Multi-Sense: Commonsense Enrichment through Multi-Task Learning

Anonymous authors

Abstract

A recent discovery showed that large pre-trained language models (LM) are capable of encoding knowledge into their parameters, as a result, serving as powerful zero-shot/few-shot learners of knowledge-dependent tasks. However, the commonsense knowledge of such models is relatively under-explored despite their importance on various natural language tasks. In this paper, we propose Multi-Sense model that is enriched with commonsense by multi-task learning (MTL) on commonsense question-answering and natural language inference (NLI) tasks. We hypothesize that supervision from tasks that require commonsense reasoning ability will implicitly help strengthen the commonsense representation within the model parameters. Empirical results demonstrate that the multi-task commonsense enrichment step is helpful in downstream tasks (i.e., reading comprehension, fact-checking). In addition, we demonstrate the strength of Multi-Sense in low resource settings by conducting a few-shot learning analysis.

1. Introduction

Natural language inference and commonsense reasoning are important requirement for achieving the human-level text comprehension of machine learning (ML) models. Many progress has been made from the NLP community, yet, there still remains room for improvement.

The most common take on dealing with commonsense reasoning tasks are to leverage external sources of knowledge to provide the necessary prior knowledge to the ML model. For example, [Lv et al., 2020, Zhou et al., 2019] use knowledge graphs and [Lv et al., 2020] use Wikipedia to serve as this external knowledge source. However, such dependence on external knowledge adds additional computational and resource burden to the model. Considering the current trend of rapidly increasing size of NLP models, the need to deal with additional large knowledge bases will act as a critical barrier; for instance, it will be difficult to fit a whole model into the GPU, and the training/inference time will increase.

Recently, Petroni et al. demonstrated that large pre-trained language models (LM), such as BERT [Devlin et al., 2019] can encode some factual knowledge into its parameters. Extending from this, many work has explored and demonstrated the potential of solely relying on the encoded knowledge of transformer parameters to solve various tasks such as open-domain QA (Open-QA) [Rajpurkar et al., 2016], and fact-checking [Lee et al., 2020] tasks.
Yet, little focus was made to explicitly solve the commonsense reasoning tasks with parameterized knowledge. Previous work hypothesised that the pre-trained LMs are learning both the factual and commonsense knowledge, but barely any experiment/analysis were conducted regarding the commonsense aspect. In addition, the fact that BERT-large [Devlin et al., 2019] only obtains 56% in commonsense tasks such as CommonsenseQA [Talmor et al., 2019] suggests that knowledge in pre-trained transformer-based LMs are insufficient for commonsense reasoning task. Therefore, one aim of our work is to understand the feasibility of commonsense reasoning without any external knowledge.

Moreover, we hypothesized that the unsupervised language modelling objective (i.e. next token or masked token prediction) might not be enough for the most maximal embedding/encoding of commonsense knowledge into the model parameters. The rationale is that the commonsense includes more complex forms of knowledge such as cause-and-effect or entailment relationship. We believe the model requires some supervision to learn the connection between the already-acquired-knowledge from LM pre-training step. Therefore, we propose to have an additional knowledge priming step with direct supervision from multiple commonsense dataset.

Multi-task learning and transfer learning are commonly used to overcome the limitation of train-data insufficiency. For instance, Liu et al. and Jin et al. showed impressive result in GLUE [Wang et al., 2018] and multiple choice QA tasks respectively through multi-task learning. Since single dataset cannot cover the entirety of all commonsense knowledge, we also do multi-task learning with many commonsense dataset and other related dataset in to leverage the “power of the crowd”.

We propose Multi-Sense, a pre-trained transformer model that is further enriched with commonsense knowledge through multi-task learning. Similar to any other pre-trained models, it can be further fine-tuned with other tasks that requires common-sense knowledge. In this work, we experiment our Multi-Sense model on open-domain question-answering (Open-QA), fact-checking (FEVER) and entailment tasks. We also carry out ablation studies and analysis to show that our proposed approach is helpful.

To summarize, our main contribution are the following:

- One of the first work to focus on tackling common-sense reasoning tasks with parameterized knowledge base.

- Further enrichment of knowledge encoded in pre-trained transformer model through multi-task learning.

