Rethinking Why Intermediate-Task Fine-Tuning Works

Ting-Yun Chang[†] **Chi-Jen Lu**

Institute of Information Science, Academia Sinica, Taiwan

tingyun@usc.edu   cjlu@iis.sinica.edu.tw

Abstract

Supplementary Training on Intermediate Labeled-data Tasks (STILT) is a widely applied technique, which first fine-tunes the pretrained language models on an intermediate task before on the target task of interest. While STILT is able to further improve the performance of pretrained language models, it is still unclear why and when it works. Previous research shows that those intermediate tasks involving complex inference, such as commonsense reasoning, work especially well for RoBERTa-large. In this paper, we discover that the improvement from an intermediate task could be orthogonal to it containing reasoning or other complex skills — a simple real-fake discrimination task synthesized by GPT2 can benefit diverse target tasks. We conduct extensive experiments to study the impact of different factors on STILT. These findings suggest rethinking the role of intermediate fine-tuning in the STILT pipeline.

1 Introduction

Pretrained language models (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019) have contributed to great progress in natural language understanding (NLU). STILT (Phang et al., 2018; Wang et al., 2019; Clark et al., 2019; Pruksachatkun et al., 2020; Phang et al., 2020; Vu et al., 2020) can further improve their performance on downstream NLU tasks by redesigning the training pipeline, introducing an intermediate-task fine-tuning phase before fine-tuning the pretrained models on the target task of interest (Figure 1). Nevertheless, this approach is not necessarily beneficial, and its effectiveness depends highly on the intermediate task applied.

To study *when* and *why* STILT works, Pruksachatkun et al. (2020) conduct large-scale experiments based on RoBERTa-large (Liu et al.,
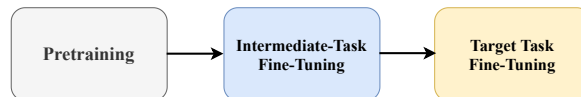

Figure 1: The pipeline of STILT.

2019) with different intermediate-target task pairs. They focus on studying what kind of intermediate tasks are helpful overall and which linguistic skills a model learns from the intermediate phase. They show the difficulty to have a generally useful intermediate task and conclude that those containing *complex reasoning and inference*, such as CosmosQA (Huang et al., 2019) and HellaSwag (Zellers et al., 2019), tend to enhance various target tasks. However, this ignores the fact that HellaSwag is a synthetic dataset, and RoBERTa tends to capture the data artifacts when fine-tuned on HellaSwag (Tamborrino et al., 2020).

In this paper, we demonstrate that intermediate tasks' enhancement could be irrelevant to providing complex reasoning or special linguistic skills — a simple real-fake discrimination task synthesized by GPT2 (Radford et al., 2019) can benefit diverse target tasks, including those commonsense reasoning tasks. These observations suggest rethinking the role of the intermediate-finetuning phase in the pipeline. Our main contributions are as follows.[1]

- We discover that a widely beneficial intermediate task is not required to provide specific linguistic or reasoning skills.

- We highlight STILT's enhancement on fine-tuning stability, providing more than 1000 experimental observations on RoBERTa-large.

- We study different factors that may influence STILT's efficacy, suggesting rethinking why it works.

[†]Work was done when the first author was a research assistant at Academia Sinica, Taiwan.

[1]Our source code is available at https://github.com/terarachang/Rethinking_STILT.git

Name	Size	Task	Input Format	Genre/Source
HellaSwag	40k	sentence continuation	multiple-choice	ActivityNet, WikiHow
HellaSwag-p	40k	real-fake discrimination	multiple-choice	ActivityNet, WikiHow
Synthesis_{GPT2}	30k	real-fake continuation	multiple-choice	Wikipedia
CoLA	8.5k	linguistic acceptability	1 sent.	linguistics publications
WiC	5.4k	word sense disambiguation	1 word; 2 sents.	WordNet, VerbNet...
RTE	2.4k	natural language inference	2 sents.	Wikipedia, news
MedNLI	11k	natural language inference	2 sents.	MIMIC-III clinical notes
SocialIQA	33k	commonsense QA	multiple-choice	crowdsourcing
$\text{WinoGrande}_{XS,M,L}$	0.2, 2.5, 10k	commonsense coreference	multiple-choice	crowdsourcing

Table 1: Overview of the tasks in our experiments. We include more descriptions in the appendix.

