Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work?

Yada Pruksachatkun[1*] **Jason Phang**[1*] **Haokun Liu**[1*] **Phu Mon Htut**[1*]
Xiaoyi Zhang[1] **Richard Yuanzhe Pang**[1] **Clara Vania**[1] **Katharina Kann**[2]
Samuel R. Bowman[1]
[1]New York University
[2]University of Colorado Boulder
{yp913,bowman}@nyu.edu

Abstract

While pretrained models such as BERT have shown large gains across natural language understanding tasks, their performance can be improved by further training the model on a data-rich *intermediate task*, before fine-tuning it on a target task. However, it is still poorly understood when and why intermediate-task training is beneficial for a given target task. To investigate this, we perform a large-scale study on the pretrained RoBERTa model with 110 intermediate–target task combinations. We further evaluate all trained models with 25 *probing tasks* meant to reveal the specific skills that drive transfer. We observe that intermediate tasks requiring high-level inference and reasoning abilities tend to work best. We also observe that target task performance is strongly correlated with higher-level abilities such as coreference resolution. However, we fail to observe more granular correlations between probing and target task performance, highlighting the need for further work on broad-coverage probing benchmarks. We also observe evidence that the forgetting of knowledge learned during pretraining may limit our analysis, highlighting the need for further work on transfer learning *methods* in these settings.

1 Introduction

Unsupervised pretraining—e.g., BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019b)—has recently pushed the state of the art on many natural language understanding tasks. One method of further improving pretrained models that has been shown to be broadly helpful is to first fine-tune a pretrained model on an *intermediate* task, before fine-tuning again on the target task of interest (Phang et al., 2018; Wang et al., 2019a; Clark et al., 2019a; Sap et al., 2019), also referred to as STILTs. However, this approach does not always improve target task performance, and it is unclear under what conditions it does.

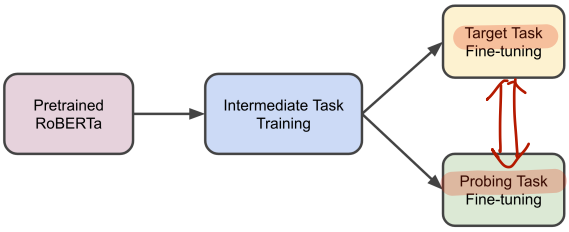


Figure 1: Our experimental pipeline with intermediate-task transfer learning and subsequent fine-tuning on target and probing tasks.

This paper offers a large-scale empirical study aimed at addressing this open question. We perform a broad survey of intermediate and target task pairs, following an experimental pipeline similar to Phang et al. (2018) and Wang et al. (2019a). This differs from previous work in that we use a larger and more diverse set of intermediate and target tasks, introduce additional analysis-oriented probing tasks, and use a better-performing base model RoBERTa (Liu et al., 2019b). We aim to answer the following specific questions:

- What kind of tasks tend to make good intermediate tasks across a wide variety of target tasks?
- Which linguistic skills does a model learn from intermediate-task training?
- Which skills learned from intermediate tasks help the model succeed on which target tasks?

The first question is the most straightforward: it can be answered by a sufficiently exhaustive search over possible intermediate–target task pairs. The second and third questions address the *why* rather than the *when*, and differ in a crucial detail: A

*Equal contribution.

model might learn skills by training on an intermediate task, but those skills might not help it to succeed on a target task.

Our search for intermediate tasks focuses on natural language understanding tasks in English. In particular, we run our experiments on 11 intermediate tasks and 10 target tasks, which results in a total of 110 intermediate–target task pairs. We use 25 *probing tasks*—tasks that each target a narrowly defined model behavior or linguistic phenomenon—to shed light on which skills are learned from each intermediate task.

Our findings include the following: (i) Natural language inference tasks as well as QA tasks which involve commonsense reasoning are generally useful as intermediate tasks. (ii) SocialIQA and QQP as intermediate tasks are not helpful as a means to teach the skills captured by our probing tasks, while finetuning first on MNLI and CosmosQA result in an increase in all skills. (iii) While a model's ability to learn skills relating to input-noising correlate with target task performance, low-level skills such as knowledge of a sentence's raw content preservation skills and ability to detect various attributes of input sentences such as tense of main verb and sentence length are less correlated with target task performance. This suggests that a model's ability to do well on the masked language modelling (MLM) task is important for downstream performance. Furthermore, we conjecture that a portion of our analysis is affected by catastrophic forgetting of knowledge learned during pretraining.

2 Methods

2.1 Experimental Pipeline

Our experimental pipeline (Figure 1) consists of two steps, starting with a pretrained model: *intermediate-task training*, and *fine-tuning* on a *target* or *probing* task.

Intermediate Task Training We fine-tune RoBERTa on each intermediate task. The training procedure follows the standard procedure of fine-tuning a pretrained model on a target task, as described in Devlin et al. (2019). We opt for single intermediate-task training as opposed to multi-task training (cf. Liu et al., 2019a) to isolate the effect of skills learned from individual intermediate tasks.

Target and Probing Task Fine-Tuning After intermediate-task training, we fine-tune our models on each target and probing task individually. Target tasks are tasks of interest to the general community, spanning various facets of natural language, domains, and sources. Probing tasks, while potentially similar in data source to target tasks such as with CoLA, are designed to isolate the presence of particular linguistic capabilities or skills. For instance, solving the target task BoolQ (Clark et al., 2019a) may require various skills including coreference and commonsense reasoning, while probing tasks like the SentEval probing suite (Conneau et al., 2018) target specific syntactic and metadata-level phenomena such as subject-verb agreement and sentence length detection.

2.2 Tasks

Table 1 presents an overview of the intermediate and target tasks.

2.2.1 Intermediate Tasks

We curate a diverse set of tasks that either represent an especially large annotation effort or that have been shown to yield positive transfer in prior work. The resulting set of tasks cover question answering, commonsense reasoning, and natural language inference.

QAMR The Question–Answer Meaning Representations dataset (Michael et al., 2018) is a crowdsourced QA task consisting of question–answer pairs that correspond to predicate–argument relationships. It is derived from Wikinews and Wikipedia sentences. For example, if the sentence is "*Ada Lovelace was a computer scientist.*", a potential question is "*What is Ada's last name?*", with the answer being "*Lovelace.*"

CommonsenseQA CommonsenseQA (Talmor et al., 2019) is a multiple-choice QA task derived from ConceptNet (Speer et al., 2017) with the help of crowdworkers, that is designed to test a range of commonsense knowledge.

SciTail SciTail (Khot et al., 2018) is a textual entailment task built from multiple-choice science questions from 4th grade and 8th grade exams, as well as crowdsourced questions (Welbl et al., 2017). The task is to determine whether a hypothesis, which is constructed from a science question and its corresponding answer, is entailed or not (neutral) by the premise.

