Making Sense of Common Sense

How can language models perform better at commonsense reasoning?

Sonali Serro, Haerang Lee
{sonaliserro,haeranglee}@berkeley.edu

Abstract

Commonsense reasoning has long been acknowledged as an essential component of natural language understanding. While humans have an innate ability to acquire and develop commonsense reasoning skills, it is still a challenge for machines. CommonsenseQA (Talmor et al., 2019) is a benchmark dataset for commonsense question answering that continues to be difficult even for large-scale pretrained language models.

Our research investigates two methods to improve a language model's performance on CommonsenseQA—**intermediate task transfer learning** and **generative data augmentation**. We find positive results from the transfer learning approach as development accuracy increases by 2.3% from the baseline. Additionally, we observe improved model performance in especially low-resource settings. Our error analysis reveals some of the commonsense reasoning skills acquired from the intermediate task fine-tuning. The generative data augmentation method also shows encouraging results as we find that the language model converges faster, potentially reducing training time by half.

1 Introduction

He is beginning work at his first job. What is he feeling? Where in a house would a table be most likely placed? What do parents encourage their children to do when they experience boredom? Humans find questions like these relatively trivial to answer, but even state-of-the-art language models often trail behind human accuracy. This gap has highlighted the importance of commonsense reasoning in natural language processing, one that parallels our ability to reason about people's behavior and intentions, and our understanding of the physical world.[1]

[1] https://en.wikipedia.org/wiki/Commonsense_reasoning

Commonsense reasoning is a challenging NLP task, frequently attributed to reporting bias, as common assumptions are often left unspoken. As a result, our research aims to incorporate additional *commonsense knowledge* into the language model to improve performance on a commonsense reasoning task.

2 Methods

2.1 Intermediate Task Transfer Learning

Fine-tuning a pretrained model on an intermediate task before fine-tuning it on a target task has been an active area of research for improving a model's general language understanding abilities (Pruksachatkun et al., 2020). This approach of transfer learning has shown broad improvements, especially in low resource settings.

Identifying the *relevant* set of intermediate tasks for a particular target task can be challenging. It requires a good understanding of what linguistic skills and abilities would help improve a model's performance on the target task. Additionally, the ability to probe the skills a model learns as a result of the intermediate task training can guide the selection of those intermediate tasks.

However, due to time and resource constraints, our approach to identifying a small set of intermediate tasks is primarily guided by the analysis of the CommonsenseQA dataset, including the distribution of the question concepts, and also relevant research in the field that identifies tasks as resources of commonsense transfer learning. Additionally, we opt for *single* intermediate task training instead of multi-task training to isolate the effects of the skills learned from each intermediate task and to avoid any potential task interference or bias arising from disproportionate dataset sizes.

Task Name	Train	Dev	Type	source
CommonsenseQA	9,741	1,221	multiple-choice QA	ConceptNet
SocialIQA	33,410	1,954	multiple-choice QA	ATOMIC, crowdsourcing
CosmosQA	25,588	3,000	multiple-choice QA	personal narrative corpa, blogs
HellaSwag	39,905	10,042	sentence completion	video captions, ActivityNet
CommonGen	67,389	4,018	generative reasoning	caption corpa, crowdsourcing

Table 1: An overview of the target, intermediate, and generative data augmentation tasks in our experiments.

2.2 Generative Data Augmentation

Inspired by a very recent paper (Lin et al., 2020), our data augmentation pipeline uses context generated by a model trained on a generative commonsense reasoning task, CommonGen,[2] to augment the CommonsenseQA dataset.

Specifically, we first fine-tune a `T5` (Raffel et al., 2020) text-to-text pretrained model with the CommonGen dataset. We then extract the *nouns and verbs* for each question-answer pair from CommonsenseQA to build concept sets as inputs to the trained CommonGen model. We use the resulting model-generated sentences to augment our target CommonsenseQA dataset, and performance of a model fine-tuned using the augmented dataset is compared to the baseline.