2 A *Good* Intermediate Task

We first define what a good intermediate task means. STILT is known for two benefits (Phang et al., 2018): 1) improving target tasks' *best* performance, and 2) stabilizing the fine-tuning process of the target tasks, notably reducing the degenerate fine-tuning runs (Devlin et al., 2019; Dodge et al., 2020; Mosbach et al., 2020). While Pruksachatkun et al. (2020) only focus on the first property, we study both benefits by extensive hyperparameter trials. Summarized in Table 1, we experiment on diverse, commonly used natural language understanding tasks, from word sense disambiguation to commonsense reasoning. A *good* intermediate task should provide both benefits to these tasks.

Note that the definition of *stability* could be controversial. Here, we follow previous work (Phang et al., 2018; Mosbach et al., 2020; Dodge et al., 2020) in this research line and refer to "improving stability" as "reducing the variance of the validation performance", which is strongly related to "reducing the occurrence of degenerate runs over multiple hyperparameters trials" as the variance in performance is often dominated by degenerate runs.

3 Rethinking: Two Simple Baselines

HellaSwag (Zellers et al., 2019) is a commonsense reasoning multiple-choice task, which contains a premise narrating an event and four plausible next scenarios (options) in each data example (Figure 4). All its negative options are generated by the machine given the premises; consequently, this dataset is known to contain superfluous artifacts (Tamborrino et al., 2020).

Despite the limitation, Pruksachatkun et al. (2020) show that HellaSwag is one of the most potent intermediate tasks for RoBERTa-large in their large-scale experiments[2]. They then attribute

such wide improvement on target tasks to the complex commonsense reasoning it requires. On the contrary, we first ablate the common sense from HellaSwag, seeking to understand if simple intermediate tasks are enough to enhance the performance of various target tasks.

We propose two baselines as intermediate tasks. The first one is to *remove the premises from HellaSwag*, denoted as HellaSwag-p, so that each data example only contains four options without contexts. Therefore, the model does not require common sense and reasoning skills to predict the follow-up anymore. It only needs to identify which option is *not* generated by the machine.

Secondly, we *build a synthetic dataset that mimics the creation of HellaSwag*, denoted as Synthesis_{GPT2}. The main difference is that, unlike HellaSwag, our premises and the correct endings are not from particular sources containing commonsense. We use Wikipedia as the source corpus since it has already been seen by the model in the pretraining phase. Specifically, given a sentence from Wikipedia, we split it into two parts. The first half becomes the premise, and the last half becomes the positive choice. We then use pretrained GPT2 to generate three negative choices conditioned on the premise. The decoding strategy is nucleus (top-p) sampling (Holtzman et al., 2019), where $p = 0.9$. More descriptions can be found in the supplementary materials.

Our goal is to use these two simple baselines to point out some underestimated factors when studying why STILT works. While previous work (Pruksachatkun et al., 2020) attempts to relate the linguistic skills between the intermediate and target task, we suspect that the linguistic knowledge further provided by the intermediate task could be less contributive than previous belief, as plenty of research

combinations and show that in many cases, the intermediate tasks are not helpful, or even hurtful in some cases.

---

[2] They study RoBERTa with 110 intermediate-target task

Intermediate	Target Accuracy (%)	
	RTE	**WiC**
None	83.5 / 85.6	70.5 / 71.8
HellaSwag	88.3 / 88.4	70.6 / 73.7

Table 2: Comparing Pruksachatkun et al./our performance on the same intermediate-target task pairs.

Interm.	Δ Mean/Best Accuracy (%)		
	MedNLI	**WiC**	**WinoG$_M$**
CoLA	+1.3 / +0.2	+3.9 / -0.3	-1.6 / -3.8
Hella-sh	+1.4 / ±0.0	+2.8 / -0.9	-1.7 / -3.5

Table 3: The effect of using other true-false tasks, CoLA and Hella-sh, as the intermediate tasks.

on probing pretrained language models has shown that diverse linguistic skills are already learned in the pretrained models' representations (Peters et al., 2018; Tenney et al., 2019b,a; Talmor et al., 2019). Please note that instead of challenging common sense and complex reasoning can be good properties for an intermediate task, the proposed two baselines are meant to raise the need of rethinking other important aspects of what a beneficial intermediate task offers.