Cosmos QA Cosmos QA is a task for a commonsense-based reading comprehension task

	Name	|Train|	|Dev|	task	metrics	genre/source
Intermediate Tasks	CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
	SciTail	23,596	1,304	natural language inference	acc.	science exams
	Cosmos QA	25,588	3,000	question answering	acc.	blogs
	SocialIQA	33,410	1,954	question answering	acc.	crowdsourcing
	CCG	38,015	5,484	tagging	acc.	Wall Street Journal
	HellaSwag	39,905	10,042	sentence completion	acc.	video captions & Wikihow
	QA-SRL	44,837	7,895	question answering	F1/EM	Wikipedia
	SST-2	67,349	872	sentiment classification	acc.	movie reviews
	QAMR	73,561	27,535	question answering	F1/EM	Wikipedia
	QQP	363,846	40,430	paraphrase detection	acc./F1	Quora questions
	MNLI	392,702	20,000	natural language inference	acc.	fiction, letters, telephone speech
Target Tasks	CB	250	57	natural language inference	acc./F1	Wall Street Journal, fiction, dialogue
	COPA	400	100	question answering	acc.	blogs, photography encyclopedia
	WSC	554	104	coreference resolution	acc.	hand-crafted
	RTE	2,490	278	natural language inference	acc.	news, Wikipedia
	MultiRC	5,100	953	question answering	$F1_\alpha$/EM	crowd-sourced
	WiC	5,428	638	word sense disambiguation	acc.	WordNet, VerbNet, Wiktionary
	BoolQ	9,427	3,270	question answering	acc.	Google queries, Wikipedia
	CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
	Cosmos QA	25,588	3,000	question answering	acc.	blogs
	ReCoRD	100,730	10,000	question answering	F1/EM	news (CNN, Daily Mail)

Table 1: Overview of the intermediate tasks (top) and target tasks (bottom) in our experiments. EM is short for Exact Match. The F1 metrics for MultiRC is calculated over all answer-options.

formulated as multiple-choice questions (Huang et al., 2019). The questions concern the causes or effects of events that require reasoning not only based on the exact text spans in the context, but also wide-range abstractive commonsense reasoning. It differs from CommonsenseQA in that it focuses on causal and deductive commensense reasoning and that it requires reading comprehension over an auxiliary passage, rather than simply answering a freestanding question.

SocialIQA SocialIQA (Sap et al., 2019) is a task for multiple choice QA. It tests for reasoning surrounding emotional and social intelligence in everyday situations.

CCG CCGbank (Hockenmaier and Steedman, 2007) is a task that is a translation of the Penn Treebank into a corpus of Combinatory Categorial Grammar (CCG) derivations. We use the CCG supertagging task, which is the task of assigning tags to individual word tokens that jointly determine the parse of the sentence.

HellaSwag HellaSwag (Zellers et al., 2019) is a commonsense reasoning task that tests a model's ability to choose the most plausible continuation of a story. It is built using adversarial filtering (Zellers et al., 2018) with BERT to create challenging negative examples.

QA-SRL The question-answer driven semantic role labeling dataset (QA-SRL; He et al., 2015) for a QA task that is derived from a semantic role labeling task. Each example, which consists of a set of questions and answers, corresponds to a predicate-argument relationship in the sentence it is derived from. Unlike QAMR, which focuses on all words in the sentence, QA-SRL is specifically focused on verbs.

SST-2 The Stanford sentiment treebank (Socher et al., 2013) is a sentiment classification task based on movie reviews. We use the binary sentence classification version of the task.

QQP The Quora Question Pairs dataset[1] is constructed based on questions posted on the community question-answering website Quora. The task is to determine if two questions are semantically equivalent.

MNLI The Multi-Genre Natural Language Inference dataset (Williams et al., 2018) is a crowdsourced collection of sentence pairs with textual entailment annotations across a variety of genres.

2.2.2 Target Tasks

We use ten target tasks, eight of which are drawn from the SuperGLUE benchmark (Wang et al., 2019b). The tasks in the SuperGLUE benchmark

[1] http://data.quora.com/First-Quora-DatasetRelease-Question-Pairs

cover question answering, entailment, word sense disambiguation, and coreference resolution and have been shown to be easy for humans but difficult for models like BERT. Although we offer a brief description of the tasks below, we refer readers to the SuperGLUE paper for a more detailed description of the tasks.

CommitmentBank (**CB**; de Marneffe et al., 2019) is a three-class entailment task that consists of texts and an embedded clause that appears in each text, in which models must determine whether that embedded clause is entailed by the text. **Choice of Plausible Alternatives** (**COPA**; Roemmele et al., 2011) is a classification task that consists of premises and a question that asks for the cause or effect of each premise, in which models must correctly pick between two possible choices. **Winograd Schema Challenge** (**WSC**; Levesque et al., 2012) is a sentence-level commonsense reasoning task that consists of texts, a pronoun from each text, and a list of possible noun phrases from each text. The dataset has been designed such that world knowledge is required to determine which of the possible noun phrases is the correct referent to the pronoun. We use the SuperGLUE binary classification cast of the task, where each example consists of a text, a pronoun, and a noun phrase from the text, which models must classify as being coreferent to the pronoun or not. **Recognizing Textual Entailment** (**RTE**; Dagan et al., 2005, et seq) is a textual entailment task. **Multi-Sentence Reading Comprehension** (**MultiRC**; Khashabi et al., 2018) is a multi-hop QA task that consists of paragraphs, a question on each paragraph, and a list of possible answers, in which models must distinguish which of the possible answers are true and which are false. **Word-in-Context** (**WiC**; Pilehvar and Camacho-Collados, 2019) is a binary classification word sense disambiguation task. Examples consist of two text snippets, with a polysemous word that appears in both. Models must determine whether the same sense of the word is used in both contexts. **BoolQ** (Clark et al., 2019a) is a QA task that consists of passages and a yes/no question associated with each passage. **Reading Comprehension with Commonsense Reasoning** (**ReCoRD**; Zhang et al., 2018) is a multiple-choice QA task that consists of news articles. For each article, models are given a question about each article with one entity masked out and a list of possible entities from the article, and the goal is to correctly identify the masked entity out of the list.

Additionally, we use **CommonsenseQA** and **Cosmos QA** as target tasks, due to their unique combination of small dataset size and high level of difficulty for high-performing models like BERT from our set of intermediate tasks.

2.2.3 Probing Tasks

We use well-established datasets for our probing tasks, including the edge-probing suite from Tenney et al. (2019b), function word oriented tasks from Kim et al. (2019), and sentence-level probing datasets (SentEval; Conneau et al., 2018).

