Multi-Cast Attention Networks for Retrieval-based Question Answering and Response Prediction

Yi Tay
Nanyang Technological University
Singapore
ytay017@e.ntu.edu.sg

Luu Anh Tuan
Institute for Infocomm Research
Singapore
at.luu@i2r.a-star.edu.sg

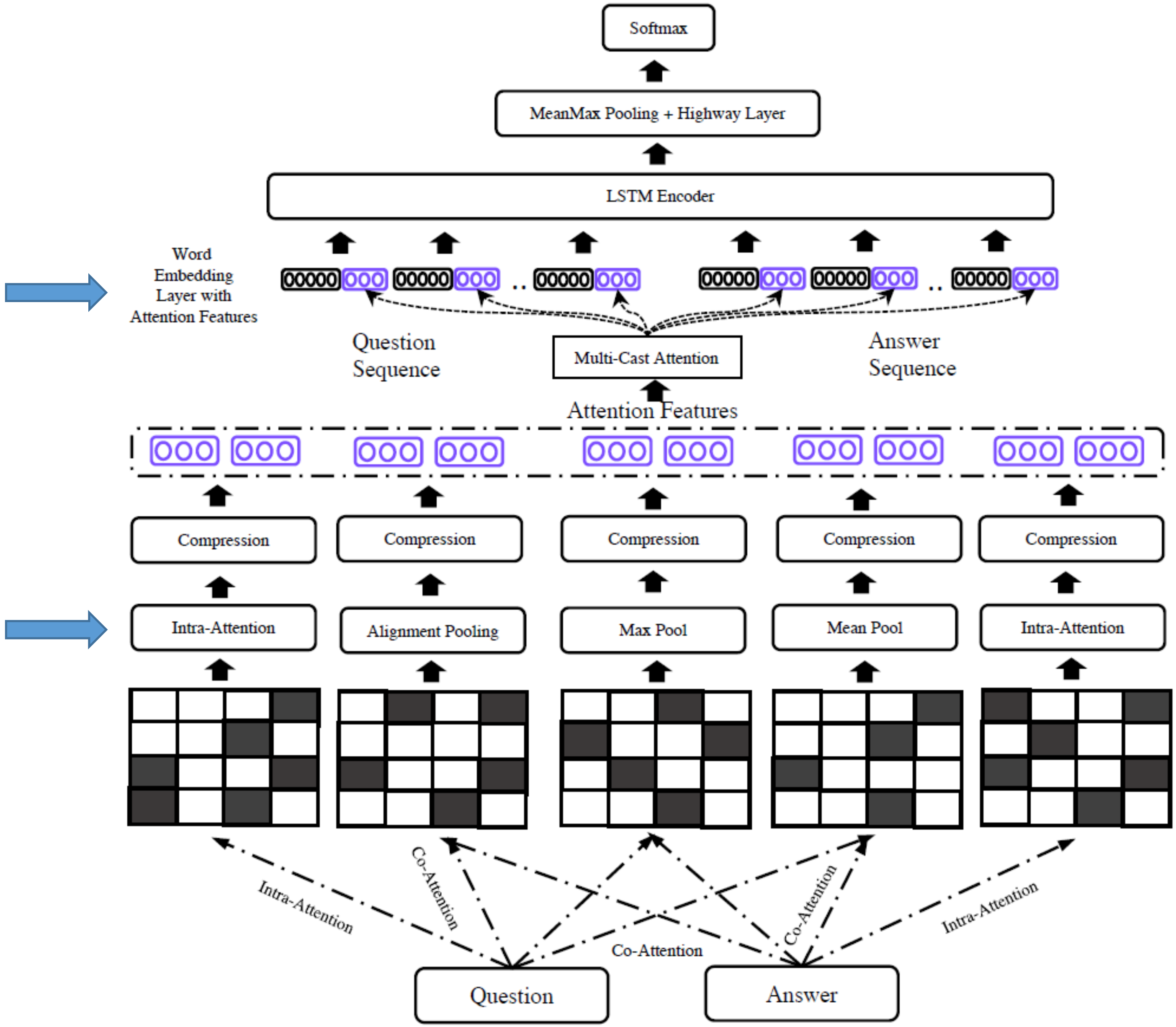
Siu Cheung Hui
Nanyang Technological University
Singapore
asschui@ntu.edu.sg