4 Experiments

4.1 Setup

Following Pruksachatkun et al. (2020), we study the powerful pretrained model RoBERTa-large in all experiments. For each intermediate task, we perform a hyperparameter sweep over the learning rate in $\{5e - 6, 1e - 5, 2e - 5\}$, the effective batch size in $\{8, 16, 32\}$, the warmup ratio in $\{0, 0.2\}$, and the random seed in $\{12, 42\}$[3] on every target task.[4] That is, for each intermediate-target task pair, we conduct $3 \times 3 \times 2 \times 2 = 36$ experiments to study STILT's stability. We follow the preprocessing of previous work. Due to some nuances in the setup and implementation, we first compare with Pruksachatkun et al. (2020) on the overlapped experiments in Table 2, showing that our results are consistent with theirs.

4.2 Results

Figure 2 shows the experimental results on all the target tasks, where we use violinplot[5] to demonstrate results of all hyperparameters. Each subplot contains four methods (light blue *violins*):

- **None**: not using any intermediate task, i.e., the standard, vanilla RoBERTa fine-tuning.

- **HellaSwag**: using HellaSwag as the intermediate task.

- **HellaSwag-p**: using the first proposed baseline, which ablates HellaSwag's premises.

- **Syn_GPT2**: using the second proposed intermediate task, which is synthesized by GPT2.

We observe that **HellaSwag** does have generally[6] positive effects compared with **None**, including enhancing the best performance and significantly reducing the degenerate runs on the various target tasks. To study what RoBERTa learns after fine-tuning on HellaSwag, we first test if it learns to select the endings according to the premises by removing all the premises in HellaSwag's dev set. The moderate drop in performance, from 84.8% to 65.0%, where random guessing is only 25%, suggests that to some extent, it uses unwanted features in the machine-generated endings to make predictions. Also, its zero-shot performance on the dev set of Synthesis$_{GPT2}$ is as high as 75.3%. Thus, it is in doubt whether we can attribute **HellaSwag**'s improvement over **None** to offering RoBERTa commonsense reasoning skills.

Meanwhile, the proposed baselines show a competitively positive effect across all target tasks, including those in specific domains, such as SocialIQA (commonsense) and MedNLI (medical). As our simple baselines do not contain knowledge in these fields or other linguistic skills[7] besides real-fake discrimination, the overall improvement requires a careful rethinking of why STILT works.

5 Analysis

In this section, we further study different factors that may influence the effectiveness of STILT.

5.1 Intermediate Tasks

While we demonstrate the efficacy of the two simple baselines in the previous section, here we in-

---

[3]These two are the seeds recommended by Dodge et al. (2020) and used by `Huggingface` in default, respectively.
[4]We use `Huggingface transformers` toolkit.
[5]`matplotlib.axes.Axes.violinplot`

