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

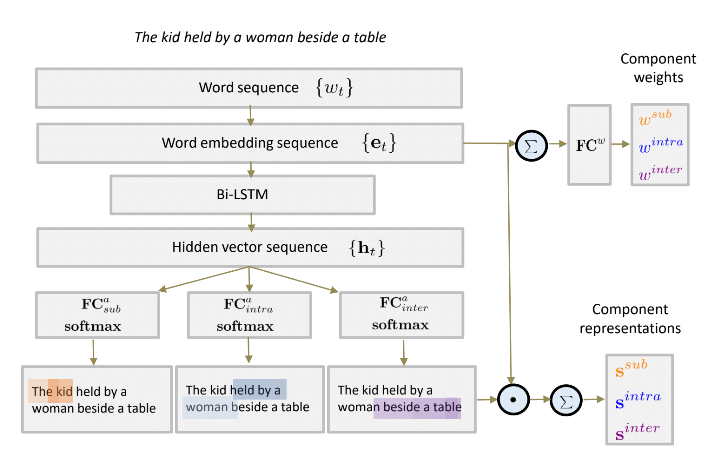
吴琦和虞立成两位老师的报告主要讲述将图像处理、自然语言处理和语音处理等人工智能技术与机器人有机结合起来的最新科研成果，将多模态与机器人相结合，可以使机器人具备像人一样的感知、推理和决策的能力，可以通过人机交互使机器人动作、完成相关任务。报告也介绍了与之相关的图像问答（VQA）、图像描述、图像对话、VLN（viusal-and-language navigation）、EQA（embodied question answering）等科研成果，并介绍了相关模型、数据集、目标函数等知识。

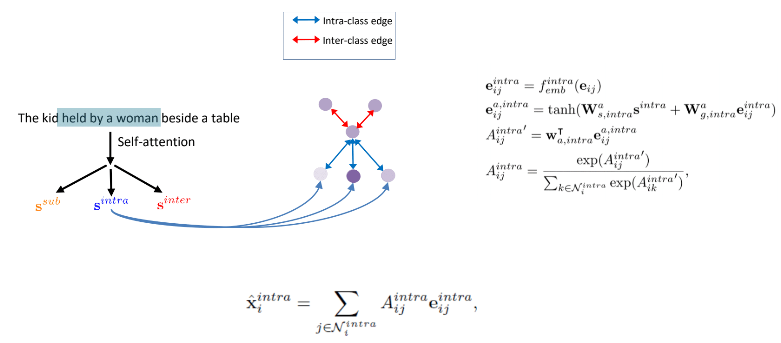
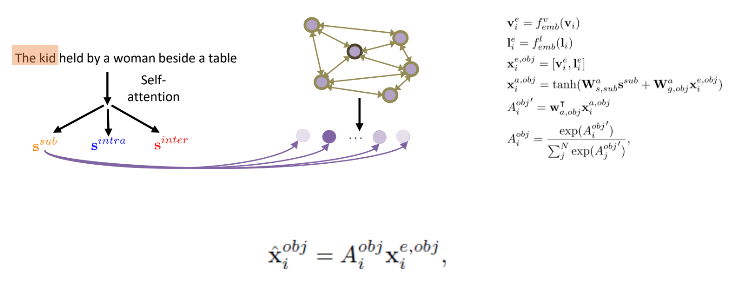
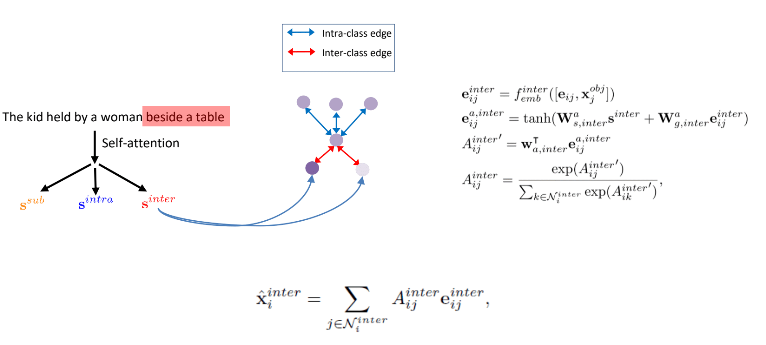
其中，吴琦老师主要讲述了他的三篇论文，虞立成博士从locate、tell、answer、act四个方面介绍了他所做的工作。

