Exploring and Exploiting Latent Commonsense Knowledge in Pretrained Masked Language Models

Anonymous EMNLP submission

Abstract

Pretrained masked language models (PLMs) were shown to be inheriting a considerable amount of relational knowledge from the source corpora. In this paper, we present an in-depth analysis concerning eliciting relational commonsense knowledge already present in PLMs from the perspective of network pruning. We show that it is possible to find subnetworks capable of representing grounded commonsense relations at non-trivial sparsity meanwhile being generalizable to downstream commonsense knowledge base completion and commonsense reasoning tasks.

1 Introduction

The past few years have witnessed the revolution of NLP methods with the advent of pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019a). These bi-directional Transformer (Vaswani et al., 2017) encoders are first pretrained on vast amount of unlabeled text corpora and then fine-tuned on task-specific data, offering a surge of improvements on a wealth of downstream NLP tasks. Although this transfer learning paradigm has become the de-facto standard, we know very little about *what* and *how much* knowledge embedded in PLMs actually contributes to the success. Notable endeavors toward this understanding focus on probing linguistic knowledge therein. They demonstrated that pretraining did impart useful linguistic abstraction about syntax and semantics into PLMs (Peters et al., 2018; Goldberg, 2019; Tenney et al., 2019).

More recently, several works are presenting intriguing results examining the relational knowledge within PLMs. Relational knowledge (Speer and Havasi, 2012; Vrandečić and Krötzsch, 2014) is typically defined as describing the abstract relationship between a pair of concepts or entities, which is crucial for facilitating language understanding. Petroni et al. (2020) first posed the LAMA probe, an English benchmark comprising multiple sets of prompts. Each prompt is a cloze-like sentence transformed from a relational knowledge triple:

Knowledge Triple: *<bus, HasA, ?>*
Object Label: seats.
Sentence: you are likely to find _ in a bus.

By substituting _ with a special [MASK] token and reusing the masked language modeling (MLM) head, prompt-based relational knowledge probing provides an estimated lower bound of what PLMs know without training additional layer as in the previous linguistic probe. They showed that, even without grounded supervision, PLMs capture such relational knowledge at a level competitive to supervised alternatives. Subsequent works further showed that some specific prompts, acquired either through heuristical mining (Jiang et al., 2020) or gradient-guided search (Shin et al., 2020), can better trigger the models to correctly predict the missing object.

Despite the mounting evidence for the existence of relational knowledge in PLMs, it remains unclear how such knowledge are represented internally. It hinders the utility of PLMs being extended to more structured tasks, such as knowledge base completion. In light of this, we raise the core question in this paper: *Can we disentangle PLMs into relation-specific knowledge models and take advantage of them in a more flexible way?*

We approach this question by first drawing inspiration from recent findings (Saunshi et al., 2020; Lee et al., 2020; Zhang and Hashimoto, 2021): *the more cloze-like MLM pretraining simulates downstream task, the more successful the transfer will be.* For example, filling in *like* or *hate* into a cloze like *I [MASK] this film, it's great.* provide a clear way in which the model can implicitly learn to perform sentiment classification. Similarly, we hypothesis that MLM pretraining on clozes expressing certain relation r between masked word and remaining context would lead to good performance on

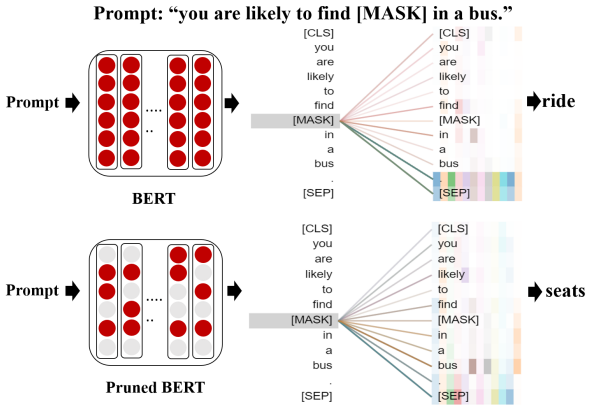

Figure 1: Querying original/pruned BERT-BASE with prompts of relation *HasA*. The color spectrum indicates the 12 attention heads in the last layer (Vig, 2019).

knowledge probing that targets relation r. Instead of designing cloze instances resembling specific downstream tasks, we exploit such correlation conversely: *the more successful the transfer is, the more cloze-like MLM pretraining simulates downstream tasks*. Specifically, we propose **weakly supervised weights pruning**, an end-to-end differentiable procedure to search for different subnetworks within PLMs that target zero-shot knowledge probing of different relations.

We show in experiment that it is possible to find subnetworks capable of representing grounded commonsense relations at non-trivial sparsity. Figure 1 exemplifies a cloze prompt where the identified subnetwork produces the valid answer *seats* by attending to relevant context, i.e., bus, while the full-scale BERT fails. We further investigate the possibility of repurposing these subnetworks on various downstream tasks. Experimental results on commonsense knowledge base completion show that the identified subnetworks perform on par with or even better than strong supervised knowledge base completion methods. We also explore the usage of the subnetworks on multiple commonsense reasoning tasks and empirically find that, when combined properly, the subnetworks can outperform the original PLMs in both many-shot and zero-shot settings.

In summary, our **contributions** include: (i) We present a novel way of eliciting relational knowledge hidden in PLMs from the perspective of network pruning (Section 2.3). (ii) Grounding on a concept-centric relation schema, we show that the proposed pruning procedure successfully identified sparse subnetworks specializing in miscellaneous commonsense knowledge remarkably better than their full-scale counterparts (Section 3.1). (iii) We showcase the effectiveness of these subnetworks on commonsense knowledge base completion tasks (Section 3.2) as well as a heuristic application on multiple commonsense reasoning tasks, gleaning insight on the transformation from language representation to knowledge representation.

We release code and all versions of our pruned PLMs at `https://anonymous.4open.science`.

