Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers

Anne Lauscher[♣] **Olga Majewska**[♠] **Leonardo F. R. Ribeiro**[♢]
Iryna Gurevych[♢] **Nikolai Rozanov**[♠] **Goran Glavaš**[♣]

[♣]Data and Web Science Group, University of Mannheim, Germany
[♠]Wluper, London, United Kingdom
[♢]Ubiquitous Knowledge Processing (UKP) Lab, TU Darmstadt, Germany
{anne,goran}@informatik.uni-mannheim.de
{olga,nikolai}@wluper.com
www.ukp.tu-darmstadt.de

Abstract

Following the major success of neural language models (LMs) such as BERT or GPT-2 on a variety of language understanding tasks, recent work focused on injecting (structured) knowledge from external resources into these models. While on the one hand, joint pretraining (i.e., training from scratch, adding objectives based on external knowledge to the primary LM objective) may be prohibitively computationally expensive, post-hoc fine-tuning on external knowledge, on the other hand, may lead to the catastrophic forgetting of distributional knowledge. In this work, we investigate models for complementing the distributional knowledge of BERT with conceptual knowledge from ConceptNet and its corresponding Open Mind Common Sense (OMCS) corpus, respectively, using *adapter training*. While overall results on the GLUE benchmark paint an inconclusive picture, a deeper analysis reveals that our adapter-based models substantially outperform BERT (up to 15-20 performance points) on inference tasks that require the type of conceptual knowledge explicitly present in ConceptNet and OMCS. We also open source all our experiments and relevant code under: https://github.com/wluper/retrograph.

1 Introduction

Self-supervised neural models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019; Liu et al., 2019b), GPT (Radford et al., 2018, 2019), or XL-Net (Yang et al., 2019) have rendered language modeling a very suitable pretraining task for learning language representations that are useful for a wide range of language understanding tasks (Wang et al., 2018, 2019). Although shown versatile w.r.t. the types of knowledge (Rogers et al., 2020) they encode, much like their predecessors – static word embedding models (Mikolov et al., 2013; Pennington et al., 2014) – neural LMs still only "consume" the distributional information from large corpora. Yet, a number of structured knowledge sources exist – knowledge bases (KBs) (Suchanek et al., 2007; Auer et al., 2007) and lexico-semantic networks (Miller, 1995; Liu and Singh, 2004; Navigli and Ponzetto, 2010) – encoding many types of knowledge that are underrepresented in text corpora.

Starting from this observation, most recent efforts focused on injecting factual (Zhang et al., 2019; Liu et al., 2019a; Peters et al., 2019) and linguistic knowledge (Lauscher et al., 2019; Peters et al., 2019) into pretrained LMs and demonstrated the usefulness of such knowledge in language understanding tasks (Wang et al., 2018, 2019). *Joint pretraining models*, on the one hand, augment distributional LM objectives with additional objectives based on external resources (Yu and Dredze, 2014; Nguyen et al., 2016; Lauscher et al., 2019) and train the extended model from scratch. For models like BERT, this implies computationally expensive retraining from scratch of the encoding transformer network. *Post-hoc fine-tuning* models (Zhang et al., 2019; Liu et al., 2019a; Peters et al., 2019), on the other hand, use the objectives based on external resources to fine-tune the encoder's parameters, pretrained via distributional LM objectives. If the amount of fine-tuning data is substantial, however, this approach may lead to catastrophic forgetting of distributional knowledge obtained in pretraining (Goodfellow et al., 2014; Kirkpatrick et al., 2017).

In this work, similar to the concurrent work of Wang et al. (2020), we turn to the recently proposed *adapter-based fine-tuning* paradigm (Rebuffi et al., 2018; Houlsby et al., 2019), which remedies the shortcomings of both joint pretraining and standard post-hoc fine-tuning. Adapter-based training injects additional parameters into the encoder and only tunes their values: original transformer parameters are kept fixed. Be-