*in KDD 2018

Motivation

- Information Retrieval: modeling textual relevance between document-query pairs.
- Scoring + Ranking: neural network models.
- Attention is employed as feature extractor/pooling, and mostly, is applied once/one type.
- Different attention variants provide different views – architectural engineering.
- This paper: treat attention as a form of feature augmentation, and concurrently use multiple attention calls.

Multi-Cast Attention Network



Co-Attention

- A pairwise attention between document (D) and query (Q).
- Similarity matrix:

$$s_{ij} = F(q_i)^T F(d_j)$$

alignment score

- Max-Pooling:

$$q' = Soft(\max_{col}(s))^\top q \quad \text{and} \quad d' = Soft(\max_{row}(s))^\top d$$

re-weighting

- Mean-Pooling:

$$q' = Soft(\underset{col}{mean}(s))^\top q \quad \text{and} \quad d' = Soft(\underset{row}{mean}(s))^\top d$$

- Alignment-Pooling

$$d'_i := \sum_{j=1}^{\ell_q} \frac{exp(s_{ij})}{\sum_{k=1}^{\ell_q} exp(s_{ik})} q_j \quad \text{and} \quad q'_j := \sum_{i=1}^{\ell_d} \frac{exp(s_{ij})}{\sum_{k=1}^{\ell_d} exp(s_{kj})} d_i$$

realigning

Multi-Cast Attention

- Intra-Attention:

$$x'_i := \sum_{j=1}^{\ell} \frac{exp(s_{ij})}{\sum_{k=1}^{\ell} exp(s_{ik})} x_j$$

  where x is either q or d.

- Casted Attention:

$$f_c = F_c([\bar{x}; x])$$
$$f_m = F_c(\bar{x} \odot x)$$
$$f_s = F_c(\bar{x} - x)$$

  where $\bar{x}$ is the representation of x after applying attention.

- Modeling the influence of co-attention by comparing representations before and after co-attention.

Multi-Cast Attention

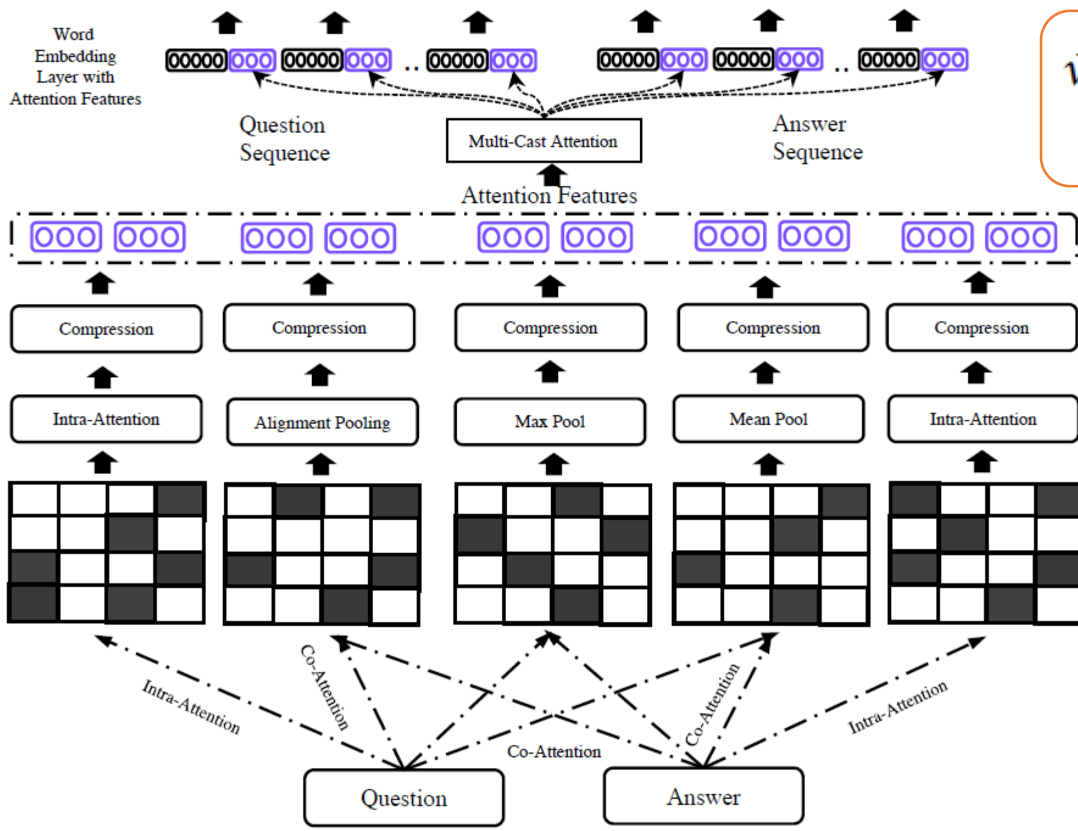
- Compression Function: reduce a vector to a scalar.

