Specializing Pre-trained Language Models for Better Relational Reasoning via Network Pruning

Siyu Ren **Kenny Q. Zhu***
Shanghai Jiao Tong University
Shanghai, China
roy0702@sjtu.edu.cn, kzhu@cs.sjtu.edu.cn

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 and comprehensive study concerning specializing PLMs into relational models 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 while being more generalizable than original PLMs in scenarios requiring knowledge of single or multiple commonsense relations.

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). They are first pretrained on vast amount of unlabeled text corpora using masked language modeling (MLM) objective and then fine-tuned on task-specific data, offering a surge of improvements on a wealth of NLP tasks. However, we know very little about *what* and *how much* knowledge embedded in PLMs actually contributes to the success. Notable endeavors (Peters et al., 2018; Goldberg, 2019; Tenney et al., 2019) toward this understanding focus on probing linguistic knowledge therein. They demonstrated that pretraining did impart useful linguistic abstraction about syntax and semantics into PLMs.

More recently, several works presented intriguing results examining relational knowledge within PLMs. Relational knowledge (Speer and Havasi, 2012; Vrandečić and Krötzsch, 2014) is typically defined as describing the abstractive 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.
Prompt: you are likely to find _ in a bus.

By substituting _ with a special [MASK] token and reusing the MLM head, prompt-based relational knowledge probing provides an estimated lower bound of what PLMs know without training an additional layer which was used in the previous linguistic probes. 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 is represented internally. In light of this, we raise the core question in this paper: *Given the general language representation space modeled by a PLM, can we extract its latent representation subspaces for different relations and specialize the PLM into relation-specific knowledge models?* These subspaces exclusively represent knowledge inherited from different subset of MLM data, expressing different relations between masked word and remaining context, thus can potentially benefit applications where knowledge of certain relations are explicitly required.

We study this question by first drawing inspiration from recent findings (Saunshi et al., 2020; Lee et al., 2020; Zhang and Hashimoto, 2021): *the more MLM pretraining simulates downstream task, the more successful the transfer will be.* For example, filling *like* or *hate* into a cloze like "*I [MASK] this film, it's great.*" provide a clear way in which

* The corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2195 - 2207

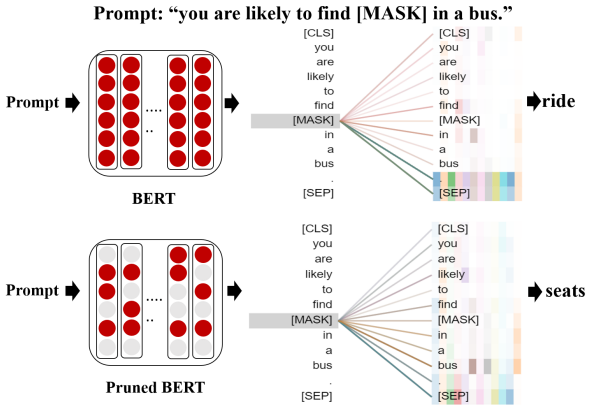

Figure 1: Querying original/pruned BERT-BASE with prompts of relation *HasA*. The color spectrum indicates the 12 attention heads in the last layer (Vig, 2019).

the model can implicitly learn to perform sentiment classification. Similarly, we hypothesize that training on MLM data expressing certain relation r between masked word and remaining context would lead to effective transfer on knowledge probing that targets relation r.

By exploiting this correlation conversely, we offer a new way for extracting representation subspaces responsible for different relational knowledge based on their transfer performance on knowledge probing task. Instead of introducing additional parametric transformation upon the original space, we restrict these subspaces to have correspondence with subnetworks of PLMs and propose an end-to-end differentiable weights pruning framework. Our experiments show that it is possible to find subnetworks capable of representing grounded commonsense relations at non-trivial sparsity while also exhibiting evident disentanglement. 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 original BERT fails. We then examine the knowledge transfer ability of these subnetworks in scenarios requiring knowledge of single or multiple commonsense relations for reasoning. Experiment on commonsense knowledge base completion show that the identified subnetworks even outperform strong supervised knowledge base completion methods. These subnetworks also outperform the original PLMs in both many-shot and zero-shot commonsense question answering tasks, when combined properly.

Code and all pruned subnetworks are open-sourced at https://github.com/DRSY/LAMP.