Proceedings of Deep Learning Inside Out (DeeLIO):
The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 43–49

cause of this, adapter training preserves the distributional information obtained in LM pretraining, without the need for any distributional (re-)training. While (Wang et al., 2020) inject factual knowledge from Wikidata (Vrandečić and Krötzsch, 2014) into BERT, in this work, we investigate two resources that are commonly assumed to contain *general-purpose* and *common sense* knowledge:[1] ConceptNet (Liu and Singh, 2004; Speer et al., 2017) and the Open Mind Common Sense (OMCS) corpus (Singh et al., 2002), from which the ConceptNet graph was (semi-)automatically extracted. For our first model, dubbed CN-ADAPT, we first create a synthetic corpus by randomly traversing the ConceptNet graph and then learn adapter parameters with masked language modelling (MLM) training (Devlin et al., 2019) on that synthetic corpus. For our second model, named OM-ADAPT, we learn the adapter parameters via MLM training directly on the OMCS corpus.

We evaluate both models on the GLUE benchmark, where we observe limited improvements over BERT on a subset of GLUE tasks. However, a more detailed inspection reveals large improvements over the base BERT model (up to 20 Matthews correlation points) on language inference (NLI) subsets labeled as requiring World Knowledge or knowledge about Named Entities. Investigating further, we relate this result to the fact that ConceptNet and OMCS contain much more of what in downstream is considered to be factual world knowledge than what is judged as common sense knowledge. Our findings pinpoint the need for more detailed analyses of compatibility between (1) the types of knowledge contained by external resources; and (2) the types of knowledge that benefit concrete downstream tasks; within the emerging body of work on injecting knowledge into pretrained transformers.

2 Knowledge Injection Models

In this work, we are primarily set to investigate if injecting specific types of knowledge (given in the external resource) benefits downstream inference that clearly requires those exact types of knowledge. Because of this, we use the arguably most straightforward mechanisms for injecting the ConceptNet and OMCS information into BERT, and leave the exploration of potentially more effective knowledge injection objectives for future work. We inject the external information into adapter parameters of the adapter-augmented BERT (Houlsby et al., 2019) via BERT's natural objective – masked language modelling (MLM). OMCS, already a corpus in natural language, is directly subjectable to MLM training – we filtered out non-English sentences. To subject ConceptNet to MLM training, we need to transform it into a synthetic corpus.

Unwrapping ConceptNet. Following established previous work (Perozzi et al., 2014; Ristoski and Paulheim, 2016), we induce a synthetic corpus from ConceptNet by randomly traversing its graph. We convert relation strings into NL phrases (e.g., `synonyms` to *is a synonym of*) and duplicate the object node of a triple, using it as the subject for the next sentence. For example, from the path "*alcoholism* $\xrightarrow{causes}$ *stigma* $\xrightarrow{hasContext}$ *christianity* $\xrightarrow{partOf}$ *religion*" we create the text "*alcoholism causes stigma. stigma is used in the context of christianity. christianity is part of religion.*". We set the walk lengths to 30 relations and sample the starting and neighboring nodes from uniform distributions. In total, we performed 2,268,485 walks, resulting with the corpus of 34,560,307 synthetic sentences.

Adapter-Based Training. We follow Houlsby et al. (2019) and adopt the adapter-based architecture for which they report solid performance across the board. We inject *bottleneck adapters* into BERT's transformer layers. In each transformer layer, we insert two bottleneck adapters: one after the multi-head attention sub-layer and another after the feed-forward sub-layer. Let $\mathbf{X} \in \mathbb{R}^{T \times H}$ be the sequence of contextualized vectors (of size H) for the input of T tokens in some transformer layer, input to a bottleneck adapter. The bottleneck adapter, consisting of two feed-forward layers and a residual connection, yields the following output:

$$Adapter(\mathbf{X}) = \mathbf{X} + f\left(\mathbf{X}\mathbf{W}_d + \mathbf{b}_d\right)\mathbf{W}_u + \mathbf{b}_u$$

where $\mathbf{W}_d$ (with bias $\mathbf{b}_d$) and $\mathbf{W}_u$ (with bias $\mathbf{b}_u$) are adapter's parameters, that is, the weights of the linear down-projection and up-projection sub-layers and f is the non-linear activation function. Matrix $\mathbf{W}_d \in \mathbb{R}^{H \times m}$ compresses vectors in $\mathbf{X}$ to the *adapter size* $m < H$, and the matrix $\mathbf{W}_u \in \mathbb{R}^{m \times H}$ projects the activated down-projections back to transformer's hidden size H. The ratio H/m determines how many times fewer

[1] Our results in §3.2 scrutinize this assumption.

parameters we optimize with adapter-based training compared to standard fine-tuning of all transformer's parameters.

