On Commonsense Cues in BERT for Solving Commonsense Tasks

Leyang Cui[△†], Sijie Cheng[‡], Yu Wu[◇], Yue Zhang[†*]
[△] Zhejiang University
[†]School of Engineering, Westlake University
[‡]Fudan University [◇]Microsoft Research Asia
cuileyang@westlake.edu.cn sjcheng20@fudan.edu.cn
Wu.Yu@microsoft.com yue.zhang@wias.org.cn

Abstract

BERT has been used for solving commonsense tasks such as CommonsenseQA. While prior research has found that BERT does contain commonsense information to some extent, there has been work showing that pre-trained models can rely on spurious associations (e.g., data bias) rather than key cues in solving sentiment classification and other problems. We quantitatively investigate the presence of structural commonsense cues in BERT when solving commonsense tasks, and the importance of such cues for the model prediction. Using two different measures, we find that BERT does use relevant knowledge for solving the task, and the presence of commonsense knowledge is positively correlated to the model accuracy.

1 Introduction

Pre-trained language models (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019b) give competitive results on a variety of NLP tasks (Zhou and Zhao, 2019; Joshi et al., 2019; Liu and Lapata, 2019; Cui et al., 2020). It has been shown that they can effectively capture syntactic features (Goldberg, 2019), semantic information (Liu et al., 2019a) and factual knowledge (Petroni et al., 2019), which provides support for the success in downstream tasks.

Recently, there has been some debate about whether commonsense knowledge can be learned by a language model trained on large corpora. While Davison et al. (2019), Bosselut et al. (2019) and Rajani et al. (2019) argue that pre-trained language models can directly identify commonsense facts, Lin et al. (2019) and Klein and Nabi (2019) believe that structured commonsense knowledge is not captured well.

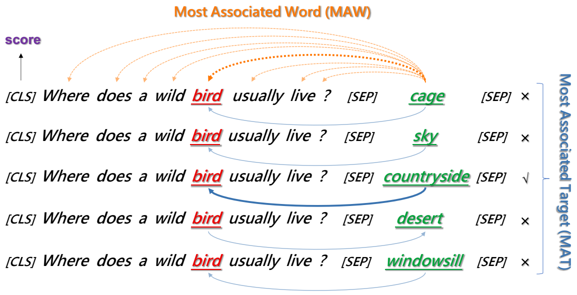
Most Associated Word (MAW)
score
[CLS] Where does a wild bird usually live ? [SEP] cage [SEP] ×
[CLS] Where does a wild bird usually live ? [SEP] sky [SEP] ×
[CLS] Where does a wild bird usually live ? [SEP] countryside [SEP] √
[CLS] Where does a wild bird usually live ? [SEP] desert [SEP] ×
[CLS] Where does a wild bird usually live ? [SEP] windowsill [SEP] ×
Most Associated Target (MAT)

Figure 1: Two methods used to study structured commonsense knowledge in pre-trained Transformer. Commonsense link is drawn from the Target Concept (Answer Concept) to the Source Concept (Question Concept).

Pre-trained language models have achieved empirical success when fine-tuned on specific commonsense tasks such as COSMOS QA (Huang et al., 2019), SWAG (Zellers et al., 2018), and CommonsenseQA (Talmor et al., 2019). One possible reason of the high performance is that there exist *superficial cues* or *spurious associations* in the dataset, which enables models to answer questions without understanding the task (Niven and Kao, 2019; Yu et al., 2020; Kaushik et al., 2020). For example, a model can choose the spurious cue word "meadow" as a feature for positive reviews simply because "meadow" occurs frequently in positive documents. It remains an interesting research question whether commonsense knowledge plays a central role among statistical cues that BERT has when solving commonsense tasks. In other words, we are interested in investigating whether BERT solves commonsense tasks using commonsense knowledge.

We try to provide quantitative answers by mainly using the CommonsenseQA dataset, which asks a model to solve a multiple-choice problem. As shown in Figure 1, given a question and five candidate answers, a model should select one candidate answer as the output. The current state-of-the-art

[*]Corresponding Author

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 683–693
August 1–6, 2021. ©2021 Association for Computational Linguistics

pre-trained language models solve the problem by representing the question jointly with each candidate answer (we call such a question-answer pair a *sentence* thereafter), and using pre-trained language models as the main encoder. Scoring of each sentence is based on a sentence-level hidden vector, and the candidate answer that corresponds to the highest-scored sentence is taken as the output.

