来，以 马尔可夫决策过程（MDP）为严格数学基础的RL算法成为ML领域一个新的突破。算法中的几个关键因素为：状态集、动作集、状态转移分布、贴现因子、奖励函数。常用的优化方法有：值迭代和策略迭代。其中，值迭代更常用一些。
7. 深度学习（DL）：本质上是具有多层隐藏层的神经网络，事实上，4层以上的神经网络就可称之为DL。其基本思想是通过多层的网络结构和非线性变换，组合低层特征，形成抽象的、易于区分的高层表示，以发现数据的分布式特征表示。其避免了手动提取特征的过程。近年来，DL逐渐与其他ML方法结合形成了功能更强的高级混合ML算法。周志华教授提出的“深度森林”未来有可能取代DNN。
8. 迁移学习（TL）：强调的是在不同但相似的领域、任务和分布之间进行知识的迁移。
9. 平行学习（PL）：把“小数据”炒成“大数据”，再把“大数据”提炼成解决具体问题的精准知识的“小智能”的过程，即“小数据、大定律”到“大数据、小定律”的过程（即从牛顿定律到默顿定律的升华），使得ML从已知训练样本集（有限小数据）走向了自我探索获取海量假想训练样本（无限大数据）的时代，这也是AI超越人类智力的分水岭。这样能在一定程度上增大数据集，促进AI的发展。使用预测学习解决如何随时间发展对数据进行探索，使用集成学习解决如何在空间分布上对数据进行探索，使用指示学习解决如何探索数据生成的方向。
10. 混合学习（HL）: 将至少两种优势互补的ML方法集成或综合为一种性能更强的ML。将DL作为“感知”与RL作为“决 策”结合形成HL的一对“黄金组合”。
11. 对抗学习（AL）或生成式对抗网络（generative adversarial network，GAN）：通过构造相互竞争的生成器和判别器来提高学习的效率。
12. 集成学习（EL）: 严格意义上说并不算一种ML算 法，而更像是一种优化手段或策略，其通常结合多个简单的弱ML算法有策略地生成一些基础模型，然后有策略地将它们集成模型组，去更可靠地完成学习任务并做出最终决策。注意其与混合学习的差别。
13. 发展瓶颈：数据，需要大量数据；泛化能力，有时实际泛化能力比训练时泛化能力明显下降；能耗，程序需长时间运行，比较耗电；语义鸿沟，缺乏语言理解能力；可解释性，现有AI过于依赖训练数据，缺乏深层次数据语义挖掘，导致可解释性差；可靠性，现有 AI系统可靠性较差，某些错误识别结果可能会带来致命后果,其为黑箱模型，内部结构难解释。
14. 启发

通读全文让我对机器学习有了初步了解。同时，也让我意识到“没有免费的午餐”定理，每种算法都有自己的优点与不足，找不到通用的算法，要根据实际问题选择合适的算法，也可以考虑把几种算法融合起来，取长补短。算法在不断优化，朝着更好、更快、操作更简单的方向发展。然后阐述了NLP面临的最大的几个开放性问题。

1. 《Frontiers of Natural Language Processing》总结
2. 内容概述

本文是一篇会议记录，会议主要围绕以下三个问题展开：1）什么是NLP？最近几年NLP的重要发展有哪些？2）NLP中最大的开放性问题是什么？3）鼓励从业者了解各地社区并加强合作。首先回顾了最近几年（2001年-2018年）NLP的发展历史。最后回答了2个问题：1）是什么导致该领域走向了错误的方向？2）对于开始学习NLP的在读研究生，有什么建议？

1. 总结及理解
2. 神经语言模型（2001年）：

* 语言模型：通过前一个词预测下一个词。
* 经典语言模型：光滑的n-grams。举例来说：将“informationretrieval”视为一段文本，它的5-grams的items依次为：

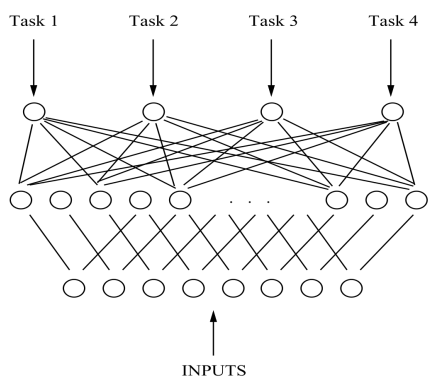
infor,nform,forma,ormat,rmati,matio,ation,tion,ionr,onre,nret,retr,retri,etrie,triev,rieva,ieval

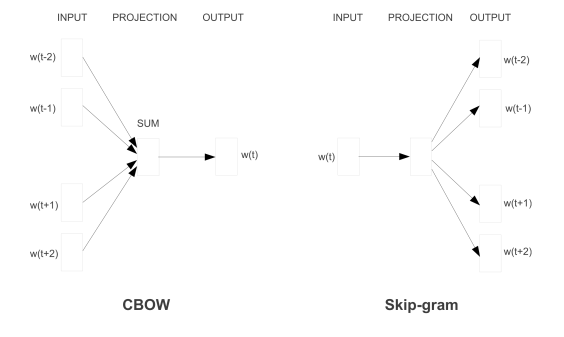
有时为了便于分析，还会在前面加空格，这就多出4个items：

\_\_\_\_i,\_\_\_in,\_\_inf,\_info。

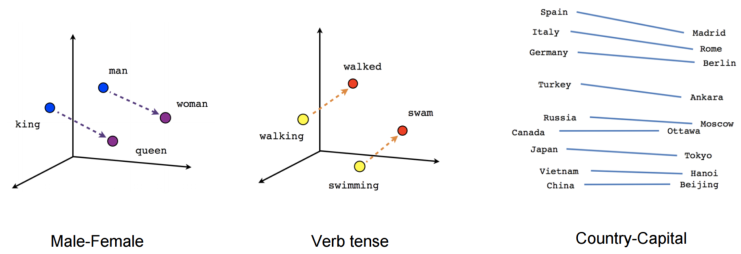
参考网站：<https://www.cnblogs.com/cdsj/p/5720391.html>

* 第一个神经语言模型：基于前n个词的前馈神经网络。
* 因为每个词对应于初始查找层的权重矩阵的一个向量，所以我们经常把这个矩阵作为词嵌入矩阵。
* 之后的语言模型有：RNN、LSTM。

1. 多任务学习（2008年）：模型在不同任务的训练中共享权重，如下图所示，各任务神经元相互连接。
2. 词嵌入（2013年）：自然语言处理（NLP）中语言模型与表征学习技术的统称。概念上而言，它是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。

* 主要创新点是：可在很大的未标记数据集通过预训练得到词嵌入矩阵。
* CBOW和skip-gram：两者模型如下图所示。CBOW模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。Skip-Gram模型和CBOW的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。CBOW对小型数据库比较合适，而Skip-Gram在大型语料中表现更好。

参考网站：<https://www.jianshu.com/p/471d9bfbd72f>

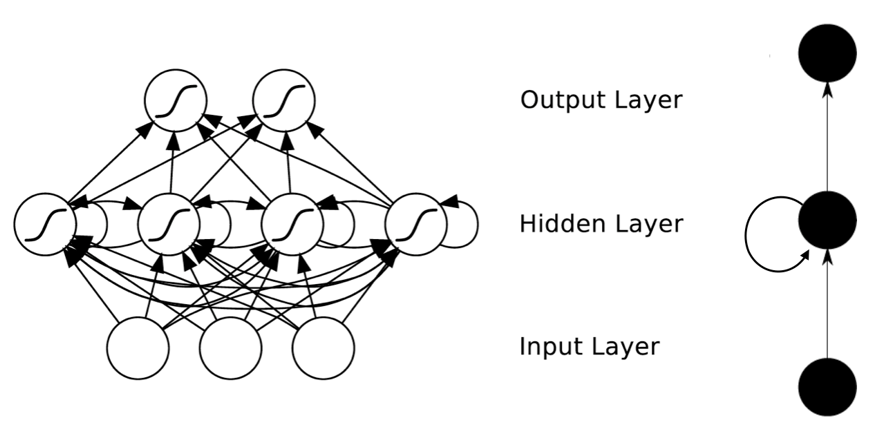
* 词嵌入可以通过对无标记数据集的预训练发掘词之间的某些关系，如下图所示：

1. NLP相关神经网络（2013年）：

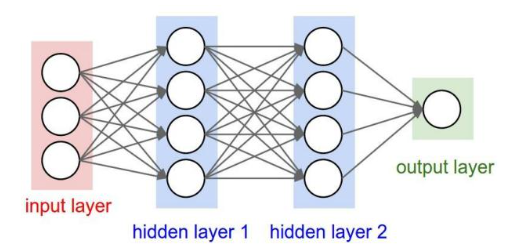
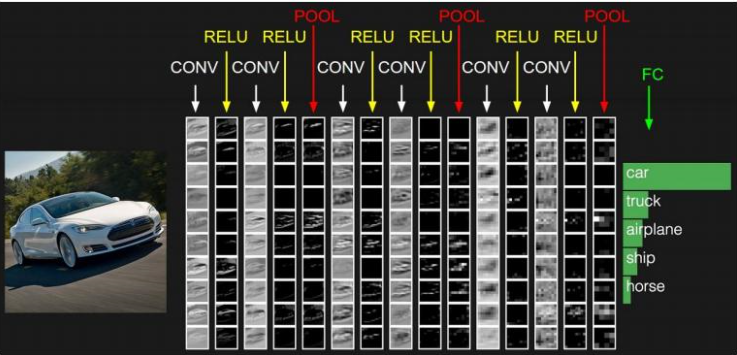
* 主要挑战：处理动态输入序列。
* 三个主要模型：
* 循环神经网络（Recurrent neural networks，RNN）：结构如下图，与CNN不同的是：不仅考虑前一时刻输入,而且赋予了网络对前面的内容的一种'记忆'功能。

之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。

参考网站：<https://blog.csdn.net/qq_32241189/article/details/80461635>



* 卷积神经网络（Convolutional neural networks，CNN）：结构如下图，CNN一般包含卷积层、激励层、池化层。卷积层中每个神经元可看成一个滤波器，只关注输入的某个特征；激励层采用函数一般为ReLU，其收敛快，求梯度简单，但较脆弱；池化层用于压缩数据和参数的量，用于图像处理的主要作用是压缩图像。

参考网站：<https://www.cnblogs.com/skyfsm/p/6790245.html>

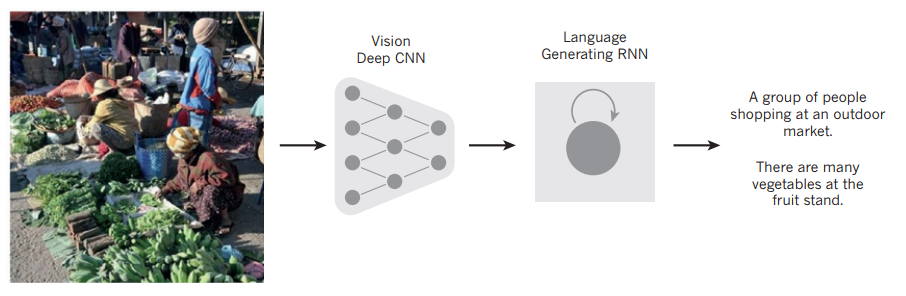
* 递归神经网络（Recursive neural networks，RNN）: 为了能够让模型区分出两个不同的意思，我们的模型必须能够按照树结构去处理信息，而不是序列，这就是递归神经网络的作用。当面对按照树/图结构处理信息更有效的任务时，递归神经网络通常都会获得不错的结果。

尽管递归神经网络具有更为强大的表示能力，但是在实际应用中并不太流行。其中一个主要原因是，递归神经网络的输入是树/图结构，而这种结构需要花费很多人工去标注。想象一下，如果我们用循环神经网络处理句子，那么我们可以直接把句子作为输入。然而，如果我们用递归神经网络处理句子，我们就必须把每个句子标注为语法解析树的形式，这无疑要花费非常大的精力。很多时候，相对于递归神经网络能够带来的性能提升，这个投入是不太划算的。

参考网站：<https://blog.csdn.net/u010164190/article/details/78172289>

* 长短时记忆网络（LSTM）：它在原始RNN的基础上增加了单元状态c来保存长期的状态。解决了梯度消失和梯度爆炸的问题。它用两个门来控制单元状态的内容，一个是遗忘门（forget gate），它决定了上一时刻的单元状态有多少保留到当前时刻；另一个是输入门（input gate），它决定了当前时刻网络的输入有多少保存到单元状态。LSTM用输出门（output gate）来控制单元状态有多少输出到LSTM的当前输出值。

参考网站：<https://zybuluo.com/hanbingtao/note/581764>

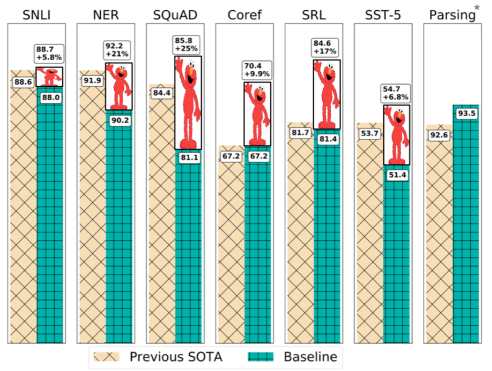
1. 序列-序列模型（2014年）：其主要应用是机器翻译。除了句子-句子的模型，还有图像-句子的模型，如下图所示，其中CNN用作编码器，RNN用作解码器。
2. Attention（2015年）：Attention-based Model其实就是一个相似性的度量，当前的输入与目标状态越相似，那么在当前的输入的权重就会越大，说明当前的输出越依赖于当前的输入。严格来说，Attention并算不上是一种新的model，而仅仅是在以往的模型中加入attention的思想，所以Attention-based Model或者Attention Mechanism是比较合理的叫法，而非Attention Model。

参考网站：<https://blog.csdn.net/u014665013/article/details/82619808>

<https://blog.csdn.net/hahajinbu/article/details/81940355>

1. 基于记忆的神经网络（2015年）：attention具有模糊的记忆，为了弥补这个缺陷，提出了具有更加清晰记忆的模型：神经图灵机、记忆网络、end-to-end记忆网络、动态记忆网络、神经可微计算机、循环实体网络等。一个记忆网络包括了记忆m，还包括以下4个组件：I (input feature map)、G (generalization)、O (output feature map)、R (response)。

参考网站：<https://blog.csdn.net/u014248127/article/details/84894739>

1. 预训练语言模型（2018年）：它仅用于初始化的第一层，在大数据集的预训练中可以获取很多附带的信息。如下图所示，应用预训练模型，在处理很多不同的任务中获得很大的提升。
2. NLP面临的最大问题：自然语言理解；低资源场景的应用；大文本或多文本的推理；数据集、定义问题和评估方法。
3. 启发

跟着会议记录的时间线阅读，让自己对NLP近几年的发展有了初步了解。会议最后的两个问题也给了自己较大的启发：要敢于尝试不同的模型和方法，成功和失败的案例将一起推动NLP螺旋式发展；可以通过阅读大量机器学习、深度学习、强化学习文章学习相关知识，并将它们与NLP有机结合起来；要勤于动手实践，学会调节模型和设计评估标准，用代码来验证理论的可行性；不要过分关注某篇文章，若不能验证其可行性的情况下，可等待更多相关论文的发表。

第3周汇报

——张溢炉

1. V3A-Visual-based Asking, Answering and Acting（吴琦，阿德莱德大学）和Locate, Tell, Answer and Act: Connecting Vision, Language and Beyond（虞立成，UNC-Chapel Hill）总结
2. 内容概述

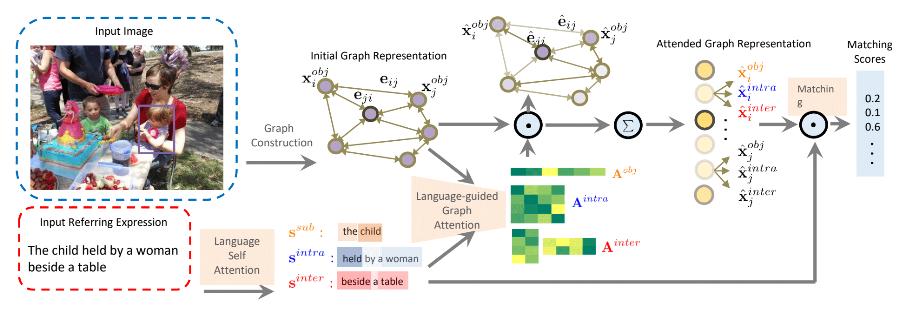
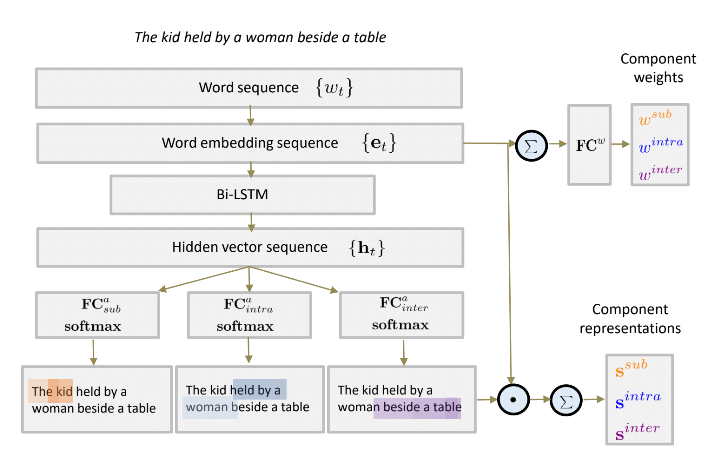
吴琦和虞立成两位老师的报告主要讲述将图像处理、自然语言处理和语音处理等人工智能技术与机器人有机结合起来的最新科研成果，将多模态与机器人相结合，可以使机器人具备像人一样的感知、推理和决策的能力，可以通过人机交互使机器人动作、完成相关任务。报告也介绍了与之相关的图像问答（VQA）、图像描述、图像对话、VLN（viusal-and-language navigation）、EQA（embodied question answering）等科研成果，并介绍了相关模型、数据集、目标函数等知识。