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

Given a sequence of tokens $\boldsymbol{w} = [w_1, w_2, ..., w_n]$ with length n, the model outputs a sequence of hidden states $\boldsymbol{h} = [h_1, h_2, ..., h_n]$ corresponding to each token. In standard MLM pretraining, h_i is fed into a 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 all other unmasked tokens. We denote the pretrained masked language model $\mathcal{LM}$ with parameter θ as $\mathcal{LM}_\theta$ in the 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 means 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, and two of which are place-holders 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 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 $\mathcal{LM}_\theta$ carries the knowledge if $\hat{obj} = obj$.

2.3 Extracting Representation Subspaces by Weights Pruning

Extracting representation subspaces for different roles/functionalities has been explored by prior works, such as attaining disentangled subspaces of style and semantic content in text style transfer task (John et al., 2019). The typical approach is to apply parametric transformation function upon

the original space and optimize through end-to-end downstream fine-tuning. However, it induces additional parameters and cannot faithfully reveal the relational knowledge originally present in PLMs since such knowledge can be stored in the newly introduced parameters outside of PLMs. We circumvent this issue by focusing on the representation spaces modeled by subnetworks of $\mathcal{LM}_\theta$. A subnetwork $\mathcal{LM}_{\theta_r}$ of relation r is obtained by setting certain dimensions of θ to zero.

The next step is to identify $\mathcal{LM}_{\theta_r}$ such that the representation space it corresponds to inherits its knowledge from the MLM data expressing relation r. Based on the previous evidence (Zhang and Hashimoto, 2021) that shows positive correlation between downstream performance and task similarity with MLM data, we propose to estimate representation space for relation r by searching for the $\mathcal{LM}_{\theta_r}$ that is the most predictive of the prompts expressing relation r. Specifically, for each weight matrix W^l from the set of all weight matrices $\boldsymbol{W}^l$ in the l-th transformer layer, we assign a 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. 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 the 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 a 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 by Li et al. (2018) to approximate $m_{i,j}^l$ with another differentiable 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).

Consequently, 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 denote 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 a 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 (Hubara et al., 2016; Zhao et al., 2020) 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 elucidate the loss function $\mathcal{L}_r$ w.r.t. relation r in the next section.

2.3.3 Training and Inference

The resultant 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 $\mathcal{LM}_\theta$. Hence 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:

$$\mathcal{L}_r = -\mathbb{E}_{(subj,T_r,obj)\in D_r}[\log P_{\mathcal{LM}_{\theta_r}}(obj|subj,T_r)] \tag{7}$$

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. Finally, we obtain a set of trained $\{\boldsymbol{G}_r\}_{r\in\mathcal{R}}$ for the designated pretrained model $\mathcal{LM}_\theta$.

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.

3 Experiments

We present our pruning setup and detailed analysis in Section 3.1. Then we compare the knowledge transfer ability of pruned subnetworks against original PLMs in scenarios requiring knowledge of single or multiple relations for reasoning in Section 3.2 and Section 3.3 respectively.

3.1 Pruning & Analysis

Data Split	# Relations	# Prompts
Train	16	20,841
Validation	16	5,955
Test	16	2,978

Table 1: Statistics of C-LAMA.

Dataset. We use the ConceptNet (Speer and Havasi, 2012) subset of LAMA benchmark for both pruning and evaluation, denoted as C-LAMA. C-LAMA contains commonsense facts from the English part of ConceptNet that has single-token objects covering 16 relations. These facts are extracted from Open Mind Common Sense (OMCS). Since C-LAMA has no official data splits, we construct the train/validation/test splits with a ratio of 7:2:1. Detailed statistics are listed in Table 1.

Models. We experiment with the DistilBERT-base (Sanh, 2019), BERT-base, RoBERTa-base (Liu et al., 2019a), and the more recent MPNet-base (Song et al., 2020) as the choices of $\mathcal{LM}$. After pruning, each $\mathcal{LM}$ will have 16 pruned subnetworks corresponding to the 16 commonsense relations. As a straightforward comparison, for each $\mathcal{LM}$ we also obtain 16 fine-tuned models corresponding to the same 16 relations. Precision P@K averaged across all relations is used to evaluate the prompt filling performance.

Setup. The prior distribution $\phi(\cdot)$ is a Gaussian $\mathcal{N}(\mu, 1)$ where μ is the mean that controls the initial sparsity of the pruned model (e.g., $\mu = 0$ indicates 50% initial sparsity). We set l_t to be the top layer of a given model and choose l_b from $\{3, 4\}$ for DistilBERT, $\{6, 7, 8, 9\}$ 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}. We run all experiments with three different random seeds and report the averaged results. All experiments are conducted on a GTX 1080 Ti GPU with 11GB RAM.