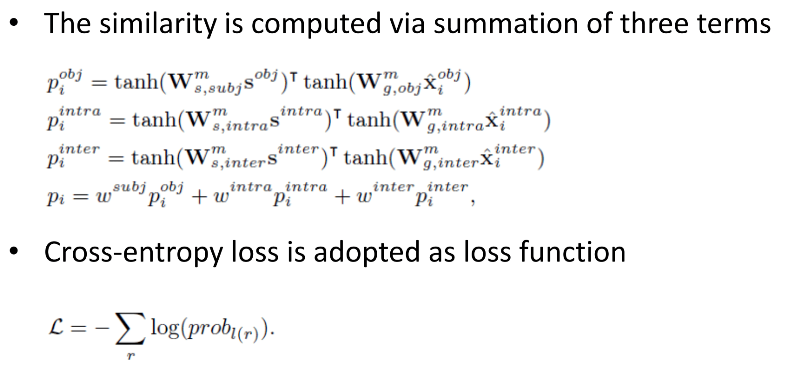
1. 总结和理解
2. 吴琦老师PPT链接：

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

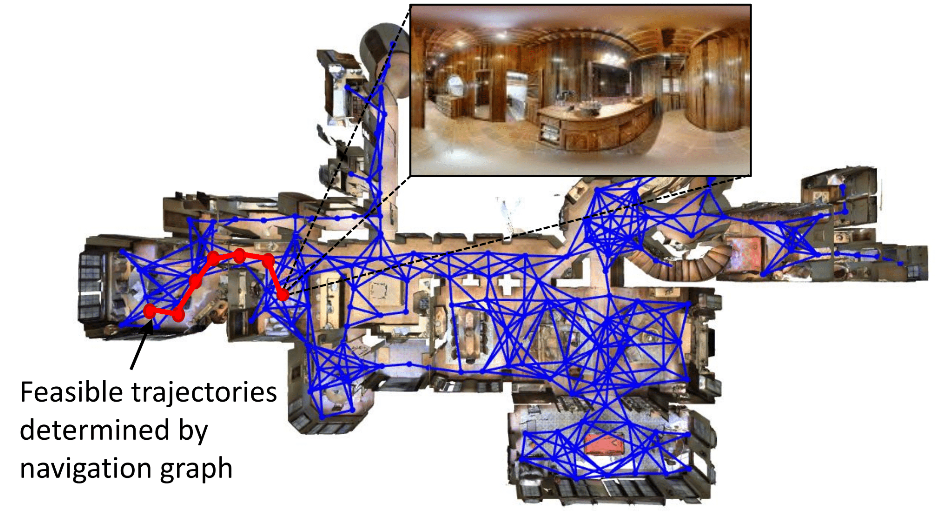
1. 吴琦老师的《Neighbourhood Watch: Referring Expression Comprehension via Language guided Graph Attention Networks》总结：

* 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

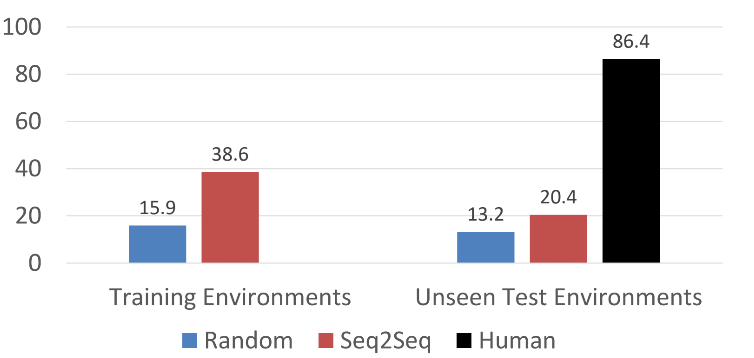


* matching module：计算表达和对象的匹配分数。公式如下：
* 实验
* 任务：模型能根据参考语言表达识别出图片中对应对象
* 数据集：RefCOCO、RefCOCO+、RefCOCOG
* 结果：较之前最好模型有1-4个百分点提升

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

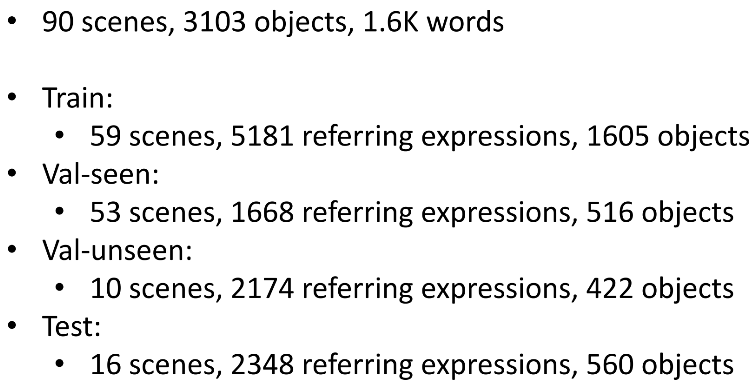
* VLN和VQA的区别：VQA和VLN输入都是图片和句子，但VQA输出是句子，VLN输出是一系列动作
* Matterport3D仿真环境：90个建筑里的10800张全景图，包含194400张RGBD图片，环境复杂多变。
* 路径：路径是通过环境中设定节点连接而成，不能在环境中随意行走，在每个节点上会有一张360度的全景图。
* Room-To-Room实验
* 任务：机器人根据自然语言指导指令（如离开卧室，进到厨房。直走，沙发的地方左转，停在窗口旁）在仿真环境中寻找目的位置
* 评价标准：单次测试运行；机器人必须自己选择停止；如果机器人最后停在了目的位置3m范围内表示成功，是否在3m范围内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

