Intermediate-task training has been shown to substantially improve pretrained model performance on many language understanding tasks, at least in monolingual English settings. Here, we investigate whether English intermediate-task training is still helpful on non-English target tasks in a zero-shot cross-lingual setting. Using a set of 7 intermediate language understanding tasks, we evaluate intermediate-task transfer in a zero-shot cross-lingual setting on 9 target tasks from the XTREME benchmark. Intermediate-task training yields large improvements on the BUCC and Tatoeba tasks that use model representations directly without training, and moderate improvements on question-answering target tasks. Using SQuAD for intermediate training achieves the best results across target tasks, with an average improvement of 8.4 points on development sets. Selecting the best intermediate task model for each target task, we obtain a 6.1 point improvement over XLM-R Large on the XTREME benchmark, setting a new state of the art. Finally, we show that neither multitask intermediate-task training nor continuing multilingual MLM during intermediate-task training offer significant improvements.

1 Introduction

Zero-shot cross-lingual transfer involves training a language-encoding model on task data in one language, and evaluating the tuned model on the same task in other languages. This format evaluates the extent to which task-specific knowledge learned in one language generalizes across languages. Transformer models such as mBERT (Devlin et al., 2019a), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2019) that have been pretrained with a masked language modeling (MLM) objective on large corpora of multilingual data have shown remarkably strong results on zero-shot cross-lingual transfer, and show promise as a way of facilitating the construction of massively multilingual language technologies.

Intermediate-task training (STILTs; Phang et al., 2018) is the simple strategy of fine-tuning a pretrained model on a data-rich intermediate task, ideally related to the target task, before fine-tuning a second time on the downstream target task. Despite its simplicity, this two-phase training setup has been shown to be beneficial across a range of Transformer models and tasks (Wang et al., 2019a; Pruksachatkun et al., 2020), at least with English intermediate and target tasks.

In this work, we investigate whether intermediate training on English language tasks can also improve performance in a zero-shot cross-lingual transfer setting. Starting with a pretrained multilingual language encoder, we perform intermediate-task training on one or more English language tasks, then fine-tune on the target task in English, and finally evaluate zero-shot on the same task in other languages.

Intermediate-task training on English data introduces a potential issue: we train the pretrained multilingual model extensively on only English data before attempting to use it on non-English target task data, leaving open the possibility that the model will lose the knowledge of other languages that it acquired during pretraining (Kirkpatrick et al., 2017; Yogatama et al., 2019). To attempt to mitigate this, we experiment with mixing in multilingual MLM training updates during the intermediate-task training.

Concretely, we use the pretrained XLM-R (Conneau et al., 2019) as our multilingual language encoder as it currently achieves state-of-the-art performance on many zero-shot cross-lingual transfer tasks. We perform experiments on 9 target tasks from the recently introduced XTREME benchmark (Hu et al., 2020), which aims to evaluate zero-shot cross-lingual performance across diverse tar-
get tasks across up to 40 languages each. We investigate how intermediate-task training on 7 different tasks, including question answering, sentence tagging/completion, paraphrase detection, and natural language inference, impacts zero-shot cross-lingual transfer performance. We find:

- Applying intermediate-task training to BUCC and Tatoeba, the two sentence retrieval target tasks that have no training data, yields dramatic improvements with almost every intermediate training configuration.
- The question-answering target tasks show consistent smaller improvement with many intermediate tasks.
- Evaluating our best performing models for each target task on the XTREME benchmark yields an average improvement of 6.1 points, setting the state of the art as of writing.
- Neither continuing multilingual MLM training during intermediate-task training nor multi-task training meaningfully bolster transfer performance.

2 Approach

We propose a three-phase approach to training, illustrated in Figure 1: (i) we use a publicly available model pretrained on raw multilingual text using MLM; (ii) we perform intermediate-task training on one or more English intermediate tasks; (iii) we fine-tune the model on English target task data, before evaluating it on task data in multiple languages.

During phase (ii), all our intermediate tasks have English labeled data only. We experiment with performing intermediate-task training on single tasks individually and as well as a multi-task format. Intermediate tasks are described in detail in Section 2.1. We use target tasks from the recent XTREME benchmark, whose goal is to evaluate zero-shot cross-lingual transfer, i.e. training on English target-task data and evaluating the model on target-task data in different languages. Target tasks are described in Section 2.2.

2.1 Intermediate Tasks

We study the effect of intermediate-task training (STILTs; Phang et al., 2018) on zero-shot cross-lingual transfer into multiple target tasks and languages. We experiment with seven different intermediate tasks, all of them with English labeled data, as illustrated in Table 1.