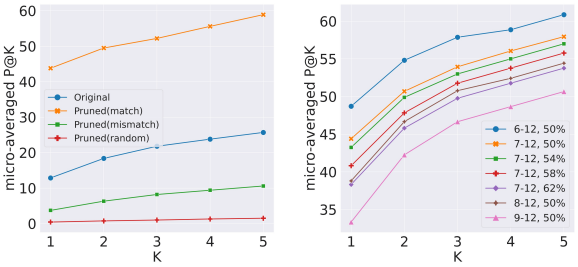

Figure 2: Ablation on the pruning masks (left) and effect of initial sparsity and pruned layers (right).

Factors impacting performance. To investigate how μ and l_b affect the performance, we perform a preliminary experiment by applying deterministic pruning on BERT-base with l_b in $\{6, 7, 8, 9\}$ and the initial sparsity in $\{50\%, 54\%, 58\%, 62\%\}$. Figure 2 (right) shows that (i) increasing the number of pruned layers helps distill more knowledge; and (ii) larger initial sparsity is more likely to prune away weights that are important to certain knowledge and cannot be recovered in the later training process. In general, we find an initial sparsity around 50% yields a decent performance both in probing and downstream applications. We adopt this setting in the rest of this paper unless otherwise stated.

How specialized are these subnetworks? Ideally, specialized subnetworks are expected to perform poorly on relations other than their associated ones. We verify this by instantiating the pruning mask upon BERT-base with a set of mismatched masks. Specifically, we corrupt the correspondence of relation between masks and prompts by shuffling the 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. As indicated by the green curve in the left part of Figure 2, if we apply the mismatched masks from other relations, the P@1 score significantly drops to 3.8 from 43.8, a performance even worse than the original model. It suggests that the representation spaces for different commonsense relations modeled by these subnetworks are highly disentangled.

Model	P@1 (%)	P@2 (%)	P@3 (%)	Sparsity	l_b-l_t	# Param.
DistilBERT-base	11.4	16.6	19.9	0%	-	66M
DistilBERT-base w/ fine-tuning	27.9	36.3	41.4	0%	-	66M
DistilBERT-base w/ stochastic pruning	14.8	21.5	26.3	∼30%	4-6	66M
DistilBERT-base w/ deterministic pruning	**34.0**	**41.8**	**46.0**	∼50%	4-6	56M
BERT-base	12.9	18.4	21.8	0%	-	110M
BERT-base w/ fine-tuning	29.2	37.4	41.3	0%	-	110M
BERT-base w/ stochastic pruning	17.2	25.1	29.6	∼30%	7-12	110M
BERT-base w/ deterministic pruning	**43.8**	**49.5**	**52.2**	∼50%	7-12	88M
RoBERTa-base	15.4	21.2	24.6	0%	-	125M
RoBERTa-base w fine-tuning	11.7	14.4	16.4	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	14.8	20.7	24.0	0%	-	110M
MPNet-base w/ fine-tuning	23.8	30.9	36.3	0%	-	110M
MPNet-base w/ stochastic pruning	19.8	27.9	33.2	∼30%	7-12	110M
MPNet-base w/ deterministic pruning	**47.9**	**52.8**	**55.6**	∼50%	7-12	88M

Table 2: Relational knowledge probing results on C-LAMA test set.

Non-triviality of subnetworks. We also examine the non-triviality of subnetworks by initializing the masks with a Bernoulli distribution $B(0.5)$ and averaging the results from 5 different random seeds. If we apply such random masks with sparsity comparable to the learned ones, the P@1 drops drastically to 0.4 (red curve in the left part of Figure 2). This notable difference proves that the effective subnetworks cannot be trivially identified through random weights pruning.

Test set results. Table 10 summarizes the test set results. Among all original PLMs, 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 coincides with previous finding (Zhang and Hashimoto, 2021) that the uniform masking adopted by PLMs is biased towards extracting statistical dependencies. Comparing the results for each pair of original and pruned models, we consistently observe a significant increase in pruned models, especially for deterministically pruned ones (27.4% on average). The pruned models also surpass their fine-tuned counterparts, which is likely due to that fine-tuning makes aggressive updates to parameters and overfits to the training set by memorizing spurious features.

To sum up, the substantial performance gains provide new evidences for the existence of sparse latent relational knowledge structures in PLMs. These structures are previously weakened by other pretrained weights *reserved* for more general-purpose use and are *exposed* by the proposed pruning method. It is worth noting that deterministic pruning excels by a big margin compared to stochastic one. It implies that successful extraction of relation-specific representation space relies on ignoring the information in the input that is irrelavant to the relation. Therefore, we focus our analysis on deterministically pruned PLMs and denote them by pruned in the rest of this paper.

