Commonsense Knowledge Transfer for Pre-trained Language Models

Anonymous ACL submission

Abstract

Despite serving as the foundation models for a wide range of NLP benchmarks, pre-trained language models have shown limited capabilities of acquiring implicit commonsense knowledge from self-supervision alone, compared to learning linguistic and factual knowledge that appear more explicitly in the surface patterns in text.

In this work, we introduce *commonsense knowledge transfer*, a framework to transfer the commonsense knowledge stored in a neural commonsense knowledge model to a general-purpose pre-trained language model. It first exploits general texts to form queries for extracting commonsense knowledge from the neural commonsense knowledge model and then refines the language model with two self-supervised objectives: *commonsense mask infilling* and *commonsense relation prediction*, which align human language with the underlying commonsense knowledge.

Empirical results show that our approach consistently improves the model's performance on downstream tasks that require commonsense reasoning. Moreover, we find that the improvement is more significant in the few-shot setting. This suggests that our approach helps language models better transfer to downstream tasks without extensive supervision by injecting commonsense knowledge into their parameters.

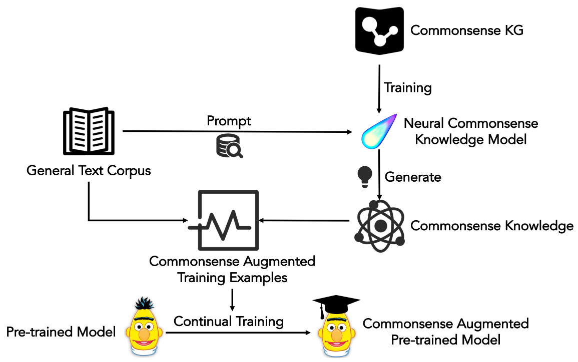

Figure 1: Illustration of the commonsense knowledge transfer framework. We first extract commonsense knowledge related to sentences in general text corpus from a neural commonsense knowledge model. We then use natural texts and the extracted commonsense knowledge to form self-supervised training data to refine a pre-trained model with commonsense knowledge.

1 Introduction

Recent advances in pre-trained language models have transformed the landscape of natural language processing. Self-supervised pre-training objectives including masked language modeling (Devlin et al., 2019) and masked span infilling (Lewis et al., 2020) enable pre-trained models to acquire linguistic (Hewitt and Manning, 2019; Manning et al., 2020) and factual knowledge (Petroni et al., 2019) by modeling the distribution of naturally occurring texts. However, most of these objectives are limited to exploiting the surface form of human language, and the lack of grounded supervision calls into question how well these representations can ever capture meaning (Bender and Koller, 2020), not to mention the underlying commonsense knowledge which is often reasoned implicitly and does not appear in the surface form of human language (Merrill et al., 2021; Zhou et al., 2020a; Hwang et al., 2021). On the other hand, commonsense reasoning is important for building generalizable models because it enables the model to reason about a great number of events, causes, and effects, while observing only a small fraction of them. The ineffectiveness of self-supervised language model pre-training on acquiring commonsense knowledge makes them require a relatively large number of labeled examples to succeed in a downstream task and prune to overfit task-specific correlations (Tu et al., 2020).

Therefore, equipping pre-trained language models with commonsense reasoning ability has attracted much attention. To this end, two distinct lines of research focus on improving commonsense reasoning ability of pre-trained language models. The first one focuses on incorporating external commonsense knowledge graph for commonsense reasoning (Lin et al., 2019; Liu et al., 2021; Cui and Chen, 2021) while the other attempts to inject commonsense knowledge into the parameters of pre-trained models (Li et al., 2019; Zhou et al., 2021; Klein and Nabi, 2021). In this work we focus on the second type of method because it alleviates the need for external knowledge bases for training and inference on downstream tasks, thus simpler, more efficient, and not limited by the coverage issue of external knowledge bases.

Prior art inject commonsense knowledge into pre-trained models either on symbolic commonsense knowledge graphs with manually defined rules (Li et al., 2019) or masked language modeling (Hosseini et al., 2021) or on general text corpus with concept-centric self-supervised objectives (Zhou et al., 2021). The former method is limited by the coverage of knowledge graphs and human-written rules. It also fails to make use of large scale diverse natural text corpus. Therefore, the training is limited on short and synthetic commonsense tuples, which affects its generalization ability on diverse downstream tasks. The latter method, however, only captures surface-level order relations between concepts and fail to learn commonsense relations between concepts such as cause, effect, intent, requirement, etc., which are crucial for commonsense reasoning but often implicitly reasoned, thus do not appear in the surface form of natural language.