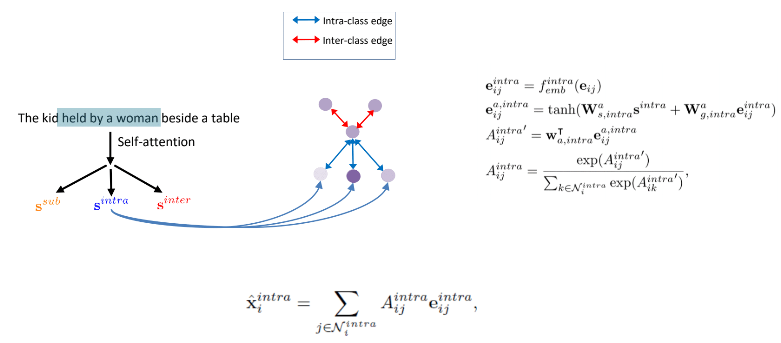
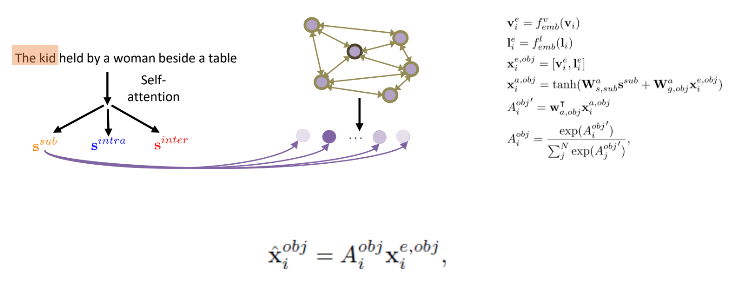
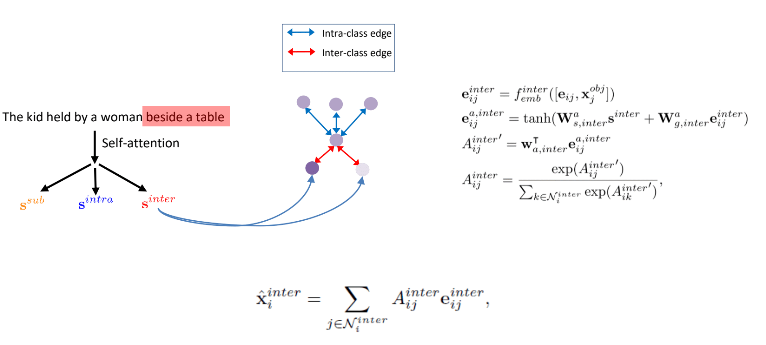
其中，吴琦老师主要讲述了他的三篇论文，虞立成博士从locate、tell、answer、act四个方面介绍了他所做的工作。

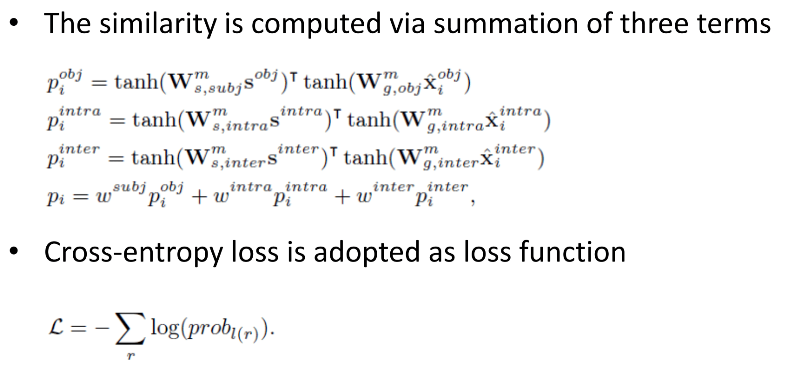
1. 总结和理解
2. 吴琦老师PPT链接：

<https://www.dropbox.com/s/int0huvbebiw46m/VALSE_online.pdf?dl=0>

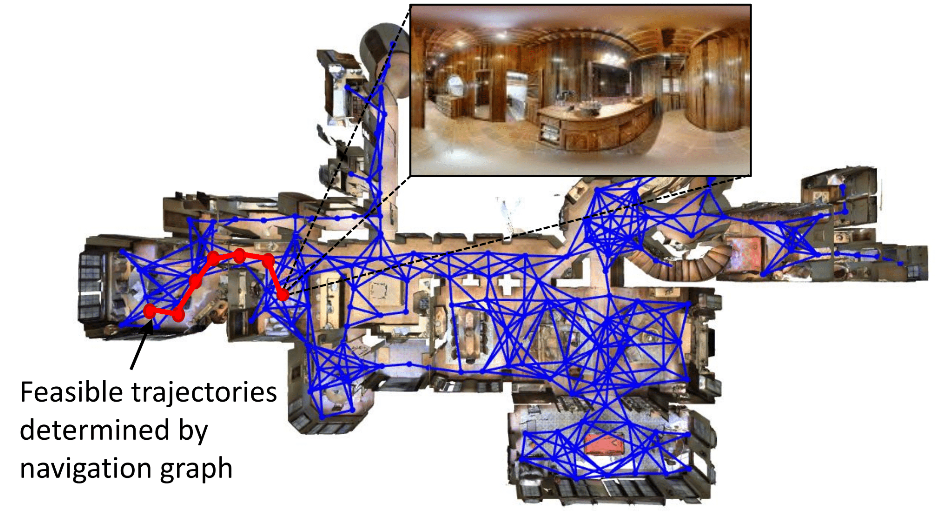
1. 吴琦老师的《Neighbourhood Watch: Referring Expression Comprehension via Language guided Graph Attention Networks》总结：

* 动机：参考表达不仅描述了指示物的属性还有它与邻居的关系；图像可灵活地捕获对象关系；语言引导注意力机制能突出相关信息。
* 模型：如下图所示，其中
* language self-attention module将语言表达分成描述主体，类内关系和类间关系三部分，模型如下图；
* language-guided graph attention module：在候选对象上构建有向图，分别在s\_sub，s\_intra和s\_inter的指导下突出显示节点（对象，如孩子），类内边缘（同一类别的对象之间的关系，如女人和孩子的关系）和类间边缘（来自不同类别的对象之间的关系，如孩子和桌子的关系），模型及公式如下图

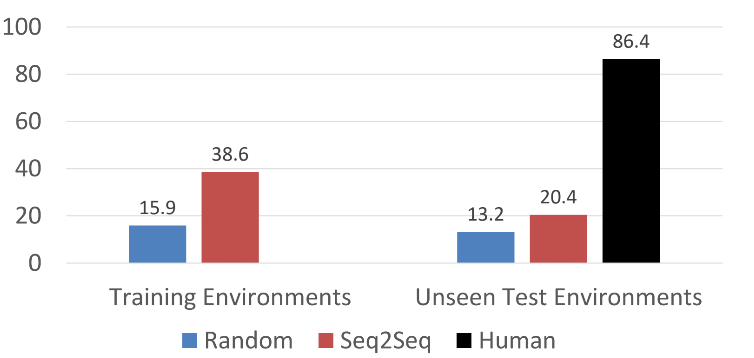


* matching module：计算表达和对象的匹配分数。公式如下：
* 实验
* 任务：模型能根据参考语言表达识别出图片中对应对象
* 数据集：RefCOCO、RefCOCO+、RefCOCOG
* 结果：较之前最好模型有1-4个百分点提升

1. 吴琦老师《Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments》总结

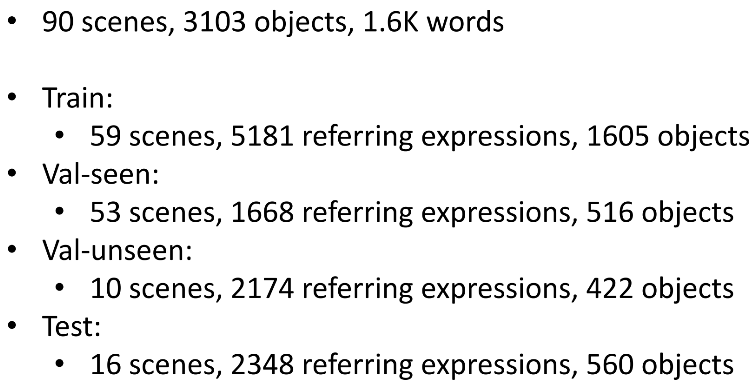
* VLN和VQA的区别：VQA和VLN输入都是图片和句子，但VQA输出是句子，VLN输出是一系列动作
* Matterport3D仿真环境：90个建筑里的10800张全景图，包含194400张RGBD图片，环境复杂多变。
* 路径：路径是通过环境中设定节点连接而成，不能在环境中随意行走，在每个节点上会有一张360度的全景图。
* Room-To-Room实验
* 任务：机器人根据自然语言指导指令（如离开卧室，进到厨房。直走，沙发的地方左转，停在窗口旁）在仿真环境中寻找目的位置
* 评价标准：单次测试运行；机器人必须自己选择停止；如果机器人最后停在了目的位置3m范围内表示成功，是否在3m范围内由另一个人（非下达指令的人）来评判。

启发：这种交叉验证的方法比较科学。首先可以判断命令是否明确，有无歧义，使验证方法更加科学。

* 指令数据集：使用AMT收集了21567条导航指令，指令平均长度是29个词，路径平均长度约10m。
* 挑战：测试时有一些未出现的场景、新的概念、新的词汇等
* 结果如下图，准确率很低，在未见过的环境中更低
* 发展的四个阶段：R2R（点到点）导航、寻找可视目标、寻找隐藏目标、机器人能通过询问他人寻找目标。

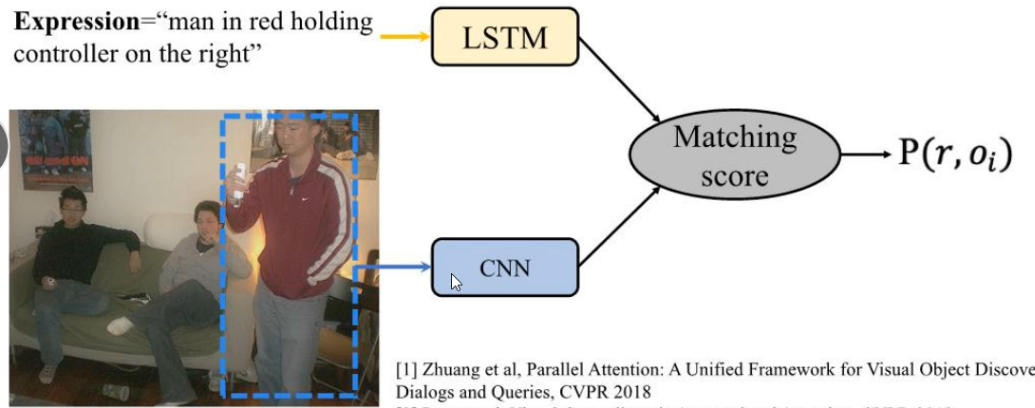
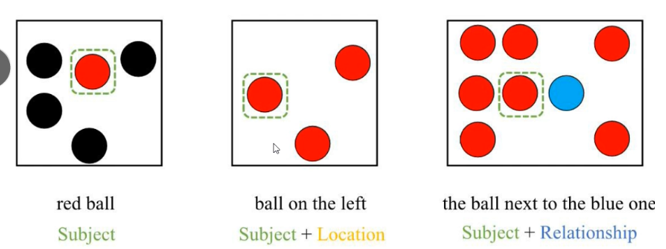
1. 吴琦老师《RERERE：remote embodied referring expression in real indoor environment》总结

* 长期目标：制作可以观察周围环境、执行命令，并能与人类交流的机器人。给机器人基于语言的导航指令，使机器人能根据指令选择最佳路径达到目的地，实验使用真实的3D场景（扫描真实建筑所得），机器人不仅要识别关键的场景、物体，还要做出相应的路径规划。
* 其指令集较之前RE和VLN数据集，更接近于人类平时的交流，如帮我把卧室的杯子拿来
* 主要工作：
* 引入新的视觉和语言任务
* 建立RERERE数据集
* 提出navigator-pointer model作为首个基准
* 挑战
* 明显的外观变化
* 高级描述的复杂多样
* 指令很难理解
* 仿真环境：Matterport3D
* RERERE数据集如下

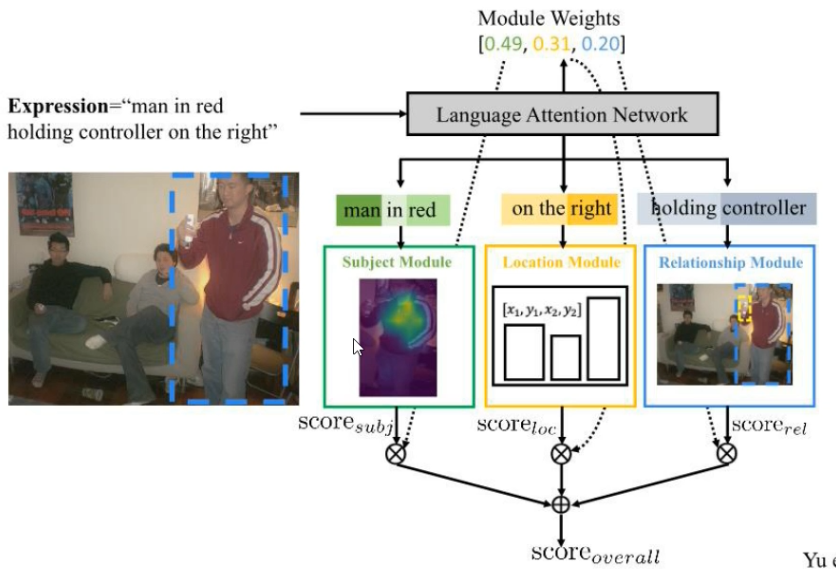


* 实验结果准确率都比较低（20%左右），表明该研究方向还有很大提升空间，其中机器人通过36张图片确定当前位置（联系slam相关知识）

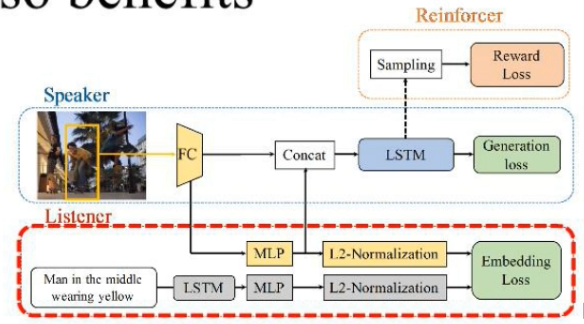
1. 虞立成博士讲座总结

* Locate：通过语言中的表达寻找图中对应物体，如Referring Expression。Tell：用语言描述图片中物体，通过观察图片生成对图片目标的描述。Answer：根据图片信息来回答问题。Act：机器人根据指令作出相应动作。
* RE模型改进步骤如下：
* 对于RE问题，首先想到简单模型如下：
* 但作者产生疑问：用统一的encoding各式各样的句子正确吗？通过观察下图可知：在不同场景中对相同物体的描述的语言结构不一样。注意location和relationship区别：前者句中只有一个实体，后者有2个实体。
* 通过思考，作者提出如下MAttNet（Modular Attention Network）模型，其中提到：

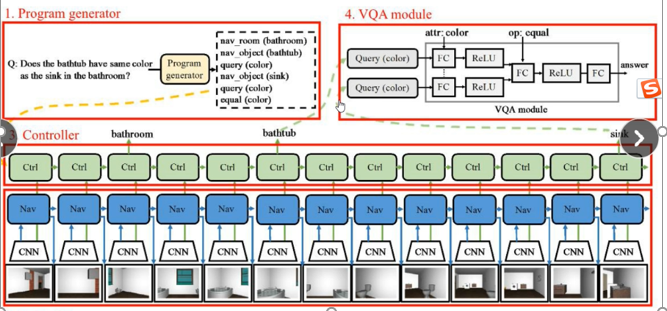
Language parallel模型可能会将误差传下去，导致错误。代码链接：

<https://github.com/lichengunc/MAttNet>

* RefCOCO数据集采集游戏描述：首先由一个人描述图片中的某个物体，再由另一个人根据描述选择对应的物体，如果选择正确，双方都能获得奖赏。游戏的设计要求双方相互合作来获得更多奖赏。根据游戏的设置可得到如下模型，很好的的将comprehension和generation结合起来，speaker和listener的交互会使两边都有提升：