Acceptability Judgment Tasks This set of binary classifications tasks was designed to investigate if a model can judge the grammatical acceptability of a sentence. We use the following five datasets: **AJ-CoLA** is a task that tests for a model's understanding of general grammaticality using the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019b), which is drawn from 22 theoretical linguistics publications. The other tasks concern the behaviors of specific classes of function words, using the dataset by Kim et al. (2019): **AJ-WH** is a task that tests a model's ability to detect if a wh-word in a sentence has been swapped with another wh-word, which tests a model's ability to identify the antecedent associated with the wh-word. **AJ-Def** is a task that tests a model's ability to detect if the definite/indefinite articles in a given sentence have been swapped. **AJ-Coord** is a task that tests a model's ability to detect if a coordinating conjunction has been swapped, which tests a model's ability to understand how ideas in the various clauses relate to each other. **AJ-EOS** is a task that tests a model's ability to identify grammatical sentences without indicators such as punctuation marks and capitalization, and consists of grammatical text that are removed of punctuation.

Edge-Probing Tasks The edge probing (EP) tasks are a set of core NLP labeling tasks, collected by Tenney et al. (2019b) and cast into Boolean classification. These tasks focus on the syntactic and semantic relations between spans in a sentence. The first five tasks use the OntoNotes corpus (Hovy et al., 2006): **Part-of-Speech tagging** (**EP-POS**) is a task that tests a model's ability to predict the syntactic category (noun, verb, adjective, etc.) for each word in the sentence. **Named entity recognition** (**EP-NER**) is task that tests a model's abil-

ity to predict the category of an entity in a given span. **Semantic Role Labeling** (**EP-SRL**) is a task that tests a model's ability to assign a label to a given span of words that indicates its semantic role (agent, goal, etc.) in the sentence. **Coreference** (**EP-Coref**) is a task that tests a model's ability to classify if two spans of tokens refer to the same entity/event.

The other datasets can be broken down into both syntactic and semantic probing tasks. **Constituent labeling** (**EP-Const**) is a task that tests a model's ability to classify a non-terminal label for a span of tokens (e.g., noun phrase, verb phrase, etc.). **Dependency labeling** (**EP-UD**) is a task that tests a model on the functional relationship of one token relative to another. We use the English Web Treebank portion of Universal Dependencies 2.2 release (Silveira et al., 2014) for this task. **Semantic Proto-Role labeling** is a task that tests a model's ability to predict the fine-grained non-exclusive semantic attributes of a given span. Edge probing uses two datasets for SPR: SPR1 (**EP-SPR1**) (Teichert et al., 2017), derived from the Penn Treebank, and SPR2 (**EP-SPR2**) (Rudinger et al., 2018), derived from the English Web Treebank. **Relation classification** (**EP-Rel**) is a task that tests a model's ability to predict the relation between two entities. We use the SemEval 2010 Task 8 dataset (Hendrickx et al., 2009) for this task. For example, the relation between "Yeri" and "Korea" in "Yeri is from Korea" is ENTITY-ORIGIN. The **Definite Pronoun Resolution** dataset (Rahman and Ng, 2012) (**EP-DPR**) is a task that tests a model's ability to handle coreference, and differs from OntoNotes in that it focuses on difficult cases of definite pronouns.

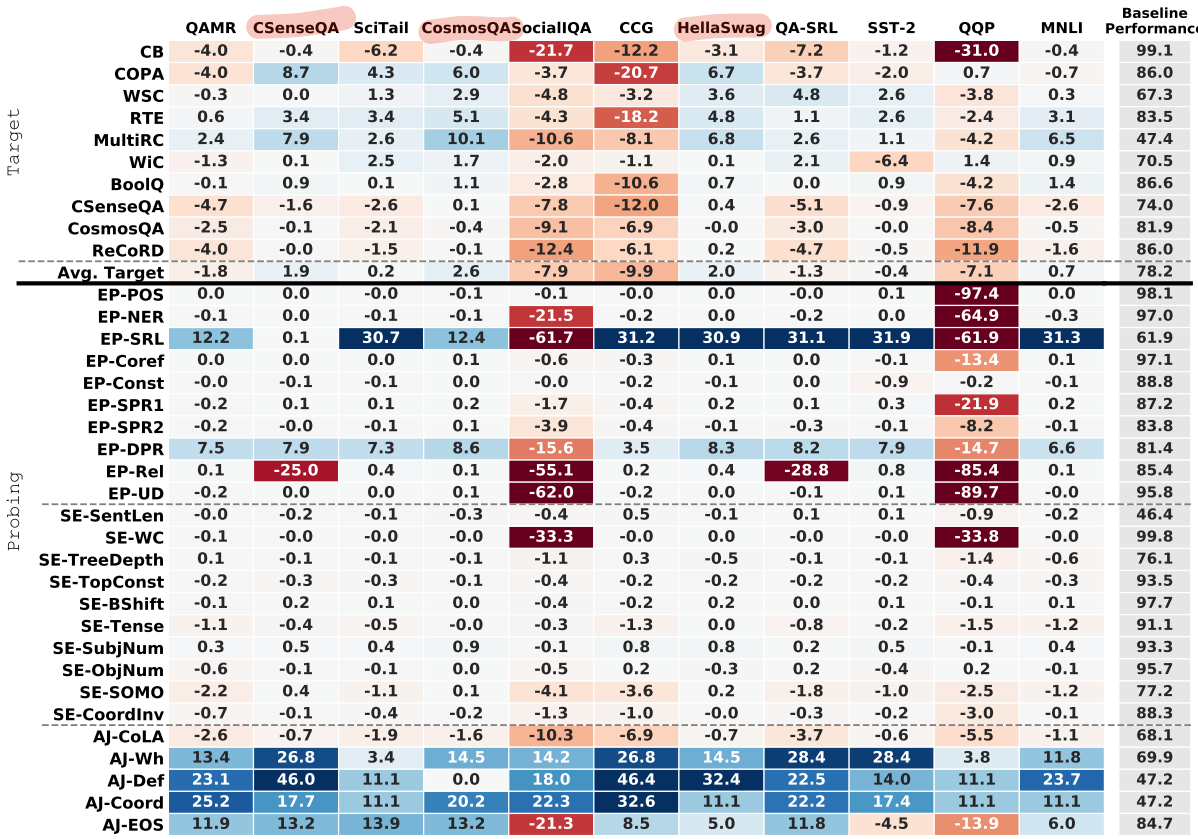
SentEval Tasks The SentEval probing tasks (SE) (Conneau et al., 2018) are cast in the form of single-sentence classification. **Sentence Length** (**SE-SentLen**) is a task that tests a model's ability to classify the length of a sentence. **Word Content** (**SE-WC**) is a task that tests a model's ability to identify which of a set of 1,000 potential words appear in a given sentence. **Tree Depth** (**SE-TreeDepth**) is a task that tests a model's ability to estimate the maximum depth of the constituency parse tree of the sentence. **Top Constituents** (**SE-TopConst**) is a task that tests a model's ability to identify the high-level syntactic structure of the sentence by choosing among 20 constituent sequences (the 19 most common, plus an *other* category). **Bigram Shift** (**SE-BShift**) is a task that tests a model's ability to classify if two consecutive tokens in the same sentence have been reordered. **Coordination Inversion** (**SE-CoordInv**) is a task that tests a model's ability to identify if two coordinating clausal conjoints are swapped (*ex:* "he knew it, and he deserved no answer."). **Past-Present** (**SE-Tense**) is a task that tests a model's ability to classify the tense of the main verb of the sentence. **Subject Number** (**SE-SubjNum**) and **Object Number** (**SE-ObjNum**) are tasks that test a model's ability to classify whether the subject or direct object of the main clause is singular or plural. **Odd-Man-Out** (**SE-SOMO**) is a task that tests the model's ability to predict whether a sentence has had one of its content words randomly replaced with another word of the same part of speech.