We investigate the presence of commonsense cues in the BERT representation of a sentence by examining **commonsense links** from the **answer concepts** to its related contextualized **question concepts**. Figure 2 shows one example, where the question concept is "bird", and the correct answer is the answer concept connected through an ATLOCATION link in the CONCEPTNET knowledge graph. Such related concepts are *not* explicitly used in a BERT model for making prediction, and therefore its strength in the BERT representation is not necessarily optimized in task fine-tuning. We call such cues *structured commonsense*, which is a source of knowledge that we can explicitly measure. We take two methods for measuring structured commonsense in BERT, including directly measuring the attention weights (Clark et al., 2019) and measuring attribution scores by considering gradients (Mudrakarta et al., 2018).

We conduct two sets of experiments to quantitatively measure commonsense links in different situation. In the first set, we examine the presence of commonsense links directly in the BERT representation both before and after fine-tuning (Section 5). In the second set of experiments, we investigate the correlation between commonsense links with model predictions (Section 6). While the former can serve as a probing task for understanding commonsense learned by pre-training, the latter can serve as a means for understanding whether a model learns to make better use of commonsense knowledge through supervised fine-tuning.

Results suggest that BERT does have commonsense knowledge from pre-training, just as syntactic and word sense information. In addition, through fine-tuning, BERT relies more on commonsense cues in making prediction. The evidence is quantitatively demonstrated by stronger commonsense links in the representation, and a salient correlation between model predictions and commonsense link strengths, despite the fact that neither the answer concept nor the related question concept in a commonsense link is directly connected to the output layer. Interestingly, results also indicate that the stronger the structured commonsense knowledge is, the more accurate the model is. In addition to CommonsenseQA dataset, we observe similar phenomenon on Wikipedia and OMCS, demonstrating the generalization of our findings. To our knowledge, we are the first to investigate key statistical cues when BERT solves the CommonsenseQA task, providing several evidences that commonsense knowledge is indeed made use of. We release our code at `https://github.com/Nealcly/commonsense`.

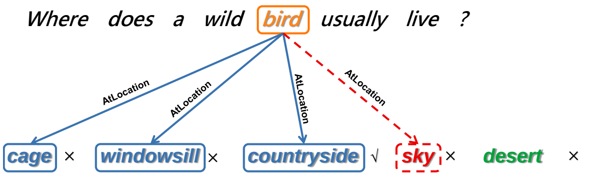

Figure 2: From CONCEPTNET to CommonsenseQA.

2 Related Work

There has been much recent work exploiting the underlying knowledge embedded in BERT representations. Peters et al. (2018) find that lower layers and higher layers in ELMo contain more syntactic and semantic information, respectively. Tenney et al. (2019), Liu et al. (2019a) and Jawahar et al. (2019) use probing models on hidden states to analyze linguistic information within pre-trained language models. Goldberg (2019) assess BERT's syntactic abilities by masking the verb, and comparing the prediction probability of the original verb with incorrect verbs. Our method is similar to Clark et al. (2019) and Htut et al. (2019), who focus on attention heads. The difference lies in that our primary goal is to investigate what information is learned and made use of when solving commonsense tasks. Therefore, our investigation is *task-centered*.

There has also been work investigating data bias and spurious associations. Gururangan et al. (2018) and Poliak et al. (2018) show that classifiers achieve accuracies around 69% on SNLI (Bowman et al., 2015) by using partial input. Kaushik et al. (2020) demonstrate BERT solve sentiment analysis and NLI by heavily relying on spurious associations. Our work is in line in investigating statistical cues. Different from the above investigations, we use probing methods to verify the presence and importance of the key feature, namely commonsense knowledge, in solving commonsense QA, rather than focusing on adversarial cases.

Commonsense reasoning is a challenging task in natural language processing. Traditional methods rely heavily on hand-crafted features (Rahman and Ng, 2012; Bailey et al., 2015) and external knowledge bases (Schüller, 2014). With recent advances in deep learning, pre-trained language models have been used as a powerful method for such tasks. Trinh and Le (2018) use a pre-trained language model to score candidate sentences for the Pronoun Disambiguation and Winograd Schema Challenge (Levesque et al., 2012). Klein and Nabi (2020) use a sentence-level loss to enhance commonsense knowledge in BERT. Mao et al. (2019) demonstrate that pre-trained language models fine-tuned on SWAG (Zellers et al., 2018) are able to provide commonsense grounding for story generation. For commonsense question answering, pre-trained language models with fine-tuning give the state-of-the-art performance (Zellers et al., 2018; Huang et al., 2019; Talmor et al., 2019). Though the above work show usefulness of BERT on comonsense tasks, little work has been done investigating the mechansim for BERT solving the tasks. Our work thus complements existing research in this aspect.

There is also a line of work leveraging ConceptNet to enhance model's commonsense reasoning ability. Lin et al. (2019) inject path information from question concepts to answer concepts to a model. Ye et al. (2019) use ConceptNet to construct pre-training dataset for BERT. Lv et al. (2019) extract evidence from ConceptNet and Wikipedia to build a relational graph for CommonsenseQA. We use ConceptNet for measuring commonsense knowledge in BERT.

