Using Common Sense Reasoning for Question Answering via Transfer Learning

Asawaree Bhide
GTID: 903461453

Rohit Gajawada
GTID: 903511115

Sahith Dambekodi
GTID: 903542538

Abstract

Current state-of-the-art question answering models are trained purely on the dataset of the task and with no external knowledge. Recent work in the area has started incorporating external knowledge dubbed ”common-sense” that gives models additional information that it should know regardless of the data its being trained on. For example, the question ”Does a boat travel on water or land?” is a very basic question that humans can answer and should be similarly easy for trained models. On a similar note, just as humans can derive information from different tasks to help with their current one, trained models should be able to share information across tasks. Our goal is to implement these two concepts (commonsense reasoning and multitask learning) across different question answering datasets and analyze the advantages and disadvantages of each approach. We have used COMET to generate a knowledge graph for the commonsense reasoning and used BERT for 3 task specific models. We plan to augment the task specific and multitask models with this commonsense reasoning and observe change in performance.

1 Goal

The goal for our project is to be able to incorporate common sense into existing QA models without negatively affecting the performance of the QA on that task. While we intend to use only QA formats for our analysis, we will apply our method to different tasks that have been framed in a question answering setting. For example, sentiment analysis can be framed as a question of whether the sentence is positive or negative. This will help in pursuing our second goal where we also want to apply these methods in both a single task setting and multi-task setting and analyze whether the inclusion of commonsense knowledge helps in better generalization across tasks. We are not aiming to solve Commonsense QA tasks as it will be infeasible given our resources and instead aim to analyze the effect of common sense knowledge on existing simpler question answering tasks.

The changes from our original goal consist of a reduction on the focus of zero-shot and few-shot methodology. We instead put an emphasis on applying commonsense in a multi-task setting. We are also reducing the chat bot aspect of our original goal to a stretch goal that we will pursue if time permits.

2 Overview of Complete Method

The goal of our project requires multiple different models to be trained and contrasted. We are doing this training in multiple steps.

1. Train 3 different models that each perform a specific task (question answering, multi-genre natural language inference and sentiment analysis) and measure the performance of each model. We have completed this step using the datasets mentioned below.

2. Train a commonsense model to augment each task-specific model with commonsense reasoning and measure performance. Compare the performance of task-specific models incorporating commonsense to performance of the task-specific models without commonsense. We hypothesize that models with commonsense should perform better for some of the 3 tasks. As each of the 3 tasks can be modelled as a question-answering problem, we want to see if the question can be used to select the top relevant inferences made by a commonsense model to generate more meaningful answers.

3. Train a multitask model that performs all 3 tasks (with and without commonsense inference). The idea of multitask is that doing

multiple tasks at the same time would benefit each individual task due to shared knowledge. We reason that using common sense would provide a good prior knowledge to benefit multitask models more. We plan to train the BERT based model with separate heads for each task. The BERT base of the model would be shared across each task (hard parameter sharing). The tasks would be sampled in a round robin fashion to prevent catastrophic forgetting. Once this is done, we would like to experiment with two more multitask learning strategies (9; 10) and see if common sense would be useful in these setups.

We will use the DecaNLP (8) challenge scheme for training/testing the different models. While not all the tasks in this challenge are traditional question answering (for example, sentiment analysis has been framed as a question answering task), getting good results on the data set will be the final goal.

3 Progress made

3.1 Data

We have identified three datasets (SQuAD, SST-2, MNLI) for the three different tasks we will be using. Eventually all of these tasks/datasets will be framed in a question answer format so that we can more smoothly apply the same model to all three tasks in a multi-task setting. These datasets are specifically for the single-task setting. We chose these three datasets to be distinct so that we can get varying results for the application of common-sense to each one. Also, since these tasks do not completely overlap, there is scope for greater variance in the multi-task results which we can analyze.

1. Question answering: (13)

2. Sentiment Analysis: SST-2 (14)

3. Multi-Genre Natural Language Inference: MNLI (15)

Examples of the data:

1. SQuAD:
   Question: What is a major importance of Southern California in relation to California and the US?
   Context:. . . Southern California is a major economic center for the state of California and US...
   Answer: major economic center

2. SST-2:
   Input: A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.
   Answer: positive

3. MNLI:
   Input: Hypothesis: Product and geography are what make cream skimming work. Premise: Conceptually cream skimming has two basic dimensions — product and geography.
   Answer: Entailment

We fine-tuned the BERT model on the GLUE benchmark for the SST-2 and MNLI tasks.
In addition we used the ATOMIC(12) dataset for training our commonsense model. ATOMIC contains a variety of social commonsense knowledge around specific event prompts like event causes, effect of the agent's actions and 7 more prompts which are all distinct.