Visualization of attention 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.

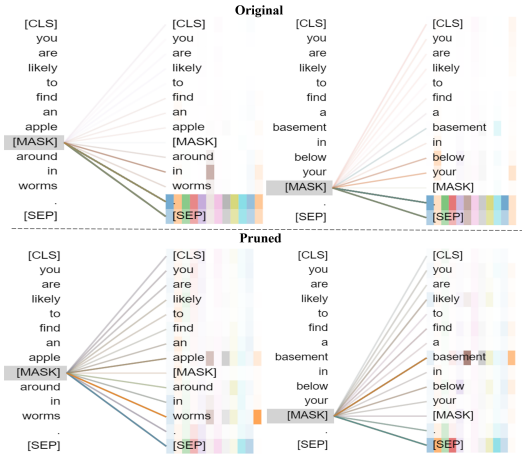

Figure 3: Attention weight visualization. *AtLocation*/*PartOf* is required in the left/right column.

Specifically, we focus on the attention weights between [MASK] and other tokens in the prompt. A first glance of change of attention pattern was given earlier in Figure 1, and now we show more examples of other ConcetpNet relations in Figure 3.

Method	Development Set				Test Set			
	MRR (%)	P@1 (%)	P@2 (%)	P@3 (%)	MRR (%)	P@1 (%)	P@2 (%)	P@3 (%)
Supervised KB completion models								
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 (Balažević et al., 2019)	17.3	10.9	14.8	18.8	21.6	14.0	20.4	24.0
SACN (Shang et al., 2019)	21.2	13.2	19.8	23.2	**24.2**	14.4	22.1	**28.0**
InteractE (Vashishth et al., 2020)	19.8	11.2	17.3	21.2	23.3	**14.9**	21.9	26.5
Unsupervised PLMs								
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: One-hop link prediction results. Best results are marked with **bold** font and second best with underline.

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 correct object. We also use t-SNE (van der Maaten and Hinton, 2008) to visualize the last 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 mixed together. In contrast, the pruned subnetwork can extract meaningful and disentangled representations for different commonsense relations.

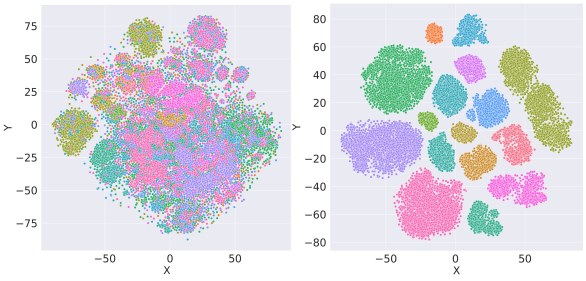

Figure 4: t-SNE visualization of [CLS] representation from original (left) and pruned (right) BERT-base.

3.2 Single-Relation Scenario

In this section, we compare the transfer ability of pruned subnetworks against original PLMs in a scenario that explicitly requires knowledge of single commonsense relation, i.e., commonsense knowledge base completion (CKBC), which aims at populating a CKB with valid relational triples. We use the ConceptNet-100K benchmark provided by Li et al. (2016). To ensure a fair evaluation, we manually create a subset of ConceptNet-100K consisting of triples with single-token subject/object. We also ensure that its dev/test set has **no overlap** with C-LAMA. Each relation is associated with a sentence template (provided in Appendix B) of which the wording is distinct from those in C-LAMA. The resulting dataset contains 17,891 training instances, 349 development instances, and 446 test instances.

One-hop link prediction. We first formulate CKBC as a link prediction task, i.e., predicting the missing object given the subject and relation. We regard the pruned subnetworks and the original PLMs as the unsupervised off-the-shelf neural knowledge base and include the results of several strong KB completion methods for reference.

Table 3 shows 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 the state-of-the-art supervised models, which shows the potential of language models as neural knowledge base that is underestimated by previous studies. Surprisingly, 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.

Two-hop extrapolation. Sometimes, a pair of commonsense relations can be combined to derive a new relational knowledge triple. For example, the two triples $\langle s1, IsA, o1\rangle$ and $\langle s2, HasProperty, o2\rangle$, where o1 equals to s2, can be combined into a new test triple $\langle s1, HasProperty, o2\rangle$. Based on this rule, we construct 5,151 new test triples (absent in

ConceptNet-100K) with *HasProperty* relation, which allows us to compare the two-hop extrapolation ability of pruned subnetworks with original PLMs. Table 4 shows that pruned subnetworks exhibit significantly better ability of extrapolating from known relational knowledge to novel knowledge, with an average improvement of 23.2/24.8/24.0 in terms of P@1/P@2/P@3.