2 Methodology

We first provide background on pretrained masked language models and the formulation of cloze prompt for querying these models, then we proceed to elaborate our proposed pruning procedure.

2.1 Pretrained Masked Language Models

Formally, given a sequence of tokens $\boldsymbol{w} = [w_1, w_2, ..., w_n]$, where n is its length, the model outputs a sequence of fixed-size hidden representations $\boldsymbol{h} = [h_1, h_2, ..., h_n]$ for each token. In standard MLM pretraining, the corresponding representation h_i is fed into a designated MLM head for computing the reconstruction probability $P(w_i|\boldsymbol{w}_{-i})$ of the masked i-th token w_i, where $\boldsymbol{w}_{-i}$ are the remaining unmasked tokens. We denote the original pretrained model $\mathcal{LM}$ with unpruned parameter θ as $\mathcal{LM}_\theta$ in following sections.

2.2 Knowledge Probing with Cloze Prompts

The natural language cloze prompts, such as "*you are likely to find a basement in below your [MASK]*", offer a straightforward mean of querying pretrained masked language models that conform to their interfaces.

We follow the formulation of Petroni et al. (2020), where relational knowledge is in the form of triplets $\langle subj, r, obj \rangle$. Here $subj$ refers to the subject, obj refers to the object, and r indicates their corresponding relation. To query a model $\mathcal{LM}_\theta$, each relation r is associated with a set of cloze template prompts T_r, each of which consists of a sequence of tokens, two of which are placeholders for $subj$ and obj (e.g., "*you are likely to find* $[subj]$ *in* $[obj]$"). We can check the existence of the knowledge in $\mathcal{LM}_\theta$ by substituting the $[subj]$ place-holder with the surface form of real subject and asking the model to predict the missing

object:

$$\hat{obj} = \arg\max_{w \in \mathcal{V}} P_{\mathcal{LM}_\theta}([obj] = w | subj, T_r)$$

where $\mathcal{V}$ is the vocabulary of $\mathcal{LM}_\theta$. We say that $\mathcal{LM}_\theta$ grasps the knowledge if $\hat{obj} = obj$.

2.3 Weakly Supervised Weights Pruning

Given a pretrained masked language model $\mathcal{LM}$ and the associated set of pretrained parameters $\theta \in R^d$, where d is the dimensionality, we are interested in finding the subnetwork $\mathcal{LM}_{\theta_r}$ that is maximally predictive of prompts of relation r. The intuition is that if a subnetwork specializes exclusively on relation r, the parameters it reserves should inherit the corresponding knowledge from MLM pretraining on cloze instances of r.

Similar to Zhao et al. (2020), for each weight matrix W^l from the set of all weight matrices $\boldsymbol{W}^l$ in the l-th transformer layer, we assign learnable pruning mask generator G_r^l that is element-wise initialized from a prior distribution $\phi(\cdot)$. Each entry $g_{i,j}^l \in G_r^l$ is a real-valued scalar that determines whether its corresponding weight $w_{i,j}^l \in W^l$ should be pruned. To investigate if $w_{i,j}^l$ should be softly scaled or entirely removed to effectively recover $\mathcal{LM}_{\theta_r}$,we explore two different schemes of converting G_r^l into a masking matrix M_r^l from a probabilistic view.

2.3.1 Stochastic Pruning

The first variant is to establish a probabilistic formulation for determining the importance of individual weights. Formally, $g_{i,j}^l$ is taken as input to a sigmoid function for parametrizing a Bernoulli distribution $B(\sigma(g_{i,j}^l))$, from which a binary masking random variable $m_{i,j}^l$ is sampled:

$$m_{i,j}^l \sim B(\sigma(g_{i,j}^l)) \tag{1}$$

where $m_{i,j}^l \in M_r^l$. The resulting masking matrix M_r^l can then be used to select weights within original linear layer W^l by Hadamard product:

$$W_r^l = W^l \odot M_r^l \tag{2}$$

Due to the non-differentiability introduced by sampling, the gradient w.r.t. loss function (described in Section 2.3.3) cannot be back-propagated to $g_{i,j}^l$. As a remedy, we use the re-parametrization technique (Li et al., 2018) to approximate $m_{i,j}^l$ with another differentialble variable $\tilde{m}_{i,j}^l$:

$$\tilde{m}_{i,j}^l = \sigma\left(\frac{g_{i,j}^l + \log U - \log(1-U)}{\tau}\right) \tag{3}$$

where $U \sim Uniform(0,1)$ and τ is a small positive temperature parameter. As τ approaches zero, $\tilde{m}_{i,j}^l$ will match sampled $m_{i,j}^l$ more accurately (detailed proof can be found in Appendix A).

In this way, Eq. (2) becomes:

$$W_r^l = W^l \odot \tilde{M}_r^l \tag{4}$$

2.3.2 Deterministic Pruning

While our first probabilistic pruning formulation considers flexible weights combination, the second proposed variant utilizes a hard thresholding function to directly generate the masking matrix.

Let t denotes the predefined thresholding hyperparameter ranging from 0 to 1, then we have:

$$\hat{m}_{i,j}^l = \begin{cases} 1, & \sigma(g_{i,j}^l) \geq t, \\ 0, & otherwise. \end{cases} \tag{5}$$

where σ is the sigmoid function. Similar to Section 2.3.1, the resulting binary masking matrix $\hat{M}_r^l$ is then used to select weights relevant to relation r by Hadamard product:

$$W_r^l = W^l \odot \hat{M}_r^l \tag{6}$$

Note that the hard thresholding operation in Eq. (5) also blocks the gradient propagation to $g_{i,j}^l$. Here we employ the Straight-Through gradient estimator (Bengio et al., 2013; Hubara et al., 2016) and use $\frac{\partial \mathcal{L}_r}{\partial \hat{m}_{i,j}^l}$ as a proxy of $\frac{\partial \mathcal{L}_r}{\partial g_{i,j}^l}$. We elaborate on the loss function $\mathcal{L}_r$ w.r.t relation r in the next section.

