Memory Networks: Architectures, Attention Mechanisms, and Applications.

This project aims to introduce the basic idea of memory and the underlying challenges behind it.

这份项目主要讲memory 的基本思路和现在面临的问题，总的来说比较简单

符尧 francis_yao@pku.edu.cn 北大信科/ 头条AI实验室

We will cover: 主要会提到

- Representation of memory, memory 的表示
- Interaction of memory, decoder 和memory 的交互

We will not cover: 不会提到的内容

- How to write the memory， 怎么去写一个memory
- Advanced topics，更多模型

Table of content 目录:

Paper list

- Memory， 与memory有关的内容
  - Jason Weston, Sumit Chopra, Antoine Bordes, *Memory Networks*, ICLR 2015
    - Memory networks origin，memory最开始的文章
    - The bAbI QA dataset，bAbI问答数据集，用来测试模型的推理能力
  - Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus, *End-To-End Memory Networks*, NIPS 2015

        - Multi-top attention，这篇文章主要提出了multihop attention
- After these two papers, the representation of memory and interactions with memory are two important aspects in this field. 在上面两篇文章之后，如何表示memory，以及如何与memory交互，成为了比较重要的研究方向
    - Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, Jason Weston, *Key-Value Memory Networks for Directly Reading Documents*, EMNLP16
        - Key-value representation for structured knowledge. 如果memory包含的是结构化的知识，那么KV是一个比较有效的表示方法
- Attention，与attention注意力机制有关的内容
    - Effective Approaches to Attention-based Neural Machine Translation
        - Multiple Attention implementations 各种attention的实现方式
        - Local attention 局部attention，让decoder只去看memory的一个点
    - Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, Richard Socher, *Ask Me Anything: Dynamic Memory Networks for Natural Language Processing*
        - Use GRU to compute attention -- sequential attention 使用GRU来计算attention，把顺序的信息加入了attention的计算
        - Multi-hop attention, again, use GRU to connect hops 在multihop attention的时候，使用GRU来连接hop，而不是简单加起来
    - Jörg Bornschein Andriy Mnih Daniel Zoran Danilo J. Rezende, *Structured Attention Networks*, ICLR 2017
        - Attention to a memory span 一次attend到memory中的一块，而不是单个单词
    - Abigail See, Peter J. Liu, Christopher D. Manning, *Get To The Point: Summarization with Pointer-Generator Networks*, SIGDIAG 2017
        - Attention history helps reduce repetition. 对decoder的历史输出做attention，考虑这部分信息，可以减少decoder重复输出

Memory Architectures

In this section we introduce different memory architectures. 这部分主要讲各种memory的架构

Memory Networks

- *Memory Networks*, Jason Weston, Sumit Chopra, Antoine Bordes, ICLR 2015
- Original Memory Networks. In short: add a memory component into a network. 这篇是memory networks 第一篇文章，简单来说，就是给一个网络增加一个memory 模块
- RNNs are known to have difficulty in performing memorization -- please note the difference between memorization and long-term dependency -- both are not fully solved 写这篇文章的动机是，RNN对于信息的记忆，以及对于长时间信息依赖的建模能力是有待商榷的，并且这两个问题都没有被完全的解决 -- 其实直到现在，还是没有被很好地解决
- General framework/ schema -- you can design/ interprete your own memory 这篇文章是提出了一个框架，你也可以根据自己的需要往这个框架之中填入自己的模型结构

- given an input x (a word) 给出一个输入，这个输入可能是一个词
- $I(x)$ representation of x (a question embedding)，I这个模块的作用是把一个输入做一个表示
- $m_i = G(m_i, I(x), m)$ Update the memory (given a statement, store it)，G这个模块的作用是根据输入，对memory进行更新。在这篇文章里，所谓的更新，就是简单地把一个新的句子填入memory的一个slot里面
- $o = O(I(x), m)$ compute output features (largest score between statements and the question) -- origin of multi-hop attention，O这个模块的作用从memory得到一个输出
- $r = R(o)$ decode outputs (largest score between word and o)，R这个模块的作用是把输出转为我们需要的记过
- Typically, o and r are the two most nontrival task，一般而言，O和R这两个模块是比较难的，接下来的讨论也都是说这两个模块具体应该怎么做