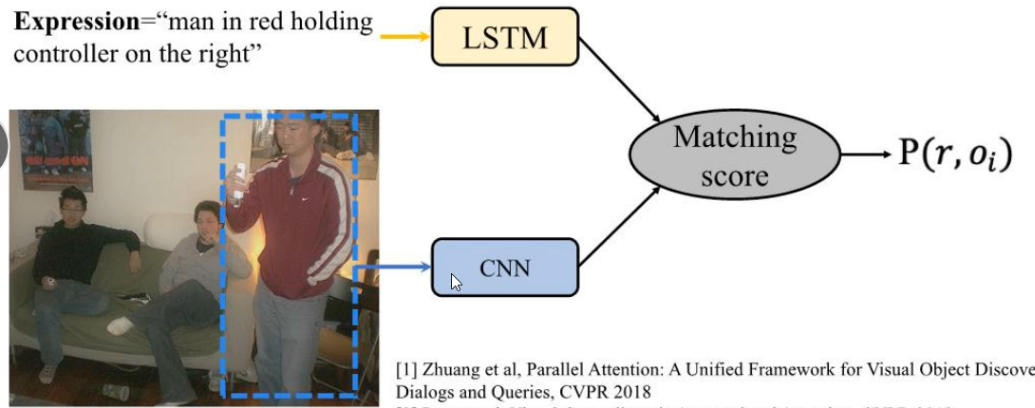
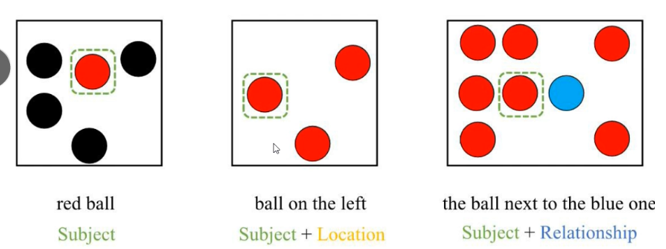
1. 吴琦老师《RERERE：remote embodied referring expression in real indoor environment》总结