**ANLI + MNLI + SNLI (ANLI+)** The Adversarial Natural Language Inference dataset (Nie et al., 2019) is collected adversarially using a human and model in the loop as an extensions of the Stanford Natural Language Inference (SNLI; Bowman et al.,...
| Name                  | Train | Dev   | Test          | Task                      | Genre/Source            |
|----------------------|-------|-------|---------------|---------------------------|-------------------------|
| Intermediate tasks   |       |       |               |                           |                         |
| ANLI+                | 1,104,934 | 22,857 | N/A          | natural language inference | Misc.                   |
| QQP                  | 363,846 | 40,430 | N/A          | paraphrase detection      | Quora questions         |
| SQuAD                | 87,599 | 34,726 | N/A          | span extraction           | Wikipedia                |
| HellaSwag            | 39,905 | 10,042 | N/A          | sentence completion       | Video captions & Wikihow |
| CCG                  | 38,015 | 5,484  | N/A          | tagging                   | Wall Street Journal      |
| Cosmos QA            | 25,588 | 3,000  | N/A          | question answering        | Blogs                    |
| CommonsenseQA        | 9,741 | 1,221  | N/A          | question answering        | Crowdsourced responses  |
|                       |        |        |              |                           |                         |
| Target tasks (XTREME Benchmark) |       |       |               |                           |                         |
| XNLI                 | 392,702 | 2,490  | 5,010        | natural language inference | Misc.                   |
| PAWS-X               | 49,401 | 2,000  | 2,000        | paraphrase detection      | Wiki/Quora              |
| POS                  | 21,253 | 3,974  | 47–20,436    | tagging                   | Misc.                   |
| NER                  | 20,000 | 10,000 | 1,000–10,000 | named entity recognition  | Wikipedia                |
| XQuAD                | 87,599 | 34,726 | 1,190        | question answering        | Wikipedia                |
| MLQA                 | 87,599 | 34,726 | 4,517–11,590 | question answering        | Wikipedia                |
| TyDiQA-GoldP         | 3,696 | 634    | 323–2,719    | question answering        | Wikipedia                |
| BUCC                 | –     | –      | 1,896–14,330 | sentence retrieval        | Wiki / news             |
| Tatoeba              | –     | –      | 1,000        | sentence retrieval        | Misc.                   |

Table 1: Overview of the intermediate tasks (top) and target tasks (bottom) in our experiments. EM is short for Exact Match. For target tasks, Train and Dev correspond to the English training and development sets, while Test shows the range of sizes for the target-language test sets for each task. XQuAD, TyDiQA and BUCC do not have separate held-out development sets.

2015) and the Multi-genre Natural Language Inference (MNLI; Williams et al., 2018) corpora. We follow Nie et al. (2019) and use the concatenated ANLI, MNLI and SNLI training sets, which we refer to as ANLI+.

CCG CCGbank (Hockenmaier and Steedman, 2007) is a translation of the Penn Treebank into Combinatory Categorial Grammar (CCG) derivations. The CCG supertagging task that we use consists of assigning lexical categories to individual word tokens which roughly determine a full parse.¹

CommonsenseQA CommonsenseQA (Talmor et al., 2019) is a multiple-choice QA dataset generated by crowdworkers based on clusters of concepts from ConceptNet (Speer et al., 2017).

Cosmos QA Cosmos QA is multiple-choice commonsense-based reading comprehension dataset (Huang et al., 2019b) generated by crowdworkers, with a focus on the causes and effects of events.

HellaSwag HellaSwag (Zellers et al., 2019) is a commonsense reasoning dataset framed as a four-way multiple choice task, where examples consist of a text and four choices of spans, one of which is a plausible continuation of the given scenario. It is built using adversarial filtering (Zellers et al., 2018; Bras et al., 2020) with BERT.

QQP The Quora Question Pairs² is a paraphrase detection dataset constructed from questions posted on Quora, a community question-answering website. Examples in the dataset consist of two questions, labeled for whether they are semantically equivalent.

SQuAD Stanford Question Answering Dataset (Rajpurkar et al., 2016) is a question-answering dataset consisting of passages extracted from Wikipedia articles and crowd-sourced questions and answers. Each example consists of a context passage and a question, and the answer is a text span from the context. We use SQuAD version 1.1.