* Task: the bAbI QA task -- to test **reasoning ability** (reason over multiple facts) 这个模型的任务是做bAbI QA，这个数据集主要的目标是去测试模型做推理的能力，特别是根据多条论据来做推理（这个任务是问答/阅读理解上比较难的一个任务）
    - 1 Joe went to the kitchen.
    - 2 Fred went to the kitchen.
    - 3 Joe picked up the milk.
    - 4 Joe travelled to the office.
    - 5 Joe left the milk.
    - 6 Joe went to the bathroom.
    - Where is the milk now? A: office -- 345, reason over multiple facts 注意要回答这个问题，模型需要根据第345句来做推理
    - Where is Joe? A: bathroom -- 6 要回答这个问题，不需要推理，只需要来做匹配就好了
    - Where was Joe before the office? A: kitchen -- 4,2 这个问题需要从第四和第二句来做推理
    - 注意：其实345这三个句子，虽然说是根据三句话来做推理，但是如果把这三句话接起来的话，还是有可能通过匹配来回答出来的
    - 更难的例子是，当345这三句没有连在一起，而是分散在不同的位置上的时候，模型是否依然能够抽出来这三句，然后回答问题
* Reasoning ability: reason over **verbs**. 在这个任务中，主要是对动词的影响进行推理
* Training: Making word embeddings and questions closer! 训练的过程实际上就是让word embedding 和question 更近
    - $loss1 = r - s(x, m_{o1}) + s(x, f_1)$ hop1 第一次hop，x是问题，m_{o1}是与问题相关的第一个句子
    - $loss2 = r - s([x, m_{o1}], m_{o2}) + s([x, m_{o1}], f_2)$ hop2 m_{o1}是与问题相关的第二个句子
    - $loss3 = r - s([x, m_{o1}, m_{o2}], r) + s([x, m_{o1}, m_{o2}], f_3)$ output 最终输出
    - Three embedding matrix: x, m_{o1}, m_{o2}
    - 在这上面的f是负样本，整个的训练思路就是让memory的中，正样本离query更近，负样本离query更远
    - *Question Answering with Subgraph Embeddings* EMNLP14
    - *Translating Embeddings for Modeling Multi-relational Data* NIPS 13
* Suspect to pattern matching: because we may simply concat these facts and perfrom pattern matching! 这个任务可能会沦为一个匹配任务，而不是推理任务，以为我们可以直接把与问题相关的句子先接起来，然后再去与问题做匹配。同时，这个训练的过程也是匹配局导向的

- Question: A dataset that tests more reasoning ability? -- we need to define **reasoning**: 那么，什么样的数据集可以更好的测试推理能力呢？在这之前，我们需要去定义好什么叫推理能力
    - In classical Criticla Thinking definition: reasoning = finding evidences to support a claim (pattern matching like) + evaluate confidence of the evidence (seems that no model doing this explicitly ...) 在经典批判性思维中，认为推理能力 = 寻找证据证明论点 + 分析证据有多可信
    - Human behavoir: more focus on evaluation of the evidences, if an evidence is not so valid, find more evidences. 人类的行为更多是在分析证据有多可信，如果证据有缺陷的话，人们会寻找补充证据来补足这个缺陷，而机器更多地是在寻找证据（匹配），却比较少地去分析证据有多可信
    - Evaluation of evidence - **induction chain**: is this evidence directly support this claim? if not, what other evidences are needed? 如果证据有缺陷，那么需要寻找补充证据，这个是一个链式的过程，重复此过程，形成推理链条 -- 机器很难做到这点
    - Datasets that are more sophisticated to test reasoning capability: The NarrativeQA Reading Comprehension Challenge 这个数据集更多的试图去测试一个模型的推理能力
    - Is SQuAD aimed to test reasoning alibity? -- probably not! -- We still have a long way toward reading comprehension! 注意：SQuAD更多是在测试匹配能力，而不是推理能力
    - An introduction of different Reading Comprehension / QA datasets
- If memory is very large -- hashing -- but not an differential operation 回到memory，如果memory很大，那么用hashing，但是此操作不可微
- This work is submitted to Arxiv as the same time as Attention and NTM Neural turing Machines -- and NTM later evolved to be DNC Differential Neural Computers 这篇文章在Attention 和NTM的同时被交上了Arxiv

The NarrativeQA Reading Comprehension Challenge

- In short: what the bAbI QA task want to achieve - reasoning. 这个数据集更多想要测试模型的推理能力
- Do not want to question to be answerable by 不希望：
    - shallow pattern matching 直接通过匹配得到答案
    - guessing based on global salience 给一个全局关键词，猜出来答案
- Want to question to be answered after 希望：
    - integrate information distributed across different parts of the document 从文章里不同的地方提取证据
    - higher-level relations between entities, places, and events 分析文中实体，地点，事件之间的关系
- The formation of this task is still a challenging topic. 如何去定义推理任务，本身就是一个比较难的问题

End to End Memory Networks

- In short:
    - Train a memory network end2end 这篇文章端到端地训练一个memory network
    - Multiple hops yields improved results 同时，这篇文章指出，multi-hop attention对于模