The underlying idea behind this approach is that a successful CommonGen model will generate more *reasonable* and *likely* sentences for the correct answer choices, providing helpful hints to the model in making the right prediction.

2.3 Target Task

CommonsenseQA is a multiple-choice question answering dataset that aims to capture commonsense reasoning beyond associations (Talmor et al., 2019). Every question has one correct and four distractor answers. Derived from ConceptNet (Speer et al., 2017), each question concept is associated with the multiple answer concepts via the same semantic relation (*AtLocation, Causes, CapableOf, etc.*). The authors report that fine-tuning BERT-Large (Devlin et al., 2018) on CommonsenseQA achieves an accuracy of 55.9%, substantially lower than human performance, 88.9%. However, recent leaderboard results[3] indicate that this gap is slowly closing, with the state-of-the-art performance achieved by the UnifiedQA (Khashabi et al., 2020) model achieving a test accuracy of 79.1%.

[2]https://inklab.usc.edu/CommonGen/

[3]https://www.tau-nlp.org/csqa-leaderboard

2.4 Intermediate Tasks

SocialIQA (Sap et al., 2019) is a benchmark dataset for commonsense reasoning about *social situations*. The dataset contains questions that probe emotional and social intelligence from various everyday situations, and has been shown as a resource for transfer learning of commonsense knowledge, achieving state-of-the-art performance on multiple commonsense reasoning tasks (Winograd Schemas, COPA).

CosmosQA (Huang et al., 2019) is a largescale dataset that tests commonsense reading comprehension in a multiple-choice question format. The questions require interpretation of the likely *causes and effects of events*, and rely on a wide range of commonsense reasoning beyond the text presented in the context.

HellaSwag (Zellers et al., 2019) is a *commonsense natural language inference* task, generated via adversarial filtering that tests a machine's ability to choose the most plausible continuation to a sequence of events or steps. The dataset proves to be easy for humans (95.6%), yet challenging even for state-of-the-art models (<50%).

2.5 Data Augmentation Task

CommonGen (Lin et al., 2020) is a constrained text generation benchmark dataset that aims to evaluate a model's generative commonsense reasoning abilities. Given a set of common concepts (e.g., {*dog,frisbee,catch,throw*}), the task is to generate a sentence describing an everyday scenario using the concepts (e.g., {*a man throws a frisbee and his dog catches it.*}). The model is expected to construct a grammatical sentence while reasoning over the commonsense relations between the concepts. The authors achieve the best performance (`31.6` `SPICE` metric) using T5.

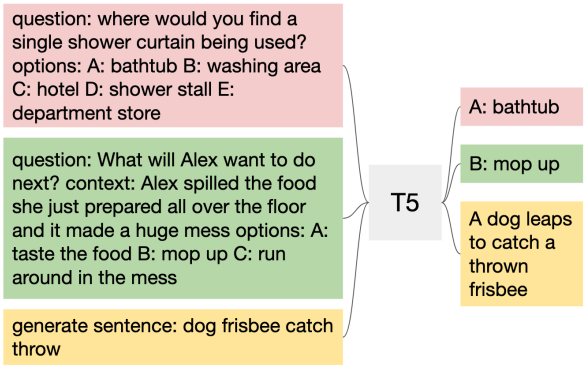

Figure 1: T5 input and output formats for CommonsenseQA, SocialIQA, and CommonGen.

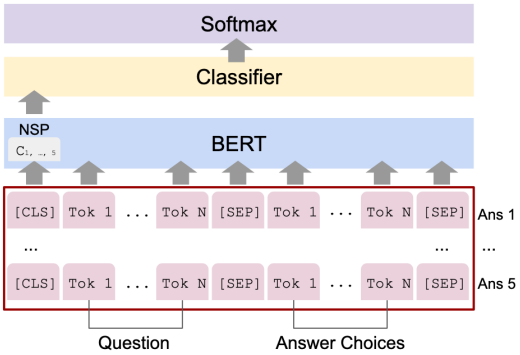

Figure 2: BERT architecture, and input and output formats for CommonsenseQA.

