Iterative Multi-document Neural Attention for Multiple Answer Prediction

URANIA Workshop
Genova (Italy), November, 28th, 2016

Claudio Greco, Alessandro Suglia, Pierpaolo Basile, Gaetano Rossiello and Giovanni Semeraro

Work supported by the **IBM Faculty Award "Deep Learning to boost Cognitive Question Answering"**
Titan X GPU used for this research donated by the **NVIDIA Corporation**

Overview

1. Motivation
2. Methodology
3. Experimental evaluation
4. Conclusions and Future Work
5. Appendix

Motivation

Motivation

- People have information needs of varying complexity such as:
  - simple questions about common facts (*Question Answering*)
  - suggest movie to watch for a romantic evening (*Recommendation*)
- An intelligent agent able to answer questions formulated in a proper way can solve them, eventually considering:
  - user context
  - user preferences

Idea

In a scenario in which the **user profile** can be represented by a **question**, intelligent agents able to answer questions can be used to find the most appealing items for a given user

Motivation

Conversational Recommender Systems (CRS)

Assist online users in their *information-seeking* and *decision making* tasks by supporting an *interactive process* [1] which could be goal oriented with the task of starting general and, through a series of interaction cycles, narrowing down the user interests until the desired item is obtained [2].

[1]: T. Mahmood and F. Ricci. "Improving recommender systems with adaptive conversational strategies". In: Proceedings of the 20th ACM conference on Hypertext and hypermedia. ACM. 2009.

[2]: N. Rubens et al. "Active learning in recommender systems". In: Recommender Systems Handbook. Springer, 2015.

Methodology

Building blocks for a CRS

According to our vision, to implement a CRS we should design the following building blocks:

1. Question Answering + recommendation
2. Answer explanation
3. Dialog manager

Our work called “Iterative Multi-document Neural Attention for Multiple Answer Prediction” tries to tackle building block 1.

Iterative Multi-document Neural Attention for Multi Answer Prediction

The key contributions of this work are the following:

1. We extend the model reported in [3] to let the inference process exploit evidences observed in multiple documents
2. We design a model able to leverage the attention weights generated by the inference process to provide multiple answers
3. We assess the efficacy of our model through an experimental evaluation on the *Movie Dialog* [4] dataset

---

[3]: A. Sordoni, P. Bachman, and Y. Bengio. "Iterative Alternating Neural Attention for Machine Reading". In: arXiv preprint arXiv:1606.02245 (2016)

[4]: J. Dodge et al. "Evaluating prerequisite qualities for learning end-to-end dialog systems". In: arXiv preprint arXiv:1511.06931 (2015).

Iterative Multi-document Neural Attention for Multi Answer Prediction

Given a query q, $\psi : Q \rightarrow D$ produces the set of documents relevant for q, where Q is the set of all queries and D is the set of all documents.

Our model defines a workflow in which a sequence of inference steps are performed:

1. Encoding phase
2. Inference phase
    - Query attentive read
    - Document attentive read
    - Gating search results
3. Prediction phase

Encoding phase

Both queries and documents are represented by a sequence of words $X = (x_1, x_2, \ldots, x_{|X|})$, drawn from a vocabulary V. Each word is represented by a continuous d-dimensional word embedding $\mathbf{x} \in \mathbb{R}^d$ stored in a word embedding matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d}$.

Documents and query are encoded using a *bidirectional recurrent neural network* with *Gated Recurrent Units (GRU)* as in [3].

Differently from [3], we build a unique representation for the whole set of documents related to the query by stacking each document token representations given by the *bidirectional GRU*.

---

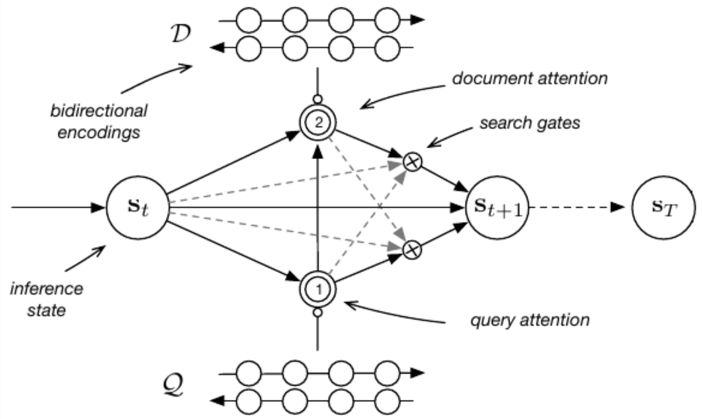
[3]: A. Sordoni, P. Bachman, and Y. Bengio. "Iterative Alternating Neural Attention for Machine Reading". In: arXiv preprint arXiv:1606.02245 (2016)

Inference phase

This phase uncovers a possible inference chain which models meaningful relationships between the query and the set of related documents. The inference chain is obtained by performing, for each timestep $t = 1, 2, \ldots, T$, the attention mechanisms given by the *query attentive read* and the *document attentive read*.