Model	P@1	P@2	P@3
DistilBERT-base	10.6	17.8	23.0
DistilBERT-base (pruned)	**28.2**	**36.5**	**44.6**
BERT-base	11.9	18.5	21.9
BERT-base (pruned)	**42.9**	**52.1**	**58.4**
RoBERTa-base	16.8	24.1	28.0
RoBERTa-base (pruned)	**21.5**	**28.7**	**31.0**
MPNet-base	16.6	24.8	29.3
MPNet-base (pruned)	**56.2**	**64.2**	**67.7**

Table 4: Two-hop extrapolation results (%).

Model	< linen, IsA, ? >	< sing, Causes, ? >
DistilBERT	surname commodity profession	death disease trouble
DistilBERT(pruned)	cloth fabric garment	happiness joy peace

Table 5: Top-3 predictions of two triples sampled from CKBC dev set. The predictions are ranked by probability in descending order with ground-truth marked in green and other plausible answers underlined.

Case Study. We present a case study of link prediction examples from both pre-trained and pruned DistilBERT in Table 5. In both examples, the pre-trained DistilBERT makes completely unrelated predictions (e.g.,*surname*, *profession*, and *death*) and only two predictions (i.e., *commodity* and *trouble*) can be hardly considered as plausible. In contrast, the model pruned for IsA/Causes are specialized in accurately representing these relations and can even produce reasonable answers in addition to the ground-truth.

3.3 Multi-relation Scenario

In this section, we compare the transfer ability of pruned subnetworks against original PLMs in a scenario that implicitly requires knowledge of multiple commonsense relations, i.e., commonsense question answering (CQA).

Stand-alone fine-tuning. We conduct stand-alone fine-tuning using BERT/DistilBERT and their pruned subnetworks on 6 widely adopted CQA tasks: RTE (Dagan et al., 2009), COPA (Roemmele et al., 2011), CommonsenseQA (Talmor et al.), SWAG (Zellers et al., 2018), aNLI (Bhagavatula et al., 2020) and CosmosQA (Huang et al., 2019). For each task, we identify the commonsense knowledge that might be required with a simple yet effective heuristic: we obtain the five most frequent relations measured by how many times the subject and object holding certain relation in ConceptNet appear in the context or the answer. Then we take the union of masks for each relation and apply the resultant mask to BERT/DistilBERT as the initialization for fine-tuning. Intuitively, such union operation would preserve all relational knowledge of interest. We repeat training three times with different random seeds for each task using hyperparameters suggested in the original papers. The detail of mask combination for each task is in Appendix B.

Task	BERT	DistilBERT
RTE	69.2±2.3⇒**69.8**±2.0	61.2±1.2⇒**62.1**±1.2
COPA	62.4±5.0⇒**63.1**±4.7	54.0±2.0⇒**56.0**±2.0
CommonsenseQA	53.1±0.6⇒**54.1**±0.7	47.9±0.7⇒**48.6**±0.7
SWAG	73.9±0.3⇒**74.2**±0.1	70.1±0.3⇒**70.4**±0.1
aNLI	63.7±0.4⇒**64.0**±0.4	60.1±0.4⇒**60.4**±0.4
CosmosQA	61.3±1.0⇒**61.8**±0.2	56.4±0.8⇒**57.2**±0.4

Table 6: Stand-alone fine-tuning accuracy (original ⇒ pruned) of BERT and DistilBERT.

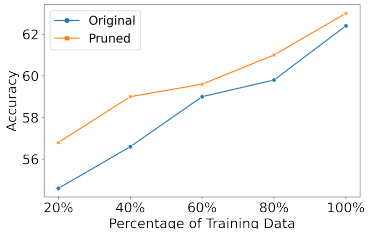

Figure 5: Results on COPA with varying portion of data.

Table 8 shows that, when initialized with proper subnetworks, the model can better transfer to commonsense question answering tasks via more relevant *prior* knowledge. We further analyze the change of performance under the low-resource regime. Figure 5 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 is reduced but still significant.

Integrated fine-tuning. We also integrate the pruned subnetworks to QA-GNN (Yasunaga et al.,

Model	COPA-Tra.	COPA-Val.	CSQA	CA	WSC	SM	ARCT1	ARCT2	Avg.	Δ
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**	+3.7
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**	+3.3
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**	+3.9
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**	+3.2

Table 7: Zero-shot accuracy (%) for commonsense question answering. Better results of each pair is in **bold**.