- Achieve competitive results across many dataset/tasks. For example, for DREAM task, compare to the baseline model [Clark et al., 2020] we gain absolute 4.1% in accuracy.

- We will release the pre-trained Multi-Sense model for others to use in closely related tasks.

2. Related Work

Parameterized Knowledge Pre-trained language models have recently lead to significant advancements in a wide variety of NLP tasks, including question-answering, commonsense reasoning, and semantic relatedness [Devlin et al., 2019, Radford et al., 2019, Peters et al., 2018, Radford et al., 2018]. These models are typically trained on a large corpus of documents, which often includes Wikipedia and other websites crawled from the web. Recently, a number of works have found
that LMs store a surprising amount of world knowledge, focusing particularly on the task of open-domain question answering [Petroni et al., 2019, Roberts et al., 2020]. Going further, Guu et al., Roberts et al. show that task-specific fine-tuning of LM can achieve impressive results, proving the power of LMs. Although some of these works mention “commonsense” knowledge, most of the actual evaluation and experiments were conducted on tasks that require more “factual” knowledge.

**Commonsense Knowledge and Reasoning**  External knowledge is widely used in commonsense reasoning tasks, such as commonsense QA where some key information is omitted in the context of the data samples. In many existing works, a commonsense knowledge base, such as ConceptNet [Speer et al., 2017] is retrieved in the symbolic space to form a subgraph and convert to vector representations using Graph Convolution Networks and various attention mechanisms [Bauer et al., 2018, Lin et al., 2019, Lv et al., 2020]. Our work mainly differs from these works for the fact that our model does not depend on any external commonsense knowledge base and graphs.

Another line of work attempts to combine external knowledge with neural models for commonsense reasoning or question answering. Zhong et al. proposes to learn a scoring function that captures the relationship between commonsense concepts that can be integrated into existing neural models. [Xia et al., 2019] improves reading comprehension performance by integrating some external knowledge through multi-tasking learning and improving the modeling of the interactions between reading-passage and candidate answer options. They use two auxiliary relation-aware tasks that aim to predict (i) the existence of any commonsense relation and (ii) the type of relationship between two words. These work demonstrate promising results, but they are confined to one specific task. Our work instead focuses on trying to obtain a pre-trained model that can easily be used for various tasks.

Multi-task learning is leveraged as one of the common training strategies to combine external knowledge with neural models for commonsense reasoning. In [Xia et al., 2019], two relation-aware tasks are designed to induce the model to learn the relational information between concepts. In the training procedure, Xia et al. adopt multi-task learning on the main task and the two auxiliary tasks for sake of better interactions between the context and the answers.

There is a line of work that leverages large-scale multi-task learning (MTL) to solve NLP problems. Liu et al. propose a MT-DNN model that uses MTL to learn a general feature representations that is generalizable across multiple natural language understanding tasks (i.e., GLUE [Wang et al., 2018]). Similarly, Jin et al. also utilizing MTL to train a model that specializes in the multiple-choice question-answering task.

Khashabi et al. propose UnifiedQA. Khashabi et al. unify 4 formats of 17 QA datasets and train all QA datasets simultaneously in the same format. UnifiedQA model improves the performance on QA tasks. However, training data of UnifiedQA model is only QA tasks, it shows reasoning skills of the model can be improved with QA datasets.

Lin et al. leverages ConceptNet to retrieve related knowledge and obtain graph representations with a novel knowledge-aware graph network module. Lv et al. propose Graph-based reasoning over heterogeneous external knowledge for commonsenseQA. Both ConceptNet and the Wikipedia plain text are extracted into graph structure and encoded by a graph-based model. Relational commonsense knowledge is also unitized to enhance the QA process [Zhong et al., 2019a, Xia et al., 2019] as a source of supportive information.
3. The Proposed Multi-Sense Model

In this regard, we combine commonsense datasets and NLU datasets together to learn general knowledge and reasoning in the multi-task learning stage.

**Training Phases** There are two phases involved in our work. The first phase is to obtain knowledge-enriched shared Multi-Sense Encoder through multi-task learning. The second phase is to fine-tune the Multi-Sense model on the target dataset with a task-specific layer on the top of the a Multi-sense encoder. The results reported in Table 2 are from these classifiers.