型效果有好处

- Multi-hop attention -- an implicit KV fashion 这篇文章在实现attention的时候实际上是实现了一个KV版本的attention，但是没有明说，我们等下会说KV attention
    - recall: standard attention
        - $o = attn(q, M)$
        - o output, q query, $M = \{m_1, m_2 \ldots m_n\}$ memory 一个标准的attention是给一个query和一个memory，还给你一个context vector
    - multi-hop:
        - $o_t = attn(q_t, M)$
        - $q_{t+1} = o_t + q_t$ multi-hop的一个基本想法是，把从query得到的context vector 和query加起来，再去做query，如此续行
        - $t \in \{1, 2, \ldots, n\}$ n number of hops
        - $a_i = q_i \cdot mk_i$
        - $e_i = \frac{exp(a_i)}{\sum exp(a_j)}$
        - $o = \sum e_i \cdot mv_i$
        - The original paper uses A and C instead of mk and mv, but essentially this is a **key-value memory**
- Why multi-hop? because want to perform **reasoning** over a memory -- effectiveness of multi-hop attention，multihop的原因是希望从memory中抽取多条内容，根据这些内容来做推理
    - bAbI QA dataset: more hops, less errors 越多hop，效果越好
    - PTB language model: more hops, less PPL 在PTB 语言模型上，也是越多hop，越小混乱度
    - MT: more hops, more BLEU 机器翻译也一样
        - Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin, *Convolutional Sequence to Sequence Learning*
    - In practice, just use the context vector to query the memory again 实际工程中，最简单的multihop的做法，就是把query 和context vector加起来，再做一次attention
- A potential issue: Gradient explosion/ vanishing[1]? 但是这样做的一个可能的问题是梯度消失或者梯度爆炸，文献3中有讨论这个问题
    - This type of attention operation is suspect to gradient vanishing
    - "To aid training, we apply ReLU operations to half of the units in each layer." -- Does "aid training" mean gradient explostion? 实际上在实现的过程中加了ReLU
    - **This question is not clearly answered.** 这个问题在文章中并没有被明确回答，而经验上，当hop到78次的时候，类似于网络到了78层，这样可能会有梯度消失
- Task: Language Modeling (we skip the bAbI task part)
    - PTB dataset and Text8 dataset. Note: the result is far from state of the art! 同时需要注意的是，在语言模型上，End2End MemNN远不是最好的，在文章1中perp就有了68.67，在文献2中达到了58.0
    - PTB: best config gives 111 test perp. Note on RNN Regularization (dropout paper): 68.67 perp (ICLR15)[2], and DeepMind[3] ICLR 18 paper: 58.0

Key-Value Memory Networks

- In short: structured knowledge, query on keys, outputs from values. 这篇文章是说，对于结构化的数据，query对key求一个分布，再把分布作用到value上得到输出
- Alexander H. Miller. Adam Fisch. Jesse Dodge. Amir-Hossein Karimi. Antoine Bordes. Jason Weston, *Key-Value Memory Networks for Directly Reading Documents*, EMNLP 2016
- In short:
  - read KB/ IE/ DOC with key-value memory: how to organize information from different sources into a KV representation 这篇文章提出了如何对各种类型的数据建立一个KV Mem
    - KB& IE: key = subject + predicate, value = object 对于从KB/ IE出来的数据，把主语和谓语作为key，把宾语作为value
    - DOC: key = center word + window, value = center word 对于从文章中的数据，把中心词和它的周边词作为key，再把中心词做value
  - performance: KB > IE > DOC -- the more clean the memory is, the better the performance 越是结构化，效果越好
- An empirical conclusion is that, KVMems may be useful for structured knowledge 一个经验是，KVMem比较适合结构化的数据
  - However, the structure it can model is quite shallow, for deeper structure modeling, an example is Percy Liang's Recursive NN[4] paper using a recursive RNN to model subject-predicate-object relations. 但是，这个模型对于结构化的数据的表示还非常地浅，既无法表示一个知识图谱中各个SPO三元组之间的关系，也无法保证在下游模型中能够保持这个关系
  - 那么怎么对结构化的数据更好地建模呢？文献5给出了一种使用 Recursive NN的方法

Different ways to compute Attention

Attention is a effective (and a only) way to let the downstream task (a decoder) to interact with the memory. 下面的内容讲如不同计算attention的方法（注意与memory的交互）

Effective Approaches to Attention-based Neural Machine Translation

- In short: Different ways to compute attention score 提出了三种不同计算attention的方法
  - $a_i = q \cdot m_i$ - Tensorflow Luong Attention `tf.contrib.seq2seq.LuongAttention`, tf1.5 两个向量点乘，有tensorflow 的实现
  - $a_i = v \cdot tanh(Wq + m_i)$ - Tensorflow Bahdanau Attention `tf.contrib.seq2seq.BahdanauAttention`, tf1.5 对query加一个线性映射，然后再加上一个双曲正切
  - $a_i = v \cdot tanh(W_1 q + W_2 m_i)$ - Attention described in *Grammar as a Foreign Language*, implemented in tf1.2 seq2seq tutorial 两者都加一个线性映射，这种attention是GOOGLE一开始的实现方法，并且在tf1.2 seq2seq的教程中有提到
  - Empirically, the third one is slightly better than the first two, but **the performance may vary from task to task**. 经验的结果是最后一种效果最好