3 Experiments

3.1 BERT Baseline

To establish a baseline in BERT (Devlin et al., 2018), we use the `BertForMultipleChoice` model from the HuggingFace transformers.[4] It takes the pretrained Bert Base Uncased model, which has 110 million parameters, and adds a classification head for multiple choice question-answering tasks.

We preprocess the data by converting each example into an array of five sub-examples, each featuring two sentences (the question and one of the five answer choices). We then further process the example by inserting special Bert tokens and converting it into standard Bert input features as shown in Figure 2.

3.2 T5 Baseline

For the T5 baseline we use the large-scale pretrained text-to-text `T5-Base` model which has about `220` million parameters.[5] Our implementation is based on PyTorch (Paszke et al., 2019), and the HuggingFace Transformers (Wolf et al., 2020). In addition, we use the HuggingFace Trainer API,[6] which provides a relatively simple interface for training and evaluation.

Given that `T5` is a text-to-text model, we format the model input and expected output with inspiration from the original `T5` paper. The input formats influence model performance, which suggests that some representations of multiple-choice datasets are more efficient than others and should be further explored. For CommonsenseQA, we format the model input as *"question: ... options: A: ... B: ... E: ..."*, and the corresponding output as *"A: ..."*.

We fine-tune our models using a single NVIDIA Tesla K80 GPU, and perform limited hyperparameter tuning to find the best performing model across the following settings, learning rate: `{1e-4,5e-5,1e-5}`, batch size: `{4,8,16}`, and number of epochs: `{3,10}`. We use the Adam optimizer with linear learning rate schedule, and gradient accumulation steps=4. Model performance is evaluated using *accuracy* against the official development set.[7]

T5 versus BERT We opt to use `T5-Base` as the baseline model for the remainder of this study as `T5`'s text-to-text format is more amenable to the varying dataset formats in our transfer learning pipeline. Additionally, `T5-Base` is a better performing model that achieves an accuracy of 62.40%, as compared to 56.34% by `BERT-Base` (3-epoch evaluation).

3.3 Intermediate Task Transfer Learning

The first step in the pipeline is fine-tuning a `T5-Base` model on each of the intermediate tasks. We perform limited hyperparameter tuning based on available research and identify the model with the highest accuracy on the tasks' official development set. In the absence of established baselines for each of our intermediate tasks on `T5-Base`, we compare our model performance with published `BERT-Base` or `BERT-Large` results.

Over-fitting on the intermediate task may impair target task performance, possibly because of catastrophic forgetting. Consequently, we fine-tune the model on the intermediate task for just 3 or 4

[4] https://huggingface.co/transformers/model_doc/bert.html

[5] https://github.com/sonaliserro/w266-commonsenseqa/tree/master/T5

[6] https://huggingface.co/transformers/main_classes/trainer.html

[7] Limited test set evaluation performed by the leaderboard.

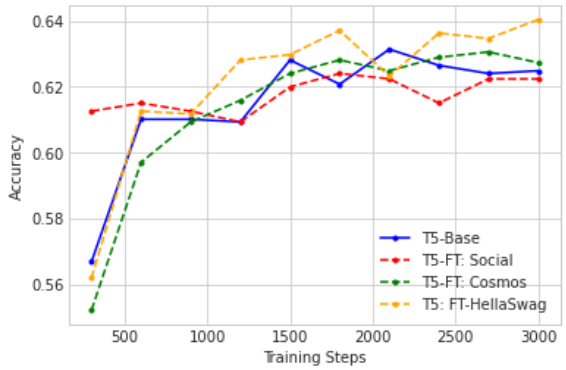

Figure 3: Accuracy over training steps over 10 epochs.