3.1 Train Datasets/Tasks

In this section, we list the dataset used for enrichment step of Multi-Sense. Note that since single dataset cannot cover the entirety of all commonsense knowledge, we leverage multiple commonsense/NLI dataset to rely on the “power of the crowd”.

We leverage multiple commonsense and NLI dataset for enriching the Multi-Sense model to rely on the “power of the crowd” because a single dataset is not enough to cover the entirety of all commonsense knowledge. We used four Natural Language Inference (NLI) tasks from GLUE [Wang et al., 2018] which are MNLI [Williams et al., 2017], QNLI, WNLI, and RTE. The goal of NLI tasks is to predict the entailment relationship of given hypothesis-premise sentence pair (i.e., entail, neutral, or contradict). In addition, we leverage four commonsense question answering tasks which are Swag [Zellers et al., 2018], CommonsenseQA [Talmor et al., 2019], CosmosQA [Huang et al., 2019] and SocialIQA [Sap et al., 2019].

3.2 Input

We follow the input representation of BERT [Devlin et al., 2019]. The start, end, and separation of text pairs are indicated with special tokens [CLS] and [SEP]. For example, all NLI-related tasks have \{Premise, Hypothesis\} text pairs, formatted to be “[CLS] Premise [SEP] Hypothesis [SEP]”.

3.3 Model

Multi-Sense Model consists of a shared encoder layer and task-specific classification layers.

**Multi-Sense Encoder** The shared encoder is initialized with the large pre-trained language model ELECTRA [Clark et al., 2020]. We utilize ELECTRA because it was one of the strongest models at the start of this project. Note that our methodology is model agnostic.

**Task Specific Classifiers** We add a Multi-layer perceptron (MLP) on top of the Multi-Sense Encoder to fine-tune for each of specific tasks.

3.4 Loss Functions

For the loss function, we normally leverage cross-entropy function. In this case, let \( \hat{y} \) be the predictions, and let \( y \) be the labels. Then the loss value of cross entropy function \( Loss_{CE} \) for the task model is

\[
Loss_{CE} = - \sum y \log \hat{y}
\]  

(1)

However, we leverages mean square error functions for regression tasks. In this case, let \( \hat{y} \) be the predictions, and let \( y \) be the labels. Then the loss value of mean square error function \( Loss_{mse} \) for
Figure 1: Overall framework of Multi-Sense model. Different inputs from multiple datasets are forwarded to the shared transformer layers. Based on the inputs, model arranges different task-specific layers.

The task model is

$$Loss_{MSE} = \frac{1}{2} \sum (y - \hat{y})^2 \tag{2}$$

Let the loss from $n$ tasks be $loss_{task1}, \ldots, loss_{taskn}$, then the final loss $loss_{final}$ actually back propagate to the model is

$$Loss_{final} = \frac{loss_{task1} + \ldots + loss_{taskn}}{n} \tag{3}$$

### 3.5 Training steps

Details of the training steps are mentioned below:

1. Prepare batch, merge multiple inputs from multiple datasets.
2. Feed-forward to shared layer model.
3. Arrange contextual outputs from the shared layer to the task-specific layers accordingly.
4. Gather all outputs and losses from the every task layers and back-propagates to the model.

We briefly introduce format of input in entailment type, multiple choice question answering type, text similarity type, and text classification type. For the entailment task, we have premise sentence and hypothesis sentence and judge whether two sentences entail or not. We link the
Table 1: Training parameters

<table>
<thead>
<tr>
<th>Hyper parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>epoch</td>
<td>4</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.5e-5</td>
</tr>
<tr>
<td>Max Length</td>
<td>128</td>
</tr>
<tr>
<td>batch size</td>
<td>16</td>
</tr>
</tbody>
</table>

give premise and hypothesis into the [CLS] and [SEP] token. In addition, we input sequences [CLS]premise[SEP]hypothesis[SEP] and separate the premise and the hypothesis with segmentation embedding. For the multiple choice question answering type, we have question and multiple options and choice only one answer among the options. We link the give question and option into the [CLS] and [SEP] token. In addition, we input sequences [CLS]question[SEP]option[SEP] and separate the question and the options with segmentation embedding. After the linear layer we get all scores from the each options, then we leverage soft-max function to compare options and select the most likely one.