Task Selection Criteria We choose these tasks based on both the diversity of task formats and evidence of positive transfer from literature. Pruskachtkun et al. (2020) shows that MNLI (which is a subset ANLI+), CommonsenseQA, Cosmos QA and HellaSwag are good candidates for intermediate tasks with positive transfer to a range of downstream tasks. CCG involves token-wise prediction and is similar to the POS and NER target tasks. SQuAD is a widely-used question-answering task, while QQP is semantically similar to sentence retrieval.²

¹If a word is tokenized into sub-word tokens, we use the representation of the first sub-word token for the tag prediction for that word.

²http://data.quora.com/First-Quora-DatasetRelease-Question-Pairs
retrieval target tasks (BUCC and Tatoeba) as well as PAWS-X, another paraphrase-detection task.

2.2 Target Tasks

We use the 9 target tasks from the XTREME benchmark (Hu et al., 2020), which span 40 different languages (hereafter referred to as the target languages): The Cross-lingual Question Answering (XQuAD; Artetxe et al., 2019) dataset, a cross-lingual extension of the SQuAD dataset (Rajpurkar et al., 2016); the Multilingual Question Answering (MLQA; Lewis et al., 2019) dataset; the Typologically Diverse Question Answering (TyDiQA-GoldP; Clark et al., 2020) dataset; the Cross-lingual Natural Language Inference dataset (XNLI; Conneau et al., 2018), a cross-lingual extension of MNLI (Williams et al., 2018); the Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X; Yang et al., 2019) dataset; the Universal Dependencies v2.5 (Nivre et al., 2018) POS tagging dataset; the Wikiann NER dataset (Pan et al., 2017); the BUCC dataset (Zweigenbaum et al., 2017, 2018), which involves identifying parallel sentences from corpora of different languages; and the Tatoeba (Artetxe and Schwenk, 2019) dataset, which involves aligning pairs of sentences.

Among the 9 tasks, BUCC and Tatoeba are sentence retrieval tasks do not include training sets, and are scored based on the similarity of learned representations (see Appendix). XQuAD, TyDiQA and Tatoeba do not include development sets separate from the test sets.\footnote{UDPOS also does not include development set data for a small number of languages: Kazakh, Thai, Tagalog and Yoruba.} For all XTREME tasks, we follow the training and evaluation protocol described in the benchmark (Hu et al., 2020) and their sample implementation.\footnote{https://github.com/google-research/xtreme} Intermediate and target task statistics are shown in Table 1.

2.3 Multilingual Masked Language Modeling

Our setup requires that we train the pretrained multilingual model extensively on English data before using it on a non-English target task, which can lead to the catastrophic forgetting of other languages acquired during pretraining. We investigate whether continuing to train on the multilingual MLM pretraining objective while fine-tuning on an English intermediate task can prevent catastrophic forgetting of the target languages and improve downstream transfer performance.

We construct a multilingual corpus across the 40 languages covered by the XTREME benchmark using Wikipedia dumps from April 14, 2020 for each language and the MLM data creation scripts from the jiant library (Wang et al., 2019c). In total, we use 2 million sentences sampled across all languages according to the sampling ratio of Conneau and Lample (2019) with the $\alpha = 0.3$.

3 Experiments and Results

3.1 Models

We use the pretrained XLM-R Large model (Conneau et al., 2019) as a starting point for all our experiments.\footnote{XLM-R (Conneau et al., 2019) is a 550m-parameter variant of the RoBERTa masked language model (Liu et al., 2019b) trained on a cleaned version of CommonCrawl on 100 languages.} Details on the intermediate and target task training procedures can be found in the Appendix.

XLM-R As our baseline, we directly fine-tune the pretrained XLM-R model on each target task’s English training data (if available) and evaluate zero-shot on non-English data, following the protocol for the XTREME benchmark (Hu et al., 2020).

XLM-R + Intermediate-Task Training We include an additional intermediate-task training phase, and first fine-tune the pretrained XLM-R on an English intermediate task before we train/evaluate on the target tasks as described above.

We also experiment with multi-task training on all available intermediate tasks but SQuAD.\footnote{Excluded in this draft due to implementation issues.} We follow Raffel et al. (2019a) and sample batches of examples for each task with probability $r_m = \min(e_m, K) / \sum_{m} \min(e_m, K)$, where $e_m$ is the number of examples in task $m$ and $K = 2^{17}$ is a size limit constant.

XLM-R + Intermediate-Task Training + MLM We consider a variant of intermediate-task training where the pretrained XLM-R model is trained on both an intermediate task as well as multilingual MLM simultaneously. We treat multilingual MLM as an additional task in the intermediate training phase, and use the same multi-task sampling strategy as above.
**3.2 Results**

**Intermediate-Task Training** As shown in Table 2, no single intermediate task yields positive transfer across all target tasks. The target tasks TyDiQA, BUCC and Tatoeba see consistent gains from most or all the intermediate tasks. In particular, BUCC and Tatoeba, the two sentence retrieval tasks with no training data, benefit universally from intermediate-task training. PAWS-X, XQuAD and MLQA also exhibit gains with the additional intermediate-training on some intermediate tasks. On the other hand, we find generally flat or negative transfer to XNLI, POS and NER target tasks.

