Exploring and Exploiting Latent Commonsense Knowledge in Pretrained Masked Language Models

Anonymous ACL-IJCNLP 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 reasoning and commonsense knowledge base completion 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., 2018) and RoBERTa (Liu et al., 2019a). These high-capacity 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 sequential transfer learning paradigm has become the de-facto standard on most NLP tasks, we know very little about *what* and *how much* knowledge embedded in PLMs actually contributes to the success. Recent endeavors toward this understanding are a body of works 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. Petroni et al. (2020) first posed the LAMA probe, an English benchmark comprising multiple sets of cloze-style natural language prompts (e.g., Hearing a joke would make people want to [MASK] with the masked object *laugh* and underlying commonsense relation *CausesDesire*.). By 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, albeit without grounded supervision, PLMs capture such relational knowledge at a level competitive to supervised alternatives. Subsequent works further showed that some specific prompts, either through heuristical mining (Jiang et al., 2019) or gradient-guided search (Shin et al., 2020), can better triggering the models to correctly predict the missing object. It reveals that certain prompts might be sub-optimal because PLMs learned target knowledge from substantially different contexts.

Despite the mounting evidence that manifests the existence of relational knowledge in PLMs, in this paper we ask two questions: (i) "Can we disentangle the pretrained general-purpose language representations into relation-specific knowledge representations?" (ii) "Can we exploit such relation-specific knolwedge in downstream knowledge-intensive scenarios?". We approach the first question by starting with inspecting the knowledge probing procedure that previous works have taken for granted. Particularly, we postulate that there is a nuanced ***discrepancy*** between language model pretraining and language model as neural knowledge base. During pretraining, given the large quantity of text corpus, PLMs have the potential to learn a wide variety of relational knowledge spanning facts and commonsense using the MLM objective. However, without being explicitly informed of the type of knowledge that each instance embodies, PLMs cannot effectively build correlation between the specific type of knowledge and structures in the model (e.g., different submodules for each type). Consequently, what PLMs have developed after

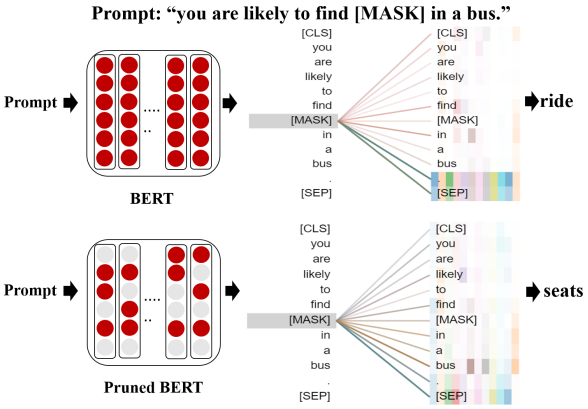

Figure 1: Querying original/pruned BERT-base with prompts of relation *HasA*. The color spectrum indicates 12 different attention heads in the last layer (Vig, 2019).

pretraining is essentially an implicit and entangled knowledge storage mechanism in the shared parameter space. Such mechanism can be considered tailor-made for minimizing the MLM loss in mini-batch mode during pretraining, but at the price of being suboptimal when serving as knowledge base with relation specification.

Based on the above assumption, we make the first attempt to recover the latent parameter structure in PLMs w.r.t. different types of relational knowledge. Specifically, we propose weakly supervised weights pruning, an end-to-end differentiable procedure to search for subnetworks within general-purpose PLMs in which specific relational knowledge is better elicited. With preliminary focus on concept-centric commonsense knowledge, we show 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 while the full-scale model fails. As an attempt to answer the second question, we explore various means of repurposing the pruned models on commonsense reasoning and commonsense knowledge base completion tasks.

The **contributions** of this paper are threefold: (i) We present a novel way of eliciting relational knowledge embosked in PLMs from the perspective of network pruning (Section 2.3). (ii) Grounding on external concept-centric relation schema, we show that the proposed procedure successfully identified sparse subnetworks specializing in miscellaneous commonsense knowledge at a level remarkably better than their full-scale counterparts (Section 3.1). (iii) We showcase several effective exploitations of the subnetworks on commonsense reasoning (Section 3.2) and commonsense knowledge base completion tasks (Section 3.3), gleaning insight of the transformation from language representation to knowledge representation.