3 Experiments

Training and Optimization We use the large-scale pretrained model RoBERTa$_{\text{Large}}$ in all experiments. For each intermediate, target, and probing task, we perform a hyperparameter sweep, varying the peak learning rate $\in \{2 \times 10^{-5}, 1 \times 10^{-5}, 5 \times 10^{-6}, 3 \times 10^{-6}\}$ and the dropout rate $\in \{0.2, 0.1\}$. After choosing the best learning rate and dropout rate, we apply the best configuration for each task for all runs. For each task, we use the batch size that maximizes GPU usage, and use a maximum sequence length of 256. Aside from these details, we follow the RoBERTa paper for all other training hyperparameters. We use NVIDIA P40 GPUs for our experiments.

A complete pipeline with one intermediate task works as follows: First, we fine-tune RoBERTa on the intermediate task. We then fine-tune copies of the resulting model separately on each of the 10 target tasks and 25 probing tasks and test on their respective validation sets. We run the same pipeline three times for the 11 intermediate tasks, plus a set of baseline runs without intermediate training. This gives us $35 \times 12 \times 3 = 1260$ observations.

We train our models using the Adam optimizer (Kingma and Ba, 2015) with linear decay and early stopping. We run training for a maximum of 10 epochs when more than 1,500 training examples are available, and 40 epochs otherwise to ensure models are sufficiently trained on small datasets. We use the `jiant` (Wang et al., 2019c) NLP toolkit, based on PyTorch (Paszke et al., 2019), Hugging Face Transformers (Wolf et al., 2019), and AllenNLP (Gardner et al., 2017), for all of our



		QAMR	CSenseQA	SciTail	CosmosQA	SocialIQA	CCG	HellaSwag	QA-SRL	SST-2	QQP	MNLI	Baseline Performance
Target	CB	-4.0	-0.4	-6.2	-0.4	-21.7	-12.2	-3.1	-7.2	-1.2	-31.0	-0.4	99.1
	COPA	-4.0	8.7	4.3	6.0	-3.7	-20.7	6.7	-3.7	-2.0	0.7	-0.7	86.0
	WSC	-0.3	0.0	1.3	2.9	-4.8	-3.2	3.6	4.8	2.6	-3.8	0.3	67.3
	RTE	0.6	3.4	3.4	5.1	-4.3	-18.2	4.8	1.1	2.6	-2.4	3.1	83.5
	MultiRC	2.4	7.9	2.6	10.1	-10.6	-8.1	6.8	2.6	1.1	-4.2	6.5	47.4
	WiC	-1.3	0.1	2.5	1.7	-2.0	-1.1	0.1	2.1	-6.4	1.4	0.9	70.5
	BoolQ	-0.1	0.9	0.1	1.1	-2.8	-10.6	0.7	0.0	0.9	-4.2	1.4	86.6
	CSenseQA	-4.7	-1.6	-2.6	0.1	-7.8	-12.0	0.4	-5.1	-0.9	-7.6	-2.6	74.0
	CosmosQA	-2.5	-0.1	-2.1	-0.4	-9.1	-6.9	-0.0	-3.0	-0.0	-8.4	-0.5	81.9
	ReCoRD	-4.0	-0.0	-1.5	-0.1	-12.4	-6.1	0.2	-4.7	-0.5	-11.9	-1.6	86.0
	Avg. Target	-1.8	1.9	0.2	2.6	-7.9	-9.9	2.0	-1.3	-0.4	-7.1	0.7	78.2
Probing	EP-POS	0.0	0.0	-0.0	-0.1	-0.1	-0.0	0.0	-0.0	0.1	-97.4	0.0	98.1
	EP-NER	-0.1	0.0	-0.1	-0.1	-21.5	-0.2	0.0	-0.2	0.0	-64.9	-0.3	97.0
	EP-SRL	12.2	0.1	30.7	12.4	-61.7	31.2	30.9	31.1	31.9	-61.9	31.3	61.9
	EP-Coref	0.0	0.0	0.0	0.1	-0.6	-0.3	0.1	0.0	-0.1	-13.4	0.1	97.1
	EP-Const	-0.0	-0.1	-0.1	0.0	-0.0	-0.2	-0.1	0.0	-0.9	-0.2	-0.1	88.8
	EP-SPR1	-0.2	0.1	0.1	0.2	-1.7	-0.4	0.2	0.1	0.3	-21.9	0.2	87.2
	EP-SPR2	-0.2	-0.0	-0.1	0.1	-3.9	-0.4	-0.1	-0.3	-0.1	-8.2	-0.1	83.8
	EP-DPR	7.5	7.9	7.3	8.6	-15.6	3.5	8.3	8.2	7.9	-14.7	6.6	81.4
	EP-Rel	0.1	-25.0	0.4	0.1	-55.1	0.2	0.4	-28.8	0.8	-85.4	0.1	85.4
	EP-UD	-0.2	0.0	0.0	0.1	-62.0	-0.2	0.0	-0.1	0.1	-89.7	-0.0	95.8
	SE-SentLen	-0.0	-0.2	-0.1	-0.3	-0.4	0.5	-0.1	0.1	0.1	-0.9	-0.2	46.4
	SE-WC	-0.1	-0.0	-0.0	-0.0	-33.3	-0.0	0.0	-0.0	-0.0	-33.8	-0.0	99.8
	SE-TreeDepth	0.1	-0.1	-0.1	-0.1	-1.1	0.3	-0.5	-0.1	-0.1	-1.4	-0.6	76.1
	SE-TopConst	-0.2	-0.3	-0.3	-0.1	-0.4	-0.2	-0.2	-0.2	-0.2	-0.4	-0.3	93.5
	SE-BShift	-0.1	0.2	0.1	0.0	-0.4	-0.2	0.2	0.0	0.1	-0.1	0.1	97.7
	SE-Tense	-1.1	-0.4	-0.5	-0.0	-0.3	-1.3	0.0	-0.8	-0.2	-1.5	-1.2	91.1
	SE-SubjNum	0.3	0.5	0.4	0.9	-0.1	0.8	0.8	0.2	0.5	-0.1	0.4	93.3
	SE-ObjNum	-0.6	-0.1	-0.1	0.0	-0.5	0.2	-0.3	0.2	-0.4	0.2	-0.1	95.7
	SE-SOMO	-2.2	0.4	-1.1	0.1	-4.1	-3.6	0.2	-1.8	-1.0	-2.5	-1.2	77.2
	SE-CoordInv	-0.7	-0.1	-0.4	-0.2	-1.3	-1.0	-0.0	-0.3	-0.2	-3.0	-0.1	88.3
	AJ-CoLA	-2.6	-0.7	-1.9	-1.6	-10.3	-6.9	-0.7	-3.7	-0.6	-5.5	-1.1	68.1
	AJ-Wh	13.4	26.8	3.4	14.5	14.2	26.8	14.5	28.4	28.4	3.8	11.8	69.9
	AJ-Def	23.1	46.0	11.1	0.0	18.0	46.4	32.4	22.5	14.0	11.1	23.7	47.2
	AJ-Coord	25.2	17.7	11.1	20.2	22.3	32.6	11.1	22.2	17.4	11.1	11.1	47.2
	AJ-EOS	11.9	13.2	13.9	13.2	-21.3	8.5	5.0	11.8	-4.5	-13.9	6.0	84.7

