Optimized Transformer Models for FAQ Answering

Sonam Damani, Kedhar Nath Narahari, Ankush Chatterjee, Manish Gupta, Puneet Agrawal

{sodamani, kedharn, anchatte, gmanish, punagr}@microsoft.com
Microsoft, India
Overview

- Both enterprises and users find chat bots super useful.
- We discuss the problem of FAQ answering which is central to designing a retrieval-based informational chatbot.
  - Given a set of FAQ pages $s$ for an enterprise, and a user query, we need to find the best matching question-answer pairs from $s$.
- We experiment with multiple transformer based deep learning models, and also propose a novel MT-DNN-based architecture, which we call Masked MT-DNN (or MMTDNN).
- Further, we propose an improved knowledge distillation component to achieve $\sim 2.4x$ reduction in model-size and $\sim 7x$ reduction in runtime while maintaining similar accuracy.
- On SemEval 2017 dataset, our approach provides an NDCG@1 of 83.1.
- On another $\sim 281K$ instances dataset, our distilled 174 MB model provides an NDCG@1 of 75.08 with a CPU runtime of mere 31 ms establishing a new state-of-the-art for FAQ answering.
Motivation

• Why enterprises need chat bots?
  • Reducing agent costs in the call center
  • FAQ pages are difficult to extract info quickly
    • User has to scan through a long list of QA pairs.
    • FAQs in a list may be poorly organized and not semantically grouped.
    • Multiple FAQs may answer the query, and the user must look out for a QA pair that answers the question with the right level of specificity.
    • An FAQ list may sometimes be scattered over several documents.

• Why users like chat bots?
  • Poorly managed call center or mismatching working hours for global customers, could lead to long wait times
  • Alternatively, users pose such queries on community question answering (cQA) forums, or contact businesses over slow media like emails or phone calls

• To provide correct information instantly at much lower operating costs, retrieval based chatbots that can match user queries with content on FAQ pages are highly desirable.
Problem Definition

- Given a set D of question answer pairs and a user query q, our goal is to rank question-answer pairs in D.
- Top K QA pairs with high scores are returned to the user.
Related Work

- **Data Mining for FAQ Web Pages**
  - FAQ mining using list detection algorithms [11, 14]
  - answering questions using FAQ web pages [8, 11, 22, 28]
    - traditional feature engineering for surfacing statistical/semantic similarities between query and questions (FAQ-Finder [8], Auto-FAQ [28], [2, 11, 13, 21, 23])
      - considered similarity between query and questions and not query-answer similarity
    - deep learning based methods
      - attention-based Question Condensing Networks (QCN) to align a question-answer pair where the question is composed of a subject and a body [29]
      - SymBiMPM (BiLSTMs with multiperspective matching blocks) [7]
  - navigational interface for Frequently Asked Question (FAQ) pages [20]
  - Completeness of FAQ pages [3]

- **Applications of Transformer Models**
  - several architectures like BERT [6], MT-DNN [15] etc.
  - Several NLP tasks like text classification, textual entailment, machine translation, word sense disambiguation, etc.

- **Model Compression**
  - pruning, quantization, knowledge distillation, and low rank factorization.
Approach

• Given a question-answer (QA) database, when a user query \( q \) arrives, we first compute a list of candidate QA pairs which have high BM25 score with respect to the query.

• These candidate QA pairs, along with the original query, are scored using various methods.

• Top K QA pairs with high scores are returned to the user.
QA pair scoring methods

(A) BiLSTMs with attention
(B) SymBiMPM (adapted from [7])
(C) BERT/MT-DNN
(D) MMT-DNN

SymBiMPM = Symmetric Bilateral Multi-Perspective Matching Block
Proposed Methods

- BERT [6]
  - transformer encoder with 12 layers.
  - pretrained on Books Corpus and Wikipedia using the MLM (masked language model) and the next sentence prediction (NSP) loss functions.
  - The query, question and answer are concatenated into a sequence and are separated with a special “SEP” token.
  - The sequence is prepended with a “CLS” token.
  - The representation C for the “CLS” token from the last encoder layer is used for classification by connecting it to an output softmax layer.
  - We finetune the pre-trained model using labeled training data for the FAQ answering task.
Proposed Methods

- MT-DNN [15]
  
  - extends BERT by further pre-training it with large amounts of cross-task data.
  
  - 12 layer transformer encoder where the BERT model has been further pre-trained using single sentence classification, text similarity, pairwise text classification and relevance ranking tasks.
  
  - The representation C for the “CLS” token from the last encoder layer is used for classification by connecting it to an output softmax layer.
  