In this work, we propose *commonsense knowledge transfer*, an alternative framework to refine a general purpose pre-trained model's commonsense reasoning ability. In contrast to previous work, it aims to transfer the commonsense knowledge stored in a neural commonsense knowledge model (e.g., COMET (Bosselut et al., 2019)) to a general purpose pre-trained model on large scale general text corpus. In this way, our approach combines the best of both worlds from prior art: the dense and informative commonsense knowledge from commonsense knowledge graphs and the accessibility of large scale diverse general corpus.

Commonsense knowledge transfer is conceptually related to knowledge distillation (KD) (Hinton et al., 2015) since they both aim to transfer knowledge from a knowledge-rich model to another model that lacks it. However, different from conventional KD, in commonsense knowledge transfer, the source model (i.e., neural commonsense model) and the target model (i.e., pre-trained model) are heterogeneous. Moreover, instead of simply mimicking the teacher model, commonsense knowledge transfer requires the target model to learn specialized knowledge from the source model while retaining its own capability. This poses unique challenges since the knowledge transfer can not be accomplished by simply matching the logits or feature distribution from the student and the teacher. To this end, we propose to first extract commonsense knowledge in textual form from the source model, and then exploits the extracted knowledge to form self-supervised training data for the target model. As illustrated in Figure 1, commonsense knowledge transfer first exploits general texts to form queries for retrieving commonsense knowledge from the neural commonsense knowledge model. Then it refines a pre-trained model with two self-supervised objectives that align surface form of human language with its underlying commonsense inference: *commonsense text infilling* and *commonsense relation prediction*. The former objective concatenates natural text with its commonsense inference to form an input example, mask certain spans in it, and train the model to reconstruct the original input. The latter method instead trains the model to distinguish valid commonsense inference from carefully constructed spurious commonsense inference given the original text and commonsense relation. Refining a pre-trained model by multi-tasking on both generation (former) and understanding (latter) tasks enables the model to better adapt to different kinds of downstream tasks.

We refine T5 (Raffel et al., 2020) with commonsense knowledge transfer and fine-tune the resulting model downstream tasks requiring commonsense reasoning ability in both the fully supervised setting and few-shot settings where only a percentage of labeled examples are available. Experimental results show substantial improvements on downstream tasks requiring commonsense reasoning, especially in the few-shot setting, demonstrating the effectiveness of our approach.

2 Methodology

Our proposed commonsense knowledge transfer framework consists of a neural commonsense knowledge model (e.g., COMET) and a pre-trained model (e.g., T5). The goal of commonsense knowledge transfer is to transfer the commonsense knowledge from the neural commonsense knowledge model (i.e., source model) to the pre-trained model (i.e., target model) so that it can generalize better to downstream tasks requiring commonsense reasoning ability.

Compared to conventional knowledge transfer methods such as knowledge distillation, commonsense knowledge transfer faces a unique challenge: the source model and the target model are heterogeneous because they are trained on different data with different objectives. As such, we can not simply feed a batch of data to both of the models and train the target model to match the source model's logits or feature distribution. To alleviate this problem, we propose a two-stage knowledge transfer scheme as illustrated in Figure 1. To be specific, we first use natural texts to form queries for retrieving commonsense knowledge (in text form) from the neural commonsense knowledge model. We then construct training data with two novel commonsense-related self-supervised objectives based on the retrieved commonsense knowledge and the corresponding natural text. Finally, we train the target model on the constructed training data to inject commonsense knowledge retrieved from the source model. We describe our method to extract commonsense knowledge from a neural commonsense knowledge model and the proposed commonsense-related self-supervised objectives in detail in this section.

2.1 Commonsense Knowledge Extraction

We first describe the source model, i.e., neural commonsense knowledge model, in the commonsense knowledge transfer framework. It is a transformer (Vaswani et al., 2017a) language model trained on commonsense knowledge graphs like ATOMIC (Sap et al., 2019a) and ConceptNet (Speer et al., 2017) with the objective of predicting the object (i.e., commonsense inference) with the subject (i.e., natural text) and relation as input. For example, given a commonsense tuple (s="take a nap", r=Causes, o="have energy"), the neural commonsense knowledge model is trained to generate o given s and r as inputs. After training, it can generate accurate, representative knowledge for new, unseen entities and events.