2021), a state-of-the-art hybrid question answering system in which a PLM and GNN are employed for joint reasoning over text and knowledge graph. We follow their official implementation with the only modification on the PLM part. With the same set of masks as in the stand-alone fine-tuning on CommonsenseQA, the pruned BERT achieves an accuracy of 60.9% versus 60.1% of original model, suggesting a generally stronger knowledge transfer ability not only in stand-alone but also in the integrated settings.

Zero-shot learning. We then assess the ability of pruned subnetworks to perform zero-shot commonsense reasoning, a setting where the knowledge relied on to complete the task is solely determined by the model itself. We focus on the following 8 multiple-choice CQA datasets: training set of COPA (COPA-Tra.), validation set of COPA (COPA-Val.), CommonsneseQA, Conjunction Acceptability (CA) (Zhou et al., 2020), 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 $\{[CLS]\ context\ [SEP]\ choice_i\ [SEP]\}_{i=1}^{N}$, where N is the number of choices. We compute the plausibility score of each choice using the MLM head. The choice of the maximum plausibility score is chosen as the answer.

Since multiple types of knowledge are typically required for effective commonsense reasoning, We employ the same heuristic used in many-shot learning setting for determining the set of the most important relations for each task as well as the same mask union operation to obtain parameter initialization. It can be observed from Table 7 that the pruned models can actually surpass their full-size counterparts in all tasks considered in our experiments. Our explanation is that knowledge irrelevant to the specific task in the original PLMs hurt the in-domain zero-shot reasoning capability. It also demonstrates that the most important relational knowledge vary from task to task.

4 Related Work

Since the emergence of large pre-trained language models, much work has focused on understanding their internal contextual representations. Most prior work (Shi et al., 2016; Belinkov et al., 2017) pays 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 (Goldberg, 2019; Tenney et al., 2019) have shown the existence of linguistic knowledge (e.g., syntax) in various but generally lower layers of pre-trained transformers.

To shed more light on how PLMs memorize abstract knowledge rather than 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 latent relational knowledge to the surface and unleashing more potential for better relation reasoning.

Another relevant line of research is network pruning (Liu et al., 2019b; Lin et al., 2020) and 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 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 that specializes on commonsense knowledge tailored for downstream tasks rather than focusing on good global initialization, and achieve good results.

5 Conclusion

This study investigated specializing PLMs for better relational reasoning via network pruning. In the pilot experiment we find evidence of latent sparse subnetworks capable of representing grounded commonsense relations in various PLMs. Further experiments revealed that such subnetworks possess stronger relational reasoning capability than original PLMs. Our work provides a new vantage point about the internal mechanism as well as practical utilization of relational knowledge in PLMs, opening up avenues to better understanding and adapting pretrained language representations.

Acknowledgement

This research is partially supported by NSFC Grant No. 91646205, and SJTU-CMBCC Joint Research Scheme.

References

Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of EMNLP*, pages 5185–5194.

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.

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*.

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 *Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020*.

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. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of AAAI 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.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019*.

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*, pages 1930–1940.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of EMNLP*, pages 2391–2401.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks. In *Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016*, 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.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

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

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.

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, 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*.

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*.

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.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of EMNLP*, pages 1499–1509, Brussels, Belgium.

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*, pages 2463–2473.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of EMNLP*, pages 3208–3229.

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*.

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, Honolulu, Hawaii, January 27 - February 1, 2019*, pages 3060–3067.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of EMNLP*, pages 1526–1534, Austin, Texas.

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 EMNLP*, pages 4222–4235.

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. 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.

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, May 6-9, 2019*.

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.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Nilesh Agrawal, and Partha Talukdar. 2020. Interacte: Improving convolution-based knowledge

graph embeddings by increasing feature interactions. In *Proceedings of AAAI2020*, pages 3009–3016. AAAI Press.

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.

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, May 7-9, 2015, Conference Track Proceedings*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP*, pages 93–104, Brussels, Belgium.

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 EMNLP*, pages 2226–2241.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, February 7-12, 2020*, pages 9733–9740. AAAI Press.

A Derivation for Stochastic Pruning

To re-parametrize the discrete binary Bernoulli variable $m^l_{i,j} \sim B(\sigma(g^l_{i,j}))$, denote the approximate differentiable variable as $\tilde{m}^l_{i,j} = \sigma(\frac{g^l_{i,j}+\log U-\log(1-U)}{\tau})$ where τ is a real-valued temperature value, we have the following derivation holds for arbitrary $\epsilon \in (0, 0.5)$:

$$P(m^l_{i,j}=1) - P(\tilde{m}^l_{i,j} \geq 1-\epsilon) \leq (\frac{\tau}{4}) \log \frac{1}{\epsilon}$$

Specifically, when temperature τ approaches 0, $\tilde{m}^l_{i,j} = m^l_{i,j}$.