3 Task and Model

We review the main experimental dataset CommonsenseQA (Section 3.1), before showing the structure of a state-of-the-art model (Section 3.2).

3.1 Dataset

CommonsenseQA (Talmor et al., 2019) is a multiple-choice question answering dataset constructed based on the knowledge graph ConceptNet (Speer et al., 2017), which is composed of a large set of triples taking the form ⟨source concept, relation, target concept⟩, such as ⟨Bird, AtLocation, Countrysides⟩. Given a source concept Bird and the relation type AtLocation, there are three related target concepts Cage, Windowsill and Countryside in ConceptNet.

As shown in Figure 2, in the development of the CommonsenseQA dataset, crowd-workers are requested to generate question and candidate answers based on the *source concept* and three related target concepts in ConceptNet, respectively. Following Talmor et al. (2019), we call the source concept in the question as *question concept*, and the target concept in the answer as *answer concept*. Each question corresponds to only one correct *answer concept* among the three related ConceptNet target concepts. In addition, two more incorrect answer concepts are added, which do not correlate with the question concept in ConceptNet, resulting in 5 candidate answers for each question. We define *commonsene link* as the link from the answer concept to the question concept.

The CommonsenseQA dataset is designed to avoid salient bias in surface patterns. First, the lexical overlap between the correct answer and the question is similar to that between the question and incorrect candidates. Second, commonsense links are not superficial patterns that can be learned from training data. In particular, the percentage of answer-question-concept pairs in test examples that also exist in the gold training examples is 15.78%, which suggests that the main source of strong commonsense links, if observed, comes mainly from the pre-trained BERT model itself.

In order to analyze implicit structured commonsense knowledge, which is based on the link from the answer concept to the question concept, we filter out questions which do not contain explicit mentions to the *question concept* in its ConceptNet form (e.g. paraphrase). The resulting dataset CommonsenseQA* contains 74 fewer instances.

3.2 Model

We adopt the method of Talmor et al. (2019), using BERT (Devlin et al., 2019). In particular, given a question q and 5 candidate answers $a_1, ..., a_5$, we concatenate the question with each answer to obtain 5 question-answer pair sequences (i.e. sentences) $s_1, \ldots, s_5$, respectively. In each sentence, we use a special symbol `[CLS]` in the beginning, a symbol `[SEP]` between the question and the candidate answer, a symbol `[SEP]` in the end.

BERT uses L stacked Transformer layers (Vaswani et al., 2017) to encode each sentence. The last layer hidden state of the `[CLS]` token is used for linear scoring with softmax normalization. The candidate among $s_1, \ldots, s_5$ with the highest score is chosen as the prediction. More details of our implementation are shown in Appendix C.

4 Analysis Methods

As mentioned earlier, we analyze commonsense links using the attention weight (Clark et al., 2019) and the corresponding attribution score (Sundararajan et al., 2017; Mudrakarta et al., 2018). We report results in one random execution for each experiment. We additionally tried five runs for each experiments, and found that the result variation is small (Appendix B).

4.1 Attention Weights

Given a sentence, attention weights in Transformer can be viewed as the relative importance weight between each token and the other tokens when producing the next layer representation (Kovaleva et al., 2019; Vashishth et al., 2020). In particular, given a sequence of input vectors $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{|H|}]$, its self-attention representation uses each vector as a query to retrieve all context vectors in $\mathbf{H}$, yielding a matrix of attention weights $\alpha \in \mathbb{R}^{|H|\times|H|}$.

The value of α is computed using the scaled dot-product of the query vector of representation $\mathbf{Q} = \mathbf{W}^Q\mathbf{H}$ and the key vector of representation $\mathbf{K} = \mathbf{W}^K\mathbf{H}$, followed by softmax normalization

$$\alpha = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}), \qquad (1)$$

where d_k is the dimension size of the key vector $\mathbf{K}$. $\alpha_{i,j}$ represents the attention strength from $\mathbf{h}_i$ to $\mathbf{h}_j$. For multi-head attention, $\mathbf{H}$ is linearly projected T times to find T sets of queries, keys, and values, where T is the number of heads. The attention operation of each head is performed in parallel, with the results being concatenated. We use $\alpha^{m,n}$ to denote the n-th attention head in the m-th layer. The attention weights $\alpha^{m,n}$ are used as a first measure of commonsense link strengths.

4.2 Attribution Scores

Kobayashi et al. (2020) point out that analyzing only attention weights can be insufficient for investigating the behavior of attention heads. As a supplement, gradient-based feature attribution methods have been studied to interpret the contribution of each input feature to the model prediction in back-propagation (Baehrens et al., 2010; Mudrakarta et al., 2018; Hao et al., 2020). Analysis of both attention weights and the corresponding attribution scores allows us to more comprehensively understand commonsense links in BERT.