To extract commonsense knowledge stored in a neural commonsense knowledge model, we use a natural sentence as the subject s (e.g., he wants to cook a meal) and concatenate it with a randomly selected commonsense relation r (e.g., xNeed) from a pre-defined set to form a prompt (e.g., he wants to cook a meal xNeed). We then feed the prompt to the neural commonsense knowledge model and use it to generate a commonsense inference (e.g., to buy ingredients). In this way, the commonsense knowledge generation process resembles the way in which the neural commonsense knowledge model is trained. As such, we can get commonsense inferences of relatively high qualities.

Using a neural commonsense knowledge model as knowledge source has two advantages. On one hand, compared to the previous method (Li et al., 2019) using a symbolic commonsense knowledge graph, a neural commonsense knowledge model can generalize to unseen subjects, thus enabling us to refine the target pre-trained model on large-scale natural text corpus together with its commonsense inferences. As such, the resulting model can better adapt to downstream tasks which are formulated in diverse natural texts. On the other hand, compared to another method (Zhou et al., 2021) that only uses plain text and thus limited to the surface form of naturally occurring text, the use of a neural commonsense knowledge model provides much denser commonsense knowledge including a diverse set of commonsense relations between natural texts and the underlying commonsense knowledge.

2.2 Commonsense Knowledge Injection

After commonsense knowledge extraction, we need to inject the extracted commonsense knowledge into the target model. A straightforward solution is to use sequence-level knowledge distillation (Kim and Rush, 2016) and continually train the student to generate retrieved commonsense inference given the original text and commonsense relation. However, this can be sub-optimal due to the domain discrepancy between commonsense knowledge and natural text, which introduces the catastrophic forgetting problem (Kirkpatrick et al., 2017) and hurts the performance on downstream tasks, which is also recently confirmed by Cui and Chen (2021).

To better inject the extracted commonsense knowledge into a pre-trained model without suffering from catastrophic forgetting so that its capability on general NLP tasks is retained (or even improved), we propose two commonsense-related self-supervised objectives: *commonsense text infilling* and *commonsense relation prediction*. The former objective is generative while the latter is a discriminative objective. We refine the pre-trained model by multi-tasking on both the objective so

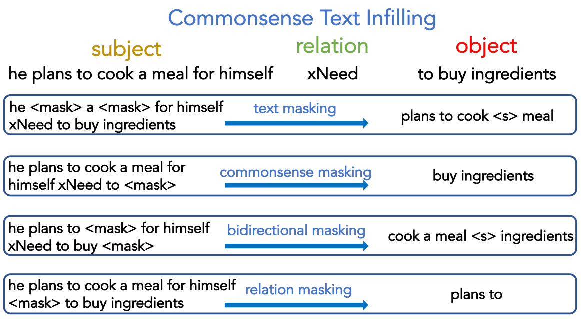

Figure 2: Illustration of the commonsense text infilling objective. Given a commonsense tuple constructed in the commonsense knowledge retrieval phase, we randomly mask text spans in the commonsense tuple following different patterns and train the pre-trained model to reconstruct the masked spans.

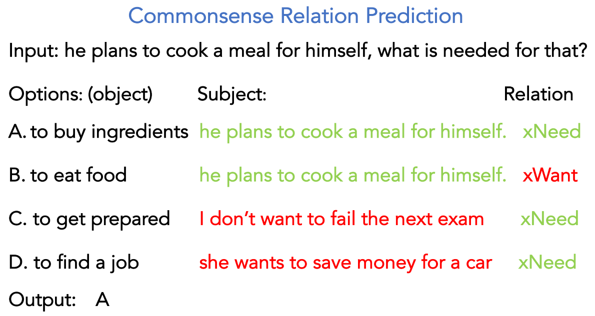

Figure 3: Illustration of the commonsense relation prediction objective. We train the pre-trained model to predict the correct commonsense inference given the subject and relation from three distractors generated with either different subjects or relations as inputs.

that the model can better adapt to tasks requiring either generative or discriminative commonsense reasoning ability.