Dataset	Train	Test	Metric
SQuAD	105k	5.4k	Accuracy
SST-2	67k	1.8k	Accuracy
MNLI	393k	20k	Accuracy
ATOMIC	710k	87k	Perplexity

For further experiments with regards to the multi task setting we will be using the DecaNLP dataset which frames all these tasks in a question answer format. This will give us a setting where common-sense reasoning can be easily applied to all three tasks.

3.2 Method and Model Overview

We are using BERT for the task-specific models on each of the 3 tasks. SST-2 and MNLI models have been fine tuned on the GLUE benchmark. Using BERT, a QA model can be trained by learning two extra vectors that mark the beginning and the end of the answer for marking the answer in the context.
The incorporation of common-sense knowledge will be done in two distinct ways. A naive method and a graph reasoning method. Given a question and a context, the context will be sent as input to the commonsense model which will be COMET. COMET will generate inferences based on each of the 9 relation types present in ATOMIC. In the

naive method, these inferences are then concatenated to the original context and sent as input to the BERT model that has been trained on the particular question answering task. The graph reasoning method will be applied by taking the generated knowledge graph from the COMET output and passing it through a network as outlined in KAGNET. KAGNET consists of a Graph Convolution Network with an LSTM with the knowledge graph as input. The output from this is then concatenated with the BERT encoder output and sent to a final MLP layer to evaluate the answer with the highest weight. The graph reasoning method will be done as a stretch goal as it is the most complex.

Our method combines a dynamically generated knowledge graph as additional input context to a BERT encoder model for question answering. To our knowledge, this particular method has not been applied to these specific datasets.

3.3 Preliminary Results

The final model that we are aiming to train will incorporate common-sense knowledge into a BERT based model that has been fine tuned for question answering. Currently we have trained the commonsense knowledge graph completion model and the task-specific models separately.

For the common-sense reasoning aspect of the project we have used the COMET model. COMET is based on the transformer language architecture, specifically GPT, and was trained on the ATOMIC dataset. The training process is exactly the same as described in the original paper and uses the code provided by the authors in the publicly available repository. Specifically we use a GPT model and initialize COMET with 12 layers, 768-dimensional hidden states, and 12 attention heads. We use a dropout rate of 0.1 and use GeLU units as activation functions. During training, our batch size is 64.

Example Output of COMET:
Input Event: A stirring, funny transporting reimagining of Beauty and the Beast and 1930s horror film.
Output with Relation oReact (How others react to this event):

- People were happy.
- People were scared.
- People were amused.

We use the BERT base model as the base for all our tasks. This model has 12 Transformer blocks, hidden size of 768 and 12 self-attention heads. In the last layer of BERT, a separate head is added for each task and the training occurs in a round robin fashion. Once we finish this multi task training setup, the three heads would become one head when we port our model to the decaNLP setup where the multi task effectively becomes multi domain.

3.3.1 Success metrics and baseline models

- Accuracy achieved on SQuAD was 81.42.
- Accuracy achieved on the Sentiment Analysis task using SST-2 dataset was 91.62.
- Accuracy achieved on MNLI was 80.61.
- Perplexity Per Example achieved on COMET trained on ATOMIC = 11.32 . The result is comparable to the the result of 11.14 reported in the paper.

3.4 Work Division

Asawaree and Rohit: Data collection, coding/training all the task specific models and setting up multi-task learning setup.
Sahith: Coding and training commonsense models along with evaluation of task specific models.

4 Future Tasks

4.1 Short Term Tasks

- As mentioned originally we will be testing commonsense in a multi-task setting and to make the process easier we will be using the DecaNLP dataset. We will need to port all our existing single task models into the DecaNLP structure where training occurs in a round robin fashion and evaluate. We will need to achieve similar results as our original models.
- COMET has been trained on the ATOMIC dataset which has input in the format of ”PersonX pushes the car of PersonY”. The nouns have been replaced by generic PersonX and PersonY. To ensure that results are consistent we will need to replace the nouns in the questions and context of the data input with these generic words. We will initially do this in a naive way by identifying the proper nouns by their capital letter and doing the replacement.

We will later transition into using a named entity recognition model to identify the words and automatically do the replacement as this will be more robust. This will be part of the pre-processing step before the input is sent to COMET.

- Once the pre-processing step is complete we will be testing the incorporation of COMET inferences with the single task models. For the initial set of experiments we will be using these inferences in a naive manner as detailed in the model overview. This will allow us to obtain results quickly which we can use for comparison with the multi-task settings.