* EQA：机器人在未知环境中以第一视角来回答问题。刚开始模型问题一般只包含单个目标物体，但作者的所做问题中包含了多个目标物体。
* MT-EQA模型如下，将问题拆解成子问题的思路，注意控制器及其实现方法。相关陈述：
* 实时性：RE的实时性不错，而MT-EQA在未知环境中2、3千个问题2、3个小时，大概一个问题10s,0.2s~0.3s一张图。
* 路径规划和机器人差不多,包含一些slam的方法。



1. 其他

* 图像和语言之间存在客观联系，其中图像是客观的，语言是主观的，人们对相同图像的解读（自然语言）都不一样，我们要想办法打破这个界限。
* 过拟合表现：在训练房间效果好，在未见过房间性能下降很快。
* Attention：知道结果后再返回推关注哪部分，不是推理，是一种借口，和推理有区别，推理是通过逻辑、证据实现推理，如query graph中每个节点形成逻辑链之后得到答案，现在有clever和GQA可以实现推理（不太确定记录是否准确）
* 机器人不可能什么都知道，要让机器人懂的更多，首先可以借助知识库，更重要的是机器人得清楚自己知道什么、不知道什么，可以通过询问的方式解决不知道的，更新自己的知识库，这也可以反过来辅助VQA
* 图像对话是图像问答多轮次的版本
* VQA包含内容很多，很难再训练集中包含所有情况
* 用游戏方式采集数据集，让更多人参与，如讲座中游戏设计很有意思，两边为了得到钱会认真参与，不像简单的问卷调查。
* 虞博陈述（记录大概意思）：早期时也是看到任务就去找各种模型，找到最好模型后就想着出论文，感觉这种方法不好，一上来就从模型的角度和写论文的角度出发不太好，后面发现这个切入点太糟糕，后期会关注dataset，百分之四十时间花在数据集分析（和吴恩达视频不谋而合），然后去寻找目前模型的缺憾，模型的缺点，会很有帮助。大概花了三年半在RE 。

在embodied中visual离解决还差很远，现在没有边走边建slam、深度信息和map信息，主要问题objection navigation和room navigation，之后和language结合的工作会做得更好。

* 谷歌bert：用bert做VQA效果很好
* 对小白建议：多看paper多coding，多花时间，不贪多，专一个方向，常锻炼，身体也很重要
* 创造未见过环境的方法：改变path、生成指令，加入数据集；直接改变环境，这个较难；抹去一些特征（feature dropout），如把所有椅子和桌子去掉。

1. 疑问

* 模型的公式如何推导的？
* 如何用代码实现模型？模型之间是如何连接的？attention的实现原理和方法？
* 怎么优化目标函数？
* 数据集中目标对象是如何选择的？怎么确定目标对象的在图像中的范围？如何标定图像中目标对象的关系？
* 参考描述设计有什么技巧或格式？
* 如何分析数据集，选择更好的模型？
* 后期如何解决实际应用时的实时性、噪声等问题？

1. 参考文献见最后
2. 启发

前段时间刚刚接触自然语言处理(NLP),自己萌生了将多模态（视觉、语音、文本等）、机器人相关技术（slam（同时定位与地图构建）、导航、路径规划等）结合起来的想法，如制作家庭服务机器人。没想到自己刚刚萌生的想法已经做得这么好了，通过这个视频了解到现在已经有人在做这个方向了，如VLN和EQA等，给了自己很大的学习动力。由于这涉及知识面太广，自己还是得通过研究生三年努力把自然语言处理（NLP）学好，争取早出成果。

1. 《KGDevReport知识图谱发展报告2018》总结
2. 内容概述

全文首先对知识图谱进行了简单概述，然后介绍了知识表示与建模、知识表示学习、实体识别与链接、实体关系学习、知识存储与查询、知识推理、通用和领域知识图谱、语义集成、语义搜索、基于知识的问答。

1. 总结和理解
2. 知识图谱（Knowledge Graph）以结构化的形式描述客观世界中概念、实体及

其关系，将互联网的信息表达成更接近人类认知世界的形式，提供了一种更好地

组织、管理和理解互联网海量信息的能力。

1. 知识图谱在知识融合、语义搜索和推荐、问答和对话系统、大数据分析与决策中凸显出越来越重要的应用价值。
2. 知识图谱技术分为三个部分：知识图谱构建技术、知识图谱查询和推理技术，以及知识图谱应用。
3. 知识图谱的构建：知识表示与建模、知识表示学习、实体识别与链接、实体关系学习、事件知识学习。
4. 知识图谱的的查询和推理计算：知识存储和查询、知识推理。
5. 知识图谱的应用：通用和领域知识图谱、语义集成、语义搜索、基于知识的问答。
6. 知识表示与建模：机器必须要掌握大量的知识，特别是常识知识才能实现真正类人的智能。表示学习的发展，以及自然语言处理领域词向量等嵌入（Embedding）技术手段的出现，启发了人们用类似于词向量的低维稠密向量的方式表示知识的研究。

RDFS（RDF Schema）主要用于定义术语集、类集合和属性集合，弥补了RDF（Subject-主语,Predicate-谓语,Object-宾语）缺少类、属性等 Schema 层的定义手段的缺陷。

1. 知识表示学习：知识表示学习将实体和关系表示为稠密的低维向量, 实现了对实体和关系的分布式表示， 可以高效地对实体和关系进行计算,、缓解知识稀疏、有助于实现知识融合，已经成为知识图谱语义链接预测和知识补全的重要方法。

近年来，以深度学习[Bengio, et al., 2009]为代表的表示学习[Bengio, et al.,

2013]技术异军突起。表示学习得到的分布式表示，则能够高效地实现语义相似度计算等操作，显著提升计算效率。将大量对象投影到统一空间的过程，能够将高频对象的语义信息用于帮助低频对象的语义表示，提高低频对象的语义表示的精确性。

1. 实体识别与链接：实体是文本中承载信息的重要语言单位，实体也是知识图谱的核心单元。实体识别是识别文本中指定类别的实体。实体链接是识别出文本中提及实体的词或者短语（称为实体提及），并与知识库中对应实体进行链接。实体识别与链接是知识图谱构建、知识补全与知识应用的核心技术。实体识别技术可以检测文本中的新实体，并将其加入到现有知识库中。实体链接技术通过发现现有实体在文本中的不同出现，可以针对性的发现关于特定实体的新知识。

实体链接主要解决实体名的歧义性和多样性问题，是指将文本中实体名指向其

所代表的真实世界实体的任务，也通常被称为实体消歧。

1. 实体关系学习：实体关系描述客观存在的事物之间的关联关系，定义为两个或多个实体之间的某种联系，实体关系学习就是自动从文本中检测和识别出实体之间具有的某种语义关系，也称为关系抽取。分为预定义关系抽取和开放关系抽取。实体关系学习方法分为基于表示学习的方法和基于图特征的方法。
2. 事件知识学习：事件是发生在某个特定时间点或时间段、某个特定地域范围内，由一个或者多个角色参与的一个或者多个动作组成的事情或者状态的改变。事件的四种属性：极性，时态，指属、形态。
3. 知识存储与查询：目前知识图谱多以三元存在的 RDF 形式进行存储管理，对知识图谱的查询支持 SPARQL 查询。
4. 知识推理：知识推理从给定的知识图谱推导出新的实体跟实体之间的关系。知识图谱推

理可以分为基于符号的推理和基于统计的推理。

1. 通用和领域知识图谱：通用知识图谱可以形象地看成一个面向通用领域的“结构化的百科知识库”，其中包含了大量的现实世界中的常识性知识，覆盖面极广。领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个“基于语义技术的行业知识库”。从无到有构建过程：知识建模、知识获取、知识融合、知识存储、知识计算、知识应用。
2. 语义集成：常见流程包括：输入、预处理、匹配、 知识融合和输出。语义集成的输入包括待集成的若干个知识库以及配置、外部资源等。语义集成的输出是一个统一的、一致的、简洁的知识库。
3. 语义搜索：传统搜索技术仍以关键词匹配、倒排索引和网页的链接结构为搜索依据，其查全率和查准率均无法满足用户日益提高的标准。与传统搜索技术不同，语义搜索是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句背后的真实意图，并依此来进行搜索，从而更准确地向用户  
   返回最符合其需求的搜索结果。
4. 基于知识的问答：问答系统（Question Answering, QA）是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。

主要面临三个关键科学问题：问句语义解析、大规模知识推理、异构知识（多种语言、多种领域、多种模态）关联。

1. 启发

在常识、百科知识、文本、新闻等包含大量知识的库中，存在很多规范和不规范的知识。在构建知识图谱时，一般规范的知识表达比较好处理，不规范的知识表达经常需要转换成规范的知识表达来处理。个人理解知识图谱将建立一个像搜索引擎背后那样庞大的知识库，但是与之前搜索不一样的是：知识图谱以图的形式存储，可直接回答问题，而不是单一的文本匹配。在如此庞大的知识图谱的构建中，一定得以借助已有结构化和非结构化知识自动构建为主，人工构建为辅。自动构建的库一定会有错误，可以通过人工来纠错。

1. 《Attention Is All You Need》总结
2. 内容概述

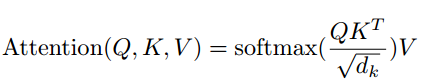
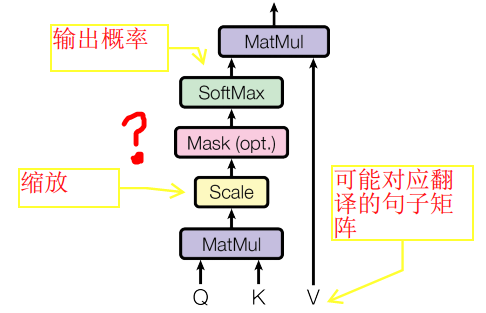
本文提出了一个新的神经网络模型-Transformer。它是完全基于注意力机制的。

翻译链接：<https://blog.csdn.net/DarrenXf/article/details/88556369>

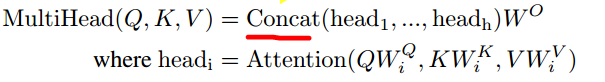
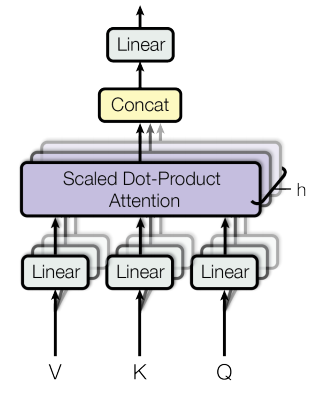
1. 总结及理解
2. 背景：之前，常用RNN、LSTM、GRNN来完成序列类及翻译任务。注意力机制的出现，改变了这一状况，它在很多任务中都有较好的表现。其中，本文提出的Transformer允许更多的并行化，并且在8个P100 GPU上经过长达12小时的培训后，可以达到翻译质量的最新技术水平。
3. Transformer模型如下图所示，在这里，编码器（图的左边）将符号表示的输入序列（x1,…,xn)映射为连续表示序列z = (z1,…,zn)。 给定z,解码器（图的右边）然后一次生成一个符号的输出序列(y1,…,ym)。在每个步骤中，模型都是自动回归的，在生成下一个时，将先前生成的符号作为附加输入使用。



* 编码器：由一组n=6个相同层组成。每层有两个子层。第一层是一个multi-head self-attention mechanism,第二层是一个简单position-wise fully，全连接于feed-forward network。
* 解码器：同样由一组n=6个相同层组成。除了每个编码器中的两个子层外，解码器还插入第三个子层，该子层在编码器堆栈上的输出上执行multi-head attention。
* 注意力机制
* Scaled Dot-Product Attention：模型图和公式如下，相比



* Multi-Head Attention：模型和公式如下图，有图可见它又3层Scaled Dot-Product Attention级联而成，输入、输出加了线性变换。它被用在了Transformer模型中的三个地方，注意每个地方的Q、K、V。



1. Additive attention 使用具有单个隐藏层的前馈网络计算兼容性函数。虽然两者在理论上复杂度是相似的，但在实践中，dot-product attention速度更快，空间效率更高，因为它可以使用高度优化的矩阵乘法代码来实现。对于较小的dk值，这两种机制的性能相似，在不缩放较大的dk值时，additive attention 优于dot product attention。

作者怀疑，对于较大的dk值，dot products的增长幅度较大，将softmax函数推到具有极小梯度的区域。为了抵消这种影响，我们缩放点积。

1. Position-wise Feed-Forward Networks：公式如下，其中使用了ReLU激活函数。
2. Positional Encoding：由于我们的模型不包含循环和卷积，为了使模型能够利用序列的顺序，我们必须注入一些关于序列中标记的相对或绝对位置的信息。
3. 使用self-attention的目标：每层的总计算复杂性；可以并行化的计算量，用所需的最小顺序操作数来衡量；网络中长期依赖项之间的路径长度。
4. 疑问：

* 在翻译任务中的，输入、输出是什么？输入、输出的表示形式是怎么样的？
* Q、K、V值有什么含义吗？在翻译任务中有什么含义吗？
* 自回归怎么实现的？
* additive attention和Dot-product attention等attention机制更深层原理？
* 级联的作用？
* 优化和正则化的原理及作用？

1. 启发

作者描述了设计模型过程式用了2个“countless”，表明作者花费了很多时间和精力设计、调节模型。这也启发自己在做科研的过程中要沉下心花费时间精力专研，要勇于尝试、直面失败。从table3中可以看出作者是有策略的调参，虽然模型的可解释性比较差，但是可以通过有策略的调参使模型泛化能力更好。

虞立成博士参考文献

[1] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, Dhruv Batra, “Multi-Target Embodied Question Answering”, in CVPR 2019.

[2] Hao Tan, Licheng Yu, Mohit Bansal, “Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout”, in NAACL 2019.

[3] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Mohit Bansal, Tamara L. Berg, “MAttNet: Modular Attention Network for Referring Expression Comprehension”, in CVPR 2018.

[4] Licheng Yu, Hao Tan, Mohit Bansal, Tamara L. Berg, “A Joint Speaker-Listener-Reinforcer Model for referring expressions.”, in CVPR 2017.

[5] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg, “Modeling Context in Referring Expressions”, in ECCV, 2016.

吴琦老师参考文献：

Image Captioning

–Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, Anthony Dick. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? CVPR’16

–Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel, Image Captioning and Visual Question Answering Based on Attributes and Their Related External Knowledge. TPAMI

•VQA

–Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. CVPR’16

–Peng Wang\*, Qi Wu\*, Chunhua Shen, Anton van den Hengel. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. CVPR’17

–Damien Teney, LingqiaoLiu, Anton van den Hengel, Graph-Structured Representations for Visual Question Answering.CVPR’17

–Peng Wang\*, Qi Wu\*, Chunhua Shen, Anton van den Hengel, Anthony Dick. Explicit Knowledge-based Reasoning for Visual Question Answering.IJCAI’17

–Peng Wang\*, Qi Wu\*, Chunhua Shen, Anton van den Hengel, Anthony Dick. FVQA: Fact-based Visual Question Answering. TPAMI

–Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. Visual question answering: A survey of methods and datasets. CVIU

–Damien Teney, Qi Wu, Anton van den Hengel. Visual Question Answering: A Tutorial. IEEE Signal Processing Magazine.

–Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, Ian Reid. Visual Question Answering with Memory-Augmented Networks. CVPR’18

–Damien Teney, Peter Anderson, XiaodongHe, Anton van den Hengel, Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. CVPR’18

Visual Dialog

–Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, Anton van den Hengel. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. CVPR’18 [oral]

•Visual Question Generation

–Junjie Zhang\*, Qi Wu\*, Chunhua Shen, Jian Zhang, Anton van den Hengel. Asking the Difficult Questions: Goal-Oriented Visual Question Generation via Intermediate Rewards. ECCV’18

–Ehsan Abbasnejad, Qi Wu, Javen Shi, Anton van den Hengell. What's to know? Uncertainty as a Guide to Asking Goal-oriented Questions. CVPR’19

•Referring Expression/Visual Grounding

–Bohan Zhuang\*, Qi Wu\*, Chunhua Shen, Ian Reid, Anton van den Hengel. Parallel Attention: A Unified Framework for Visual Object Discovery through Dialogs and Queries. CVPR’18

–ChaoruiDeng\*, Qi Wu\*, FuyuanHu, Fan Lv, MingkuiTan, QingyaoWu. Visual Grounding via Accumulated Attention.CVPR’18

–Peng Wang, Qi Wu, JieweiCao, ChunhuaShen, LianliGao, Anton van den Hengel. Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks. CVPR’19

•Image-Sentence Matching

–Yan Huang, Qi Wu, Liang Wang. Learning Semantic Concepts and Order for Image and Sentence Matching. CVPR’18

–Yan Huang, Qi Wu, Wei Wang, Liang Wang. Image and Sentence Matching via Semantic Concepts and Order Learning. IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI),

•Language-guided Navigation

–Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Snderhauf, Ian Reid, Stephen Gould, Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR’18

•Visual Relationship Detection

–Bohan Zhuang\*, Qi Wu\*, Ian Reid, Chunhua Shen, Anton van den Hengel. HCVRD: a benchmark for large-scale Human-CenteredVisual Relationship Detection. AAAI’18

第4周汇报

——张溢炉

1. 《Deep Reinforcement Learning for NLP》总结
2. 内容摘要

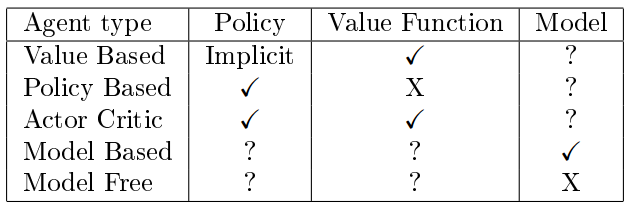
本文是对Association for Computational Linguistics一次会议的简单介绍，介绍了DRL在NLP上的应用。

1. 总结及理解
2. 最近DRL在NLP的生成、对话和信息抽取方面用的比较多
3. DRL（Deep Reinforcement Learning）在NLP应用中最大的挑战是模型的设计。
4. DRL中几个关键问题：动作空间，状态空间和奖励功能的高效实用设计；探索和利用之间的权衡；将语言结构纳入DRL
5. 启发

本文系统、全面地介绍了DRL在NLP上应用的相关作者和文章，具有很好的指导作用，可以根据自己需要展开阅读。

1. 《Deep Reinforcement Learning：A brief survey》总结
2. 内容摘要

本文首先简单介绍了DRL。然后介绍了DRL的一些方法及优化策略。

1. 总结及理解
2. 主要算法：deep Q-network (DQN)（学习玩游戏）, trust region policy optimization (TRPO), and asynchronous advantage actor critic。
3. 以前RL的缺陷：缺乏可拓展性和只能在低维应用。这是由存储复杂性，计算复杂性，样本复杂性造成的。具有强大函数近似和表示学习功能的DL出现，给这些问题提供了新的工具。DL的函数近似功能可以用于建立环境的仿真模型，可以避免机器人在真实环境中训练，大大减小的机械磨损。DL的表示学习功能可以将高维降成低维，如可以将高维的图像通过卷积神经网络降为低维，使机器人通过摄像头感知真实环境成为了可能。DRL在一些游戏中的表现已经超过人类，如AlphaGo。
4. RL可以分为如下表的几类，其中，Model-based和model-free是指agent是否有模型，模型是环境的动态和奖励的数学描述。partially observable和fully observable的区别：前者是指agent的observation只是部分状态，后者是指agent的observation包括所有的状态，其中observation是指agent通过探索观察到的状态
5. RL通过与环境交互，像人一样在试错中成长，更新自己的对环境的认知，通过奖励函数来调整自己的动作，然后形成动作集，一系列动作组成策略。当关于环境的观察满足马尔可夫性质（转换矩阵分布只与当前状态和当前的动作有关）时，RL问题被表述为MDP,其中5个重要因素是{动作集A，状态集S，状态转换分布P，贴现因子，奖励函数R}。常用回报公式如下，其中r为每步的回报，为贴现因子，取（0,1]，随着时间流逝，回报的系数会越来越小，这使得决策过程中会尽量用最少时间来完成任务。因为当其等于1时且状态无限多时，会出现回报的计算值超出计算机的计数范围，所以一般小于1。

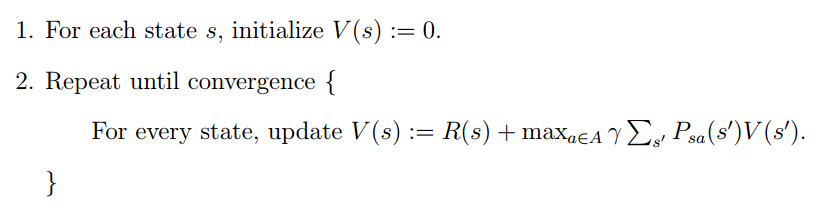
与环境交互过程中，RL会通过奖励函数不断优化状态转换分布P，从而获得最优策略，使期望回报最大化。可见，奖励函数是RL的关键。

1. MDP中的基本思想是基于下一个状态的回报作为当前状态决策的依据，这需要遍历整个环境才能知道所有状态，似乎有些不切实际。而partially observable MDPs(POMDPs)可以解决这个问题，它状态转换分布依赖于当前状态和之前的动作，其中用到了RNN。
2. RL面临的挑战：

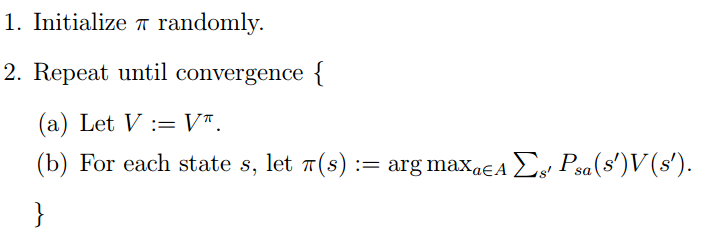
* 最优的策略必须通过与环境的交互、试错之后获得，agent收到的信息只有回报。
* agent的观察依赖于它的动作，采取不同的动作会看到不同的环境，具有强时间相关性。
* agent必须处理长时间依赖，当前的动作造成的后果经过一系列状态转换之后才显现出来。

1. 解决有限状态MDP问题的方法Dynamic programming：一般，策略迭代比较适合低维情况，但现实生活中，很多环境都是高维的，因而值迭代用的比较多。

* 值迭代：迭代步骤和优化各状态动作概率公式（Q\_learning）如下，其中Y为当前状态下的期望回报，是当前状态动作对应策略的概率分布。从式中可以看出，某个状态的回报会影响在它之前两个状态的状态转换分布概率，在点到点的任务中，当agent在地图中随意探索到目的地后，就可以通过回报反向传递，增大对应动作的概率。同时，贴现因子会使agent寻找最短路径。





* 策略迭代：策略迭代步骤如下，内循环重复计算当前策略的值函数，然后使用当前值函数更新策略

1. 文中还介绍了很多方法，自己在文中用红色波浪线做了标注，之后可以根据需要阅读文章给出参考文献深入了解，如：

* Deep Q-network(DQN),trust region policy ptimization(TRPO),and

asynchronous advantage actor critic

* heuristic search algorithm、Exhaustive Search
* partially observable MDPs(POMDPs)
* Bellman equation
* state-action-reward-state-action (SARSA)
* 策略迭代的的更新方法：先通过最小化TD误差方程更新状态转移分布概率，再遍历所有策略获得最优策略？
* 优化状态转移矩阵方法sampling：平均多个策略来设计期望回报Y？
* advantage function
* Policy search：gradient-based or gradient-free optimization
* deterministic approximations、stochastic approximations
* Actor-critic methods
* DRL误差的反向传播

1. 启发

本文让自己对强化学习和深度强化学习有了初步了解。RL可以使agent在未知环境通过学习完成相关任务。DL可以建立环境的仿真模型、将高维降为低维和通过学习近似值函数。在仿真模型中进行强化学习，可以避免机器人与真实环境交互而造成的机械磨损，同时也可添加一些先验知识。将高维降到低维，可以让图像特征更少，去掉冗余特征，让机器人识别真实环境成为可能。RL和监督学习不同的是，监督学习需要通过带标记的数据学习作预测，使其具有一定的泛化能力，但RL可以在未知环境中探索和利用已有知识选择最好动作、作出正确的决定。现实生活中的环境都是连续的，在对环境进行离散化处理之后，会产生无数的状态，这时采用暴力搜索的方法显然是不合理的，所以我们就需要综合agent长短时记忆，借助DRL作出正确决定，在无限状态的环境中更具挑战。总之，由DL和RL组成的DRL具有巨大的潜力，值得深挖。

1. 《DEEP REINFORCEMENT LEARNING: AN OVERVIEW》总结
2. 内容摘要

本文从背景、核心元素、重要机制、应用等等方面对深度强化（DRL）做了全面的介绍，对很多方法没有详细介绍，但给出了一些参考文献。

1. 总结及理解
2. DL可以用于逼近RL的值函数、状态转换分布、策略和环境模型，也常用SGD来优化DRL。DQN和AlphaGo很好的将off-policy, function approximation，bootstrapping结合了起来，避免了不稳定和分歧（divergence）的问题。
3. 探索和利用的矛盾：只利用目前已知的最优选择，可能学不到最优解，收敛到局部最优，而加入探索又降低了学习效率，但有机会找到最优解。在任务中会对两者进行折中，如-贪心算法：以概率进行探索，1-概率进行利用。其优点是直接了当，速度快，劣势是不一定找到最优策略。
4. on-policy 与off-policy的区别：

* on-policy：生成样本的policy（value function）跟网络更新参数时使用的policy（value function）相同。如SARAS算法
* off-policy：生成样本的policy（value function）跟网络更新参数时使用的policy（value function）不同。如Q-learning算法

1. DQN（Deep Q-Learning）是将深度学习deeplearning与强化学习reinforcement learning相结合，实现了从感知到动作的端到端的革命性算法。DQN非常擅长玩游戏。它的衍生版DOUBLE DQN在Atari games中能找到更好的策略。
2. Actor-Critic方法合并了Value-based(比如 Q\_learning) 和 Policy-based (比如 Policy Gradients) 两类强化学习算法。Actor 基于概率选行为, Critic 基于 Actor 的行为评判行为的得分, Actor 根据 Critic 的评分修改选行为的概率。其优缺点如下：

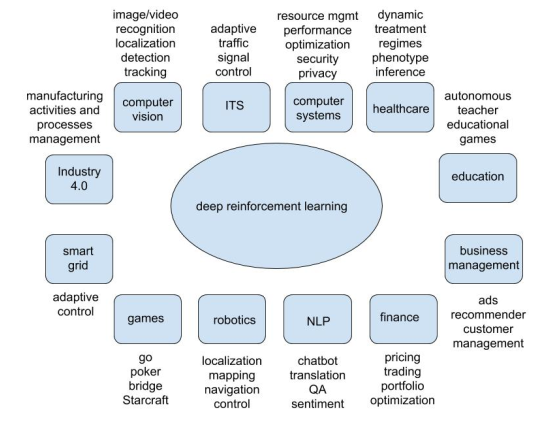
* 优点：可以进行单步更新, 相较于传统的PG回合更新要快.
* 缺点：Actor的行为取决于 Critic 的Value，但是因为 Critic本身就很难收敛和actor一起更新的话就更难收敛了。
* 为了解决收敛问题，Deepmind 提出了 Actor Critic 升级版 Deep Deterministic Policy Gradient，后者融合了 DQN 的优势, 解决了收敛难的问题

1. Deep Q-learning from Demonstrations (DQfD)结合了TD，监督学习和正则化，比DQN具有更好的初始性能，更多的平均奖励，并且学习速度更快
2. Model-free的方法需要agent自己通过采样识别环境，需要花费比较多时间，也可能物理上无法实现，Model-based的方法以数据有效的方式学习价值函数和/或策略，然而，它们可能遭受模型识别的问题，使得估计的模型可能不准确，并且性能受到估计的模型的限制。
3. 知识对于RL的进一步发展至关重要。 知识可以通过值，奖励，策略，模型，探索策略等以各种方式纳入RL。将知识加入RL，RL可以将像人一样拥有一定的常识。
4. DRL的几个重要机制

* attention and memory：注意力是一种关注突出部分的机制。 内存长时间提供数据存储，注意力是一种内存寻址方法。如DNC（Differentiable neural computers），其网站链接为：

<https://deepmind.com/blog/differentiable-neural-computers/>

* unsupervised learning：目前深度学习中的无监督学习主要分为两类，一类是确定型的自编码方法及其改进算法，其目标主要是能够从抽象后的数据中尽量无损地恢复原有数据，一类是概率型的受限波尔兹曼机及其改进算法，其目标主要是使受限玻尔兹曼机达到稳定状态时原数据出现的概率最大。DRL用到的方法有：Horde、unsupervised auxiliary learning、generative adversarial networks
* transfer learning：借助其他领域的知识来辅助学习
* multi-agent reinforcement learning：多智能体系统由一群有自主性的，可互相交互的实体组成，它们共享一个相同的环境，通过感知器感知环境并通过执行器采取行动。多智能体在现实生活中已有应用，如机器人战队，分布式控制和资源管理。
* hierarchical RL：是一种通过多层次的时空抽象来学习，规划和表示知识的方法。 分层RL是一种稀疏奖励和/或长视角问题的方法
* learning to learn：就是让机器能快速的通过学习完成新任务

1. AlphaGo的底层算法有DL、RL、MCTS(Monte Carlo tree search)
2. LEARN TO NAVIGATE可能有机会更换流行的SLAM,方法见下载文献DRL中的《LEARNING TO NAVIGATE IN COMPLEX ENVIRONMENTS》。
3. DRL的应用领域如下图所示
4. 文章最后有DRL相关参考资料，可根据需要查阅
5. 启发

本文让自己对DRL有了初步了解，它与NLP、机器人、多智能体等很多技术的结合展现出很好的发展前景。RL与DL等技术的融合能有效增强算法的学习能力，本文在differentiable neural computer (DNC)模型中使用了controller有似乎与上周MT-EQA模型中的controller存在某种联系，值得之后继续学习一下。给我的直观感觉controller像中央控制器一样，能够通过学习协调不同机制，使模型发挥更强的泛化能力，使模型具备了自我调节的能力。比如可以将controller加入GRU作为gate可能会获得更好的结果。总之，RL与其他技术的结合具备广阔的发展前景。

第5周汇报

——张溢炉

1. 《Multi-Target Embodied Question Answering》总结
2. Background

人工智能的一大挑战是建立智能agent，通过视觉感知周围环境，通过自然语言与人类交流，并在所处环境中行动以完成任务。Embodied Question Answering (EQA)是一个与之相关比较新的任务，它的任务是通过agent观察周围环境的第一视角来回答固定形式的问题，提问的内容一般是对象的属性，如“what color is the car?”，在这之前主要是只包含单个目标的EQA。

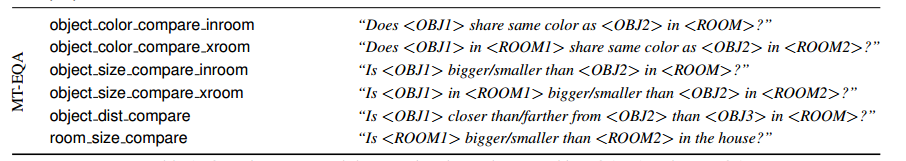
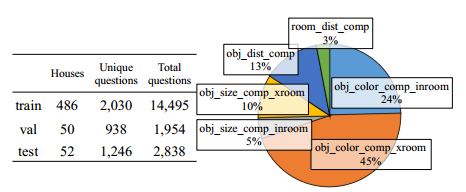
1. Motivation

之前单个目标的EQA会限制agent的能力，故本文提出了Multi-Target EQA(MT-EQA)，在其问题中包含多个目标，可以用来比较相同房间或不同房间中的两个目标的大小、颜色以及距离，如“Is the dresser in the bedroom bigger than the oven in the kitchen?”。

1. 3D environments

文中使用的仿真环境是House3D，另外还有ViZDoom，TorchCraft，DeepMind Lab，HoME，MINOS，Gibson，AI2THOR。

1. Dataset

文中所用数据集问题类型及各类型所占比例分别如下图，其中包括了6种类型的问题。数据集包含588个环境中的19287个问题，涉及8个房间中的61种对象。在设计集过程中用到一个比较重要的指标Intersection-Over Union (IOU)来选择观察一个对象的最佳位置。

1. Model

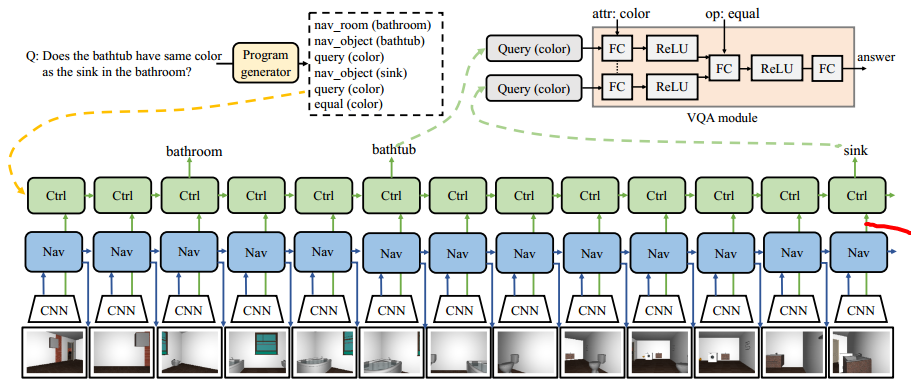
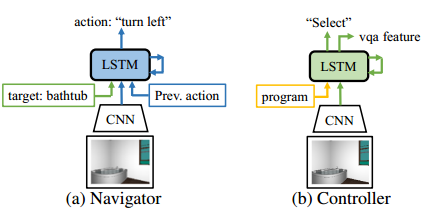
模型如下图，其采用《Visual question answering A survey of methods and datasets》

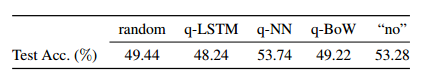
中提到的组合模型（Compositional models）的方法，包括question-to-program generator, navigator, controller, VQA module四个模块，program generator为下图的左上部分，VQA module为下图的右上部分，navigator, controller在图中为缩写形式。