Commonsense Text Infilling Commonsense text infilling is a simple extension to the conventional text infilling objective used for pre-training BART and T5. It transforms each sentence to a commonsense tuple similar to that in a commonsense knowledge graph by appending the commonsense relation and the generated commonsense inference. We then mask text spans in the commonsense tuple by randomly select one masking scheme among *text masking*, *commonsense masking*, *bidirectional masking*, and *relation masking*. As illustrated in Fig 2, these masking strategies selectively mask different components in the input commonsense tuple and lead to different optimization objectives. Specifically, these masking schemes masks either spans in natural text (P($s|\tilde{s}, r, o$)), commonsense inference (P($o|s, r, \tilde{o}$)), natural text/commonsense inference (P($s, o|\tilde{s}, r, \tilde{o}$)), or commonsense relation (P($r|s, \tilde{r}, o$)), respectively. We then train the model to predict the masked spans autoregressively. The diverse masking strategies provide more diverse training signals compared to randomly masking, thus enabling the model to better align the surface form of human language and the underlying commonsense knowledge.

In addition, unlike conventional practice in masked span infilling objective that randomly mask text spans with a same probability, we propose to mask text spans including concepts (tokens recognized as nouns or verbs by a Spacy POS tagger) with a higher probability so that the model will be trained to predict concepts more frequently compared to non-content words that are generally not related to commonsense reasoning.

Commonsense Relation Prediction While the commonsense text infilling objective encourages the pre-trained model to align natural texts and their commonsense inferences, it is always trained on *valid* commonsense tuples. This can be suboptimal because we also want the model to be capable of discriminating invalid commonsense inferences, which is important for many commonsense-related downstream tasks.

To this end, we introduce a commonsense relation prediction task which trains the model to distinguish the correct commonsense inference corresponding to the input sentence and the commonsense relation from distractors. To be specific, the commonsense relation prediction objective is formulated as a multi-choice QA problem with an input sentence as the context, a commonsense relation as the question, and a set of four commonsense inferences as options. The set of options consists of one correct commonsense inference, which is generated by the neural commonsense model with the input sentence and commonsense relation as input, and three carefully curated distractors (i.e., negative examples) generated by the same neural commonsense knowledge model with different inputs. As illustrated in Figure 3, among the three distractors, one is generated with an input composed by the same sentence and a different commonsense relation, and another two are generated with an input composed by different sentences with the same commonsense relation. In this way, the model learns to align the natural texts with valid commonsense knowledge while also distinguishing commonsense inferences that do not make sense.

Methods	CSQA	OBQA	PIQA	aNLI	SOCIALIQA	COPA
BERT-base	53.08(±0.16)	57.60(±0.8)	64.86(±0.52)	61.88(±0.56)	64.3(±0.4)	67.3(±0.4)
ERNIE-base	54.06(±0.12)	58.90(±0.9)	66.47(±0.58)	63.04(±0.46)	65.1(±0.4)	68.9(±0.4)
KnowBERT	53.88(±0.15)	58.50(±0.8)	66.61(±0.63)	63.18(±0.52)	65.4(±0.5)	69.4 (±0.4)
T5-base	61.88(±0.08)	58.20(±1.0)	68.14(±0.73)	61.10(±0.38)	65.1(±0.5)	71.4 (±0.7)
T5-base + TI	62.05(±0.17)	58.43(±0.8)	68.32(±0.66)	61.42(±0.32)	65.3(±0.4)	71.8 (±0.8)
T5-base + SSM	62.37(±0.25)	58.60(±0.9)	68.48(±0.65)	61.57(±0.44)	65.5(±0.5)	72.1 (±0.6)
T5-base + CSKG (TI)	60.22(±0.40)	56.17(±0.8)	66.51(±0.57)	59.92(±0.47)	62.7(±0.7)	68.5 (±1.1)
T5-base + CSKG (Rule)	63.10(±0.35)	57.97(±0.8)	68.27(±0.71)	60.15(±0.51)	65.7(±0.4)	72.4 (±0.9)
T5-base + KD	61.83(±0.42)	56.54(±0.7)	67.35(±0.63)	60.94(±0.66)	64.8(±0.5)	71.0 (±1.0)
CALM	63.32(±0.35)	60.90(±0.4)	71.01(±0.61)	63.20(±0.52)	66.0(±0.5)	72.2 (±0.8)
CKT-base	**64.11(±0.31)**	**61.58(±0.5)**	**72.26(±0.61)**	**64.37(±0.49)**	**67.3(±0.4)**	**73.4 (±0.5)**

Table 1: **Experimental results on base-size models.** Best models are bold and second best ones are underlined within each metric. Mean and standard deviation of 3 different runs with different random seeds are reported. TI denotes the text infilling objective and SSM denotes the salient span masking objective.