We employ an attribution technique called *Integrated Gradients* (Sundararajan et al., 2017). Intuitively, integrated gradients works by simulating the process of pruning the specific attention head (from the original attention weight α to a zero vector α'), and calculating the integrated gradients in back-propagation. The attribution score directly reflects how much a change of attention weights affects model outputs. A higher attribution score represents more importance of the corresponding individual attention weight. Suppose that $F(x)$ represents the BERT model output for CommonsenseQA given an input x. The attribution of attention head $\alpha_t t \in [1, \ldots, T]$ in a Transformer layer can be computed by comparing with a set of baseline weights α':

$$Atr(\alpha^t) = (\alpha^t - \alpha'^t) \otimes \int_{x=0}^{1} \frac{\partial F(\alpha' + x(\alpha - \alpha'))}{\partial \alpha^t} dx \qquad (2)$$

where $\otimes$ denotes element-wise multiplication, $\alpha = [\alpha^1, \ldots, \alpha^T]$. Intuitively, $F(\alpha' + x(\alpha - \alpha'))$ is closer to $F(\alpha')$ when x is closer to 0, and closer to α when x is closer to 1. Therefore, $\int_{x=0}^{1} \frac{\partial F(\alpha' + x(\alpha - \alpha'))}{\partial \alpha^t} dx$ gives the amortized gradient with all different x. $Atr(\alpha^t) \in \mathbb{R}^{n\times n}$ denotes the attribution score which corresponding to the attention weight α^t. $Atr(\alpha^t_{i,j})$ is represented for the interaction from token $\mathbf{h}_i$ to $\mathbf{h}_j$. We set the uninformative baseline α' as zero vector. Following Sundararajan et al. (2017), we approximate $Atr(\alpha^t)$ via a gradient summation function,

$$Atr(\alpha^t) ::= (\alpha^t - \alpha'^t) \otimes \sum_{i=1}^{s} \frac{\partial F(\alpha' + i/s(\alpha - \alpha'))}{\partial \alpha'^t} \times \frac{1}{s}, \qquad (3)$$

where s is the number of approximation steps for computing integrated gradients. We set s to 20 based on the empirical results.

5 The Presence of Knowledge

We first conduct a set of experiments to investigate commonsense link strengths in BERT representations of question-answer pairs (i.e. sentences). Intuitively, if the link weight from the answer concept to the question concept is higher than those from the answer concept to other question words, then

	Statistics		MAW Accuracy				Random
			Max		Avg		
Dataset	# Instances	# Avg Length	BERT	BERT-FT	BERT	BERT-FT	
CommonsenseQA*	1,147	13.18	46.82	49.22	12.38	17.35	10.53
OMCS	37,895	7.63	88.48	89.14	37.82	39.52	24.11
Wikipedia	176,449	16.40	40.24	43.53	13.19	13.48	6.22

Table 1: Average and maximum MAW accuracies across three datasets. BERT-FT model denotes the BERT model with fine-tuning on CommonsenseQA training set.

Relation Type	Max	Avg	L-H	# Ins
Random	10.53	10.53	-	-
OverAll	49.22	17.35	8-7	-
AtLocation	55.85	18.42	8-7	574
Causes	55.93	18.91	8-7	162
CapableOf	47.88	14.71	8-1	104
Antonym	52.53	10.97	4-3	83
HasPrerequisite	54.15	18.93	9-8	41
HasSubevent	55.29	18.74	9-0	34
Desires	40.00	7.92	8-1	27
CausesDesire	48.89	14.28	4-0	27
PartOf	59.09	18.56	9-0	22
HasProperty	54.00	15.12	9-1	20
MotivatedByGoal	75.56	24.31	9-7	18
HasA	68.89	22.10	8-1	9
RelatedTo	62.22	18.44	9-0	9

Table 2: The average and maximum MAW accuracies of BERT-FT for different commonsense relations. We exclude the relation types with frequencies of occurrence less than 9. L-H represents the best performing attention head for each relation.

we have evidence of BERT using commonsense cues according to ConceptNet. As mentioned earlier, rather than the question concept, the representation of the [CLS] token is directly connected to the output layer for candidate scoring. Hence there is no direct supervision signal from the output layer to the link weight during fine-tuning, and better prediction does not necessarily indicate strong commonsense links.

5.1 Probing Task

Without losing generality, we call both attention weights in Section 4.1 and attribution weights in Section 4.2 *link weight*. We evaluate link weights by calculating the **most associated word** (MAW), namely the question concept word that receives the maximum link weight from the answer concept among all question words. MAW is measured for each individual attention head in each layer.

