Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng,
Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, Xuedong Huang
Microsoft Corporation
{yicxu,chezhu,shuowa,siqi.sun,chehao,
xiaodl,jfgao,penhe,nzeng,xdh}@microsoft.com

Abstract

Most of today's AI systems focus on using self-attention mechanisms and transformer architectures on large amounts of diverse data to achieve impressive performance gains. In this paper, we propose to augment the transformer architecture with an external attention mechanism to bring external knowledge and context to bear. By integrating external information into the prediction process, we hope to reduce the need for ever-larger models and increase the democratization of AI systems. We find that the proposed external attention mechanism can significantly improve the performance of existing AI systems, allowing practitioners to easily customize foundation AI models to many diverse downstream applications. In particular, we focus on the task of Commonsense Reasoning, demonstrating that the proposed external attention mechanism can augment existing transformer models and significantly improve the model's reasoning capabilities. The proposed system, Knowledgeable External Attention for commonsense Reasoning (KEAR), reaches human parity on the open CommonsenseQA research benchmark with an accuracy of 89.4% in comparison to the human accuracy of 88.9%.

1 Introduction

Transformers (Vaswani et al., 2017) have revolutionized many areas of AI with state-of-the-art performance in a wide range of tasks (Devlin et al., 2018; Dosovitskiy et al., 2020). The most notable and effective component in a Transformer model is the self-attention mechanism, which enables the model to dynamically leverage different parts of the input for computation with no information loss for even the most distant parts in input. With the success of pre-trained models (Devlin et al., 2018; Liu et al., 2019), the Transformer and its self-attention mechanism have been widely adopted as the cornerstone of foundation models trained on huge amounts of data (Bommasani et al., 2021).

One phenomenon found during the development of Transformer models is that models with larger size tend to have better learning abilities, especially when combined with large-scale data (Kaplan et al., 2020). This has prompted the recent boom of super large Transformer models, ranging from BERT (Devlin et al., 2018) with 110 million parameters, to GPT-3 (Brown et al., 2020) with 175 billion parameters. Nevertheless, numerous studies have shown that the corresponding understanding and generation capabilities of these huge models are still behind humans (Bommasani et al., 2021). Furthermore, the sheer size of these models already poses serious practical challenges in utilization, deployment, interpretation, and environmental impact (Patterson et al., 2021). Thus, the recent "scaling-up" approach to Transformer-based NLP modeling is unsustainable and has been questioned in recent studies (Bommasani et al., 2021).

In this paper, we take a step back and examine the mechanism of current Transformer-based models. Self-attention was designed to allow the model to better analyze the inner structure of input data, and the model is trained to have its parameters grasp and memorize all the content and patterns of the training data. When the model is given a novel input X, the implicitly stored knowledge in the parameters about related information is activated to facilitate the analysis of X. This could partly explain why larger models pre-trained with more data have an advantage in performance.

While Transformer models process input by looking *inward* via self-attention, we propose to make the model look *outward* by providing it with related context and knowledge from various sources. We then let the model conduct self-attention on the input while also computing external attention to the knowledge (Figure 1). As the context and knowledge can usually be stored in an non-parametric and symbolic way (e.g., plain text, knowledge graph and dictionary entries), even moderately-sized Transformer models can perform exceptionally well on NLP tasks. This approach allows one to shrink the size of Transformer-based foundation models, which is critical to the accessibility and democratization of AI technology. This approach is also analogous to the way humans conduct intelligence; we often resort to search engines, dictionaries, or information from other people in order to navigate the world.

Another benefit of the external attention is that, as the related knowledge is stored outside of the model, practitioners can easily update the knowledge source to change the behavior of their models. For example, one could add or delete entries from a knowledge graph or rewrite certain paragraphs in Wikipedia. By explicitly representing knowledge, the decision process of the model becomes much more transparent and explainable.