* 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

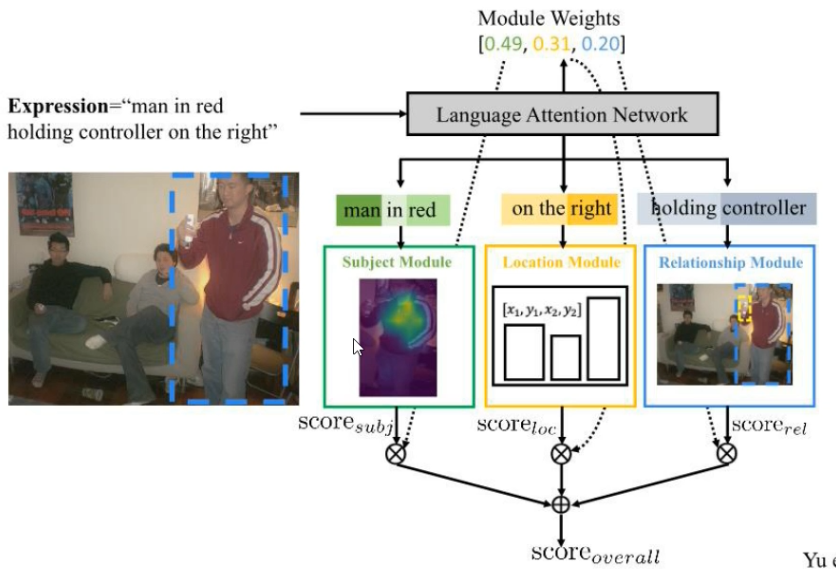


* 实验结果准确率都比较低（20%左右），表明该研究方向还有很大提升空间，其中机器人通过36张图片确定当前位置（联系slam相关知识）

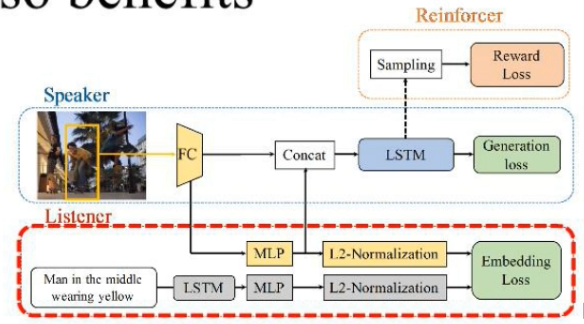
1. 虞立成博士讲座总结

* 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：

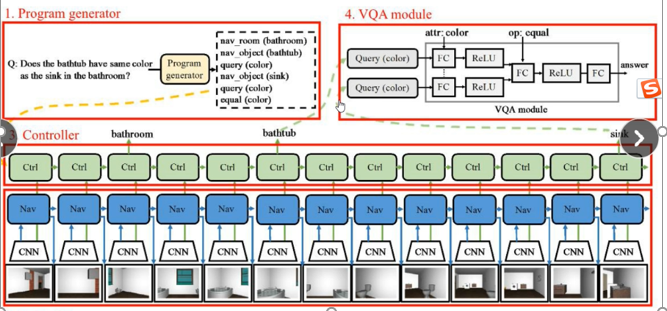
Language parallel模型可能会将误差传下去，导致错误。代码链接：

<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

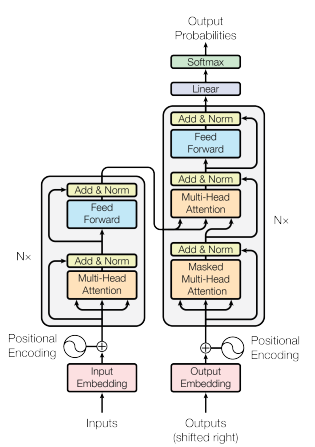
在常识、百科知识、文本、新闻等包含大量知识的库中，存在很多规范和不规范的知识。在构建知识图谱时，一般规范的知识表达比较好处理，不规范的知识表达经常需要转换成规范的知识表达来处理。个人理解知识图谱将建立一个像搜索引擎背后那样庞大的知识库，但是与之前搜索不一样的是：知识图谱以图的形式存储，可直接回答问题，而不是单一的文本匹配。在如此庞大的知识图谱的构建中，一定得以借助已有结构化和非结构化知识自动构建为主，人工构建为辅。自动构建的库一定会有错误，可以通过人工来纠错。

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

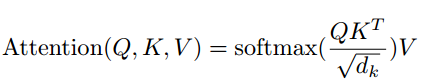
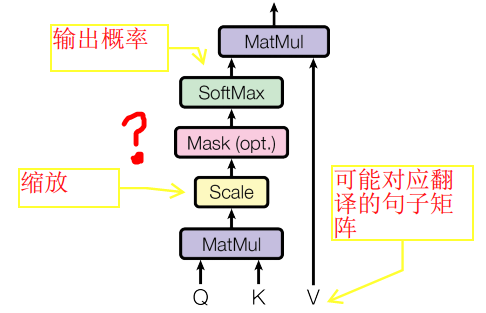
本文提出了一个新的神经网络模型-Transformer。它是完全基于注意力机制的。

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

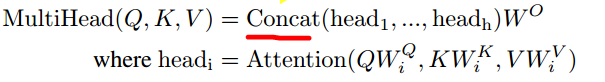
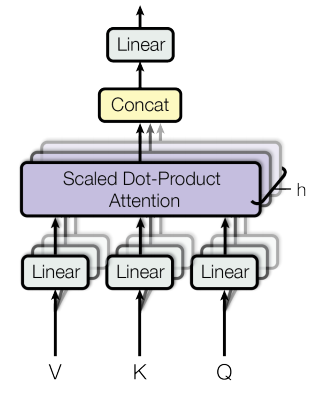
1. 总结及理解
2. 背景：之前，常用RNN、LSTM、GRNN来完成序列类及翻译任务。注意力机制的出现，改变了这一状况，它在很多任务中都有较好的表现。其中，本文提出的Transformer允许更多的并行化，并且在8个P100 GPU上经过长达12小时的培训后，可以达到翻译质量的最新技术水平。
3. Transformer模型如下图所示，在这里，编码器（图的左边）将符号表示的输入序列（x1,…,xn)映射为连续表示序列z = (z1,…,zn)。 给定z,解码器（图的右边）然后一次生成一个符号的输出序列(y1,…,ym)。在每个步骤中，模型都是自动回归的，在生成下一个时，将先前生成的符号作为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



* 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