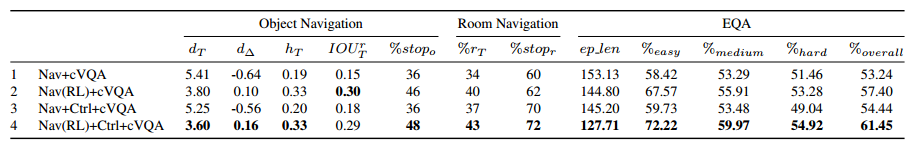
由图可以发现program generator和VQA module作为输入、输出模块相对独立，program generator将question分解为几个子程序，VQA module则用于比较controller两个输出属性产生answer。navigator, controller模块详图分别如图（a）和图（b）。controller比较当前第一视角图像（经CNN处理后）与子程序中所要求图像进行比对，判断是否到达目标房间或找到目标，逐步完成子程序所要求的任务，并给navigator模块提供当前子程序的目标。Navigator模块则根据controller输出的target、当前第一视角图像（经CNN处理后）以及上一个动作产生相应的动作，寻找controller给定target。

过程中，寻找房间和物体的过程是不一样的，房间是通过12张全景图来分辨的，而物体是通过语义信息来分辨的，CNN是预训练好的。另外，navigator的动作为左转30度、右转30度和前进。作者发现较EQA-v1转角的增大，减少的动作的次数，使得agent更容易训练。

对于整个模型的训练过程，作者没有细说，只是简单提到可以使用了模仿学习（Imitation Learning IL)或强化学习（RL）

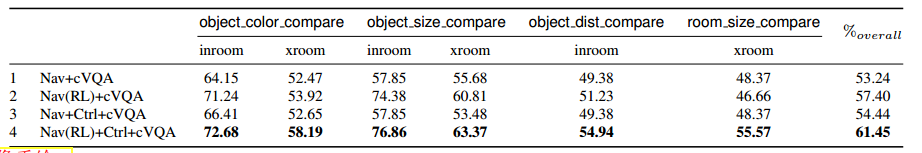


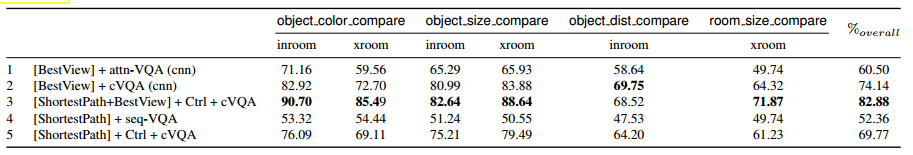
1. Experiments
2. Baselines如下图所示，它是由以下5种问答模型直接输入数据集的question得到answer的准确率
3. 针对不同模块组合的对象/房间导航和EQA准确性实验如下表所示。其中除、ep\_len以外，其他指标越大越好。由图可知，RL训练方法和Ctrl模块都有助于性能的提升。需要行动距离越短（越简单）的问题，准确率越高。



1. Oracle Comparisons如下两表所示

* 比较第二张表1、2行可以得出，本文的VQA模型优于attention-based VQA
* 比较第二张表2、3行可以得出，controller模块提取的图像特征更有助于准确率的提高
* 比较第二张表3、5行，第一张表4行与第二张表3行可以得出，controller模块作出的决定会显著影响准确率





1. Others
2. CNN的预训练是怎么进行->可阅读文献[8]
3. 怎么用IL和RL来训练网络->可日后学会IL和RL再回头思考
4. 准确率提高1-2个百分点会不会是数据集等其他因素引起而并非模型的改善？
5. 对于不同问题、不同初始位置是否都需要重新训练？
6. 各模块推导公式及优化方法，Controller和navigator模块更具体的结构、模块个数，寻找房间和对象的controller和navigator的结构是否一样->这些问题日后决定深入研究再解决
7. 这技术尝试着将图像、NLP与机器人的动作相结合，离落地还有很多问题需要解决，比如在到达目的地后拾取、放下目标物体等一系列操作；实际环境是会经常变化的需要实时三维建图；真实环境中会有很多干扰，如光照、摄像头畸变、测量误差等；机器人本体的直立行走是很难做的，目前波斯顿动力的几款机器人在这方面有了很大突破，可以过独木桥、跳跃、后空翻等，但是体型比较大，还需小型化，另外也可以采用轮式的，但不能上下楼梯等；真实运用对实时性要求也比较高，不能让主人等太久；大规模应用还得考虑成本问题；得把AI和传统技术起来，如AI用于问答和导航，拾取物品使用传统技术，目前来说机器人导航和定位主要用GPS、SLAM等。
8. 《Knowledge Acquisition for Visual Question Answering via Iterative Querying》总结
9. Background

到目前为止，基于深度学习的模型主导了标准的VQA基准。在这些模型中，使用CNN对图像进行编码，使用LSTM对单词进行编码是最受欢迎的方法之一。 此外，许多表现良好的模型采用了注意力机制来获得更好的结果。 最近Jabri等人提出了一种新的替代模型，即一种双层MLP，它将答案作为输入并进行二元预测。 与其他更复杂的架构相比，这个简单的网络展示出了较好的结果。

1. Motivation

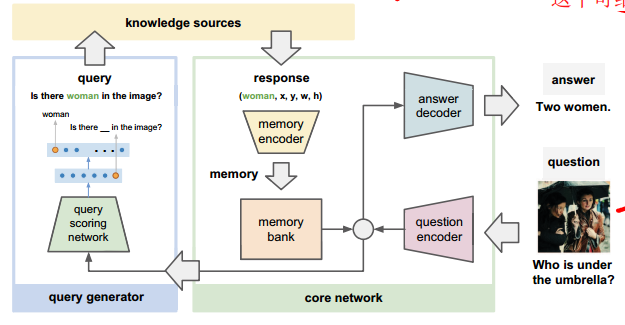
为了弥补之前方法没有充分利用知识库和基于图片的信息的缺陷，提出了利用问题产生模块通过多次迭代询问的方式从已有知识获取答案的线索的方法，本文中已有知识是指来自人类标注或算法生成的数据源的视觉信息

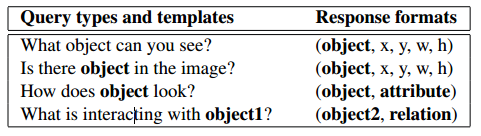
1. Dataset

本文用到的两个数据集是Visual7W、VQA Real Multiple Choice

1. Model

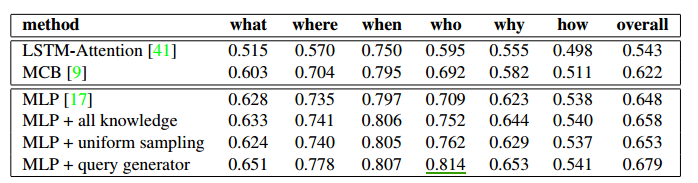
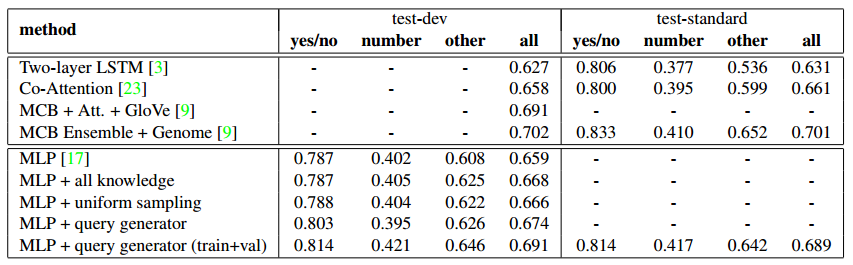
模型如下图所示，又Core Network、Query Generator、knowledge sources三个模块组成，模型输入图片和问题，输出答案。模型采用与EM相似方法来训练，将Core Network与Query Generator逐次循环训练，其中Query Generator使用Monte-Carlo方法来训练。模型的优点是结构简单，可以获取外部知识，可以展示模型的迭代推理过程。缺点是模型也可能被外部知识误导。



* Core Network
* 对图像-问题对进行编码
* 对knowledge sources信号进行编码并将其存入memory bank，其采用堆栈存储
* 由memory bank信息和image-question编码信息产生Query Generator的输入
* 由answer encoder对memory bank信息和image-question编码信息进行译码产生answer
* Query Generator：综合图像、问题和memory bank产生query到knowledge sources寻找信息，其中由两层softmax层来选择query types和object，如下图所示。
* knowledge sources：知识来自人类标注或算法生成的数据源的视觉信息，输入query返回一定形式的response，如下图所示。

1. Experiments

visual7W和VQA数据集上实验结果如下图所示，由图可以看出，在visual7W和VQA问题query generator的加入对MLP有优化的作用。在visual7W中对who问题的提升最明显，总体提升3个百分点。VQA实验结果与目前最好模型结果相当。



1. Others

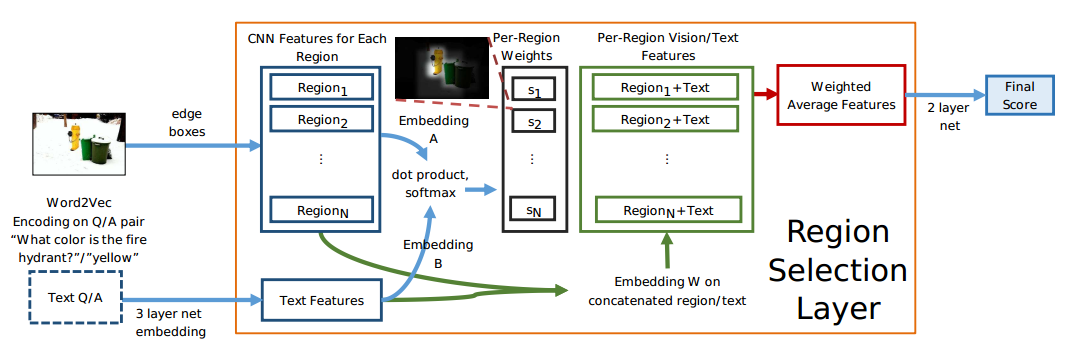
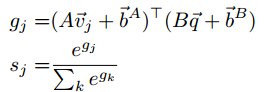
* 可不可以尝试把图像输入knowledge sources模块，将knowledge sources模块改成另一个问答模块，相当于由两个人来完成同一个问题
* 有空阅读文献3、9、17、23了解其他VQA模型
* 在训练模型时能否将语义部分打印出来，方便更好理解模型
* 模型的设计有点类似于人类的思考方式，如本文模型通过借助外部知识来提升性能

1. 《Where To Look: Focus Regions for Visual Question Answering》总结
2. Motivation

本文主要是为了使模型能通过问题、回答和图像来选择图像中对于回答问题比较重要的部分

1. Model

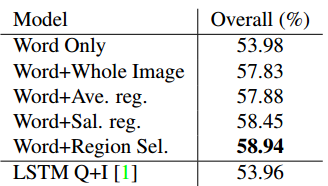
模型如下图所示，输入图像、问题和答案，输出是分数。其中，图像通过ImageNet预训练好的CNN网络来提取特征，使用Stanford Parser方法表示问题，再与答案的表示组成15000的向量，这种方法可以将不一样长度的句子变换成固定长度。训练时发现，低初始值可以防止选择用的softmax过早飙升以及高维视觉向量过早占主导地位。



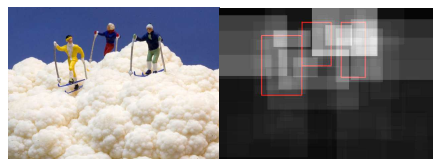
将两个向量串联得W

1. Experiments

* 实验结果如下表，比较第一行和第五行可以发现，相对纯文本问答，图像信息的加入会有5个百分点的提升，提升较大。第二行为简单将图像特征与问题-答案特征串联起来替代选择环节的方法，比较第二行和第五行发现作者方法只有1个百分点的提升，说明选择环节对性能改善效果并不是很明显。



* 作者后面用图像展示模型根据问题-答案选择图像对应部分的能力，如下图，其中问题是“Are the people real?”，红框是经过训练后权重较大部分。综合文中其它几个图片可以发现，该模型能将问题-答案与图像对应起来，有点类似于文本问答的attention机制。



1. Others

* 文章的图9说明了权重越大图像区域对准确率影响越大，也说明了权重排在前K的图像区域对准确率具有决定性作用，而且选择合适K值准确率可能高于选择全部图像区域
* 组合2种类型向量时可以考虑用内积或串联的方法

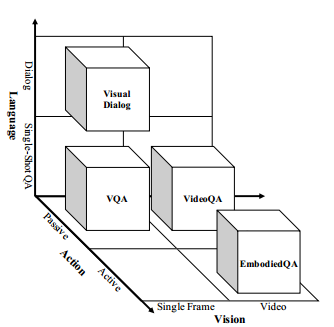
第6周汇报

——张溢炉

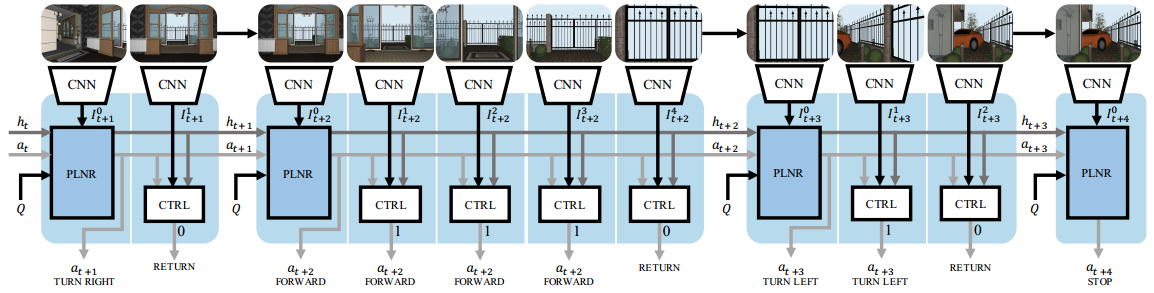
1. 《Embodied Question Answering》总结
2. 内容摘要

文章提出的EQA在视觉、NLP的基础上加了action，相对VQA而言，EQA任务还需通过探索在房间中找到目标物体，然后回答主人的问题。如果再加上拿放物体等动作，可以应用到家庭服务机器人。

1. 总结及理解

* EQA融合了几乎所有AI技术，如CV、NLP、IL、RL、知识图谱等，目的是使agent具有像人一样的感知、理解、决策的能力，最终应该会与传统的定位、机械手等技术相结合构成人形机器人，但这似乎还很遥远，涉及了太多技术，这里充满了机遇和挑战，需要团队合作进行整合各方面技术。
* 这里面涉及很多技术，一个人做的话工作量太大，需要团队合作；另外，3D环境、CV、NLP、RL、IL等整合的计算量很大，目前VQA需要的运算时间已经挺长了（在1080上训练简单模型遍历一遍VQA数据集需要2小时左右，复杂的需要5、6个小时），那么，如果训练集很大，EQA的训练会很长
* 对于解决实际问题，有些场景可能使用传统的slam等技术能更好的解决问题，如机器人可能在某个特定环境下，room不会改变，只会改变一些摆设，room navigator时可以尝试使用传统技术来导航，导航到对应房间之后再采用AI技术。但是科研就是需要不断尝试新的方法，即使它目前效果并不是那么好。
* 下图直观的展示了visual dialog、VQA、videoQA、EQA之间的关系，videoQA、EQA正上方是空白的，可看成是Video Dialog、Embodied Dialog，有一定的挖掘价值。
* 可以通过多任务的预训练来共享权重，使一个模型具有读取多种隐性特征的能力，如文中CNN的预训练
* 模型如下图，其中PLNR模块产生动作，动作集是{前进，左转（9度），右转（9度），停止}，CTRL只决定是否继续执行当前动作。另外question经过编码后输如，answer模块通过最后5帧图像和问题产生答案。

它的训练过程如下：先通过模仿学习预训练PLAR、CTRL和answer模块，再使用强化学习的方法微调PLAR、CTRL模块。



1. 《Interactions in Multiagent Systems: Fairness, Social Optimality and Individual Rationality》总结
2. 内容摘要