[6]Note that *generally* does not mean *universally* works well.
[7]One could argue that the two baselines still include some *reasoning* skills, which may lead to a long debate, depending on the definition of *reasoning*. Another debatable issue is that whether our baselines are really *simple*.
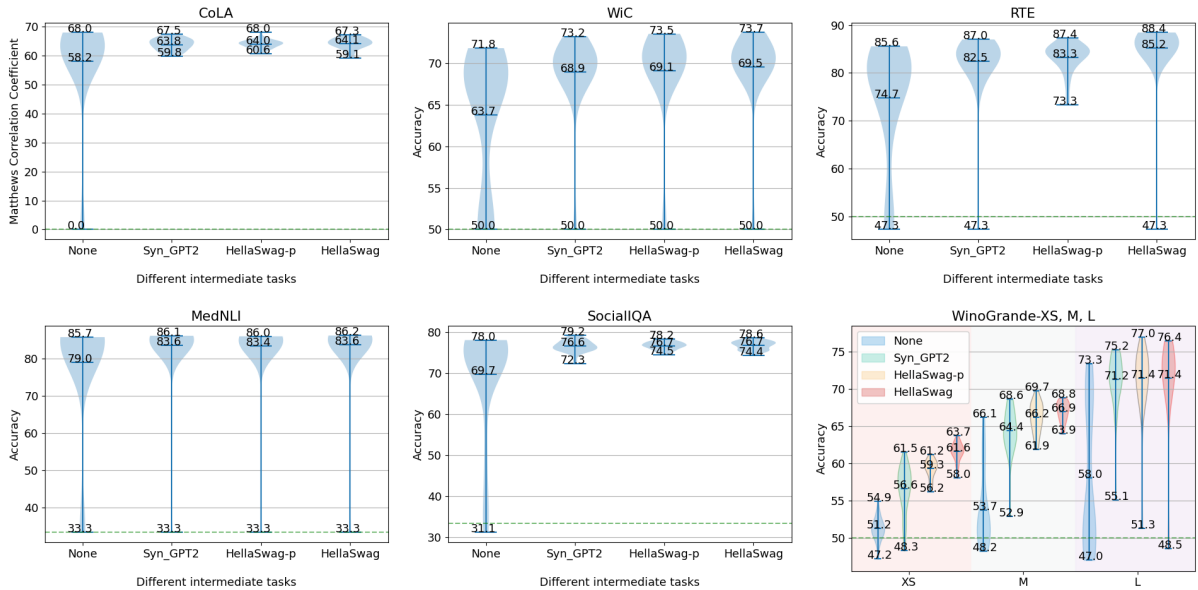
Figure 2: Results across different target tasks. Each *violin* contains 36 hyperparameter trials of an intermediate-target pair, where the 3 annotated values correspond to the min, mean, and best performance within. The scores are $100\times$ for better visualization. The green dash line shows the random guessing performance on each task.

vestigate if other true-false intermediate tasks also work, including 1) CoLA, a task of grammaticality, and 2) Hella-sh, a dummy task created by shuffling the words in the fake endings of HellaSwag. Table 3 shows that they both contribute *negatively* (red-colored) in many cases. We suppose that a true-false intermediate task works widely when it provides RoBERTa general, high-level overlaps with target tasks. For example, focusing on summarizing the semantic-level information to the [CLS] token so that the classifier atop can make decisions easier since RoBERTa's pretraining only applies mask language modeling (Liu et al., 2019). On the contrary, leaning toward learning specific rules or skills such as linguistic acceptability (CoLA) cannot benefit diverse target tasks. In this paper, we raise the need for rethinking by showing the different efficacy of some related intermediate tasks and leave it for future work to provide a more convincing explanation on why or why not they work.

5.2 Target Training Size

Previous work (Phang et al., 2018; Pruksachatkun et al., 2020; Vu et al., 2020) has found that STILT works especially well on limited labeled target tasks. Here, we study the impact of target-task size on WinoGrande$_{XS,M,L}$ (Sakaguchi et al., 2019), since the dataset contains different training sizes. Figure 2 shows that RoBERTa can barely learn from the 160 training data of WinoGrande$_{XS}$ with

vanilla fine-tuning. We observed that the training loss was about constant during the entire fine-tuning phase. At this point, introducing the intermediate tasks notably enhances the model's stability and its best performance. However, when we increase the training size, the improvements on the best performance dwindle. On the other hand, the average-performance improvements remain significant, mainly because RoBERTa still suffers from a few degenerate runs. This section shows that the target-task size has a strong influence on STILT's effectiveness, especially when the pretrained model struggles to learn from the sparse training signals, where STILT can help converge better.