Moreover, this objective is formulated as a multi-choice QA task which closely resembles several downstream commonsense-related tasks such as CommonsenseQA and SOCIALIQA, thus enabling easier transfer especially when labeled training examples are scarce.

3 Experiments

3.1 Experimental Settings

Models In our experiments we apply commonsense knowledge transfer to refine T5 (Raffel et al., 2019), a popular model pre-trained with the text infilling objective. We experiment with both T5-base and T5-large, which consist of 220 million and 774 million parameters respectively, as the target model in the commonsense knowledge transfer framework. We use COMET-ATOMIC$^{20}_{20}$, a state-of-the-art neural commonsense knowledge model that can generate accurate, representative knowledge for new, unseen entities and events, as the source model. It is initialized with BART and continually trained on ATOMIC$^{20}_{20}$ (Hwang et al., 2021), a new general purpose commonsense knowledge graph.

Data We randomly sample a subset consisting of 10 million sentences from the English Wikipedia and the BookCorpus (Zhu et al., 2015), which is used for pre-training BERT and its variants. We select a set of representative commonsense relations including intent, reason, effect, need, want, and react from relations used to train COMET-ATOMIC$^{20}_{20}$. For each sentence, we randomly sample two relations and retrieve the corresponding commonsense explanation from COMET$^{20}_{20}$. We randomly select one relation-explanation pair to form the input example and leave another as the distractor for the commonsense relation prediction objective.

Training We refine the pre-trained models on the self-supervised examples constructed with the sampled 10 million sentences for 100k steps with a batch size of 1024, a maximum sequence length of 256, and a learning rate of 5e-5/2e-5 for base-size and large-size models respectively with a linear warm-up for the first 8,000 updates. After knowledge transfer, we fine-tune the models on downstream tasks by formulating the tasks into text-to-text problems. Pre-training and fine-tuning details are included in the Appendix.

Evaluation We evaluate the continual pre-trained models on downstream tasks that require commonsense reasoning including **CommonsenseQA** (Talmor et al., 2018), **OpenbookQA** (Mihaylov et al., 2018), **PIQA** (Bisk et al., 2020), **aNLI** (Bhagavatula et al., 2019), **COPA** (Roemmele et al., 2011), and **SOCAILIQA** (Sap et al., 2019b) In addition to the conventional fully supervised setting, we also test our approach in the few-shot setting by varying the percentage of labeled examples from the original training set used for fine-tuning. The idea is that limited labeled examples can only help the model understand the task but are insufficient for the model to acquire enough commonsense knowledge to solve the task. As such, it requires the model to store enough commonsense knowledge in its parameters to succeed in the few-shot setting. For both the settings, we report the results on the official development set and tune the hyperparameters based on the models' performance on an in-house split dev set. We report the mean and variance of 3 individual runs with different random seeds because most datasets are relatively small, which makes the variance in results non-negligible.

Methods	CSQA	OBQA	PIQA	aNLI	SocialIQA	COPA
BERT-large	57.06(±0.12)	60.40(±0.6)	67.08(±0.61)	66.75(±0.56)	69.5(±0.4)	82.8(±0.8)
T5-large	69.81(±1.02)	61.40(±1.0)	72.19(±1.09)	75.54(±1.22)	71.3(±0.8)	83.6(±1.1)
CALM-large	71.31(±0.04)	66.00(±1.0)	75.11(±1.65)	77.12(±0.34)	72.7(±0.7)	84.9(±1.0)
CKT-large	**72.15(±0.61)**	**66.70(±1.1)**	**76.07(±0.95)**	**77.94(±0.59)**	**73.8(±0.8)**	**86.0(±1.2)**

Table 2: **Experimental results on large-size models.** Best models are bold and second best ones are underlined within each metric. Mean and variance of 3 different runs with different random seeds are reported.