In this paper, we use the commonsense reasoning task CommonsenseQA (Talmor et al., 2019) as a case study in leveraging external attention to obtain and integrate information related to the input. Given a commonsense question and a choice, we retrieve knowledge from three external sources: a knowledge graph (ConceptNet), dictionary (Wiktionary) and labeled training data (CommonsenseQA and 16 related QA datasets). The retrieved knowledge is directly appended to the input and sent to the language model with no revision to the underlying architecture. We show that with the proposed external attention, the accuracy of commonsense reasoning using a DeBERTa-xxlarge model (He et al., 2020) can be significantly boosted from 83.8% to 90.8% on the dev set, while fine-tuned large-scale models like GPT-3 can only achieve 73.0%. The ensembled version of our model, Knowledgeable External Attention for commonsense Reasoning (KEAR), reaches an accuracy of 93.4% on the dev set and 89.4% on the test set, surpassing human performance (88.9%) for the first time (Talmor et al., 2019).

The benefits of our approach extend beyond commonsense reasoning. First, the external attention dramatically reduces our system's dependence on large-scale models, i.e., achieving human parity with models up to 1.5B parameters. Second, the external information is obtained via computationally efficient methods, such as information retrieval and word matching, adding little computational cost to the main model. Third, the text-level concatenation of input and knowledge leads no change to the Transformer model, enabling existing systems to easily adopt this new external attention mechanism.

2 Method

We first describe our external attention framework in Sec 2.1. Next, we describe our external knowledge sources in Sec 2.2. Last, we present additional modeling techniques for improving commonsense reasoning in Sec 2.3. We present empirical results of our techniques in Sec 3.

Problem Formulation. We focus on the multiple-choice question answering task in this paper, where the goal is to select the correct answer from a given list $c_1, c_2, ..., c_n$ for a commonsense question q. The output of the model is a distribution $\mathcal{P}$ on $\{1, 2, ..., n\}$.

2.1 Attention

Self Attention. The majority of recent language models are based on the Transformer architecture (Vaswani et al., 2017). One of the most important components in Transformer is the self-attention mechanism, which can be formulated as

$$\boldsymbol{Q} = \boldsymbol{H}_l \boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{H}_l \boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{H}_l \boldsymbol{W}_v,$$
$$\boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}, \boldsymbol{H}_{l+1} = \text{softmax}(\boldsymbol{A})V, \quad (1)$$

where $\boldsymbol{H}_l \in \mathbb{R}^{N \times d}$ is the input hidden vectors to the l-th Transformer layer, $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ are projection matrices, N is the input length and d is the hidden vector's dimension. The inputs to the first Transformer layer are usually the embeddings of the tokenized input text, denoted as $\boldsymbol{H}_0 = \boldsymbol{X} = [x_1, x_2, ..., x_N]$[1]. In the multi-choice question answering context, the input text is a concatenation of the question and a specific choice.

External Attention. For commonsense question answering, the required information needed to answer the question is usually absent from the input. Thus, we need to integrate external knowledge into the model. In this work, we denote the extra knowledge in text format as $\boldsymbol{K} = [x_1^K, x_2^K, ..., x_{N_k}^K]$. There are many ways to integrate the external knowledge into the model. In this paper we concatenate the knowledge to the input text: $\boldsymbol{H}_0 =$

[1] We do not differentiate between tokens and their embeddings in the following discussion. Following previous work, we prepend a `[CLS]` token to the input.