- *query attentive read*: performs an attention mechanism over the query at inference step t conditioned by the inference state
- *document attentive read*: performs an attention mechanism over the documents at inference step t conditioned by the refined query representation and the inference state
- *gating search results*: updates the inference state in order to retain useful information for the inference process about query and documents and forget useless one

Inference phase



[3]: A. Sordoni, P. Bachman, and Y. Bengio. "Iterative Alternating Neural Attention for Machine Reading". In: arXiv preprint arXiv:1606.02245 (2016)

Prediction phase

- Leverages document attention weights computed at the inference step t to generate a relevance score for each candidate answer
- Relevance scores for each token coming from the l different documents D_q related to the query q are accumulated

$$score(w) = \frac{1}{\pi(w)} \sum_{i=1}^{l} \phi(i, w)$$

where:

- $\phi(i, w)$ returns the score associated to the word w in document i
- $\pi(w)$ returns the frequency of the word w in D_q

Prediction phase

- A 2-layer feed-forward neural network is used to learn latent relationships between tokens in documents
- The output layer of the neural network generates a score for each candidate answer using a *sigmoid* activation function

$$\mathbf{z} = [score(w_1), score(w_2), \dots, score(w_{|V|})]$$
$$\mathbf{y} = \text{sigmoid}(\mathbf{W}_{ho}\ \text{relu}(\mathbf{W}_{ih}\mathbf{z} + \mathbf{b}_{ih}) + \mathbf{b}_{ho})$$

where:

- u is the hidden layer size
- $\mathbf{W}_{ih} \in \mathbb{R}^{u \times |V|}, \mathbf{W}_{ho} \in \mathbb{R}^{|A| \times u}$ are weight matrices
- $\mathbf{b}_{ih} \in \mathbb{R}^{u}$, $\mathbf{b}_{ho} \in \mathbb{R}^{|A|}$ are bias vectors
- $\text{sigmoid}(\text{x}) = \frac{1}{1+e^{-x}}$ is the *sigmoid* function
- $\text{relu}(\text{x}) = max(0, x)$ is the *ReLU* activation function

Experimental evaluation

Movie Dialog

bAbI Movie Dialog [4] dataset, composed by different tasks such as:

- factoid QA (QA)
- top-n recommendation (Recs)
- *QA*+recommendation in a dialog fashion
- Turns of dialogs taken from *Reddit*

[4]: J. Dodge et al. "Evaluating prerequisite qualities for learning end-to-end dialog systems". In: arXiv preprint arXiv:1511.06931 (2015).

Experimental evaluation

- Differently from [4], the relevant knowledge base facts, represented in triple from, are retrieved by ψ implemented using *Elasticsearch* engine
- Evaluation metrics:
    - *QA* task: HITS@1
    - *Recs* task: HITS@100
- The optimization method and tricks are adopted from [3]
- The model is implemented in *TensorFlow* [5] and executed on an *NVIDIA TITAN X* GPU

---

[3]: A. Sordoni, P. Bachman, and Y. Bengio. "Iterative Alternating Neural Attention for Machine Reading". In: arXiv preprint arXiv:1606.02245 (2016)

[4]: J. Dodge et al. "Evaluating prerequisite qualities for learning end-to-end dialog systems". In: arXiv preprint arXiv:1511.06931 (2015).

[5]: M. Abadi et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: CoRR abs/1603.04467 (2016).

Experimental evaluation

METHODS	QA TASK	RECS TASK
QA SYSTEM	*90.7*	N/A
SVD	N/A	19.2
IR	N/A	N/A
LSTM	6.5	27.1
SUPERVISED EMBEDDINGS	50.9	29.2
MEMN2N	79.3	28.6
JOINT SUPERVISED EMBEDDINGS	43.6	28.1
JOINT MEMN2N	83.5	26.5
OURS	*86.8*	**30**

Table 1: Comparison between our model and baselines from [4] on the *QA* and *Recs* tasks evaluated according to *HITS@1* and *HITS@100*, respectively.

[4]: J. Dodge et al. "Evaluating prerequisite qualities for learning end-to-end dialog systems". In: arXiv preprint arXiv:1511.06931 (2015).

Inference phase attention weights

Question:
what does Larenz Tate act in ?

Ground truth answers:
The Postman, A Man Apart, Dead Presidents, Love Jones, Why Do Fools Fall in Love, The Inkwell

Most relevant sentences:

- The Inkwell starred actors Joe Morton , Larenz Tate , Suzzanne Douglas , Glynn Turman
- Love Jones starred actors Nia Long , Larenz Tate , Isaiah Washington , Lisa Nicole Carson
- Why Do Fools Fall in Love starred actors Halle Berry , Vivica A. Fox , Larenz Tate , Lela Rochon
- The Postman starred actors Kevin Costner , Olivia Williams , Will Patton , Larenz Tate
- Dead Presidents starred actors Keith David , Chris Tucker , Larenz Tate
- A Man Apart starred actors Vin Diesel , Larenz Tate

Figure 1: Attention weights computed by the neural network attention mechanisms at the last inference step T for each token. Higher shades correspond to higher relevance scores for the related tokens.