Baselines We compare our approach with methods that continual train a pre-trained model with different objectives. We divide the baselines into two categories based on the source of their supervision. The first category include methods that only exploit general text corpus, including (1) **T5 + TI** that continually pre-trains the public checkpoint of T5 with the same text infilling objective for more steps, (2) **T5 + SSM** that also continual pre-trains T5 with the text infilling objective, but use salient span masking (Roberts et al., 2020) instead of random masking for data construction, and (3) **CALM** (Zhou et al., 2021) that uses novel self-supervised objectives to construct concept-centric self-supervision from general text corpus. The second category instead exploit CSKG, including (4) **T5 + CSKG (TI)** train T5 with the text infilling objective on tuples in a CSKG, and (5) **T5 + CSKG (Rule)** (Li et al., 2019) that use manually defined rules to construct training examples from a CSKG and continually pre-train T5 with these examples. We also include a baseline method using sequence-level knowledge distillation (Kim and Rush, 2016) (**T5 + KD**). For fair comparison, we use the same data and training steps compared to our approach for baselines from the first category, and use ATOMIC$^{20}_{20}$, on which the teacher model in our framework is pretrain on, as the commonsense knowledge graph and train until convergence. For reference, we also include some popular knowledge-enhanced pre-trained model including ERNIE (Zhang et al., 2019) and KnowBERT (Peters et al., 2019).

3.2 Fully-supervised Results

We first present results in the fully-supervised setting. Results on base-size models are presented in Table 1. We can see that our approach yields significant improvement compared to the T5 baseline (up to 4 absolute scores) and consistently outperform CALM, the state-of-the-art method on injecting commonsense knowledge into PTLMs.

In addition, we observe that simply using continual training with the original text infilling objective or its variant with salient span masking only marginally improves the performance. Surprisingly, training with text infilling on a commonsense knowledge graph leads to degraded performance compared to the T5 baseline. We suspect this is because the commonsense tuples in commonsense knowledge graphs are generally too short and simple, making the pre-trained model unable to reason within relatively long contexts which is crucial for most downstream tasks. Moreover, we find that continually pre-training with training data constructed with commonsense tuples in a commonsense knowledge graph following manual designed rules leads to improvements in certain tasks. However, the improvement is inconsistent across different tasks and it even hurts the performance on certain tasks, which may because the rules for constructing training data are tailored for certain tasks like CSQA. The inferior performance of using commonsense knowledge graphs as data sources also confirms the need of using natural text corpus during continual pre-training for better adapting to diverse downstream tasks. Moreover, directly applying sequence-level KD and train the student to mimic the teacher on the commonsense tuple generation task fails to improve the performance because the task is too narrow and thus cannot transfer to diverse downstream tasks well.

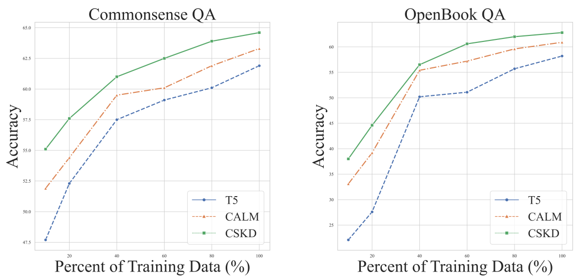

Figure 4: Performance of compared base-size models fine-tuned with different fraction of the datasets.

To further confirm the effectiveness of commonsense knowledge transfer, we apply it on T5-large and compare it to competitive baselines in the base-

Methods	CSQA	OBQA	PIQA	aNLI	SIQA	COPA
T5-base	61.88	58.20	68.14	61.10	65.1	71.4
CKT-base	64.57	62.77	73.26	64.75	68.3	73.4
Objective Analysis						
CKT-base w/o CSTI	62.58	60.97	70.61	62.11	66.5	72.0
CKT-base w/o text masking	62.98	61.74	72.55	63.81	67.7	72.8
CKT-base w/o commonsense masking	63.61	62.03	72.83	64.40	67.5	72.7
CKT-base w/o bidirectional masking	63.52	62.11	72.30	64.24	67.6	72.9
CKT-base w/o relation masking	64.12	62.48	73.31	64.57	67.4	72.7
CKT-base w/o CSRP	63.12	62.07	72.44	64.11	67.5	72.6
CKT-base w/ random distractors	64.04	62.29	72.95	64.48	68.0	73.1
Multi-task versus Sequential Transfer						
CKT-base (CSTI → CSRP)	64.69	62.51	73.35	64.11	67.9	73.5
CKT-base (CSRP → CSTI)	63.49	61.33	71.54	63.41	67.0	72.0
Corpus Size						
CKT-base w/ 10% data	64.18	62.21	71.86	64.31	67.7	73.1
CKT-base w/ 50% data	64.45	62.66	73.10	64.72	68.2	73.4

Table 3: Analysis of the proposed commonsense knowledge transfer framework. CSTI and CSRP denote the commonsense text infilling objective and the commonsense relation prediction objective, respectively. CSTI → CSRP means first continual pre-training using CSTI and then switch to the CSRP objective, and vice versa.

size experiments. The results are presented in Table 2. We can see that our approach consistently outperforms T5-large and CALM-large. This suggests that our approach can successfully generalize to large-size pre-trained models.