Figure 2: Transfer learning results between intermediate and target/probing tasks. Baselines (rightmost column) are models fine-tuned without intermediate-task training. Each cell shows the difference in performance (delta) between the baseline and model with intermediate-task training. We use the macro-average of each task's metrics as the reported performance. Refer to Table 1 for target task metrics.

experiments.

4 Results and Analysis

4.1 Investigating Transfer Performance

Figure 2 shows the differences in target and probing task performances (deltas) between the baselines and models trained with intermediate-task training, each averaged across three restarts. A positive delta indicates successful transfer.

Target Task Performance We define good intermediate tasks as ones that lead to positive transfer in target task performance. We observe that tasks that require complex reasoning and inference tend to make good intermediate tasks. These include MNLI and commonsense-oriented tasks such as CommonsenseQA, HellaSWAG, and Cosmos QA (with our poor performance with the similar SocialIQA serving as a suprising exception). SocialIQA, CCG, and QQP as intermediate tasks lead to negative transfer on all target tasks and the majority of probing tasks.

We investigate the role of dataset size in the intermediate tasks with downstream task performance by additionally running a set of experiments on varying amounts of data on five intermediate tasks, which is shown in the Appendix. We do not find differences in intermediate-task dataset size to have any substantial consistent impact on downstream target task performance.

In addition, we find that smaller target tasks such as RTE, BoolQ, MultiRC, WiC, WSC benefit the most from intermediate-task training.[2] There are no instances of positive transfer to CommitmentBank, since our baseline model achieves 100% accuracy.

Probing Task Performance Looking at the probing task performance, we find that intermediate-task training affects performance

[2]The deltas for experiments with the same intermediate and target tasks are not 0 as may be expected. This is because we perform both intermediate and target training phases in these cases, with reset optimizer states and stopping criteria in between intermediate and target training.

on low-level syntactic probing tasks uniformly across intermediate tasks; we observe little to no improvement for the SentEval probing tasks and higher improvement for acceptability judgment probing tasks, except for AJ-CoLA. This is also consistent with Phang et al. (2018), who find negative transfer with CoLA in their experiments.

Variation across Intermediate Tasks There is variable performance across higher-level syntactic or semantic tasks such as the Edge-Probing and SentEval tasks. SocialIQA and QQP have negative transfer for most of the Edge-Probing tasks, while CosmosQA and QA-SRL see drops in performance only for EP-Rel. While we do see that intermediate-task trained models improve performance on EP-SRL and EP-DPR across the board, there is little to no gain in SentEval probing tasks from any intermediate tasks. Additionally, tasks that increase performance in the most number of probing tasks perform well as intermediate tasks.

Degenerate Runs We find that the model may not exceed chance performance in some training runs. This mostly affects the baseline (no intermediate training) runs on the acceptability judgment probing tasks, excluding AJ-CoLA, which all have very small training sets. We include these degenerate runs in our analysis to reflect this phenomenon. Consistent with Phang et al. (2018), we find that intermediate-task training reduces the likelihood of degenerate runs, leading to ostensibly positive transfer results on those four acceptability judgment tasks across most intermediate tasks. On the other hand, extremely negative transfer from intermediate-task training can also result in a higher frequency of degenerate runs in downstream tasks, as we observe in the cases of using QQP and SocialIQA as intermediate tasks. We also observe a number of degenerate runs on the EP-SRL task as well as the EP-Rel task. These degenerate runs decrease positive transfer in probing tasks, such as with SocialIQA and QQP probing performance, and also decrease the average amount of positive transfer we see in target task performance.

4.2 Correlation Between Probing and Target Task Performance

Next, we investigate the relationship between target and probing tasks in an attempt to understand *why* certain intermediate-task models perform better on certain target tasks.

We use probing task performance as an indicator of the acquisition of particular language skills. We compute the Spearman correlation between probing-task and target-task performances across training on different intermediate tasks and multiple restarts, as shown in Figure 3. We test for statistical significance at $p = 0.05$ and apply Holm-Bonferroni correction for multiple testing. We omit correlations that are not statistically significant. We opt for Spearman and not Pearson correlation because of the wide variety of metrics used for the different tasks.[3]

We find that acceptability judgment probing task performance is generally uncorrelated with the target task performance, except for AJ-CoLA. Similarly, many of the SentEval tasks do not correlate with the target tasks, except for Bigram Shift (SE-BShift), Odd-Man-Out (SE-SOMO) and Coordination Inversion (SE-CoordInv). These three tasks are input noising tasks—tasks where a model has to predict if a given input sentence has been randomly modified—which are, by far, the most similar tasks we study to the masked language modeling task that is used for training RoBERTa. This may explain the strong correlation with the performance of the target tasks.

We also find that some of these strong correlations, such as with SE-SOMO and SE-CoordInv, are almost entirely driven by variation in the degree of negative transfer, rather than any positive transfer. Intuitively, fine-tuning RoBERTa on an intermediate task can cause the model to forget some of its ability to perform the MLM task. Thus, a future direction for potential improvement for intermediate-task training may be integrating the MLM objective into intermediate-task training or bounding network parameter changes to reduce *catastrophic forgetting* (Kirkpatrick et al., 2016; Chen et al., 2019).

Interestingly, while intermediate tasks such as SocialIQA, CCG and QQP, which show negative transfer on target tasks, tend to have negative transfer on these three probing tasks, the intermediate tasks with positive transfer, such as CommonsenseQA tasks and MNLI, do not appear to adversely affect the performance on these probing tasks. This asymmetric impact may indicate that, beyond the similarity of intermediate and target tasks, avoiding catastrophic forgetting of pretrain-

[3]Full correlation tables across all target and probing tasks with both Spearman and Pearson correlations can be found in the Appendix.