Among the intermediate tasks, we find that SQuAD performs best; in addition to having positive transfer to BUCC and Tatoeba, it also leads to improved performance across the three question-answering target tasks. Additionally, CommonsenseQA and CosmosQA lead to slightly improved performance across the question-answering tasks.

ANLI\(^+\) does not lead to improved performance on XNLI as we first expected, despite the additional training data. This is consistent with Nie et al. (2019), who showed that adding ANLI data to SNLI and MNLI training sets did not improve performance on NLI benchmarks. QQP significantly improves sentence retrieval tasks performance but has broadly negative transfer to the remaining target tasks. CCG also has relatively poor transfer performance, consistent with Pruksachatkun et al. (2020).

One might find it surprising that intermediate-task training on SQuAD improves XQuAD and MLQA performance, which also use SQuAD as training data. We follow the sample implementation for target task training in the XTREME benchmark, which trains on SQuAD for only 2 epochs. This may explain why an additional phase of SQuAD training can improve performance.

**MLM and Multi-Task Training** Incorporating MLM during intermediate-task training shows no
clear trend. It reduces negative transfer, as seen in the cases of CommonsenseQA and QQP, but it also tends to reduce positive transfer, for instance as seen with intermediate-training on SQuAD. Both TyDiQA and PAWS-X see a small improvement from incorporating multilingual MLM training—in particular, every combination of intermediate tasks and MLM has strictly positive transfer to TyDiQA with an improvement of at least 2 points. The multi-task intermediate-training transfer results are on par with the average of the single intermediate-task transfer results, showing no clear benefit from using the additional tasks.

**XTREME Benchmark Results** At the bottom of Table 2, we show results obtained by XLM-R on the XTREME benchmark as reported by Hu et al. (2020), results obtained with our reimplementation of XLM-R (i.e. our baseline), and results obtained with our best models, which use intermediate-task training are selected according to development set performance on each target task. Scores on the XTREME benchmark are computed based on the respective target-task test sets where available, and based on development sets for target tasks without separate held-out test sets.7 We are generally able to replicate the best reported XLM-R baseline results, except for: Tatoeba, where our implementation significantly underperforms the reported scores in Hu et al. (2020); and TyDiQA, where our implementation outperforms the reported scores. On the other hand, our best models show gains in 8 out of the 9 XTREME tasks relative to both baseline implementations, attaining an average score of 74.2 across target tasks, a 6.1 point improvement over the previous best reported average score of 68.1.

### 4 Related work

Sequential transfer learning using pretrained Transformer-based encoders (Phang et al., 2018) has been shown to be effective for many text classification tasks. This setup generally involves fine-tuning on a single task (Pruksachatkun et al., 2020; Vu et al., 2020) or multiple tasks (Liu et al., 2019a; Wang et al., 2019b; Raffel et al., 2019b), occasionally referred to as the intermediate task(s), before fine-tuning on the target task. Our work builds upon this line of work, focusing on intermediate-task training for improving cross-lingual transfer.

Early work on cross-lingual transfer mostly relies on the availability of parallel transfer, where one can perform translation (Mayhew et al., 2017) or project annotations from one language into another (Hwa et al., 2005; Agić et al., 2016). When there is enough data in multiple languages with consistent annotations, training multilingual models becomes an option for cross-lingual transfer (Plank et al., 2016; Cotterell et al., 2017; Zeman et al., 2017).

For large-scale cross-lingual transfer, Johnson et al. (2017) train a multilingual neural machine translation system and perform zero-shot translation without explicit bridging between the source and target languages. Recent works on extending pretrained Transformer-based encoders to multilingual settings show that these models are effective for cross-lingual tasks and competitive with strong monolingual models on the XNLI benchmark (Devlin et al., 2019b; Conneau and Lample, 2019; Conneau et al., 2019; Huang et al., 2019a). More recently, Artetxe et al. (2020) showed that cross-lingual transfer performance can be sensitive to translation artifacts arising from a multilingual datasets’ creation procedure.

Finally, Pfeiffer et al. (2020) propose adapter modules that learn language and task representation for cross-lingual transfer, which also allow adaptation to languages that are not observed in the pretraining data.