2.3.3 Training and Inference

The resultant pruned model (i.e., subnetwork) $\mathcal{LM}_{\theta_r}$ is expected to behave like a specialized neural knowledge base. That is, given a prompt requiring knowledge about relation r, $\mathcal{LM}_{\theta_r}$ should be able to fill in the missing object more accurately than its full-scale counterpart $\mathcal{LM}_\theta$. To this end, the learning objective for pruning mask generator $\{\boldsymbol{G}_r^l\}_{l_b \leq l \leq l_t}$, where l_b and l_t indicate the range of transformer layers, is to find the subnetwork $\mathcal{LM}_{\theta_r}$ that minimizes the following objective:

$$\mathcal{L}_r = -\mathbb{E}_{(subj,T_r,obj)\sim D_r}[\log P_{\mathcal{LM}_{\theta_r}}(obj|subj,T_r)]$$

where D_r is the collection of prompts under relation r. The training procedure is conducted for each relation $r \in \mathcal{R}$ of interest and finally, we acquire a set of trained $\{\boldsymbol{G}_r\}_{r\in\mathcal{R}}$ for the designated pretrained model $\mathcal{LM}$.

During inference, for deterministic pruning, M_r is obtained from G_r by Eq. (5). For stochastic pruning, M_r is obtained by taking the expectation value (i.e., $\sigma(G_r)$) of Bernoulli variables.

Model	P@1 (%)	P@2 (%)	P@3 (%)	Sparsity	$l_b - l_t$	# Param.
DISTILBERT-BASE w/o pruning	11.4	16.6	19.9	0%	-	66M
DISTILBERT-BASE w/ stochastic pruning	14.8	21.5	26.3	∼30%	4-6	66M
DISTILBERT-BASE w/ deterministic pruning	44.1	52.9	57.6	∼50%	4-6	56M
BERT-BASE w/o pruning	12.9	18.4	21.8	0%	-	110M
BERT-BASE w/ stochastic pruning	17.2	25.1	29.6	∼30%	7-12	110M
BERT-BASE w/ deterministic pruning	57.6	63.8	67.2	∼50%	7-12	88M
ROBERTA-BASE w/o pruning	15.4	21.2	24.6	0%	-	125M
ROBERTA-BASE w/ stochastic pruning	16.6	22.2	25.8	∼30%	7-12	125M
ROBERTA-BASE w/ deterministic pruning	38.3	42.8	44.6	∼50%	7-12	100M
MPNET-BASE w/o pruning	14.8	20.7	24.0	0%	-	110M
MPNET-BASE w/ stochastic pruning	19.8	27.9	33.2	∼30%	7-12	110M
MPNET-BASE w/ deterministic pruning	62.7	68.7	71.4	∼50%	7-12	88M

Table 1: Relational knowledge probing results on C-LAMA. We show one representative pruning configuration for each type of model and relegate the complete results to Appendix due to space limits.

3 Experiments

We first expound our pruning setting and provide evidences of its ability to identify relation-specific subnetworks in PLM. Then we experiment on several commonsense-intensive scenarios to seek good practices for using these subnetworks.

3.1 Disentangling PLMs into Relation-specific Subnetworks

Corpus	#Facts	#Rels	#Sentences
ConceptNet	11,458	16	29,774

Table 2: Statistics of C-LAMA.

Dataset. We use the ConceptNet (Speer and Havasi, 2012) subset of LAMA benchmark as weak supervision, denoted as C-LAMA. C-LAMA contains facts from the English part of ConceptNet that has single-token objects covering 16 relations. Prompts that will be used for pruning and probing are extracted from Open Mind Common Sense (OMCS). Statistics is listed in Table 2. Since our goal is to explore the existence of specialized subnetworks within general-purpose PLMs, we utilize the whole dataset to fully exploit the high-precision knowledge therein. This naturally deviates from the conventional supervised experimental train/valid/test setting.

Models. For the choices of $\mathcal{LM}$, we consider the 6-layer DISTILBERT-BASE (Sanh, 2019), 12-layer BERT-BASE, 12-layer ROBERTA-BASE (Liu et al., 2019a). We also include the more recent powerful 12-layer MPNET-BASE (Song et al., 2020) model. All models are implemented with HuggingFace's transformers (Wolf et al., 2019) library.

Setup. The prior distribution $\phi(\cdot)$ is a Gaussian $\mathcal{N}(\mu, 1)$ where μ is the mean controlling initial sparsity of pruned model (e.g., $\mu = 0$ indicates 50% initial sparsity). We set l_t to be the top layer of a given model and set l_b to be within $3-4$ for DISTILBERT, $5-9$ for BERT, ROBERTA, and MPNET. The temperature τ is fixed as 0.1. The threshold t is fixed as 0.5. We use Adam (Kingma and Ba, 2015) with a batch size of 32 and a linear warm-up scheduler with 0.1 warm-up ratio for training the mask up to 6 epochs. The learning rate is fixed as 3×10^{-4}. All experiments are conducted on a GTX 1080 Ti GPU with 11G RAM.

To what extent can we achieve disentanglement? As stated in Section 2.3, for each model, our pruning procedure will identify 16 subnetworks for 16 commonsense relations. We record the P@K (K=1,2,3) scores of each subnetwork $\mathcal{LM}_{\theta_r}$ evaluated on its corresponding subset D_r of C-LAMA and shows the macro-averaged results in Table 1.

Among all models without pruning, ROBERTA achieves the highest P@1 score of 15.4 while DISTILBERT gets the lowest 11.4. It indicates that while PLMs are shown to be helpful for downstream learning, they cannot accurately complete cloze-like prompts that require commonsense relation knowledge. This observation also coincide with previous finding (Zhang and Hashimoto, 2021) that the uniform masking adopoted by PLMs is biased towards extracting statistical and syntactic dependencies. Comparing the results for each pair of original and subnetworks, we consistently observe a surprisingly significant increase (37.0 on average), especially for deterministically pruned ones. This large performance gap provides unique

new evidence of sparse latent relational knowledge structures in PLMs, which are weakened by pretrained weights that are *reserved* for more general-purpose use.