本文主要讲述了多智能体，内容较多，自己浏览了一遍，还未仔细看文中公式，对多智能体有了初步的了解。

1. 总结及理解

* 类似于人类决策的3个假设：智能体会根据自身经验进行行推理论证，根据不同环境作出有利于自身的决定；智能体也会综合其他智能体的利益来作决定，就像人遵从个人利益、集体利益和社会利益那样；另外，本文还利用了认知心理学的知识，如人们会对对自己友善的人友善。
* 在多智能体系统中会面临利益冲突、信息有限和回报等问题，多智能体环境可分为合作环境和竞争环境。在合作环境下，智能体注重集体利益和社会利益，在竞争环境下智能体注重个体利益。在多智能体系统中需要均衡Fairness, Social Optimality和Individual Rationality。
* 纳什平衡（Nash equilibrium）：为了使利益最大化，应该在对手出不同策略的时候我们的收益都相等（不然在游戏中，对方可以改变不同策略的概率让我们的期望收入减少）。这种方法的一个好处是不需要沟通，可以自己做决策。纯纳什平衡有时并不能公平对待每个agent，如果要获得fairness，得采用混合纳什平衡，给不同策略赋予一定的概率。
* 传统的由superagent作决策的方法有以下几个缺点：
* 因为计算量会随着agent的增加而不断增大，不利于扩展
* 单个agent都必须与superagent通讯，通讯花费大，也可能相互干扰
* 一旦superagent出现故障，会造成整个系统崩溃
* 文中主要以两个agent的游戏来介绍算法，涉及只玩一次和不断玩下去的情况。玩一次的游戏意味着大家只有一次机会，大家都会比较在乎结果，容易引发矛盾；而不断玩下去的游戏，大家对一次游戏的结果并不是那么在乎，可以寻找一个策略使得大家总的回报差不多。
* 第3章加入了很多人类的情绪在模型中，比如厌恶不公平现象、人不犯我我不犯人和人若犯我我必犯人的思想、惩罚违背集体利益的人等，感觉将想法转换成数学表达是一个有趣的过程，自己会在之后的学习中好好体会，比如调整下系数改变对某个方面的关注程度。
* 公平的策略有利于解决各个agent 之间利益冲突的问题，从而增进合作
* 第4章讲述合作环境中的social optimality，我的理解是将共同利益最大化。各个agent有着共同的利益和相同的回报函数，个体回报的增加会促使集体回报的增加，但有时个体回报的增加会导致其他agent回报的损失，从而导致集体回报的减少，所以单个agent在行动时除了考虑自身回报，还得考虑集体回报。
* 在合作环境中，更偏向于合作，主要考虑其中的fairness、social optimality；在竞争环境中，更偏向于竞争，主要考虑其中的individual rationality、social optimality。这两者之间也没有明显的界限，有时也可能相互转变。如当前的中美贸易战，从之前更偏向于合作的合作关系演变成互相伤害的竞争关系。
* 文章主要以两者游戏的方式来讲述这些方法，多智能体系统也可以通过随机配对的方式简化为两个智能体的问题。其中主要运用到强化学习、博弈论、认知心理学等方法，可以通过改变不同策略或action的概率来改变回报。
* 以后继续看需要注意的方面
* 如何设计多智能体系统环境并以数学模型来表示？是否类似于强化学习的隐马尔科夫模型？
* 如何表示及实现Fairness, Social Optimality和Individual Rationality均衡？通过调节公式中不同项的系数？
* 如何学习策略？
* 如何实现智能体之间有效的沟通？减小通讯的信息量？如何实现各智能体独立作出决策？如：未来自动驾驶技术及各种无人驾驶的落地，如果各智能体间需要通讯才能作出决策，那么很可能出现通讯混乱的情况，最好各智能体能独立作出决策，在决策过程中也需考虑Fairness, Social Optimality和Individual Rationality等因素。

1. 《A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input》总结
2. 总结及理解

* 此篇论文是《deep learning-based models have dominated standard VQA benchmarks》中提到的标准VQA基准测试之一，自己粗略读了一下，了解一下以前的方法，有需要再细读
* 模型如下图所示，上半部single-world approach是之前的方法，其在图像分割S只产生一种对图像的理解W，单一的理解势必在回答问题时会有些局限。于是作者提出了下半部的muti-world approach。这两个方法将当前NLP和CV结合起来，是打破客观（CV）与主观（自然语言）的一种尝试。图像分割会提取图像中的object及其属性、位置等信息。根据x、y、z三维坐标可以判断各个object之间的位置关系，如y轴方向为从左到右，那么y值较小的物体在左边。
* VQA的问题和答案具有主观性，不同的人可能有不同的问题、答案及表示方式，而图像具有客观性，但现实中图像可能会有干扰和被遮挡的情况，而且图像一般是二维的。这给VQA任务带来了很多挑战。如
* 同一个物体可能有多个名称，这个问题可以利用词向量来解决，在词向量空间中词义相近的词的坐标相距比较近
* 图像存在被遮挡和截断的情况，这不利于物体识别
* 人类回答某个问题的句子可能比较长，这会增加产生答案的难度
* 物体的一些状态比较难检测，如门是否开着
* 疑问
* 产生问题、图像和答案表示的具体模型？
* 怎么获得图像中物体的三维坐标？

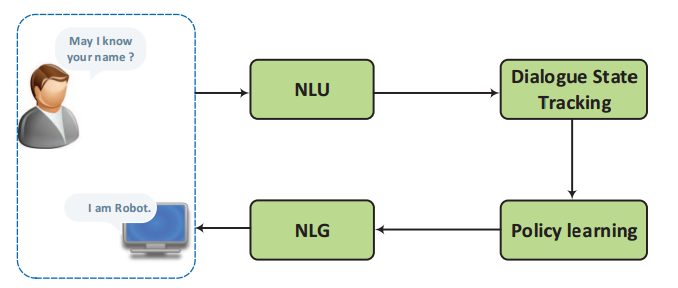
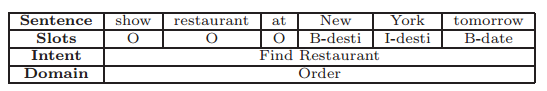
第7周汇报

——张溢炉

1. 《A Survey on Dialogue Systems:Recent Advances and New Frontiers》总结
2. 内容摘要

本文概述了对话系统最新发展，提到了一些主要论文。

1. 总结及理解

* 对话系统可分任务型和非任务型。任务型有pipeline（如下图）和端到端的方法，非任务型有生成式、检索式和混合式。
* Pipeline方法简介
* 自然语言理解（NLU）需标记语句中各词的slots标记，对语句的intent、domain进行分类，如下表所示
* 对话状态追踪（Dialog state tracking）：它管理每个回合的输入以及对话历史并输出当前对话状态。之前是采用规则、条件随机场、最大熵模型、网络式排名等方法，最近引入了深度学习的模型
* 对话策略学习（Policy learning）：根据当前状态采取下一步动作。其可以使用监督学习或强化学习来训练。
* 自然对话生成（NLG）：根据选择的动作生成答案
* 对话是多轮问答，且多轮是围绕某个topic展开的，多轮之间相互联系

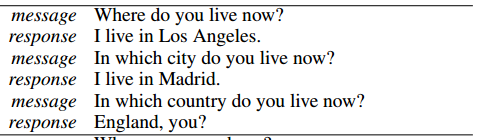
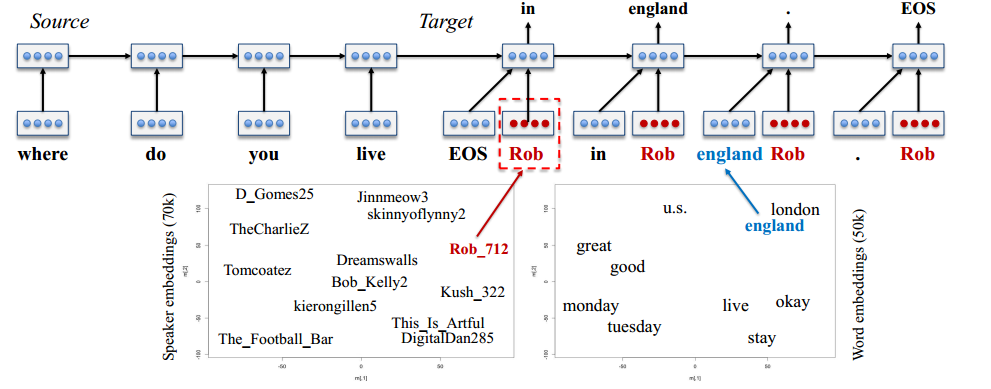
1. 启发

文章使我初步了解了对话系统，打算阅读一些文章后，复现一些经典论文。

1. 《A Persona-Based Neural Conversation Model》总结
2. 内容摘要

本文主要介绍了Speaker Model和Speaker-Addressee Model两种模型，Speaker Model能使agent具有一定的特征和说话风格，Speaker-Addressee Model能是agent能像人一样对不同的人有不同的态度。

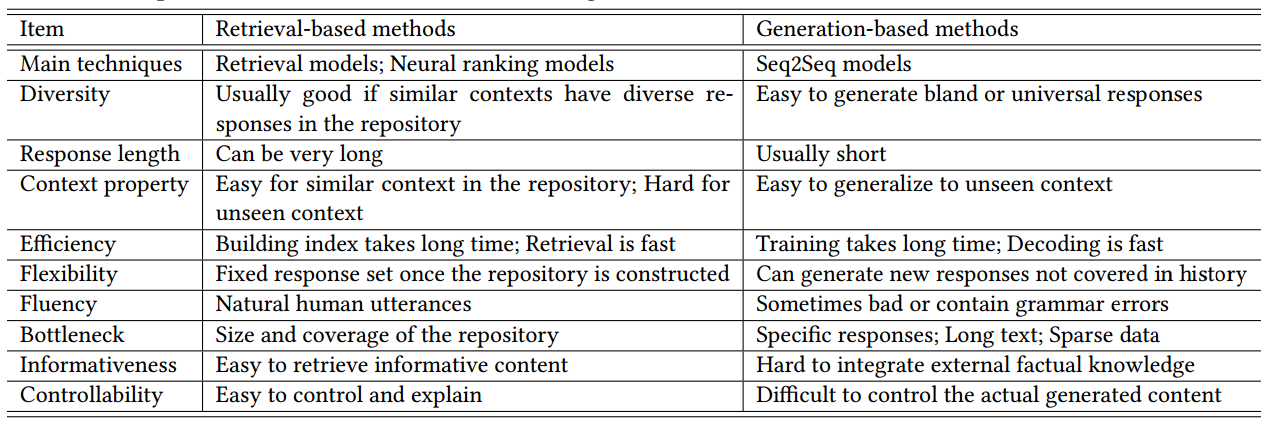
1. 总结及理解

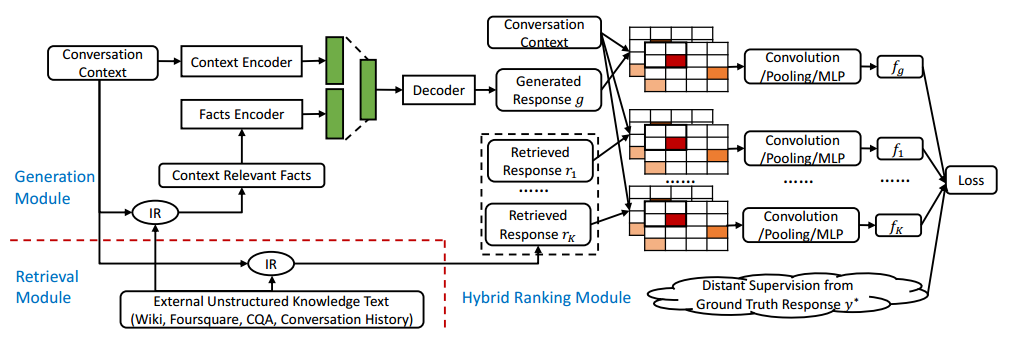
* 动机：文中模型旨在解决一致性的问题，将任务特征，如身份、年龄等作为输入，使模型具备一定的风格，从而对问题的回答更具一致性，避免出现以下现象
* 作者的相关工作主要是NLG上，与上一篇文章提到的一样，NLG由基于规则的生成转向端到端的多层神经网络。因为人工成本过高、时间花费也多，一些数据集的生成、标记也都向着自动生成发展
* Speaker Model模型如下图所示，模型的亮点是将speaker嵌入到同一个向量空间，向量接近的speaker具有相似的说话风格。模型中也将speaker向量加入到解码过程，使模型的回答具有speaker的风格，相当于将回答限制于与speaker有关的内容，使模型不再基于整个数据集对相似问题产生截然不同的回答：一会出生于美国，一会出生于英国
* 数据集
* Twitter Persona Dataset：用于Speaker Model
* Twitter Sordoni Dataset：用于(non-persona)LSTM model，在BLEU的评价中，具有10条参考答案的分数一般比具有1条参考答案的分数高
* Television Series Transcripts：用于Speaker-Addressee Model
* seq2seq常用评价指标或目标函数：perplexity、BLEU、maximum likelihood (MLE) 、maximum mutual information (MMI)

1. 《A Hybrid Retrieval-Generation Neural Conversation Model》总结
2. 内容摘要

本文介绍了一个混合生成式和检索式的模型

1. 总结及理解

* 动机：为了结合基于检索和基于生成的两种的优点，其中，检索式的回答具有多样性、信息丰富和流畅，但是它的回答会局限于历史数据库，缺乏灵活性。生成式能生成新的连贯的回答，具有灵活性，但是它的回答可能是与问题的背景无关的一些通用简单回答，如“I don’t know”，有时也可能有语法错误，下图是两者较详细对比图。文中的混合式模型能结合两者的优点。
* 模型如下图所示，模型的前半部分相当于生成式模型和检索式模型的并行网络，和以前论文模型相似，没有什么创新。该模型的主要创新点是后半部分排序模块，从生成式模型和检索式模型产生的答案中选择最好的答案，采用了distant supervision的方法来训练网络。作者在最后提到了将来可能会用强化学习作排序，我在想是否可以将生成式和检索式做更深度融合，而不是简单并行后从所有答案中做选择。



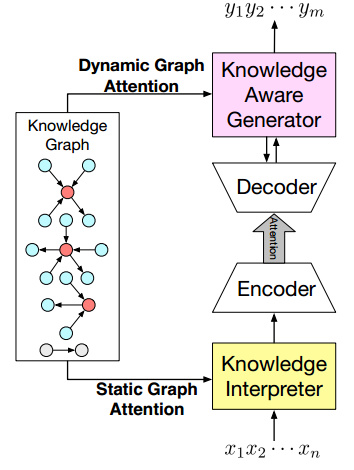
* 文中提到了很多生成式、检索式以及将强化学习与之结合的文章，后面可以选择性阅读。
* 数据集：Twitter（文献8）、Foursquare（<https://foursquare.com/>）
* 评估方法：
* BLEU和ROUGE-L：用于评估模型给出答案与参考答案的相似度
* 人工评测：这种评测成本比较高而且花时间多，但是接近于图灵测试。方法是人在不知道是哪个模型的情况下，比较两个模型的Appropriateness和Informativeness，为了更客观，可选择多个评估者。
* Distinct-1和Distinct-2：用于测试回答的信息量和多样性。基于上下文、事实和外部知识库的检索式模型具有较大信息量和多样性，而生成式则会产生一些能在很多场景中试用的回答，如“I don’t know”。

1. 《Commonsense Knowledge Aware Conversation Generation with Graph Attention》总结
2. 内容摘要

* 本文介绍了一种利用常识来优化生成式对话的方法，提出了CCM（commonsense knowledge aware conversational model）模型，其中使用了多个attention机制。

1. 总结及理解

* 动机：为了充分利用外部知识，解决之前模型存在的2个问题：1）高度依赖于非结构化文本的质量，或受到小规模、特定领域知识的限制；2）通常分开、独立地利用知识三元组（实体）。
* seq2seq模型GRU、LSTM模型虽然具有一定记忆功能，但它终究是一堆数学公式，没有精确记忆。对话并不像翻译，两种语言之间存在明显的映射关系，相同语义的词会在数据集的输入输出反复出现，看起来更有可能训练出具体数学公式，而各种对话问题和答案没有规则的对应关系，似乎很难形成一个从上下文到答案的具有数学关系的映射。又因为场景和常识的缺乏，seq2seq产生的回答一般不能让人满意。我觉得对话系统应该使用知识图谱等知识库实现准确的记忆，利用attention、LSTM等模型来捕获潜在规律，从知识库检索相关内容，利用NLG（可加入RL）来组织语言生成答案。
* 模型如下，模型使用了外部知识图谱，在输入、输出两次分别使用static graph attention和dynamic graph attention引入外部知识图谱，这点和上一篇文章使用的IR方法有所不同。模型中在编码器和译码器之间也加入了一个attention来改善模型。作者提供了比较详细的公式推导并开源了源码（https://github.com/tuxchow/ccm），之后会考虑复现论文，深入理解各个attention所发挥的作用。



* 数据集
* 知识库：ConceptNet（<https://conceptnet.io>）
* 对话：

<https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/>

<http://coai.cs.tsinghua.edu.cn/hml/dataset/#commonsense>

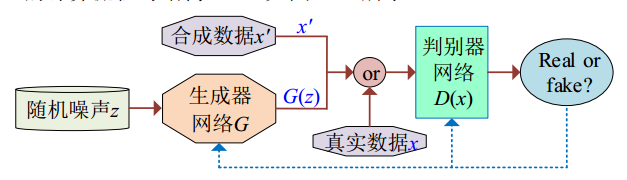
* 评估指标：
* 自动：perplexity
* 人工：与上一篇论文相似

1. 《Adversarial Learning for Neural Dialogue Generation》总结
2. 内容摘要

本文介绍了使用对抗学习来优化生成式对话的模型

1. 总结及理解

* 动机：作者的灵感来自于图灵测试，文中判别器就是模仿一个评价者的行为，模型目的是让判别者分不清生成的回答与人类的回答，解决一般生成器生成回答枯燥、通用、重复、浅显等问题
* 书中没有给出模型图，按自己理解文中模型与下图来自（来自程乐峰《机器学习在能源与电力系统领域的应用和展望》）相似，GAN模型的思想与actor-critic RL model相似。其中，随机噪声Z对应输入的上下文，G(z)对应生成器输出回答。真实数据x为数据集中真实回答。Real or fake对应奖励。文中生成器和判别器是预训练好的，然后通过奖励来微调它们的参数、是使用强化学习中的policy gradient方法来训练。