3.3 Few-shot Results

Injecting commonsense knowledge into pre-trained models is important because it enables the model to reason and generalize to unseen examples while observing only a few labeled examples. To this end, we fine-tune the compared models with different fractions of labeled training data to investigate the transition of the behavior of our model and baselines from the low-resource regime to the fully-supervised setting (Fig. 4). We observe that the performance improvement of our approach compared to the baselines is more significant in the low-resource regime. This shows that commonsense knowledge transfer can successfully transfer commonsense knowledge into pre-trained models so that they can generalize well while seeing only a small part of training data. This may also help the model reduce the risk/tendency of fitting the spurious correlations in the annotated datasets and thus generalize better.

3.4 Analysis

To better understand the proposed commonsense knowledge transfer framework and the role of its different components, we conduct an ablation study about the impact of different proposed objectives, the impact of multi-tasking the commonsense-related self-supervised objective versus sequentially training, and the impact of the size of natural text corpus used for transfer (see Table 3).

Impact of Objectives We find that both the proposed objectives contribute to the performance improvement of our approach. The commonsense text infilling objective is shown to be more critical than the commonsense relation prediction task. We suspect this is because commonsense text infilling resembles the vanilla text infilling objective with which the T5 models are pre-trained, thus preventing the model from catastrophic forgetting. In addition, all of the four masking strategies are beneficial, and their contribution varies for different downstream tasks. This confirms the necessity of a diverse masking scheme. Moreover, our strategy for constructing distractors outperforms the random counterpart, demonstrating the necessity of hard negative examples for the commonsense relation prediction task.

Multi-task versus Sequential Transfer As for the training order between the two objectives, we find that starting from the commonsense text infilling task and then switching to the commonsense relation prediction task performs similarly with our multi-tasking strategy while significantly outperforming its counterpart training with the reverse direction. We think this is because the commonsense text infilling objective resembles the original pre-training while the commonsense relation prediction is more similar to downstream tasks. We opt to the multi-tasking strategy because of its simplicity.

Impact of Corpus Size We find that commonsense knowledge transfer significantly outperforms both the T5 baseline and the competitive CALM method with only 10 percent of the full data used for distillation. Nevertheless, the performance improvement also confirms that our approach can benefit from the accessibility of large-scale natural texts. For base-size models, the performance improvements seem to saturate after 10 million sentence pairs. However, we anticipate that larger-size models may still benefit from a larger amount of data, and leave this for future work.

4 Related Work

SSL for NLP Recently, the pre-training then fine-tuning paradigm has become a common practice in NLP. Large scale language models based on transformer architecture (Vaswani et al., 2017b) pre-trained with self-supervised objectives including mask language modeling objective (Devlin et al., 2018; Liu et al., 2019; Lan et al., 2019) and text infilling objective (Lewis et al., 2019; Raffel et al., 2019) have advanced the state of the art on multiple NLU and NLG tasks.

Knowledge-augmented Pre-trained Models A number of recent works have examined the problem of incorporating world knowledge with the pre-trained models. A number of works utilizes an external knowledge base to incorporate entity knowledge with pre-trained models (Zhang et al., 2019; Peters et al., 2019; Wang et al., 2020; Liu et al., 2020). However, these approaches require specialized resources like knowledge bases which are non-trivial to seek, thus limiting the domain they can be applied to. Xiong et al. (2020) proposed a novel entity replacement detection objective which incorporates Wikipedia to encode world knowledge into a BERT-like pre-trained model. The aforementioned approaches generally focus on factual knowledge of entities while our work mainly focuses on commonsense knowledge.