We also observe that the deterministic pruning excels by a huge margin across all models, which suggests the importance of removing noisy transformation of input representations to regain the expressivity of specific commonsense knowledge. Another advantage of deterministic pruning in memory footprint is that only sets of 1-bit masks rather than 32-bits float parameters need to be saved for solving multiple tasks. For the above reasons, we focus our analysis on and use pruned to denote deterministically pruned PLMs throughout this paper henceforth.

How specialized are these subnetworks? We next investigate alternative ways of building the subnetworks. We attempted to instantiate the pruning mask upon BERT-BASE via either creating a set of mismatched masks or assigning a randomly initialized mask with comparable sparsity. For *mismatched* one, we corrupt the correspondence of relation between masks and prompts by shifting the order of masks 15 times, as there are 16 relations in total. Then we calculate the micro-averaged P@K for each shift and average the results. For *random* one, we initialize the mask element-wise with a Bernoulli $B(0.5)$ and average the results from 5 different random seeds. The results are shown in

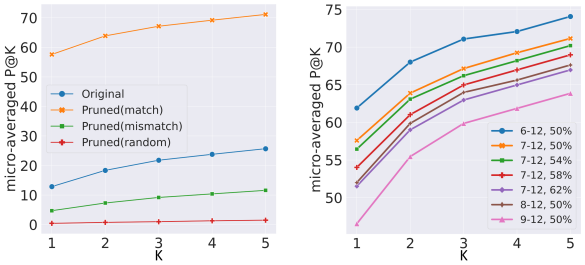

Figure 2: Ablation on the pruning masks (left) and effect of initial sparsity and pruned layers (right).

Figure 2 (left). If we apply the random masks with sparsity comparable to learned ones, the P@1 drops drastically to 0.4. This notable gap proves that the effective subnetworks cannot be trivially identified through random weights sampling. If we apply the mismatched masks from other relations, the P@1 also significantly drops to 4.8, even inferior to the unpruned one. It shows that the latent structure for different types of commonsense knowledge exhibits remarkably distinct geometry.

Factors impacting performance. We also investigate how initial masking sparsity and choice of layers to prune influence the probing performance. We experiment on BERT-BASE with l_b in $\{6, 7, 8, 9\}$ and initial sparsity in $\{50\%, 54\%, 58\%, 62\%\}$. Figure 2 (right) shows that (i) increasing the number of pruned layers helps distill more knowledge. (ii) larger initial sparsity is more likely to prune away weights important to certain knowledge and cannot be recovered in the later training process. In general, we find an initial sparsity around 50% yields decent performance both in probing and downstream applications (see Section 3.3 and Section 3.2).

Visualization of attention weights and representations. To explain how the subnetworks accommodate more accurate commonsense knowledge despite having far fewer weights than the full-scale models, we randomly sample several prompts that the subnetworks correctly answered but the full-scale model (BERT-BASE) failed to and visualize the attention patterns in the last layer. Specifically,

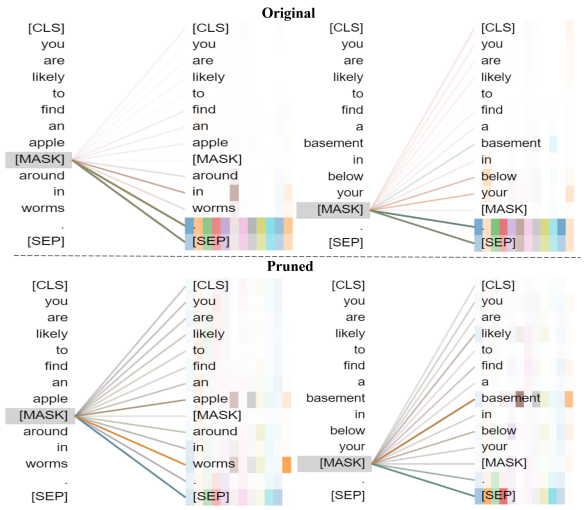

Figure 3: Attention weight visualization. *AtLocation* is required for prompt in the left column and *PartOf* is required for prompt in the right column.

we focus on the attention weights between [MASK] token and other tokens in the prompt. A first glance of change of attention pattern is given in Figure 1 and we show more examples of other ConceptNet relations in Figure 3. We observe that while the original pretrained model tends to attend to special tokens like period and [SEP], the subnetwork successfully grasps the relevant concepts (i.e., apple, worms, and basement) in the prompt hence produces the right object. We also use t-SNE (van der Maaten and Hinton, 2008) to visualize the last

Model	Development Set				Test Set			
	MRR (%)	P@1 (%)	P@2 (%)	P@3 (%)	MRR (%)	P@1 (%)	P@2 (%)	P@3 (%)
Supervised								
DISTMULT (Yang et al., 2015)	8.5	4.2	6.6	8.3	10.5	5.4	8.4	10.9
COMPLEX (Trouillon et al., 2016)	10.7	6.5	9.0	11.0	13.6	8.2	12.4	15.7
CONVE (Dettmers et al., 2018)	18.9	11.5	16.6	19.0	21.9	13.5	18.9	24.0
TUCKER (Balazevic et al., 2019)	17.3	10.9	14.8	18.8	21.6	14.0	20.4	24.0
CONVTRANSE (Shang et al., 2019)	19.8	13.2	17.8	21.3	24.0	**15.6**	21.9	26.5
SACN (Shang et al., 2019)	21.2	13.2	19.8	23.2	**24.2**	14.4	22.1	**28.0**
Unsupervised								
DISTILBERT-BASE	9.0	3.1	6.9	10.3	10.8	5.8	9.6	11.2
BERT-BASE	12.4	7.2	10.0	13.7	14.3	8.3	13.7	16.6
ROBERTA-BASE	8.3	4.2	6.0	7.1	9.4	5.1	7.1	9.3
MPNET-BASE	11.7	7.2	9.4	11.1	11.1	6.0	9.9	11.7
DISTILBERT-BASE (pruned)	**24.1**	**15.8**	**24.1**	26.4	23.4	14.8	**22.2**	26.5
BERT-BASE (pruned)	23.7	15.5	22.1	**27.0**	22.8	14.3	20.9	26.0
ROBERTA-BASE (pruned)	9.0	4.9	7.1	8.9	9.5	6.1	7.6	11.4
MPNET-BASE (pruned)	22.1	12.9	21.2	25.5	20.0	11.4	18.8	22.9

Table 3: Link prediction results. Best results are marked with **bold** font and second best with underline.

layer's representation of [CLS] for each prompt. From Figure 4, the representations computed by original pretrained model are hardly separable as different types of knowledge are intermingled. In contrast, the pruned subnetwork can extract meaningful and disentangled representations for different commonsense relations.