$$F(x) = \sum_{i}^{n} x_i \ , \qquad F(x) = ReLU(W_c(x) + b_c). \qquad F(x) = w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle \, x_i \, x_j$$

three ways: SUM, NN, and FM.

- The model:

$$f_c = F_c([\bar{x}; x])$$
$$f_m = F_c(\bar{x} \odot x)$$
$$f_s = F_c(\bar{x} - x)$$



$$\bar{w}_i = [w_i; z_i]$$
$$z \in \mathbb{R}^{12}$$

Dialogue Prediction

- Predict the next reply in conversations.

	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
MLP	0.651	0.256	0.380	0.703
DeepMatch	0.593	0.345	0.376	0.693
ARC-I	0.665	0.221	0.360	0.684
ARC-II	0.736	0.380	0.534	0.777
CNTN	0.743	0.349	0.512	0.797
MatchPyramid	0.743	0.420	0.554	0.786
LSTM	0.725	0.361	0.494	0.801
AP-LSTM	0.758	0.381	0.545	0.801
MV-LSTM	0.767	0.410	0.565	0.800
KEHNN	0.786	0.460	0.591	0.819
MCAN (SM)	0.831	0.548	0.682	0.873
MCAN (NN)	<u>0.833</u>	<u>0.549</u>	**0.686**	**0.875**
MCAN (FM)	**0.834**	**0.551**	<u>0.684</u>	**0.875**

Table 1: Performance Comparison on Ubuntu Dialogue Corpus. Best result is in boldface and second best is underlined.

Question Answering

- Factoid Question Answering

Model	MAP	MRR
QA-LSTM (dos Santos et al.)	0.728	0.832
AP-CNN (dos Santos et al.)	0.753	0.851
LDC Model (Wang et al.)	0.771	0.845
MPCNN (He et al.)	0.777	0.836
HyperQA (Tay et al.)	0.784	0.865
MPCNN + NCE (Rao et al.)	0.801	0.877
BiMPM (Wang et al.)	0.802	0.899
IWAN (Shen et al.)	0.822	0.889
MCAN (SM)	0.827	0.880
MCAN (NN)	<u>0.827</u>	<u>0.890</u>
MCAN (FM)	**0.838**	**0.904**

Table 2: Performance Comparison on TrecQA (*clean*) dataset. Best result is in boldface and second best is underlined.