### 5 Conclusion and Future work

In this paper, we conduct a large-scale study on the impact intermediate-task training has on zero-shot cross-lingual transfer. We evaluate 7 intermediate tasks and investigate how intermediate-task training impacts the zero-shot cross-lingual transfer to the 9 target tasks in the XTREME benchmark.

Overall, intermediate-task training significantly improves the performance on BUCC and Tatoeba, the two sentence retrieval target tasks in the XTREME benchmark, across almost every intermediate-task configuration. Our best models obtain 5.9 and 25.5 point gains on BUCC and Tatoeba, respectively, in the leaderboard when compared to the best available XLM-R baseline scores (Hu et al., 2020). We also observed gains in question-answering tasks, particularly using SQuAD as an intermediate task, with absolute gains of 0.6 F1 for XQuAD, 0.7 F1 for MLQA, and 5.7 for F1 TyDiQA, again over the best available baseline scores. We improve over XLM-R by...
6.1 points on average on the XTREME benchmark, setting a new state of the art. However, we found that neither incorporating multilingual MLM into the intermediate-task training phase nor multi-task training led to improved transfer performance. While we have explored the extent to which English intermediate-task training can improve cross-lingual transfer, an obvious next avenue of investigation for future work is whether multilingual or target-language intermediate-task training can further bolster performance, and how the relationship between the intermediate- and target-task languages influences transfer.

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., by NVIDIA Corporation (with the donation of a Titan V GPU), by Google (with the donation of Google Cloud credits). IC has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 838188. This project has benefited from direct support by the NYU IT High Performance Computing Center.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. Transactions of the Association for Computational Linguistics, 4:301–312.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation Artifacts in Cross-lingual Transfer Learning. arXiv e-prints, page arXiv:2004.04721.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. arXiv preprint arXiv:2002.04108.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32 (NeurIPS), pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, pages 1–30, Vancouver. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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
Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Hrut, 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.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. 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, 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.

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, 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, Ian 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. 2019b. 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, 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 Hrut, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohanayan, 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.

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.

Yinfeng Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. arXiv preprint arXiv:1901.11373.

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
A Implementation details

A.1 Intermediate Tasks

For intermediate-task training, we use a learning rate of 1e-5 without MLM, and 5e-6 with MLM. Hyperparameters in the Table 3 were chosen based on an initial search based on intermediate task validation performance. We use a warmup of 10% of the total number of steps, and perform early stopping based on the first 500 development set examples of each task. For CCG, where tags are assigned for each word, we use the representation of first sub-word token of each word for prediction.

| Task       | Batch size | # Epochs |
|------------|------------|----------|
| ANLI       | 24         | 2        |
| CCG        | 24         | 15       |
| CommonsenseQA | 4         | 10       |
| Cosmos QA  | 4          | 15       |
| HellaSwag  | 24         | 7        |
| QQP        | 24         | 3        |
| SQuAD      | 8          | 3        |
| MLM        | 8          | -        |
| Multi-task | Mixed      | 3        |

Table 3: Intermediate-task training configuration.

A.2 Target Tasks / XTREME Benchmark

We follow the sample implementation for the XTREME benchmark unless otherwise stated. We use a learning rate of 3e-6, and use the same optimization procedure as for intermediate tasks. Hyperparameters in the Table 4 follow the sample implementation. For POS and NER, we use the same strategy as for CCG for matching tags to tokens. For BUCC and Tatoeba, we extract the representations for each token from the 13th self-attention layer, and use the mean-pooled representation as the embedding for that example, as in the sample implementation. Similarly, we follow the sample implementation and set an optimal threshold for each language sub-task for BUCC as a similarity score cut-off for extracting parallel sentences based on the development set and applied to the test set.

We randomly initialize the corresponding output heads for each task, regardless of the similarity between intermediate and target tasks (e.g. even if both the intermediate and target tasks train on SQuAD, we randomly initialize the output head in between phases).
| Task             | Batch size | # Epochs |
|------------------|------------|----------|
| XNLI (MNLI)      | 4          | 2        |
| PAWS-X           | 32         | 5        |
| XQuAD (SQuAD)    | 16         | 2        |
| MLQA (SQuAD)     | 16         | 2        |
| TyDiQA           | 16         | 2        |
| POS              | 32         | 10       |
| NER              | 32         | 10       |
| BUCC             | -          | -        |
| Tatoeba          | -          | -        |

Table 4: Target-task training configuration.

### A.3 Software

Experiments are run using the NLP Runners\(^8\) library, based on PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2019).

\(^8\)https://github.com/zphang/nlprunners