5.3 Intermediate Training Size

Finally, we study the influence of the intermediate-task training size on three target tasks: WinoGrande$_M$, RTE, and WiC. Figure 3 shows that fine-tuning on a few data (2000) of the intermediate task, **HellaSwag-p**, already leads to noticeable improvements on these target tasks. We suggest rethinking what the intermediate task provides under such a few-resource circumstance. We believe that instead of providing RoBERTa more linguistic knowledge related to the target tasks, the intermediate task offers some high-level guidance to bridge the gaps between the pretraining and fine-tuning phase.
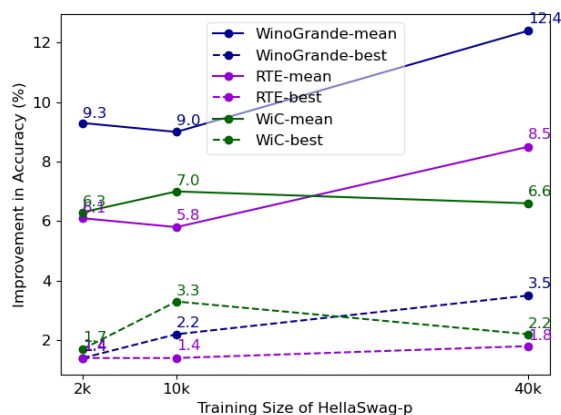
709

Figure 3: The influence of the intermediate-task training size (2k, 10k, 40k). We run a hyperparameter sweep for each size and report the best (solid lines) and mean (dashed lines) improvements over **None**.

6 Conclusion

We discover that a generally beneficial intermediate task to RoBERTa can be as simple as a synthetic real-fake discrimination task, and provide observations on different factors that influence STILT's best and mean effectiveness. Therefore, we suggest rethinking why intermediate-task fine-tuning works, particularly under low-resource settings.

7 Acknowledgments

We would like to thank Ting-Rui Chiang, Ta-Chung Chi, and Jian-Jia Su for in-depth discussions. We are thankful to the anonymous reviewers for their insightful comments on the paper.

References

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimiciii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

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*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint arXiv:2005.13013*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273.

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 models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

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

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Alexandre Tamborrino, Nicola Pellicano, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. *arXiv preprint arXiv:2004.14074*.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. *arXiv preprint arXiv:2004.14975*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, et al. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

A Datasets Details

A.1 Intermediate Tasks

In this section, we include more details about the HellaSwag dataset and how we construct our second baseline, Synthesis$_{GPT2}$.

HellaSwag is a commonsense reasoning task that tests a model's ability to choose the most plausible continuation of an event. The premises and the correct options are derived from ActivityNet Captions (Krishna et al., 2017) and WikiHow to include commonsense knowledge, while its negative options are GPT-generated. Adversarial Filtering (Zellers et al., 2018, 2019) are applied against BERT to create more challenging options. Figure 4 illustrates an example in HellaSwag.

In our Synthesis$_{GPT2}$ baseline, we mimicked HellaSwag's creation process to build a sentence continuation task *without* commonsense knowledge for the ablation study. We chose Wikipedia as the source of the premises and correct answers, while the negative options are generated by GPT2-medium. We did not apply Adversarial Filtering.

Please note that we only run hyperparameter sweeps on the target tasks, not on the intermediate tasks, as we believe that a handy intermediate task should not require the resource-consuming hyperparameter search. For intermediate tasks, we simply use the central hyperparameters in the span (learn rate=1e-5, batch size=16, random seed=42).

A.2 Target Tasks

Here, we make a brief introduction about the target tasks we evaluate on.

CoLA The Corpus of Linguistic Acceptability (Warstadt et al., 2019) is a binary classification task containing sentences labeled as either grammatical or ungrammatical. Performance on CoLA is reported in Matthew's correlation coefficient (MCC). CoLA is a task in GLUE benchmark.

RTE Recognizing Textual Entailment (Dagan et al., 2005) is a textual entailment task. We use the binary sentence classification version of the task. Each example contains a premise and a hypothesis sentence. Performance on RTE is reported in accuracy. RTE is a task in both GLUE and SuperGLUE benchmarks.

WiC Word-in-Context (Pilehvar and Camacho-Collados, 2019) is a binary classification word sense disambiguation task. Each example consists of two sentences and a polysemous word that appears in both sentences, asking whether the word

Figure 4: An example in HellaSwag dataset.

has the same sense in both. Performance on WiC is reported in accuracy. WiC is a task in SuperGLUE benchmark.