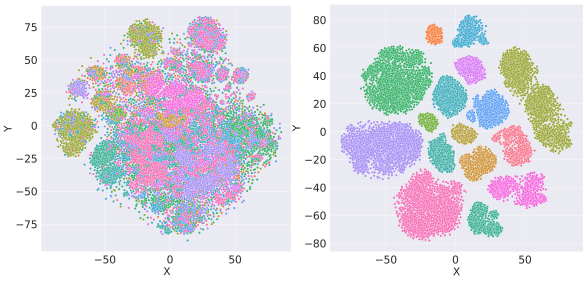

Figure 4: t-SNE visualization of [CLS]'s representation from original (left) and pruned (right) BERT-BASE.

3.2 Commonsense Knowledge Base Completion (CKBC)

We evaluate the utility of identified relation-specific subnetworks on CKBC in an unsupervised manner. Specifically, we use the ConceptNet-100K benchmark provided by Li et al. (2016). To allow a fair evaluation, we manually create a subset of ConceptNet-100K consisting of triples with single-token subject/object, of which the dev/test set have **no overlap** with C-LAMA. Each relation is associated with a sentence template (provided in Appendix) (Kwon et al., 2019) of which the wording is distinct from those in C-LAMA. We acknowledge that these sentence templates might be suboptimal for certain relations, but prompt optimization is out of the scope of this paper. The resulting dataset contains $17,891$ training instances, 349 development instances, and 446 test instances.

Link prediction. We first formulate CKBC as a link prediction task and compare subnetworks (i.e., $\mathcal{LM}_{\theta_r}$ is queried to predict missing link for instance of relation r) as well as original PLMs against strong supervised KB completion mothods.

Table 3 shows the results. Most of the supervised models outperform full-scale PLMs by a large margin, which suggests the inefficacy of directly using PLMs to perform link prediction. However, the subnetworks identified by our pruning procedure can acquire performance on par with or better than state-of-the-art supervised models. Surprisingly, the pruned DISTILBERT get the highest MRR, outperforming other larger and more advanced PLMs. ROBERTA struggles to predict correct objects, perhaps due to its larger vocabulary size compared to WordPiece ($50,265$ vs $30,522$) and less lexicon overlap (53% vs 59%) with the dataset.

Triple classification We can also formulate CKBC as a triple classification task. Following Davison et al. (2019), we use estimated point-wise mutual information (PMI) computed by pretrained language model as a surrogate of a triple's validity. An expectation-maximization-based Gaussian mixture clustering method is used and instances in the cluster with higher mean PMI are labeled as valid. In our preliminary experiments, we found that the model pruned by the mask of a single relation might not be robust for PMI estimation and generally performed inferior to the intact model. In the same spirit as model ensembling, we then perform grid search over combinations of multiple knowledge, which is similar to what we did in zero-shot commonsense reasoning. For all four PLMs considered in Table 4, we observe that there exists

Model	F1 Score
DistilBERT-base	74.1
DistilBERT-base (pruned)	**76.3**
BERT-base	73.7
BERT-base (pruned)	**76.7**
RoBERTa-base	74.8
RoBERTa-base (pruned)	**76.9**
MPNet-base	76.5
MPNet-base (pruned)	**78.0**

Table 4: Triple classification on ConceptNet-100K.

one or multiple knowledge combinations delivering F1 score higher than the original models.

Triple extraction. We then investigate the ability of specialized subnetworks to extract novel commonsense knowledge triples absent from the dataset. We randomly sample 100 triples from the test set of ConceptNet-100K and for each sample use top-K predictions from pruned DistilBERT-base as candidate objects. Three human annotators are asked to first determine the correctness of each candidate object and further determine their novelty (i.e., not present in any of train/validation/test set) if deemed to be correct. The Fleiss Kappa inter-annotator agreement κ is 0.66/0.65 for precision and novelty, respectively. Figure 5 shows the change of precision-novelty

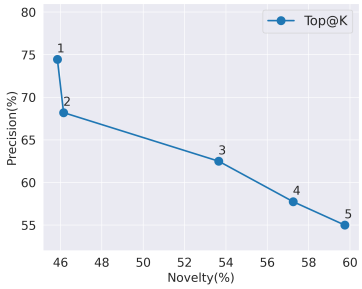

Figure 5: Precision-novelty curve with varied K.

with varied K. We observe a clear trade-off between the validity and novelty of triples extracted by the pruned model. As expected, a large K inevitably makes noisy predictions but is more likely to extract unseen knowledge. For the purpose of knowledge enrichment, one might choose a large K to ensure a desirable recall. We list the obtained novel triples in the Appendix D due to space limits.

3.3 Commonsense Reasoning (CSR)

After identifying sparse subnetworks within PLMs that specialize in different commonsense knowledge, we now evaluate their generalization ability

Task	Original	Pruned	p-value
RTE	69.2±2.3	69.8±2.0	0.12
COPA	62.4±5.0	63.0±4.7	0.33
CommonsenseQA	53.1±0.6	54.1±0.7	0.08
SWAG	73.9±0.3	74.2±0.1	0.09
HellaSWAG	38.9±0.4	39.1±0.5	0.32
aNLI	63.7±0.4	64.0±0.4	0.19
CosmosQA	61.3±1.0	61.8±0.2	0.26

Table 5: Fine-tuning results of BERT-base for CSR.