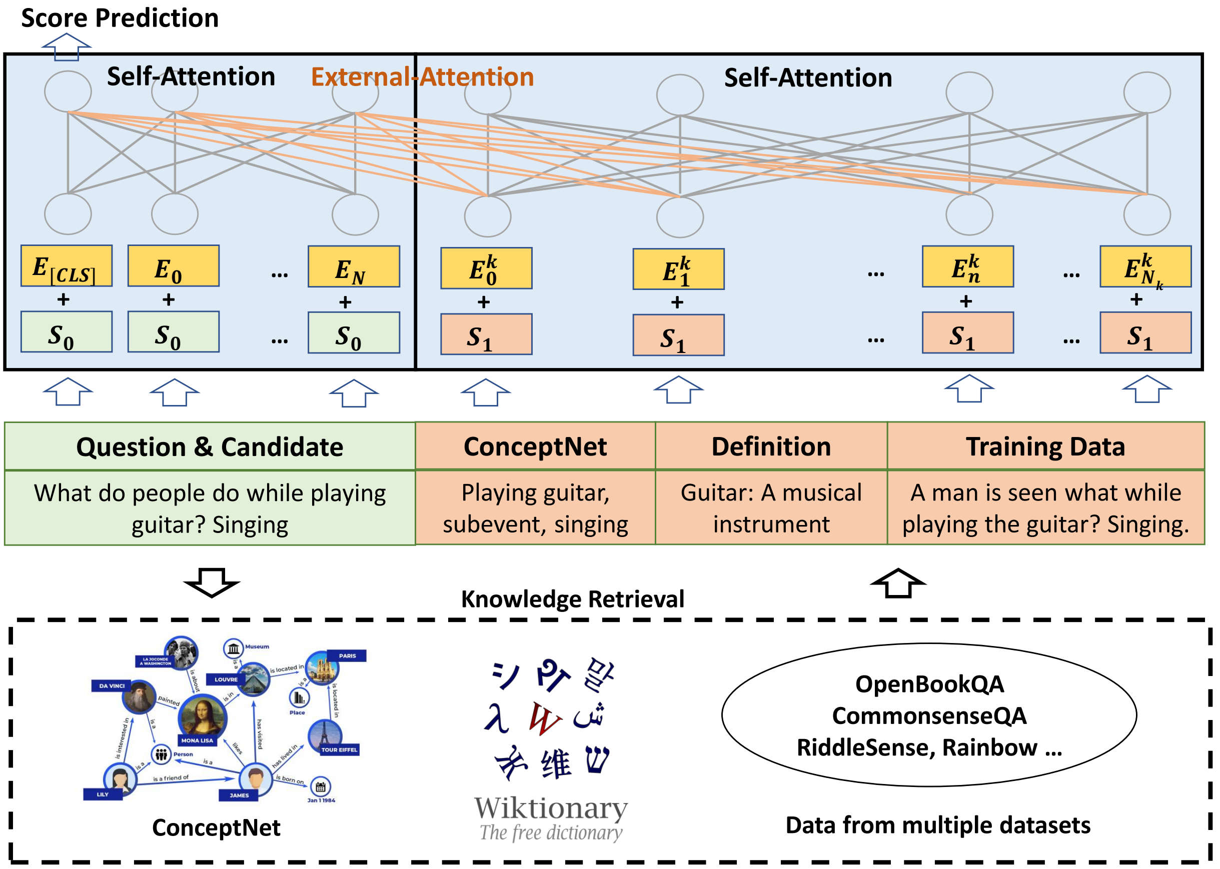

Figure 1: Our proposed method of Knowledgeable External Attention for commonsense Reasoning (KEAR). Related knowledge is retrieved from external sources, e.g., knowledge graph, dictionary and training data, using the input as key and then integrated with the input. While additional external attention layers can be added to the Transformer blocks, we adopt text-level concatenation for external attention, incurring no structural change to the model architecture.

$[\boldsymbol{X};\boldsymbol{K}] = [x_1, ..., x_N, x_1^K, ..., x_{N_k}^K]$. The advantage of this input-level integration is that the existing model architecture does not need to be modified. Then, applying self-attention on $\boldsymbol{H}_0$ can make the model freely reason between the knowledge text and the question/choices, therefore equipping the model with enhanced reasoning capacity.

2.2 Knowledge Retrieval

The knowledge to append to the input for external attention is crucial for getting the correct prediction. For commonsense reasoning, we collect three external knowledge sources to complement the input questions and choices.