3 Evaluation

We first briefly describe the downstream tasks and training details, and then proceed with the discussion of results obtained with our adapter models.

3.1 Experimental Setup.

Downstream Tasks. We evaluate BERT and our two adapter-based models, CN-ADAPT and OM-ADAPT, with injected knowledge from ConceptNet and OMCS, respectively, on the tasks from the GLUE benchmark (Wang et al., 2018):

CoLA (Warstadt et al., 2018): Binary sentence classification, predicting grammatical acceptability of sentences from linguistic publications;

SST-2 (Socher et al., 2013): Binary sentence classification, predicting binary sentiment (positive or negative) for movie review sentences;

MRPC (Dolan and Brockett, 2005): Binary sentence-pair classification, recognizing sentences which are are mutual paraphrases;

STS-B (Cer et al., 2017): Sentence-pair regression task, predicting the degree of semantic similarity for a given pair of sentences;

QQP (Chen et al., 2018): Binary classification task, recognizing question paraphrases;

MNLI (Williams et al., 2018): Ternary natural language inference (NLI) classification of sentence pairs. Two test sets are given: a matched version (MNLI-m) in which the test domains match the domains from training data, and a mismatched version (MNLI-mm) with different test domains;

QNLI: A binary classification version of the Stanford Q&A dataset (Rajpurkar et al., 2016);

RTE (Bentivogli et al., 2009): Another NLI dataset, ternary entailment classification for sentence pairs;

Diag (Wang et al., 2018): A manually curated NLI dataset, with examples labeled with specific types of knowledge needed for entailment decisions.

Training Details. We inject our adapters into a BERT Base model (12 transformer layers with 12 attention heads each; $H = 768$) pretrained on lowercased corpora. Following (Houlsby et al., 2019), we set the size of all adapters to $m = 64$ and use GELU (Hendrycks and Gimpel, 2016) as the adapter activation f. We train the adapter parameters with the Adam algorithm (Kingma and Ba, 2015) (initial learning rate set to $1e^{-4}$, with 10000 warm-up steps and the weight decay factor of 0.01). In downstream fine-tuning, we train in batches of size 16 and limit the input sequences to $T = 128$ wordpiece tokens. For each task, we find the optimal hyperparameter configuration from the following grid: learning rate $l \in \{2 \cdot 10^{-5}, 3 \cdot 10^{-5}\}$, epochs in $n \in \{3, 4\}$.

3.2 Results and Analysis

GLUE Results. Table 1 reveals the performance of CN-ADAPT and OM-ADAPT in comparison with BERT Base on GLUE evaluation tasks. We show the results for two snapshots of OM-ADAPT, after 25K and 100K update steps, and for two snapshots of CN-ADAPT, after 50K and 100K steps of adapter training. Overall, none of our adapter-based models with injected external knowledge from ConceptNet or OMCS yields significant improvements over BERT Base on GLUE. However, we observe substantial improvements (of around 3 points) on RTE and on the Diagnostics NLI dataset (Diag), which encompasses inference instances that require a specific type of knowledge.

Since our adapter models draw specifically on the conceptual knowledge encoded in ConceptNet and OMCS, we expect the positive impact of injected external knowledge – assuming effective injection – to be most observable on test instances that target the same types of conceptual knowledge. To investigate this further, we measure the model performance across different categories of the Diagnostic NLI dataset. This allows us to tease apart inference instances which truly test the efficacy of our knowledge injection methods. We show the results obtained on different categories of the Diagnostic NLI dataset in Table 2. The improvements of our adapter-based models over BERT Base on these phenomenon-specific subsections of the Diagnostics NLI dataset are generally much more pronounced: e.g., OM-ADAPT (25K) yields a 7% improvement on inference that requires factual or common sense knowledge (KNO), whereas CN-ADAPT (100K) yields a 6% boost for inference that depends on lexico-semantic knowledge (LS). These results suggest that (1) ConceptNet and OMCS do contain the specific types of knowledge required for these inference categories and that (2) we managed to inject that knowledge into BERT by training