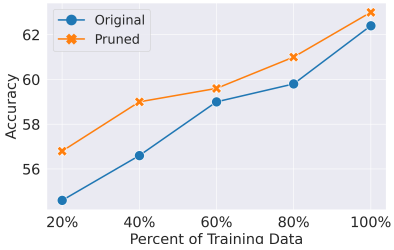

Figure 6: BERT-base results on COPA.

in the context of commonsense reasoning.

Many-shot learning. We experiment with BERT-base and its deterministically pruned version using supervised fine-tuning on 7 datasets: RTE (Dagan et al., 2009), COPA (Roemmele et al., 2011), CommonsenseQA (Talmor et al., 2019), SWAG (Zellers et al., 2018), HellaSWAG (Zellers et al., 2019), aNLI (Bhagavatula et al., 2020) and CosmosQA (Huang et al., 2019). For each task, we heuristically identify the commonsense knowledge it might requires. If multiple types of knowledge are required, we simply take the union of all masks and apply the resultant mask to the pretrained model as initialization for finetuning. We repeat training three times with different random seeds for each task. The choice of mask combination for each task can be found in Appendix B.

The results in Table 5 shows that, when initialized with proper weights, the model can be better fine-tuned on downstream commonsense reasoning tasks via more useful *prior* knowledge. We further analyze the change of performance under the low-resource regime on COPA dataset. Figure 6 shows that the pruned BERT exhibits a notable advantage when training data is extremely scarce. As more training data is seen, the benefit of the pruned model becomes less prominent, i.e., $p > 0.05$.

Zero-shot learning. We next assess the ability of specialized subnetworks to perform zero-shot commonsense reasoning, a setting where the knowledge relied on to complete the task is solely determined by the model parameters. Here we focus on: COPA, CommonsneseQA, Conjunction Accept-

Model	COPA (Tra.)	COPA (Dev.)	CSQA	CA	WSC	SM	ARCT1	ARCT2	Average
DistilBERT-base	58.3	60.0	29.6	84.6	53.3	71.6	48.6	50.4	57.0
DistilBERT-base (pruned)	**61.5**	**69.0**	**31.5**	**89.6**	**56.9**	**72.1**	**53.4**	**51.6**	**60.7**
BERT-base	60.2	54.0	26.5	89.0	57.3	69.7	46.8	50.3	56.7
BERT-base (pruned)	**63.0**	**64.0**	**28.5**	**91.8**	**59.0**	**71.7**	**50.0**	**52.0**	**60.0**
RoBERTa-base	60.7	59.0	39.9	90.1	61.8	73.1	48.6	53.1	60.7
RoBERTa-base (pruned)	**65.3**	**72.0**	**40.4**	**93.4**	**62.9**	**74.4**	**53.2**	**55.1**	**64.6**
MPNet-base	66.5	69.0	40.0	94.5	64.3	75.8	52.9	56.7	64.9
MPNet-base (pruned)	**71.0**	**74.0**	**41.7**	**97.3**	**66.4**	**77.5**	**56.1**	**57.7**	**67.7**

Table 6: Zero-shot results of accuracy (%) on commonsense reasoning tasks. Better results of each pair is in **bold**.

ability (CA) (Zhou et al., 2019), Winograd Schema Challenge (WSC) (Levesque et al., 2012), SenseMaking (SM) (Wang et al., 2019), ARCT1 (Habernal et al., 2018) and ARCT2 (Niven and Kao, 2019). Each sample in the above datasets can be formulated as *[CLS] premise [SEP] hypothesis$_i$ [SEP]*, where i is the subscript. Hypothesis with the highest plausibility scored by PLMs is the predicted answer.

Since multiple types of knowledge are typically required for effectively reasoning over concepts, for each task, we perform grid search over combinations of 3-4 different commonsense knowledge out of the 16 total types and reported the best accuracy in Table 6. We put the best combination for each model on each task in Appendix B for space constraints. By combining multiple commonsense knowledge useful for the task, we show that the pruned models can actually surpass their full-scale version in all tasks considered in our experiments. The most likely explanation is that knowledge irrelevant to the specific task in the original models hurt the in-domain zero-shot reasoning capability. It also manifests that the most important reasoning skills vary from task to task.

4 Related Work

Since the emergence of large pretrained language models, much work has focused on understanding internal contextual representations produced by such models. Most prior work (Shi et al., 2016; Belinkov et al., 2017) pays special attention to either using extraneous probing tasks to examine whether certain linguistic properties can be identified from those representations or ablating the models to observe how behavior changes. More recently, some studies (Peters et al., 2018; Goldberg, 2019; Tenney et al., 2019) have shown the existence of linguistic knowledge (e.g., syntax) in various but generally lower layers of pretrained transformers.

It is worthy of exploring and shedding more light on how PLMs memorize abstract knowledge rather than trivial statistical co-occurrence patterns. We extend previous probe (Petroni et al., 2020) on relational knowledge. Specifically, we are concerned with commonsense knowledge that is grounded on ConceptNet relations. Our work differs in that we focus on not only probing but also bringing potentially more implicit commonsense knowledge to the surface and unleashing more potential in knowledge-intensive applications.

Another line of researches relevant to our work is network pruning (Liu et al., 2019b; Lin et al., 2020) and the lottery ticket hypothesis (Frankle and Carbin, 2019; Prasanna et al., 2020; Chen et al., 2020). The former aims at reducing the size of model parameters without compromising test accuracy and the latter reveals subnetworks whose initializations made them capable of being trained effectively comparable to the original model. In contrast, we seek to uncover subnetworks in overparametrized PLMs that specializes on commonsense knowledge tailored for downstream tasks rather than focusing on good global initialization, and achieve good results.