4.2 Long Term Tasks

Once we have the results for common sense incorporation for task specific and multi task models, we would like to understand on which setups common sense seems to work. We would like to incorporate common sense in different ways like directly adding it to context, concatenating common sense features within the model, etc.

4.3 Additional Future Tasks

- COMET can also be trained using ConceptNet instead of ATOMIC. This would provide us a different set of inferences based on the context and can be used to either replace the previous COMET model trained on ATOMIC or augment the inferences that are already being produced. However, ConceptNet requires more pre-processing in terms of the input required and we will need to come up with the correct methods so that we can use it.

- If we have completed our previous tasks we will also attempt to apply our model to a dataset that is designed for CommonSense Question Answering. We will apply it to STORYCOMMONSENSE Dataset(16). This dataset consists of short 5-sentence stories with annotated motivations and emotional responses whose labels are drawn from classical theories of psychology. The task will be to predict these labels.

- Use the graph reasoning approach detailed in the model overview. We will implement KAGNET and replace the knowledge graph being used in the paper (which was ConceptNet) with the dynamically generated knowledge graph from COMET.

4.4 Timeline for future tasks

April 3rd - April 10th: Finish multi task learning setup and incorporating common sense into task specific models
April 10th - April 15th: Incorporating common sense into multi-task models
April 15th - April 20th: Perform graph reasoning on the common sense inferences and feed this into our previously created models
April 20th - April 23rd: Final Presentation and Report

4.5 Project changes in case of failure

If the incorporation of common sense into the task models does not result in an improvement in performance, we will do zero shot learning without the task specific BERT based models like in ”Dynamic Knowledge Graph Construction for Zero-shot Commonsense Question Answering[2]”. The dataset we will use is STORYCOMMONSENSE dataset as detailed previously. Here, we would do graph based reasoning on the common sense inference graph created via COMET by using the question as a relation and the context as the question. At every level in the graph created, we would calculated scores based on their softmax probabilities weighted by the current score and the score at the previous level.

4.5.1 Division of tasks and labor for future tasks

Asawaree: Porting our task-specific models for each of the 3 tasks into DecaNLP structure.
Rohit and Sahith: Coding and training multitask model approaches. Incorporating commonsense inferences with the existing task specific and multitask models along with evaluation of these models.

References

[1] Bosselut, Antoine, et al. ”Comet: Commonsense transformers for automatic knowledge graph construction.” arXiv preprint arXiv:1906.05317 (2019).

[2] Bosselut, Antoine, and Yejin Choi. ”Dynamic Knowledge Graph Construction for Zero-shot Commonsense Question Answering.” arXiv preprint arXiv:1911.03876 (2019).

[3] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[4] Radford, Alec, et al. ”Language models are unsupervised multitask learners.” OpenAI Blog 1.8 (2019): 9.

[5] Lin, Bill Yuchen, et al. ”Kagnet: Knowledge-aware graph networks for commonsense reasoning.” arXiv preprint arXiv:1909.02151 (2019).

[6] Lv, Shangwen, et al. ”Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering.” arXiv preprint arXiv:1909.05311 (2019).

[7] Tafjord, Oyvind, et al. ”Quarel: A dataset and models for answering questions about qualitative relationships.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[8] McCann, Bryan, et al. ”The natural language decathlon: Multitask learning as question answering.” arXiv preprint arXiv:1806.08730 (2018).

[9] Augenstein, Isabelle, Sebastian Ruder, and Anders Søgaard. ”Multi-task learning of pairwise sequence classification tasks over disparate label spaces.” arXiv preprint arXiv:1802.09913 (2018).

[10] Sanh, Victor, Thomas Wolf, and Sebastian Ruder. ”A hierarchical multi-task approach for learning embeddings from semantic tasks.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[11] Liu, X., He, P., Chen, W., Gao, J. (2019). Multitask deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504.

[12] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., ... Choi, Y. (2019, July). Atomic: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 3027-3035).

[13] Rajpurkar, Pranav, et al. ”Squad: 100,000+ questions for machine comprehension of text.” arXiv preprint arXiv:1606.05250 (2016).

[14] A. Radford, R. Józefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. CoRR, abs/1704.01444, 2017.

[15] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. CoRR, abs/1704.05426, 2017

[16] Rashkin, H., Bosselut, A., Sap, M., Knight, K., Choi, Y. (2018). Modeling naive psychology of characters in simple commonsense stories. arXiv preprint arXiv:1805.06533.