Model	CoLA MCC	SST-2 Acc	MRPC F1	STS-B Spear	QQP F1	MNLI-m Acc	MNLI-mm Acc	QNLI Acc	RTE Acc	Diag MCC	Avg –
BERT Base	52.1	93.5	**88.9**	85.8	71.2	**84.6**	83.4	90.5	66.4	34.2	75.1
OM-ADAPT (25K)	49.5	93.5	88.8	85.1	71.4	84.4	83.5	**90.9**	67.5	35.7	75.0
OM-ADAPT (100K)	**53.5**	93.4	87.9	**85.9**	71.1	84.2	**83.7**	90.6	68.2	34.8	75.3
CN-ADAPT (50K)	49.8	**93.9**	**88.9**	85.8	**71.6**	84.2	83.3	90.6	**69.7**	37.0	**75.5**
CN-ADAPT (100K)	48.8	92.8	87.1	85.7	71.5	83.9	83.2	90.8	64.1	**37.8**	74.6

Table 1: Results on test portions of GLUE benchmark tasks. Numbers in brackets next to adapter-based models (25K, 50K, 100K) indicate the number of update steps of adapter training on the synthetic ConceptNet corpus (for CN-ADAPT) or on the original OMCS corpus (for OM-ADAPT). **Bold**: the best score in each column.

Model	LS	KNO	LOG	PAS	All
BERT Base	38.5	20.2	26.7	39.6	34.2
OM-ADAPT (25K)	39.1	**27.1**	26.1	39.5	35.7
OM-ADAPT (100K)	37.5	21.2	27.4	41.0	34.8
CN-ADAPT (50K)	40.2	24.3	30.1	**42.7**	37.0
CN-ADAPT (100K)	**44.2**	25.2	**30.4**	41.9	37.8

Table 2: Breakdown of Diagnostics NLI performance (Matthews correlation), according to information type needed for inference (coarse-grained categories): Lexical Semantics (LS), Knowledge (KNO), Logic (LOG), and Predicate Argument Structure (PAS).

Model	CS	World	NE
BERT Base	**29.0**	10.3	15.1
OM-ADAPT (25K)	28.5	25.3	31.4
OM-ADAPT (100K)	24.5	17.3	22.3
CN-ADAPT (50K)	25.6	21.1	26.0
CN-ADAPT (100K)	24.4	**25.6**	**36.5**

Table 3: Results (Matthews correlation) on Common Sense (CS), World Knowledge (World), and Named Entities (NE) categories of the Diagnostic NLI dataset.

adapters on these resources.

Fine-Grained Knowledge Type Analysis. In our final analysis, we "zoom in" our models' performances on three fine-grained categories of the Diagnostics NLI dataset – inference instances that require Common Sense Knowledge (CS), World Knowledge (World), and knowledge about Named Entities (NE), respectively. The results for these fine-grained categories are given in Table 3. These results show an interesting pattern: our adapter-based knowledge-injection models massively outperform BERT Base (up to 15 and 21 MCC points, respectively) for NLI instances labeled as requiring World Knowledge or knowledge about Named Entities. In contrast, we see drops in performance on instances labeled as requiring common sense knowledge. This initially came as a surprise, given the common belief that OMCS and ConcepNet contain the so-called *common sense* knowledge. Manual scrutiny of the diagnostic test instances from both CS and World categories uncovers a noticeable mismatch between the kind of information that is considered common sense in KBs like ConceptNet and what is considered common sense knowledge in the downstream. In fact, the majority of information present in ConceptNet and OMCS falls under the World Knowledge definition of the Diagnostic NLI dataset, including factual geographic information (`stockholm [partOf] sweden`), domain knowledge (`roadster [isA] car`) and specialized terminology (`indigenous [synonymOf] aboriginal`).