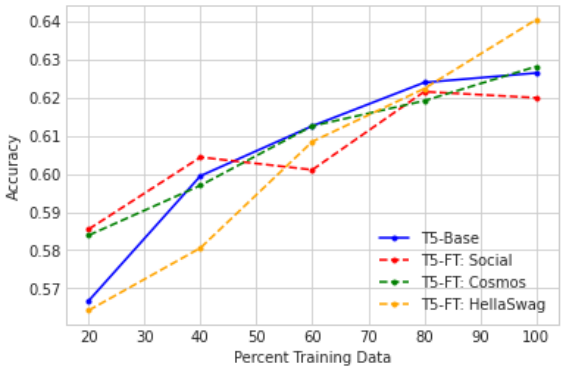

Figure 4: Accuracy per fraction of training data.

Model	Dev Acc (%)	LR	WS
T5-Base	62.57	1e-4	0
T5-FT$_{\text{Social}}$	62.00	1e-4	200
T5-FT$_{\text{Cosmos}}$	62.82	5e-5	0
T5-FT$_{\text{HellaSwag}}$	**64.04**	1e-4	200

Table 2: Intermediate Task Transfer Learning Results. Model accuracy is measured against the official CommonsenseQA development set. The notation `T5-FT` indicates a `T5-Base` model fine-tuned on the named intermediate task. Batch size=8 and epochs=10.

epochs. See Appendix A for the experiment results.

After intermediate task training, we fine-tune each resulting model on CommonsenseQA for up to 10 epochs, but with potentially different hyperparameters from the baseline in Sec 3.2. We evaluate model performance and also checkpoint the model every 300 steps during training. In some cases, using a small fraction of training steps as *warmup* helps the model adapt to the new dataset. All fine-tuned models[8] and formatted datasets[9] are available for download from Google Storage.

3.4 Generative Data Augmentation

In order to generate choice-specific context for the CommonsenseQA dataset, we first fine-tune a `T5-Base` model on the CommonGen dataset. The resulting model after 3 epochs achieves a `ROUGE-L` score of `32.30` using `nlg-eval` (Sharma et al., 2017).

We then extract the nouns and verbs for each question-choice pair from CommonsenseQA using the spaCy NLP library.[10] For instance, for the example, *"What do people aim to do at work? A: learn from each other B: wear hats"*, we extract the following concept sets: {*people, aim, work, learn, from, each, other*} and {*people, aim, work, wear, hats*}. Using these concepts as input, the CommonGen model generates the following sentences, *"people work together to learn from each other"* and *"people are wearing hats to work"*.

The last step is to augment the CommonsenseQA dataset with this generated context.[11] We extend each answer choice in the model input to also include the generated sentence separated by the pipe (" | ") character as follows, *"question: ... options: A: ... | ..."*. We suggest further exploration for a potentially more efficient input format to augment the dataset with the generated context.

We then fine-tune a `T5-Base` model with this augmented dataset for up to 10 epochs, and compare model performance with the established baseline from Sec. 3.2.

4 Experimental Results

4.1 Transfer Learning Experiments

Performance over Training Steps After 10 epochs of fine-tuning on CommonsenseQA, T5-FT$_{\text{HellaSwag}}$ achieves the highest dev accuracy as shown in Table 2. Figure 3 shows the learning curve across all models, evaluated every 300 steps. T5-FT$_{\text{Social}}$ converges much more quickly than T5-Base, achieving an accuracy of over 61% at 300 steps. T5-FT$_{\text{Cosmos}}$ converges more slowly but also more smoothly, and exceeds T5-Base accuracy by 0.25 percentage points at the end of 10 epochs. T5-FT$_{\text{HellaSwag}}$ starts surpassing all other models as early as 900 steps, and makes the most gains in the final 500 steps of training.