We release code and all versions of our pruned PLMs at https://anonymous.4open.science/r/9be5104a-bb0a-489a-beeb-b8814a2be3cb/ to facilitate future research.

2 Methodology

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

2.1 Pretrained Masked Language Models

Pretrained masked language models have proven to be effective at extracting contextualized representations. Formally, given a sequence of tokens $\boldsymbol{w} = [w_1, w_2, ..., w_n]$, where n is the total 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. We denote the original pretrained model LM with unpruned parameter θ as LM_θ in following sections.

2.2 Knowledge Probing with Cloze Prompts

While it is infeasible to probe PLMs with structured query defined by the KB schema and query language, the natural language prompts, such as "*you are likely to find a basement in below your [MASK]*", offer a mean of querying language models that conforms to their interface.

We follow the formulation of (Petroni et al., 2020), where relational knowledge is in the form of triplets $(subj, r, obj)$. Here $subj$ refers to the subject, obj refers to the object, and r indicates their corresponding relation. To query a model LM, 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 place-holders for $subj$ and obj (e.g., "*you are likely to find* $[subj]$ *in* $[obj]$"). The existence of the knowledge in M is gauged by substituting the $[subj]$ place-holder with the surface form of real subject and ask model

M to predict the missing object:

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

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

2.3 Weakly Supervised Weights Pruning

Although it is practically impossible to perfectly recover the latent parameters θ_r corresponding to relation r out of an off-the-shelf model LM_θ, we might still be able to alleviate the noise and redundancy brought by over-parametrization.

Given a pretrained language model LM and the associated set of pretrained parameters $\theta \in R^d$, where d is the dimensionality, we are interested in finding the subnetwork LM_{θ_r} such that LM_{θ_r} is maximally predictive of prompts of relation type 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 an equal-size 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 kept or pruned.

Based on G_r^l, we explore two variants of converting G_r^l into a binary masking matrix M_r^l.

2.3.1 Stochastic Pruning

The first variant is to establish a probabilistic formulation for determining the importance of individual weight. 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 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 trick (Li et al., 2018) to approaximate $m_{i,j}^l$ 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 block 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) LM_{θ_r} is expected to behave like a specialized neural knowledge base. That is, given a prompt requiring knowledge about relation r, LM_{θ_r} should be able to fill in the missing object more accurately than its full-scale counterpart M_θ. 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 LM_{θ_r} that minimizes the following objective:

$$\mathcal{L}_r = -E_{(subj,T_r,obj)\sim D_r}[\log P_{LM_{\theta_r}}(obj|subj, T_r)]$$

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 LM.

In inference, for deterministic pruning, M_r is obtained from G_r according to 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				∼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	∼%	7-12	110M
MPNet-base w/ deterministic pruning	62.7	68.7	71.4	∼50%	7-12	88M
BERT-base-finetuned-CoNLL03 w/o pruning	0.0	0.0	0.0	0%	-	110M
BERT-base-finetuned-CoNLL03 w/ deterministic pruning	27.1	37.7	43.1	∼50%	7-12	88M
BERT-base-finetuned-SQuAD w/o pruning	0.0	0.0	0.0	0%	-	110M
BERT-base-finetuned-SQuAD w/ deterministic pruning	22.5	32.4	37.5	∼50%	7-12	88M

Table 1: Rank-based metrics on LAMA. We show one representative pruning configuration for each type of model and relegate the complete results into Appendix due to space limits.

3 Experiments

In this section, we first expound the resource used for model pruning as well as conducting detailed analysis. Then we experiment on several commonsense-intensive scenarios to seek best practices of using the pruned models.

3.1 Exploring Latent Commonsense Knowledge in PLMs

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

Table 2: Statistics about ConceptNet subset of LAMA.

Dataset We use the ConceptNet (Speer and Havasi, 2012) subset of the LAMA benchmark as weak supervision, which contains facts from the English part of ConceptNet that have 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. And this naturally departs from the conventional supervised experimental train/valid/test setting.