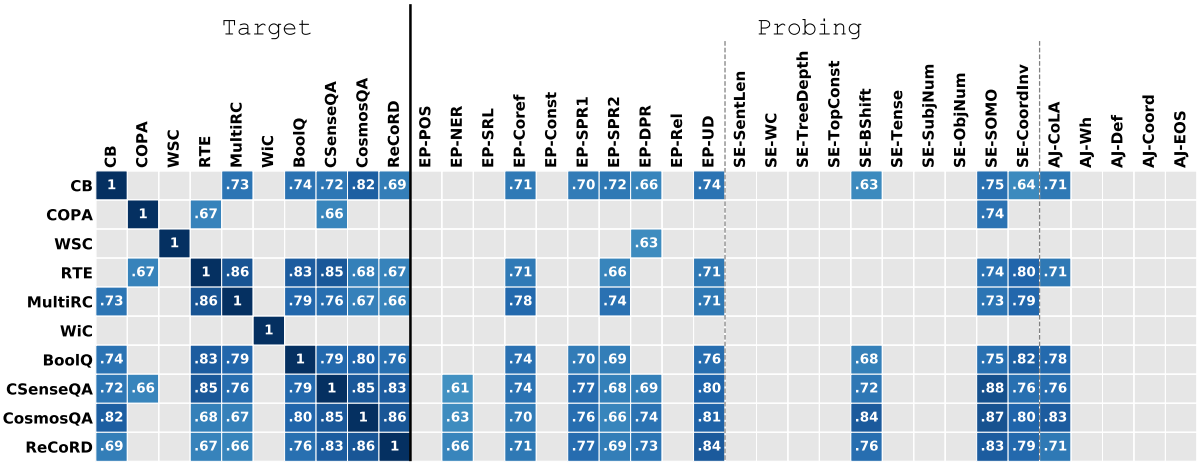


Figure 3: Correlations between probing and target task performances. Each cell contains the Spearman correlation between probing-task and target-task performances across training on different intermediate tasks and random restarts. We test for statistical significance at $p = 0.05$ with Holm-Bonferroni correction, and omit the correlations that are not statistically significant.

ing is critical to successful intermediate-task transfer.

The remaining SentEval probing tasks have similar delta values (Figure 2), which may indicate that there is insufficient variation among transfer performance to derive significant correlations. Among the edge-probing tasks, the more semantic tasks such as coreference (EP-Coref and EP-DPR), semantic proto-role labeling (EP-SPR1 and EP-SPR2), and dependency labeling (EP-Rel) show the highest correlations with our target tasks. As our set of target tasks is also oriented towards semantics and reasoning, this is to be expected.

On the other hand, among the target tasks, we find that ReCoRD, CommonsenseQA and Cosmos QA—all commonsense-oriented tasks—exhibit both high correlations with each other as well as a similar set of correlations with the probing tasks. Similarly, BoolQ, MultiRC, and RTE correlate strongly with each other and have similar patterns of probing-task performance.

5 Related Work

Within the paradigm of training large pre-trained Transformer language representations via intermediate-stage training before fine-tuning on a target task, positive transfer has been shown in both sequential task-to-task (Phang et al., 2018) and multi-task-to-task (Liu et al., 2019a; Raffel et al., 2019) formats. Wang et al. (2019a) perform an extensive study on transfer with BERT, finding language modeling and NLI tasks to be among the most beneficial tasks for improving target-task performance. Talmor and Berant (2019) perform a similar cross-task transfer study on reading comprehension datasets, finding similar positive transfer in most cases, with the biggest gains stemming from a combination of multiple QA datasets. Our work consists of a larger, more diverse, set of intermediate task–target task pairs. We also use probing tasks to shed light on the skills learned by the intermediate tasks.

Among the prior work on predicting transfer performance, Bingel and Søgaard (2017) is the most similar to ours. They do a regression analysis that predicts target-task performance on the basis of various features of the source and target tasks and task pairs. They focus on a multi-task training setting without self-supervised pretraining, as opposed to our single-intermediate task, three-step procedure.

Similar work (Lin et al., 2019b) has been done on cross-lingual transfer—the analogous challenge of transferring learned knowledge from a high-resource to a low-resource language.

Many recent works have attempted to understand the knowledge and linguistic skills BERT learns, for instance by analyzing the language model surprisal for subject–verb agreements (Goldberg, 2018), identifying specific knowledge or phenomena encapsulated in the representations learned by BERT using probing tasks (Tenney et al., 2019b,a; Warstadt et al., 2019a; Lin et al., 2019a; Hewitt and Manning, 2019; Jawahar et al., 2019), analyzing the attention heads of BERT (Clark et al., 2019b;

Coenen et al., 2019; Lin et al., 2019a; Htut et al., 2019), and testing the linguistic generalizations of BERT across runs (McCoy et al., 2019). However, relatively little work has been done to analyze fine-tuned BERT-style models (Wang et al., 2019a; Warstadt et al., 2019a).

6 Conclusion and Future Work

This paper presents a large-scale study on when and why intermediate-task training works with pretrained models. We perform experiments on RoBERTa with a total of 110 pairs of intermediate and target tasks, and perform an analysis using 25 probing tasks, covering different semantic and syntactic phenomena. Most directly, we observe that tasks like Cosmos QA and HellaSwag, which require complex reasoning and inference, tend to work best as intermediate tasks.

Looking to our probing analysis, intermediate tasks that help RoBERTa improve across the board show the most positive transfer in downstream tasks. However, it is difficult to draw definite conclusions about the specific skills that drive positive transfer. Intermediate-task training may help improve the handling of syntax, but there is little to no correlation between target-task and probing-task performance for these skills. Probes for higher-level semantic abilities tend to have a higher correlation with the target-task performance, but these results are too diffuse to yield more specific conclusions. Future work in this area would benefit greatly from improvements to both the breadth and depth of available probing tasks.

We also observe a worryingly high correlation between target-task performance and the two probing tasks which most closely resemble RoBERTa's masked language modeling pretraining objective. Thus, the results of our intermediate-task training analysis may be driven in part by *forgetting* of knowledge acquired during pretraining. Our results therefore suggest a need for further work on efficient transfer learning mechanisms.

Acknowledgments

This project has benefited from support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), by Samsung Research (under the project *Improving Deep Learning using Latent Structure*), by Intuit, Inc., and by NVIDIA Corporation (with the donation of a Titan V GPU).

References

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 1906–1916. Curran Associates, Inc.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019a. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. Unpublished manuscript available on arXiv.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. Unpublished manuscript available on arXiv.