[8] gs://w266-commonsenseqa/models/T5
[9] gs://w266-commonsenseqa/data
[10] https://spacy.io/
[11] https://github.com/sonaliserro/w266-commonsenseqa/tree/master/T5/common_gen

Model	Dev Accuracy (%)	
	3 epochs	10 epochs
T5-Base	61.83	**62.73**
T5-Base$_{CG}$	**62.48**	62.48

Table 3: Generative Data Augmentation Results using CommonGen. Model accuracy is measured against the official CommonsenseQA development set. Batch size=8 and lr=1e-4.

Performance in Low-Resource Settings Figure 4 shows the results of evaluating the effectiveness of transfer learning with varying fractions of the CommonsenseQA training data.

When only 20% of training examples are available, we observe that T5-FT$_{\text{Social}}$ and T5-FT$_{\text{Cosmos}}$ outperform T5-Base, by up to 1.88 percentage points. However, when data availability increases, the initially estimated benefits quickly dwindle. Hence, these datasets can serve as a useful—albeit limited—supplementary resource when labeled, commonsense data is scarce.

We observe T5-FT$_{\text{HellaSwag}}$ has a consistently lower accuracy per resource level, but exceeds T5-Base at 100%. The HellaSwag dataset format is longer and more complex in comparison to the other intermediate tasks. Therefore, one possible explanation for this observation is that the model requires a lot more training examples to get accustomed to the new format. However, once the model has had the opportunity to see all training examples, the commonsense knowledge from the intermediate task leads to a positive transfer.

4.2 Data Augmentation Experiments

Results from the generative data augmentation can be found in Table 3. We find that a model fine-tuned with the CommonGen generated context converges considerably faster, and can potentially speed up training by a factor of 2. This result aligns with the findings from the original paper (Lin et al., 2020). Despite not being able to exceed our baseline performance, we recommend further exploration of this approach to improving a model's commonsense reasoning abilities.

5 Error Analysis

5.1 Sampling Errors Unique to Each Model

We compare the errors made by our four intermediate task transfer learning models on the CommonsenseQA development set. Each model makes between 440-465 incorrect predictions, out of which 207 question-answer pairs are common to all models. Figure 5 shows how many errors are common or unique to the models. We filter for the examples that one model predicts incorrectly while all other models do not. From this subset, we sample 20 errors per model for a detailed error analysis.

5.2 Quality of Predictions

We annotate the commonsense reasoning quality of the incorrect answer choices. A **high-quality** error is synonymous with the target or is a sensible alternative that a human respondent might choose. A **medium-quality** error is wrong but semantically related to the question or the target. A **low-quality** error is the opposite of the target or is non-sensical. See Appendix A for examples.

We observe that transfer learning tends to improve the quality of predictions (Figure 6). Fine-tuning on HellaSwag reduces both the proportion of low-quality errors and increases that of high-quality ones, suggesting that there is transferable commonsense knowledge. CosmosQA also appears helpful, as it reduces low-quality errors.

T5-FT$_{\text{Social}}$ is less likely to make both low- and high-quality responses than T5-Base. This analysis sheds light on why fine-tuning on SocialIQA does not improve the main task accuracy, despite our best efforts at hyperparameter-tuning.

5.3 Commonsense Reasoning Skills Required

For the same sample of errors, we annotate the type of commonsense reasoning skill a human would require to solve them. We reference the categories used in the original CommonsenseQA paper. See Appendix A for the categories.

Our T5-FT models show the greatest improvement in the questions requiring *activity* and *cause*

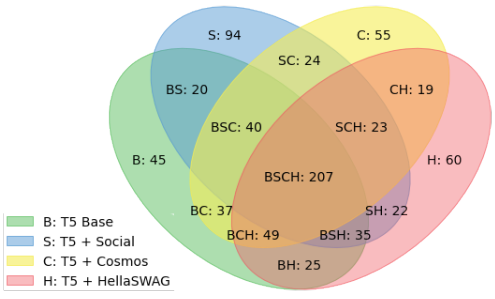

Figure 5: Venn diagram of incorrect predictions. The key of letters represent the collection of models that share the same question-answer pairs.