  - We finetune the pre-trained model using labeled training data for the FAQ answering task.
Proposed Methods

- **MMT-DNN**
  - $\text{Encoder}_1$ consists of $l$ encoder layers, while $\text{encoder}_2$ contains $12-l$ layers. $l$ is a hyper-parameter tuned on validation data.
  - Both $\text{encoder}_2$ blocks share weights.
  - The C token from both these $\text{encoder}_2$ blocks are concatenated and connected to an output softmax layer.
Proposed Methods

• Knowledge Distillation

To facilitate gradual transfer of knowledge, the distillation from MMT-DNN-12 to MTDNN-3 is done in a chain of steps where MMT-DNN-12 is first distilled to a MT-DNN9, then to MT-DNN-6 and finally to an MT-DNN-3 student model (DSM)

• TVM compiler optimizations [4]
Datasets

- SemEval-2017
  - Task 3 data had the QA pairs grouped by search query terms.
  - Transformed into FAQ Retrieval format where FAQs are ranked for a query are classified as Good, Average or Bad
  - Standard train, dev, test splits provided by the task organizers.

- FAQ Search Dataset (FSD)
  - Created using ~30K Bing queries leading to clicks to FAQ pages.
  - Query was compared to all QA pairs extracted from clicked FAQ pages using BM25 to extract a max of top 15 QA pairs.
  - Judged into 3 classes (Good, Average or Bad) with three-way redundancy.
  - Multiple domains like airports, banks, supermarkets, tourism and administrative bodies
  - Queries and QA pairs of various sizes are considered and FAQ pages with varying number of QA pairs are included.

Dataset statistics (train / dev / test)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SemEval-2017</th>
<th>FSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Queries</td>
<td>266/72/70</td>
<td>20242/1966/7478</td>
</tr>
<tr>
<td>#Question Answer pairs</td>
<td>6711/1575/2313</td>
<td>1630/477/649</td>
</tr>
<tr>
<td>#Data points</td>
<td>9977/1851/2767</td>
<td>202969/22549/55751</td>
</tr>
<tr>
<td>Avg length of queries</td>
<td>41.2/37.9/43.7</td>
<td>7.2/9.3/9.5</td>
</tr>
<tr>
<td>Avg length of questions</td>
<td>50.4/47.2/49.1</td>
<td>7.7/9.8/7.9</td>
</tr>
<tr>
<td>Avg length of answers</td>
<td>48.9/45.1/46.3</td>
<td>61.4/55.3/57.9</td>
</tr>
</tbody>
</table>

http://alt.qcri.org/semeval2017/task3/
Experimental Settings

- GloVe embeddings for the query, question and answer.
- 4 Tesla V100-SXM2-32GB GPUs.
- Metric: NDCG@K
- For BiLSTMs in all baseline methods, the hidden layer size was 300.
- For transformer-based methods, the embedding size was fixed to 30522 and the input sequence length was fixed to 512 tokens.
Accuracy comparison across various methods

<table>
<thead>
<tr>
<th></th>
<th>SemEval-2017</th>
<th>FSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NDCG@1</td>
<td>NDCG@5</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTMs</td>
<td>36.62/38.83</td>
<td>38.43/43.17</td>
</tr>
<tr>
<td>SymBiMPM [7]</td>
<td>34.21/34.00</td>
<td>38.55/38.59</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT (pre-trained)</td>
<td>63.38/65.39</td>
<td>61.85/68.41</td>
</tr>
<tr>
<td>MT-DNN (pre-trained)</td>
<td>68.01/60.97</td>
<td>64.92/61.85</td>
</tr>
<tr>
<td>BERT (finetune)</td>
<td>68.01/69.22</td>
<td>65.19/68.61</td>
</tr>
<tr>
<td>MT-DNN (finetune)</td>
<td>70.22/82.49</td>
<td>67.06/81.79</td>
</tr>
<tr>
<td>MMT-DNN</td>
<td>71.03/84.71</td>
<td>70.67/82.59</td>
</tr>
</tbody>
</table>

- SymBiMPM [7] performs worse than BiLSTMs.
- For SemEval-2017 dataset, results are for two settings: (using just the query subject/using query subject+body).
  - query subject+body usually provides better accuracy
- Both BERT and MT-DNN benefit from finetuning across the two datasets.
- MMT-DNN outperforms all other methods by a significant margin
NDCG@K for the MMT-DNN approach

- With increase in K, accuracy improves for FSD
- For SemEval-2017 dataset, accuracy somewhat reduces with increase in K
  - small size of the dataset implies that there are very few good answers matching any query.
- The accuracy is better at larger values of $l$.
  - It is useful to allow attention across question and answer in the first few layers but let the query-question and query-answer attention be learned separately at higher layers.
### Top 3 QA pairs returned by MMT-DNN

#### Q1 from SemEval-2017 dataset: working permit...

1. Do I need working permit since I have residence visa in Qatar under husband sponsor? 2. Without working permit expat’s wife could not work in Qatar? ...