Knowledge Graph. Knowledge graphs (KG) contain curated facts that can help with commonsense reasoning. We follow KCR (Lin, 2020) to retrieve a relevant relation triple in the ConceptNet graph (Speer et al., 2017). Suppose the question entity is e_q and the choice contains entity e_c[2]. If there is a direct edge r from e_q to e_c in ConceptNet, we choose this triple (e_q, r, e_c). Otherwise, we retrieve all the triples originating from e_c. We score each triple j by the product of its confidence w_j (provided by ConceptNet) and the defined relation type weight t_{r_j}: $s_j = w_j \cdot t_{r_j} = w_j \cdot \frac{N}{N_{r_j}}$, where r_j is the relation type of j, N is the total number of triples originating from e_c, N_{r_j} is the number of triples with relation r_j among these triples. We then choose the triple with highest weight. Finally, if the selected triple is (e_1, r, e_2), we format the knowledge from the KG as $\boldsymbol{K}_{\text{KG}} = [e_1\ r\ e_2]$.

Dictionary. Although pre-trained language models are exposed to large-scale text data, the long tail distribution of words means that the quality of a word's representation is highly dependent on that word's frequency in the pre-training corpus.

[2] In CommonsenseQA dataset, both e_q and e_c are provided. Otherwise, we can use entity linking to find related knowledge graph nodes to the input text.

Dictionaries, on the other hand, can provide accurate semantic explanation of words regardless of their frequency in datasets. To help understand key concepts in the question and answer, we follow DEKCOR (Xu et al., 2021) to use the Wiktionary definitions of the question and answer concepts as external knowledge. For every concept, we fetch the first (most frequent) definition from Wiktionary using its closest lexical match. Let d_q be the definition text for e_q and d_c be the definition text for e_c, we format the dictionary knowledge as $\boldsymbol{K}_{\text{dict}} = [e_q : d_q; e_c : d_c]$.

Training Data. Although recent language models are giant in terms of the number of parameters, recent studies show that they cannot perfectly memorize all the details of their training data (Anonymous, 2022).

To tackle this challenge, we propose to retrieve relevant questions and answers from the training data as additional knowledge. We use BM25 (Schütze et al., 2008) to retrieve top M relevant questions and answers from the training data. We build the query and index using the concatenation of question, ConceptNet triples and Wiktionary definitions. For each retrieved question from the training data, we drop the knowledge part and employ the retrieved question and its ground-truth answer as external knowledge. During training, for query x, we filter itself from the retrieved results to avoid data leakage. Suppose the retrieved questions and answers are $\{(x_1, c_1), (x_2, c_2), ..., (x_M, c_M)\}$, we format the knowledge from training data as $\boldsymbol{K}_{\text{train}} = [x_1\ c_1; x_2\ c_2; \cdots; x_M\ c_M]$.

Different from Anonymous (2022) where the retrieval questions are only obtained from the same dataset, we experiment with three sources of training data for retrieval: i) CSQA training data, ii) CSQA+OBQA+RiddleSense, a small collection of datasets focusing on ConceptNet knowledge, and iii) a pool of 17 datasets focusing on commonsense reasoning (we describe details of these 17 datasets in the Appendix).

Finally, we concatenate the retrieved knowledge from our three sources to form a final knowledge input: $\boldsymbol{K} = [\boldsymbol{K}_{\text{KG}}; \boldsymbol{K}_{\text{dict}}; \boldsymbol{K}_{\text{train}}]$. In practice, the semicolon is replaced by the separator token (e.g., `[SEP]`). We name our knowledge retrieval and integration technology as Knowledgeable External Attention for commonsense Reasoning (KEAR), shown in Figure 1.