Lemma 1: $\sigma^{-1}(x) = \log \frac{x}{1-x}$.

Lemma 2: $\frac{\sigma(x)-\sigma(y)}{x-y} \leq \frac{1}{4}$.

Proof:

$$\begin{aligned}
&P(\tilde{m}^l_{i,j} \geq 1-\epsilon) \\
=&P(\sigma(\frac{g^l_{i,j}+\log U-\log(1-U)}{\tau}) \geq 1-\epsilon) \\
=&P(\frac{g^l_{i,j}+\log U-\log(1-U)}{\tau} \geq \log(\frac{1}{\epsilon}-1)) \\
=&P(g^l_{i,j} - \tau\log(\frac{1}{\epsilon}-1) \geq \log(\frac{1}{U}-1)) \\
=&P(e^{g^l_{i,j}-\tau\log(\frac{1}{\epsilon}-1)} \geq \frac{1}{U}-1) \\
=&P(U \geq \frac{1}{1+e^{g^l_{i,j}-\tau\log(\frac{1}{\epsilon}-1)}}) \\
=&\sigma(g^l_{i,j} - \tau\log(\frac{1}{\epsilon}-1))
\end{aligned}$$

Then:

$$\begin{aligned}
&P(m^l_{i,j}=1) - P(\tilde{m}^l_{i,j} \geq 1-\epsilon) \\
=&\sigma(g^l_{i,j}) - \sigma(g^l_{i,j} - \tau\log\frac{1}{\epsilon}-1) \\
\leq&\frac{\tau}{4}\log(\frac{1}{\epsilon}-1) \\
\leq&\frac{\tau}{4}\log\frac{1}{\epsilon}
\end{aligned}$$

The process for deriving $P(m^l_{i,j}=0) - P(\tilde{m}^l_{i,j} \leq \epsilon) \leq (\frac{\tau}{4})\log\frac{1}{\epsilon}$ can be analogously obtained. □

B Implementaiton Details

B.1 Templates

The templates we used in single-relation scenario for different commonsense relations are defined as follows:

AtLocation: Something you find at $\langle obj \rangle$ is $\langle subj \rangle$.
CapableOf: $\langle subj \rangle$ can $\langle obj \rangle$.
Causes: $\langle subj \rangle$ causes $\langle obj \rangle$.
CausesDesire: $\langle subj \rangle$ would make you want to $\langle obj \rangle$.
Desires: $\langle subj \rangle$ wants to $\langle obj \rangle$.
HasPrerequisite: $\langle subj \rangle$ requires $\langle obj \rangle$.
HasProperty: $\langle subj \rangle$ can be $\langle obj \rangle$.
HasSubevent: when $\langle subj \rangle$, $\langle obj \rangle$.
HasA: $\langle subj \rangle$ contains $\langle obj \rangle$.
IsA: $\langle subj \rangle$ is a $\langle obj \rangle$.
MadeOf: $\langle subj \rangle$ can be made of $\langle obj \rangle$.
MotivatedByGoal: you would $\langle subj \rangle$ because $\langle obj \rangle$.
NotDesires: $\langle subj \rangle$ does not want $\langle obj \rangle$.
PartOf: $\langle subj \rangle$ is part of $\langle obj \rangle$.
ReceivesAction: $\langle subj \rangle$ can be $\langle obj \rangle$.
UsedFor: $\langle subj \rangle$ may be used for $\langle obj \rangle$.

When performing CKBC task, we first fetch the template based on the relation of the triple to be complete and fill in the $\langle subj \rangle$ and let the model predict the missing $\langle obj \rangle$. Concretely, the $\langle obj \rangle$ placeholder is replaced by the mask token corresponding to different pre-trained language models.

B.2 Notation for Knowledge Type

HasSubevent: 0
MadeOf: 1
HasPrerequisite: 2
MotivatedByGoal: 3
AtLocation: 4
CausesDesire: 5
IsA: 6
NotDesires: 7
Desires: 8
CapableOf: 9
PartOf: 10
HasA: 11
UsedFor: 12
ReceivesAction: 13
Causes: 14
HasProperty: 15

In the remainder of this section, we use $\cup$ to indicate mask union operation upon multiple commonsense relations.