Denote the hidden states of the whole question, question concept and answer concept as $[\mathbf{h}_1, \dots, \mathbf{h}_{|q|}]$, $[\mathbf{h}_{b_s}, \dots, \mathbf{h}_{e_s}]$ and $[\mathbf{h}_{b_t}, \dots, \mathbf{h}_{e_t}]$, respectively. If the answer concept is composed of multiple tokens, we consider the link weight from the answer concept to the question token $\mathbf{h}_i$ $(i \in [1, |q|])$ as the mean of the link weights over all answer tokens $\alpha_i = \frac{1}{e_t - b_t}\sum_{j=b_t}^{e_t} \alpha_{j,i}$. For the n-th attention head in the m-th layer, if the question concept receives the maximum link weight from the answer concept (i.e., $\mu^{m,n} = \arg\max_i \alpha_i^{m,n}$, $\mu^{m,n} \in [b_s, e_s]$), we consider that this attention head gives the correct MAW.

We take two different measures of MAW accuracies, calculating the average accuracy among all attention heads, and the accuracy of the most-accurate head, respectively. Previous work probing syntactic information from attention head takes the second method (Clark et al., 2019; Htut et al., 2019). We additionally measure the average in order to comprehensively evaluate the prevalence of commonsense cues in BERT.

The average MAW accuracy is measured by:

$$acc^{avg} = \frac{\sum_{m=1}^{12}\sum_{n=1}^{12}\sum_{d=1}^{D}\mathbb{1}(\mu^{m,n} \in [b_s, e_s])}{12 \times 12 \times D}.$$

The maximum MAW accuracy is measured by:

$$acc^{max} = \max_{m=1}^{12}\max_{n=1}^{12}\frac{\sum_{d=1}^{D}\mathbb{1}(\mu^{m,n} \in [b_s, e_s])}{D},$$

where D represents the number of instances for evaluation.

In theory, if link weights for each attention head are randomly distributed, the average and maximum MAW accuracies should be both

$$acc^{baseline} = \frac{\sum_{d=1}^{D}\frac{e_s - b_s}{|q|}}{D},$$

which reflects the fact that the representation does not contain explicit correlation between the answer concept and its related question concept. In contrast, MAW accuracies significantly better than this baseline indicates that commonsense knowledge is contained in the representation.

5.2 Results

The results for off-the-shelf BERT (BERT) and a BERT model fine-tuned on CommonsenseQA

(BERT-FT) are shown in the first row of Table 1. First, looking at the original non-fine-tuned BERT, the *maximum* MAW accuracy of each layer significantly outperforms[1] the random baseline. This shows that commonsense links are a part of BERT representation of a sentence in general, just as syntactic (Goldberg, 2019) and semantic (Liu et al., 2019a) knowledge. Second, BERT-FT outperforms BERT in terms of both the average MAW accuracy and the maximum MAW accuracy, with a relatively large boost on the average MAW accuracy, which shows that structured commonsense features are enhanced by supervised training on commonsense tasks.

We explore the best performing attentions head for each relation type in Table 2, finding that certain attention heads capture specific commonsense relations. There is no single attention head that does well for all relation types, both with fine-tuning and without fine-tuning, which is similar to the previously finding for syntactic heads (Raganato and Tiedemann, 2018; Clark et al., 2019).

To further differentiate commonsense cues from superficial association, we measure the co-occurrence between each word in the question and answer concept in 1 million English Wikipedia documents. There is only 1.85% question concept word among the highest co-occurring words of each answer concepts, which partly shows that the strong commonsense links do not heavily rely on superficial pattern.

5.3 Additional Datasets.

Since this set of experiments concerns the representation only, we take additional two unlabeled corpora in addition to CommonsenseQA. In particular, we extract sentences from Open Mind Common Sense (OMCS) [2] and Wikipedia, if there existing one and only one source-target concept pair in this sentence, yielding two large-scale datasets. The detailed statistics are shown in Table 1. The results are consistent with the CommonsenseQA dataset, which shows the generation ability of our methods.

6 Co-relating Knowledge with Task

We further conduct a set of experiments to draw the correlation between commonsense links and model prediction. The goal is to investigate how BERT makes use of commonsense knowledge for making a prediction in the CommonsenseQA task. In particular, we compare the link weights across the five answer candidates for the same question, and find out the candidate that is the most associated with the relevant question concept. This candidate is called the **most associated target** (MAT). Correlations are drawn between MATs and the model prediction for each question. Intuitively, the more the MATs are correlated with the model predictions, the more evidence we have that the model makes use of commonsense cues in making prediction.

