Memory Networks: Architectures, Attention Mechanisms, and Applications.

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

We will cover:

- Representation of memory
- Interaction of memory

We will not cover:

- How to write the memory
- Advanced topics

Table of content:

Paper list

- Memory
    - Jason Weston, Sumit Chopra, Antoine Bordes, *Memory Networks*, ICLR 2015
        - Memory networks origin
        - The bAbI QA dataset
    - Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus, *End-To-End Memory Networks*, NIPS 2015
        - Multi-top attention
- After these two papers, the representation of memory and interactions with memory are two important aspects in this field.

    - 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.
- Attention
    - Effective Approaches to Attention-based Neural Machine Translation
        - Multiple Attention implementations
        - Local attention
    - 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
        - Multi-hop attention, again, use GRU to connect hops
    - Jörg Bornschein Andriy Mnih Daniel Zoran Danilo J. Rezende, *Structured Attention Networks*, ICLR 2017
        - Attention to a memory span
    - Abigail See, Peter J. Liu, Christopher D. Manning, *Get To The Point: Summarization with Pointer-Generator Networks*, SIGDIAG 2017
        - Attention history helps reduce repetition.

Memory Architectures

In this section we introduce different memory architectures.

Memory Networks

- *Memory Networks*, Jason Weston, Sumit Chopra, Antoine Bordes, ICLR 2015
- Original Memory Networks. In short: add a memory component into a network.
- RNNs are known to have difficulty in performing memorization -- please note the difference between memorization and long-term dependency -- both are not fully solved
- General framework/ schema -- you can design/ interprete your own memory
    - given an input x (a word)
    - $I(x)$ representation of x (a question embedding)
    - $m_i = G(m_i, I(x), m)$ Update the memory (given a statement, store it)
    - $o = O(I(x), m)$ compute output features (largest score between statements and the question) -- origin of multi-hop attention
    - $r = R(o)$ decode outputs (largest score between word and o)
    - Typically, o and r are the two most nontrival task
- Task: the bAbI QA task -- to test **reasoning ability** (reason over multiple facts)
    - 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
- Where is Joe? A: bathroom -- 6
- Where was Joe before the office? A: kitchen -- 4,2

* Reasoning ability: reason over **verbs**.
* Training: Making word embeddings and questions closer!
  - $loss1 = r - s(x, m_{o1}) + s(x, f_1)$ hop1
  - $loss2 = r - s([x, m_{o1}], m_{o2}) + s([x, m_{o1}], f_2)$ hop2
  - $loss3 = r - s([x, m_{o1}, m_{o2}], r) + s([x, m_{o1}, m_{o2}], f_3)$ output
  - Three embedding matrix: x, m_{o1}, m_{o2}
  - *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!
* Question: A dataset that tests more reasoning ability? -- we need to define **reasoning**:
  - 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 ...)
  - Human behavoir: more focus on evaluation of the evidences, if an evidence is not so valid, find more evidences.
  - Evaluation of evidence - **induction chain**: is this evidence directly support this claim? if not, what other evidences are needed?
  - Datasets that are more sophisticated to test reasoning capability: The NarrativeQA Reading Comprehension Challenge
  - Is SQuAD aimed to test reasoning alibity? -- probably not! -- We still have a long way toward reading comprehension!
  - An introduction of different Reading Comprehension / QA datasets
* If memory is very large -- hashing -- but not an differential operation
* 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

The NarrativeQA Reading Comprehension Challenge

* In short: what the bAbI QA task want to achieve - reasoning.
* Do not want to question to be answerable by
  - shallow pattern matching
  - guessing based on global salience
* Want to question to be answered after
  - integrate information distributed across different parts of the document
  - higher-level relations between entities, places, and events
* The formation of this task is still a challenging topic.

End to End Memory Networks

- In short:
  - Train a memory network end2end
  - Multiple hops yields improved results
- Multi-hop attention -- an implicit KV fashion
  - recall: standard attention
    - $o = attn(q, M)$
    - o output, q query, $M = \{m_1, m_2 \dots m_n\}$ memory
  - multi-hop:
    - $o_t = attn(q_t, M)$
    - $q_{t+1} = o_t + q_t$
    - $t \in \{1, 2, \dots, n\}$ 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
  - bAbI QA dataset: more hops, less errors
  - PTB language model: more hops, less PPL
  - MT: more hops, more BLEU
    - 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
- A potential issue: Gradient explosion/ vanishing[1]?
  - 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?
  - **This question is not clearly answered.**
- Task: Language Modeling (we skip the bAbI task part)
  - PTB dataset and Text8 dataset. Note: the result is far from state of the art!
  - 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.
- 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
        - KB& IE: key = subject + predicate, value = object
        - DOC: key = center word + window, value = center word
    - performance: KB > IE > DOC -- the more clean the memory is, the better the performance
- Mihail Eric, Lakshmi Krishnan, Francois Charette, Christopher D. Manning, Key-Value Retrieval Networks for Task-Oriented Dialogue, SIGDIAL 17
    - Store KB in to a KVMem, and apply to task-oriented dialogue
- An empirical conclusion is that, KVMems may be useful for structured knowledge
    - However, the structure it can model is quite shallow, for deeper structure modeling, an example is Percy Liang's Recursive NN paper[4] using a recursive RNN to model subject-predicate-object relations.

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.

Effective Approaches to Attention-based Neural Machine Translation

- In short: Different ways to compute attention score
    - $a_i = q \cdot m_i$ - Tensorflow Luong Attention `tf.contrib.seq2seq.LuongAttention`, tf1.5
    - $a_i = v \cdot tanh(Wq + m_i)$ - Tensorflow Bahdanau Attention `tf.contrib.seq2seq.BahdanauAttention`, tf1.5
    - $a_i = v \cdot tanh(W_1 q + W_2 m_i)$ - Attention described in *Grammar as a Foreign Language*, implemented in tf1.2 seq2seq tutorial
    - Empirically, the third one is slightly better than the first two, but **the performance may vary from task to task**.
    - 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.
- Luong's local attention: focus on a local place
    - $p_t = L \cdot sigmoid(v \cdot tanh(Wh_t))$ -- predict a source location
    - $a_s = a_s \cdot exp(-\frac{(s-p_t)^2}{2\sigma^2})$ -- use a gaussian to let the attention scores a to focus on p_t
- This attention want to focus on one single location in the source, multi-hop attention can extract different source locations (set operation).
    - 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.
  - But this may not work well! We will discuss this later.

Other attention variants

- 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
  - Tasks: bAbI QA
  - Architecture (simplified here for better understanding)
    - Sequential attention: $h_t = GRU(q, m_i, h_{t-1})$, $c = h_T$ c context vector. i.e. change softmax to GRU
    - Sequential hops $c_i = GRU(c_{i-1})$ i.e. change addition to GRU
- Yoon Kim, Carl Denton, Luong Hoang, Alexander M. Rush, *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)
  - 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.
  - Attention as pointers.

What this survey does not cover

- Advanced memory representation learning
- Advanced memory architecture and addressing
  - 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*↩