* 对抗学习在图像领域获得了巨大的成功，而因为自然语言是离散的，判别器的输出很难反向传播。文中提到几篇解决这个问题的文章，如将生成器的隐藏层向量作为判别器的输入、使用强化学习来训练。
* 结合文意和自己理解，模型把生成的一个词作为一个状态，生成词的操作可看成一个action，生成一个回答作为一个episode，也是一个策略。因为给回答中所有词相同的的奖励和分别给每个词奖励都有些不妥，所以文中提出了2中REGS（reward for every generation step）奖励策略
* 判别器模型有：SVM+Unigram、Concat Neural、Hierarchical Neural、SVM+Neural+multil-features。文中使用的是Hierarchical Neural，它是端到端的模型。
* 数据集：OpenSubtitles
* 评估指标：AdverSuc、machine-vs-random及人工评价。
* GAN和RL在生成领域具有很大的发展潜力，能使其从监督学习走向自我探索，从而具备更强的学习能力。其中，面临着一些挑战：
* 如何评价生成结果？人类语言具有主观性，回答的好坏很难量化。人工评测需花费很多时间，而且也具有主观性。那么，设计自动评价模型可能会更好，这也存在一些问题，如何评价模型的好坏？判别器和评价器是否需要不同的设计？这些问题都很难量化，相信通过大量测试能得到一个比较客观自动评估方法。
* 模型训练过程中会出现不稳定的现象，较难收敛。文中也提出了好几种训练的技巧。

第8周汇报

——张溢炉

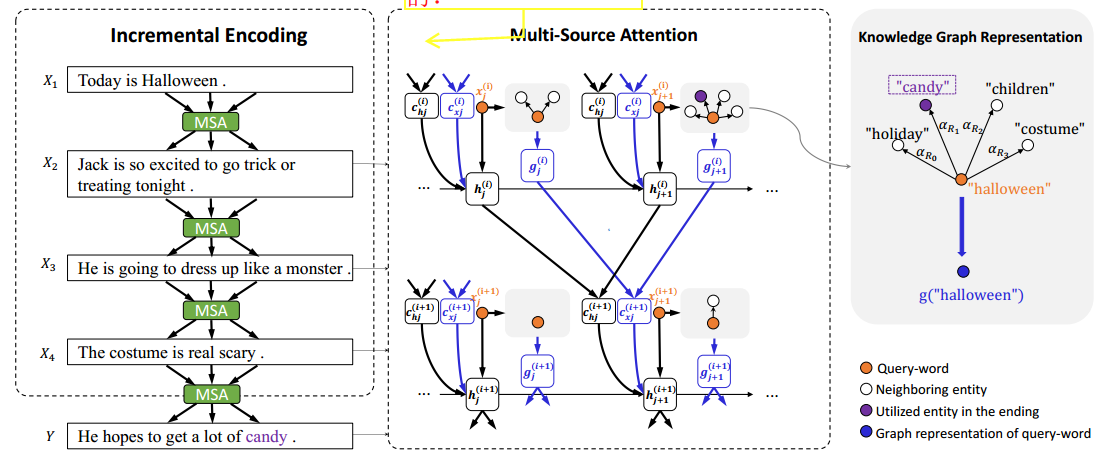
1. 《Story Ending Generation with Incremental Encoding and Commonsense

Knowledge》总结

1. 内容摘要

本文提出了一种自动生成故事结尾的模型。

1. 总结及理解

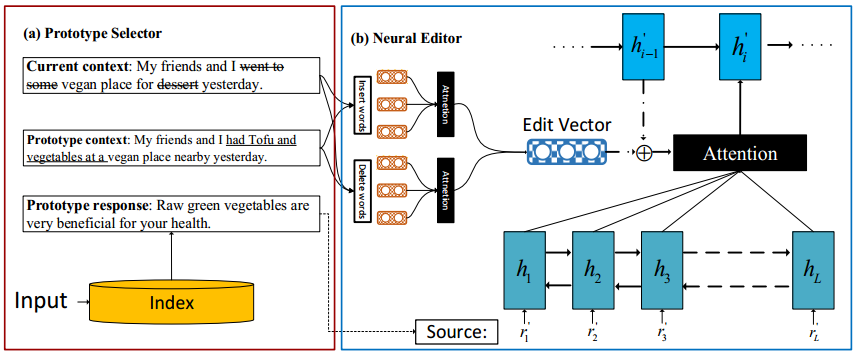
* 动机：通过分析，作者认为文章的结尾主要与文章中的线索和常识有关，线索指各时间和实体之间的上下文逻辑、因果关系。于是，作者分别设计了增量式编码方案和多头注意力机制来利用线索和常识。
* 模型如下图所示，故事和结尾的句子数是固定，分别是4句和1句。模型以句为单位进行编码，注意h由(当前词的词向量)使用LSTM编码而成，使用注意力机制存储上一层隐层h信息，使用注意力机制存储上一层隐层三元组g信息，三元组信息是通过当前词搜索知识库所得。
* 数据集
* 语料集：4句话的故事和1句话的结尾
* 知识库ConceptNet：16652条三元组关系，45种关系，关系种类有点少
* 词向量：6位小数的200维Glove词向量
* 复现代码总结
* 读代码时发现：其不能体现2个相同h、t实体间的多种关系，只会选取数据集中的第一种关系
* 语料集：训练集有90000个故事，验证集和测试集都是4081个故事
* 词向量（GloVe.6B）：下载网址: <https://github.com/stanfordnlp/GloVe>。网址中提供训练词向量的源码，可根据需要训练词向量。本文使用的是200维向量，742个词。
* CPU跑完一个epoach用了4个小时左右，GTX 1070花了不到9分钟左右
* 需要对句子进行填充，统一句子长度，可引入掩膜使填充位置乘0消除其对输出的影响
* 训练PPL（6）与测试PPL（9）相差较大，我觉得这应该是一个欠拟合的问题，要减小两者差可以尝试增加模型的复杂度（深度和宽度等）、增大训练集。其中可能原因是：任务的输出关于输出的函数极其复杂，或者说不存在能准确表达输出与输入关系的数学表达式，这与人类语言表达的多样性、随机性、主观性等有关，故不可能训练出能精确捕获复杂语言任务关系的模型
* 通过观察生成结尾及与参考结尾对比发现：生成结尾语句通顺，语法错误较少，与上下文相关，但是抓不住重点；生成结尾与参考结尾相差较大，可能因为人的思维和生成都具有随机性，还可能和数据集有关，数据集是由多个人一起收集的，这使得模型更难捕获数据集的思维（规律）
* 首先要设计好代码的整体框架；然后，要注意数据的预处理，要将文本转换成词向量，并且要与模型的输入格式对应，这其中有很多细节性问题，很可能出现一些不易察觉的逻辑错误，如将词与词向量的映射搞错。编码过程中最好分段调试，将数据打印出来观察是否出错；最后就是模型的搭建，模型常使用类来表示，复杂的模型包含很多公式，也很容易出错，这个可以利用tensorflow可视化功能，将模型以图的形式展现出来。

1. 《Response Generation by Context-aware Prototype Editing》总结
2. 内容摘要

本文提出一种可以利用检索到的原型上下文及其与输入上下文差异、原型回答来生成回答的模型

1. 总结及理解

* 动机：利用检索到的原型具有良好的语法和信息量为生成提供了良好的起点，后期编辑过程进一步提高了原型的相关性和连贯性
* 生成句子追求目标：与上下文相关性强、流畅、语法错误少、多样性、原创性，原创性是本文提出的
* 模型如下图，模型的大致思路是首先检索出与当前输入上下文相关的原型上下文，然后根据上下文差异（当前上下文没有而原型有的词（insert words）、当前上下文有而原型没有的词（delete words））通过注意力机制得到edit向量，最后将edit向量和原型回答串联作为解码器输入，输出回答。其中，测试时的检索方式与训练时的检索方式是不同的，测试时是利用Lucene方法检索上下文相似的原型；而训练时是利用回答的相似性来检索的。



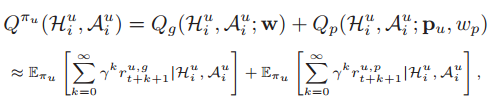
* 模型的亮点是利用了已有的数据集作为参考，提高了生成句子的质量

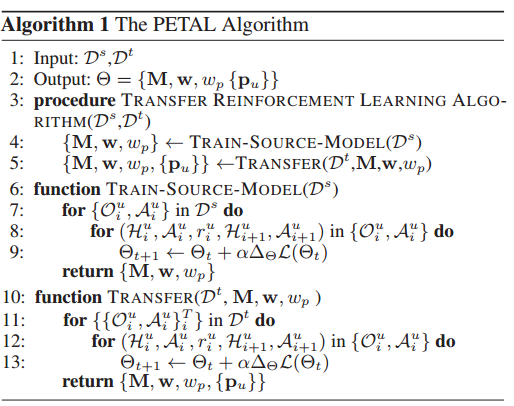
1. 《Personalizing a Dialogue System with Transfer Reinforcement Learning》总结
2. 内容摘要

本文提出了迁移强化学习的方法使对话agent的更具个性化，可以对常客做到私人订制，减少了对话轮次，有利于提升客户体验。

1. 总结及理解

* 动机：利用迁移学习可以解决关于某个人的数据少的问题，同时也解决了之前迁移忽略个体差异带来的负面影响；利用端到端的方法解决基于规则学习需手动设置规则和特征的麻烦
* 模型是基于POMDP的思想创建的，包含7个元素{S,A,O,P,R,Z,}。S为隐层不可观测的状态；A为agent生成的回答；O为顾客所说；P为状态转移矩阵；R为奖励；Z为观察函数；为贴现因子，使对话更短。个性化的Q-function如下式：其中，第一项代表所有顾客的偏好，第二项代表个人偏好，所有b为由历史信息{A，O}经矩阵M仿射而得信念状态。参数M，W，由所有顾客的数据训练而得，在本文实验中参数个数为85000；参数由某个顾客数据训练而得，在本文实验中参数个数为100。利用迁移学习解决了个人数据匮乏的问题，也利用了个体之间的差异



* 本文采用State-Action-Reward-State-Action(SARSA)方法来优化参数，步骤如下图，先用所有顾客数据训练M，W，，再用某个人数据来微调M，W，，。
* 本文实验real-world dataset，实验结果明显优于其他baseline
* 疑问：
* 如何编码实现？

第9周汇报

**这周生病了几天，头比较痛，没时间看论文，只看了一些基础性的资料。**

第10周汇报

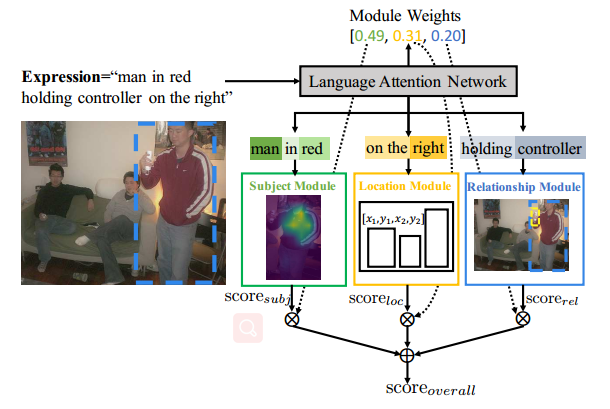
——张溢炉

1. 《MAttNet: Modular Attention Network for Referring Expression Comprehension》总结
2. 内容摘要

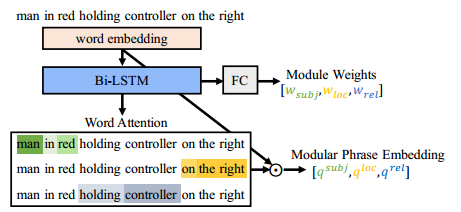
本文提出了一种多模块的模型来解决指代表述理解问题。

1. 总结及理解

* 动机：把指代表述和图片分成主体、位置、关系三个部分能够更好的捕获语言和图片中的信息
* Referring Expression Comprehension曾用模型：CNN-LSTM based frameworks、joint vision-language embedding frameworks、混合前两种模型的方法。
* 模型如下图所示，输入是图片和图中某个物体的指代表述，输出是图中物体与指代表述的匹配分数。Language Attention Network将指代表述分成subject、location、relationship三个部分，subject用于捕获主体，包括边界框内所有物体，location主要用于捕获主体位置，relationship用于捕获主体与边界框外物体间关系。Visual Modules用于提出图片中特征并得到与指代表述各个部分的匹配分数，最后分别乘module权重得到总分



* Language Attention Network：模型如下图，模型输出module weights用于最后求取匹配总分用，modular phrase embedding输出给visual module作匹配用



* 在两者关系或匹配分数不能直接算出时，可以考虑把它俩同时作为神经网络输入，通过训练来求得这种隐性关系，如文中用MLP求取匹配分数
* 物体的7种属性：类名、颜色、尺寸、绝对位置、相对位置、相关物体和通用属性

1. 《Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks》总结
2. 内容摘要

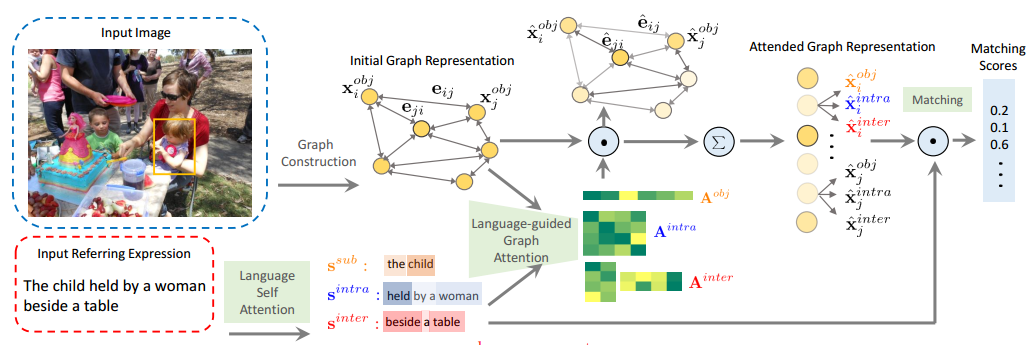
本文提出了一种用于理解指代表述的基于语言指导的图注意力模型，上一篇是本文的baseline，本文模型与上一篇模型有些相似与不同之处

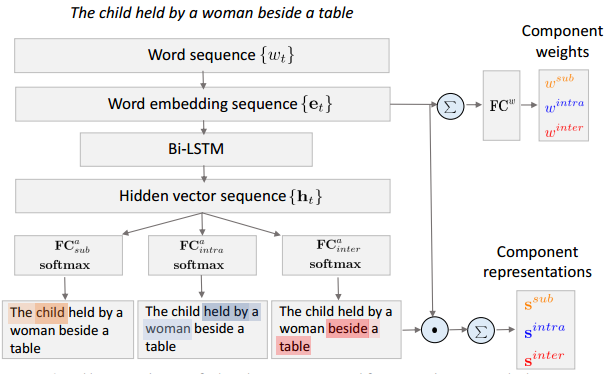
1. 总结及理解

* 动机：以前模型都未综合利用图片中物体的属性及物体之间关系，也没利用好图片与描述语言之间的对应关系。从本文模型能直观的体现出利用了这些关系
* 模型如下图所示，输入是图片和图中某个物体的指代表述，输出是图中物体与指代表述的匹配分数。图中的物体是使用预训练好的VGG16识别出来的。

模型包括语言自注意力机制、语言指导图注意力机制、匹配模块三部分。

* 语言自注意力机制将指代表述分成主体、类内关系、类间关系三部分
* 语言指导图注意力机制根据语言自注意力机制输出的三部分，使与指代表述对应的通过VGG16提取出输入图片中的物体以及类内、类间关系图部分高亮
* 匹配模块就是将经处理后的图像信息与指代表述信息使用点乘匹配，得到各个物体的匹配分数



* Language Self-Attention Module：模型如下图
* 动机：为了弥补LSTM等简单编码方式未充分利用结构信息，该模块将语句信息划分成了三部分，充分利用物体属性及类内、类间关系
* 疑问：该模块是如何训练的？是否能捕获大多数指代表述中的这三种关系？
* Language-guided Graph Attention Module：先通过卷积处理的方法得到物体的属性V和位置信息l。然后使用Language Self-Attention Module输出s\_sub，s\_intra和s\_inter得到图中各部分对应注意力值。最后将原图与注意力值相乘得到加权后的图结构
* Matching Module：计算图结构与指代表述的匹配分数
* 与上一篇baseline的相同与区别
* 两篇文中都利用模块化思想来利用主体属性、位置、物间关系
* 上一篇把指代表述分为了主体、位置、关系三部分，而本文将指代表述分为了主体、类内关系、类间关系，而把位置直接作为物体的特征，与卷积网络提取的特征串联，从结果上看，本文方法优于上一篇
* 从模型上看，上一篇在visual module中，只有提取主体特征时利用了指代表述来指导提取，而本文中三个部分都利用了指代表述来指导提取