	Attention				Attribution	
	BERT-FT		**BERT-Probing**		**BERT-FT**	
Head	MAT	MAS	MAT	MAS	MAT	MAS
1	49.00	18.92	29.21	4.01	51.61	23.54
2	49.17	19.62	20.75	10.99	27.46	24.85
3	32.00	56.23	16.04	43.85	49.17	33.83
4	41.33	16.74	32.17	9.68	22.93	47.08
5	49.96	24.32	33.91	6.28	31.04	44.29
6	45.42	13.25	34.87	4.62	34.26	20.14
7	48.39	13.33	25.72	7.41	33.83	22.67
8	54.14	13.39	28.07	3.66	25.98	49.61
9	39.67	16.74	28.86	9.50	36.97	22.84
10	38.71	13.95	24.50	18.66	52.14	21.01
11	49.17	8.89	36.88	7.15	36.79	21.19
12	53.53	11.07	30.08	3.31	25.81	26.94
Avg	45.87	18.85	28.42	10.76	35.67	29.83

Table 3: MAT^{overlap} and MAS^{overlap} in the top layer.

Both attention weights and the corresponding attribution scores are used, because now we are considering model prediction, for which gradients play a role and can be measured. For all experiments, the trend of attribute scores is consistent with that measured using attention weights.

6.1 Probing Tasks

Formally, given a question q and 5 candidate answers $a_1, \ldots, a_5$, we make comparisons across five candidate sentences $s_1, \ldots, s_5$. In each candidate sentence, we calculate the link weight from the answer concept to the question concept according to CONCEPTNET. Denote the hidden states of the question concept and the answer concept as $[\mathbf{h}_{b_s}, \ldots, \mathbf{h}_{e_s}]$ and $[\mathbf{h}_{b_t}, \ldots, \mathbf{h}_{e_t}]$, respectively. The link weight of the answer-question-concept pair (α_{a2q}) is the average between each answer concept token and each question concept token

$$\alpha_{a2q} = \frac{\sum_{i=b_s}^{e_s} \sum_{j=b_t}^{e_t} \alpha_{j,i}}{(e_s - b_s)(e_t - b_t)}$$

Among the five candidates in each instance, we take the one with the highest α_{a2q} as the most associated target MAT, denoted as $s^{\text{MAT}} \in [1, 5]$.

[1] $p \leq 0.01$ using t-test; similar for subsequent mentions.

[2] Open Mind Common Sense (OMCS) corpus is the source corpus of ConceptNet.

(a) Measured by attention weights.

#H	#Ins	Model Acc.	#H	#Ins	Model Acc.
0	158	20.89	7	69	78.26
1	135	28.15	8	63	82.54
2	119	52.10	9	57	92.98
3	132	53.79	10	47	89.36
4	93	62.37	11	44	97.73
5	106	66.04	12	36	100.00
6	88	68.18	-	-	-

(b) Measured by attribution.

#H	#Ins	Model Acc.	#H	#Ins	Model Acc.
0	89	10.11	5	171	72.51
1	114	22.81	6	119	81.51
2	148	51.35	7	85	82.35
3	156	56.41	8	43	74.72
4	207	66.67	9	13	84.62

Table 4: The relationship between the MAT head count and the model prediction accuracy. #H denotes how many heads yield the correct MAT prediction. #Ins denotes the instance number.

As a baseline for MAT, we further define **most associated sentence** (MAS) as the candidate answer that has the maximum link weight from the answer concept to the [CLS] token among the five candidates. The reason is that gradients are backpropagated from the [CLS] token rather than the question concept or the answer concept. By comparing MAT and MAS, we can have useful information on whether MAT is an influencing factor for the model decision.

We measure the correlation between MAT ($s^{\text{MAT}} \in [1,5]$), the model prediction ($s^{model} \in [1,5]$) and the gold-standard answer ($s^{golden} \in [1,5]$) by using two metrics, including the overlapping rate between MATs and model predictions, and the accuracy of MATs.

The **overlapping rate of MATs** is defined as:

$$\text{MAT}^{overlap} = \frac{\sum_{d=1}^{D} \mathbb{1}(s_d^{\text{MAT}} = s_d^{model})}{D}$$

The **accuracy of MATs** is defined as the percentage of Mats that equals the gold answer:

$$\text{MAT}^{acc} = \frac{\sum_{d=1}^{D} \mathbb{1}(s_d^{\text{MAT}} = s_d^{golden})}{D}$$

Similar to MAW, MAT and MAS can be measured for each attention head, and we calculate the average and maximum values across different heads.

6.2 Commonsense Link and Model Output

We measure the MAT performance of BERT-FT, and a BERT model that is fine-tuned for the output layer only (BERT-probing). The latter is a linear probing model (Liu et al., 2019a). Intuitively, if the probing model can solve the commonsense task accurately, then the original non-fine-tuned BERT likely encodes the rich commonsense knowledge.