Yoav Goldberg. 2018. Assessing BERT's syntactic abilities. Unpublished manuscript available on arXiv.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics*, 33(3):355–396.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? Unpublished manuscript available on arXiv.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

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

James Kirkpatrick, Razvan Pascanu, Neil C. 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. 2016. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the national academy of sciences (PNAS)*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019a. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019b. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. Unpublished manuscript available on arXiv.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. Unpublished manuscript available on arXiv.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

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. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. Unpublished manuscript available on arXiv.

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, Minneapolis, Minnesota. 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. Unpublished manuscript available on arXiv.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. 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 *2011 AAAI Spring Symposium Series*.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-Davidsonian Semantic Proto-role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4463, Hong Kong, China. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew R Gormley. 2017. Semantic proto-role labeling. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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, Florence, Italy. Association for Computational Linguistics.

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*.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. 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, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. SuperGLUE: A multi-task benchmark and analysis platform for natural language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Haokun Liu, Najoung Kim, Phu Mon Htut, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019c. `jiant` 1.2: A software toolkit for research on general-purpose text understanding models.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2870–2880, Hong Kong, China. Association for Computational Linguistics.

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

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. Unpublished manuscript available on arXiv.

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

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

4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. Unpublished manuscript available on arXiv.

A Correlation Between Probing and Target Task Performance

Figure 4 shows the correlation matrix using Spearman correlation and Figure 5 shows the matrix using Pearson correlation.

B Effect of Intermediate Task Size on Target Task Performance

Figure 6 shows the effect of dataset size on intermediate task training on downstream target task performance for five intermediate tasks, which were picked to maximize the variety of original intermediate task sizes and effectiveness in transfer learning abilities.

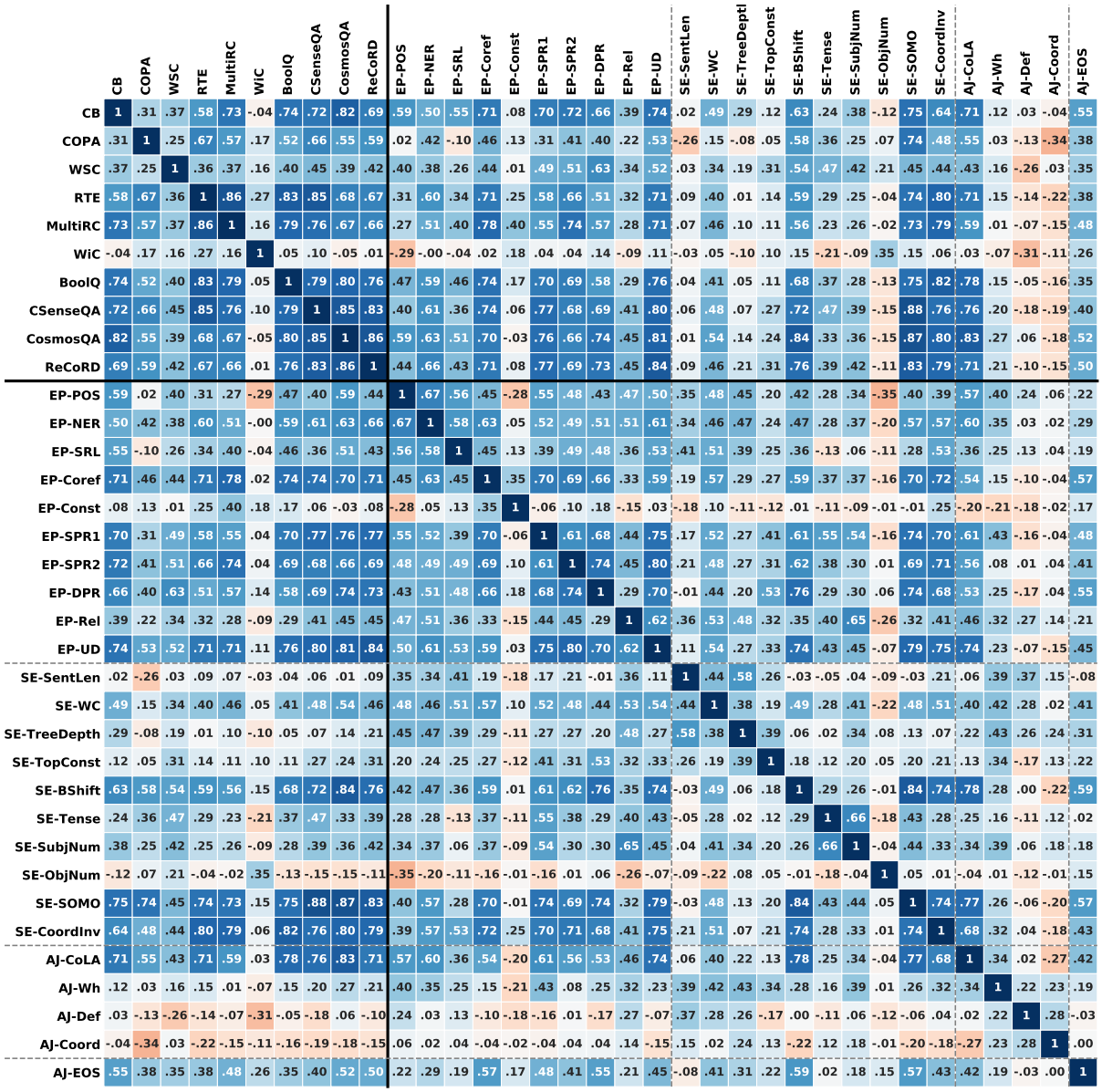


Figure 4: Correlations between probing and target task performances. Each cell contains the Spearman correlation between probing and target tasks performances across training on different intermediate tasks and random restarts.