B.3 Stand-alone Fine-tuning

For fine-tuning on commonsense reasonging tasks, we only experiments with BERT-base due and perform hyper-parameter search only in terms of batch size in the range of $\{8, 16, 32\}$ and learning rate in the range of $\{3e^{-5}, 4e^{-5}, 5e^{-5}\}$ due to computational budget. We also adopt early stopping based

Model/Task	RTE	COPA	CSQA	SWAG	HellaSWAG	aNLI	CosmosQA
BERT	0 ∪ 6 ∪ 14	5 ∪ 8 ∪ 14	3 ∪ 4 ∪ 8 ∪ 14	1 ∪ 6 ∪ 10 ∪ 11	0 ∪ 3 ∪ 8 ∪ 14	3 ∪ 5 ∪ 8 ∪ 14	3 ∪ 5 ∪ 8

Table 8: Optimal fine-tuning knowledge type combination for BERT-base on commonsense reasoning tasks.

Model/Task	COPA (Dev.)	CSQA	CA	WSC	SM	ARCT1	ARCT2
DistilBERT	1 ∪ 6 ∪ 14	2 ∪ 3 ∪ 13	0 ∪ 1 ∪ 9	6 ∪ 7 ∪ 10	2 ∪ 8 ∪ 13	2 ∪ 3 ∪ 14	1 ∪ 2 ∪ 7
BERT	4 ∪ 11 ∪ 15	1 ∪ 2 ∪ 15	6 ∪ 8 ∪ 12	2 ∪ 9 ∪ 14	6 ∪ 12 ∪ 15	1 ∪ 9 ∪ 10	1 ∪ 5 ∪ 8
RoBERTa	2 ∪ 3 ∪ 8	0 ∪ 2 ∪ 5	0 ∪ 1 ∪ 8	1 ∪ 2 ∪ 4 ∪ 5 ∪ 11	8 ∪ 11 ∪ 12	2 ∪ 5 ∪ 11 ∪ 13	0 ∪ 8 ∪ 11
MPNet	1 ∪ 6 ∪ 8	6 ∪ 12 ∪ 13	2 ∪ 3 ∪ 10	1 ∪ 3 ∪ 4	6 ∪ 10 ∪ 13	2 ∪ 5 ∪ 6	5 ∪ 6 ∪ 7

Table 9: Optimal zero-shot knowledge type combination for each PLM on each commonsense reasoning tasks.

Model	P@1	P@2	P@3	Sparsity	$l_b - l_t$	# Param.
BERT-large w/o pruning	15.1	20.9	24.6	0%	-	336M
BERT-large w/ stochastic pruning	22.1	30.1	35.4	∼30%	17-24	336M
BERT-large w/ deterministic pruning	69.2	74.1	76.3	∼50%	17-24	284M

Table 10: Macro-averaged precision metrics of BERT-large on C-LAMA test set

on accuracy on the devlopment set. The combination achieving highest accuracy is shown in Table 8.

B.4 Zero-shot Learning

In constrast with fine-tuning, zero-shot evaluation is deterministic as long as the model does not involve any stochastic module, thereby averting extensive hyperparameter tuning. Instead we perform exaustive search over knowledge combinations for each pretrained language model with number of knowledge types in $\{3, 4, 5\}$. The ConceptNet-grounded knowledge type combination achieving highest accuracy is listed in Table 9.

C Extracted Commonsense Triples

Here we present the additional experiment result of extracting novel relaitonal triples based on our specialized relation-specific knowledge models. Applying the pruned DistilBERT-base model to predict missing objects for triples in ConceptNet-100K test set, we obtain commonsense triples deemed to be novel by three human annotators with Flessi's Kappa score κ of 0.65. We further filtered out triples that are included in the training or development set of ConceptNet-100K. Here we show some representative cases categorized by their relations:

CapableOf:
$(computer, crash)$, $(computer, communicate)$
IsA:
$(sex, relationship)$, $(submarine, weapon)$, $(submarine, vessel)$
AtLocation:
$(knife, war)$, $(knife, dinner)$, $(crab, dinner)$
UsedFor:
$(stage, fun)$, $(stage, performance)$, $(literature, education)$, $(literature, research)$
HasA:
$(book, index)$, $(book, information)$
HasProperty:
$(music, loud)$

Future work involves using seed triples beyond ConceptNet-100K dataset, e.g., the whole ConceptNet knowledge graph , and mining more novel and plausible commonsense knowledge.

D Limitations

The major limitations of our work lie in how do we choose and combine the representation subspaces/subnetworks in multi-relation scenario. We proposed a simple heuristic (i.e., based on statistics of dataset and union operation upon masks) in the paper and it empirically works well, but more principled and optimal method should be further studied. Another potential limitation is that we limit our scope to commonsense relations only in this paper. We leave other binary relations (e.g., factual relations defined in WikiData) as future work.