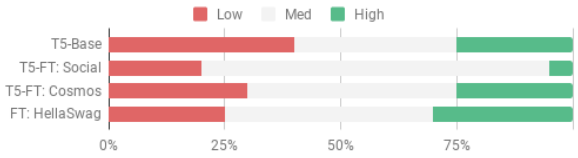

Figure 6: Distribution of answer quality over a sample of examples each model uniquely predicted incorrectly.

& effect reasoning skills. Whereas 30% to 40% of the sampled T5-Base errors require these skills, the proportion decreases to between 10% and 30% in the T5-FT models' errors.

The problems requiring *spatial* reasoning skills appear to improve substantially after transfer learning, doubling with fine-tuning on CosmosQA. The reason is likely that the questions requiring spatial reasoning skills represent 41% of the CommonsenseQA development set (Talmor et al., 2019).

T5-FT$_{\text{Social}}$ appears to make fewer mistakes requiring *social* reasoning skills than T5-Base, but T5-FT$_{\text{Cosmos}}$ makes even fewer such errors.

6 Related Work

Recent commonsense benchmarks like WinoGrande (Sakaguchi et al., 2019), HellaSwag (Zellers et al., 2019), and CommonsenseQA (Talmor et al., 2019) have highlighted the lack of common sense in language models. Consequently, there has been significant research to incorporate *"commonsense knowledge"* into language models, ranging from leveraging external commonsense knowledge bases such as ATOMIC and ConceptNet (Zhong et al., 2019; Mitra et al., 2020), to generation of human-like explanations for commonsense reasoning (Rajani et al., 2019). Furthermore, evaluating the amount of commonsense knowledge present in a language model is still an active area of exploration (Cui et al., 2020).

The state-of-the-art UnifiedQA (Khashabi et al., 2020) is a format-agnostic QA system that is based on the scaled-up `T5-11B` model pre-trained with a set of seed QA datasets. This mutli-task pretraining objective along with model size are perhaps some of the key factors of improved performance.

Our research is influenced by the large-scale study to understand the effectiveness of intermediate task transfer learning (Pruksachatkun et al., 2020), where the authors observe that tasks that require inference and reasoning abilities make good intermediate tasks. Despite the much smaller scale of our study, our results align with the original paper in that CosmosQA and HellaSwag both proved to be most successful in delivering positive commonsense knowledge transfer.

Required Skill	T5-Base	T5-FT: Social	T5-FT: Cosmos	T5-FT: HellaSwag	Total
Spatial	0.25	0.25	0.50	0.40	0.35
Cause & Effect	0.40	0.25	0.10	0.15	0.23
Activity	0.30	0.05	0.10	0.20	0.16
Purpose	0.10	0.15	0.20	0.15	0.15
Social	0.15	0.10	0.00	0.15	0.10
Definition	0.10	0.15	0.10	0.05	0.10
Preconditions	0.15	0.10	0.05	0.05	0.09
Is member of	0.05	0.05	0.00	0.05	0.04
Has parts	0.00	0.00	0.00	0.05	0.01

Figure 7: The fraction of uniquely incorrect questions requiring each commonsense reasoning skill. A question may require multiple skills.

7 Conclusion and Future Work

This study explores two different methods to incorporate "commonsense knowledge" into the language model. Our transfer learning results suggest that CosmosQA and HellaSwag can transfer a certain amount of commonsense knowledge that helps improve the model's performance on CommonsenseQA. CosmosQA and SocialIQA can also serve as useful supplemental datasets in low-resource settings, where labeled commonsense data is scarce. Even though our error analysis reveals some of the commonsense reasoning skills acquired from the transfer learning, it is difficult to draw conclusions about the specific skills that drive the positive transfer. Future work for this study would benefit from an accompanying probing analysis to evaluate the commonsense knowledge present in the language model (Petroni et al., 2019). Additionally, the results of the transfer learning may have been in part affected by catastrophic forgetting, and therefore suggests further work on potentially more efficient transfer learning mechanisms (Houlsby et al., 2019; Chen et al., 2020). Our data augmentation approach also reveals encouraging results, and points to interesting future directions toward incorporating generated commonsense knowledge into a language model.