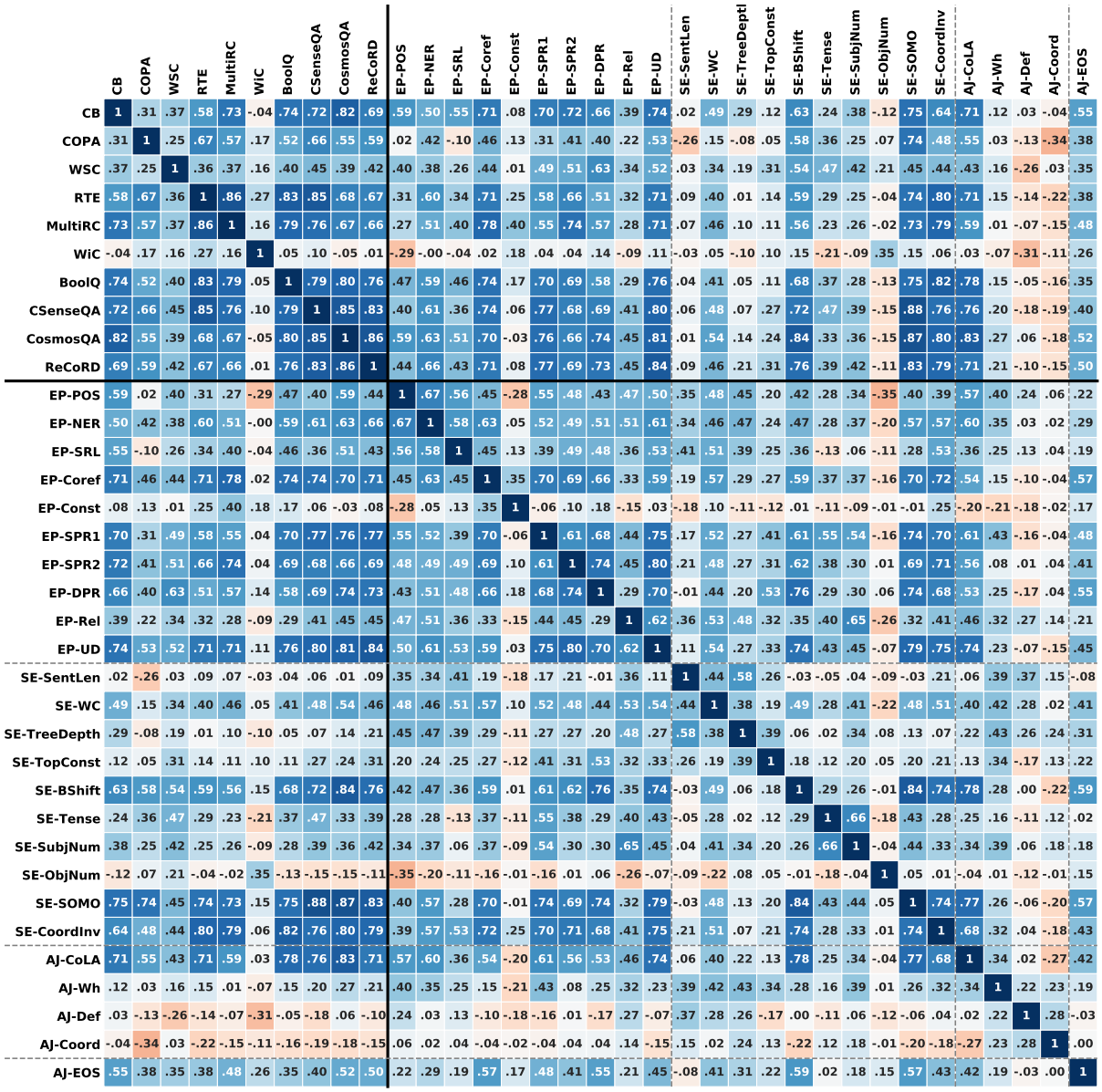


Figure 5: Correlations between probing and target task performances. Each cell contains the Pearson correlation between probing and target tasks performances across training on different intermediate tasks and random restarts.

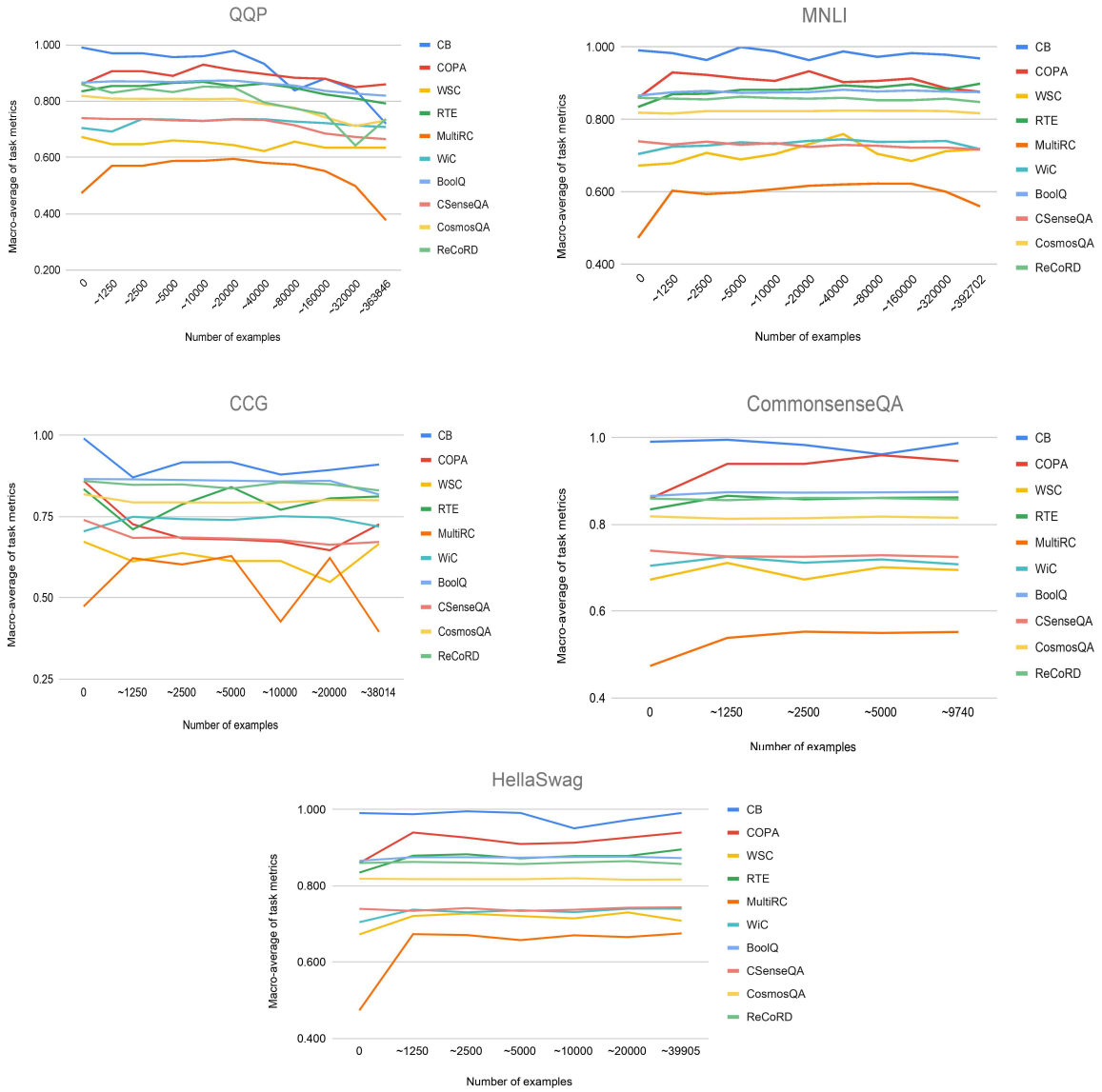


Figure 6: Results of experiments on impact of intermediate task data size on downstream target task performance. For each subfigure, we finetune RoBERTa over a variety of dataset size (sampled randomly from the dataset). We report the macro-average of each target task's performance metrics after finetuning on each dataset size split.