WinoGrande An Adversarial Winograd Schema Challenge at Scale (Sakaguchi et al., 2019) is a commonsense coreference resolution task, which improves the scale and the hardness of WSC (Levesque et al., 2012). Each example contains one sentence with a blank and two options to be filled in. We follow Sakaguchi et al. (2019) when preprocessing its input for RoBERTa. For example, an instance is formatted as *" [CLS] The trophy doesn't fit into the brown suitcase because the [SEP] _ is too large. [SEP]"*, where the blank is filled with either option1 or option2. This dataset includes different training scales, where we use the XS, M, and L versions in this paper. Performance on WinoGrande is reported in accuracy.

SocialIQA (Sap et al., 2019) is a multiple-choice commonsense question-answering dataset. Each example consists of a context, a question, and three options. The task is about commonsense reasoning that requires emotional and social intelligence in everyday situations. Performance on SocialIQA is reported in accuracy.

MedNLI (Romanov and Shivade, 2018) is a natural language inference dataset for the clinical domain, which is annotated by doctors and grounded in the medical history of patients. The premise sentences are from MIMIC-III (Johnson et al., 2016). The label classes are *entailment*, *contradiction*, and *neutral*. Performance on MedNLI is reported in accuracy. MedNLI is a task in PhysioNet (Goldberger et al., 2000).

B Future Work

Our work raises the need for rethinking why intermediate fine-tuning works. We found that in some target tasks, STILT's efficacy seems to be correlated with the phenomenon of degenerate fine-tuning runs (Devlin et al., 2019; Dodge et al., 2020;

| Intermediate | Zero-shot Acc. (%) | |
	SocialIQA	WinoGrande
None	35.0	52.3
Syn_GPT2	43.2	54.3
HellaSwag-p	44.0	56.4

Table 4: Zero-shot performance on two target tasks: SocialIQA and WinoGrande.

Mosbach et al., 2020). Thus, more research in degenerate runs may help us better understand how STILT works.

Unfortunately, we are unable to provide a convincing explanation on why and how our simple baselines work across target tasks in various domains. We suspect that after the large-scale pretraining, RoBERTa-large has already learned a certain amount of knowledge required in the downstream target tasks and that our proposed intermediate tasks work well as they help RoBERTa bridge the gaps between the pretraining and fine-tuning phases. For example, they probably help summarize the semantic-level information to the [CLS] token so that the classifier atop can make decisions easier since RoBERTa's pretraining only applies mask language modeling. However, this is just an unverified hypothesis.

We conduct an experiment related to our hypothesis. We evaluate our two baselines on the dev sets of SocialIQA and WinoGrande without fine-tuning on their training sets. We can apply such a zero-shot setting as our baselines share the same model architecture, RobertaForMultipleChoice[8], with WinoGrande and SocialIQA.[9] The results in Table 4 show that our simple intermediate fine-tuning methods, Syn_GPT2 and HellaSwag-p, have better performance over pretrained RoBERTa (None), although they can hardly provide RoBERTa with the commonsense knowledge required in SocialIQA and WinoGrande. Where does the improvement come from? Could it give credence to our hypothesis that bridging the gap between pretraining and downstream tasks?

Besides, we acknowledge that one could argue that the two baselines still include some *reasoning* skills, depending on the definition of *reasoning*.

[8] https://huggingface.co/transformers/model_doc/roberta.html#robertaformultiplechoice

[9] Similarly, we cannot conduct such experiments on other target tasks as they do not share the same architecture.

We believe that after the research community formulates clear notions and definitions on *reasoning* and *common sense*, we can have a better understanding of STILT. Similarly, another debatable issue is that whether our baselines are really *simple*. In this paper, we only meant to show that proposed strong baselines are not heavily human curated and unintuitively work well.

We leave it for future work to better understand when, why, and how STILT can help what target task. We believe that recent work in pretrained language models' transferability (Vu et al., 2020; Tamkin et al., 2020; Zhang et al., 2020; Chung et al., 2020) can provide some insights into these questions.

C Implementation Details

All our models are based on HuggingFace's transformers Pytorch toolkit. We use RobertaForSequenceClassification class for RTE, CoLA, MedNLI, and WiC; and use RobertaForMultipleChoice for WinoGrande, SocialIQA, HellaSwag, and our two baselines. The RoBERTa-large model contains 24-layer, 1024-hidden, and 16-heads, with $\sim 350M$ parameters totally.