2.3 General Methods to Improve Commonsense Reasoning

Prior works have proposed other methods to improve general NLU performance, and it is therefore natural to wonder if these methods also works for commonsense reasoning. Here, we explore two general methods for improving commonsense reasoning performance: i) using different text encoders and ii) virtual adversarial learning.

Text Encoders. Previous methods for natural language understanding (NLU) (Xu et al., 2021; Yan et al., 2020; Wang et al., 2020; Khashabi et al., 2020) have tried using BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), T5 (Raffel et al., 2019), ELECTRA (Clark et al., 2020) and DeBERTa (He et al., 2020) as the text encoder, achieving state-of-the-art performance on the GLUE benchmark (Wang et al., 2019). Thus, we evaluate these models as encoders for the commonsense reasoning task.

Virtual Adversarial Training (VAT). Previous works show that virtual adversarial training (VAT, Miyato et al. (2018)) can improve the performance for general NLU and question answering tasks (Jiang et al., 2020; Cheng et al., 2021). In the multiple-choice commonsense reasoning task, the goal is to minimize the cross-entropy loss:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\text{CE}(f(x;\theta), y)] \tag{2}$$

where f produces the model prediction (distribution $\mathcal{P}$ on the choices), θ represents the model parameters, y is the one-hot ground-truth answer vector, CE is cross entropy, and D is the empirical data distribution. VAT first finds the update δ that leads to the largest change in the predicted distribution, subject to a L_p-norm constraint. Then, a consistency regularization loss term is added to minimize the difference in the function's output when compared to the input variation δ:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\text{CE}(f(x;\theta), y) + \tag{3}$$

$$\alpha \max_{\|\delta\|_2 \le \varepsilon} \text{CE}(f(x;\theta), f(x+\delta;\theta))], \tag{4}$$

where α and ε are hyperparameters.

3 Experiments

3.1 Setup

Data. We focus on the CommonsenseQA (CSQA, Talmor et al., 2019) benchmark. CommonsenseQA is a widely used multiple-choice question answering dataset that requires commonsense knowledge. It contains 12k questions created using ConceptNet (Speer et al., 2017). For an edge (subject, relation, object) in ConceptNet, Talmor et al. (2019) retrieves other object concepts with the same subject and relation as distractors for a question. A human worker is then asked to i) write a question containing the subject and with the object as the correct answer, ii) pick the most distractive answer from the retrieved concepts, and iii) write another distractor for the question. The final question contains 5 choices, with one correct choice, two random retrieved concepts, one human-picked concept, and one human-curated answer.

Model Setup. We feed the input text into a pretrained text encoder (e.g., DeBERTa) and take the representation $\boldsymbol{v} \in \mathbb{R}^d$ of the `[CLS]` token, where d is the dimension of the encoder. We set the segment id as 0 for the question and answer text, and 1 for the appended knowledge text. The final prediction is computed via softmax($\boldsymbol{v}^T \boldsymbol{b}$), where $\boldsymbol{b} \in \mathbb{R}^d$ is a parameter vector and the softmax is computed over all five choices for a question. We then minimize the cross entropy error during training.

Implementation Details. We finetune the model using the AdamW optimizer. The batch size is set to 48 or smaller to fit the batch on to a single GPU. We train the model for 10 epochs and take the best result on the dev set. We choose best the weight decay in $\{0, 0.01, 0.1\}$. The learning rate are chosen from $\{1e-5, 2e-5, 3e-6\}$ for all encoders except for DeBERTa; following the DeBERTa paper (He et al., 2020) we use a smaller learning rate, chosen from $\{4e-6, 6e-6, 9e-6\}$. We use the DeBERTa v2 model from Huggingface Transformers (Wolf et al., 2020), and choose from the pretrained model or model finetuned on MNLI. For VAT, we choose $\alpha \in \{0.1, 1.0, 10.0\}$ and set $\varepsilon = 1e-5$. For VAT on DeBERTa-xxlarge, we follow SiFT (He et al., 2020) that normalizes the word vectors before adding the perturbation δ, and set $\varepsilon = 1e-4$. For knowledge from training data, we choose the best from the three retrieval source datasets. We run each experiment with 3 different seeds and present results from the best run.