Model	P@1	MAP
ARC-I (Hu et al.)	0.741	0.771
ARC-II (Hu et al.)	0.753	0.780
AP-CNN (dos Santos et al.)	0.755	0.771
Kelp (Filice et al.)	0.751	0.792
ConvKN (Barron Cedeno et al.)	0.755	0.777
AI-CNN (Zhang et al.)	0.763	0.792
CTRN (Tay et al.)	0.788	<u>0.794</u>
MCAN (SM)	<u>0.803</u>	0.787
MCAN (NN)	0.802	0.784
MCAN (FM)	**0.804**	**0.803**

Table 3: Performance comparison on QatarLiving dataset for community question answering. Best result is in boldface and second best is underlined.

- Community Question Answering

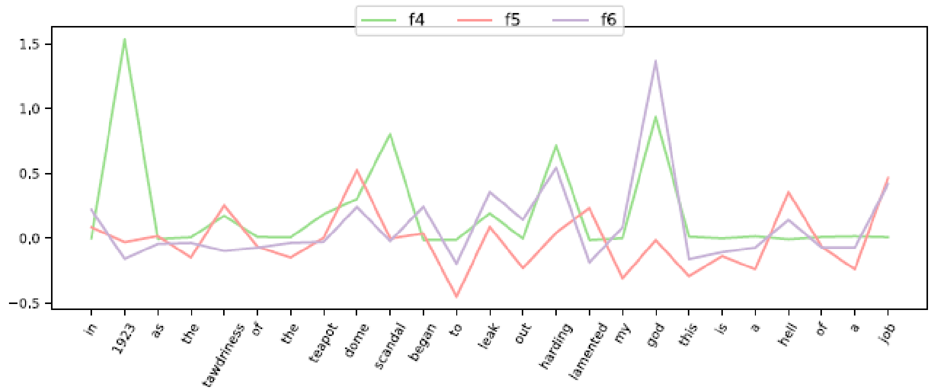
  answers are generally subjective and longer

Ablation Analysis

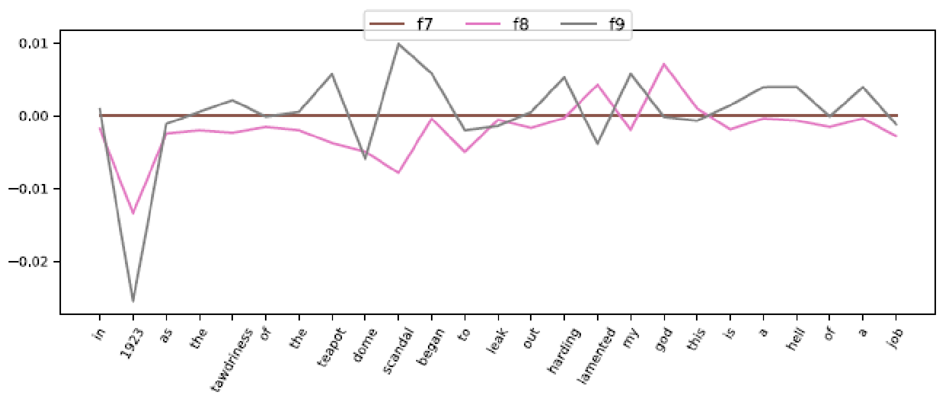
Setting	MAP	MRR
Original	**0.866**	**0.922**
(1) Remove Highway	0.825	0.863
(2) Remove LSTM	0.765	0.809
(3) Remove MCA	0.670	0.749
(4) Remove Intra	0.834	0.910
(5) Remove Align	0.682	0.726
(6) Remove Mean	0.858	0.906
(7) Remove Max	0.862	0.915

Table 5: Ablation analysis (validation set) on TrecQA dataset.

Visualization



(a) Features generated from ***max***-pool Co-Attention.



(b) Features generated from ***mean***-pool Co-Attention.

Figure 4: Differences between Max and Mean-pooled Casted Attention Features on answer text from TrecQA dataset. Diverse features are learned by different attention casts.

Summarization

- Employing multiple attention functions is beneficial.
- Solid evaluation and analysis.
- Co-attention in MT decoder?

- Ablation study on only MCA features.