<table>
<thead>
<tr>
<th>Q1</th>
<th>Work permit for husband? I am thinking of sponsoring my husband to live in Qatar. I heard that if he gets a job, he will need to get a work permit. Are husbands able to get a work permit? ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>If he is on a family visa he needs to find a job first so that the company who will hire him will be the one to process his work permit. ...</td>
</tr>
</tbody>
</table>

#### Q2 from FSD: I’ve paid for my parking but my flight is delayed

1. What happens if I exit the car park prior to my confirmed booking time? 

<table>
<thead>
<tr>
<th>Q2</th>
<th>What happens if I exit the car park prior to my confirmed booking time?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>If for whatever reason you cannot exit the car park in your confirmed booking time (e.g., you haven’t returned due a cancelled flight), the credit card or debit card that you use to exit the car park (i.e. your nominated card) will be debited with the cost of the additional time, based on the rates displayed at the entry to the car park.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q2</th>
<th>What happens if I enter the car park prior to my confirmed booking time?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>If you enter the car park before your confirmed booking time, or exit the car park later than your confirmed booking time, the credit card or debit card that you use to exit the car park (i.e. your nominated card) will be debited with the cost of the additional time, based on the rates displayed at the entry to the car park.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q3</th>
<th>How do I amend or cancel my booking?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>You may cancel your Booking, for any reason at any time up to 24 hours before the start of the Booking Period. To do this, ...</td>
</tr>
</tbody>
</table>
Attention Visualization for MMT-DNN

Visualization of a few heads for various examples for the last encoder layer of our best approach. (left): query-question, (middle): query-answer, (right): question-answer
Error Analysis

- The most confusing category is the “Average class” with lowest precision and recall.
  - This does not impact ranking significantly especially in cases where there are enough “good” QA pairs for a query.

<table>
<thead>
<tr>
<th>Category</th>
<th>Meaning</th>
<th>%</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Mismatch</td>
<td>q and Q/A refer to a different main entity</td>
<td>29</td>
<td>q: “What is best mall in Doha to buy good furniture?”, Q: “where to buy good abhaya in doha”</td>
</tr>
<tr>
<td>Generalization</td>
<td>q and Q/A have entities with “is a” relationship</td>
<td>7</td>
<td>q: “Any aquapark in Doha?”, Q: “any water theme park in qatar?”</td>
</tr>
<tr>
<td>Intent Mismatch</td>
<td>q and Q/A have different intents</td>
<td>5</td>
<td>q: “What is best mall in Doha to buy good furniture? ... showrooms ...”, Q: “Where to buy used furniture? .. cheap ...”</td>
</tr>
<tr>
<td>Negation</td>
<td>q and Q/A have opposite intents</td>
<td>7</td>
<td>q: “Is there any Carrefour which is open ??”, Q: “any other good supermarkets apart from Carrefour”</td>
</tr>
<tr>
<td>Verbose Match</td>
<td>q and Q/A match on unimportant parts</td>
<td>52</td>
<td>q: “Is it good offer? Hi Frds;i QA supervisor with 8 years exp in pharmaceutal have got job offer from Qatar pharma company; Salary which they have offered to me is 5000QAR...”, Q: “Is it a good offer? Dear all; I need your help please :) ; i got an offer from Habtoor lighton group for Planning Engineer position. They are offering 10K ...”</td>
</tr>
</tbody>
</table>
Knowledge Distillation (KD)

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>CPU Runtime</th>
<th>NDCG@1</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMT-DNN-12</td>
<td>417 MB</td>
<td>225 ms</td>
<td>75.38</td>
<td>78.59</td>
<td>80.24</td>
</tr>
<tr>
<td>MT-DNN-9</td>
<td>336 MB</td>
<td>210 ms</td>
<td>76.28</td>
<td>78.83</td>
<td>80.48</td>
</tr>
<tr>
<td>MT-DNN-6</td>
<td>255 MB</td>
<td>143 ms</td>
<td>74.55</td>
<td>77.88</td>
<td>79.76</td>
</tr>
<tr>
<td>MT-DNN-3</td>
<td>174 MB</td>
<td>68.9 ms</td>
<td>70.56</td>
<td>75.47</td>
<td>77.91</td>
</tr>
<tr>
<td>MT-DNN-3 (unlabeled data)</td>
<td>174 MB</td>
<td>68.9 ms</td>
<td>75.08</td>
<td>78.28</td>
<td>80.00</td>
</tr>
<tr>
<td>MT-DNN-3 (unlabeled data+TVM)</td>
<td>174 MB</td>
<td>31.4 ms</td>
<td>75.08</td>
<td>78.28</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Accuracy vs size and runtime latency comparison across various models for the knowledge distillation experiments (on FSD)

Initialization for KD for MT-DNN-3 model using MMT-DNN-12 layers or Random (on FSD)
Conclusion

- We proposed the use of transformer-based models like BERT and MT-DNN for solving the FAQ Answering task.
- We also proposed a novel MT-DNN architecture with masking, MMT-DNN, which establishes a new state-of-the-art for FAQ answering, as evaluated on two real world datasets.
- Further, we propose and experiment with an improved knowledge distillation strategy to reduce the model size and model runtime.
- Overall, the proposed techniques lead to models with high accuracy, small runtime, and small model size.