3.2 Effects of Individual Components

Effect of the Encoders. As shown in Table 1, there is a positive correlation between general performance on NLI tasks and commonsense reasoning abilities on CommonsenseQA. Notice that the fine-tuned GPT-3 model with 175 billion parameters could only achieve 73.0% on the dev set of CommonsenseQA. Based on these results, we choose ELECTRA-large and DeBERTa variants (He et al., 2020, 2021) as the encoders for subsequent experimentation.

Encoder	CSQA	MNLI	#Para
Fine-tuned GPT-3	73.0	82.1	175B
RoBERTa-large	76.7	90.2	355M
ALBERT-xxlarge	81.2	90.6	235M
ELECTRA-base	75.0	88.8	110M
ELECTRA-large	81.3	90.9	335M
DeBERTa-xlarge	82.9	91.7	900M
DeBERTa-xxlarge	83.8	91.7	1.5B
DeBERTaV3-large	84.6	91.8	418M
T5-11B	83.5[1]	91.3	11B

Table 1: CSQA dev set accuracy for various encoders. We append the accuracy on MNLI dataset (in-domain) for each encoder as a reference. MNLI scores are from the corresponding GitHub repositories. [1]: from Liu et al. (2021).

Method	Dev Acc(%)
Baselines	
ELECTRA-large	81.3
DeBERTa-xxlarge	83.8
DeBERTaV3-large	84.6
With VAT	
ELECTRA-large + VAT	82.1
DeBERTa-xxlarge + SiFT	84.4
DeBERTaV3-large + VAT	85.2

Table 2: Results on virtual adversarial training.

Effect of Virtual Adversarial Training. Table 2 shows that VAT can improve commonsense reasoning accuracy for of the models under consideration. ELECTRA-large exhibits the largest increase in accuracy (0.8%). Thus, we apply VAT to ELECTRA-large for the following experiments.

Effect of External Attention. As shown in Table 3, all of the proposed knowledge sources bring gains in commonsense reasoning accuracy across all base encoder models. The dictionary, knowledge graph and training data bring 0.5%, 2.1% and 2.5% improvement, respectively, when DeBERTaV3-large (He et al., 2021) is the base encoder model. We find that the best training data

Method	**E-l+V**	**D-xxl**	**DV3-l**
Base	82.1	83.8	84.6
+ KG	85.2	86.4	86.7
+ Dictionary	83.8	84.0	85.1
+ Training data	84.0	86.4	87.1

Table 3: Applying external attention to different knowledge sources. E-l+V stands for ELECTRA-large with VAT, D-xxl stands for DeBERTa-xxlarge, DV3-l stands for DeBERTaV3-large.

Method	**Dev Acc(%)**
ELECTRA-large + KEAR	88.7
DeBERTa-xlarge + KEAR	89.5
DeBERTa-xxlarge + KEAR	90.8
DeBERTaV3-large + KEAR	**91.2**
Ensemble (39 models w/ KEAR)	**93.4**

Table 4: CSQA dev set results with different encoders and ensembles.

retrieval source depends on the exact encoders, and we present a detailed comparison in the Appendix. This demonstrates the effectiveness of our proposed knowledge retrieval and concatenation methods.

Method	Single	Ensemble
BERT+OMCS	62.5	-
RoBERTa	72.1	72.5
RoBERTa+KEDGN	-	74.4
ALBERT	-	76.5
RoBERTa+MHGRN	75.4	76.5
ALBERT + HGN	77.3	80.0
T5	78.1	-
UnifiedQA	79.1	-
ALBERT+KCR	79.5	-
ALBERT + KD	80.3	80.9
ALBERT + SFR	-	81.8
DEKCOR	80.7	83.3
Human	-	88.9
KEAR (ours)	**86.1**	**89.4**

Table 5: Results on test set from the leaderboard. The human performance is ensemble of 5 workers (Talmor et al., 2019).