5 Conclusion

We apply network pruning, a novel approach to explore the latent relational knowledge hidden in PLMs. With a preliminary focus on commonsense knowledge, we find evidence of latent sparse subnetworks capable of representing grounded commonsense relations in a plethora of PLMs. Further experiments on downstream tasks showed that such subnetworks can be effectively utilized as auspicious neural knowledge bases, fine-tuning starting points, and robust zero-shot reasoners. Our work raises a new viewpoint about the inner storage scheme as well as practical utilization of relational knowledge in PLMs, opening up avenues to future work on better understanding and adapting pretrained language representations.

References

Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor factorization for knowledge graph completion. 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 5185–5194, Hong Kong, China. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(Special Issue 04):i–xvii.

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.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4107–4115.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sunjae Kwon, Cheongwoong Kang, Jiyeon Han, and Jaesik Choi. 2019. Why do masked neural language models still need common sense knowledge? *arXiv*.

Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2020. Predicting what you already know helps: Provable self-supervised learning. *CoRR*, abs/2008.01064.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Zhuohan Li, Di He, Fei Tian, Wei Chen, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Towards binary-valued gates for robust LSTM training. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3001–3010. PMLR.

Tao Lin, Sebastian U. Stich, Luis Barba, Daniil Dmitriev, and Martin Jaggi. 2020. Dynamic model pruning with feedback. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. 2019b. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. Language models as knowledge bases? *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2463–2473.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University.

Victor Sanh. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2020. A mathematical exploration of why language models help solve downstream tasks. *CoRR*, abs/2010.03648.

Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. End-to-end structure-aware convolutional networks for knowledge base completion. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3060–3067. AAAI Press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense

knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 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, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. *CoRR*, abs/1606.06357.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig. 2019. Visualizing attention in transformer-based language representation models. *CoRR*, abs/1904.02679.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

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.

Tianyi Zhang and Tatsunori Hashimoto. 2021. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. *CoRR*, abs/2104.05694.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241, Online. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating Commonsense in Pre-trained Language Models.

Review:

META-REVIEW

Comments: This paper presents a set of experiments in which pruning methods are used to identify (relatively sparse) subnetworks within pretrained MLMs that encode "commonsense" knowledge. While the practical use of this approach is not immediately clear, reviewers agreed that this is nonetheless an interesting direction and analysis, and the authors clarified key remaining questions about the work in their response. These clarifications should be integrated into future versions of the manuscript.

==========

REVIEWER #1

==========

The core review

This paper presents an in-depth analysis of the nature of relational knowledge elicited by pretrained language models. Along these lines, the authors posit that it is possible to find subnetworks at non-trivial sparsity that can generalize to downstream commonsense reasoning tasks.

Strengths:

- Understanding the nature of relational knowledge present in pretrained language models is an important challenge and this paper tackles that by identifying a smaller sparse subnetwork
- They validate this by deploying the pruned model to downstream tasks

Weaknesses:

- It is hard to understand the motivation of the paper. What does pruning really show ? Presence of relation-only subnetwork ? If yes, the authors have not made an attempt to show that such networks do not capture any additional information. From an end-task perspective, are those the only "subnetworks" that can be used to enhance performance ? If yes, is the relational knowledge distributed differently than the authors claim ?
- It is unclear how their network specifically targets "relational commonsense" knowledge. In particular, the current "knowledge probing with cloze prompts" is not sufficiently convincing that it is meant to capture relational knowledge in the first place
- The paper is fraught with terms that are either vague or not defined
    -- Line 97 - what does "embodies" mean ?
    -- Line 89 - what does "discrepancy" mean ?
    -- Line 415 - what does "special' mean ?

- The authors don't adequately motivate why the visualization of attention weights is necessary to establish any analysis.

Reasons to Accept

The paper is not ready to be accepted for scholarly publication

---

Reasons to Reject

The paper is not motivated well, and currently lacks some foundational basis albeit having a good idea to understand the relationship between pruning and knowledge.

Reviewer's Scores

Overall Recommendation: 3.5

==========

REVIEWER #2

==========

The core review

This paper addresses the question of disentangling relational knowledge from pretrained language embeddings. They propose an end-to-end weakly supervised weights pruning method to search for subnetworks within pretrained language models in which relational knowledge is elicited. They show that the proposed method identifies sparse subnetworks which performs well on several downstream tasks. For pruning they propose two different methods - (1) stochastic pruning - where the first method a binary masking variable is sampled from a Bernoulli distribution (2) deterministic pruning - where a hard thresholding function is used.

Reasons to Accept

The paper provides new insights into how relational knowledge is captured in pretrained language models. Experiments are thorough. The authors show a huge performance difference between finetuning the original language model vs finetuning the pruned model.

Reasons to Reject

There is very little discussion about the practical implications of identifying these subnetworks. Apart from the understanding that they exist and can be finetuned for downstream tasks, there is no discussion on why there is such a huge performance difference between finetuning original BERT vs finetuning pruned BERT.

Reviewer's Scores

Overall Recommendation: 3.5

Questions for the Author(s)

In Table 1, the precision is zero for models finetuned on the original bert. Any reason why pure Bert finetuning would do so poorly? Also, how does finetuning RoBERTA (without pruning) perform for this task?

==========

REVIEWER #3

==========

The core review

In this paper, the authors probe into the common sense knowledge already present in the pre-trained masked language models by using weakly supervised weight pruning techniques. The idea is to find subnetworks within PLMs in which relational knowledge is better represented. The authors show that grounding on external relation schema successfully identifies sparse subnetworks. They specifically investigate two questions (a) if they can disentangle the pre-trained general-purpose knowledge representation into a relation-specific knowledge representation, (b) if the relation-specific knowledge can be used for downstream knowledge-intensive tasks.

Strengths:
- The paper proposes novel pruning strategies to find subnetworks that take into account relational knowledge.
- Empirical evaluation is conducted on multiple PLMs and tested on 7 common sense reasoning tasks.