For the text similarity task, we have two different texts to compare with, we link text1 and text2 into the [CLS] and [SEP] token. Then we input sequences [CLS]text1[SEP]text2[SEP]. Also, we separate two texts with the segment embedding. Output of task specific layer is the similarity score of the two texts.

For the text classification task such as FEVER [Thorne et al., 2018] and CoLA [Wang et al., 2018]. We have only one text to input, hence we concatenate [CLS] and [SEP] token to the text. Then we classify input sentence. Overall framework and general information of our model is described in Figure 1. Details of hyper parameters in training procedure is described in Table 1.

4. Experiment and Analysis

4.1 Test datasets

The Multi-Sense model is tested on three different types of tasks, which all require knowledge and reasoning/inference ability. The first task is commonsense question answering of which the goal is to answer questions based on commonsense knowledge and reasoning ability - Swag [Zellers et al., 2018]. commonsenseQA [Talmor et al., 2019]. The second task is reading comprehension that requires general world knowledge, logical inference, and human level reading comprehension skills - DREAM [Sun et al., 2019] and SQuAD [Rajpurkar et al., 2016]. The third task is fact-checking which requires knowledge and reasoning ability to determine the veracity of the fact - [Thorne et al., 2018].

- DREAM: The Dialogue-based REAding comprehension exaMination dataset, DREAM [Sun et al., 2019] is collected from the English language examination for college entrance in China. The dataset consists of 10,197 questions for 6,444 multi-turn, multi-party dialogues written by educational experts in order to evaluate high school level Chinese student’s English comprehension skills [Sun et al., 2019]. DREAM covers many situations and scenarios of daily lives. In order to solve this dataset [Sun et al., 2019], general world knowledge, logical infer-
sense, and human level reading comprehension skills are required. An example from DREAM dataset is demonstrated in Table 2.

- **FEVER**: The Fact Extraction and VERification dataset [Thorne et al., 2018] is one of the most widely studied fact-checking dataset. It consist of 185,445 claims that is labeled as 'Supported’, ‘Refuted’, and ‘Not enough information.’ Most of related work contains knowledge bases such as Wikipedia, and document retrieval etc.

### 4.2 Training Details

**Multi-Sense Encoder**  For training the Multi-Sense Encoder, we used learning rate of $5e^{-6}$, batch size of 16, and maximum input sequence length of 128. We evaluated the performance on the validation set and stopped the training on the first performance drop (i.e., training epoch of 4). Note that we did not have enough computational power to do any parameter sweeping for the Multi-Sense training step. We used two 1080Ti GPUs and each epoch took approximately 20 hours.

**Task-Specific Classifier**  We fine-tune each of Multi-Sense-based task-specific classifiers by trying out the following hyper-parameter choices: learning rates of {$1e^{-5}, 5e^{-6}$}, batch sizes of {16, 24} and input sequence size of {128, 340}. There are a variety of optimal hyper-parameter choices for each of the tasks due to their different natures. For instance, a machine-reading comprehension task requires a larger input sequence length to fit the whole reading passage. We report the best performing results from these classifiers in Table 2.

### 4.3 Baseline Models

We compare to the following models to evaluate our Multi-Sense model:

- **SoTA models**: We report SoTA results from the following papers. [Liu et al., 2019] for Swag, [Khashabi et al., 2020b] for CommonsenseQA, [Wan, 2020] for DREAM, [Zhong et al., 2019b] for FEVER, and [Zhang et al., 2020] for SQuAD. Note that each of these SoTA models are specifically designed for each task.

- **ELECTRA**: strong baseline model obtained from fine-tuning pretraining ELECTRA [Clark et al., 2020] model with MLP classification head on top. Note that it has the same model architecture and setting as the Multi-Sense model, except for the commonsense enhancement step.

### 4.4 Results

**Multi-Sense vs. ELECTRA**  We compare our Multi-Sense with ELECTRA baseline to analyze the effectiveness of our proposed methodology in controlled setting (Table 2). Recall that ELECTRA and Multi-Sense are identical model with same setup except for the additional multi-task knowledge enrichment step applied to Multi-Sense. The consistent performance gain by Multi-Sense throughout all tasks ($2 - 6\%$ gain) empirically proves that the knowledge enrichment step was helpful to downstream task performance.