  - Many important details in tf1.2 are hidden in tf1.5, good for quick prototyping, bad for research, if you want to use tf1.5 for research purpose (e.g. implement attention on a KV memory, or implement attention in the following sections), it is recommended that you read the source code. 如果你想要迅速实现模型，那么建议直接使用tf现在的接口，如果你想要做一些更细粒度的研究，那么建议去看tf1.5 源码，或者tf1.2 的教程
- Luong's local attention: focus on a local place 除了上面提出的三个attention之外，这里还提出了一个local attention：让attention去关注memory的一个点
  - $p_t = L \cdot sigmoid(v \cdot tanh(Wh_t))$ -- predict a source location，L是memory的长度，因为sigmoid取值为0到1，那么这个函数去预测了一个memory的位置
  - $a_s = a_s \cdot exp(-\frac{(s-p_t)^2}{2\sigma^2})$ -- use a gaussian to let the attention scores a to focus on p_t 把原先计算出来的attention score加上一个高斯分布，把高斯中心的score调高
- This attention want to focus on one single location in the source, multi-hop attention can extract different source locations (set operation). 这种attention的方法是想要侧重在单个memory 位置上，multihop 的方法想要读取多个位置
  - Note that we may simply change the softmax $e_i = softmax(a_i)$ into sigmoid $e_i = sigmoid(a_i)$ to let the attention to retrieval information from different locations. 注意：如果我们把计算attention时求分布的softmax改成sigmoid，这样也可以提取多个位置的信息
  - But this may not work well! We will discuss this later. 但这不一定有效

Other attention variants 其他attention的衍生

- Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, Richard Socher, *Ask Me Anything: Dynamic Memory Networks for Natural Language Processing*, ICML 2016
  - When there is order, need sequential attention/ multihop attention improves reasoning 这个模型的假设是，memory的内部是有顺序的，同时，对memory 做attention也是有顺序的，因此它使用两个GRU来对这两种顺序建模
  - Tasks: bAbI QA 同样是facebook的bAbI QA数据集
  - Architecture (simplified here for better understanding) 这里为了更好地理解这个模型，我们做一点简化
    - Sequential attention: $h_t = GRU(q, m_i, h_{t-1})$, $c = h_T$ c context vector. i.e. change softmax to GRU 在对memory做attention的时候，把softmax 改成了GRU，然后取GRU的最后一个输出作为context vector
    - Sequential hops $c_i = GRU(c_{i-1})$ i.e. change addition to GRU 在做multihop attention的时候，把hop与hop之间连接改成了GRU，原先只是简单地加起来
- Jörg Bornschein Andriy Mnih Daniel Zoran Danilo J. Rezende, *Structured Attention Networks*, ICLR 2017
  - Add a linear chain CRF on attention scores, extend attention from **a single word** to **a span of words**. (skip details here) 这个文章用了一些latent variable来控制attention score，这些latent variable之间形成一个线性条件随机场，这个线性条件随机场的参数是模型学习出来的

- Note: this paper needs the CRF math prior, here is a tutorial[5]

- Abigail See, Peter J. Liu, Christopher D. Manning, *Get To The Point: Summarization with Pointer-Generator Networks*
  - Add attention history to prevent repetition. 这篇文章在做attention的时候，增加了历史attention的信息来防止模型重复输出
  - Attention as pointers. 同时，也把attention当做pointer来用

What this survey does not cover 这篇综述没有涉及到的内容

- Advanced memory representation learning 更加复杂的表示方法
  - Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings
- Advanced memory architecture and addressing 更加复杂的architecture和 addresing 方法
  - Hierarchical Memory Networks
  - Memory Augmented Neural Networks with Wormhole Connections
  - Dynamic Neural Turing Machine with Continuous and Discrete Addressing Schemes
  - Unbounded cache model for online language modeling with open vocabulary
- Differential Neural Computers, Neural Turing Machine
  - Hybrid computing using a neural network with dynamic external memory(DNC)
  - Neural Turing Machines

---

1. On the Difficulty of Training Recurrent Neural Networks↩
2. Recurrent Neural Network Regularization↩
3. On The State Of The Art Of Evaluation In Neural Language Models↩
4. He He, Anusha Balakrishnan, Mihail Eric, Percy Liang, *Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings*, ACL 2017↩
5. Charles Sutton and Andrew McCallum, *An Introduction to Conditional Random Fields*↩