1. 《VQA: Visual Question Answering》总结
2. 内容摘要

本文2015一篇比较系统阐述VQA任务的文章，发布在arXiv上的原文包含更多内容。提供了VQA数据集并作了较全面的分析

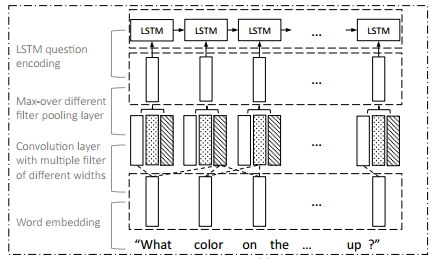
1. 总结及理解

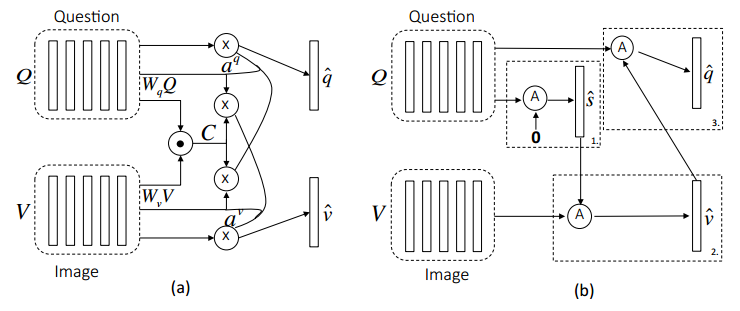
* 动机：相对于图像描述，VQA可以更好的理解图片，结合外部知识回答相关问题。VQA可以帮助视觉障碍人士和小孩更好的理解真实世界
* 数据集包括Real Images dataset（MS COCO dataset）和abstract scenes dataset，其中包括对应的开放性和多选项问题-回答对，以及对图片的描述。下图展示了问题的种类，通过前四个词来判断问题种类，从图中可看出各种问题的大概分布以及问题的多样性，而1-3个词的回答在Real Images dataset（MS COCO dataset）和abstract scenes dataset中分别为89.32%, 6.91%, 2.74%和90.51%, 5.89%, 2.49%
* 文中展示了一些简单的模型的测试结果作为baseline，准确率较低，还有很大的提升空间
* 疑问：
* 怎么让模型发现自己回答错了？用户反馈或自我判断？
* 发现之后如何调整参数？反向传播或者使用强化学习？

1. 《Hierarchical Question-Image Co-Attention for Visual Question Answering》总结
2. 内容摘要

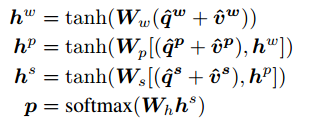
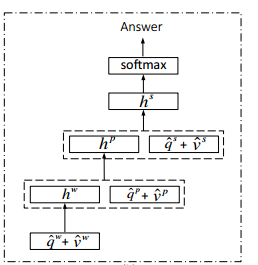
对于VQA任务，本文提出了Hierarchical Question-Image Co-Attention，对图片、问题都采用了注意力机制

1. 总结及理解

* 动机：以前模型只采用了图片注意力机制，为了更有效利用问题，本文模型不仅使用了图片注意力，还使用了问题注意力机制
* 模型输入图片和问题，输出为训练、验证集中TOP1000的回答，覆盖了86.54%的回答，可看成一个分类问题，模型可分成如下三部分
* 模型将问题分为词级、短语级、句子级三种情况，三种情况经过co-attention独立计算结果q^、v^。词级用词向量表示，短语级采用如下图1-D的CNN（unigram、bigram、trigram）模型编码而得，句子级用LSTM编码而得。短语级模型设计比较新颖，但是从消融实验结果来看，其对性能提升影响不大
* co-attention：其有两种方案（Parallel co-attention mechanism，Alternating co-attention mechanism），分别如下图所示，在本文实验中，前者表现优于后者。具体公式看原文，（b）公式中，先将q、0赋给X、g得到s^，再将v、s^赋给X、g得到v^,最后将q、v^赋给X、g得到q^



* 最后通过如下模型和公式综合计算三种问题分级情况得到分类结果



* 数据集：VQA dataset、COCO-QA
* 待解决问题：
* 学习VGG、Resnet等卷积神经网络

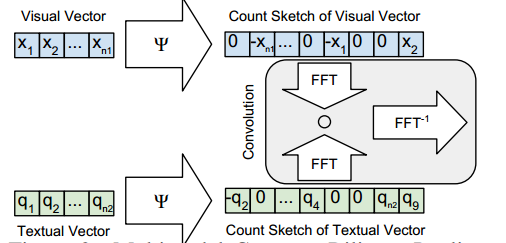
第11周汇报

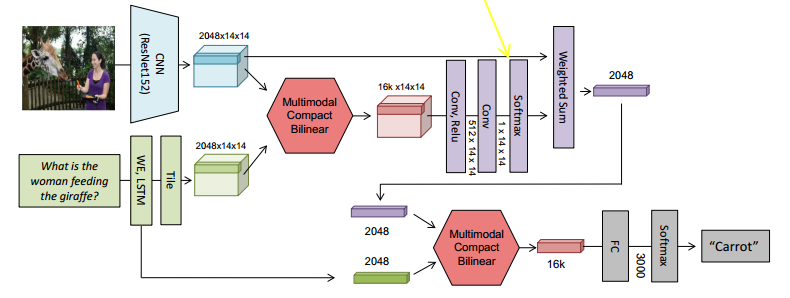
——张溢炉

1. 《Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding》总结
2. 内容摘要

本文介绍了一种能有效结合两种模态表示向量的方法：MCB

1. 总结及理解

* 动机：组合两种模态表示向量的方法：内积、和、串联、外积，本文假设外积更具表达性，但高维时计算外积不可行，参数太多，故本文提出Multimodal Compact Bilinear pooling (MCB)方法来组合多模态表示向量，此方法更有效、更具表达性
* MCB模型如下图下图所示，输入为图片表示向量和语言表示向量，输出为两者外积的count sketch，解决了外积在高维时计算量大、不可行的问题。其中FFT表示傅里叶变换，这里涉及卷积和傅里叶变换关系的公式：x0 ∗ q0=FFT−1(FFT(x0) FFT(q0))。其中Count Sketch projection function Ψ和相关证明见文章文献Charikaret al., 2002和Pham and Pagh (2013)。
* 文中提到了应用在VQA任务上的注意力机制模型，如下图所示。模型将图片划分为14\*14的区域，使用MCB模型有效利用图片和问题得到每个区域的注意值，然后通过加权和得到2048D向量。再将该向量和语言表示向量经过MCB得到16K维向量，最后通过全连接层得到输出。



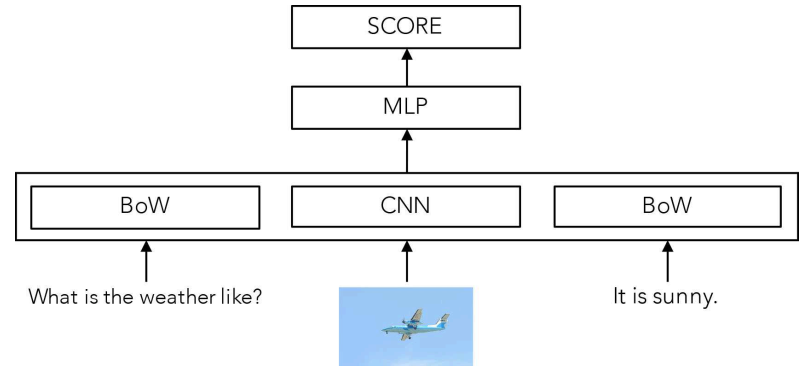
* 从实验结果来看，MCB的表现相对点积、求和串联等方式具有明显的提升，将该方法用到第4篇中能否带来性能提升？

1. 《Revisiting Visual Question Answering Baselines》总结
2. 内容摘要

本文提出了一种判断图片-问题-回答三元组是否正确的模型，用于多选项的VQA任务。文中对产生的错误结果进行了分析，发现当前VQA模型的一个关键问题在于缺乏对问题和答案中出现的概念的视觉基础。另外，当前VQA模型并没有明显优于简单利用数据集偏差的模型

1. 总结及理解

* 模型及对应公式如下图所示，是一个二分类模型，可见模型十分简单。输入为图片-问题-回答三元组，三种特征采用串联得形式组合，其中图片特征使用预训练好的卷积网络提取，问题和回答的特征分别取所有词向量和的平均值。输出为{正确，错误}，输出层采用sigmoid函数σ(x) = 1/(1 + exp(−x))来计算二分类结果

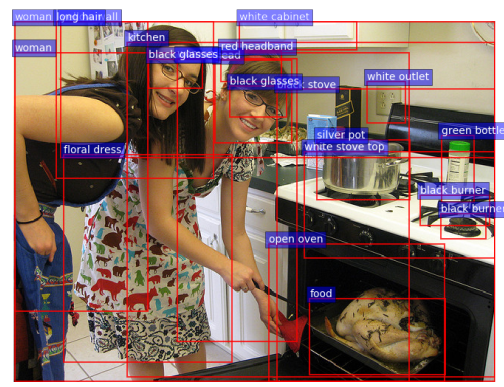
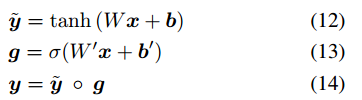


* 本文设计了相关实验对模型表现好坏的原因进行了分析，得出如下结论：
* 将问题作为输入能捕获不同回答相似性
* 不同卷积网络对结果影响较小
* 使用循环神经网络编码问题和回答在不同数据集影响不一样，可能是因为其在较小数据集上容易过拟合
* 迁移学习后微调参数能提升模型
* 文中通过对比输入为A+Q和A+Q+I时的结果，分析了图片的加入对颜色、形状、计数、空间推理、动作、因果关系六种问题影响，从结果来看，本文模型并不能很好理解图片信息
* 疑问
* 模型真的能学会推理吗?能理解where等意思吗？以后在复现论文时注意观察测试结果，观察会不会存在一些不太正常现象，如where的问题回答为数字

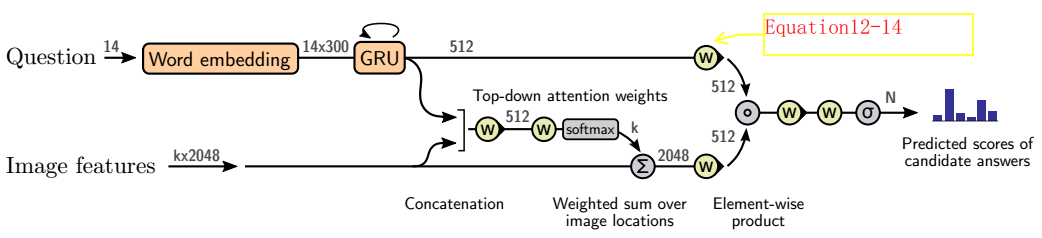
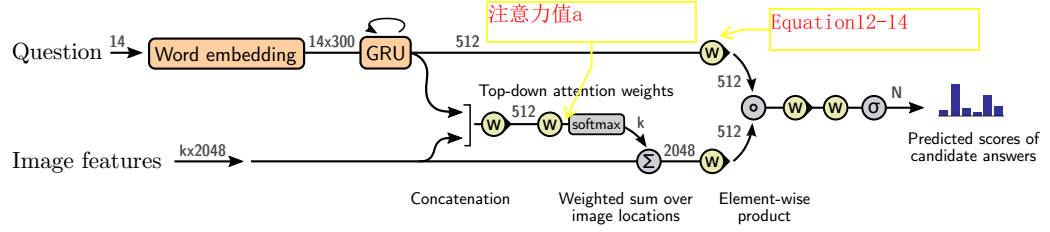
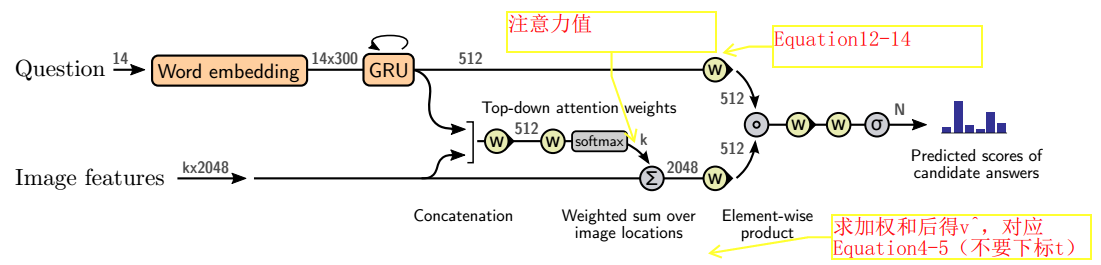
1. 《Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering》总结
2. 内容摘要

本文是2017年VQA比赛中获得冠军的模型，其结合了Bottom-Up Attention和Top-Down Attention，前者通过预训练好CNN（本文用Faster R-CNN）提取图片中的物体和显著区域，分别用相同维度特征向量表示，后者用于确定各个特征向量的权重。它们在图像描述和VQA任务中都有很好表现

1. 总结及理解

* Faster R-CNN bottom-up attention模型包括两部分：Region Proposal Network (RPN)和region of interest (RoI)池化层。第一层提取object proposals，然后通过一个intersection-over-union (IoU)阈值得到RoI输出给第二层，第二层通过池化得到固定尺寸的特征。模型的结构和实现还需阅读文中提到的相关文献进一步理解，模型的输出结果如下图所示，输出包含了物体类别和属性
* VQA模型如下图所示，输出的回答是在训练集中出现过8次以上的。模型多次使用了‘gated tanh’ layers，对应图中带尾巴的W。具体公式如下

另外，将回答作为简单分类，不能很好理解答案语义，能否将回答的词向量以合适的方式加入到模型中提升性能？



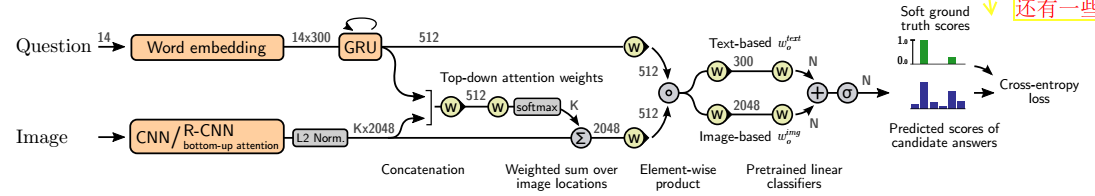
* 默认一般是列向量，加T转置变成行向量。外积：将两向量所有元素分别相乘得矩阵，注意和几何叉积的区别；element-wise product：两张量对应元素相乘；串联各个向量；累加：将所有向量对应元素相加；element-wise sum：两向量对应元素相加

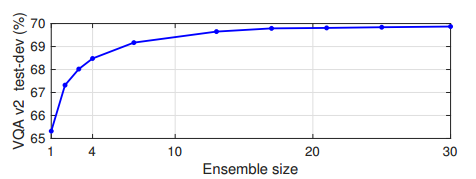
1. 《Tips and Tricks for Visual Question Answering:Learnings from the 2017 Challenge》总结
2. 内容摘要

本文讲述的是2017年VQA比赛的冠军模型，里面提到了很多模型设计、训练测试的一些重要细节，作了大量的对比实验并作了分析，值得以后自己在实验时回过头来细读

1. 总结及理解

* 作者在选择模型和超参数时花费了3000GPU-hours，这说明VQA研究需要花费大量时间和精力，从文后实验可见作者做了大量的对比实验来论证自己观点
* 模型如下图所示，模型中的top-down attention也叫image attention，是典型的question-guided attention，自己对question-guided attention的理解是：question特征和image特征通过一定计算后得到注意力值，然后使用这个注意力值作用于image 特征，因而称为question-guided image attention，同样地，可以将此理解拓展到其他注意力机制的命名。从图可以看出，本文模型与上一篇模型大体一致，只是在模型最后变成两个text-based和image-based分支对应元素相加的形式。整体模型看起来比较简单，但是在比赛中取得的效果是最好的，作者把它归功于模型结构和超参数的选择。包括以下几个方面：
* 用sigmoid output替代常用的softmax output
* 对参考答案选择了柔性化评分，不是一个问题只有一个正确的答案，而是一个问题可以对应1个或多个评分为[0,1]的答案，这也像现实生活中的一些问题，不同的人有不同的解答
* ‘gated tanh’激活函数，从实验结果来看，相对于ReLU、tanh和gated ReLU有0.6-1.5个百分点的提升
* 用bottom-up attention提取图像特征，相对ResNet-200 7×7 features，有3.8个百分点的提升
* 使用预训练好的参数赋给，具体方式还需结合代码进一步理解
* 选择了较大的小批量训练数据，本文选择了一种比较有目的性的打乱方式：让相同问题对应的不同图片和不同答案不要出现在同一个小批量训练数据中。作者目的是为了避免在一次迭代中梯度下降往两个不同方向跑



* 从下图可以看出集成（ensemble）对性能提升的显著影响，本文使用了相同模型，只是采用了不同的初始化参数，最后将各个模型预测分数相加作为预测依据