Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention

Chen Zhao
University of Maryland, College Park

Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, Saurabh Tiwary
Microsoft AI and Research
• Input sentence(s) in sequential form
• (New) standard solutions with pre-trained Transformers (e.g., BERT)
• (New) standard solutions with pre-trained Transformers (e.g., BERT)

Transformer in Machine Reading

- Contextualized Embedding
- Transformer Multi-Head Attention
- Embedding

Span Prediction

P(Start)   P(End)

K Layers
Facebook was founded by Mark Zuckerberg, along with fellow Harvard College students and roommates.

Zuckerberg built a website called "Facemash" in 2003 while attending Harvard University. The site

Harvard University is a private Ivy League research university in Cambridge, Massachusetts, with about 6,800 undergraduate
This Work

• Transformer-XH for *structured* text:
  • Transformer with eXtra Hop attentions
  • Global representations of *multiple connected* text pieces

• Strong performance on different multi-evidence reasoning tasks
  • Multi-hop QA (Hotpot QA)
  • Multi-evidence Fact Verification (FEVER)
In sequence ($\tau$) attention in layer $l$, token $i$:

$$h_{\tau,i}^l = \sum_j \text{Softmax}_j \left( \frac{q_{\tau,i}^T \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}$$
Transformer-XH: Extra Hop Attention

In sequence ($\tau$) attention in layer $l$, token $i$:

$$h^l_{\tau,i} = \sum_j \text{Softmax}_j \left( \frac{q^T_{\tau,i} \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}$$

Extra hop attention between sequences ($\tau \rightarrow \eta$):

$$\hat{h}_{\tau,0}^l = \sum_{\eta; \ e_{\tau \eta}=1} \text{Softmax}_\eta \left( \frac{\hat{q}^T_{\tau,0} \cdot \hat{k}_{\tau,0}}{\sqrt{d_k}} \right) \cdot \hat{v}_{\eta,0}$$
In sequence (\(\tau\)) attention in layer \(l\), token \(i\):

\[
h^l_{\tau,i} = \sum_j \text{Softmax}_j \left( \frac{q^T_{\tau,i} \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}
\]

Extra hop attention between sequences (\(\tau \to \eta\)):

\[
\hat{h}^l_{\tau,0} = \sum_{\eta; e_{\tau\eta}=1} \text{Softmax}_\eta \left( \frac{\hat{q}^T_{\tau,0} \cdot \hat{k}_{\tau,0}}{\sqrt{d_k}} \right) \cdot \hat{v}_{\eta,0}
\]

Combine two attentions:

\[
\tilde{h}^l_{\tau,0} = \text{Linear} \left( [h^l_{\tau,0} \circ \hat{h}^l_{\tau,0}] \right)
\]

\[
\tilde{h}^l_{\tau,l} = h^l_{\tau,i}; \; \forall i \neq 0
\]
Transformer-XH in Multi-hop QA

**Input Question:**
In which city was Facebook launched?

**Evidence graph construction**
**Input Question:**
In which city was Facebook launched?

- Evidence graph construction
- Transformer-XH

---

**Transformer-XH in Multi-hop QA**

**Retrieval**
- Facebook
  - Facebook was founded at Harvard...
- Mark Zuckerberg
  - Zuckerberg attended Harvard...

**Entity Linking**

**Wiki Hyperlink**
- Harvard University
  - Harvard University is a private Ivy League...
- Social Media
  - Social media are ...

**Initial Documents**

**Connected Documents**

---
In which city was Facebook launched?

Harvard University is a private Ivy League research university in Cambridge, Massachusetts...

• Evidence graph construction
• Transformer-XH
• Task specific Layers
Results on Hotpot QA Full-Wiki Test Set

Public STOA: BERT + GCN

Answer Exact Match
- Official Baseline: 0.24
- Cognitive Graph QA: 0.371
- SR-MRS: 0.453
- Transformer-XH: 0.516

Answer F1
- Official Baseline: 0.329
- Cognitive Graph QA: 0.489
- SR-MRS: 0.573
- Transformer-XH: 0.641
Results on Hotpot QA Full-Wiki Test Set

Public STOA: BERT + GCN
Contemporary work: Better IR

Answer Exact Match

Contemporary work: Better IR

Answer F1

Official Baseline  Cognitive Graph QA  SR-MRS  Transformer-XH
Results on Hotpot QA Full-Wiki Test Set

Official Baseline
Cognitive Graph QA
SR-MRS
Transformer-XH

Public STOA: BERT + GCN
Contemporary work: Better IR
Transformer-XH: Simpler & Stronger

Answer Exact Match
Answer F1

0.24
0.371
0.24
0.329
0.453
0.516
0.329
0.573
0.516
0.641

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Official Baseline
Cognitive Graph QA
SR-MRS
Transformer-XH
Results on FEVER 1.0

Contemporary work:
Top leaderboard
Transformer-XH: Simple but powerful adaption of Transformer for structured text

- It learns better representations via extra hop attention
- Strong performance on multiple tasks

Thanks!