Commonsense Reasoning for NLP Several recent studies (Talmor et al., 2018; Sap et al., 2019c; Zhou et al., 2020b; Lin et al., 2020; Xu et al., 2021) evaluate the performance of several pre-trained language models on tasks that require commonsense reasoning and find that it is still very hard for pre-trained language models to match or exceed human-level performance even fine-tuned on many labeled examples. Therefore, approaches to improve the commonsense reasoning ability of pre-trained language models has attracted much attention. The approaches for improving the commonsense reasoning ability of pre-trained models can be divided into two categories. The first category focuses on incorporating an external commonsense knowledge graph for commonsense reasoning. For example, Lin et al. (2019), Cui and Chen (2021), and Liu et al. (2021) propose to exploit structured symbolic commonsense knowledge graphs to perform commonsense reasoning. The second one instead attempts to inject commonsense knowledge into the parameters of pre-trained models. For example, Li et al. (2019) proposed to use manually designed rules to construct commonsense related training examples from commonsense knowledge graphs. Zhou et al. (2021) instead only relies on general text corpus and proposed two concept-centric self-supervised objectives to refine pre-trained models with commonsense knowledge. Concurrently to our work, Hosseini et al. (2021) propose to verbalize commonsense knowledge graphs into a text corpus and continually train BERT with the masked language modeling objective on it.

5 Conclusion

We introduce commonsense knowledge transfer, a framework to transfer the commonsense knowledge stored in a neural commonsense knowledge model into a general-purpose pre-trained model. Our method first extracts commonsense knowledge from the source model and then uses the extracted knowledge to construct self-supervised training data for the target model. Empirical results show that our approach outperforms previous methods that exploit either symbolic knowledge graphs or texts alone. Moreover, our proposed approach may also be generalized to transfer other types of knowledge (e.g., factual knowledge) from specific knowledge models to general-purpose models.

Ethical Considerations

Our work focuses on improving the commonsense reasoning ability of pre-trained language models. It probably does not introduce extra ethical concerns. However, in commonsense knowledge extraction, the neural commonsense knowledge model may generate unexpected (e.g., biased) commonsense inferences and training with these inferences may lead to additional bias in the pre-trained model. Nevertheless, all pre-trained language models contain bias and should be examined.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1908.05739*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL (1)*, pages 4762–4779. Association for Computational Linguistics.

Wanyun Cui and Xingran Chen. 2021. Enhancing language models with plug-and-play large-scale commonsense. *CoRR*, abs/2109.02572.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

WA Falcon. 2019. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Pedram Hosseini, David A. Broniatowski, and Mona T. Diab. 2021. Commonsense knowledge-augmented pretrained language models for causal reasoning classification. *CoRR*, abs/2112.08615.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, pages 6384–6392. AAAI Press.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*, pages 1317–1327. The Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Tassilo Klein and Moin Nabi. 2021. Towards zero-shot commonsense reasoning with self-supervised refinement of language models. *CoRR*, abs/2109.05105.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, and Dian Yu. 2019. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach. *CoRR*, abs/1909.09743.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP/IJCNLP (1)*, pages 2829–2839. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1823–1840. Association for Computational Linguistics.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2020. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning. In *AAAI*, pages 6418–6425. AAAI Press.

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

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. USA*, 117(48):30046–30054.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *CoRR*, abs/2104.10809.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019c. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Trans. Assoc. Comput. Linguistics*, 8:621–633.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *NIPS*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian J. McAuley, and Furu Wei. 2021. Blow the dog whistle: A chinese dataset for cant understanding with common sense and world knowledge. In *NAACL-HLT*, pages 2139–2145. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *ICLR*. OpenReview.net.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020a. Evaluating commonsense in pre-trained language models. In *AAAI*, pages 9733–9740. AAAI Press.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020b. Evaluating commonsense in pre-trained language models. In *AAAI*, pages 9733–9740.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Pre-training and Fine-tuning Details

A.1 Pre-Training Details

We implement our models using Pytorch-lightning (Falcon, 2019) and Hugginface's Pytorch Transformers (Wolf et al., 2019). For pre-training phase, we use the AdamW optimizer with maximum sequence length 256, train batch size 8, gradient accumulation 8, warmup steps 8000, weight decay 0.01 and adam epsilon 1e-6. We train the models with 8 V100 GPUs and FP32 precision. The model is pre-trained for 10 epochs. We searched for the best learning rate for our model out of [5e-6, 2e-5, 5e-5, 1e-4].

A.2 Fine-Tuning Details

For fine-tuning, we use 4 V100 GPUs and use FP32. For all tasks, we use the AdamW optimizer with learning rate from [1e-5, 2e-5, 5e-5, 1e-4, 2e-4], maximum sequence length 256, batch size from [4, 8, 16, 32]. For all tasks, we use a warmup fraction of 0.01, and max epoch of 20.