3.3 Combining the Techniques

Table 4 shows the results of KEAR, which combines the best techniques in previous experiments, i.e., VAT and external attention to all knowledge sources, to further boost the performance. The best single model (DeBERTaV3-large + KEAR) achieves 91.2% accuracy on the dev set. We further ensemble 39 models with 12 ELECTRA models, 12 DeBERTaV3 models, 11 DeBERTa-xxlarge models and 4 DeBERTa-xlarge models. Our ensemble model reaches 93.4% accuracy on the dev set. Table 5 shows the official leaderboard result on the hidden test set. Our ensemble model exceeds the previously best DEKCOR model by over 6% and exceeds the human performance (88.9%) by 0.5%.

4 Related work

Many previous works have proposed ways of incorporating external knowledge sources into Transformer architectures. For commonsense question answering, specialized knowledge graphs like ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a) are the most popular choices for external knowledge source. Lin et al. (2019) constructs a scheme graph from concepts in the question and choices, and uses an LSTM to reason on paths between question and choice concepts. Feng et al. (2020) further proposes the multi-hop graph relation network (MHGRN) for reasoning on paths between concepts. Yasunaga et al. (2021) constructs a joint graph containing the QA context and KG, then use graph neural networks to reason over the two knowledge sources.

Another line of work explores less structured knowledge such as Wikipedia and dictionaries for commonsense reasoning (Xu et al., 2021; Chen et al., 2020; Lv et al., 2020). Bhakthavatsalam et al. (2020) combine the knowledge from ConceptNet, WordNet and other corpora to form 3.5M generic statements and show that this knowledge can help boost accuracy and explanation quality.

Recently, there are approaches to generate facts from pretrained language models to complement missing facts in the external knowledge source. Bosselut et al. (2019) and Hwang et al. (2020) finetunes a pretrained model on ATOMIC for commonsense knowledge graph completion. Liu et al. (2021) directly prompts the GPT-3 model (Brown et al., 2020) to get knowledge for reasoning.

Beyond commonsense reasoning, external knowledge can also help boost performance on other language processing tasks like open domain question answering (Yu et al., 2021), relation classification (Yu et al., 2020a), dialog response generation (Ghazvininejad et al., 2018), conversational QA (Qin et al., 2019), multilingual NLU (Fang et al., 2021) and text generation (Yu et al., 2020b). Compared with prior work that uses extra modules (e.g., GNNs) or extra models (e.g., GPT-3), our external attention framework is extremely lightweight. It operates via a combination of non-parametric retrieval and text concatenation, which we show is highly effective, able to surpass human parity on the CommonsenseQA task.

5 Conclusion

We propose external attention as a lightweight framework for retrieving and integrating external knowledge for language understanding. Compared with self-attention which benefits from ever increasing model sizes, external attention can bring related information from external sources to supplement the input. We demonstrate that this strategy can lead to considerable gains in performance with little additional computational cost. By leveraging knowledge from knowledge graphs, dictionaries and training data, we show that our technology, KEAR, achieves human parity on the CommonsenseQA benchmark task for the first time. For future work, we will apply the technique to other NLP tasks to improve language model performance with external knowledge.

Acknowledgement

We thank the anonymous reviewers for their comments on the paper of our earlier system DECKOR. We thank Reid Pryzant for proof-reading the paper.

References

Anonymous. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. Under review.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69.

Qianglong Chen, Feng Ji, Haiqing Chen, and Yin Zhang. 2020. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2583–2594, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao Cheng, Xiaodong Liu, Lis Pereira, Yaoliang Yu, and Jianfeng Gao. 2021. Posterior differential regularization with f-divergence for improving model robustness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1078–1089, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2019. From'f'to'a'on the ny regents science exams: An overview of the aristo project. *arXiv preprint arXiv:1909.01958*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuwei Fang, Shuohang Wang, Yichong Xu, Ruochen Xu, Siqi Sun, Chenguang Zhu, and Michael Zeng. 2021. Leveraging knowledge in multilingual commonsense reasoning. *arXiv preprint arXiv:2110.08462*.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*.