**Multi-Sense vs. SoTA**  From Table 2, we can observe that our Multi-Sense model outperforms SoTA models in both commonsense question-answering datasets (i.e. Swag, CommonsenseQA).
<table>
<thead>
<tr>
<th>Model / Task</th>
<th>Swag</th>
<th>CommonsenseQA</th>
<th>DREAM</th>
<th>FEVER</th>
<th>SQuAD (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELECTRA</td>
<td>93.2%</td>
<td>81.8%</td>
<td>82.7%</td>
<td>62.1%</td>
<td>91.3%</td>
</tr>
<tr>
<td>Multi-Sense</td>
<td>95.6%</td>
<td>84.2%</td>
<td>86.8%</td>
<td>68.2%</td>
<td>94.1%</td>
</tr>
<tr>
<td>SoTA</td>
<td>91.2%</td>
<td>79.1%†</td>
<td>91.8%‡</td>
<td>79.2%§</td>
<td>94.9%∥</td>
</tr>
</tbody>
</table>

Table 2: Comparison of Multi-Sense model, ELECTRA and State-of-the-art (SoTA) for five separate tasks that all require knowledge and reasoning/inference ability. The results are accuracy values except for SQuAD, which is F1 result. Note that all tasks except DREAM do not publicly release test-set. Thus, we report experimental results on the dev-set. The SoTA models for each of the task are from ¶ Liu et al. [2019], † Khashabi et al. [2020b], ‡ Wan [2020], § Zhong et al. [2019b] and ∥ Zhang et al. [2020] in an order from left to right.

<table>
<thead>
<tr>
<th>Few-shot Data %</th>
<th>DREAM</th>
<th>FEVER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELECTRA</td>
<td>Multi-Sense</td>
</tr>
<tr>
<td>1%</td>
<td>70.0%</td>
<td>74.5%</td>
</tr>
<tr>
<td>5%</td>
<td>78.3%</td>
<td>82.5%</td>
</tr>
<tr>
<td>10%</td>
<td>79.6%</td>
<td>83.2%</td>
</tr>
<tr>
<td>100%</td>
<td>82.7%</td>
<td>86.8%</td>
</tr>
</tbody>
</table>

Table 3: Few-shot learning accuracy result comparison between ELECTRA and Multi-Sense; Comparison of the before (ELECTRA) and after (Multi-Sense) the proposed multi-task knowledge enrichment step.

However, it underperforms the SoTA models in other tasks, especially in DREAM and FEVER. We believe this is because only 33% of the DREAM test set are commonsense-type questions, with the rest being other types such as arithmetic-type. Also, FEVER is a Wikipedia-based fact-checking dataset that is designed to rely not only on commonsense but mostly on factual knowledge from Wikipedia. Also, unlike our simple MLP classification layer, SOTA models have additional task-specific attention modules to aid the final classification. For example, DREAM SOTA [Wan, 2020] uses Dual Multi-head layer [Zhu et al., 2020], which has 13.5M parameters, before the classification layer.

5. Analysis

5.1 Error Analysis

One of the obvious ways to prove the Multi-Sense model’s enhanced knowledge is to leverage a task that involves external knowledge, such as fact-checking. By leveraging the FEVER task-specific Multi-Sense model, our model classifies a claim sentence to "Supports", "Refutes", or "Not Enough Info". Table 4 describes the result of the fact-checking classification. We extract the first 3 sentences from the FEVER dataset and manually generate the others. Classification results in Table
<table>
<thead>
<tr>
<th>Claim</th>
<th>Electra</th>
<th>Multi-Sense</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Homo sapiens live on the fourth planet from the Sun.&quot;</td>
<td>Supports</td>
<td>Refutes</td>
<td>Refutes</td>
</tr>
<tr>
<td>&quot;Uranium is a country.&quot;</td>
<td>Not Enough Info</td>
<td>Refutes</td>
<td>Refutes</td>
</tr>
<tr>
<td>&quot;Kung Fu Panda made more than $1 million on opening day.&quot;</td>
<td>Supports</td>
<td>Supports</td>
<td>Supports</td>
</tr>
<tr>
<td>&quot;Kung Fu Panda made less than $1 million on opening day.&quot;</td>
<td>Supports</td>
<td>Refutes</td>
<td>Refutes</td>
</tr>
<tr>
<td>&quot;Donald john trump is the 44th president of the united states.&quot;</td>
<td>Not Enough Info</td>
<td>Refutes</td>
<td>Refutes</td>
</tr>
<tr>
<td>&quot;Donald john trump is the 45th president of the united states.&quot;</td>
<td>Supports</td>
<td>Supports</td>
<td>Supports</td>
</tr>
<tr>
<td>&quot;Donald john trump is the 46th president of the united states.&quot;</td>
<td>Supports</td>
<td>Refutes</td>
<td>Refutes</td>
</tr>
</tbody>
</table>