Table 3 shows the relative strengths of MATs and MASs according to the 12 attention heads in the top Transformer layer. First, for both models, the overlapping rates of MATs are significantly ($p \leq 0.01$) larger than that with MASs. This suggests that the link weight from the answer concept to the question concept is more closely-related to the model prediction as compared to the link weight from the answer concept to the [CLS] token, despite that model output scores are calculated on the [CLS] token. The results give strong evidence that commonsense cues from BERT are relied on for model decision. Second, when fine-tuned with training data, the model gives an even stronger correlation between MAT and the model prediction. This suggests that the model can *learn* to make use of commonsense cues for making prediction, which partly shows how a BERT model solves CommonsenseQA.

Figure 3 shows the overlapping rate between MAT and model prediction at each Transformer layer. Both the maximum and the average overlapping rates across the 12 layers are shown. The random overlapping rate of 20% is drawn as a reference. It can be seen from the figure that the maximum overlapping rate of BERT-probing is significantly larger than the random baseline, which shows that the model prediction is associated with the relevant structured commonsense cues. In addition, after fine-tuning, the BERT-FT model shows a tendency of weakened maximum MAT overlapping rate on lower Transformer layers and much strengthened MAT overlapping rate on higher layers, and in particular the top layer. The trend of MAT measured by attribution score is consistent with attention weights. This suggests that fine-tuned model relies more on the commonsense structure in the top layer for making prediction.

We compare the co-occurrence between question concepts and candidate answer concepts in 1 million English Wikipedia documents, and find that only 18.2% gold answers has the most co-occurrence with the question concept among 5 answer candidates, which is even lower than the random baseline (20%), showing that CommonsenseQA cannot be solved by solely relying on superficial patterns.

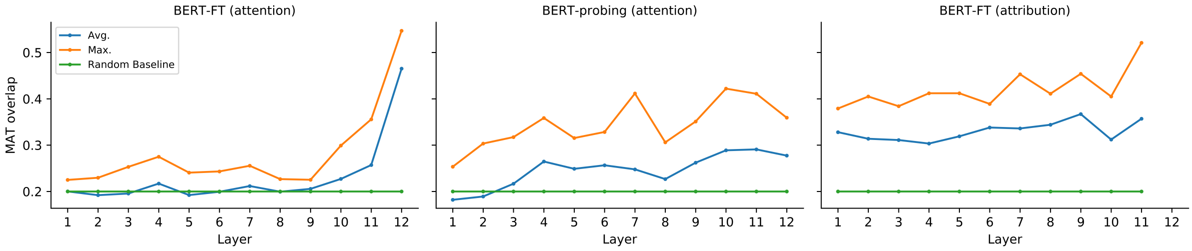

Figure 3: MAT^{overlap} across different layers.

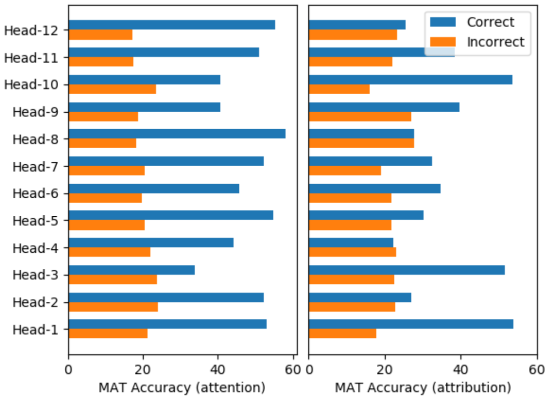

Figure 4: MAT^{acc} of each attention head in the top layer with correct and incorrect model predictions. "Red" and "Blue" indicate the model performance if attention head-n gives correct and incorrect prediction, respectively.

6.3 Commonsense Link and Model Accuracy

Table 4 shows the correlation between MAT accuracies and model prediction accuracies. Each row shows a different number of heads in the top layer for which the MAT corresponds to the correct answer candidate, together with the number of test instances for such cases, and the model prediction accuracy on the instances. There is an obvious trend where increased MAT accuracies correspond to increased model prediction accuracies, which shows that making use of structured commonsense cues leads to better model prediction.

Figure 4 shows the MAT accuracies of each attention head in the top layer for the test instances with correct and incorrect model predictions, respectively. The MAT accuracies of correctly predicted instances are larger than those of incorrectly predicted instances by a large margin. The finding is consistent with Table 4, which shows that structured commonsense cues are a key factor in BERT making the correct decision.

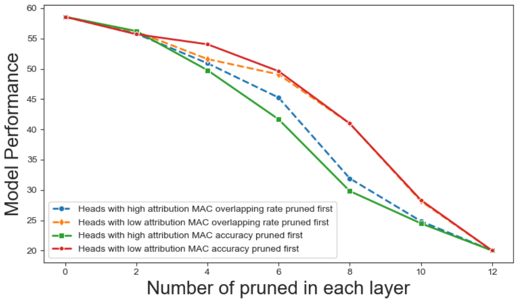

Figure 5: Model performance on the CommonsenseQA development set when different heads are pruned.