Liye Fu, Jonathan P Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. *arXiv preprint arXiv:1904.01587*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021): Findings*. To appear.

Jession Lin. 2020. Knowledge chosen by relations. `https://github.com/jessionlin/csqa/blob/master/Model_details.md`.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, William B Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *EMNLP 2019*.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge University Press Cambridge.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI conference on artificial intelligence (AAAI)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Peifeng Wang, Nanyun Peng, Pedro A. Szekely, and X. Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Association for Computational Linguistics (ACL)*.

Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2020. Learning contextualized knowledge structures for commonsense reasoning. *arXiv preprint arXiv:2010.12873*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qagnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2021. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *arXiv preprint arXiv:2110.04330*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020a. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv preprint arXiv:2010.00796*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020b. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Datasets

We use a combination of 17 datasets for our largest-scale training data retrieval. The datasets include αNLI (Bhagavatula et al., 2019), SWAG (Zellers et al., 2018), RACE (Lai et al., 2017) (we only use the middle-school subset), CODAH (Chen et al., 2019), RiddleSense (Lin et al., 2021), SciTail (Khot et al., 2018), Com2Sense (Singh et al., 2021), AI2 Science Questions (Clark et al., 2019), WinoGrade (Sakaguchi et al., 2019), CommonsenseQA (Talmor et al., 2019), CommonsenseQA2.0 (Talmor et al., 2021), ASQ (Fu et al., 2019), OBQA (Mihaylov et al., 2018), PhysicalIQA (Bisk et al., 2020), SocialIQA(Sap et al., 2019b), CosmosQA (Huang et al., 2019) and HellaSWAG (Zellers et al., 2019). We present details of the datasets that we use for training data retrieval in Table 6.

Dataset	**Task**	**#Train**	**#Label**
αNLI	NLI	170k	2
SWAG	MC	73.5k	4
RACE-Middle	MRC	87.9k	4
CODAH	MC	1672	4
RiddleSense	MC	3512	5
SciTail	NLI	23.6k	2
Com2Sense	MC	808	2
AI2Science	MC	1232	4
WinoGrade	CoRef	40.4k	2
CSQA	MC	9741	5
CSQA2.0	CLF	9264	2
ASQ	MC	8872	2
OBQA	MC	4960	4
PhysicalIQA	MC	16.1k	2
SocialIQA	MC	33.4k	3
CosmosQA	MRC	25.3k	4
HellaSWAG	NSP	39.9k	4

Table 6: The datasets used for training data retrieval. NLI stands for natural language inference, MC is multiple choice, MRC is machine reading comprehension, CLF is classification, NSP is next sentence prediction.

B Retrieval Sources

We present results comparing different sources of training data retrieval in Table 7. The best choice of retrieval source varies with encoders and the techniques applied; in general the 17-dataset pool achieve the best performance for DeBERTa, but for ELECTRA retrieving from the CSQA training set alone can get the best performance.

Model	**CSQA**	**3-Data**	**17-Data**
E-l+V	**84.0**	82.9	82.8
D-xxl	86.2	86.1	**86.4**
DV3-l	87.0	**87.1**	**87.1**
E-l+V, best	**88.5**	88.2	87.1
D-xxl, best	89.8	90.5	**90.8**
DV3-l, best	91.0	**91.2**	**91.2**

Table 7: Performance on CSQA dev set of model w.r.t source of training data retrieval. “Best” means our best model combining all the techniques. E-l+V stands for ELECTRA-large with VAT, D-xxl stands for DeBERTa-xxlarge, DV3-l stands for DeBERTaV3-large.