Setup For choices of LM, we consider the 6-layer DistilBERT-base (Sanh, 2019), 12-layer BERT-base, 12-layer RoBERTa-base (Liu et al., 2019a), as well as two fine-tuned models: BERT-base-finetuned-CoNLL03 and BERT-base-finetuned-SQuAD. We also include the more recent powerful MPNet (Song et al., 2020) model. All models are implemented with HuggingFace transformers (Wolf et al., 2019).

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 always set l_t to be the top layer of a given model and set l_b to be within $3 - 4$ for DistilBERT-base, $5 - 9$ for BERT-base, RoBERTa-base, and MPNet-base. The temperature τ is fixed as 0.1. We use Adam optimizer (Kingma, 2014) 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 0.0003 and other hyperparameters remain default.

How more knowledgeable are PLMs than we thought? Table 1 shows the aggregated results. Among all models without pruning, RoBERTa-base achieves the highest P@1 score of 15.4 while DistilBERT gets the lowest 11.4, which is in line with their reported performance when fine-tuned on downstream NLU tasks (Sanh, 2019). Signally, both of the two fine-tuned BERT-base models appear commonsense-oblivious, which demonstrates that fine-tuning on downstream tasks brings aggressive change of weights in order to acommodate varied task-dependent specificities.

Comparing the results for each pair of original and pruned models, we consistently observe a surprisingly significant increase, especially for deterministically pruned ones. Furthermore, the results of fine-tuned models shows that the NER task requires relatively more commonsense knowledge than the extractive question answering on SQuAD (Rajpurkar et al., 2018). This large perfor-

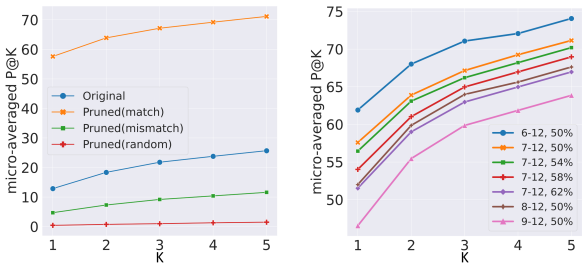

Figure 2: Ablation on the pruning masks (left) and effect of initial sparsity and pruned layers (right).

mance gap provides unique new evidence of sparse latent relational knowledge structures in PLMs , which are embosked 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 implies 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 in the remainer of this paper.

Are the subnetworks non-trivial to find? With curiosity on the emergence of such knowledgeable subnetworks, we then examine their non-trivialities by ablating instantiation of 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 (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 implies that the latent structure for different types of commonsense knowledge exhibits remarkably distinct geometry.

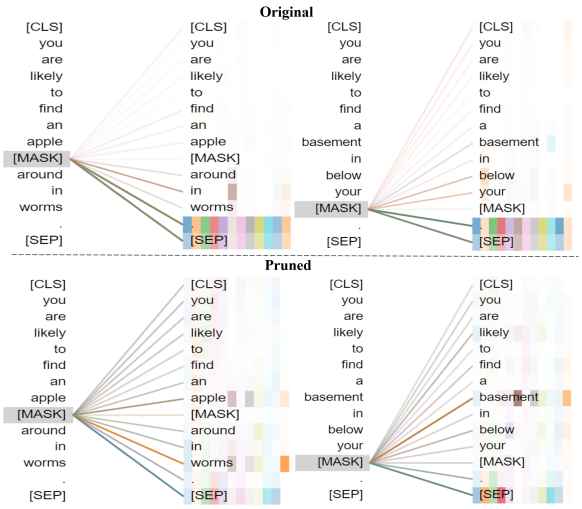

Figure 3: Attention weight visualization. *AtLocation* is required for prompt in the left column and *PartOf* is required for prompt in the right column.

Effect factors 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 3.2 and Section 3.3).

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 and visualize the attention patterns in the last layer. Specifically, 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 ConcetpNet relations in Figure 3. We observe that while the original pretrained model tends to attend to special tokens like period and [SEP], the subnetwork we found successfully grasp the relevant concepts (i.e., apple, worms, and basement) in the prompt hence producing the right object. We also use t-SNE (van der Maaten and Hinton, 2008) to visualize the last layer's representation of [CLS] for

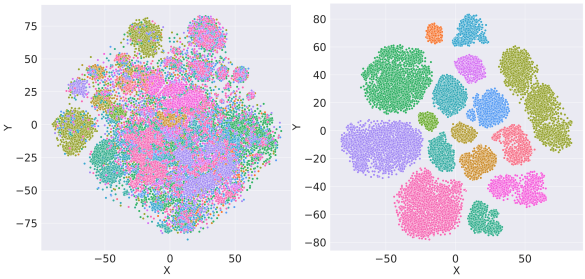

Figure 4: t-SNE visualization of [CLS] token. Each color represents 1 out of 16 commonsense relations.