	BERT-FT			BERT-probing		
	MAT^{overlap}		Model	MAT^{overlap}		Model
L	Max	Avg	Acc	Max	Avg	Acc
12	54.14	45.87	58.59	36.88	28.42	39.23
11	46.56	26.65	56.50	37.66	27.11	35.48
10	37.40	27.86	53.36	39.84	28.50	33.74
9	34.61	24.01	51.53	30.08	24.76	32.52
8	31.82	21.39	49.35	25.81	21.53	33.57
7	31.73	24.40	48.74	37.05	24.04	32.96
6	31.56	23.64	45.95	31.21	24.02	32.00
5	34.44	25.01	44.99	33.39	24.03	32.43
4	44.73	34.13	40.28	41.06	27.67	33.83
3	44.20	32.48	37.58	25.81	21.02	21.88
2	23.71	19.47	26.68	23.63	20.74	20.40
1	23.45	19.50	23.02	20.58	18.81	19.27

Table 5: Performance of MAT^{overlap} across different layers. L-n represents adding the output classifier on the hidden state of layer-n. Our BERT-FT model (layer-11) gives 58.15% accuraies, which is slightly higher than the reported results of 55.57% on Lin et al. (2019). It achieves 58.59% on our dataset CommonsenseQA*.

We further evaluate the model performance after pruning specific heads. We sort all the attention heads in each layer according to their MAT performance by attribution scores, and then prune these heads in order. Following Michel et al. (2019), we replace the pruned head with zero vectors. Figure 5 shows the model performance on the development set. As the number of pruned heads increases, the model performance decreases, which conforms to intuition. In addition, the model performance drops much more rapidly when the attention heads with higher MAT performances are pruned first, which demonstrates that capturing commonsense features is crucial to strong model prediction.

6.4 Commonsense Link and BERT Layer

We further investigate two specific questions on the commonsense knowledge usage. First, which layer does BERT rely on the most for making its decision. Second, does the commonsense knowledge that BERT uses come more from pre-training or fine-tuning. We compare 12 model variations by connecting the output layer on each of the Transformer layer, respectively. Table 5 shows the model accuracies and the MAT overlapping rates. First, BERT-probing gives the best performance when prediction is made on the top layer, and the accuracy generally decreases as the layer moves to the bottom. This indicates that relevant commonsense knowledge is more heavily distributed towards higher layers during pre-training. Our experimental settings here are the same as the probing task for syntactic information by Liu et al. (2019a), who find that syntactic information is distributed more heavily towards lower BERT layers.

With fine-tuning, we observe stronger improvements of both model accuracies and MAT overlaps on higher layers when comparing BERT-FT and BERT-probing. This demonstrates that commonsense knowledge on higher layers is more useful to the CommonsenseQA task. Interestingly, comparing layer 11 and layer 10, the model accuracy after fine-tuning is similar, but the MAT overlap of layer 11 is significantly larger. This can suggest that the structured commonsense knowledge that we probe attributes only partly to the overall useful knowledge for CommonsenseQA.

7 Conclusion

We conducted quantitative analysis to investigate how BERT solves the CommonsenseQA task, aiming to gain evidence on the key source of information involved in the disambiguation process. Empirical results demonstrated that BERT encodes structured commonsense knowledge, and is able to leverage such cues on the downstream CommonsenseQA task. Our analysis has further revealed that with fine-tuning, BERT learns to make better use of commonsense features on higher layers. These suggest that BERT can learn to make use of truly relevant commonsense cues rather than superficial patterns for CommonsenseQA.

References

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831.

Daniel Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. The winograd schema challenge and reasoning about correlation. In *AAAI Spring Symposia*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Self-attention attribution: Interpreting information interactions inside transformer.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies?

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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *EMNLP-IJCNLP-19*, Hong Kong, China. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention module is not only a weight: Analyzing transformers with vector norms. *ArXiv*, abs/2004.10102.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, pages 552–561.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *NAACL-2019*, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP-2019*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. 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. 2019. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. Improving neural story generation by targeted common sense grounding. In *Proceedings of EMNLP-IJCNLP-2019*, pages 5988–5993, Hong Kong, China. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *acl-19*, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-2018*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP-2019*, Hong Kong, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Peter Schüller. 2014. Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *KR*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL-2019*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2020. Attention interpretability across {nlp} tasks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP-18*.

Junru Zhou and Hai Zhao. 2019. Head-driven phrase structure grammar parsing on Penn treebank. In *ACL-19*, Florence, Italy. Association for Computational Linguistics.