Kenny: In the following, if you are going to follow any of the suggestions by the reviewer, u can say "it's a very good idea that ..." or "it is true that...". Even if u are not going to follow the advise, as long as the comment has some merit, u can paise the reviewers in some appropriate way first, like flattering them. The purpose is not be too confrontational but be more consultational.

Reviewer #1

Q1:" For pruning layer selection, ... not examining lower layers ($l_b < 6$)?"
R1: Recent works (Jawahar, Sagot, and Seddah 2019; Rogers, Kovaleva, and Rumshisky 2020) have shown that lower layers (< 6) of PLMs capture basic linguistic knowledge like constituency syntax, chunking, and word order, while the semantic knowledge is generally learned in higher layers. The commonsense relational knowledge we focus on in this work is one kind of semantic knowledge. Thus, we choose the upper layers of PLMs for pruning.
Q2:" In Table 2, ... which doesn't make sense?"
R2: We have carefully checked our implementation, including different tokenization between WordPiece (used by DistilBERT, BERT, and MPNet) and BPE (used by RoBERTa) and optimization setting. However, we observe that RoBERTa consistently shows worse results. We posit that the large vocabulary size (50,265) of RoBERTa causes the training process to memorize surface co-occurrence patterns on this small training set of C-LAMA.
Q3:" From Figure 2 (right), $l_b = 6$ ... "
R3: Setting l_b to 6 removes more weights and produce more specialized subnetwork for a specific commonsense relation, compared to $l_b > 6$. However, it shows slightly worse performance on multi-relation scenario. We show the results of $l_b = 6$ because it shows the best trade-off.
Q4:" In Table 3, it is surprising ... "
R4: We haven't tried this setting, but it's an interesting direction that we would like to explore in the future.
Q5:" In Table 3, it is also surprising ... "
R5: DistilBERT is less over-parametrized than other larger PLMs. After pruning under the same sparsity regime, its subnetworks have the least amount of remaining weights, making them more specialized for different relations thus giving good results on KBC. QA tasks have more requirements beyond single commonsense relation, and the larger over-parametrization of models like BERT is preferred.

We show one case of KBC in R1 to Reviewer #3. The novel triple extraction part (Section 3.2) involves human evaluation. More error analysis will be added in the final version. Kenny: If you are putting anything in the final version, u better say what error analysis in more details.

Reviewer #2

Thanks for your feedback. We would like to reiterate the contribution and novelty of this paper: a) we are the first to ask the question of whether one can extract relation-specific commonsense knowledge from a general PLM; b) we propose a novel pruning framework to show that we can; and c) we conduct in-depth analysis (both qualitatively and quantitatively) and comprehensive experiments (12 datasets in total) to show that the extracted subnetworks can generalize better.

Reviewer #3

Q1:" It would be helpful to see some qualitative ..."
R1: One example on link prediction: For a triple <*classroom*, AtLocation, ?> from KBC test set, the top-3 predicted by original DistilBERT: *home*, *school*, *night* and those by pruned one: *school*, *college*, *university*. Some examples of extracted novel triples are listed in Appendix. We will include more multi-relation examples in Appendix.
Q2:" are there any ... consistently modeled than others"
R2: The most popular relations ranked by the number of appearances in the experiments in Section 3.3 are: IsA, CausesDesire, Desires, MotivatedByGoal, which are mostly about *causal* knowledge. Details about optimal relation sets for each task are provided in Appendix.
Q3:" It is mentioned that the code and pruned ..."
R3: The URL was purposely anonymized by a special tool for blind review. Our code and models have already been uploaded to the real URL and we will show it in the final version.

Reviewer #4

Q1:" The method ... a way of task-specific compression."
R1: We would like to clarify that we are not doing task-specific compression. Task-specific compression is to produce different sparse networks for different downstream tasks like was done in the reference provided by the reviewer. While in our study, our research question is whether we can transform a general PLM into dedicated knowledge models that inherit different relational commonsense knowledge from pretraining. This transformation should not introduce new knowledge (hence new parameters). We propose an effective pruning framework to identify subnetworks as the hidden knowledge models for various PLMs in Section 3.1. We then examine the knowledge transfer ability (Section 3.2 and 3.3) of these subnetworks on *real downstream tasks* by either zero-shot or standard fine-tuning.
Q2:" It's non-surprising that the general LM can be compressed or fine-tuned for a specific task."
R2: Please refer to R1 for our clarification on this.
Q3:" The comparisons between original and pruned in Figure 6 and Table 6 are not very meaningful, if we regard pruning as one way of fine-tuning ."
R3: Figure 6 and Table 6 show that our identified relation-specific subnetworks provide more task-relevant prior knowledge hence delivering better performance on those tasks.
Q4:" How about the comparisons between pruning (-fine-tuning) and other parameter-efficient-fine-tuning methods (such as adapter)?"
R4: Parameter-efficient methods like Adapter and Prefix-Tuning are motivated by alleviating catastrophic forgetting issues and high computational cost in standard fine-tuning,

and they both introduce additional parameters. Please refer to R1 for a detailed explanation.

References

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8: 842–866.