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.

3.2 Commonsense Reasoning (CSR)

After identifying sparse subnetworks within PLMs that specialize in different commonsense knowledge, we now evaluate their generalization ability in the context of commonsense reasoning.

Fine-tuning 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., 2019) 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 masks for each task as well as training details can be found in Appendix.

The results in Table 3 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 low-resource regime on COPA dataset. Figure 5 shows that the pruned BERT exhibits a notable advantage when training data is extremely scarce. As more training data are 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

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 3: Fine-tuning results of BERT-base on CSR.

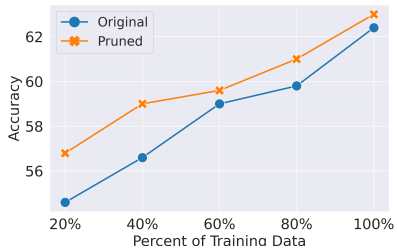

Figure 5: BERT-base results on COPA.

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 Acceptability (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). Instance of the form *[CLS] premise [SEP] hypothesis$_i$ [SEP]* 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 4. We put the best combination for each model on each task in Appendix due to space constraint.

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 could be 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.

3.3 Commonsense Knowledge Base Completion (CKBC)

While most prior work on CKBC (Li et al., 2016; Saito et al., 2018) are performed in a supervised setting, here we explore the effectiveness of using

Model	COPA (Tra.)	COPA (Dev.)	CSQA	CA	WSC	SM	ARCT1	ARCT2	Avg.
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)	**63.8**	**69.0**	**40.4**	**93.4**	**62.2**	**74.4**	**53.2**	**55.1**	**63.9**
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 4: Zero-shot results on commonsense reasoning tasks.

Model	Validation set				Test set			
	MRR	Hits@1	Hits@2	Hits@3	MRR	Hits@1	Hits@2	Hits@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., 2017)	19.9	12.5	17.6	20.0	23.9	15.5	20.9	26.0
TuckER (Balazevic et al., 2019)	18.3	11.9	15.8	19.8	23.6	16.0	22.4	26.0
ConvTransE (Shang et al., 2019)	21.8	15.2	19.8	23.3	26.0	17.6	23.9	28.5
SACN (Shang et al., 2019)	23.2	15.2	21.8	25.2	27.2	17.4	25.1	31.0
InteractE (Vashishth et al., 2019)	21.5	13.2	19.3	23.2	**27.8**	17.9	**26.6**	**32.0**
Unsupervised								
DistilBERT-base	10.4	4.3	8.4	12.3	11.9	6.7	9.9	12.6
BERT-base	14.3	8.8	11.8	15.7	14.9	8.4	13.4	17.3
RoBERTa-base	9.3	5.2	7.3	8.9	10.2	5.4	8.6	11.8
MPNet-base	13.0	8.2	11.4	13.2	11.9	6.9	10.4	12.5
DistilBERT-base (pruned)	**26.2**	**17.5**	**25.7**	28.5	27.2	**18.0**	25.6	30.1
BERT-base (pruned)	25.7	16.2	24.6	**30.0**	26.9	**18.0**	25.2	30.3
RoBERTa-base (pruned)	9.7	5.6	8.2	9.5	10.7	6.9	8.8	12.1
MPNet-base (pruned)	23.7	14.5	22.2	27.4	23.4	14.0	22.6	26.9

Table 5: Link prediction results. Best results are noted with **bold** font and second best with underline.

PLMs as well as their deterministically pruned versions to accomplish the task in an unsupervised manner. Specifically, we use the ConceptNet-100K benchmark provided by Li et al. (2016).