In contrast, many of the CS inference instances require complex, high-level reasoning, understanding metaphorical and idiomatic meaning, and making far-reaching connections. We display NLI Diagnostics examples from the World Knowledge and Common Sense categories in Table 4. In such cases, explicit conceptual links often do not suffice for a correct inference and much of the required knowledge is not explicitly encoded in the external resources. Consider, e.g., the following CS NLI instance: [`premise:` *My jokes fully reveal my character* `;` `hypothesis:` *If everyone believed my jokes, they'd know exactly who I was* `;` `entailment`]. While ConceptNet and OMCS may associate *character* with *personality* or *personality* with *identity*, the knowledge that the phrase *who I was* may refer to *identity* is beyond the explicit knowledge present in these resources. This sheds light on the results in Table 3: when the knowledge required to tackle the inference problem at hand is available in the external resource, our adapter-based knowledge-injected models significantly outperform the baseline transformer; otherwise, the benefits of knowledge injection are neg-

Knowledge	Premise	Hypothesis	ConceptNet?
World	*The sides came to an agreement after their meeting in **Stockholm**.*	*The sides came to an agreement after their meeting in **Sweden**.*	`stockholm [partOf] sweden`
	*Musk decided to offer up his personal Tesla **roadster**.*	*Musk decided to offer up his personal **car**.*	`roadster [isA] car`
	*The Sydney area has been inhabited by **indigenous** Australians for at least 30,000 years.*	*The Sydney area has been inhabited by **Aboriginal** people for at least 30,000 years.*	`indigenous [synonymOf] aboriginal`
Common Sense	*My jokes fully reveal my character.*	*If everyone believed my jokes, they'd know exactly who I was.*	
	The systems thus produced are incremental: dialogues are processed word-by-word, shown previously to be essential in supporting natural, spontaneous dialogue.	*The systems thus produced support the capability to interrupt an interlocutor mid-sentence.*	
	He deceitfully proclaimed: "This is all I ever really wanted."	*He was satisfied.*	

Table 4: Premise-hypothesis examples from the diagnostic NLI dataset tagged for commonsense and world knowledge, and relevant ConceptNet relations, where available.

ligible or non-existent. The promising results on *world knowledge* and *named entities* portions of the Diagnostics dataset suggest that our methods does successfully inject external information into the pretrained transformer and that the presence of the required knowledge for the task in the external resources is an obvious prerequisite.

4 Conclusion

We presented two simple strategies for injecting external knowledge from ConceptNet and OMCS corpus, respectively, into BERT via bottleneck adapters. Additional adapter parameters store the external knowledge and allow for the preservation of the rich distributional knowledge acquired in BERT's pretraining in the original transformer parameters. We demonstrated the effectiveness of these models in language understanding settings that require precisely the type of knowledge that one finds in ConceptNet and OMCS, in which our adapter-based models outperform BERT by up to 20 performance points. Our findings stress the importance of having detailed analyses that compare (a) the types of knowledge found in external resources being injected against (b) the types of knowledge that a concrete downstream reasoning tasks requires. We hope this work motivates further research effort in the direction of fine-grained knowledge typing, both of explicit knowledge in external resources and the implicit knowledge stored in pretrained transformers.

Acknowledgments

Anne Lauscher and Goran Glavaš are supported by the Eliteprogramm of the Baden-Württemberg Stiftung (AGREE grant). Leonardo F. R. Ribeiro has been supported by the German Research Foundation as part of the Research Training Group AIPHES under the grant No. GRK 1994/1. This work has been supported by the German Research Foundation within the project "Open Argument Mining" (GU 798/25-1), associated with the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). The work of Olga Majewska was conducted under the research lab of Wluper Ltd. (UK/ 10195181).

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

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.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ian J Goodfellow, Mehdi Mirza, Aaron Courville Da Xiao, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *In Proceedings of International Conference on Learning Representations (ICLR*. Citeseer.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*.

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. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of ACL*, pages 454–459.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 701–710, New York, NY, USA. Association for Computing Machinery.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word

representations. 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 43–54.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *CVPR*.

Petar Ristoski and Heiko Paulheim. 2016. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

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.

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*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Blacbox NLP Workshop*, pages 353–355.

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

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.