Conclusions and Future Work

Pros and Cons

Pros

- Huge gap between our model and all the other baselines
- Fully general model able to extract relevant information from a generic document collection
- Learns latent relationships between document tokens thanks to the feed-forward neural network in the prediction phase
- Provides multiple answers for a given question

Cons

- Still not satisfying performance on the *Recs* task
- Issues in the *Recs* task dataset according to [6]

[6]: R. Searle and M. Bingham-Walker. “Why “Blow Out”? A Structural Analysis of the Movie Dialog Dataset”. In: ACL 2016 (2016)

Future Work

- Design a ψ operator able to return relevant facts recognizing the most relevant information in the query
- Exploit user preferences and contextual information to learn the user model
- Provide a mechanism which leverages attention weights to give explanations [7]
- Collect dialog data with user information and feedback
- Design of a framework for dialog management based on *Reinforcement Learning* [8]

---

[7]: B. Goodman and S. Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: arXiv preprint arXiv:1606.08813 (2016).

[8]: R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. Vol. 1. 1. MIT press Cambridge, 1998

Appendix

Recurrent Neural Networks

- **Recurrent Neural Networks (RNN)** are architectures suitable to model variable-length sequential data [9];
- The connections between their units may contain *loops* which let them consider past states in the learning process;
- Their roots are in the *Dynamical System Theory* in which the following relation is true:

$$s^{(t)} = f(s^{(t-1)}; x^{(t)}; \theta)$$

where $s^{(t)}$ represents the current system state computed by a generic function f evaluated on the previous state $s^{(t-1)}$, $x^{(t)}$ represents the current input and θ are the network parameters.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Tech. rep. DTIC Document, 1985

RNN pros and cons

Pros

- Appropriate to represent sequential data;
- A versatile framework which can be applied to different tasks;
- Can learn short-term and long-term temporal dependencies.

Cons

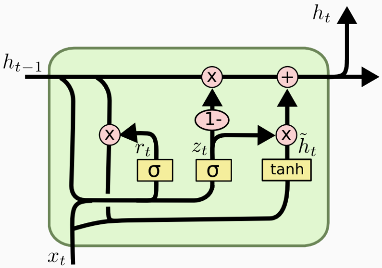
- **Vanishing/exploding gradient problem** [10, 11];
- Difficulties to reach satisfying minima during the optimization of the loss function;
- Difficult to parallelize the training process.

---

[10] Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: Neural Networks, IEEE Transactions on 5 (1994)

[11] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 2001.

Gated Recurrent Unit (GRU) [12] is a special kind of RNN cell which tries to solve the vanishing/exploding gradient problem.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

GRU description taken from `https://goo.gl/gJe8jZ`.

[12] K. Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: arXiv preprint arXiv:1406.1078 (2014).

Attention mechanism

- Mechanism inspired by the way the human brain is able to focus on relevant aspects of a dynamic scene and supported by studies in visual cognition [13];
- Neural networks equipped with an *attention mechanism* are able to *learn* relevant parts of an input representation for a specific task;
- Attention mechanisms in Deep Learning techniques has incredibly boosted performance in a lot of different tasks such as *Computer Vision* [14–16], *Question Answering* [17, 18] and *Machine Translation* [19].

References

[1] T. Mahmood and F. Ricci. “Improving recommender systems with adaptive conversational strategies”. In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM. 2009, pp. 73–82.

[2] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan. “Active learning in recommender systems”. In: *Recommender Systems Handbook*. Springer, 2015, pp. 809–846.

[3] A. Sordoni, P. Bachman, and Y. Bengio. “Iterative Alternating Neural Attention for Machine Reading”. In: *arXiv preprint arXiv:1606.02245* (2016).

[4] J. Dodge et al. “Evaluating prerequisite qualities for learning end-to-end dialog systems”. In: *arXiv preprint arXiv:1511.06931* (2015).

[5] Martín Abadi et al. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”. In: *CoRR* abs/1603.04467 (2016). URL: `http://arxiv.org/abs/1603.04467`.

[6] R. Searle and M. Bingham-Walker. “Why “Blow Out”? A Structural Analysis of the Movie Dialog Dataset”. In: *ACL 2016* (2016), p. 215.

[7] Bryce Goodman and Seth Flaxman. “European Union regulations on algorithmic decision-making and a” right to explanation””. In: *arXiv preprint arXiv:1606.08813* (2016).

[8] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998.

[9] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by error propagation*. Tech. rep. DTIC Document, 1985.

[10] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. “Learning long-term dependencies with gradient descent is difficult”. In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[11] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

[12] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[13] Ronald A Rensink. “The dynamic representation of scenes”. In: *Visual cognition* 7.1-3 (2000), pp. 17–42.

[14] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. “Learning where to attend with deep architectures for image tracking”. In: *Neural computation* 24.8 (2012), pp. 2151–2184.

[15] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *arXiv preprint arXiv:1502.03044* 2.3 (2015), p. 5.

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2204–2212.

[17] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. "End-to-end memory networks". In: *Advances in neural information processing systems*. 2015, pp. 2440–2448.

[18] Alex Graves, Greg Wayne, and Ivo Danihelka. "Neural turing machines". In: *arXiv preprint arXiv:1410.5401* (2014).

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).