Link prediction We first formulate CKBC as a link prediction task and compare both original and pruned PLMs (i.e., LM_{θ_r} is queried to predict missing link for instance with relation r.) against several supervised high performing KB completion models. To allow a fair comparison, we manually create a subset of ConceptNet-100K by keeping only triples within the 16 relations included in LAMA while having single-token subject/object. Each relation is associated with a sentence template (provided in Appendix) (Kwon et al., 2019) of which the wording is diverse from those in LAMA. Note that the sentence templates used by PLMs are surely suboptimal for certain relations, but prompt optimization is out of scope of this paper. The resulting dataset contains 17,891 training instances, 463 development instances and 594 test instances.

Table 5 shows the results. Most of the supervised models outperform full-scale PLMs by a large margin, which implies the inefficacy of directly using PLMs to perform link prediction. However, after pruned by mask of the designated relation,

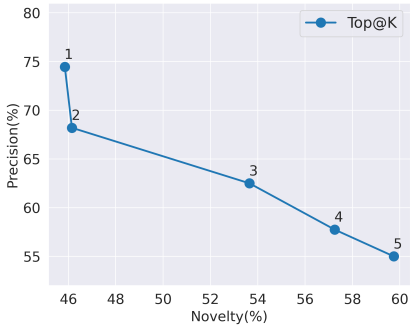

Figure 6: Precision-novelty curve with varied K.

PLMs are able to 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 likely due to its larger vocabulary size compared to WordPiece (50265 vs 30522) and less lexicon overlap (53% vs 59%) with the dataset.

Triple extraction We then investigae 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

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
RoBERT-base (pruned)	
MPNet-base	76.5
MPNet-base (pruned)	**78.0**

Table 6: Triple classification on ConceptNet-100K.

we use top-K predictions from pruned DistilBERT as candidate objects for a given $(subj, r)$. Three human annotators are asked to first determine the correctness of each candidate object and further determine its 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 6 shows the change of precision-novelty 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 prediction but is more likely to extract unseen knowledge. For the purpoe of knowledge enrichment, one might choose a large K to ensure a desirable recall. We list the obtained novel triples in the Appendix due to space limits.

Triple classification Following Feldman et al. (2020), 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 single relation might not be robust for PMI estimation and generally performed inferior to the intact model. In analogy with 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 6, we observe that there exists one or multiple knowledge combinations delivering F1 score higher than the intact models.

4 Related Work

Ever since the emergence of large-scale pretrained language models, many works have focused on the understanding of internal contextual representations produced by such models. Most prior works (Shi et al., 2016; Belinkov et al., 2017) pay 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, a body of studies (Peters et al., 2018; Goldberg, 2019; Tenney et al., 2019) have shown the existence of linguistic knowledge (e.g., syntactic and semantic) 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 into 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 over-parametrized PLMs with specialization on chiseled commonsense knowledge rather than good initialization for specific downstream tasks.

5 Conclusion

We apply network pruning, a novel perspective to explore the latent relational knowledge embosked in PLMs. With 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 fine-tuning starting points, robust zero-shot reasoners, and auspicious neural knowledge bases. Our work raises 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 M. Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. *CoRR*, abs/1901.09590.

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, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *CoRR*, abs/1908.05739.

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.

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.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2017. Convolutional 2d knowledge graph embeddings. *CoRR*, abs/1707.01476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2020. Commonsense knowledge mining from pretrained models. *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 1173–1178.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*.

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*, volume 29, pages 4107–4115. Curran Associates, Inc.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know? *CoRR*, abs/1911.12543.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

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

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, Lim Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3:1445–1455.

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.

Tao Lin, Sebastian U. Stich, Luis Barba, Daniil Dmitriev, and Martin Jaggi. 2020. Dynamic model pruning with feedback. In *International Conference on Learning Representations*.

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 *ICLR*.

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

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *CoRR*, abs/1907.07355.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

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.

Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense knowledge base completion and generation. *CoNLL 2018 - 22nd Conference on Computational Natural Language Learning, Proceedings*, (CoNLL):141–150.

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

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. *AAAI*.

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, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts.

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. *CoRR*, abs/1905.06316.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Nilesh Agrawal, and Partha P. Talukdar. 2019. Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions. *CoRR*, abs/1911.00219.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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

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, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

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? *CoRR*, abs/1905.07830.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an Efficient Alternative to Finetuning for Pretrained Language Models. (i).

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