Weakness:
- As Petroni et. al 2019 and authors mention PLMs are parameterized to have a knowledge representation that is entangled in the shared parameter space. However, when fine-tuning the general knowledge representation is tuned towards a specific task. I wonder if the results presented by the pruning methods proposed would be significant after fine-tuning to the task.
- Some of the experiments and results are unclear. Why BERT-BASE-FINETUNED-SQuAD does not perform on the LAMA task (0.0 for P@1) as the SQuAD dataset is also a subset of LAMA?

Reasons to Reject

- It is understandable that subnetworks of PLMs are useful for representing knowledge. However, it is unclear their usefulness after fine-tuning PLMs for a task. A fine-tuned model would ideally capture the required common sense relational representation for the specific task.

Reviewer's Scores

Overall Recommendation: 3.5

Rebuttal:

Response to Review #1:
Thanks for your valuable review. Regarding your questions:
Q1: "...What does pruning really show..."
A1: The pruning technique learns parts of a PLM that are specicialized for a certain type of commonsense relation by keeping the relevant parameters and reseting the rest of the paramenters. The resulting subnetwork can accurately model P(object|subject, relation) for a particular relation. This is shown in Section 3.1 line 420-425, 444-447, as well as in Figure 2.
The main motivation of the paper is to show that PLM can be partitioned into these relation-specific subnetworks, and not to improve the performance of any particular down-stream tasks, although we have successfully shown that some of the down-stream tasks can indeed benefit from
the PLM pruning, as a by-product. In contrast, Chen et al.[1] prunes BERT to obtain a subnetwork that targets a specific application and not a relation like in our work, which is more general.

References:
[1]. The Lottery Ticket Hypothesis for Pretrained BERT Networks. NIPS 2020.

Q2: "knowledge probing with cloze prompts" is not convincing...
A2: The cloze promts are written by human annotators in OMCS. Each prompt describes a commonsense relation that can be formulated as a subject-preditcate-object triple. The blank is carefully chosen by human that is either the subject or the object of the triple.
We use cloze prompts here because they well unifies the human language model and structured knowledge.

Q3: "The paper is fraught with...":
A3:
1. "embody" means some MLM pretraining instances require commonsense knowledge of certain relations to correctly predict masked word. Done
2. "discrepancy" means the difference between MLM pretraining formulation P(word_masked|word_unmasked) and relational knowledge evaluation formulation P(object|subject, relation). Done
3. "special" means the uniqueness of a particular subnetwork for representing a specific commonsense relation. Done
We will correct some of the language uses according to your suggestion.

Q4: "no motivation of visualization of attetion weights...":
A4: Before pruning, the attention between words in the example cloze prompt shows no particular interesting pattern; after pruning, there is clearly strong attention between the subject and the object words for the same example. This is a good evidence that our technique strengthens the PLM representation of the relation depicted in the prompt.

Response to Review #2:
Thanks for your valuable review. Regarding your questions:
Q1: "the precision is zero for models finetuned on ..." Done
A1: First of all, all the P@1 scores of Table 1 are results of LAMA, which tests the model's understanding of structured knowledge, and it's not a downstream task. Second, fine-tuning BERT on downstream tasks is essentially adapting the pretrained representations to capture task-specific features, usually by heavily modifying the outer-most layers of BERT. It has been shown that large language models, when fine-tuned, are prone to exploiting superficial statistical cues (Naik et al.[1], Sanchez et al.[2], McCoy et al.[3]) in the dataset, e.g., trigger word or n-gram indicative for certain label, to achieve high performance without truly utilizing "knowledge". That is why BERT fine-tuned by CONLL and SQuAD suffers the most on the LAMA test, because it "forgets" some of its previously acquired knowledge. The same goes for fine-tuned RoBERTa. The precisions for RoBERTa fine-tuned on SST-2 and SQuADv1.0 are both near zero.
When we apply pruning of fine-tuned BERT, Table 1 shows that we can recover the representation of
some knowledge and the P@1 is improved to 27.1 and 22.5. But some of the knowledge maybe permanently
lost, so the precision is still much lower than BERT without finetuning at 57.6.

References:
[1]. Stress Test Evaluation for Natural Language Inference. COLING 2019.
[2]. Behavior Analysis of NLI models: Uncovering the Influence of Three Factors on Robustness. NAACL 2018.
[3]. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019.

Q2: "huge performance difference between finetuning original BERT vs finetuning pruned BERT" Done
A2: Please refer to the previous answer. Notice in Table 1, "BERT-BASE-FINETUNED-CONLL03 w/ deterministic pruning" means finetune BERT first and then do deterministic pruning, not the other way around.
This also applies to other fine-tune + pruning experiments.

Response to Review #3:
Thanks for your valuable review. Regarding your questions:
Q1: "...the SQuAD dataset is also a subset of LAMA?" Done
A1: In this paper we only used the subset of LAMA exclusively targeting relational commonsense knowledge (line 331-335). It has nothing to do with SQuAD.

Q2: "...results of the pruning will hold even after comparing with fine-tuned models?"
A2: We think you are referring to Table 1, which only includes comparisons with BERT fine-tuned on CONLL03 and SQuAD. We agree that it is conceivable that BERT fine-tuned on commonsense related tasks
in Table 3 might strengthen the representation of commonsense knowledge and improve the

its performance on LAMA. But in reality, that's the case. Our experiments on BERT fine-tuned on SWAG and aNLI shows that the P@1 on LAMA is only slightly better than zero, significantly lower than 12.9. The reason is given in A1 to Review #2. With your permission we can add this discussion into the revised version.

Q3: "What is the value of t in the deterministic pruning?" Done
A3: The pruning threshold value t is 0.5 for all models.

Q4: "unclear how a specific combination of these relations is useful for a task?"
A4: This framework provides a mask for each relation type. One can try to union the masks of a set of relation types to prune the model. Table 2 of Appendix shows the best combination of relation masks
for different PLMs and different tasks.
When all relations are used, the results on the downstream commonsense tasks deteriorate. This is because most tasks use only a few relations. A model masked by all relations behaves toward the original network without pruning.