Table 4: Fact-checking results of Multi-Sense model and ELECTRA [Clark et al., 2020] model

4 show how the Multi-Sense model outperforms the baseline model. Not only does the Multi-Sense model correctly classifies some of the knowledge required claims that the baseline model fails to classify, but also successfully classifies all alternative and converse claims such as claim No 4, and 5. The baseline language model fails to identify converse sentence for the claim no 3 and 5. One possible explanation is that the pre-trained language model tends to classify claims not based on the knowledge itself, but based on the likelihood of the sentence. Hence, the words "Kung Fu Panda” and "$1 million on opening day” seem very likely, the baseline model classifies both claim 3 and 4 as ”supports”.

5.2 Ablation Study

We conduct a brief ablation study to understand how different tasks/datasets contribute to the performance (Table 5). When only one of either commonsense or GLUE-NLI tasks are added, the performance changes are very minor - minor drop/rose in the performance. However, when both are added together, there are visible performances improvements in both Swag and CommonsenseQA tasks. We believe there are two explanations for this observation. First, there was a regularization effect from the joint training of two different tasks. Second, two different but closely related tasks helped each other and resulted in a synergy effect. Especially the logical inference aspect of the NLI tasks would have been extremely beneficial to the commonsense reasoning aspect.

5.3 Few-shot Analysis

We conjecture the proposed Multi-Sense model will produce richer representation that benefits the downstream tasks. To evaluate this aspect, we conducted few-shot learning experiments to compare the performance before and after our knowledge enrichment step (Table 3). The most notable observation is that Multi-Sense only requires 5% of DREAM and FEVER train data to achieve comparable performance to the ELECTRA model fully fine-tuned on 100% of data (refer to the gray
Table 5: Ablation Study of the Multi-task Knowledge Enrichment Step (Accuracy). “ELECTRA” refers to the vanilla ELECTRA model being used as a shared encoder. “+ Commonsense only” and “GLUE-NLI only” refer to the ELECTRA model additionally trained in the multi-task setting with the aforementioned commonsense and GLUE-NLI datasets respectively. “Both” refers to our proposed Multi-Sense model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Swag</th>
<th>CommonsenseQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELECTRA</td>
<td>93.2%</td>
<td>81.8%</td>
</tr>
<tr>
<td>+ Commonsense only</td>
<td>92.7%</td>
<td>80.8%</td>
</tr>
<tr>
<td>+ GLUE-NLI only</td>
<td>93.5%</td>
<td>82.6%</td>
</tr>
<tr>
<td>+ Both (Multi-Sense)</td>
<td><strong>95.6%</strong></td>
<td><strong>84.2%</strong></td>
</tr>
</tbody>
</table>

Moreover, we would like to highlight the fact that Multi-Sense-based classifiers were faster to converge than the Electra-based classifiers. This implies that the learnt representation from Multi-Sense encoder were more effective than vanilla Electra encoder for the knowledge-related tasks used in our experiments.

6. Conclusion

In this paper, we present Multi-Sense, a pre-trained model with enhanced commonsense knowledge from multi-task learning. Empirical results show the effectiveness of the proposed approach in aiding the downstream tasks such as reading comprehension and fact-checking. Moreover, the Multi-Sense model showed strength in a low resource setting. By leveraging Multi-Sense model, one can data-efficiently train task-specific classifiers. This means that Multi-Sense model enables data-efficient training of downstream tasks, lifting the requirement for expensive and time-consuming large labeled data creation.

Looking forward, we are interested in work involving machine reading comprehension tasks in its multi-task learning procedure. Since machine reading comprehension tasks require long sequence lengths for the input, it is hard to train multi-task models with longer sequences due to memory issues.

References