Acknowledgments

We thank Joachim Rahmfeld and Daniel Cer from the UC Berkeley School of Information for advising us on the research methods.

References

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does bert solve commonsense task via commonsense knowledge?

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*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *arXiv:1909.00277v2*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single qa system. In *Findings of EMNLP*.

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. *Findings of EMNLP*. To appear.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2020. How additional knowledge can improve natural language commonsense question answering?

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQA: Commonsense reasoning about social interactions. In *EMNLP*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

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

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Improving question answering by commonsense-based pre-training.

A Appendices

A.1 Results of Hyperparameter Tuning on CommonsenseQA

LR	BS	Epochs	Dev Acc (%)
1e-4	8	3	62.40
1e-4	16	3	61.26
1e-4	4	3	61.58
1e-5	8	3	58.55
5e-5	8	3	59.37

Table 4: An overview of the results of fine-tuning T5-Base on CommonsenseQA, with the specified hyperparameters.

A.2 Intermediate Task Fine-Tuning Details

Task	LR	BS	Epochs	Dev Acc (%)
SocialIQA	1e-4	8	4	66.58
CosmosQA	1e-4	8	3	66.53
HellaSwag	1e-4	4	3	46.78

Table 5: An overview of the results of fine-tuning T5-Base on each intermediate task, evaluated on the tasks' development set.

A.3 CommonsenseQA Concepts

Category	Definition	%
Spatial	Concept A appears near Concept B	41
Cause & Effect	Concept A causes Concept B	23
Has parts	Concept A contains Concept B as one of its parts	23
Is member of	Concept A belongs to the larger class of Concept B	17
Purpose	Concept A is the purpose of Concept B	18
Social	It is a social convention that Concept A correlates with Concept B	15
Activity	Concept A is an activity performed in the context of Concept B	8
Definition	Concept A is a definition of Concept B	6
Preconditions	Concept A must hold true in order for Concept B to take place	3

Figure 8: CommonsenseQA question categories, as described in COMMONSENSEQA: A Question Answering Challenge Targeting Commonsense Knowledge (Talmor et al., 2019)

A.4 Sample Errors and Quality

Questions	Answer Choices (Green for target, *red* for incorrect prediction) A	B	C	D	E
HIGH-QUALITY					
Reading newspaper one of many ways to practice your what?	**literacy**	*knowing how to read*	money	buying	money bank
There was more than one bum asking for change or a ticket, it was the cheapest way to travel so it was no surprise sight at the what?	*train station*	beach	**bus depot**	bridge	stumblebum
MEDIUM-QUALITY					
When we are running what are we doing?	stretches	*running from police*	learn to walk	**go quickly**	get out of bed
What does drinking alcohol lead to?	have fun	**intoxication**	vomiting	drinking more alcohol	*nausea*
LOW-QUALITY					
What are people likely to do when an unexpected decent outcome occurs?	kill each other	**thank god**	*experience pain*	hatred	talk to each other
What would use a musical instrument?	guitar	music room	**orchestra**	case	*movie*

Figure 9: Sample errors by assessed response quality.

A.5 Leaderboard Results

We submitted our test predictions for T5-Base and T5-FT$_{\text{HellaSwag}}$ to the leaderboard, with the following results. The results align with an expected 2-6% drop in accuracy scores between development and test from other researchers in the community. We can only hypothesize the reasons for no improvement in test scores from our baseline to be related to evidence of catastrophic forgetting, or perhaps the possibility of overfitting the development dataset.

Model	Test Acc (%)
T5-Base	55.96
T5-FT$_{\text{HellaSwag}}$	55.35

Table 6: Leaderboard results for T5-Base and T5-FT$_{\text{HellaSwag}}$.